THESE LINEAR ALGEBRA NOTES ARE POISONOUS

TO LEARN FROM THIS IS TO MUDDLE YOUR VIEW OF LINEAR ALGEBRA FOREVER! ABANDON HOPE ALL YE WHO READ THESE!

BY YILONG YANG

A SLAVE TO HIS WIFE AND TWO KIDS, ITS MATH AND MADNESS! MADNESS, I TELL YOU!

Picture above by Richard Schwartz

PUBLISHER 404 NOT FOUND
PUBLISHED IN THE WILD

Contents

0.1 A philosophical discourse on the philosophy of math and learning

Math is not an object-oriented subject. Rather, it is more relation-oriented. When we learn that $1 + 1 = 2$, the symbols 1 and 2 here do not refer to some genuin objects somewhere in the universe. Rather, the equation is suppose to points out a universal relation about quantities in this world. If you have one apple and another apple, then you have two apples. You can easily find infinite examplifications of such a relation. As math students, we see this quantity relation, and we write down the equation $1 + 1 = 2$ to capture this relation.

Similarly, linear algebra uses matrices. But what are matrices? They are not actual objects. They are relations of some kind. The object-oriented view versus relation-oriented view will lead to very different teaching practice and learning practice.

Suppose I am teaching my children (2 year old and 3 year old) the concept of cars. From an object-oriented view, I would do the following: I show them a million pictures, and say "This is a car" or "This is not a car". Then after enough practices, they are taught the concept of cars. Perhaps your education before college is exactly like this. Do the same problem enough times, your exam scores will then improve, right?

However, from a relational point of view, I would do things differently. I would say the following: "A car has a steering wheel, and you can turn it." "A car has a gas paddle to run faster, and a break to go slower." "Yesterday we went to the park by a car." "This red car is our car, and that blue car is our neighbor's car." "The policeman use a police car to catch bad guys." "Let us sing a song about cars." "Let us play with this toy car." I bring out the relation of cars with everything else in my children's life. Then with enough connections, their knowledge of car is now solidified in the web of its relations with regards to everything else. In particular, they now understand the concept of cars, even though they have not seen millions of pictures of cars.

Now suppose a person A learned about cars in the object-oriented way. And a person B learned about cars in the relation-oriented way. If I give them exams ("which of the following picture is a car?"), both should perform adequately well. Maybe A would perform even better, because of the extra practices. However, which one is more likely to enjoy cars when they grow up? Which one is going to have a better intuition about car mechanics? Which one is more likely to be a better driver? I would guess that B is going to be better in these senses.

Think about learning English. You can memorize words and grammar rules to your heart's content, and you might still freeze when you try to speak English out loud. Instead, if you go live in an English environment, then you'll be able to speak English easily in no time, because now you have connected English with various aspects of your life.

So I hope dearly that in your own studies, try to focus NOT on the WHAT, but on the HOW; NOT on the concepts, but on the CONNECTIONS between concepts. Everyone can search for the definition of some concepts online, but only YOU can connect the concept with YOUR own prior knowledge, in ways comfortable to YOU. These web of relations is what you learn.

Mere recitations will make you STOP learning, while making connections is how you START learning.

Chapter 1

Introduction

1.1 What is Linear Algebra

1.1.1 Linear Combinations

In my personal view, the study of Linear Algebra is essentially the study of ***linear combinations***. But what is this? Is this something we can eat? Is it something we wear to keep us warm during winter?

Succintly put, I think a linear combination is a phenomenon in our daily lives. Think about the following examples:

Example 1.1.1.

1. For breakfast I like to eat apples, bananas and cherries. Today I ate 2 apples, 3 bananas and 4 cherries. Then I can say that I ate a linear combination of apples, bananas and cherries, with coefficients $2, 3, 4$. I can also write that I ate

$$(2 \times \text{apples} + 3 \times \text{bananas} + 4 \times \text{cherries}).$$

2. Each of your class must use linear algebra at least once, to calculate your grade for the class. Your grade in a linear algebra class might be determined as the following: homework takes 20 percent, midterm takes 30 percent, and the final exam takes 50 percent. Then I can say that your grade is a linear combination of your homework, midterm and final with coefficients $0.2, 0.3, 0.5$. I can also write that your final grade is

$$(0.2 \times \text{Homework} + 0.3 \times \text{Midterm} + 0.5 \times \text{Final}).$$

3. Say I traveled to the USA. At the airport, I need to change yuans into dollars. Say I exchanged 12 yuan for 2 dollar. Then what's the change in my wallet? Well, it is a linear combination of yuans and dollars, with coefficients -12 and 2. I can also write that the change in my wallet is

$$(-12 \times \text{yuan} + 2 \times \text{dollar}).$$

4. My household is made of myself, my wife, two sons and no pet. You can say that my household is a linear combination of me, wife, sons and pets, with coefficients $1, 1, 2, 0$ respectively.

   You can also say that my household is a linear combination of me, wife and sons, which is also correct. However, it would be WRONG to say that my house hold is a linear combination of me, sons and pets, because no matter how you combine them, you will always miss out my wife, which is an important part of my household. (THE most important part. My wife might read this.)

☺

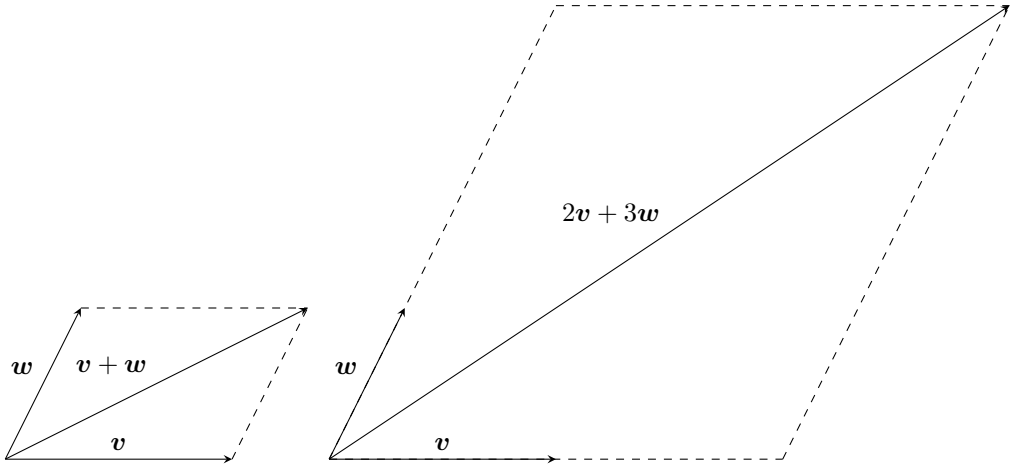
Remark 1.1.2. *As you can see from the examples above, the coefficients (or* ***scalars****, since they "scale" the corresponding item to some multiples of them.) here can be integers, non-integers, positive numbers, negative numbers, and even zero. All real numbers are good.*

Later we shall also introduce complex coefficients. But for the most part of this class, you can stay comfortably in the world of real numbers.

Here are some more abstract, mathematical structures, on which we can also do linear combinations.

Example 1.1.3.

1. In high school, we define "vectors" as some sort of arrows. We say that it is something with direction and magnitude. For example, if $\boldsymbol{v}, \boldsymbol{w}$ are two arrows, then we can perform "arrow additions" like $\boldsymbol{v}+\boldsymbol{w}$, or even linear combinations like $2\boldsymbol{v} + 3\boldsymbol{w}$, by drawing the corresponding parallelogram and find the diagonal arrow.



2. Given two functions $f(x), g(x)$, defined from real numbers to real numbers, we can do their linear combinations like $2f(x) + 3g(x)$, which will still be a function.

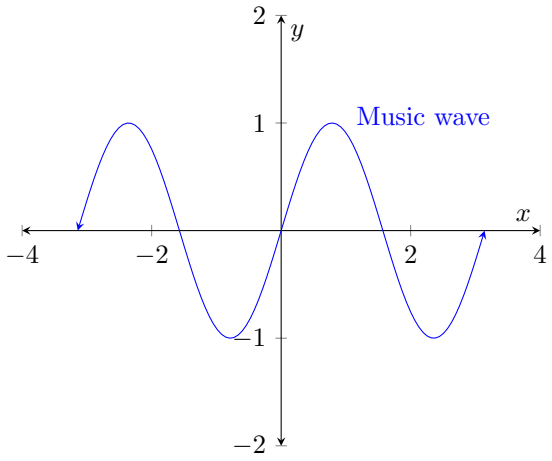
☺

Now, as long as we are doing linear combinations, here is a fun problem to comtemplate. It will also serve as a major indication of how one should learn mathematics in general.

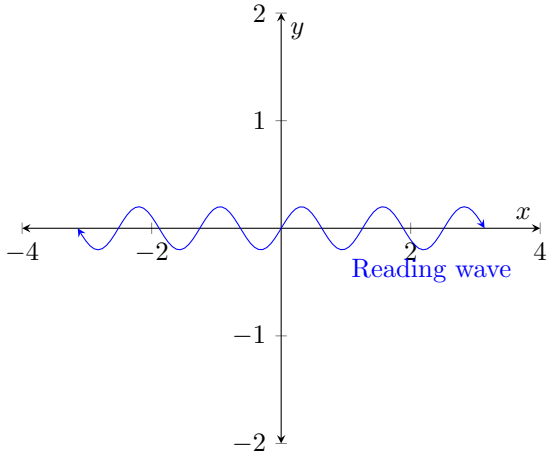
Example 1.1.4 (Grinfeld's Question)**.** It is crutial how we interpred our linear combinations. The following question is raised by Pavel Grinfeld.

We all have songs stored in our computer or ipod or smart phones or any music players. If you look into a music file, you see a bunch of numbers, the data of your file, recording the soundtracks.

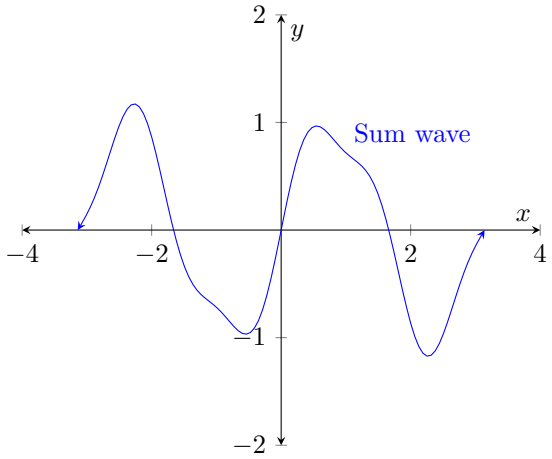
Say you open a soft music file, and for simplicity, suppose the sound track is like the following:



Say you open a file of someone reading a book, and for simplicity, suppose the sound track is like the following:



If you simply add of the two sound tracks, "book+music" will be like the following:



What does this sum sound like? Well, actually this is very straight forward: you will simultaneously hear both. You will hear the reading of the book, and with music in the background. Simple yes?

Now think about this: what if we do substraction "book-music"? What does it sound like now? Is the music played in the reversed order? Maybe the high pitches and low pitches are inverted? Or is it going to be gibberish?

Think about this for a while and then check the answer on the next page. ☺

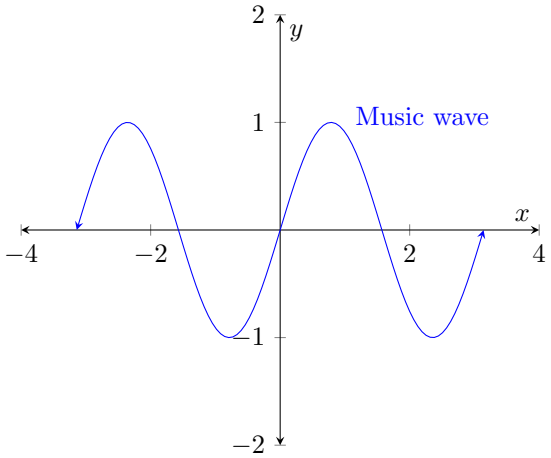
Here I leave a page break, so you can think about it a bit without looking at the answer on the next page.

Example 1.1.5 (Grinfeld's Question Answer)**.** It turns out that "book-music" sounds exactly the same as "book+music"! Surprised?

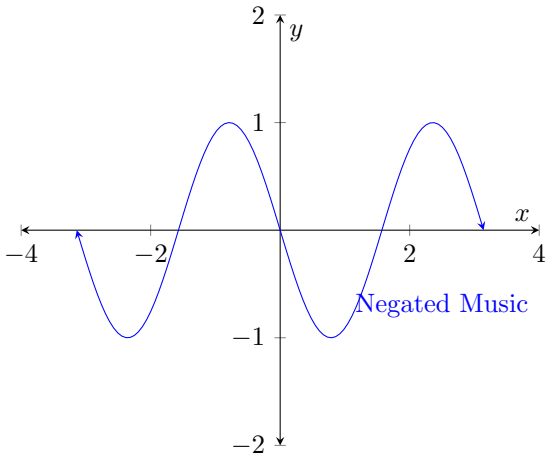
The key lies in interpretation. For simplicity, let us consider a simple sound wave $A\sin(\frac{2\pi t}{f})$. Here A is a real number that means amplitude, and it records how high the "belly" of your waves are. And f is an integer that means frequence, i.e, how many periods has we gone through when t goes from 0 to 1.

Given a sound wave, the amplitude would determine how loud we hear, while the frequence would determine the pitch of the sound. For example, a pitch may vibrate 440 cycles per second, or 440Hz in short, then it sounds like the standard "La" sound in most modern musics. And the amplitude simply determines how loud the sound is.

Now suppose we have a sound wave "music" which is the following:



Then the "minus music" is actually the following:



Wait! Even though the waves are not exactly the same, nevertheless they have the SAME amplitude and SAME frequency! So "-music" sounds exactly the same as "music", as it has the same pitch and same loudness. (To be more precise, the negated sound wave is in fact the original sound wave shifted by half a period, so they sounds exactly the same in our ears.)

In particular, "book+music" and "bood-music" will sound exactly the same. ☺

Remark 1.1.6. *Note that if "book"="music" are exactly the same, then book* $-$ *music* $= 0$ *while book* $+$ *music* $= 2$*book, hey, surely they must sound different now! However, by Fourier analysis, such a thing could*

only happen if the book soundwaves and the music soundwaves coincide completely on some short interval. That is almost never the case in real life.

Also be careful here. I am NOT claiming that the soundwave for "book+music" and "bood-music" are exactly the same. I am merely saying that they sounds the same to our ears. $\sin(x)$ *and* $\sin(x+1)$ *are different curves, yet they have the same frequency and amplitude, so our ear cannot tell the difference.*

The point of the above example is the following: simply knowing the math will not help you at all. What WILL help you is the interpretation of the math. For every mathematical concepts, try to gather at least one, and hopefully many interpretations. It could be interpretations in terms of applications, or in terms of other mathematical concepts, or it could just be some mental picture of sorts. The more the merrier. These interpretations will give you intuitions, and thus tell you where to go in your own problem solving endeavors.

When you look at a problem and you go blank, having no clue what to do, then it simply means you have not gathered enough intuitions. Go explore the definitions some more, play with more examples, talk to other people. Keep finding more interpretations, until you have enough intuition to guide you.

And while you are gathering these intuitions, keep in mind that they could be wrong. Anything that is not a rigorous proof could very well be wrong. But it does not mean your old interpretations are useless. Build intuitions, find the mistakes in your intuitions, and revise and improve your intuitions. This is simply how learning should be done.

1.1.2 Linear Maps

So, intuitively at least, linear combination simply means we are combining things by multiplying each with a number, and add them together. As you can see, I'm teaching you nothing new so far. I'm merely pointing out a structure with which you are already familiar, since they are everywhere in your life.

But what's the point of having structures? Well, structures are meant to be related to each other. Consider the following examples:

Example 1.1.7.

1. Say I go shopping and buy apples, bananas and cherries, and the prices for each is 3 yuan, 1 yuan and 2 yuan respectively.

   I intend to buy the following:

   $$(2 \times \text{apples} + 3 \times \text{bananas} + 4 \times \text{cherries}).$$

   When I check out, the linear combination of fruits will transform into the same linear combination of prices of each fruit, which is what I must pay:

   $$(2 \times \text{apple price} + 3 \times \text{banana price} + 4 \times \text{cherry price}) = 17\text{yuan}.$$

   In this example, the inputs are linear combinations of fruits, and the outputs are real numbers. In mathematical words, we have a function "CheckOut" that sends linear combination of fruits to their total price. In this case, we have

   $$\text{CheckOut}(2 \times \text{apples} + 3 \times \text{bananas} + 4 \times \text{cherries}) = 17\text{yuan}.$$

   You can easily see that, for any real number x, y, and any linear combination of fruits $\boldsymbol{a}, \boldsymbol{b}$, we should always have

   $$\text{CheckOut}(x\boldsymbol{a} + y\boldsymbol{b}) = x\text{CheckOut}(\boldsymbol{a}) + y\text{CheckOut}(\boldsymbol{b}).$$

   In particular, even though we are changing fruits into numbers, the structure of linear combination is in fact preserved.

2. I put a bunch of hens and rabbits into a cage. Each hen has 1 head, 2 legs. Each rabbit has 1 head and 4 legs. Then if I put a linear combination $3 \times \text{hens} + 5 \times \text{rabbits}$ into the cage, then in my cage, I would have the linear combination

$$3(1 \times \text{head} + 2 \times \text{leg}) + 5(1 \times \text{head} + 4 \times \text{legs}) = 8 \times \text{head} + 26 \times \text{legs}.$$

In this example, the inputs are linear combinations of animals, and the outputs are linear combinations. In mathematical words, we have a function "Count", that sends linear combinations of animals to linear combinations of body parts.

If $\boldsymbol{a}, \boldsymbol{b}$ are two linear combination of animals, then $\text{Count}(x\boldsymbol{a} + y\boldsymbol{b}) = x\text{Count}(\boldsymbol{a}) + y\text{Count}(\boldsymbol{b})$.

☺

As you can see, at the center of this is a kind of functions with a special property. These are called ***linear maps***. In mathematical words, a linear map is a function f that preserves the structure of linear combinations, i.e., if $\boldsymbol{a}, \boldsymbol{b}$ are possible inputs, and we combine them into a new input $x\boldsymbol{a}+y\boldsymbol{b}$, then we always have $f(x\boldsymbol{a} + y\boldsymbol{b}) = xf(\boldsymbol{a}) + yf(\boldsymbol{b})$.

1.2 The spaces $\mathbb{R}^n$

1.2.1 Definition of $\mathbb{R}^n$

Now, I hope that you acknowledge that linear combinations and linear maps are omnipresent in our lives. However, the job of mathematics is to do abstractions, i.e., focus only on the structure and ignore the irrelavant details.

Say I want to talk about fruit combinations of apples, bananas and cherries. Maybe I'm tired of keep saying $(2 \times \text{apples} + 3 \times \text{bananas} + 4 \times \text{cherries})$. Instead, I may simply write a sequence of real numbers, i.e., $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$. This saves time, so why not?

Note that the ORDER of these real numbers are vital. If I have $\begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$, then it will NOT be $(2 \times \text{apples} + 3 \times \text{bananas} + 4 \times \text{cherries})$. Instead, it is $(3 \times \text{apples} + 2 \times \text{bananas} + 4 \times \text{cherries})$.

Definition 1.2.1. *1. We define $\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ or simply $\mathbb{R}^n$ to be the set of ordered list of n elements of $\mathbb{R}$, i.e., $\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ such that all these "coordinates" are in $\mathbb{R}$. (Traditionally people also write $(a_1, ..., a_n)$ to denote this ordered list of elements. But in our class, we uses the vertical notation. Don't ask why, it is just a tradition for linear algebra, probably for aethetic reasons that you shall see later, when we multiply matrices to vectors.)*

*2. For elements of $\mathbb{R}^n$, which we shall call **real vectors**, we define vector addition as $\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ \vdots \\ a_n + b_n \end{bmatrix}$.*

3. *For any real number* $k \in \mathbb{R}$ *and a real vector* $\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^n$*, we define scalar multiplication as* $k \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} ka_1 \\ \vdots \\ ka_n \end{bmatrix}$.

4. *We say the dimension of* $\mathbb{R}^n$ *is the number* n.

In our class, if we refer to $\mathbb{R}^n$, we shall always use the vector addition and scalar multiplication defined above.

Example 1.2.2.

In the breakfast example, everything seems to be in $\mathbb{R}^3$. Of course, one need to have an open mind about "decimal quantities" of fruits, or "negative quantities" of fruits. Well, maybe I cut an apple in half to have $0.5 \times$ apple. Or maybe instead of eating an apple for breakfast, I vomit out an apple, so today I had $(-1) \times$ apple.

We all know negative energy and negative mass could be useful in physics, so why not negative apples? I'm simply pleading the fair readers to keep a flexible mind about things such as these. (Mathematical rigor is never my main concern. I did call these sets of notes POISONOUS, because traditional math teachers will probably not stand for such things....) ☺

Example 1.2.3.

When given a plane, say we are doing some high school geometry problem, then we can draw a coordinate chart. By doing so, we have declared that each point on the plane is now essentially a pair of real numbers. What is this? This is $\mathbb{R}^2$.

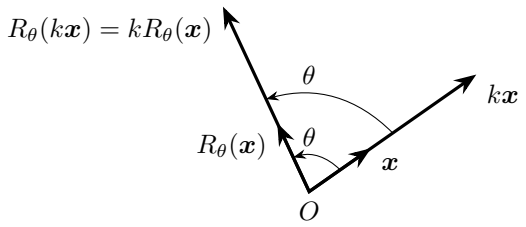
So in a sense, we can think of $\mathbb{R}^2$ as a plane with a coordinate chart. Similarly, one can think of $\mathbb{R}^3$ as the space with a coordinate chart. In this fashion, one can see that, the so called "n-dimensional space" is simply $\mathbb{R}^n$ if we draw a coordinate chart. We may not see higher dimensional space, but we can certainly calculate. Just treat it as $\mathbb{R}^n$ and we can happily compute away whatever we come across. ☺

What about linear maps? Recall that a function is a linear map if it sends linear combinations of many inputs to the SAME linear combination of respective outputs, i.e., $f(\sum a_i \boldsymbol{v}_i) = \sum a_i f(\boldsymbol{v}_i)$.

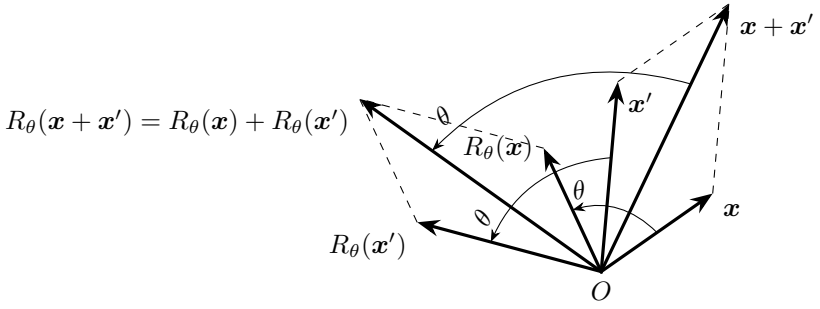
Definition 1.2.4. *A function* $f : \mathbb{R}^n \to \mathbb{R}^m$ *is said to be a* ***linear map*** *if* $f(\sum a_i \boldsymbol{v}_i) = \sum a_i f(\boldsymbol{v}_i)$ *for all* $a_i \in \mathbb{R}, \boldsymbol{v}_i \in \mathbb{R}^n$.

Example 1.2.5. To show that a map f is linear, it is enough to check that $f(k\boldsymbol{v}) = kf(\boldsymbol{v})$ and $f(\boldsymbol{v} + \boldsymbol{w}) = f(\boldsymbol{v}) + f(\boldsymbol{w})$. These will obviously allow you to show that f respect all linear combinations.

Consider a map $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$. For each arrow vector $\boldsymbol{v} \in \mathbb{R}^2$, R_θ simply rotate it by θ counter clockwise. Then this map is linear, as can be seen in the graphs below.



(a) Scalar multiplication



(b) Vector addition

Figure 1.2.1: Rotation is linear

☺

Keep in mind of a very important property of linear maps. Since the output is always proportional to the input, a linear map should always send no input to no output.

Proposition 1.2.6. *Given a linear map $f : \mathbb{R}^n \to \mathbb{R}^m$, we must have $f(\mathbf{0}) = \mathbf{0}$.*

Here $\mathbf{0}$ refers to the zero vector $\mathbf{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$, or the ***origin*** of the space $\mathbb{R}^n$ or $\mathbb{R}^m$. In the formula $f(\mathbf{0}) = \mathbf{0}$, the input $\mathbf{0}$ is the origin of $\mathbb{R}^n$, while the output $\mathbf{0}$ is the origin of $\mathbb{R}^m$, so technically they might not be the same zero vector. (They might have different number of coordinates, even though all coordinates are zero.)

Proof. We just calculate this

$$f(\mathbf{0}) = f(2\mathbf{0}) = 2f(\mathbf{0}).$$

Now substract both sides by $f(\mathbf{0})$, we see that $\mathbf{0} = f(\mathbf{0})$. □

Example 1.2.7. In particular, a translation map (shift all inputs by the same vector), say $T : \mathbb{R}^2 \to \mathbb{R}^2$ where $T(\boldsymbol{v}) = \boldsymbol{v} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, is NOT linear. This is because $T(\mathbf{0})$ is no longer $\mathbf{0}$.

This map is in fact an ***affine maps***, which is basically a linear map but you can then add a constant vector afterwards. So functions $f : \mathbb{R} \to \mathbb{R}$ such as $f(x) = 2x$ are linear, and functions such as $f(x) = 2x + 1$ are affine. We do not do affine maps in this class. (But affine maps can still be studied in similar manners.) ☺

1.2.2 Coordinates and Matrices

Now the single most important property of $\mathbb{R}^n$ is a basis.

Definition 1.2.8. *For the vectors* $\begin{bmatrix}1\\0\\0\\\vdots\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\\vdots\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\\vdots\\0\\1\end{bmatrix}$*, we usually denote them as* $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$*, and we call these the standart basis vectors for* $\mathbb{R}^n$*.*

Proposition 1.2.9. *Every vector in* $\mathbb{R}^n$ *is a UNIQUE linear combination of the standard basis.*

Proof. The proof is trivial, but let us do it to see some bigger picture. "Unique" means exactly one. To show that something is unique, you need to show that there is at least one, and then you need to show that there is at most one. This is the standard procedure to prove uniqueness.

Every vector is AT LEAST one linear combination of them:

We have $\begin{bmatrix}a_1\\\vdots\\a_n\end{bmatrix} = \sum a_i \boldsymbol{e}_i$. Honestly, this is how you should think of coordinates. They are just coefficients telling you how to combine the standard basis vectors.

Every vector is AT MOST one linear combination of them:

Suppose $\begin{bmatrix}a_1\\\vdots\\a_n\end{bmatrix} = \sum b_i \boldsymbol{e}_i$ is another linear combination. Then note that $\sum b_i \boldsymbol{e}_i = \begin{bmatrix}b_1\\\vdots\\b_n\end{bmatrix}$. So we see that $\begin{bmatrix}a_1\\\vdots\\a_n\end{bmatrix} = \begin{bmatrix}b_1\\\vdots\\b_n\end{bmatrix}$, which by definition of $\mathbb{R}^n$ indicates that $a_i = b_i$ for all i. So it is NOT another linear combination after all, it is the same one. □

The idea of the standard basis is implicit in many applications already. For example, if we were doing linear combinations of apples, bananas and cherries, then the standard basis is simple the following three vectors: a single apple, a single banana, and a single cherry.

The following is an example that reveals how we could exploit these structures to our benefit.

Example 1.2.10. On a plane, each vector is given by two coordinates. (Note that in our class, we usually write vectors vertically.) Given a vector $\boldsymbol{v} = \begin{bmatrix}x\\y\end{bmatrix}$, we can rotate it counterclockwise by θ. What is the coordinate of the result in terms of x, y, θ?

On the face, this appears to be a rather annoying problem. Sure, in the traditional sense, you can draw graphs, draw auxilliary lines (Fu Zhu Xian), analyze triangles, use trigonometries (San Jiao Han Shu), and eventually find the answer. But here we shall present another proof, which is both straightforward and elementary, if we start with exploiting the linear structures. THIS IS HOW WE SHOULD THINK IN THIS CLASS!

Consider the rotation as a map R_θ sending vectors to vectors. Then this is a linear map!

Then in particular, we should have the following computation:

$$R_\theta(\begin{bmatrix}x\\y\end{bmatrix}) = R_\theta(x\begin{bmatrix}1\\0\end{bmatrix} + y\begin{bmatrix}0\\1\end{bmatrix}) = xR_\theta(\begin{bmatrix}1\\0\end{bmatrix}) + yR_\theta(\begin{bmatrix}0\\1\end{bmatrix}).$$

In particular, we will figure out the rotation formula as soon as we find out the value of $R_\theta(\begin{bmatrix}1\\0\end{bmatrix})$ and $R_\theta(\begin{bmatrix}0\\1\end{bmatrix})$! In fact, it is easy to see the following:

$$R_\theta(\begin{bmatrix}1\\0\end{bmatrix}) = \begin{bmatrix}\cos\theta\\\sin\theta\end{bmatrix};$$

$$R_\theta(\begin{bmatrix}0\\1\end{bmatrix}) = \begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}.$$

As a result, we see that

$$R_\theta(\begin{bmatrix}x\\y\end{bmatrix}) = x\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix} + y\begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix} = \begin{bmatrix}x\cos\theta - y\sin\theta\\ x\sin\theta + y\cos\theta\end{bmatrix}.$$

☺

As you can see from the example above, the only computations we did are the rotations of only two vectors. By understanding the rotation of two vectors (the two standard basis vectors), we now understand the rotation of all vectors.

The key to our simplicity is the fact that EVERY vectors here can be UNIQUELY expressed as the linear combination of $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$. This means that by studying only two vectors, we manage to study ALL vectors (as far as linear things are concerned)!

And this leads to a very central concept of study in this course, the infamous ***matrices***. As we have seen, the image of the standard basis vectors $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ via R_θ, i.e., $\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}$ and $\begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}$, would completely determine the entire formula for R_θ!

So we say that the linear map R_θ is represented by the matrix $\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}$, where the first colum records the image of the first standard basis vector $\begin{bmatrix}1\\0\end{bmatrix}$, and the second column records the image of the second standard basis vector $\begin{bmatrix}0\\1\end{bmatrix}$. As we have seen, together they record ALL info about this linear map.

Definition 1.2.11. *1. An m by n **real matrix** (plural of matrix is matrices) is a rectangular array of real numbers with m rows and n columns, like $\begin{bmatrix}a_{11} & \dots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1} & \dots & a_{mn}\end{bmatrix}$. We usually denote these matrices (plural form of "matrix") with capital letters like A, B, C, D, M, X, Y. For example, we might say $A = \begin{bmatrix}a_{11} & \dots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1} & \dots & a_{mn}\end{bmatrix}$.*

2. *We call the number in the i-th row and j-th column of this matrix the (i, j) entry of this matrix.*
3. *Given a linear map $f : \mathbb{R}^n \to \mathbb{R}^m$, its **matrix** is $M_f := \begin{bmatrix}f(\boldsymbol{e}_1) & \dots & f(\boldsymbol{e}_n)\end{bmatrix}$, so the columns are exactly images of the standard basis vectors.*

Remark 1.2.12. *Note that we sometimes write matrices as a capital letter A, sometimes as its entries $\begin{bmatrix}a_{11} & \dots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1} & \dots & a_{mn}\end{bmatrix}$, sometimes as its column vectors $\begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}$, and sometimes as its row vectors $\begin{bmatrix}\boldsymbol{r}_1^{\mathrm{T}}\\ \dots\\ \boldsymbol{r}_n^{\mathrm{T}}\end{bmatrix}$. (Here the letter T in the exponent means this is a horizontal vector, instead of a vertical vector.) Despite different notations, it is easy to see that they are all about rectangular arrays of numbers.*

Think about how this works. If a linear map f has a matrix $M_f = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}$, then it implies that $f(\boldsymbol{e}_i) = \boldsymbol{a}_i$. So when we do calculations, we see that $f(\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}) = f(\sum x_i\boldsymbol{e}_i) = \sum x_i f(\boldsymbol{e}_i) = \sum x_i\boldsymbol{a}_i$. So

we merely need do a linear combination of columns of M_f, with coefficients according to the inputting coordinates.

Example 1.2.13 (Nutrition table is a linear map)**.** Suppose $f : \mathbb{R}^3 \to \mathbb{R}^2$ is some unknown linear map, whose matrix is $\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}$. Then to find $f(\begin{bmatrix}2\\2\\3\end{bmatrix})$, we have

$$\begin{aligned} f(\begin{bmatrix}2\\2\\3\end{bmatrix}) =& f(2\begin{bmatrix}1\\0\\0\end{bmatrix} + 2\begin{bmatrix}0\\1\\0\end{bmatrix} + 3\begin{bmatrix}0\\0\\1\end{bmatrix}) && \text{express the input in the standard basis}\\ =& 2f(\begin{bmatrix}1\\0\\0\end{bmatrix}) + 2f(\begin{bmatrix}0\\1\\0\end{bmatrix}) + 3f(\begin{bmatrix}0\\0\\1\end{bmatrix}) && \text{use linearity of } f\\ =& 2\begin{bmatrix}1\\4\end{bmatrix} + 2\begin{bmatrix}2\\5\end{bmatrix} + 3\begin{bmatrix}3\\6\end{bmatrix} && \text{use the matrix of } f\\ =& \begin{bmatrix}15\\36\end{bmatrix} && \text{calculate.} \end{aligned}$$

Maybe, say, the domain $\mathbb{R}^3$ represent the "fruit space", where $\begin{bmatrix}2\\2\\3\end{bmatrix}$ means two apples, two bananas and three cherries. While the codomain $\mathbb{R}^2$ here represent the "nutrition space", where $\begin{bmatrix}1\\4\end{bmatrix}$ means one unit of fiber and four units of sugar.

Now suppose our matrix is actually the nutrition table

$$\begin{bmatrix} & \text{apple} & \text{banana} & \text{cherry}\\ \text{fiber} & 1 & 2 & 3\\ \text{sugar} & 4 & 5 & 6 \end{bmatrix}.$$

Then what is the meaning of the f? It is simply a function sending any combination of fruits to the contained nutrition!

(I made up these numbers arbitrarily. They in no way reflect the real nutrition value in these fruits....) ☺

Definition 1.2.14. *Given an m by n matrix $A = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}$, and a vector $\boldsymbol{x} = \begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix}$, we define their multiplication to be*

$$A\boldsymbol{x} = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}\begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix} = \sum x_i \boldsymbol{a}_i.$$

Corollary 1.2.15. *If a linear map f is represented by a matrix A, then $f(\boldsymbol{v}) = A\boldsymbol{v}$.*

We have an big formula for matrix-vector multiplication below. Try to wrap your mind around it at least. However, in practice I would highly recommand you to take the column view presented above, or the row view that we shall introduce in the future.

$$\begin{bmatrix}a_{11} & \dots & a_{1n}\\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn}\end{bmatrix}\begin{bmatrix}x_1\\ \vdots \\ \vdots \\ x_n\end{bmatrix} = \begin{bmatrix}\sum a_{1i}x_i\\ \vdots \\ \sum a_{mi}x_i\end{bmatrix}.$$

Remark 1.2.16. *Note that the input vector might NOT be as "tall" as the matrix. But the output vector MUST be as "tall" as the matrix. Think about why.*

We may sometimes think about matrices and linear maps interchangably. Given a linear map, the first thing we do should be finding out its matrix. Given a matrix, the best way to interpret this array of numbers is to think of it as a linear map. Matrices are for calculation, and maps are for interpretation. They serve each other very well. Let us see some more examples.

Example 1.2.17.

Say I go shopping and buy apples, bananas and cherries, and the prices for each is 3 yuan, 1 yuan and 2 yuan respectively. When I check out, I send fruit combinations to their total cost. Now, what's vital to this "CheckOut" function are the unit prices, the prices for a single apple, a single banana, or a single cherry. Why is that? because it is precisely the value of "CheckOut" on the standard basis! Essentially, the unit prices tell us $\text{CheckOut}(\boldsymbol{e}_1) = 3, \text{CheckOut}(\boldsymbol{e}_2) = 1, \text{CheckOut}(\boldsymbol{e}_3) = 2$. As a result,

$$\begin{aligned}&\text{CheckOut}(\begin{bmatrix}a\\b\\c\end{bmatrix})\\=&\begin{bmatrix}3 & 1 & 2\end{bmatrix}\begin{bmatrix}a\\b\\c\end{bmatrix}\\=&3a+b+2c.\end{aligned}$$

We see that the matrix for check out is simply a "row vector" $\begin{bmatrix}3 & 1 & 2\end{bmatrix}$. ☺

Example 1.2.18 (Hens and Rabbits in a Cage)**.**

Ah, the famous Ji Tu Tong Long problem.

I put a bunch of hens and rabbits into a cage. Each hen has 1 head, 2 legs and 2 wings. Each rabbit has 1 head and 4 legs. When we attempt to do the counting, we see that $\text{Count}(\boldsymbol{e}_1) = \begin{bmatrix}1\\2\end{bmatrix}$, and $\text{Count}(\boldsymbol{e}_2) = \begin{bmatrix}1\\4\end{bmatrix}$. So if we input x hens and y rabbits, we have

$$\begin{aligned}&\text{Count}(\begin{bmatrix}x\\y\end{bmatrix})\\=&\begin{bmatrix}1 & 1\\2 & 4\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}\\=&\begin{bmatrix}x+y\\2x+4y\end{bmatrix}.\end{aligned}$$

Now, in light of this, what is the traditional hen-rabbit cage problem? Well, it states that given the number of heads and legs, say 6 heads and 20 legs, can we deduce the number of hens and rabbits?

So given unknown vector $\boldsymbol{x}$, we want to solve it from the equation $A\boldsymbol{x} = \begin{bmatrix}6\\20\end{bmatrix}$, where $A = \begin{bmatrix}1 & 1\\2 & 4\end{bmatrix}$ is the matrix here. (This is a standard pre-image problem.)

The question is equivalent to the following: how can we combine $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$ into $\begin{bmatrix}6\\20\end{bmatrix}$?

The linear way of thinking might goes like this. Instead of getting $\begin{bmatrix}6\\20\end{bmatrix}$, let us first try to get $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ instead. Just by staring at these vectors, and try some values, it is easy to realize the following:

1. To create $\begin{bmatrix}1\\0\end{bmatrix}$ from $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$, I would need multiples of $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$ to cancel each other's second coordinate. So we can see that
$$2\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}1\\4\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}.$$

2. To create $\begin{bmatrix}0\\1\end{bmatrix}$ from $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$, I would need multiples of $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$ to cancel each other's first coordinate. So we can see that

$$\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}1\\4\end{bmatrix} = \begin{bmatrix}0\\-2\end{bmatrix}.$$

Thus we have

$$-\frac{1}{2}\begin{bmatrix}1\\2\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\4\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix}.$$

Note that $f(\begin{bmatrix}1\\0\end{bmatrix}) = \begin{bmatrix}1\\2\end{bmatrix}$ and $f(\begin{bmatrix}0\\1\end{bmatrix}) = \begin{bmatrix}1\\4\end{bmatrix}$. Substituting this, we get

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

$$-\frac{1}{2}f(\begin{bmatrix}1\\0\end{bmatrix}) + \frac{1}{2}f(\begin{bmatrix}0\\1\end{bmatrix}) = \begin{bmatrix}0\\1\end{bmatrix}.$$

Using linearity of f, we see that

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

$$f(\begin{bmatrix}-\frac{1}{2}\\\frac{1}{2}\end{bmatrix}) = \begin{bmatrix}0\\1\end{bmatrix}.$$

Now, our desire is to get a right hand side of $\begin{bmatrix}6\\20\end{bmatrix}$. So simply add six copies of the first equation and twenty copies of the second equation above, and we get

$$6f(\begin{bmatrix}2\\-1\end{bmatrix}) + 20f(\begin{bmatrix}-\frac{1}{2}\\\frac{1}{2}\end{bmatrix}) = \begin{bmatrix}6\\20\end{bmatrix}.$$

Use linearity again, we get our answer

$$f(\begin{bmatrix}2\\4\end{bmatrix}) = \begin{bmatrix}6\\20\end{bmatrix}.$$

So the answer is 2 hens and 4 rabbits.

Note that, our original problem is $f(\boldsymbol{x}) = \boldsymbol{b}$ for some constant vector $\boldsymbol{b}$. Therefore, ideally, IF we have an inverse function f^{-1}, then one would simply have $\boldsymbol{x} = f^{-1}(\boldsymbol{b})$. How would one figure out this inverse function? Again, it is enough to figure it out on the standard basis. If you look at our arguments above, you will realize that we are essentially proving that $f^{-1}(\begin{bmatrix}1\\0\end{bmatrix}) = \begin{bmatrix}2\\-1\end{bmatrix}$ and $f^{-1}(\begin{bmatrix}0\\1\end{bmatrix}) = \begin{bmatrix}-\frac{1}{2}\\\frac{1}{2}\end{bmatrix}$. Hence the inverse map have a matrix of $A^{-1} = \begin{bmatrix}2 & -\frac{1}{2}\\-1 & \frac{1}{2}\end{bmatrix}$. What does this mean? It means that 2 hens and -1 rabbit will give 1 head and no leg, while $-\frac{1}{2}$ hen and $\frac{1}{2}$ rabbit will give no head and 1 leg.

Now given any constant vector $\boldsymbol{b}$, to solve $\boldsymbol{x}$ from $A\boldsymbol{x} = \boldsymbol{b}$, we simply have $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Hence simpy multiplying the matrix $\begin{bmatrix}2 & -\frac{1}{2}\\-1 & \frac{1}{2}\end{bmatrix}$ with $\boldsymbol{b}$ would give us the answer. We have solved ALL possible hen-rabbit cage problems in one go. ☺

1.2.3 (Optional) Modular Examples

The idea of linear combinations and basis extends far beyond just linear algebra. Here is an ancient Chinese problem on mathematics. (Sun Zi Ding Li, or Chinese Remainder Theorem.)

Example 1.2.19. Suppose I have an integer. When divided by 3, the remainder is 2. When divided by 5, then remainder is 3. When divided by 7, the remainder is 4. Find all such integers.

In the ancient text of Sun Zi Suan Jing, the standard solution is this: multiply the remainder for 3 by 70. Multiply the remainder for 5 by 21. Multiply the remainder for 7 by 15. Now the answer is the sum of these numbers plus an arbitrary multiple of 105.

Why does this work? The first realization is the following: for each integer $n \in \mathbb{Z}$, we can send it to $r(n)$, a triple of remainders $\begin{bmatrix} n \mod 3 \\ n \mod 5 \\ n \mod 7 \end{bmatrix}$. For example, for the number 73, we have $r(73) = \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix}$. You can easily check that $r(a+b) = r(a) + r(b)$ and $r(ab) = a \cdot r(b)$, where addition and multiplication is done with respect to the modular arithmetic of each coordinate. In some sense, this "remainder map" r is "linear".

Now, the algorithm described by Sun Zi Suan Jing is merely pointing out that $r(70) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, r(21) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, r(15) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. So in this sense the numbers 70, 21 and 15 form a "basis". Now to seek a number with remainders $a \mod 3, b \mod 5, c \mod 7$, one obvious candidate would then be $70a + 21b + 15c$, because it would yield

$$\begin{aligned} r(70a + 21 + 15c) =& a \cdot r(70) + b \cdot r(21) + c \cdot r(15) \\ =& a \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ =& \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \end{aligned}$$

So $70a + 21b + 15c$ is a number with the desired remainders.

Now if n is another number with the desired remainders, then $n - (70a + 21b + 15c)$ would have all remainders zero. So it must be a multiple of 3, 5, and 7, and hence it is a multiple of 105. So n is $70a + 21b + 15c$ with an arbitrary multiple of 105.

As you can see, at the heart of this problem is the idea of linear combinations, and expressions from a collection of "basis" elements. ☺

Technically, the example above is NOT considered a problem of linear algebra, because the three coordinates take values in DIFFERENT sets. To be precise, this is actually a $\mathbb{Z}$-module (because you can multiply $\begin{bmatrix} a \mod 3 \\ b \mod 5 \\ c \mod 7 \end{bmatrix}$ by any integer).

However, the idea is completely about linear combinations, as you can clearly see from the examples. And as such it opens up a lot of applications of ideas we are going to use.

Part I

The Space $\mathbb{R}^n$

Chapter 2

Systems of Linear Equations

2.1 Preliminary on Maps

We need to get some terminology straight. First we need sets. What is a set? Intuitively, a ***set*** is a collection of stuff, and these stuff inside a set are called ***elements***. This intuitive picture will serve us just fine, so let us NOT dive into the set theory stuff for the moment. We only focus on what we would need.

Remark 2.1.1. *This intuitive understanding of set might run into trouble such as Russel's paradox. Learn set theory if you'd like to know more.*

Example 2.1.2. Say $S = \{1, 2, \clubsuit\}$ and $T = \{2, 3\}$. Then we can create the ***union*** of the sets $S \cup T = \{1, 2, 3, \clubsuit\}$. We can also create the ***intersection*** of the sets $S \cap T = \{2\}$.

We can also create the ***product*** (or more specifically a ***Cartesian product***) of the sets as $S \times T = \{\begin{bmatrix} s \\ t \end{bmatrix} : s \in S \text{ and } t \in T\}$, or more specifically,

$$S \times T = \{\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} \clubsuit \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} \clubsuit \\ 3 \end{bmatrix}\}.$$

Note that if $|S|$ means the number of elements in a set S, then we have $|S \times T| = |S| \times |T|$ for finite sets S, T.

Finally, we write $S \times \cdots \times S$ (n times) simply as S^n, as we did in the notation of $\mathbb{R}^n$. Note that we have $|S^n| = |S|^n$ for a finite set S. ☺

Now we need maps to connect various sets.

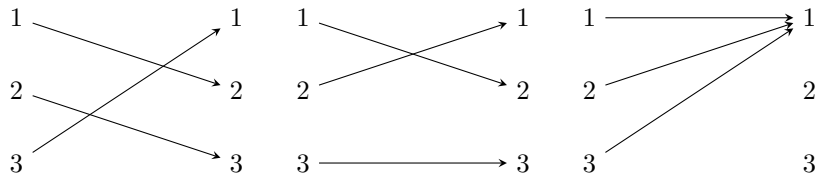
1. A ***map*** (or sometimes called a ***function***) is a machine that takes an element (or "input") from one set, and transform it into an element (or "output") in anther set.
2. We usually write $f : X \to Y$, if f is sending elements of X to elements of Y.
3. We call the set of inputs the ***domain*** of our map, and the set of outputs the ***codomain*** of our map.
4. If the domain and codomain are the same, we also say that this map is a ***transformation***.
5. If f sends an input x to y, we write $f(x) = y$ or $f : x \mapsto y$. We say y is the ***image*** of x, and x is a ***pre-image*** of y.

   (IMPORTANT DETAIL: be careful of the difference of "a" and "the". When I write "the image", I am implying that such image is UNIQUE, that there is only one possible image for x. When I write "a pre-image", I am saying that maybe y has more than one pre-image. We simply don't know and make no assumptions.)

Example 2.1.3. 1. A human is a map. If we think of X as the set of all foods, and Y as the set of all, well, you know. Then $human : X \to Y$ is a map. Each particular food is transformed by a human into a particular piece of, well, you know.

2. In the world of Harry Potter, a wand is a function that sends incantations into magical spells. (Incidentally, "f" looks like a wand in shape.)

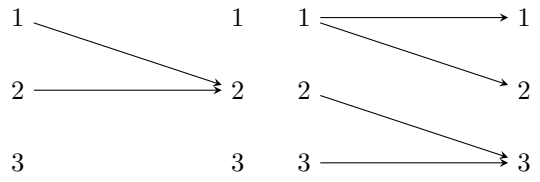
3. Let $X = \{1, 2, 3\}$. Then $f : X \to X$ with $f(1) = 2, f(2) = 3, f(3) = 1$ is a map, and $g : X \to X$ with $g(1) = 2, g(2) = 1, g(3) = 3$ is a map, and $h : X \to X$ with $h(x) = 1$ for all x is a map. In fact, let us draw some diagrams to see them.



(a) Map f (b) Map g (c) Map h

4. Of course we also know many maps from $\mathbb{R}$ to $\mathbb{R}$. Say $f(x) = x + 1, f(x) = 2x, f(x) = e^x, f(x) = \ln(x), f(x) = \sin x$, etc..

5. What is NOT a function? Well, here are two cases where $f : \{1, 2, 3\} \to \{1, 2, 3\}$ is NOT a function.



(a) 3 has no defined value (b) 1 has too many defined value

☺

Remark 2.1.4. *Many academic text in English uses the following shorthands all the time. Might as well get used to them. (They all come from Latin.)*

Some useful shorthands in any English text:

1. *The word "etc." means "and so on...". For example, "I ate some apples, bananas, cherries, etc.."*
2. *The word "i.e." means "that is", and it usually means the next words are explanations of the previous words. For example, "I went to my dream university, i.e., Tsingua University."*
3. *The word "e.g." means "for example". For example, "I like all fruits, e.g., apples, bananas, cherries, etc.."*
4. *This is not Latin, but it is VERY important. If I write "**iff**", it does not mean "if", and it is NOT a typo. It means "if and only if".*

Now a central theme of mathematics is to try to figure out inverse maps, if possible.

Example 2.1.5. 1. My son had red poop one day. I was super scared for a minute, but then I was reminded that he had dragon fruits for lunch. If we think of my son as a map sending foods into poops, then given the red poop, I have successfully figured out the pre-image.

2. Say $f(x) = 2x + 5$. How to solve for $f(x) = 3$? Well, $2x + 5 = 3$, and then we see that $x = -1$.

3. Say $f(x) = e^x$. How to solve for $f(x) = 1$? Well, $e^x = 1$, so we have $x = \ln 1 = 0$.

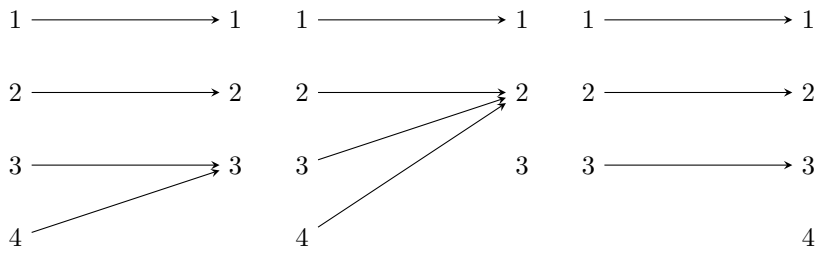
4. Say $f(x) = x^2$. How to solve for $f(x) = -1$? Well, you cannot....

5. Say $f(x) = x^2$. How to solve for $f(x) = 1$? Well, there are two solutions $x = \pm 1$. There is no UNIQUE solution!

☺

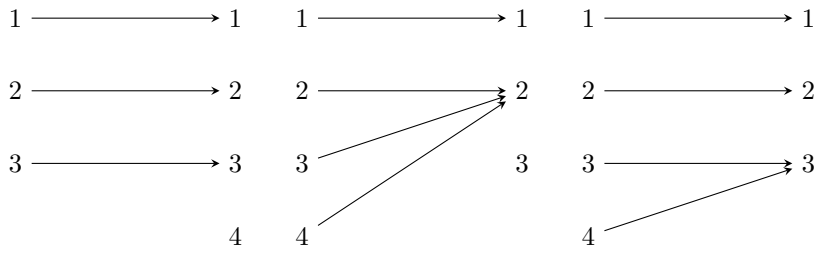
Obviously the last two scenarios are less desirable. Preferably, we hope our maps would have these kinds of properties:

1. If every element in the codomain has at least one preimage, then we say the map is a ***surjective*** map. Or simply say our map is a ***surjection***. (Everything is the codomain is hit by some arrow in the diagram.)



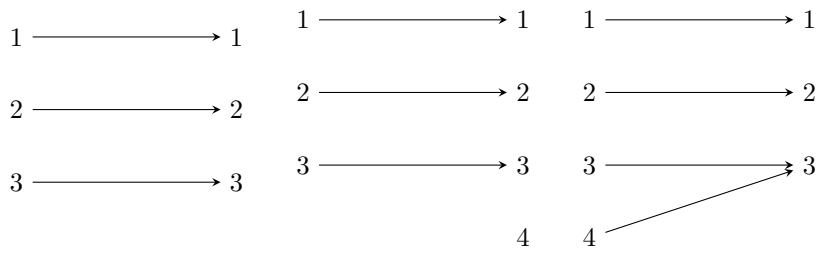
(a) Surjection (b) Not a surjection (c) Not a surjection

2. If every element in the codomain has at most one preimage, then we say the map is a ***injective*** map. Or simply say our map is an ***injection***. (Arrows never collide to the same targets in the diagram.)



(a) Injection (b) Not an injection (c) Not an injection

3. If every element in the codomain has a UNIQUE preimage (exactly one), then we say the map is a ***bijective*** map. Or simply say our map is a ***bijection***. (Arrows give a one-to-one correspondence in the diagram.)



(a) Bijection (b) Not a bijection (c) Not a bijection

Remark 2.1.6.

Bijection = each element in the codomain has exactly one pre-image.

Surjection = each element in the codomain has at least one pre-image.

Injection = each element in the codomain has at most one pre-image.

Since exactly one = "at least one" +"at most one", we immediately see that bijection = surjection + injection.

Proposition 2.1.7 (Inverse map)**.** *If $f : X \to Y$ is a bijection, then there is a UNIQUE map $g : Y \to X$ such that $f(x) = y$ if and only if $g(y) = x$.*

Proof. ***Existence of*** g***:*** Draw diagram with arrows as above. Reverse all arrows. Done.

As an extra remark, look at what would happen if f is not bijective. If f is not injective, then we have $f(a) = f(b) = c$ for some $a \neq b$. Then reversing the arrows would require $g(c) = a$ AND $g(c) = b$, contradiction.

If f is not surjective, then there is some $y \in Y$ with no pre-image. So reversing the arrows, $g(y)$ is still undefined.

In conclusion, bijectivity of f is central here.

Uniqueness of g***:*** How to show that something is unique? Well, you simply show that any two possible things must actually be the same, and therefore there is only one thing after all.

Suppose g, h both satisfy the requirements, we aim to show that $g = h$. How to show that two maps are the same? Well, if they send the same inputs to the same outputs, then they are the same as maps. Pick any $y \in Y$. But then since f is a bijection, there is a UNIQUE x such that $f(x) = y$. Then by requirements we must have $g(y) = x$ and $h(y) = x$. So we see that $g(y) = h(y)$ for all y, and thus they are the same as maps. □

In this case, we say g is the ***inverse map*** of f, and write $g = f^{-1}$.

Proposition 2.1.8. *If $f : X \to Y$ has an inverse map $g : Y \to X$, then both are bijections.*

Proof. For each $y \in Y$, then $g(y) \in X$ is a pre-image of y for f. Hence f is surjective.

Suppose $f(x) = f(x')$. Let y represent this element. Then $f(x) = y$ implies that $x = g(y)$, while $f(x') = y$ implies that $x' = g(y)$. Hence $x = x'$. So f is injective.

The case of g is identical. □

Remark 2.1.9. *This is the standard way to prove surjectivity and injectivity. To prove that some $f : X \to Y$ is surjective, for each $y \in Y$, simpy find a pre-image. To prove that f is injective, then assume $f(x_1) = f(x_2)$, and try to deduce $x_1 = x_2$ from that.*

Example 2.1.10. Here are some examples. Pay special attention to the codomain.

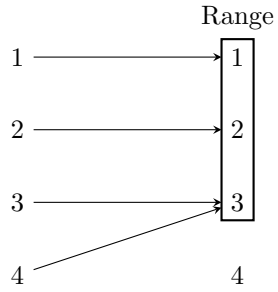
1. The map $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = e^x$ has NO inverse. It is injective but NOT surjective, since it can never take non-positive value.

2. The map $f : \mathbb{R} \to \mathbb{R}^+$ with $f(x) = e^x$ has an inverse. Here $\mathbb{R}^+$ refers to the set of positive real numbers. Its inverse is $f^{-1} : \mathbb{R}^+ \to \mathbb{R}$ with $f^{-1}(x) = \ln(x)$.
3. The map $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is NEITHER surjective NOR injective.
4. The map $f : \mathbb{R} \to [-1, 1]$ with $f(x) = \sin x$ is surjective but NOT injective.
5. The map $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = \frac{1}{x}$ is NOT really a map, since $f(0)$ is not defined.
6. But the map $f : \mathbb{R} - \{0\} \to \mathbb{R} - \{0\}$ with $f(x) = \frac{1}{x}$ is a bijection. It is its own inverse.

☺

As you can see, the codomain MATTERS! Let us have a definition here to differentiate the idea between "the whole codomain" and "the codomain that the map actually touched". For any map $f : X \to Y$, we say the ***range*** or ***image*** of the map is the set $\{f(x) \mid x \in X\}$. We may write it as $\mathrm{Ran}(f), \mathrm{Im}(f)$ or simply $f(X)$ sometimes. Obviously, f is surjective iff $f(X) = Y$.

Here is an example of a function $f : \{1, 2, 3, 4\} \to \{1, 2, 3, 4\}$ where its range is different from its codomain. Therefore it is NOT surjective.



Remark 2.1.11. *In fact, for any subset of the domain $S \subseteq X$, we can define $f(S) = \{f(x) \mid x \in S\}$. Conversely, for any subset of the codomain $S \subseteq Y$, we can define $f^{-1}(S) = \{x \mid f(x) \in S\}$, EVEN when the map is NOT invertible! Here f^{-1} does NOT refer to the inverse map. It is simply a symbolic convenience, referring to the pre-images of f.*

One last thing, composition. Given a map $f : X \to Y$ and $g : Y \to Z$, their ***composition*** is a map $g \circ f : X \to Z$, such that $g \circ f(x) = g(f(x))$.

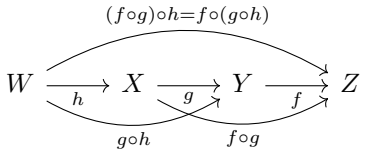
Proposition 2.1.12. *Map composition is associative, i.e., $(f \circ g) \circ h = f \circ (g \circ h)$ for maps $f : Y \to Z, g : X \to Y, h : W \to X$.*

Proof. Just compute directly by definition.

$$(f \circ g) \circ h(x) = (f \circ g)(h(x)) = f(g(h(x))) = f((g \circ h)(x)) = f \circ (g \circ h)(x).$$

□

Graphically, map composition is obviously associative. Just look at this graph:



So we have associativity checked. What about other properties? Say, do we have commutativity? The answer is NO.

Example 2.1.13. Here are some examples.

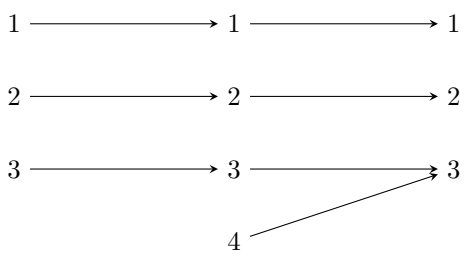
1. Say $f, g : \mathbb{R} \to \mathbb{R}$ with $f(x) = e^x, g(x) = 2x$. You can easily verify that $f \circ g$ is NOT the same as $g \circ f$, since $f \circ g(x) = e^{2x}$ while $g \circ f(x) = 2e^x$. Composition is NOT commutative.
2. Say $f, g : \mathbb{R} \to \mathbb{R}$ with $f(x) = 2x + 1, g(x) = 3x + 2$. Compute $f \circ g$ and $g \circ f$, what would you see? What if we change g into $g(x) = 3x + 1$?
3. What condition on the injectivity or surjectivity of f, g would guarantee that $f \circ g$ is injective? surjective? bijective? Find proofs and counter examples if you like.

☺

Example 2.1.14. Here are some food for thought. Try to prove the following statements yourself.

1. If f, g are both injective, then $f \circ g$ is injective.
2. If f, g are both surjective, then $f \circ g$ is surjective.
3. If f, g are both bijective, then $f \circ g$ is bijective.
4. If $f \circ g$ is injective, then g is injective. (But f may not be injective.)
5. If $f \circ g$ is surjective, then f is surjective. (But g may not be surjective.)
6. If $f \circ g$ is bijective, then g is injective and f is surjective. (But we know no more.)

For the last three statements, here is an example to think about.



☺

Remark 2.1.15. *Here we provide proof to two statements that will be useful later.*

If $f \circ g$ is surjective, then let us show that f is surjective.

The standard proof for surjectivity is like this: you pick an arbitrary element of the codomain, and then you go find a pre-image. That's it. (So everything has a pre-image.)

For each y in the codomain of f (which is also the codomain of $f \circ g$), we can find a pre-image for the map $f \circ g$, say x. Then $f(g(x)) = y$. But then $g(x)$ is a pre-image of y for the map f. So we have found a pre-image for arbitrary y. So f is surjective.

If $f \circ g$ is injective, then let us show that g is injective.

The standard proof for injectivity is like this: you show that any two inputs with the same image must actually be the same input. (So different inputs must go to different images.)

Suppose $g(x) = g(y)$ for two inputs x, y, and our goal is to prove $x = y$. Now we apply f to both sides of $g(x) = g(y)$, and we have $f \circ g(x) = f \circ g(y)$. But since $f \circ g$ is injective, $f \circ g(x) = f \circ g(y)$ means that $x = y$. This is what we want, so we are done! This shows that g is injective.

Now, in the example above, we see that if $f \circ g$ is bijective, then we can NOT guarantee that f, g are bijective. However, in some special cases, this can be done.

Proposition 2.1.16. *If X, Y are finite sets with the same amount of element, then any map $f : X \to Y$ is injective iff surjective iff bijective.*

Proof. Suppose X, Y both have n elements.

Suppose f is injective. Let elements of X be $x_1, \ldots, x_n$, and let $y_i = f(x_i)$. Since f is injective, and $x_1, \ldots, x_n$ are all distinct, therefore their images $y_1, \ldots, y_n$ are all distinct. But Y only has n elements! Therefore $Y = \{y_1, \ldots, y_n\}$, and f is surjective.

Now suppose f is surjective. Let elements of Y be $y_1, \ldots, y_n$, and for each y_i pick any pre-image x_i (which we can do because f is surjective). Now, since $x_1, \ldots, x_n$ all goes to distinct images, they must themselves be distinct elements. But X only has n elements! Therefore $X = \{x_1, \ldots, x_n\}$, and f is injective.

In conclusion, f is injective iff surjective, and hence iff bijective. □

Remark 2.1.17. *The porposition above can also be thought about in an intuitively manner. Suppose X, Y have the same size. Then if f is not injective, then it will squeeze elements somehwere, hence its range will be smaller in size. So it cannot be surjective.*

Conversely, if f is not surjective, then it will have smaller range than domain. So something must squeeze. So it cannot be injective.

Corollary 2.1.18. *If X, Y, Z are finite sets, all with the same number of elements. Then for any $g : X \to Y$ and $f : Y \to Z$, if $f \circ g$ is bijective, then both f and g are bijective.*

Proof. If $f \circ g$ is bijective, then g is injective and f is surjective. But since X, Y, Z all have the same number of elements, g is injective iff bijective, and f is surjective iff bijective. So f, g are both bijective. □

Now that we have all the terminology, let us look at my favorite map. (I'm sure it will be your favorite as well.) It is going to be the ***identity map***.

Definition 2.1.19. *A map id_X is an identity map for the set X if its domain and codomain are both X, and we have $\mathrm{id}_X(x) = x$ always. Sometimes we write id for short if the underlying set is obvious.*

Proposition 2.1.20.

1. *The identity map is always bijective, and it is its own inverse map.*
2. *For any set, its identity map is UNIQUE.*
3. *$\mathrm{id} \circ g = g$ whenever the composition is defined, and $g \circ \mathrm{id} = g$ whenever the composition is defined.*
4. *If a map f satisfies the condition that $f \circ g = g$ whenever the composition is defined, or that $g \circ f = g$ whenever the composition is defined, then f must be the identity map for its domain. (This is a necessary and sufficient condition.)*

Proof. Only the last one is non-trivial. Suppose the domain of f is X and the codomain is Y. Pick $g = \mathrm{id}_X$, then $f \circ g$ is defined, and $f \circ g = g$ as required. However, since g is the identity map on X, we also have $f \circ g = f$. Hence $f = g$. Done. □

Proposition 2.1.21. *If we have a pair of maps $f : X \to Y$ and $g : Y \to X$, such that $f \circ g = \mathrm{id}_Y, g \circ f = \mathrm{id}_X$, then f, g are inverse of each other.*

Proof. If $f(x) = y$, then $g(y) = g \circ f(x) = \mathrm{id}_X(x) = x$.

If $g(y) = x$, then $f(x) = f \circ g(y) = \mathrm{id}_Y(y) = y$.

So $f(x) = y$ iff $g(y) = x$. □

Remark 2.1.22. *However, note that we need BOTH $f \circ g = \mathrm{id}_Y$ AND $g \circ f = \mathrm{id}_X$. Merely one of them is not enough.*

For example, $f : \{0\} \to \mathbb{R}$ such that $f(0) = 0$, and $g : \mathbb{R} \to \{0\}$ such that $g(x) = 0$ for all $x \in \mathbb{R}$. Then $g \circ f = \mathrm{id}_{\{0\}}$. Yet, f, g are not inverse of each other. (Because $f \circ g \neq \mathrm{id}_{\mathbb{R}}$).

The identity map is the nicest and most important map ever. In particular it is also linear.

Proposition 2.1.23. *Consider the identity map* $\mathrm{id} : \mathbb{R}^n \to \mathbb{R}^n$. *Then* id *is linear, and its matrix is the* $n \times n$ *matrix* $\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$. *(Here empty entries means zero.)*

Proof. Recall that the identity map simply does nothing to the input. So it is linear because for all $s, t \in \mathbb{R}$ and $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we have

$$\mathrm{id}(s\boldsymbol{a} + t\boldsymbol{b}) = s\boldsymbol{a} + t\boldsymbol{b} = s(\mathrm{id}(\boldsymbol{a})) + t(\mathrm{id}(\boldsymbol{b})).$$

Now let us find its matrix. Since $\mathrm{id}(\boldsymbol{e}_i) = \boldsymbol{e}_i$, we see that the i-th column is simply $\boldsymbol{e}_i$. This results in the pattern as described. □

2.2 A Single Linear Equation: Dot Product, Projection, Hyperplanes

Our goal is to understand the pre-image problem in linear algebra: given a known vector $\boldsymbol{b}$ and a matrix A, solve $\boldsymbol{x}$ from $A\boldsymbol{x} = \boldsymbol{b}$. (I.e., find a pre-image of $\boldsymbol{b}$ under the linear map A.)

We start by looking at the simplest case, when A has a single row. It turns out that there is a great geometric interpretation in this case.

To start, if we have $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$, then the codomain is really just $\mathbb{R}$, and $A\boldsymbol{x} = \boldsymbol{b}$ is now

$$\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = b.$$

Or in high school terms, it is the linear equation

$$a_1 x_1 + \cdots + a_n x_n = b.$$

Focus on the left hand side for now. Hey, this is something we are familiar with. In high school, we have talked about dot products (or scalar products) between vectors, and they look exactly like the left hand side!

Definition 2.2.1. *Given two vectors* $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$ *in* $\mathbb{R}^n$, *we define their* ***dot product*** *or* ***scalar product*** *to be* $\sum v_i w_i$.

Conventionally, we write $\boldsymbol{v}^{\mathrm{T}}$ to represent a "row vector" which is essentially $\boldsymbol{v}$ written horizontally. In particular, $\boldsymbol{v}^{\mathrm{T}}$ is a matrix with only one row. Then we see that $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{w}$. So the meaning of the matrix $\boldsymbol{v}^{\mathrm{T}}$ is simply "taking a dot product with $\boldsymbol{v}$". As a linear map, $\boldsymbol{v}^{\mathrm{T}}$ sends $\boldsymbol{w}$ to $\boldsymbol{v} \cdot \boldsymbol{w}$.

Proposition 2.2.2. *The dot product is bilinear, symmetric and positive definite.*
Bilinear means $(a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2) \cdot \boldsymbol{w} = a_1(\boldsymbol{v}_1 \cdot \boldsymbol{w}) + a_2(\boldsymbol{v}_2 \cdot \boldsymbol{w})$ *and* $\boldsymbol{w} \cdot (a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2) = a_1(\boldsymbol{w} \cdot \boldsymbol{v}_1) + a_2\boldsymbol{w} \cdot \boldsymbol{v}_2)$.
Symmetric means $\boldsymbol{v} \cdot \boldsymbol{w} = \boldsymbol{w} \cdot \boldsymbol{v}$. *(So using this, we only need one of the two bilinearity requirements.)*
Positive definite means $\boldsymbol{v} \cdot \boldsymbol{v} \geq 0$ *always, and* $\boldsymbol{v} \cdot \boldsymbol{v} = 0$ *iff* $\boldsymbol{v} = \boldsymbol{0}$.

Proof. Straightforward computation. □

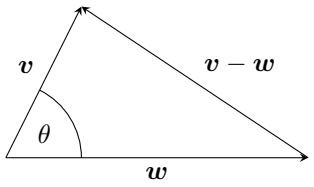
The idea of a dot product is super useful, because it is very geometric. In particular, we can use it to detect orthogonality and compute angles and projections. Here is some brief review + generalization of facts from high school.

Definition 2.2.3. *The **length** of a vector* $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ *is defined as* $\|\boldsymbol{v}\| := \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}}$*, or* $\sqrt{v_1^2 + \cdots + v_n^2}$ *in coordinates.*

*The **angle** between two vectors* $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ *is defined as* $\theta(\boldsymbol{v}, \boldsymbol{w}) = \arccos(\frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|})$*. We say* $\boldsymbol{v}$ *and* $\boldsymbol{w}$ *are **perpendicular** or **orthogonal** if the angle between them is* $\frac{\pi}{2}$*, or equivalently, if* $\boldsymbol{v} \cdot \boldsymbol{w} = 0$*.*

Remark 2.2.4. *We write the angle formula here as a definition for convenience, but technically one can deduce it instead. If you are curious, it goes like this.*

Given $\boldsymbol{v}, \boldsymbol{w}$*, then together with the vector* $\boldsymbol{v} - \boldsymbol{w}$*, they form a triangle.*



So by the cosine law, we know that

$$\|\boldsymbol{v} - \boldsymbol{w}\|^2 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - 2\|\boldsymbol{v}\|\|\boldsymbol{w}\| \cos\theta.$$

Therefore we have

$$\cos\theta = \frac{\|\boldsymbol{v} - \boldsymbol{w}\|^2 - \|\boldsymbol{v}\|^2 - \|\boldsymbol{w}\|^2}{2\|\boldsymbol{v}\|\|\boldsymbol{w}\|} = \frac{(\boldsymbol{v} - \boldsymbol{w}) \cdot (\boldsymbol{v} - \boldsymbol{w}) - \boldsymbol{v} \cdot \boldsymbol{v} - \boldsymbol{w} \cdot \boldsymbol{w}}{2\|\boldsymbol{v}\|\|\boldsymbol{w}\|}.$$

Now simplify the numerator and we are done.

Note that the angle formula is very revealing. We always have $\cos\theta = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|} = (\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}) \cdot (\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|})$. In particular, given two unit vectors, their dot product is cosine of the angle between them. It reaches maximum value of 1 when the two vectors are in the same direction, and it reaches minimum value -1 when the two vectors are in the opposite direction.

For a general intuition, you might interpret dot product as a measurement of "agreeness". If two vector "agree" with each other a lot, then the dot product is big. ($\boldsymbol{v} \cdot \boldsymbol{w}$ is large when the two vectors are in similar directions.) If two vector "disagree" with each other a lot, then the dot product is very negative. ($\boldsymbol{v} \cdot \boldsymbol{w}$ is very negative when the two vectors are in almost opposite directions.)

Example 2.2.5. This is a purely computational example. Skip it if you feel like it.

Consider $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$. They each has length $\sqrt{2}$, and their dot product is 1. So their angle θ satisfies $\cos\theta = \frac{1}{2}$. So you can now compute this angle easily. Draw them out to see better. (Also note that these two vector form a triangle with $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$, and this is an equilateral triangle with all sides of length $\sqrt{2}$. So the angle is exactly as you would expect.)

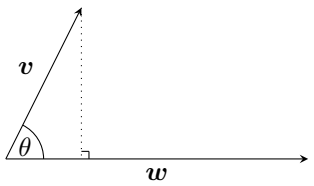
Now look at $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. By doing a dot product, we get zero. So they are perpendicular. Draw them out to see better. (Also note that these two vector form a triangle with $\begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix}$, and the three edges satisfy Pythagorean theorem. So this is a right triangle, and the angle is exactly as you would expect.) ☺

Now, since dot product gives us the angle, they also help us with computing projections.

Proposition 2.2.6. *Given two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$, where $\boldsymbol{w} \neq \mathbf{0}$ (because $\mathbf{0}$ gives no direction for projection....), then we can perform the projection of $\boldsymbol{v}$ onto the direction of $\boldsymbol{w}$. The signed-length of this projection is $\frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{w}\|}$. (Or more neatly as $\boldsymbol{v} \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, where $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ is the unit vector in the direction of $\boldsymbol{w}$.)*

(Here signed-length means this: a positive length means the projected result has the same direction as $\boldsymbol{w}$, while a negative length means the projected result has the opposite direction as $\boldsymbol{w}$.)

Proof. Let the angle between $\boldsymbol{v}, \boldsymbol{w}$ be θ. Then $\cos\theta = \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$.



By basic trigonometry, the signed-length is $\|\boldsymbol{v}\|$ times the cosine of this angle. So we have

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

So we are done. Now, is there a problem with this proof? OOPS! We are assuming $\|\boldsymbol{v}\| \neq 0$ in the above process (see the denominator), which is not guaranteed. So we have to do it separately.

IF we do have $\|\boldsymbol{v}\| = 0$, then $\boldsymbol{v} = \mathbf{0}$, and hence the projection result would have zero lengt. So the formula is still correct. Now we are finally done. □

Corollary 2.2.7. *Given two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$, where $\boldsymbol{w} \neq \mathbf{0}$ (because $\mathbf{0}$ gives no direction for projection....), then we can perform the projection of $\boldsymbol{v}$ onto the direction of $\boldsymbol{w}$. The resulting vector is $\frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\boldsymbol{w}\cdot\boldsymbol{w}}\boldsymbol{w}$.*

Proof. The resulting vector should be the signed-length $\frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{w}\|}$ times the unit vector in the direction of $\boldsymbol{w}$, i.e., $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$. The result is $\frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\boldsymbol{w}$, which is as desired. □

Now, how are these related to linear algebra, and in particular, to row vectors? Well, lo and behold.

Corollary 2.2.8. *Consider a "row vector" $\boldsymbol{u}^{\mathrm{T}}$ such that $\boldsymbol{u}$ is a unit vector (i.e., length one). Then $\boldsymbol{u}^{\mathrm{T}}$ as a linear map would send each $\boldsymbol{v}$ to the (signed-length of the) projection of $\boldsymbol{v}$ to the direction of $\boldsymbol{u}$.*

Proof. Let $\boldsymbol{u} = \begin{bmatrix} a \\ b \end{bmatrix}$. So $\|\boldsymbol{u}\| = 1$. Then the signed length of the projection of $\boldsymbol{v}$ to $\boldsymbol{u}$ is

$$\frac{\boldsymbol{v}\cdot\boldsymbol{u}}{\|\boldsymbol{u}\|} = \boldsymbol{v}\cdot\boldsymbol{u} = v_1 a + v_2 b = \begin{bmatrix} a & b \end{bmatrix}\boldsymbol{v}.$$

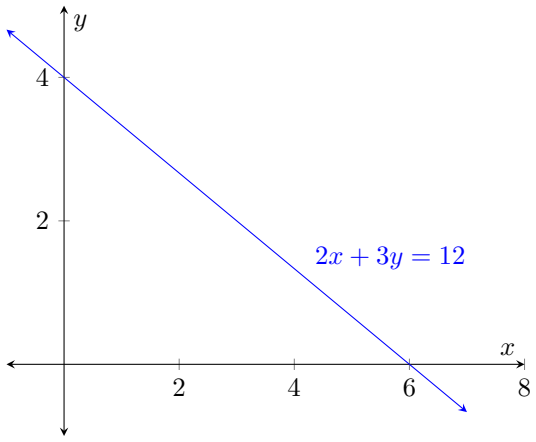
□

So, if $\boldsymbol{u}$ is a unit vector, then $\boldsymbol{u}^{\mathrm{T}}$ is just the (signed-length of the) projection to $\boldsymbol{u}$. If $\boldsymbol{u}$ is not a unit vector, then it is a multiple of a unit vector. Hence $\boldsymbol{u}^{\mathrm{T}}$ is a multiple of the (signed-length of the) projection to $\boldsymbol{u}$. So, as you can see, all row vectors should be interpreted as multiples of a (signed-length of a) projection.

Now we move on to the reverse process. Suppose A is a matrix with a single row, i.e., a multiple of a projection. Given $b \in \mathbb{R}$, how to solve for $\boldsymbol{x}$ from the equation $A\boldsymbol{x} = b$?

Consider the following two problems.

Example 2.2.9. Suppose I have 12 yuan in my wallet, and I intend to spend them all. I can buy apple with 2 yuan each, or banana with 3 yuan each. What sorts of linear combinations can I buy?

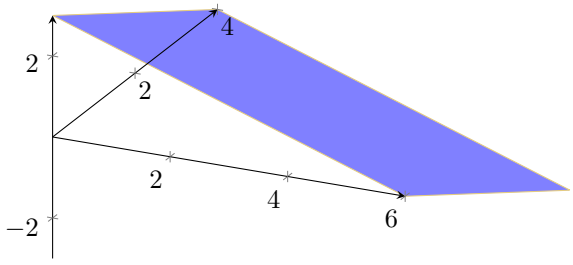


Well, suppose I buy x apples and y bananas. Then they should satisfy a constraint of $2x + 3y = 12$. The set of all possible solutions here, treated as points on the plane with coordinates (x, y), is a straight line in $\mathbb{R}^2$ going through the point $\begin{bmatrix}6\\0\end{bmatrix}$ and $\begin{bmatrix}0\\4\end{bmatrix}$.

Now let us revisit this equation $2x + 3y = 12$. Note that the left hand side is actually the dot product between two vectors. So in this sense the equation can be rewritten as $\begin{bmatrix}2\\3\end{bmatrix} \cdot \begin{bmatrix}x\\y\end{bmatrix} = 12$.

Alternatively, using the matrix notation that we have know of, we can also write $\begin{bmatrix}2 & 3\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = 12$. ☺

Example 2.2.10. Suppose now we can also buy pears with 4 yuan each. Then we have a price vector $\boldsymbol{p} = \begin{bmatrix}2\\3\\4\end{bmatrix}$, and the linear combinations that we can purchase could be any solution $\boldsymbol{x}$ to the equation $\boldsymbol{p} \cdot \boldsymbol{x} = 12$. Or alternatively, writing $\boldsymbol{p}$ as a row vector $\boldsymbol{p}^{\mathrm{T}}$, we want solutions to $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{x} = 12$.

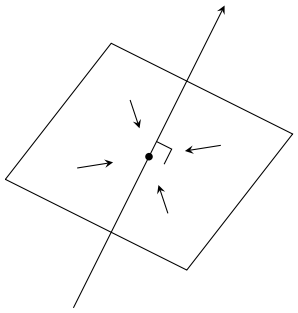


This would be a plane in $\mathbb{R}^3$ with equation $2x + 3y + 4z = 12$. It is the plane that went through points $\begin{bmatrix}6\\0\\0\end{bmatrix}, \begin{bmatrix}0\\4\\0\end{bmatrix}, \begin{bmatrix}0\\0\\3\end{bmatrix}$. ☺

Either way, we are interested in the solution of the following problem: Given a vector $\boldsymbol{p} \in \mathbb{R}^n$ and a constant b, what are all possible solutions to the equation $\boldsymbol{p} \cdot \boldsymbol{x} = b$?

Example 2.2.11. What is the solution to $\boldsymbol{p} \cdot \boldsymbol{x} = b$?

Well, it means for any solution $\boldsymbol{x}$, its projection to $\boldsymbol{p}$ must be the vector $\frac{b}{\|\boldsymbol{p}\|^2}\boldsymbol{p}$, a fixed vector independent of $\boldsymbol{x}$. So geometrically, this is the collection of all vectors that got projected to the SAME PLACE in the direction of $\boldsymbol{p}$.



In $\mathbb{R}^2$, this would yield a line perpendicular to $\boldsymbol{p}$ and its distance to the origin is exactly $\frac{|b|}{\|\boldsymbol{p}\|}$. In $\mathbb{R}^3$, this would yield a plane perpendicular to $\boldsymbol{p}$ and its distance to the origin is exactly $\frac{|b|}{\|\boldsymbol{p}\|}$. You can easily guess the pattern here. ☺

On $\mathbb{R}^2$, we see that the solution is a line, while in $\mathbb{R}^3$ we see that the solution is a plane. In general, we may have the following definition.

Definition 2.2.12. *A subset of $\mathbb{R}^n$ is a called a **hyperplane** if it is the set of all possible solutions $\boldsymbol{x}$ for an equation $\boldsymbol{n} \cdot \boldsymbol{x} = b$ for fixed $\boldsymbol{n} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. We also require that $\boldsymbol{n} \neq \boldsymbol{0}$.*

*Sometimes we call any vector that is a non-zero multiple of $\boldsymbol{n}$ a **normal vector** to that hyperplane, and we say $\boldsymbol{n} \cdot \boldsymbol{x} = b$ to be the **linear equation** for this hyperplane. (Typically this looks like something similar to $ax + by + cz = d$ for constants a, b, c, d and unknowns x, y, z.)*

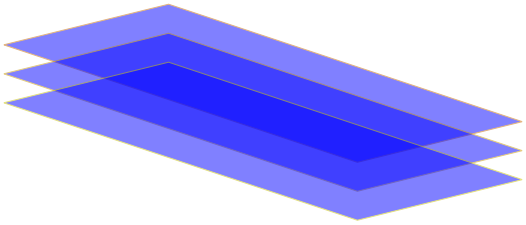
So basically we can define a hyperplane as the solution set with a single linear constraint (e.g., your budget b). In $\mathbb{R}^n$, we can intuitively see that this is some $n-1$ dimensional thingy that is perpendicular to the direction of $\boldsymbol{p}$ with a distance $\frac{d}{\|\boldsymbol{p}\|}$ to the origin.

Remark 2.2.13. *Now what does this perpendicularity mean? Intuitively, any "arrow" vector INSIDE this hyperplane should be perpendicular to $\boldsymbol{p}$. Take any points $\boldsymbol{x}_1, \boldsymbol{x}_2$ in this hyperplane, then $\boldsymbol{x}_1 - \boldsymbol{x}_2$ would represent an arrow vector from $\boldsymbol{x}_2$ to $\boldsymbol{x}_1$ inside this hyperplane.*

Here we see that if $\boldsymbol{x}_1, \boldsymbol{x}_2$ are both solutions to $\boldsymbol{p} \cdot \boldsymbol{x} = b$, then we necessarily have $\boldsymbol{p} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$. See? Perpendicularity.

Next let us explore some relations among these hyperplanes.

Example 2.2.14. Suppose we fix $\boldsymbol{p}$ and vary b. How would the hyperplane $\boldsymbol{p} \cdot \boldsymbol{x} = b$ changes?



Draw $2x + 3y + 4z = 12, 13, 14, 15$, and you shall see that they are all parallel planes in $\mathbb{R}^3$, since they must all be perpendicular to the same $\boldsymbol{p}$. In particular, in our case, the larger is b, the more the plane will move in the direction of $\boldsymbol{p}$. ☺

Example 2.2.15. Suppose we fix b and double $\boldsymbol{p}$. How would the hyperplane $\boldsymbol{p} \cdot \boldsymbol{x} = b$ changes?

Well, note that $(2\boldsymbol{p}) \cdot \boldsymbol{x} = b$ iff $\boldsymbol{p} \cdot \boldsymbol{x} = \frac{1}{2}b$. So it is as if we never changed $\boldsymbol{p}$, and simply change b into half of b. ☺

Example 2.2.16. Suppose we have two hyperplanes given by $\boldsymbol{p} \cdot \boldsymbol{x} = b$ and $\boldsymbol{q} \cdot \boldsymbol{x} = c$. Then one might simply add the two equations and get $(\boldsymbol{p} + \boldsymbol{q}) \cdot \boldsymbol{x} = b + c$. Are the three hyperplanes related?

Consider the case of planes $x = 1$ and $y = 2$ in $\mathbb{R}^3$. These two plane intersect at a line perpendicular to the xy-plane and hit the xy-plane at $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$. Points on this line typically would have coordinates $\begin{bmatrix} 1 \\ 2 \\ t \end{bmatrix}$ for some $t \in \mathbb{R}$.

Now let us add the two plane equations and get $x + y = 3$. Immediately one may check that this plane still contains the same line! The planes $x = 1, y = 2$ and $x + y = 3$ have a line as their common intersection!

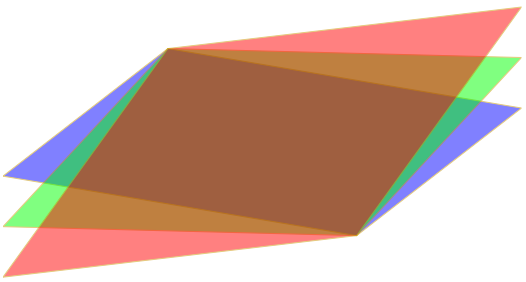


Figure 2.2.1: red equation + blue equation = green equation

In general, if you have one linear equation, then adding another linear equation to it is like rotating (or more precisely, shearing) the hyperplane along the common intersection. See if you can prove this yourself. Alternatively, you can think of adding two linear equations as taking some "weighted averate" of the two planes.

So try to see if you can answer the following question: Suppose we fixed b and all but the first coordinate of $\boldsymbol{p}$. Say we increase the first coordinate of $\boldsymbol{p}$ by 1. How would the hyper plane $\boldsymbol{p} \cdot \boldsymbol{x} = b$ change? Around which intersection should it rotate? ☺

2.3 Row and Column View of Linear System

This section is NOT about solving a linear system. Rather, it is about understanding the phenomena of a linear system. You do not learn HOW to solve linear systems in this section. Your expectation should be to open your mind and connect some intuitions here.

Recall that, a map $f : \mathbb{R}^n \to \mathbb{R}^m$ is called a linear map if $f(\sum a_i \boldsymbol{v}_i) = \sum a_i f(\boldsymbol{v}_i)$. Our whole class is devoted to the study of such maps.

Also recall that any linear map could be written as a matrix, by studying how the standard basis goes via this map.

Think now. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, the domain is $\mathbb{R}^n$, and the standard basis would be $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$. So the matrix of f would be $\begin{bmatrix} f(\boldsymbol{e}_1) & \ldots & f(\boldsymbol{e}_n) \end{bmatrix}$. Note how we have n columns. The number of columns is ALWAYS the dimension of domain in this case.

What about rows? Each $f(\boldsymbol{e}_i)$ would be an element of the codomain $\mathbb{R}^m$, so it has m coordinates or m "rows". So in total, we see that $\begin{bmatrix} f(\boldsymbol{e}_1) & \ldots & f(\boldsymbol{e}_n) \end{bmatrix}$ has m rows and n columns. We say that this is an m by n matrix or $m \times n$ matrix.

Given an $m \times n$ matrix A and a vector $\boldsymbol{b} \in \mathbb{R}^m$ (Think: why not $\mathbb{R}^n$?), one might want to tries to find all solutions $\boldsymbol{x}$ to the system $A\boldsymbol{x} = \boldsymbol{b}$. Previously we have already seen this, as in the hen and rabbit problem. However, let us now look at this linear system with some fresh perspectives.

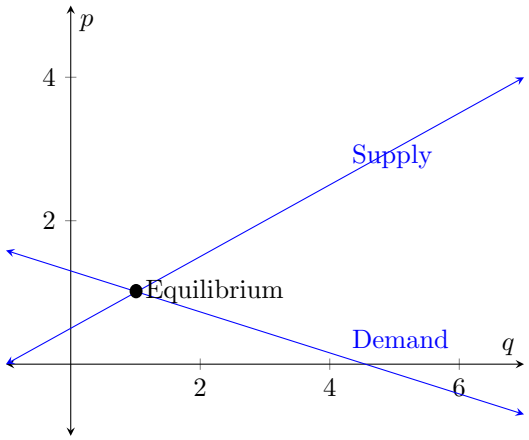
What is the MOST important feature of a matrix? Well, it has rows and columns. In fact, this seems to be the whole point of a matrix: to write numbers in rows and columns. As a result, there are always TWO ways to view a system of linear equations.

Example 2.3.1. Taking the row view, each row of $A\boldsymbol{x} = \boldsymbol{b}$ is a single linear equation, and thus corresponds to a hyperplane. We want ALL equations to hold simultaneously, so we are seeking the INTERSECTION of all these hyperplanes.

For example, in the beef market, the more expansive is the price, the less likely customers will buy beef. According to a 1924 paper in the Journal of Farm Economics (by Schultz), if the price is p, then the demand for beef is roughly $q = 1.3 - \frac{2}{7}p$. On the other hand, the higher the beef price, the more people would start supplying beef. I did not find available data here, but let us say the supply is $q = 0.5 + \frac{1}{2}p$. Note that I have used the same symbol q for both supply and demand, since they both equal to the amount of trades of beef.

So we have the following equations that must simultaneously hold.

$$\begin{cases} q + \frac{2}{7}p = 1.3, \\ q - \frac{1}{2}p = 0.5. \end{cases}$$



Or in matrix terminology, it means $\begin{bmatrix} 1 & \frac{2}{7} \\ 1 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} 1.3 \\ 0.5 \end{bmatrix}$.

Each equation (or each row) would represent a line in the plane. Graph the two lines, and their intersection point determines the amount of trade and the price of the beef. ☺

Example 2.3.2. To see the column view, review the Hen-Rabbit problem, Example 1.2.18. It boils down to the following equation:

$$\begin{cases} x + y = 6, \\ 2x + 4y = 20. \end{cases}$$

Or in matrix terminology, it means $\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 20 \end{bmatrix}$. We may also write it as $x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 20 \end{bmatrix}$.

The two columns of our matrix represent a hen and a rabbit respectively. (Btw, the two rows represent heads and legs respectively.)

In the column view, we are asking ourselves what linear combination would yield the result, while in the row view we are trying to find intersections of hyperplanes. For this specific problem, the column view is slightly more intuitive.

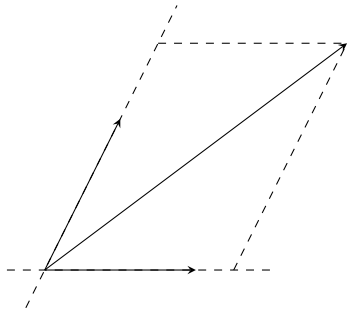


Figure 2.3.1: This is just a schematic picture. Things are not to scale.

Intuitively, we are trying to use a "mixture" of $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 4 \end{bmatrix}$ to get to $\begin{bmatrix} 6 \\ 20 \end{bmatrix}$. As you can see from a graph, too much $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ in the mixture and your resulting arrow will be too low, while too much $\begin{bmatrix} 1 \\ 4 \end{bmatrix}$ in the mixture and your resulting arrow will be too high.

A geometric solution would be like this: draw the arrow $\begin{bmatrix} 6 \\ 20 \end{bmatrix}$, and from the end point of this arrow, draw

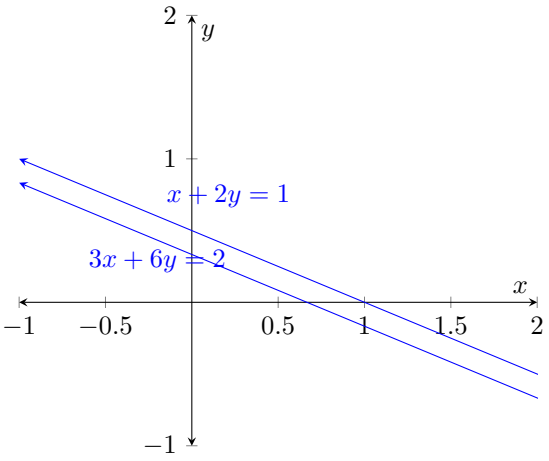
lines parallel to $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\4\end{bmatrix}$, as show in the figure above. We should be able to figure out the answer now using this parallelogram.

Let us do the same process algebraically. Suppose the correct answer requires x hens. Then we want $\begin{bmatrix}6\\20\end{bmatrix} - x\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}6-x\\20-2x\end{bmatrix}$ to be parallel to $\begin{bmatrix}1\\4\end{bmatrix}$. In particular, we want $20-2x$ to be four times of $6-x$. From this we can solve for a solution.

Note that we are trying to figure out how much hens we need, so that $\begin{bmatrix}6-x\\20-2x\end{bmatrix}$ is parallel to the rabbit vector. In primary school, we call this method "pretent the animals are all rabbits, and see how much discrepancy we have in the leg numbers." What we have seen above is simply that primary school method. ☺

Ideally, we always hope that our system would have a unique solution. However, this is not always possible. Now, with the visual picture in mind, let us look at the following linear system, which is weirdly unsolvable.

Example 2.3.3. Consider $\begin{bmatrix}1&2\\3&6\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}$. It turns out that there is no solution.



With the row view, you may graph out the two lines and try to find intersection. However, you shall see that the two lines are parallel. So there is no intersection.

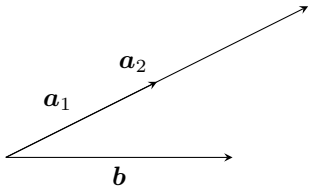


Figure 2.3.2: This is just a schematic picture, not drawn to scale.

With the column view, you may see that the two column vectors are actually ***colinear***, i.e., they are on the same line. As you can imagine, any linear combination of them must STAY on this line. Since $\begin{bmatrix}1\\2\end{bmatrix}$ is not on this line, this is impossible.

In high school, we would do the following. We realize that the first equation tripled is $3x+6y=3$, while the second equation is $3x+6y=2$, contradiction. The first-row-second-row ratio is different on the left side

and on the right side. In the row picture, it means the two planes are not coinciding. In the column picture, it means the slopes are off. ☺

Here is another situation, where we have too many solutions.

Example 2.3.4. Consider $\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 2\end{bmatrix}\begin{bmatrix}x\\ y\\ z\end{bmatrix} = \begin{bmatrix}2\\ 3\end{bmatrix}$. It turns out that there are infinitely many solutions. And if you tries to solve it the high school way, then you will fail to find a unique answer, because you don't have enough equations.

With the row view, things happen in $\mathbb{R}^3$. You may graph out the two planes in the space, and try to find intersection. However, the two planes intersect in a line. So everything on that line is a solution.

To find this line, since the line lies on both planes, it is orthogonal to both normal vectors, i.e., it is orthogonal to both $\begin{bmatrix}1\\ 0\\ 1\end{bmatrix}$ and $\begin{bmatrix}0\\ 1\\ 2\end{bmatrix}$. So, this is a line parallel to $\begin{bmatrix}1\\ 2\\ -1\end{bmatrix}$. (You can compute this via cross product. We do not really need it in this course, but feel free to search online and learn about it.)

You can also verify that the line goes through $\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$. If you like, you may now write your solution in parametrized form, i.e., $\begin{bmatrix}1\\ 1\\ 1\end{bmatrix} + \begin{bmatrix}1\\ 2\\ -1\end{bmatrix} t$ for any $t \in \mathbb{R}$. Alternatively, you may also write $\begin{bmatrix}1+t\\ 1+2t\\ 1-t\end{bmatrix}$ or, more aligned with high school tradition, as

$$\begin{cases} x = 1+t,\\ y = 1+2t,\\ z = 1-t. \end{cases}$$

With the column view, things happen now on the plane $\mathbb{R}^2$. We have have three vectors in $\mathbb{R}^2$ trying to combine into $\begin{bmatrix}2\\ 3\end{bmatrix}$, but in fact the first two would be enough. (In fact any two of the three would be enough.) We do not really need the third vector.

In fact, let us say that x, y, z are the coordinates of the solution, so $x\begin{bmatrix}1\\ 0\end{bmatrix} + y\begin{bmatrix}0\\ 1\end{bmatrix} + z\begin{bmatrix}1\\ 2\end{bmatrix} = \begin{bmatrix}2\\ 3\end{bmatrix}$. Now, if we have already found this solution, we can get naughty and try to increase z by one unit, and simultaneously decrease x by one unit and y by two units. Then the left hand side would remain unchanged after all.

In fact, I can change z by an arbitrary amound, and simply change x, y accordingly, and therefore get infinitely many solutions. Essentially, since the three columns have a relation $\begin{bmatrix}1\\ 0\end{bmatrix} + 2\begin{bmatrix}0\\ 1\end{bmatrix} - \begin{bmatrix}1\\ 2\end{bmatrix} = 0$, we see that decreasing z by t while increasing x by t and y by $2t$ will always leave the left hand side unchanged.

In effect, I have found a direction to move freely within the solution set. Given any solution, say $\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$, then I can move an arbitrary amount along the direction $\begin{bmatrix}1\\ 2\\ -1\end{bmatrix}$, and we would still be in the solution set. So again, we have found our solution set to be $\begin{bmatrix}1+t\\ 1+2t\\ 1-t\end{bmatrix}$ for arbitrary $t \in \mathbb{R}$. ☺

Here I want to draw your attention to the following phenomena, which underlies both anomalies.

Definition 2.3.5. *A collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ in $\mathbb{R}^n$ is called **linearly dependent** if one is a linear combination of the others, or if one of them is the zero vector. Otherwise, the collection is **linearly independent**.*

(Conventionally, a single non-zero vector is linearly independent. And any collection that contains a zero vector is linearly dependent.)

Example 2.3.6. Consider

$$\begin{cases} x \ \ +y = 1, \\ x+2y = 2, \\ 2x+3y = 3. \end{cases}$$

Note that the third equation is a linear combination of the first two (it is the sum of the first two). It is REDUNDANT!

Now consider

$$\begin{cases} x \ \ +y = 1, \\ x+2y = 2, \\ 2x+3y = 4. \end{cases}$$

Here we have an inconsistency. The third equation is NOT a linear combination of the previous two, but its left hand side is. In particular, the left hand side and the right hand side of the third equation are NOT the same linear combinations of the corresponding sides of the previous two equations! This creates a contradiction.

All in all, consider $A\boldsymbol{x} = \boldsymbol{b}$. If rows of A are linearly dependent, we will want to check the right hand side $\boldsymbol{b}$ and see if it follows the same linear dependency. If they have the same linear dependency, we have redundant equations. If they have distinct dependencies, then we have no solution to our system. ☺

Example 2.3.7. Again consider $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$. This is the case of $A\boldsymbol{x} = \boldsymbol{b}$ where columns of A are now linearly dependent.

Note that rows of A are describing the equations. What does columns of A describe? Well, they are each coefficients for specific variables. For example, the first column of A are all the coefficients for your first variable, and the second column for the second variable, etc..

So just as dependencies among rows usually signifies a redundancy in equations, a dependency among columns signifies a redundancies in variables. We don't actually need these much variables. In our case, since the third column is a linear combination of the first two, we can in fact set z to be ANYTHING, and work out the first two variables to compensate for that.

So if a column is a linear combination of others, then the corresponding variable is FREE! Specifically in our example, say if you want one extra copy of the third column, you can just use one less first column and two less second column.

Note that you may have no solution to start with. Say $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. Here you can freely choose the seond variable and compensate with the first vairable. HAD we a solution, then we can use one more copies of the second column by using two less first column. But it does not matter in the end, since there is no solution to start with.

Having column dependencies means you have freedom to move around your solutions. However, if you have no solution to begin with, then there is no where to start your moving from. (But if you can find a single solution, then this freedom guarantees infinitely many solutions.)

So a linear dependency among columns of A means you either have infinitely many solutions, or maybe you have no solutions. ☺

After Gaussian elimination we shall prove the following with the help of rank. For now just let me spoil the answer in advance.

1. A matrix A as a linear map is injective iff columns of A are linearly independent. In this case, a linear system $A\boldsymbol{x} = \boldsymbol{b}$ has either a UNIQUE solution, or there is no solution. (I.e., the system has at most one solution.)

2. A matrix A as a linear map is surjective iff rows of A are linearly independent. In this case, a linear system $A\boldsymbol{x} = \boldsymbol{b}$ always have at least one solution.

3. A matrix A as a linear map is bijective iff both columns and rows are linearly independent. In this case, a linear system $A\boldsymbol{x} = \boldsymbol{b}$ always have a UNIQUE solution.

2.4 Gaussian Elimination

This section is about solving a linear system.

Example 2.4.1. Let us revisit the hen rabbit problem. Say we want to solve $\begin{bmatrix}1 & 1\\ 2 & 4\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}6\\ 20\end{bmatrix}$. What should we do?

Well, the high school way is to solve by substitution. However, there is a much cooler way to do this.

Imagine that all hens and rabbits are well-trained. If they hear a whistle, they will all raise a leg. Now, we started with 6 heads and 20 legs. I whistle twice, now I have 6 heads but only 8 standing legs. But now all hens have no leg standing! All the hens must now fall on their butt, and only rabbits are left with two leg standing each. So 8 divided by 2 is the number of rabbits, 4. We have 2 hens and 4 rabbits.

This is the process of Gaussian elimination. ☺

For the ease of use, let us define the following notion.

Definition 2.4.2. *For a linear system $A\boldsymbol{x} = \boldsymbol{b}$, we say A is the **coefficient matrix** of the system, and we say $\begin{bmatrix}A & \boldsymbol{b}\end{bmatrix}$ is the **augmented matrix** of the system.*

Here the notation means we write A, but then put in $\boldsymbol{b}$ as the last column. Basically the augmented matrix is just the same as writing down the whole system, but we get too lazy to write the variables and the equality symbles and so on.

Now, in the process above, we started with the system $\begin{bmatrix}1 & 1 & 6\\ 2 & 4 & 20\end{bmatrix}$. When I whistle twice, essentially I am substracting twice the first row from the second row (we also write $r_2 \to r_2 - 2r_1$), and we get a new system $\begin{bmatrix}1 & 1 & 6\\ 0 & 2 & 8\end{bmatrix}$. (In high school terms, we are substracting multiples of one equation from the other.) However, despite being a new system, note that the solution set is the SAME as before! This perserving of solutions is the very fundation of why we did this.

Then we divide the second row by 2 (we also write $r_2 \to \frac{1}{2}r_2$), we have $\begin{bmatrix}1 & 1 & 6\\ 0 & 1 & 4\end{bmatrix}$. Pay attention now. The second row reads exactly $y = 4$. We have solved one variable.

Next we want to solve for x from the first equation. We substract the second row from the first row ($r_1 \to r_1 - r_2$), and we have $\begin{bmatrix}1 & 0 & 2\\ 0 & 1 & 4\end{bmatrix}$. Now the two lines reads $x = 2, y = 4$, which is the solution.

They key lies in the following operations, called ***elementary row operations***.

1. (Swapping) Swapping two rows. (I.e., swapping two equations.) We write $r_i \leftrightarrow r_j$.

2. (Scaling) Multiplying a row by the same scalar. (I.e., multiplying both sides of an equation by the same number.) We write $r_i \to kr_i$.

3. (Shearing) Adding a multiple of one row to another row. (I.e., adding multiples of one equation to another.) We write $r_j \to r_j + kr_i$.

(Note that, geometrically, the first row operation only re-order the hyperplanes. The second one does not change the hyperplanes. And the third one is "rotating" hyperplanes around their common intersections.)

The magic of these three operations is that they preserve solutions.

Proposition 2.4.3. *The elementary row operations on the augmented matrix would preserve the solution set.*

Proof. Swapping means simply switch the places of two equations. And this obviously has no effect on the solution set. Similarly, Scaling means simply multiplying both sides of a particular equation by some non-zero number. And this obviously has no effect on the solution set as well.

For shearing, here are two kinds of arguments. First is geometrical. Recall that changing r_j to $r_j + kr_i$ means we rotate the hyperplane for the equation r_j along the intersection of r_i and r_j. This will definitely preserve the intersection, which is what we actually need to find the solution set. Hence the solution set is preserved.

Alternatively, here is a more algebraic argument. For simplicity, consider the system with equations $r_i, r_j, r_j + kr_i$. Here the last equation $r_j + kr_i$ is obviously equivalent. So the system of $r_i, r_j, r_j + kr_i$ has the same solution set as the system of r_i, r_j.

However, look at $r_i, r_j, r_j + kr_i$ again. Now, we do NOT treat $r_j + kr_i$ as redundant. Rather, we treat r_j is redundant, since it can also be expressed as linear combination of others. I.e., $r_j = (r_j + kr_i) - kr_i$. Hence throwing away the redundant equation r_j, we see that $r_i, r_j, r_j + kr_i$ also has the same solution set as the system of $r_i, r_j + kr_i$.

Therefore, the system of r_i, r_j has the same solution set as the system of $r_i, r_j + kr_i$. □

Whatever argument you use, there is a really important hidden part of the proof: all elementary operations can be reversed. The inverse operation of $r_i \leftrightarrow r_j$ is itself, the inverse operation of $r_i \to kr_i$ is $r_i \to = \frac{1}{k} r_i$, and the inverse operation of $r_j \to r_j + kr_i$ is $r_j \to r_j - kr_i$.

So, whatever row operations you just did, you can always go back. This is vital for the preservation of the solution set.

Example 2.4.4. Consider the ILLEGAL row operation $r_j \to kr_j$ where $k = 0$. Since 0 has no inverse, this operation CANNOT be reversed. It actually would fail to preserve the solution set. For example, the solution set to the linear system of $x = 1$ (which is a hyperplane) is NOT the same as the solution set to the linear system $0 = 0$ (which is the whole space). ☺

So, ideally, if you have a linear system, you basically just use these three operations in whatever order necessary, until we reached the solution.

Remark 2.4.5. *One can similarly define elementary column operations. What would that mean?*

Well, row operations acts on equations. Column operations acts on variables. An elementary column operation would be equivalent to a change of variable.

For example, say we have a system $x + 2y = 3, x + 3y = 4$. If we substract the first column from the second column, this would mean we are doing a change of variable $x' = x + y, y' = y$, and now we have $x' + y' = 3, x' + 2y' = 4$ as desired.

In theory one can also solve equations by keep changing variables. However, it is less efficient, because once you solved the new variables, you have to then retrace and try to resolve the old variable. So as far as linear systems are concerned, we prefer row operations.

Now, given a system of linear equations $\begin{bmatrix} A & \boldsymbol{b} \end{bmatrix}$, when is the end game? When would we consider it solved?

Example 2.4.6. For the hen-rabbit problem, we have $\left[\begin{array}{cc|c} 1 & 1 & 6 \\ 2 & 4 & 20 \end{array}\right]$. We eventually got to $\left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & 4 \end{array}\right]$

Now the two equations reads $x = 2, y = 4$. So we are done. We have an unique solution. ☺

So ideally, we hope to make the coefficient matrix into what we call an identity matrix.

Definition 2.4.7. *The $n \times n$ identity matrix I_n or simply I is a matrix whose diagonal entries are all 1, and non-diagonal entries are all zero.*

So basically somethings like $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Looking at the matrix, it does not look like much. But looking at its column, we see that $I = \begin{bmatrix} \boldsymbol{e}_1 & \dots & \boldsymbol{e}_n \end{bmatrix}$. So it is basically a linear map that sends all basis vectors to themselves! In particular, it will send every vector to itself.

So, you can see why we call it the identity matrix. As a linear map, it is the identity map.

We love the identity matrix for many many reasons. To list a few,

1. For $n \times n$ identity matrix I, we have $I\boldsymbol{v} = \boldsymbol{v}$ for all vector $\boldsymbol{v} \in \mathbb{R}^n$.
2. As a linear map, I is bijective and it is its own inverse. (Since it is the identity map.)
3. Suppose we have a augmented matrix for a linear system $\begin{bmatrix} A \mid \boldsymbol{b} \end{bmatrix}$, and ideally we would have no redundant equations, no free variables, so we can just solve all variables and it would be unique. What is the end game then? Then end game would be like $x_1 = SomeNumber, x_2 = SomeOtherNumber$, etc., and in matrix form it would look like $\begin{bmatrix} I \mid AnswerVector \end{bmatrix}$. So the identity matrix is exactly what we want for the left hand side of our matrix. Also note that $\begin{bmatrix} I \mid AnswerVector \end{bmatrix}$ is automatically in RREF.

Now, this might not be possible sometimes. Consider the following example.

Example 2.4.8. Suppose we have a system $\begin{bmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$. We want to reduce the number of variables in the equations, i.e., we want as many zeroes as possible in the coefficient matrix. So let us first kill the w variable in the second equation by doing $r_2 \to r_2 - r_1$. This gives the system $\begin{bmatrix} 1 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$.

As you can see now, the second equation probably cannot simplify any further. So instead, let us use this to simplify the first equation as much as possible. Doing $r_1 \to r_1 - r_2$, we get another zero entry with $\begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.

Now the two equations reads $w + x + z = 2, y + z = 2$. For each equation, if we move all but the first variable to the right side of the equation, we have $w = 2 - x - z, y = 2 - z$.

It is now reasonably clear that we can let x, z be whatever values they like, and then deduce the corresponding w and y from it. If you like, you can also write the solution set as $\{ \begin{bmatrix} 2 - x - z \\ x \\ 2 - z \\ z \end{bmatrix} : x, z \in \mathbb{R} \}$.

☺

So, what we should do is to use all those row operations, and try to simplify the equation by introducing as much zeroes as possible. Then, hopefully we can choose some "free variables", and the other variables can just depend on them, i.e., "dependent variables".

Now, there is a special type of matrices where we can easily choose our free variables and dependent variables.

Definition 2.4.9. *We say a matrix is in* ***Row Echelon Form*** *if it satisfy the following property:*

1. *All zero rows are below all non-zero rows.*

2. *In each non-zero row, we call the first non-zero entry the* ***pivot of that row****. Then the pivots of lower rows are always to the right of previous (higher) rows.*

We say a matrix is in ***Reduced Row Echelon Form*** *if furthermore, all pivots are* 1*, and all entries above each pivot are zero.*

So a REF looks something like this:

$$\begin{bmatrix} \bullet & * & \cdots & * & * & * & \cdots & * & * & * & \cdots & * & \cdots\cdots & * & * & \cdots & * \\ & & & & \bullet & * & \cdots & * & * & * & \cdots & * & \cdots\cdots & * & * & \cdots & * \\ & & & & & & & & \bullet & * & \cdots & * & \cdots\cdots & * & * & \cdots & * \\ & & & & & & & & & & & & \ddots & \vdots & \vdots & & \vdots \\ & & & & & & & & & & & & & * & * & \cdots & * \\ & & & & & & & & & & & & & \bullet & * & \cdots & * \end{bmatrix},$$

And a RREF looks something like this:

$$\begin{bmatrix} 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots & 0 & * & \cdots & * \\ & & & & 1 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & 1 & * & \cdots & * & \cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & & & & & \ddots & \vdots & \vdots & & \vdots \\ & & & & & & & & & & & & & 0 & * & \cdots & * \\ & & & & & & & & & & & & & 1 & * & \cdots & * \end{bmatrix}.$$

Example 2.4.10. Consider the system $\begin{bmatrix} 1 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$. This is in row Echelon form.

The the first equation says w (the pivot location) has to depend on later variables in certain way. The second equation says that y (the pivot location) has to depend on later variables in certain way. And these are the ONLY requirements!

So there are only requirements on w and y, and no requirement on x, z. Hence we can pick arbitrary x, z, and get a solution. ☺

Example 2.4.11. Suppose our system is now $\begin{bmatrix} 1 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 5 \\ 0 \end{bmatrix}$. This is a reduced Echelon form. Note that the pivots are on the first, third and forth column.

Now the equations are like $x_1 + x_2 + 2x_5 = 4, x_3 + 3x_5 = 4, x_4 + 4x_5 = 5, 0 = 0$. By moving all variables other than the left-most one to the right side of the equation, we see that we can choose the x_2, x_5 to be the free variables. As you can see here, the dependent variables correspond exactly to columns with pivots, i.e., ***pivotal columns***, whereas the free variables corresponds to columns without pivots, i.e., ***free columns***.

The point is that, each pivot, as the left-most non-zero entry of the row, will corresponds to a dependent variable specified by this very row (which represents an equation). And the requirements of RREF guarantee that this dependent variable will NOT be used in any other equations, and hence we can simply read out the solution.

Intuitively, you may think of RREF is "as close as possible" to the identity matrix. So when the identity matrix is out of reach, we go for RREF. (Also note that the identity matrix itself is a special case of RREF.) ☺

Now we move on to Gaussian Elimination. The foundation of the idea is of course row operations. Starting from a system $\left[A \mid \boldsymbol{b}\right]$, we perform row operations until this augmented matrix is in reduced row echelon form. Then we are done.

However, sometimes we would use upper rows to reduce the rows below, like $r_j \to r_j + kr_i$ with $j > i$. Other times, we would use lower rows to reduce the rows above, like $r_j \to r_j + kr_i$ with $j < i$. Can we do this in a more organized way? Gaussian elimination is the attempt to always go in the following order: first, we ONLY use upper rows to reduce lower rows. We keep doing this until we reach REF. Then, we only use lower rows to reduce upper rows, and reach RREF.

Example 2.4.12. Typically Gaussian elimination works like this. We start with a system say $\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 2 & 4 & 4 & 10 \\ 3 & 4 & 6 & 13 \end{array}\right]$.

First I use ONLY the first row to reduce the rows below. The goal is to make the entry 1 in the upper left column into a pivot, so I want to kill every entry below it. We have

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 2 & 4 & 4 & 10 \\ 3 & 4 & 6 & 13 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 2 & 2 & 4 \\ 0 & 1 & 3 & 4 \end{array}\right].$$

Now the first row is done. Leave it alone forever. Next we want to move to the down-right entry of the previous pivot, and make it the new pivot. So we ONLY use the second row to reduce the rows below. We have

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 2 & 2 & 4 \\ 0 & 1 & 3 & 4 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & 2 & 2 \end{array}\right].$$

Now we are in REF and we have finished the first portion of elimination. For the second portion, we shall only use lower rows to change upper rows. First I change all pivots to ones.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & 2 & 2 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{array}\right].$$

Now I go bottom up. We start from the last row, and kill all entries above the last pivot. So we have

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right].$$

Next we look at the second last pivot, and use this row to kill all entries above this pivot. We have

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right].$$

Now we are in RREF and we can simply read out the answer. ☺

As you can see, the idea is basically to keep doing row operations to get RREF. However, we do it in an organized way: top-down first, and bottom up later.

Unfortunately, this does not always work. Look at this example.

Example 2.4.13. We start with a system say $\left[\begin{array}{ccc|c} 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 2 & 2 & 5 \end{array}\right]$.

First, I would use the first row to reduce the rows below, and make the upper left entry a pivot. Oops! This cannot be done, since the upper left entry is zero, and it has no ability to reduce anything. We failed.

Well, this is actually not a big deal. In this case, just pick any non-zero entry in the first column, and SWAP it with the first row. Then we can proceed as desired. In this case, we have

$$\left[\begin{array}{ccc|c} 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 2 & 2 & 5 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 2 & 5 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{array}\right].$$

Now we want to make the $(2,2)$ entry into a pivot again. Oops! We hit zero again. What should we do? Well, we simply swap again. Keep in mind that we ONLY swap the second row with lower rows. Row one is already done, and should be left alone ever since. So we have

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{array}\right].$$

Well, actually no more reduction is needed. We are already done with REF. Now we proceed with the bottom-up portion of elimination, and find RREF. ☺

As you can see, if you hit an obstruction during your elimination process, then don't worry. Just swap rows to make your pivot non-zero, and continue.

What if there is no proper rows to swap with? Consider this example.

Example 2.4.14. Consider the system $\left[\begin{array}{cccccc|c} 0 & 1 & 2 & 3 & 4 & 0 & 10 \\ 0 & 2 & 4 & 0 & 2 & 1 & 9 \\ 0 & 3 & 6 & 1 & 4 & 1 & 15 \end{array}\right]$.

First I would like to make the upper left entry a pivot. Oops, it is zero. No worry, I shall simply swap with... wait, the entire first column is zero. There is no one to swap with!

But actually this is no cause for worry. This simply means that the first pivot is NOT in the first column. So we start with the second column and proceed. So we have

$$\left[\begin{array}{cccccc|c} 0 & 1 & 2 & 3 & 4 & 0 & 10 \\ 0 & 2 & 4 & 0 & 2 & 1 & 9 \\ 0 & 3 & 6 & 1 & 4 & 1 & 15 \end{array}\right] \to \left[\begin{array}{cccccc|c} 0 & 1 & 2 & 3 & 4 & 0 & 10 \\ 0 & 0 & 0 & -6 & -6 & 1 & -11 \\ 0 & 0 & 0 & -8 & -8 & 1 & -15 \end{array}\right].$$

So $(1,2)$-entry is my first pivot. Now we move to the lower right entry, i.e., the $(2,3)$-entry, and try to make it a pivot. However, it is zero again, and there is nothing below to swap with. But this is fine. This simply means that we have no pivot in this column as well. We simply move to the next column, and make the $(2,4)$-entry a pivot instead. So we have

$$\left[\begin{array}{cccccc|c} 0 & 1 & 2 & 3 & 4 & 0 & 10 \\ 0 & 0 & 0 & -6 & -6 & 1 & -11 \\ 0 & 0 & 0 & -8 & -8 & 1 & -15 \end{array}\right] \to \left[\begin{array}{cccccc|c} 0 & 1 & 2 & 3 & 4 & 0 & 10 \\ 0 & 0 & 0 & -6 & -6 & 1 & -11 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} \end{array}\right].$$

And now we are in REF. Then we proceed with the bottom-up portion of elimination, and find RREF, which is $\left[\begin{array}{cccccc|c} 0 & 1 & 2 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right]$. ☺

Technically, some people only use the term Gaussian elimination to describe the process of going from an arbitrary matrix to REF. And people use the longer term Gauss-Jordan elimination to refer to the whole process of going all the way to RREF. It is NOT a very important distinction though.

Algorithm 2.4.15. *The **Gaussian Elimination** refers to the following process, where we start with a generic matrix A, and put it into REF.*

First, find a non-zero entry in the first column, and do a row swap to make that row the first row. (If the first column is entirely zero, then simply skip it, and do this to the second column, etc..)

Next, use the first row to row-reduce all lower rows, so that all other entries in the first column are zero. We are happy with our first row now. From this point on, the first row shall NEVER change.

Next, look at the next column, find a non-zero entry in it (but not in the first row), swap that row to the second row, and row-reduce lower rows, and so on. You can see how this goes.

This process will give you a REF in the end. (Note that this is largely a "downward" process. We work out the first row, then the second row, and so on.)

Algorithm 2.4.16. *The **Gauss-Jordan Elimination** refers to the following process, where we start with a generic matrix A, and put it into RREF.*

First, we use Gaussian elimination to go from A to a REF. Next, we scale the rows so that all pivots are 1. Finally, we use each pivotal rows to clear entries above the pivots. Now we have RREF and we are done. (In practice, this is an"upward" process. We usually work with the last pivot, clear all entries above it, then the second to last pivot, clear all entires above it, and so on. This way we avoided some redundant calculations.)

Proposition 2.4.17. *For any matrix A, we can transform it into RREF using elementary row operations. (We call the result **the RREF of** A or simply write $RREF(A)$.)*

Proof. Just use Gauss-Jordan elimination. □

Proposition 2.4.18. *We have the following situations for a linear system $A\boldsymbol{x} = \boldsymbol{b}$.*

1. *If the RREF of the augmented matrix has a pivot in the last column (the "augmented portion"), then there is no solution. (Because that corresponding equation would read $0 = 1$.)*
2. *If the RREF of the augmented matrix has no contradiction (no pivot in the "augmented portion"), and has the same number of pivots as variables, then there is a unique solution.*
3. *If the RREF of the augmented matrix has no contradiction (no pivot in the "augmented portion"), and has less pivots than variables, then there are infinitely many variables. (Not enough constraints to solve it.)*

2.5 Uniqueness of RREF

Freedom is not what it is. It is what we choose it to be.

I could be talking about life, but actually I am talking about variables. Recall that in reality, we can sometimes choose which variable is free. Say we only have a single equation $x + y = 1$, then we can write $x = 1 - y$, where y is free and x is dependent. Or we can also write $y = 1 - x$, and now x is free and y is dependent. In this sense, RREF is trying to move dependent variables to be as early as possible, and push free variables to be as late as possible.

However, if we choose free variables to be as late as possible, then there must be only one way to do it. In particular, it means we must have the following:

Proposition 2.5.1. *For any matrix A, its RREF is unique. (I.e., if we can use elementary row operations to reduce A to RREF, say X, and also use elementary row operations to reduce A to RREF, say Y, then $X = Y$.)*

To prove this, we make two key observations.

Lemma 2.5.2. *For any matrix A already in RREF, then the free columns are exactly the columns that are linear combinations of columns to its left.*

Proof. Look at the definition of an RREF.

$$\begin{bmatrix} 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & 1 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & 1 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & & & & & \ddots & \vdots & \vdots & & \vdots \\ & & & & & & & & & & & & & 0 & * & \cdots & * \\ & & & & & & & & & & & & & 1 & * & \cdots & * \\ & & & & & & & & & & & & & & & & \end{bmatrix}.$$

If we pick a free column, then it is in fact a linear combination of all the pivotal columns to its left! □

Lemma 2.5.3. *For any matrix A already in RREF, then the k-th pivotal column must be $\boldsymbol{e}_k$.*

Proof. Look at the definition of an RREF. This is trivial by definition.

$$\begin{bmatrix} 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & 1 & * & \cdots & * & 0 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & 1 & * & \cdots & * & \cdots\cdots\cdots & 0 & * & \cdots & * \\ & & & & & & & & & & & & \ddots & \vdots & \vdots & & \vdots \\ & & & & & & & & & & & & & 0 & * & \cdots & * \\ & & & & & & & & & & & & & 1 & * & \cdots & * \\ & & & & & & & & & & & & & & & & \end{bmatrix}.$$

□

Proof of Proposition 2.5.1. We perform induction on the number of columns of A. Suppose A has only one column. Then if $A \neq \mathbf{0}$, the RREF of A must be a single pivotal column, so the RREF must be $\boldsymbol{e}_1$. It is indeed unique. Now, if $A = \mathbf{0}$, then the RREF of A can only remain $\mathbf{0}$, so again it is unique.

Now suppose the statement is true for matrices k columns. Suppose A has $k+1$ columns, say $A = \begin{bmatrix} A_k & \boldsymbol{a} \end{bmatrix}$ where A_k is the collection of its first k columns. Suppose we can row reduce A into RREF $X = \begin{bmatrix} X_k & \boldsymbol{x} \end{bmatrix}$, and we can also row reduce A into RREF $Y = \begin{bmatrix} Y_k & \boldsymbol{y} \end{bmatrix}$. Our goal is to show that $X = Y$.

First of all, since X, Y are both in RREF, we see that X_k, Y_k are both RREF of A_k. By induction hypothesis, we have $X_k = Y_k$. So now our goal is to show that $\boldsymbol{x} = \boldsymbol{y}$.

Suppose $\boldsymbol{x}$ is a free column in the RREF X. Then it is a linear combination of columns to its left. Say $X_k = Y_k = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_k \end{bmatrix}$, and say $\boldsymbol{x} = t_1\boldsymbol{b}_1 + \cdots + t_k\boldsymbol{b}_k$. Then this implies that $X_k\boldsymbol{t} = \boldsymbol{x}$ for $\boldsymbol{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_k \end{bmatrix}$. In particular, $\boldsymbol{t}$ is a solution to the linear system with augmented matrix $\begin{bmatrix} X_k & \boldsymbol{x} \end{bmatrix}$.

Since the matrix $\begin{bmatrix} X_k & \boldsymbol{x} \end{bmatrix}$ can turn into A via row operations, which can turn into $\begin{bmatrix} Y_k & \boldsymbol{y} \end{bmatrix}$ via row operations, therefore the linear system with augmented matrix $\begin{bmatrix} X_k & \boldsymbol{x} \end{bmatrix}$ and the linear system with augmented matrix $\begin{bmatrix} Y_k & \boldsymbol{y} \end{bmatrix}$ must have identical solution set. So $\boldsymbol{t}$ must also be a solution to the linear system with augmented matrix $\begin{bmatrix} Y_k & \boldsymbol{y} \end{bmatrix}$. So $\boldsymbol{y} = Y_k\boldsymbol{t} = X_k\boldsymbol{t} = \boldsymbol{x}$, and thus we are done with $X = Y$.

Similarly, if $\boldsymbol{y}$ us a free column in the RREF of Y, we shall also get $\boldsymbol{x} = \boldsymbol{y}$ for the same reason.

Finally, suppose $\boldsymbol{x}$ is a pivotal column in X and $\boldsymbol{y}$ is a pivotal column in Y. Say $X_k = Y_k$ has t pivotal columns, then $\boldsymbol{x}$ must be the $t+1$-th pivotal column in X, hence it must be $\boldsymbol{e}_{t+1}$. And similarly, $\boldsymbol{y}$ must also be $\boldsymbol{e}_{t+1}$. So we are done. □

The key to the magic proof above, is the conversion between solutions to a linear system, and the linear relations among columns of a matrix. In particular, we can extract the following lemma.

Lemma 2.5.4. *Elementary row operations preserves linear relations among columns.*

Example 2.5.5. Say $\begin{bmatrix}0&1&2&3&4&0\\0&2&4&0&2&1\\0&3&6&1&4&1\end{bmatrix}$. Note how the third column is twice the second column. Now do all kinds of row operations to it, and see that this will always be the case. ☺

Proof of the Lemma. The key realization is the following: given a matrix A, its column linear relations are exactly solutions to the system $A\boldsymbol{x} = \mathbf{0}$!

(For example, in the $\begin{bmatrix}0&1&2&3&4&0\\0&2&4&0&2&1\\0&3&6&1&4&1\end{bmatrix}$ case, if you apply this matrix to $\begin{bmatrix}0\\-2\\1\\0\\0\\0\end{bmatrix}$, you will get $\mathbf{0}$. This is precisely because the third column is twice the second column. Recall that we DEFINE the matrix vector multiplication to be a linear combination of the columns of the matrix.)

Now row operations preserve solutions. Therefore, they preserve linear dependencies among columns. □

Example 2.5.6. Think about a RREF, say $X = \begin{bmatrix}0&1&2&0&1&0\\0&0&0&1&1&0\\0&0&0&0&0&1\end{bmatrix}$. You can easily see that pivotal columns are EXACTLY the columns that are NOT a linear combination of previous columns, where as free columns MUST BE a linear combination of previous pivotal columns, and the entries in the free column simply tells you the coefficients! (Furthermore, these "linear" statements must be simultaneously true for both X and A, because row operations preserves the linear relations among columns.)

Say $A = \begin{bmatrix}0&1&2&3&4&0\\0&2&4&0&2&1\\0&3&6&1&4&1\end{bmatrix}$. The first column of A is all zero, hence the first column is always going to be zero in RREF.

The second column of A is NOT a linear combination of previous columns. Hence it is the FIRST pivotal column, and the corresponding column in RREF must be $\boldsymbol{e}_1$.

The third column of A is twice the second column. So this must still be true in the RREF. So the third column of RREF must be $2\boldsymbol{e}_1$.

The fourth column of A is NOT a linear combination of previous columns. Hence it is the SECOND pivotal column, and the corresponding column in RREF must be $\boldsymbol{e}_2$.

The fifth column of A is the sum of the second and fourth column of A. Hence this relation must still be true in RREF, and the fifth column of RREF must be $\boldsymbol{e}_1 + \boldsymbol{e}_2$.

Finally, the last column of A is NOT a linear combination of previous columns. (One can see this by, say, noticing that all previous columns are orthogonal to $\begin{bmatrix}-1\\-4\\3\end{bmatrix}$, but not the last column.) Hence it is the THIRD pivotal column, and the corresponding column in RREF must be $\boldsymbol{e}_3$.

So the RREF of A has no choice but to be $\begin{bmatrix}0&1&2&0&1&0\\0&0&0&1&1&0\\0&0&0&0&0&1\end{bmatrix}$.

In short, the first column of A that is NOT a linear combination of previous columns must be a pivotal column, and thus must be $\boldsymbol{e}_1$ in any RREF. The second one which is NOT a linear combination of previous columns must also be a pivotal column, and thus must be $\boldsymbol{e}_2$. And so on.

And a free column MUST be a linear combination of previous pivotal columns for some specific coefficients. RREF simply records this coefficients. So any RREF must also have the same free columns.

So RREF is unique. ☺

Definition 2.5.7. *For a matrix A, we define its **rank** to be the number of pivots in the RREF of A. (Note that this is well-defined since RREF is unique. Otherwise we cannot define this.)*

So given a linear system $A\boldsymbol{x} = \boldsymbol{b}$, what is the rank of A? What is the rank of the augmented matrix $\begin{bmatrix} A & \boldsymbol{b} \end{bmatrix}$?

Well, if the two rank disagree, it means we have a contradiction. (Can you see why?)

If the two ranks agree, what is this rank? On one hand, it means after simplification, how many non-zero rows we have for our system. Note that the zero rows are all REDUNDANT equations. Therefore what's left are the effective eqations.

For example, if we have $x + y = 1, 2x + 2y = 2$, then the second equation is redundant. Even though we have two equations, but effectively it is as if we only have one equation (either one will do in this case). So, rank is the same as the number of effective equations.

On the other hand, each pivot corresponds to a dependent variable. So, rank is also the number of dependent variables.

So given a coefficient matrix A of size $m \times n$ (i.e., the system $A\boldsymbol{x} = \boldsymbol{b}$ has m equations and n variables), we have the following conclusion: (Here by rank I mean the rank of A, not the augmented matrix.)

1. rank = the number of effective equations = the number of dependent variables. (The fundamental theorem of linear algebra.)
2. m - rank = the number of redundant equations.
3. n - rank = the number of free variables (in fact, the dimension of the solution set).

Here are some really interesting corollaries.

Proposition 2.5.8. *Given an $m \times n$ matrix A of rank r, then we always have $r \leq m, n$. Furthermore, $r = n$ iff the linear map defined by A is injective (whatever $\boldsymbol{b}$, we have at most one solution to $A\boldsymbol{x} = \boldsymbol{b}$), and $r = m$ iff the linear map defined by A is surjective (whatever $\boldsymbol{b}$, we have at least one solution to $A\boldsymbol{x} = \boldsymbol{b}$).*

Proof. That $r \leq m, n$ is obvious. (Just look at the pivots in RREF....)

Suppose $r = n$. Then either we have a contradiction, or we have all variables dependent variables. There is no free variable. So the solution is unique in that case.

Conversely, suppose A is injective. Then the only solution to $A\boldsymbol{x} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$, i.e., there are NO linear dependency among the columns at all. So all columns are pivotal, and $r = n$.

Suppose $r = m$. Note that the augmented matrix also have m rows, and by adding another column, its rank cannot get smaller. So its rank is at least r and at most m. But given $r = m$, its rank must also be r. So there is no source of contradiction. We always have at least one solution.

Conversely, suppose $r < m$. This implies that on the left hand side, we have linear dependency among rows (and therefore we get redundant equations). Choose $\boldsymbol{b}$ that violates this linear dependency in its coordinates, then $A\boldsymbol{x} = \boldsymbol{b}$ will have no solution. □

Corollary 2.5.9. *The linear map for a matrix A is bijective iff A is a square matrix with full rank. (I.e., $m = n = r$.) In particular, a square matrix is injective iff surjective iff bijective.*

(Compare this with the situation of sets. If finite sets S, T have the same size, then any functions $f : S \to T$ is injective iff surjective iff bijective.)

Corollary 2.5.10. *A matrix has linearly independent columns iff its linear map is injective. It has linearly independent rows iff its linear map is surjective.*

For a square matrix, its rows are linearly independent iff its columns are linearly independent iff its corresponding linear map is bijective.

Proof. We have column independence iff all columns are pivotal, iff $n = r$.

We have row independence iff row operations cannot produce a zero row, iff $m = r$. □

The corollary above is a bit shocking for beginners. You might want to try a few to see what might happen. Say $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. Its rows are dependent, and columns are dependent. If you change the lower

right entry to 10, then rows are independent and columns are independent. You must always have both or nothing.

Corollary 2.5.11 (Dimensions are linearly well-established)**.** *There is no linear bijection between $\mathbb{R}^m$ and $\mathbb{R}^n$ when $m \neq n$. Furthermore, if $m > n$, then any linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ must not be surjective (smaller space cannot cover larger space), while any linear map from $\mathbb{R}^m \to \mathbb{R}^n$ must not be injective (larger space must squeeze to fit into a smaller space).*

Proof. For any linear map $f : \mathbb{R}^n \to \mathbb{R}^m$, then it corresponds to some $m \times n$ matrix. But if $m \neq n$, then it cannot be a bijection. (Other statements can be done similarly.) □

This last corollary is also comparable to the situation of sets. However, the requirement of linearity is important here. If one only require continuity, then there is in fact a continuous and surjective map from $\mathbb{R}$ to $\mathbb{R}^2$. Search for "space-filling curve" if you are interested.

Chapter 3

Operations on Matrices

We are doing linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$

3.1 Matrix Multiplication

Recall that we have learned about matrix-vector multipications. if $A = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix}$ and $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$, then we have

$$\begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \sum v_i \boldsymbol{a}_i.$$

So $A\boldsymbol{v}$ is simply a linear combination of columns of A with respect to the coefficients given as coordinates of $\boldsymbol{v}$.

Now consider the first coordinate of the output. Given a linear combination of columns of A, to find the first coordinate, we actually only need the first coordinate of EACH column, and combine them accordingly.

This gives rise to a different (but equivalent) formula. If $A = \begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{r}_m^{\mathrm{T}} \end{bmatrix}$, then see if you can verify the following formula

$$\begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{r}_m^{\mathrm{T}} \end{bmatrix} \boldsymbol{v} = \begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} \boldsymbol{v} \\ \vdots \\ \boldsymbol{r}_m^{\mathrm{T}} \boldsymbol{v} \end{bmatrix}.$$

So effectively, it is like we are doing a "dot product" of $\boldsymbol{v}$ with each row of A.

Example 3.1.1. Here is a simple example. Let us see the two ways of computing a matrix-vector multiplication.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + 3 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 8 \\ 18 \end{bmatrix}.$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 8 \\ 18 \end{bmatrix}.$$

Personally, I just imagine that for $A\boldsymbol{v}$, then $\boldsymbol{v}$ is a brick, and the rows of A are the faces of the people that I hate. So I take the brick, and smash it onto those faces one by one. Here "smash" would mean taking a dot product. ☺

3.1.1 Composition of Linear Maps

Recall that, the matrix $A = \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}$ represents the "counting" process in the chicken-rabbit cage problem. Given these four enties, we now know how to count. And given a vector $\boldsymbol{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ that represents x chickens and y rabbits, the vector $A\boldsymbol{v}$ represents the number of heads and leg. Note that as a linear map, we have $A : \mathbb{R}^2 \to \mathbb{R}^2$.

Now suppose we are furthermore selling these heads and legs for money, and suppose the rest of the body parts has no value. (Ma La Tu Tou and Pao Jiao Feng Zhua....) Anyway, say each head is worth 5 yuan and each leg 7 yuan. How much would x chickens and y rabbits worth in total? Note that this evaluation process is also linear, and it should be a linear map from $\mathbb{R}^2$ to $\mathbb{R}$, i.e., it should be a 1 by 2 matrix.

To do this, we need the evaluation map $B = \begin{bmatrix} 5 & 7 \end{bmatrix}$ applied to $A\boldsymbol{v}$. The evaluation map has this matrix because it sends $\boldsymbol{e}_1$, i.e., one head, to 5, while it sends $\boldsymbol{e}_2$, i.e., one leg, to 7.

So with the chicken-rabbit input of $\boldsymbol{v} = \begin{bmatrix} x \\ y \end{bmatrix}$, we should get $A\boldsymbol{v} = \begin{bmatrix} x+y \\ 2x+4y \end{bmatrix}$ of head-leg output, and the total worth of them would be $B(A\boldsymbol{v}) = 5(x+y) + 7(2x+4y) = 19x + 33y$.

In particular, we see that the composition $B \circ A$ is still linear, and it has a matrix of $\begin{bmatrix} 19 & 33 \end{bmatrix}$

Proposition 3.1.2. *The composition of linear maps is still linear.*

Proof. Say f, g are linear and has a well-defined composition. Then $f \circ g(a\boldsymbol{v} + b\boldsymbol{w}) = f(g(a\boldsymbol{v} + b\boldsymbol{w})) = f(ag(\boldsymbol{v}) + bg(\boldsymbol{w})) = af(g(\boldsymbol{v})) + bf(g(\boldsymbol{w})) = a(f \circ g)(\boldsymbol{v}) + b(f \circ g)(\boldsymbol{w})$. □

Definition 3.1.3. *Given two matrix A, B, say as linear maps we have $A : \mathbb{R}^n \to \mathbb{R}^m, B : \mathbb{R}^d \to \mathbb{R}^n$, then we can do their composition. The composition of these two linear maps is what we define to be the matrix multiplication AB.*

Proposition 3.1.4. *The matrix multiplication AB is well-defined iff the number of columns of A is the same as the number of rows of B.*

This is because the codomain of B must match with the domain of A. So when you do matrix multiplications, you should always expect to see something like below. You can almost imagine that the n here in the middle just got canceled away.

$$m\left\{\overbrace{\begin{bmatrix} & \\ & \end{bmatrix}}^{n}\right. \quad n\left\{\overbrace{\begin{bmatrix} & \\ & \end{bmatrix}}^{d}\right. = m\left\{\overbrace{\begin{bmatrix} & \\ & \end{bmatrix}}^{d}\right. .$$

$$(m \times n \text{ matrix})(n \times d \text{ matrix}) = (m \times d \text{ matrix})$$

Now we move on to the computations.

Proposition 3.1.5. *Let us write B in terms of its columns, $B = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix}$. Then $AB = \begin{bmatrix} A\boldsymbol{b}_1 & \dots & A\boldsymbol{b}_n \end{bmatrix}$.*

Proof. Note that the i-th column of AB should be $(AB)\boldsymbol{e}_i$. By definition of map composition, this is $A(B(\boldsymbol{e}_i))$, and $B\boldsymbol{e}_i$ must be the i-th column of B, i.e., $\boldsymbol{b}_i$. So $(AB)\boldsymbol{e}_i = A\boldsymbol{b}_i$ is the i-th column of AB. □

In particular, feel free to verify now that $\begin{bmatrix} 5 & 7 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 19 & 33 \end{bmatrix}$, as in our chicken-rabbit example. We also see immediately that matrix-vector multiplication is a special case of matrix-matrix multiplication, if we simply think of the vector as a matrix with a single column.

Remark 3.1.6. *As a side note, you also see that given an $m \times n$ matrix A, then you can only multiply a column vector (i.e., $n \times 1$ matrix) to the right of A, and only multiply a row vector (i.e., $1 \times m$ matrix) to the left of A.*

Example 3.1.7. To rotate $\mathbb{R}^2$ by an angle of θ counter clockwise, this linear map has a matrix of $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. To reflect $\mathbb{R}^2$ about the line $x = y$, this linear map has a matrix of $F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. (Can you see why F is this matrix?)

What if we rotate and then reflect? Then we are doing FR_θ. Doing this computation, you shall see we have $FR_\theta = \begin{bmatrix} \sin\theta & \cos\theta \\ \cos\theta & -\sin\theta \end{bmatrix}$.

What if we reflect and then rotate? Again computation gives $R_\theta F = \begin{bmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{bmatrix}$.

Can you see the relation between the two? In fact we shall always have $R_\theta F = FR_{-\theta}$ for whatever rotation R_θ and reflection F. (Here we require the rotation to be around the origin, and the reflection to be fixing the origin.)

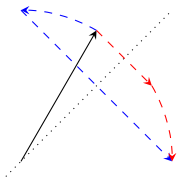


Figure 3.1.1: The blue process is FR_θ while the red process is $R_\theta F$.

Note that rotations before and after a reflecion are like mirror images of each other. So intuitively, it should be clear that rotation by θ before a reflection is the same as rotation by θ after the reflection. ☺

Above example shows that matrix multiplication (i.e., linear map composition) fails to be commutative. But we do have associativity. Warning: Associativity is the source of MANY magic in linear algebra. Keep it in mind.

Proposition 3.1.8. *We have $(AB)C = A(BC)$.*

Proof. This is because map composition is associative. End of proof just by definition.

If you want a more computational proof, here it is. Write C in terms of its columns $C = \begin{bmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_n \end{bmatrix}$. Then

$$A(BC) = A(B\begin{bmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_n \end{bmatrix}) = A\begin{bmatrix} B\boldsymbol{c}_1 & \dots & B\boldsymbol{c}_n \end{bmatrix} = \begin{bmatrix} AB\boldsymbol{c}_1 & \dots & AB\boldsymbol{c}_n \end{bmatrix} = (AB)\begin{bmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_n \end{bmatrix} = (AB)C.$$

□

Remark 3.1.9. *In the old Chinese textbooks, this was done like this. They introduce matrix multiplication without any talk of map composition, and simply throw a formula to the students as the definition of matrix multiplication.*

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \dots & b_{1d} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nd} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots + a_{1n}b_{n1} & \dots & a_{11}b_{1d} + a_{12}b_{2d} + \cdots + a_{1n}b_{nd} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{11} + a_{m2}b_{21} + \cdots + a_{mn}b_{n1} & \dots & a_{m1}b_{1d} + a_{m2}b_{2d} + \cdots + a_{mn}b_{nd} \end{bmatrix}.$$

Next, they literally compute AB and then $(AB)C$. Then they compute BC and $A(BC)$. And they manipulate the formulat to show that the two are the same.

It is my opinion that this is a counter-intuitive and counter-productive approach to prove associativity.

Example 3.1.10. From this point on, we don't need matrix-vector multiplication any more. We only need matrix multiplication. Here are somethings to think about.

Suppose $\boldsymbol{v} = \begin{bmatrix}1\\2\\3\end{bmatrix}$ and $\boldsymbol{w} = \begin{bmatrix}1\\2\\3\end{bmatrix}$. Then what is $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$? This is simply the dot product, and it yields $1+4+9=14$. Alternatively, you can also think of this as a matrix multiplication of a 1×3 matrix $\boldsymbol{v}^{\mathrm{T}}$ with a 3×1 matrix $\boldsymbol{w}$, and it gives you the 1×1 matrix which is simply the number $\begin{bmatrix}14\end{bmatrix}$.

Now consider $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$. This is the matrix multiplication of a 3×1 matrix with a 1×3 matrix, hence it gives a 3×3 matrix as a result! This would be $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}} = \boldsymbol{v}\begin{bmatrix}1 & 2 & 3\end{bmatrix}$.

Recall the formula

$$A\begin{bmatrix}\boldsymbol{b}_1 & \boldsymbol{b}_2 & \boldsymbol{b}_3\end{bmatrix} = \begin{bmatrix}A\boldsymbol{b}_1 & A\boldsymbol{b}_2 & A\boldsymbol{b}_3\end{bmatrix}.$$

So we see that

$$\boldsymbol{v}\begin{bmatrix}1 & 2 & 3\end{bmatrix} = \begin{bmatrix}\boldsymbol{v}1 & \boldsymbol{v}2 & \boldsymbol{v}3\end{bmatrix} = \begin{bmatrix}1 & 2 & 3\\2 & 4 & 6\\3 & 6 & 9\end{bmatrix}.$$

In general, $\begin{bmatrix}v_1\\\vdots\\v_m\end{bmatrix}\begin{bmatrix}w_1 & \dots & w_n\end{bmatrix} = \begin{bmatrix}v_1w_1 & \dots & v_1w_n\\\vdots & \ddots & \vdots\\v_mw_1 & \dots & v_mw_n\end{bmatrix}$. In particular, the (i,j)-entry of $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$ is simply v_iw_j. ☺

Example 3.1.11 (Projection Formula)**.** Recall scalar multiplication $k\boldsymbol{v}$ and $\boldsymbol{v}k$ for some $\boldsymbol{v}\in\mathbb{R}^n$, say $n\neq 1$. The former is not matrix multiplication, because k only has one column while $\boldsymbol{v}$ has n rows. However, the latter IS a matrix multiplication, since $\boldsymbol{v}$ only has one column, and k has one row, and the way of calculation is exactly as expected!

In particular, if you started with $(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w})\boldsymbol{u}$, then it should NOT be equal to $\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{w}\boldsymbol{u})$ (which is an illegal anyway). This is because $(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w})$ and $\boldsymbol{u}$ here are NOT doing a matrix multiplication, and this is merely a scalar multiplication. However, you can write $(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w})\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}) = (\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})\boldsymbol{w}$. Now $\boldsymbol{u}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w})$ is a legal matrix multiplciation, so we can proceed to use the associativity of matrix multiplication.

We know the projection of $\boldsymbol{v}$ to a direction $\boldsymbol{w}$ gives the vector $\frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\boldsymbol{w}\cdot\boldsymbol{w}}\boldsymbol{w}$. This is a linear map sending $\boldsymbol{v}$ to its image, i.e., $P_{\boldsymbol{w}}:\mathbb{R}^n\to\mathbb{R}^n$ such that $P_{\boldsymbol{w}}(\boldsymbol{v}) = \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\boldsymbol{w}\cdot\boldsymbol{w}}\boldsymbol{w}$. How to find the matrix $P_{\boldsymbol{w}}$?

Well, first we try to reorganize $P_{\boldsymbol{w}}(\boldsymbol{v})$ into matrix multiplications as much as possible, as $\frac{1}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}\boldsymbol{w}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})$. Next we use associativity to get $\frac{1}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\boldsymbol{v}$.

So we see that $P_{\boldsymbol{w}} = \frac{\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}$. What a pretty formula!

Be careful here. The denominator is NOT a matrix!!! You can NEVER put a matrix or a vector in the denominator. Only numbers are allowed to do so. The number here means we are dividing each entry of the matrix $\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}$ by the number $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$.

In particular, if $\boldsymbol{u}$ is a unit vector, then projection to $\boldsymbol{u}$ is $P_{\boldsymbol{u}} = \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$. ☺

3.1.2 Rows, Columns, and Entries of a matrix

We already know the following by definition of matrices as linear maps.

Proposition 3.1.12. *The i-th column of a matrix A is $A\boldsymbol{e}_i$.*

But what about rows?

Proposition 3.1.13. *The i-th row of a matrix A is $\boldsymbol{e}_i^{\mathrm{T}}A$.*

For example, calculate $\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\5 & 6 & 7 & 8\\9 & 10 & 11 & 12\end{bmatrix}$. When we apply the row vector to each column, it is as if we are taking a dot product, and we simply get the second coordinate of each column, which gives the second row.

Remark 3.1.14. *Say A is a matrix and $\boldsymbol{v}$ is a (vertical) vector, then $A\boldsymbol{v}$ is again a (vertical) vector.*

In comparison, if we have a row vector $\boldsymbol{v}^{\mathrm{T}}$, then $\boldsymbol{v}^{\mathrm{T}}A$ is again a row vector.

In fact, you may check that $A\boldsymbol{v}$ is a linear combination of columns of A according to coordinates of $\boldsymbol{v}$. Similarly, $\boldsymbol{v}^{\mathrm{T}}$ is a linear combination of rows of A according to coordinates of $\boldsymbol{v}^{\mathrm{T}}$.

In the example $\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 5 & 6 & 7 & 8\\ 9 & 10 & 11 & 12\end{bmatrix}$*, we are taking none of the first row, one copy of the second row, and none of the third row. So the result of this linear combination is simply the second row.*

Remark 3.1.15. *If the columns of A are linearly dependent, then $A\boldsymbol{v} = \boldsymbol{0}$ for some nonzero vector $\boldsymbol{v}$. Similarly, if the rows of A are linearly dependent, then $\boldsymbol{v}^{\mathrm{T}}A = \boldsymbol{0}^{\mathrm{T}}$ for some nonzero $\boldsymbol{v}^{\mathrm{T}}$.*

It all works out pretty symmetrically. Just remember, "columns are to the right of A, while rows are to the left of A".

Corollary 3.1.16. *Suppose* $A = \begin{bmatrix}\boldsymbol{a}_1^{\mathrm{T}}\\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\end{bmatrix}$*. We have matrix multiplication* $AB = \begin{bmatrix}\boldsymbol{a}_1^{\mathrm{T}}\\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\end{bmatrix} B = \begin{bmatrix}\boldsymbol{a}_1^{\mathrm{T}}B\\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}B\end{bmatrix}$.

Proof. This is because $\boldsymbol{e}_i^{\mathrm{T}}(AB) = (\boldsymbol{e}_i^{\mathrm{T}}A)B = \boldsymbol{a}_i^{\mathrm{T}}B$. □

This is the "dual" picture to our previous go to way to do matrix multiplication, i.e., $A\begin{bmatrix}\boldsymbol{b}_1 & \dots & \boldsymbol{b}_n\end{bmatrix} = \begin{bmatrix}A\boldsymbol{b}_1 & \dots & A\boldsymbol{b}_n\end{bmatrix}$.

Let us now finally shift our attention to entries.

Proposition 3.1.17. *The (i,j) entry of A is $\boldsymbol{e}_i^{\mathrm{T}}A\boldsymbol{e}_j$.*

Proof. Note that $\boldsymbol{e}_i^{\mathrm{T}}(A\boldsymbol{e}_j)$ is the i-th row of the j-th column of A, so we are done.

Also note how associativity is lurking here. We can alternatively look at $(\boldsymbol{e}_i^{\mathrm{T}}A)\boldsymbol{e}_j$, which is the j-th column of the i-th row of A, and it is the same entry. □

Proposition 3.1.18. *The (i,j) entry of AB is the dot product between the i-th row of A and the j-th column of B.*

Proof. $\boldsymbol{e}_i^{\mathrm{T}}(AB)\boldsymbol{e}_j = (\boldsymbol{e}_i^{\mathrm{T}}A)(B\boldsymbol{e}_j)$. Simple associativity. □

In particular, if we write A in rows and B in columns, we have

$$AB = \begin{bmatrix}\boldsymbol{a}_1^{\mathrm{T}}\\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_1 & \dots & \boldsymbol{b}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{b}_1 & \dots & \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{b}_n\\ \vdots & \ddots & \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{b}_1 & \dots & \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{b}_n\end{bmatrix}.$$

If you truely like things to be computational, you can further write matrix multiplication in terms of entries.

$$AB = \begin{bmatrix}a_{11} & \dots & a_{1r}\\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mr}\end{bmatrix}\begin{bmatrix}b_{11} & \dots & b_{1n}\\ \vdots & \ddots & \vdots \\ b_{r1} & \dots & b_{rn}\end{bmatrix} = \begin{bmatrix}\sum_i a_{1i}b_{i1} & \dots & \sum_i a_{1i}b_{in}\\ \vdots & \ddots & \vdots \\ \sum_i a_{mi}b_{i1} & \dots & \sum_i a_{mi}b_{in}\end{bmatrix}.$$

Example 3.1.19. Let us do a calculation for fun.

A ***lower triangular matrix*** is a matrix whose entries above the diagonal are all zero. For example, say $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$. Similarly, an ***upper triangular matrix*** is a matrix whose entries below the diagonal are all zero. For example, say $U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. (Note that row echelon forms are all upper triangular.)

Calculation gives $LU = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$. Pretty, yes?

Let us try to see another pretty sight. A Pascal's lower triangular matrix is a lower triangular matrix, where the first column has all 1, the diagonal entries are all 1, and each entry is the sum of the entry above it and the entry to its upper left. Say $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{bmatrix}$.

Similarly, a Pascal's upper triangular matrix is a lower triangular matrix, where the first row has all 1, the diagonal entries are all 1, and each entry is the sum of the entry to the left and the entry to its upper left. Say $U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Calculation gives $LU = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}$. Pretty, yes? This is the symmetric Pascal's matrix, where the first row and first column has all 1, and each entry is the sum of the entry to the left and the entry above it.

These pretty structures are of curse NOT coincidences. (Nothing in math is a coincidence.) However, their nature is more combinatorial than linear algebra, so let us leave it at that. ☺

We like triangular matrices for the following reason.

Proposition 3.1.20. *The product of two upper (lower) triangular matrices is still upper (lower) trianguler. Furthermore, the diagonal entries of the product is the entry-wise product of the diagonal entries of the two matrices. In short, we should have*

$$\begin{bmatrix} a_{11} & * & * \\ & \ddots & * \\ & & a_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & * & * \\ & \ddots & * \\ & & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & * & * \\ & \ddots & * \\ & & a_{nn}b_{nn} \end{bmatrix}.$$

Proof. Note that a triangular matrix must be a square matrix. So if two triangular matrices could multiply, then they must have the same number of columns and rows.

Suppose A, B are upper triangular, say $A = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_n^{\mathrm{T}} \end{bmatrix}$ and $B = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix}$. Then the (i, j)-entry of AB is

$\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{b}_j$. This is a dot product. Say $\boldsymbol{a}_i^{\mathrm{T}} = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 & a_i & \dots & a_n \end{bmatrix}$ and $\boldsymbol{b}_j = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_j \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

Let us consider a lower-triangular entry, which is the case when $i > j$. Since A, B are upper triangular, the first $(i-1)$ entries of $\boldsymbol{a}_i^{\mathrm{T}}$ are zero, and the last $(n-j)$-entries of $\boldsymbol{b}_j$ are zero. So we have

$$\underbrace{\begin{bmatrix} 0 & \dots & 0 & a_i & \dots & a_n \end{bmatrix}}_{i-1} \left.\begin{bmatrix} \vdots \\ b_j \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} n-j = \overbrace{a_1 b_1 + \cdots +}^{\text{First } i-1 \text{ terms are 0}} \underbrace{\cdots + \cdots + a_n b_n}_{\text{Last } n-j \text{ terms are 0}}.$$

Since $i > j$, we have $(i-1) + (n-j) \geq n$. So all terms are zero. So all (i,j)-entry of AB with $i > j$ are zero. Hence AB is still upper triangular.

When $i = j$, then we have

$$\underbrace{\begin{bmatrix} 0 & \dots & 0 & a_i & \dots & a_n \end{bmatrix}}_{i-1} \left.\begin{bmatrix} \vdots \\ b_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} n-i = \overbrace{a_1 b_1 + \dots}^{\text{First } i-1 \text{ terms are 0}} + a_i b_i + \underbrace{\cdots + a_n b_n}_{\text{Last } n-i \text{ terms are 0}} = a_i b_i.$$

So the (i,i)-entry of AB is just $a_{ii}b_{ii}$.

The case of lower triangular matrices is similar. □

Let us have one last example. What if we multiply a matrix with itself?

Definition 3.1.21. *The k-th power of a matrix A, written as A^k, is A multiplying itself k times.*

Example 3.1.22. Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. This is a very important matrix, the simplest shearing matrix.

By calculation, you can see that $A^2 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, A^3 = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$ and so on. See a pattern? Let us prove it here with mathematical induction.

I claim $A^k = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$. The base case when $k = 1$ is trivial, so we only need the inductive step now. Suppose the statement is true for k, let us prove it for $k+1$.

We have $A^{k+1} = A^k A = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & k+1 \\ 0 & 1 \end{bmatrix}$. So we are done. ☺

Finally, let us play a bit with the most important matrix, i.e., the identity matrix.

Proposition 3.1.23. *Consider the $n \times n$ identity matrix I. For all $n \times m$ matrix A, we have $IA = A$. For all $m \times n$ matrix A, we have $AI = A$.*

By convention, for any square matrix A, we usually define $A^0 = I$. (We use this convention even if A has no inverse matrix. Even if $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, we still define $A^0 = I$. This convention will ensure that $A^s A^t = A^{s+t}$ for all non-negative integer s, t and any square matrix A.)

3.1.3 Geometries of Linear Maps

Now we move on to several kinds of matrices, and hopefully also provide you with a variaty of perspectives on how to understand them. I want you to keep in mind: each specific entry is NOT important. Only collectively as a linear may, would they be important. What does the matrix behave? What is the process of this linear map? Those are the more important questions.

Let us start with some geometrically interesting maps.

Example 3.1.24. We know rotations on the plane are $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. What if you compose two rotations? We have $R_\theta R_\phi = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} = \begin{bmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\sin\theta\cos\phi - \sin\phi\cos\theta \\ \sin\theta\cos\phi + \sin\phi\cos\theta & \cos\theta\cos\phi - \sin\theta\sin\phi \end{bmatrix}$. Do these formula looks familiar to you? They SHOULD! These are basic trignometry stuff.

So upon further simplification, this gives $\begin{bmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{bmatrix} = R_{\theta+\phi}$. Huh. On second thought, this is no surprise at all if you just think about the geometry. $R_\theta R_\phi$ literally means do the rotation by ϕ, then do the rotation by θ. Obviously the result is a rotation by $\theta+\phi$. In fact, this whole process could be thought of as a proof of the trigonometry sum formula.

Btw, it should also be obvious that $R_\theta^k = R_{k\theta}$ and so on.

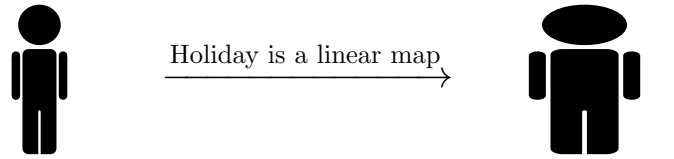
The moral of the story is this: matrix multiplication formula is fine. But sometimes, understanding the meaning of the maps can help you calculate much faster. ☺

Example 3.1.25. Flipping the plane about the line $x = y$ has a matrix of $F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. If you try to apply this to a vector, you see that it sends $\begin{bmatrix} x \\ y \end{bmatrix}$ to $\begin{bmatrix} y \\ x \end{bmatrix}$, so it literally just swap the coordinates. We have previously see that $FR_\theta = R_{-\theta}F$.

Flipping the plane about the x-axis has a matrix of $D = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. (Can you see why?) You can again verify that $DR_\theta = R_{-\theta}D$.

In fact, for any reflection F of the plane about a line through the origin, we always have $FR_\theta = R_{-\theta}F$. Intuitively, rotating by θ means in the mirror, you are rotating by $-\theta$. Try to see this geometrically, and then see if you can verify this numerically. (In general, to reflect along the line with slope angle θ, the matrix is $\begin{bmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{bmatrix}$.) ☺

Example 3.1.26. Let us stretch the plane now. Say we stretch everything in the x-axis direction by a factor of 2. This is a linear map with matrix $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$. (Can you see why?) A picture of myself will now appear twice as fat....

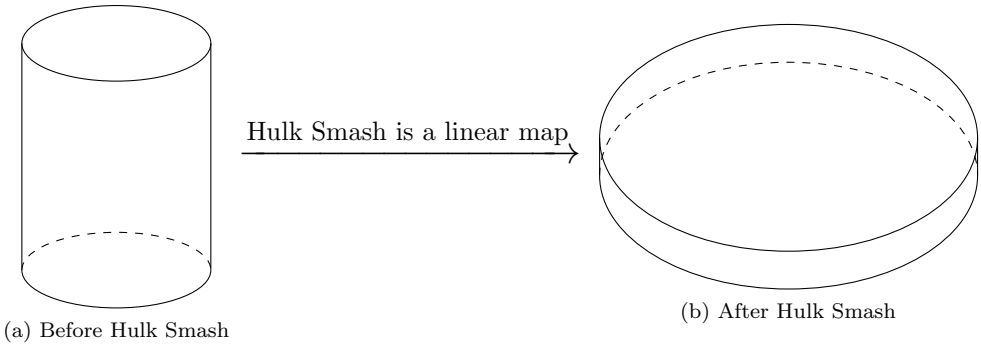


(a) Before Holiday (b) After Holiday

Figure 3.1.2: Holiday = $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

Matrices like these are called ***diagonal matrices***. They stretch things along the coordinate-axes. Say

if we have $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix}$, then this linear map is stretching the space $\mathbb{R}^3$ in the x-axis direction by a factor of 2, then in the y-axis direction by a factor of 3, and finally in the z-axis direction by a factor of $\frac{1}{6}$ (so we are squeezing in this direction).



(a) Before Hulk Smash

(b) After Hulk Smash

Figure 3.1.3: Hulk Smash= $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix}$

The identity matrix is a special case of a diagonal matrix. We are now stretching everything by a factor of 1, i.e., we leave them all unchanged. You may also think of the reflection matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ as a diagonal matrix. We are scaling things in the y-direction by a factor of -1, so things are flipped about the x-axis.

The square projection matrix (projection to the x-axis), $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, which sends $\begin{bmatrix} x \\ y \end{bmatrix}$ to $\begin{bmatrix} x \\ 0 \end{bmatrix}$, is also a special case of this. We are now squeezing everything in the y-axis direction into nothing. Hence we obtained an orthogonal projection to the x-axis.

Multiplications of diagonal matrices are always diagonal. This compares nicely with triangular matrices. (In fact, a matrix is diagonal iff it is both upper triangular and lower triangular.)

In general, for diagonal matrices we can computationally verify that $\begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix} \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_n \end{bmatrix} = \begin{bmatrix} a_1 b_1 & & \\ & \ddots & \\ & & a_n b_n \end{bmatrix}$. Here some entries are empty, which means they are zero. Can you see this geometrically?

Algebraically, also note that for diagonal matrices, we always have commutativity. ☺

Example 3.1.27. For a parallelogram, we learned long ago that its area is "base" times "height". Now, given a parallelogram and fixed base, can you draw all parallelograms with the same height?

You will end up drawing many "sheared" versions of this parallelogram.

Consider the matrix $\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$. Then for any parallelogram whose base is on the x-axis, this linear map will "shear" it into some tilted version of itself, but always with the same base and same height! In particular, a shearing always preserves their volumn. This is the shearing of the x-axis direction to the y-axis direction.

In physics, a "shearing" is a force that acts like a pair of scissors. If you observe a pair of scissors, you will realize that one blade is always on top of the other. When you use the scissors to cut things, the top blade will push things to the right, while the lower blade will push things to the left. This is exactly what $\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$ will tries to do to the plane.

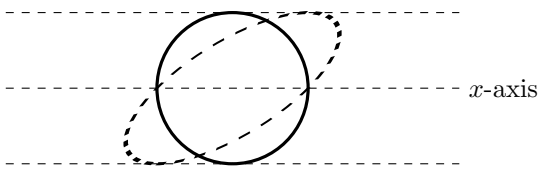


Figure 3.1.4: A shearing of a circle

In fact, a shearing will always preserve the area of whatever shape in the plane. Draw some triangles and see if you can prove that these triangles have area preserved after a shearing.

Now, consider the geometric nature of a shearing, if I do the same shearing twice, it is as if I simply sheared by twice the original amount. Hence it is geometrically very obvious that $\begin{bmatrix}1 & 1\\0 & 1\end{bmatrix}^k = \begin{bmatrix}1 & k\\0 & 1\end{bmatrix}$.

If I did some shearing and then sheared some more, then in total I simply did a bigger shear. Hence it is also geometrically very obvious that $\begin{bmatrix}1 & a\\0 & 1\end{bmatrix}\begin{bmatrix}1 & b\\0 & 1\end{bmatrix} = \begin{bmatrix}1 & a+b\\0 & 1\end{bmatrix}$. Note that all shearings of the xy-plane parallel to the x-axis will commute.

However, shearings along different directions will NOT commute. For example, $\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}$ is a shearing parallel to the y-axis. (Draw some graph and verify this yourself.)

Then we have

$$\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\0 & 1\end{bmatrix} = \begin{bmatrix}1 & 1\\1 & 2\end{bmatrix}.$$

$$\begin{bmatrix}1 & 1\\0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix} = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}.$$

So they do not commute. ☺

3.1.4 Linear Combinations of Matrices

Matrix multiplication is a very powerful tool. It allows us to compose linear maps, understand behaviors among linear maps, understand iterations of a linear transformation, and so on so forth.

Another powerful tool is to do linear combinations of matrices.

Definition 3.1.28. *Given two maps $f, g : \mathbb{R}^n \to \mathbb{R}^m$ (for this definition, they are not required to be linear), we define $f+g : \mathbb{R}^n \to \mathbb{R}^m$ to be the map such that $(f+g)(\boldsymbol{x})$ is defined as $f(\boldsymbol{x})+g(\boldsymbol{x})$, and for any number $k \in \mathbb{R}$, we define $kf : \mathbb{R}^n \to \mathbb{R}^m$ to be the map such that $(kf)(\boldsymbol{x}) = kf(\boldsymbol{x})$.*

This definition is very much standard. For example, if we are considering real functions from $\mathbb{R}$ to $\mathbb{R}$, say e^x and x^2, then the sum of e^x and x^2 is the function $e^x + x^2$. This is obviously the only sensible way to define such a sum.

In the case of matrices (i.e., linear maps), note that we require the domains and codomains to be the same. In particular, $A+B$ is ONLY defined when A, B have the same number of rows and the same number of columns.

Luckily for us, this is super easy to compute.

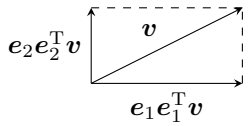
Proposition 3.1.29. *Given A, B of the same size, then the (i,j) entry of $A+B$ is simply the (i,j) entry of A plus the (i,j) entry of B. Similarly, for any $k \in \mathbb{R}$, the (i,j) entry of kA is simply k times the (i,j) entry of A.*

Proof. By definition of linear combinations of matrices, $(A+B)\boldsymbol{v} = A\boldsymbol{v} + B\boldsymbol{v}$ and $(kA)\boldsymbol{v} = k(A\boldsymbol{v})$ for all input $\boldsymbol{v}$.

Let us now figure out the j-th column of $A+B$. But we have $(A+B)\boldsymbol{e}_j = A\boldsymbol{e}_j + B\boldsymbol{e}_j$. Hence this is simply the sum of the j-th column of A and j-th column of B. Done.

Similarly, $(kA)\boldsymbol{e}_j = k(A\boldsymbol{e}_j)$. So the j-th column of kA is simply k times the j-th column of A. □

Example 3.1.30. Think about the identity map $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}$. What does this equation mean? It tells us that the identity map is the sum of the projection to x-axis and the projection to y-axis.



In particular, if we input a vector $\boldsymbol{v}$ to the equation, we have $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\boldsymbol{v} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\boldsymbol{v} + \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\boldsymbol{v}$. This means each vector is the sum of its x-axis "component vector" plus its y-axis "component vector". This orthogonal decomposition of $\boldsymbol{v}$ is a very useful tool in many problems in high school physics.

Note that $\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$ and $\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} = \boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}}$. In general, for the $n \times n$ identity matrix I, we have $I = \sum \boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}}$, and geometrically this means any vector is the sum of its "coordinate component vectors".

In fact, take any pair of orthogonal unit vectors in $\mathbb{R}^2$, say $\boldsymbol{u} = \begin{bmatrix}\frac{3}{5}\\ \frac{4}{5}\end{bmatrix}, \boldsymbol{v} = \begin{bmatrix}-\frac{4}{5}\\ \frac{3}{5}\end{bmatrix}$, and calculate $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$. What do you see? And now you know why. ☺

Example 3.1.31. Let us work out a formula for the reflection matrix in $\mathbb{R}^3$. Suppose we want to reflect about a plane through the origin. What should I do?

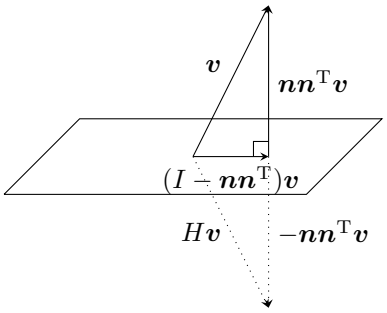


Figure 3.1.5: From $\boldsymbol{v}$ to its reflection $H\boldsymbol{v}$.

Let $\boldsymbol{n}$ be a unit normal vector. Then for any vector $\boldsymbol{v}$, we have $\boldsymbol{v} = (\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v} + (I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}$, where I is the identity matrix. You can verify that this is an orthogonal decomposition into a component parallel to $\boldsymbol{n}$, i.e., $(\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}$, and a component orthogonal to $\boldsymbol{n}$ (and hence parallel to our plane of reflection), i.e., $(I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}$.

Now for the part that is parallel to plane of reflection, it should stay unchanged. For the part that is parallel to the normal vector, it shall be reflected (i.e., negated). So all in all, we want to send $\boldsymbol{v}$ to $-(\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v} + (I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}$. Simplifying this, we want to send $\boldsymbol{v}$ to $(I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}$. Obviously this linear map is simply multiplying by the matrix $I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$.

In general, for any unit vector $\boldsymbol{n} \in \mathbb{R}^n$, for this very reason $I_n - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ is always a reflection about a hyperplane with normal vector $\boldsymbol{n}$. This gives all possible reflections in all dimensions.

Some texts also refers to these "higher dimensional reflections" as ***Householder transformations***. ☺

So far we have been trying to focus on the geometry of linear maps. Now let us also look at some analytical perspectives.

Example 3.1.32. We learn integration, and it is basically the idea of accumulation. Suppose I track my total amount of money each month. Say I start with 1 dollar, but after one month I have 3 dollars, after

two months I have 6 dollars, and after three months I have 7 dollars. I can write these data as a vector $\begin{bmatrix}1\\3\\6\\7\end{bmatrix}$.

Suppose I am NOT interested in the accumulated amount of money. Rather, I am interested in how many money can I earn each month. The " rate of increase", or in calculus terms, the "derivative" of the total amount of money. What should I do?

The key is the ***forward difference matrix***. It looks like $D = \begin{bmatrix}1 & 0 & 0 & 0\\-1 & 1 & 0 & 0\\0 & -1 & 1 & 0\\0 & 0 & -1 & 1\end{bmatrix}$. It will send $\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}$ to $\begin{bmatrix}x_1\\x_2-x_1\\x_3-x_2\\x_4-x_3\end{bmatrix}$. So it records the initial amount, and traces the increases between terms.

In our case, $D\begin{bmatrix}1\\3\\5\\9\end{bmatrix} = \begin{bmatrix}1\\2\\3\\1\end{bmatrix}$. I started with 1 dollar, and during the first month I earned 2 dollars. During the second month I earned 3 dollars. And during the third month I earned 1 dollar.

Obviously this matrix is very useful in statistics. But furthermore, keep in mind that essentially, D is just a discrete version of "taking derivatives". For example, consider $D\begin{bmatrix}1\\2\\4\\8\end{bmatrix} = \begin{bmatrix}1\\1\\2\\4\end{bmatrix}$, i.e., the derivative of a exponential function is still exponential (ignoring the first coordinates).

There are some interesting properties if you apply this D to sequences. For example, D applied to an arithmetic sequence gives $D\begin{bmatrix}1\\3\\5\\7\\9\end{bmatrix} = \begin{bmatrix}1\\2\\2\\2\\2\end{bmatrix}$. This is because the arithmetic sequence as a function has constant derivative. So after D, its coordinates are constant (ignoring the first coordinate).

Applying D to a quadratic sequence gives $D\begin{bmatrix}1\\4\\9\\16\\25\end{bmatrix} = \begin{bmatrix}1\\3\\5\\7\\9\end{bmatrix}$, an arithmetic sequence (ignoring the first coordinate), because the derivative of a degree two polynomial is a degree one polynomial.

Finally, applying D to the Fibbonacci sequence gives the sequence again $D\begin{bmatrix}1\\1\\2\\3\\5\\8\\13\end{bmatrix} = \begin{bmatrix}1\\0\\1\\1\\2\\3\\5\end{bmatrix}$, just shifted by two terms. This is because the formula for the Fibbonacci numbers is the linear combination of two exponential functions, and hence the derivative of it is also a linear combination of these two exponential functions. ☺

Example 3.1.33. Now look at $D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$. It is obviously the difference between two matrices, $D = I - J$, where I is the identity matrix, and J is the matrix with ones immediately below the diagonal, and zeroes everywhere else. What do they do?

I obviously sends a sequence to itself. The identity map never changes anything. But J is the "shift down" operator. It shifts the sequence down a term, and sends $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$. Looking at this, you can probably see why $I - J$ gives the forward difference matrix. We have the original term (preserved by I) minus the previous term (that got shifted down by J).

Similarly one can look at J^{T}, the matrix with ones immediately above the diagonal, and zeroes everywhere else. Then $I - J^{\mathrm{T}}$ is the ***backward difference matrix***. It will send $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$ to $\begin{bmatrix} x_1 - x_2 \\ x_2 - x_3 \\ x_3 - x_4 \\ x_4 \end{bmatrix}$. ☺

A very interesting property of linear combinations of matrices is the law of distribution. We look back at the definition, and we have

Lemma 3.1.34 (Matrix-Vector law of distribution)**.** *For any $\boldsymbol{v} \in \mathbb{R}^n$, any $m \times n$ matrices A, B and any $x, y \in \mathbb{R}$, we have $(xA + yB)\boldsymbol{v} = x(A\boldsymbol{v}) + y(B\boldsymbol{v})$.*

Proof. This is just repeating the definition of linear combinations of maps. □

This can immediately be generalized to a full law of distribution on any matrix multiplications.

Proposition 3.1.35 (Law of distribution)**.** *For any matrices A, B, C and any $x, y \in \mathbb{R}$, we have $(xA + yB)C = x(AC) + y(BC)$ and $C(xA + yB) = xCA + yCB$ whenever the matrix multiplications/linear combinations involved are all well-defined.*

Proof. $(xA + yB)C$ is just $xA + yB$ applied to each column of C. So this immediately follows from the lemma above.

$C(xA + yB)$ is slightly trickier, but not much. Let $\boldsymbol{a}_i$ be the i-th column of A and $\boldsymbol{b}_i$ be the i-th column of B, then we know the i-th column of $xA + yB$ must be $x\boldsymbol{a}_i + y\boldsymbol{b}_i$.

So the i-th column of $C(xA + yB)$ is $C(x\boldsymbol{a}_i + y\boldsymbol{b}_i) = xC\boldsymbol{a}_i + yC\boldsymbol{b}_i$ by linearity of C. And this is exactly the i-th column of $x(CA) + y(CB)$. □

Remark 3.1.36. *By this time it should be pretty clear that "the law of distribution" is just a special case of linearity. In real numbers, when we have the law of distribution $a(b + c) = ab + ac$, we are really just saying that "multiplication by a" is linear.*

Pay special attention to the ORDER of multiplication, because for example AB might NOT be BA. So if you write $A(xB + yC) = xBA + yCA$, then it would be the WRONG formula. As a quick memorization tip, whatever is on the left before, it will still be on the left (because the map happens later in time). Similarly, whatever is on the right before, it will still be on the right.

Corollary 3.1.37. *$(A + B)^2 = A^2 + AB + BA + B^2$, and so on. If $AB = BA$, then in fact $(A + B)^2 = A^2 + 2AB + B^2$. However, if $AB \neq BA$, then we must have $(A + B)^2 \neq A^2 + 2AB + B^2$.*

Corollary 3.1.38. *$(A + I)^2 = A^2 + 2A + I$, and so on for other polynomial calculations.*

Definition 3.1.39. *Given a polynomial $p(x) = a_n x^n + \cdots + a_1 x + a_0$ and a square matrix A, we have $p(A) = a_n A^n + \cdots + a_1 A + a_0 I$.*

Corollary 3.1.40. *Given two polynomials $p(x), q(x)$ and a square matrix A, then $p(A)q(A) = q(A)p(A)$. And if $h(x) = p(x)q(x)$, then $h(A) = p(A)q(A) = q(A)p(A)$.*

Proof. Observe that polynomial is simply a linear combination of powers. For example, $p(x) = a_n x^n + \cdots + a_1 x + a_0$ means $p(x)$ is a linear combination of $x^n, \ldots, x^1, x^0$ with coefficient a_i for x^i. And $p(A)$ is simply the corresponding linear combination of powers of A.

You can summarize this proof using only one sentence: ***Since all powers of A commute, therefore all polynomials of A (i.e., linear combinations of powers of A) commute.***

Now we begin our proof. Let $p(x) = \sum_{i=0}^{m} a_i x^i$ and $q(x) = \sum_{j=0}^{n} b_j x^j$. Then we have

$$\begin{aligned}
p(A)q(A) =& (\sum_{i=0}^{m} a_i A^i)(\sum_{j=0}^{n} b_j A^j) \\
=& \sum_{i,j} a_i b_j A^i A^j \\
=& \sum_{i,j} a_i b_j A^{i+j} \\
=& \sum_{i,j} a_i b_j A^{j+i} \\
=& \sum_{i,j} a_i b_j A^j A^i \\
=& (\sum_{j=0}^{n} b_j A^j)(\sum_{i=0}^{m} a_i A^i) = q(A)p(A).
\end{aligned}$$

Now

$$h(x) = p(x)q(x) = (\sum_{i=0}^{m} a_i x^i)(\sum_{j=0}^{n} b_j x^j) = \sum_{i,j} a_i b_j x^{i+j}.$$

So from previous calculations we have

$$h(A) = \sum_{i,j} a_i b_j A^{i+j} = p(A)q(A).$$

□

Example 3.1.41. Consider the matrix $E = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. It sends $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ to $\begin{bmatrix} x_1 + x_2 \\ x_2 \\ x_3 \end{bmatrix}$. So what it does is to add the second coordinate onto the first.

Note that this matrix E differ from I at only one entry. The difference $X = E - I$ is a matrix whose $(1, 2)$ entry is 1, and all other entries are zero. It sends $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ to $\begin{bmatrix} x_2 \\ 0 \\ 0 \end{bmatrix}$. So it simply takes the second coordinate, and stuff it in the first coordinate, and clear everything else.

In particular, it is easy to see that $I + kX = \begin{bmatrix} 1 & k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ will sends $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ to $\begin{bmatrix} x_1 + kx_2 \\ x_2 \\ x_3 \end{bmatrix}$. Furthermore, it is also easy to check that $X^2 = O$, the zero matrix. In particular, $X^k = O$ whenever $k \geq 2$.

As a result, we have

$$(I + X)^k = I + kX + \text{Higher degree terms} = I + kX.$$

So $E^k = \begin{bmatrix} 1 & k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Hey! This is exactly how shearings would behave!

This is no surprise. Imagine what would happen if we have a parallelopiped with base on the xz-plane. The base would be fixed, and it would be tilted along the x-direction but the height is preserved (the height would be in the y-direction). This E is a 3-dimensional shear, and it preserves volumn instead of area. ☺

So to end this section, let us write out all different ways to write matrix multiplication AB.

1. $A \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix} = \begin{bmatrix} A\boldsymbol{b}_1 & \dots & A\boldsymbol{b}_n \end{bmatrix}$.
2. $\begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix} B = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} B \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} B \end{bmatrix}$.
3. $\begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{b}_1 & \dots & \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{b}_n \\ \vdots & \ddots & \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{b}_1 & \dots & \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{b}_n \end{bmatrix}$.
4. $\begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}} \end{bmatrix} = \sum \boldsymbol{a}_i \boldsymbol{b}_i^{\mathrm{T}}$.

The last one is new. Can you see why?

Proposition 3.1.42. $\begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}} \end{bmatrix} = \sum \boldsymbol{a}_i \boldsymbol{b}_i^{\mathrm{T}}$.

Proof. An entry-wise calculation proof goes like this. Say A is $m \times n$ and B is $n \times d$. The (i, j) entry of AB is simply $\sum_k a_{ik} b_{kj}$. So we have

$$AB = \begin{bmatrix} \sum_k a_{1k}b_{k1} & \dots & \sum_k a_{1k}b_{kd} \\ \vdots & \ddots & \vdots \\ \sum_k a_{mk}b_{k1} & \dots & \sum_k a_{mk}b_{kd} \end{bmatrix} = \sum_k \begin{bmatrix} a_{1k}b_{k1} & \dots & a_{1k}b_{kd} \\ \vdots & \ddots & \vdots \\ a_{mk}b_{k1} & \dots & a_{mk}b_{kd} \end{bmatrix} = \sum_k \boldsymbol{a}_k \boldsymbol{b}_k^{\mathrm{T}}.$$

So we are done.

Alternatively, to use fancy entry notation, consider the (i, j) entry $\boldsymbol{e}_i^{\mathrm{T}} AB \boldsymbol{e}_j$. We have $\boldsymbol{e}_i^{\mathrm{T}} \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}} \end{bmatrix} \boldsymbol{e}_j =$

$\begin{bmatrix} \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{a}_1 & \dots & \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{a}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{e}_j \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}}\boldsymbol{e}_j \end{bmatrix}$. Now this is just a row vector with a column vector, so it calculates like a dot product and gives $\sum_k (\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{a}_k)(\boldsymbol{b}_k^{\mathrm{T}}\boldsymbol{e}_j)$. Use associativity and linearity, we have $\boldsymbol{e}_i^{\mathrm{T}}(\sum_k \boldsymbol{a}_k \boldsymbol{b}_k^{\mathrm{T}})\boldsymbol{e}_j$. So we see that AB and $\sum_k \boldsymbol{a}_k \boldsymbol{b}_k^{\mathrm{T}}$ have the same (i, j)-entry for all i, j. So they are the same matrix.

Yet alternatively, for a third proof, consider the fact that $I = \sum \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}}$. Then we have

$$AB = AIB = A(\sum \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}})B = \sum A\boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}} B = \sum (A\boldsymbol{e}_i)(\boldsymbol{e}_i^{\mathrm{T}} B) = \sum \boldsymbol{a}_i \boldsymbol{b}_i^{\mathrm{T}}.$$

Associativity is so cool.... □

3.1.5 Transpose of Matrices

Now we talk about the transpose of a matrix.

Definition 3.1.43. *Given a $m \times n$ matrix A, its **transpose** is an $n \times m$ matrix A^{T} whose (i, j) entry is the (j, i) entry of A.*

Intuitively, A^{T} is just an entry-wise "reflection" of A about the diagonal entries. Like $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$.

Proposition 3.1.44. $(A^{\mathrm{T}})^{\mathrm{T}} = A$.

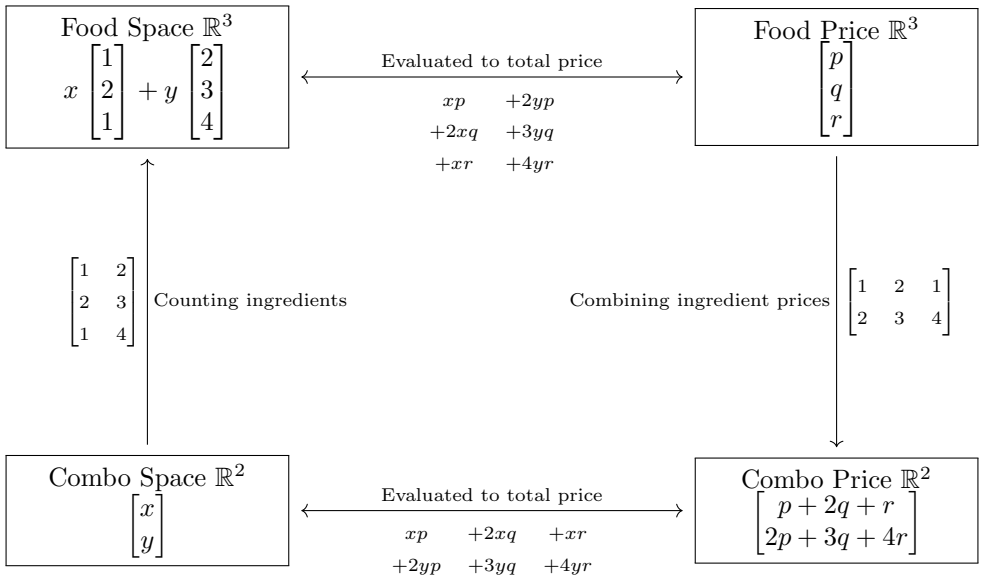
But what does this transpose mean?

It concers a deeper mathematical concept called a "dual space", which I cannot explain yet. (It is taught in next semester's class.) However, here is an example, and if you like, draw whatever intuition from it as much as you can.

Example 3.1.45. Suppose I want to burgers, chicken wings and cokes. There are two meal combos. Combo one contains 1 burger and 2 wings and 1 coke. Combo two contains 2 burger 3 wings and 4 cokes. Can you see a linear relation here?

Of course. If we have x combo one and y combo two, then we have $\begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ burgers, wings and cokes.

Now suppose the burgers, wings and cokes are p, q, r dollars each. Then the combos have prices $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix}$.



So, we are combining foods into meal combos. On the content side, we have A. On the evaluation side, we have A^{T}. "Counting ingredients" and "combining prices" are transpose of each other. ☺

Remark 3.1.46. *This remark is largely philosophical. Read it if you like, and forget about it if you cannot understand....*

Transpose is suppose to bounce between objects and evaluations of objects. To understand this, we first need a philosophical perspective: an object and its evaluation is in duality.

Mathematically speaking, given a set X, we can call a function $f : X \to \mathbb{R}$ an evaluation on X. Now let Y be the set of all evaluations on X.

Given $f \in Y$, obviously it evaluates $x \in X$ via $x \mapsto f(x)$. However, given $x \in X$, we can also use it to evaluate any $f \in Y$ via $f \mapsto f(x)$.

Consider this. Say that we have students doing two math tests. Then we have a function f sending students to their scores on the first test, and a function g sending students to their scores on the second test. So who is evaluating whom? An obvious answer is that functions are evaluating inputs. Each test evaluates students to the corresponding test score.

But one can also argue that the students are also evaluating the tests. For a fixed student, by doing both tests, this person is evaluating the differences between the tests, by giving different scores.

Such is the duality, where evaluations and "evaluatees" can switch places if we switch perspective. Transpose, philosophically speaking, is such a perspective switch, where we are switching the roles of evaluations and "evaluatees". What is a vector $\boldsymbol{v}$? Think of this as an object. What is a row vector $\boldsymbol{w}^{\mathrm{T}}$? Think of this as an evaluation of vectors. Then if A sends vectors forward to vectors, we have A^{T} sending evaluations backward to evaluations.

Now, philosophy and understanding aside, the biggest computational consequence of transpose is to make rows into columns and columns into rows. Recall that when we do AB, we are doing dot products of rows of A and columns of B. So if we do transpose, we are reversing the order of multiplication.

Proposition 3.1.47. *We have $(AB)^{\mathrm{T}} = B^{\mathrm{T}}A^{\mathrm{T}}$*

Proof. This is an entry-wise proof.

Write A in rows by $A = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix}$ and write B in columns by $B = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix}$. Then the (i, j) entry of AB is $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{b}_j$.

Now note that $A^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix}$ and $B^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_m^{\mathrm{T}} \end{bmatrix}$. So the (i, j) entry of $B^{\mathrm{T}}A^{\mathrm{T}}$ is $\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{a}_j$, which by commutativity of dot product equals to $\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{b}_i$. This is the (j, i) entry of AB. So $B^{\mathrm{T}}A^{\mathrm{T}} = (AB)^{\mathrm{T}}$.

Intuitively, since matrix multiplication uses the rows of the first matrix to dot the columns of the second matrix, by taking transpose (i.e., reversing the role of rows and columns), the order of multiplication is reversed.

(Also recall the idiom "rows to the left and columns to the right". If we switch the roles of rows and columns, then we are switching left and right. So the order of multiplication is reversed.) □

Proof. This is a cooler (albeit longer) proof. Again we uses special case that we know to be true, to establish a more general case.

First, by commutativity of dot product, we know $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}$. Note that transpose does nothing to 1 by 1 matrices. So $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})^{\mathrm{T}}$, and this is a special case of our desired statement.

Second, consider $(A\boldsymbol{x})^{\mathrm{T}}$. Say $A = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix}$, then

$$(A\boldsymbol{x})^{\mathrm{T}}$$

$$
\begin{aligned}
&=\left(\begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix} \boldsymbol{x}\right)^{\mathrm{T}} \\
&= \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} \end{bmatrix}^{\mathrm{T}} \\
&= \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{x} & \dots & \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}_1 & \dots & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}_m \end{bmatrix} \\
&=\boldsymbol{x}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_m \end{bmatrix} \\
&=\boldsymbol{x}^{\mathrm{T}} A^{\mathrm{T}}.
\end{aligned}
$$

So our statement is good when we have matrix multiplying vectors.

Finally, consider the full case. Say $B = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix}$. Then

$$
\begin{aligned}
&(AB)^{\mathrm{T}} \\
&=\left(A \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_n \end{bmatrix}\right)^{\mathrm{T}} \\
&= \begin{bmatrix} A\boldsymbol{b}_1 & \dots & A\boldsymbol{b}_n \end{bmatrix}^{\mathrm{T}} \\
&= \begin{bmatrix} (A\boldsymbol{b}_1)^{\mathrm{T}} \\ \vdots \\ (A\boldsymbol{b}_n)^{\mathrm{T}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}}A^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}}A^{\mathrm{T}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}} \end{bmatrix} A^{\mathrm{T}} \\
&=B^{\mathrm{T}}A^{\mathrm{T}}.
\end{aligned}
$$

The idea of using special case to prove a general case is a powerful one, and the "linear perspective" rather than "entry perspective" is nice, so I want you to see this proof. □

Remark 3.1.48. *Note that 1 by 1 matrices are unchanged by taking transpose. So we have* $\boldsymbol{e}_i^{\mathrm{T}}A\boldsymbol{e}_j = (\boldsymbol{e}_i^{\mathrm{T}}A\boldsymbol{e}_j)^{\mathrm{T}} = \boldsymbol{e}_j^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{e}_i$*, so the* (i,j) *entry of* A *is the same as the* (j,i) *entry of* A^{T}*, as expected.*

Another fine property of transpose is that it is linear.

Proposition 3.1.49. *We have* $(xA+yB)^{\mathrm{T}} = xA^{\mathrm{T}} + yB^{\mathrm{T}}$

Proof. This is trivial. Just name the entries and compute. Say $A = (a_{ij})_{m\times n}, B = (b_{ij})_{m\times n}$, then both sides has (i,j) entry $xa_{ji} + yb_{ji}$.

Or if you would like to practice fancy ways to write entries, you can do this, where I repeatedly used the fact that transpose of 1 by 1 matrix is itself:

$$
\boldsymbol{e}_i^{\mathrm{T}}(xA+yB)^{\mathrm{T}}\boldsymbol{e}_j = \boldsymbol{e}_j^{\mathrm{T}}(xA+yB)\boldsymbol{e}_i = x\boldsymbol{e}_j^{\mathrm{T}}A\boldsymbol{e}_i + y\boldsymbol{e}_j^{\mathrm{T}}B\boldsymbol{e}_i = x\boldsymbol{e}_i^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{e}_j + y\boldsymbol{e}_i^{\mathrm{T}}B^{\mathrm{T}}\boldsymbol{e}_j = \boldsymbol{e}_i^{\mathrm{T}}(xA^{\mathrm{T}} + yB^{\mathrm{T}})\boldsymbol{e}_j.
$$

□

3.2 Gaussian Elimination via Matrices

3.2.1 Elementary Matrices

Now it is time to review Gaussian eliminations. What are those row operations?

Consider the shear $E = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. We know it adds the second coordinate onto the first of the input vector. Now imagine applying E to a matrix on the left, EA. What would happen?

Well, say A in columns is $A = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix}$. Then $EA = E\begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix} = \begin{bmatrix} E\boldsymbol{a}_1 & \dots & E\boldsymbol{a}_n \end{bmatrix}$. So we are adding the second coordinate onto the first of EACH column vector. Effectively, we are adding the second row of A to the first row of A, i.e., $r_1 \to r_1 + r_2$. This is a row operation!

Remember how elementary operations preserve linear relations among the columns? That means exactly that these row operations are linear. So all of them are in fact matrix multiplications.

Definition 3.2.1. *The following matrices are called elementary matrices. (They corresponds to elementary row operations when we multiply them to the left of some matrix.) Here we assume $i \neq j$ are two indices.*

1. *A swap matrix is a matrix P_{ij} which has 1 on the (i,j) and (j,i) entry, 0 on the (i,i) and (j,j) entry, and otherwise identical to the identity matrix. For example, over $\mathbb{R}^7$, $P_{25} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.*
2. *A scale matrix is a diagonal matrix D with all diagonal entries non-zero. (We don't want to multiply both sides of an equation by zero....)*
3. *A shear matrix is a matrix E_{ij}^k with $i \neq j$, whose (i,j) entry is k, and is otherwise identical to the identity matrix.*

You can check yourself that they corresponds exactly to elementar row operations. The correspondence is like this:

1. $P_{ij}A$ means applying the row operation $r_i \leftrightarrow r_j$ to A.
2. DA means applying the row operation $r_i \to d_i r_i$ to A for all i, where d_i is the i-th diagonal entry of D.
3. $E_{ij}^k A$ means applying the row operation $r_i \to r_i + kr_j$ to A.

All of these should be remembered. However, if you have trouble remember them, just think this: for any elementary matrix E, we have $E = EI$. So the look of E as a matrix is exactly how the identity matrix I would be transformed, if we apply the corresponding row operation.

Remark 3.2.2. *So what are we doing when we do Gaussian elimination? We started with augmented matrix $\begin{bmatrix} A & \boldsymbol{b} \end{bmatrix}$. When we perform a row operation, we are multiplying some elementary matrix E to the left of it. So we are doing $E\begin{bmatrix} A & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} EA & E\boldsymbol{b} \end{bmatrix}$, and we now have a new augmented matrix.*

Equivalently, consider the equations $A\boldsymbol{x} = \boldsymbol{b}$. We can also multiply E from the left to both sides of the equations. This gives $EA\boldsymbol{x} = E\boldsymbol{b}$.

So as you can see, doing a row operations on the augmented matrix is EXACTLY the same as simply multiplying a matrix to both sides of the equation.

Example 3.2.3. Immediately I know that $E_{ij}^x E_{ik}^y = E_{ik}^y E_{ij}^x$. Why? Because it is obvious if you think of them as row operations. Similarly, for different indices i, j, k, l, we have $E_{ij}^x P_{kl} = P_{kl} E_{ij}^x$, because the corresponding row operations don't even touch each other. And so on so forth.

(In general, parallel things commute. Disjoint things commute. Entangled things are more likely to fail to commute.)

You can also easily see that $(E_{ij}^k)^t = E_{ij}^{kt}, E_{ij}^x E_{ij}^y = E_{ij}^{x+y}$, which justify the exponent notation. We even have E_{ij}^k and E_{ij}^{-k} as inverse map of each other. (All by simply looking at the meaning as row operations.)

In this sense, many calculations about these elementary matrices can simply be done without any entry-wise calculations. Just think what would happen as row operations. ☺

Example 3.2.4. In the example above, we have spotted many commuting behavior. However, there are also many non-commuting behaviors. Consider $E_{12} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $E_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

A calculation would reveal that $E_{12}E_{23} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$, while $E_{23}E_{12} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. Notice the difference in $(1,3)$ entry.

Essentially, we can think of it like this. Say $A = \begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} \\ \boldsymbol{r}_2^{\mathrm{T}} \\ \boldsymbol{r}_3^{\mathrm{T}} \end{bmatrix}$. $E_{12}E_{23}$ means we first do E_{23} and add the third row to the second row, so now we have $\begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} \\ \boldsymbol{r}_2^{\mathrm{T}} + \boldsymbol{r}_3^{\mathrm{T}} \\ \boldsymbol{r}_3^{\mathrm{T}} \end{bmatrix}$. Then we do E_{12} and add the second row to the first, and we have $\begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} + \boldsymbol{r}_2^{\mathrm{T}} + \boldsymbol{r}_3^{\mathrm{T}} \\ \boldsymbol{r}_2^{\mathrm{T}} + \boldsymbol{r}_3^{\mathrm{T}} \\ \boldsymbol{r}_3^{\mathrm{T}} \end{bmatrix}$. Note that the original third row is eventually carried over to the first row.

In comparison, $E_{23}E_{12}$ means we do E_{12} first and add the second row to the first, and now we have $\begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} + \boldsymbol{r}_2^{\mathrm{T}} \\ \boldsymbol{r}_2^{\mathrm{T}} \\ \boldsymbol{r}_3^{\mathrm{T}} \end{bmatrix}$. Then we do E_{23} and add the third row to the second, and now we have $\begin{bmatrix} \boldsymbol{r}_1^{\mathrm{T}} + \boldsymbol{r}_2^{\mathrm{T}} \\ \boldsymbol{r}_2^{\mathrm{T}} + \boldsymbol{r}_3^{\mathrm{T}} \\ \boldsymbol{r}_3^{\mathrm{T}} \end{bmatrix}$. As you can see, this order of operations means that the original third row never made it into the first row.

The difference in the $(1,3)$-entry between the matrices $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ means exactly this question: does the original third row make it into the first row or not? ☺

Now this is not the end of it. For an elementary matrix E, what if we do AE?

Consider AE_{12}^1 for 3 by 3 matrices. Note that $E_{12}^1 = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_1 + \boldsymbol{e}_2 & \boldsymbol{e}_3 \end{bmatrix}$. So if we write A in columns $A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_3 \end{bmatrix}$, then we haven

$$AE_{12}^1 = A \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_1 + \boldsymbol{e}_2 & \boldsymbol{e}_3 \end{bmatrix} = \begin{bmatrix} A\boldsymbol{e}_1 & A\boldsymbol{e}_1 + A\boldsymbol{e}_2 & A\boldsymbol{e}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_1 + \boldsymbol{a}_2 & \boldsymbol{a}_3 \end{bmatrix}.$$

So we are adding the first column to the second, i.e., $c_2 \to c_2 + c_1$! Remember the saying "row left column right"? Well congratulations, it just got a new meaning. For these elementary matrices, if applied to the left, then they are elementar row operations. If applied to the right, then they are in fact elementar column operations.

1. AP_{ij} means applying the row operation $c_i \leftrightarrow c_j$ to A.
2. AD means applying the row operation $c_i \to d_i c_i$ to A for all i, where d_i is the i-th diagonal entry of D.
3. AE_{ij}^k means applying the row operation $c_j \to c_j + kc_i$ to A.

Pay special attention to the shearings! For E_{ij}^k, as row operation it is $r_i \to r_i + kr_j$, where as column operations it is $c_j \to c_j + kc_i$, the meaning of the indices is reversed!

Remark 3.2.5. *There is more that is reversed. Consider the matrix multiplication of several elementary matrices, $E_1E_2 \dots E_k$. Which operation happen first?*

As row operations, we are doing $E_1E_2 \dots E_kA$, so actually E_k happens first, then E_{k-1}, and so on.

But as column operations, we are doing $AE_1E_2 \dots E_k$. So actually E_1 happens first, then E_2, and so on.

Here is an extra funny non-trivial thing to think about.

Example 3.2.6. Look at $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$. Let us first do $r_2 \to r_1 + r_2$ and have $\begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 9 \end{bmatrix}$, and then do $c_3 \to c_3 + c_2 + c_1$ and have $\begin{bmatrix} 1 & 2 & 6 \\ 5 & 7 & 21 \end{bmatrix}$.

What if we do $c_3 \to c_3 + c_2 + c_1$ first, and then do $r_2 \to r_1 + r_2$? Then we first get to $\begin{bmatrix} 1 & 2 & 6 \\ 4 & 5 & 15 \end{bmatrix}$, and then we have $\begin{bmatrix} 1 & 2 & 6 \\ 5 & 7 & 21 \end{bmatrix}$. We get to the same result.

A row op and a column op always commute. ☺

Proposition 3.2.7. *Given a matrix A, if we first do a row operation and then do a column operation, it is the same as first doing that column operation, then do that row operation.*

Proof. Say the row operation is represented by matrix E, and the column operation is represented by matrix F. Then we are saying $(EA)F = E(AF)$, which is true by associativity. □

3.2.2 Inverse Matrices

Proposition 3.2.8. *The inverse map of a bijective linear map is linear.*

Proof. Consider $f(f^{-1}(a\boldsymbol{v} + b\boldsymbol{w})) = a\boldsymbol{v} + b\boldsymbol{w} = af(f^{-1}(\boldsymbol{v})) + bf(f^{-1}(\boldsymbol{w})) = f(af^{-1}\boldsymbol{v} + bf^{-1}(\boldsymbol{w}))$. Now apply f^{-1} to both sides, and we have $f^{-1}(a\boldsymbol{v} + b\boldsymbol{w}) = af^{-1}\boldsymbol{v} + bf^{-1}(\boldsymbol{w})$ as desired. □

Definition 3.2.9. *We say a square matrix is **invertible** if it is bijective as a linear map. The matrix for its inverse map is denoted as A^{-1}, the **inverse** of A.*

In particular, a matrix is invertible if and only if it is bijective as linear maps. Since there is NO bijection between spaces of different dimension, only square matrices could have the possibility of being invertible. In fact, we already know the following when we talk about ranks.

Proposition 3.2.10. *Suppose A is a square matrix. Then TFAE*

1. *A is invertible (bijective).*
2. *A is injective.*
3. *A is surjective.*

These are enough to get us the following basic properties of matrices.

Proposition 3.2.11. *If A, B are square matrices (very important condition), then AB invertible implies that A, B are both invertible.*

Proof. If the composition AB is bijective, then B must be injective. (This is simply a property of any (not necessarily linear) maps.) But since B is square, this means that B is bijective.

Similarly, if the composition AB is bijective, then A must be surjective. (This is simply a property of any (not necessarily linear) maps.) But since A is square, this means that A is bijective. □

Corollary 3.2.12. *For square matrices A, B, if $AB = I$, then A, B are invertible and they are inverse of each other.*

Proof. I is bijective, so A, B are both invertible. Now apply A^{-1} to both sides from the left, we have $B = A^{-1}$.

Keep in mind that order of multiplications matter! If we were to apply A^{-1} to both sides from the right, we would have $ABA^{-1} = A^{-1}$, which would fail to tell us anything. □

About the order of multiplication, we have a very useful calculation law here.

Proposition 3.2.13. *If A, B are invertible (bijective), then AB is invertible. We have $(AB)^{-1} = B^{-1}A^{-1}$.*

Proof. If A, B are both bijective, then obviously their composition is also bijective.
$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$. So we are done. □

Note how the order of multiplication is reversed! This fact is true for all maps, not just linear ones. If you wear your sock and then wear your shoe, then you must take off your shoe and then take of your sock. The order is reversed. If you open a door and go out, then to undo it, you have to get in and then close the door. Again the order is reversed.

This is a very neat comparison with transpose. In fact, inverse and transpose play nicely with each other. Recall the following:

Proposition 3.2.14. *Suppose A is any matrix (maybe non-square). Then TFAE (short hand for "the followings are equivalent")*

1. *A is injective. ($\forall \boldsymbol{b}$, $A\boldsymbol{x} = \boldsymbol{b}$ has at most one solution.)*
2. *The columns of A are linearly independent.*

Proposition 3.2.15. *Suppose A is any matrix (maybe non-square). Then TFAE*

1. *A is surjective. ($\forall \boldsymbol{b}$, $A\boldsymbol{x} = \boldsymbol{b}$ has at least one solution.)*
2. *The rows of A are linearly independent.*

This immediately implies the following statement.

Corollary 3.2.16. *A is injective iff A^{T} is surjective, and A is surjective iff A^{T} is injective*

Proof. A is injective iff columns of A are linearly independent iff rows of A^{T} are linearly independent iff A^{T} is surjective.

The other one is very similar. □

Combining the two, we have the following.

Proposition 3.2.17. *A is invertible iff A^{T} is invertilbe, and $(A^{\mathrm{T}})^{-1} = (A^{-1})^{\mathrm{T}}$.*

Proof. $A^{\mathrm{T}}(A^{-1})^{\mathrm{T}} = (A^{-1}A)^{\mathrm{T}} = I^{\mathrm{T}} = I$. Yay. □

We are now at a very good place to talk about the law of cancellation. Suppose you have real numbers $x, y, z \in \mathbb{R}$, and you have an equation $xy = xz$. What would you do to simplify this? Surely you would "cancel" the x on both sides, and obtain $y = z$. This is called the law of cancellation. However, this is ONLY true if $x \neq 0$, and NOT true if $x = 0$.

Does matrix multiplications satisfy the law of cancellation? Consider these examples.

Example 3.2.18. Let $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. You can verify that $AA = O$. Yet we also have $AO = OA = O$. Thus we have $AA = AO$ and $AA = OA$. You can easily see that BOTH indicates that the law of cancellation FAILS for matrices.

In general, the term ***law of left cancellation*** refers to the phenomenon of $AB = AC$ implying $B = C$. For some A, this is true. For some other A, this is false.

Similarly, the term ***law of right cancellation*** refers to the phenomenon of $BA = CA$ implying $B = C$. For some A, this is true. For some other A, this is false. ☺

Proposition 3.2.19. *The law of left cancellation is true for A iff A as a linear map is injective. Dually, the law of right cancellation is true for A iff A as a linear map is surjective.*

Proof. Suppose the law of left cancellation is true for A. For any inputs $\boldsymbol{v}, \boldsymbol{w}$, if $A\boldsymbol{v} = A\boldsymbol{w}$, then applying left cancellation gives $\boldsymbol{v} = \boldsymbol{w}$. So A is injective.

Conversely, suppose that A is injective. Suppose $AB = AC$. Then for any input $\boldsymbol{v}$ for B and C, we would have $(AB)\boldsymbol{v} = (AC)\boldsymbol{v}$. Associativity then gives us $A(B\boldsymbol{v}) = A(C\boldsymbol{v})$. Now we use injectivity of A, which gives us $B\boldsymbol{v} = C\boldsymbol{v}$. So B and C give the same output for all inputs. Hence $B = C$. So the law of left cancellation is true for A.

So we have proven the statement about left cancellation. What about right cancellation? Well, observe that A has the law of left cancellation iff A^{T} has the law of right cancellation. At the same time, A is injective iff A^{T} is surjective. So we are done. □

As you can see, being injective, surjective or bijective will have a very tangible impact on our calculations with matrices.

Let us now see some examples of invertible matrices and transpose. Note that all matrices, even non-square matrices, have transpose. In contrast, even some square matrices has no inverse.

Example 3.2.20.

1. The inverse of the identity matrix is itself, obviously.
2. R_θ and $R_{-\theta}$ are inverse of each other. Funnily, they are also transpose of each other. Huh.
3. $I - \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$, i.e., projections to a hyperplane with unit normal vector $\boldsymbol{u}$, is not invertible. Apply this to $\boldsymbol{u}$ and you get $\boldsymbol{0}$, so this is not injective.
4. Householder transformations are inverse of themselves. (Because they are reflections.) Calculate and see: $(I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^2 = I$. (Is the inverse also the transpose?)
5. In the case of rotations and reflections, note that these maps preserves orthogonality. They send orthogonal inputs to orthogonal outputs. Later we shall see that transpose and orthogonality are very closely related.
6. Say $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$. Then it is invertible iff all d_i are non-zero, and its inverse in that case is $D^{-1} = \begin{bmatrix} d_1^{-1} & & \\ & \ddots & \\ & & d_n^{-1} \end{bmatrix}$. On the other hand, the transpose of D is just D itself.
7. Swaps are inverses of themselves. (This is in fact a special case of a Householder transformation.)
8. E_{ij}^k and E_{ij}^{-k} are inverse of each other. They are NOT transpose of each other. This is a very good place to look at the difference between transpose and inverse. E_{ij}^k as a row operation is $r_i \to r_i + kr_j$. Its transpose E_{ji}^k is $r_j \to r_j + kr_i$, whereas its inverse is E_{ij}^{-k}, which is $r_i \to r_i - kr_j$.

☺

In particular, all elementary matrices are invertible. As a result, we always have $A\boldsymbol{x} = \boldsymbol{b}$ if and only if $EA\boldsymbol{x} = E\boldsymbol{b}$ for any elementary matrix E. This explains right away why row operations fix solution set. And why don't we use column operations? Because it is hard to slide in E after A in this equation.

Now, if we are faced with a linear system $A\boldsymbol{x} = \boldsymbol{b}$, the most obvious reaction would be applying A^{-1} to both sides, and get $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. But how to find the inverse? We already know how: Gaussian elimination.

Proposition 3.2.21. *Given an invertible $n \times n$ matrix A, then its RREF is I.*

Proof. Informally this is obvious. Given any system, say the augmented matrix is $\begin{bmatrix} A & \boldsymbol{b} \end{bmatrix}$, then with A bijective, we should have a unique solution. All variables solved, which can only come from $\begin{bmatrix} I & AnswerVector \end{bmatrix}$.

Formally, consider any system $A\boldsymbol{x} = \boldsymbol{b}$. We have no free variables because A is injective, and no contradiction because A is surjective. So all n pivots are in the n columns of A. This gives no choice except that A will be row reduced to I. □

Corollary 3.2.22. *Any invertible matrix is a product of elementary matrices.*

Proof. Consider row reducing A to I. Say we used elementary row operations with matrices $E_1, \dots, E_k$. Then $E_k \dots E_1 A = I$. So $A = E_1^{-1} \dots E_k^{-1}$. □

Above corollary shows that, if B is any invertible matrix, then BA is essentially the same as applying some sequence of row operations on A. And similarly AB is the same as applying some sequence of column operations on A.

Corollary 3.2.23. *Given an invertible matrix A, consider the matrix $\begin{bmatrix} A & I \end{bmatrix}$. (This means we put A and I side to side to make a big rectangular matrix.) Then its RREF is $\begin{bmatrix} I & A^{-1} \end{bmatrix}$.*

Proof. Obviously $\begin{bmatrix} I & A^{-1} \end{bmatrix}$ is a RREF, so we only need to show that the two are row-equivalent, i.e., they can be transformed into each other using elementary row operations.

But if A is invertible, then A^{-1} is a sequence of row operations. Note that applying row operations to $\begin{bmatrix} A & I \end{bmatrix}$ is equivalent to applying the same row operations to A and to I simultaneously. So we have the following matrix multiplication calculation

$$A^{-1} \begin{bmatrix} A & I \end{bmatrix} = \begin{bmatrix} A^{-1}A & A^{-1}I \end{bmatrix} = \begin{bmatrix} I & A^{-1} \end{bmatrix}.$$

□

So this is how to find the inverse of a matrix.

Example 3.2.24. Consider the chicken rabbit matrix $\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}$. To find its inverse, we consider the augmented $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 4 & 0 & 1 \end{bmatrix}$. Now to RREF, we do $r_2 \to r_2 - 2r_1$, $r_2 \to \frac{1}{2}r_2$, and $r_1 \to r_1 - r_2$. This gives $\begin{bmatrix} 1 & 0 & 2 & -\frac{1}{2} \\ 0 & 1 & -1 & \frac{1}{2} \end{bmatrix}$. So the inverse to our original matrix is $\begin{bmatrix} 2 & -\frac{1}{2} \\ -1 & \frac{1}{2} \end{bmatrix}$. ☺

Proposition 3.2.25. *For an invertible 2×2 matrix, we have*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

So we swap the diagonal entries, and negate the non-diagonal ones. In particular, 2×2 matrix here is invertible iff $ad - bc \neq 0$.

Proof. We have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & ad \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix}.$$

So if $ad-bc \neq 0$, we see that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ multiply with $\frac{1}{ad-bc}\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ would give us the identity matrix. So this matrix is invertible and we have found its inverse.

If $ad-bc=0$, then the two column vectors are parallel. So A is not bijective. □

There is no easy formula for higher dimension cases. (Oh there is a formula alright, it is just too ugly to be useful.)

3.2.3 Inverses of Triangular Matrices

Here let us give pay special attention to triangular matrices.

Example 3.2.26. Let us calculate $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1}$. Consider $\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$. Now we do $r_2 \to r_2 - r_1, r_3 \to r_3 - r_1, r_4 \to r_4 - r_1$, and then we do $r_3 \to r_3 - r_2, r_4 \to r_4 - r_2$, and finally we do $r_4 \to r_4 - r_3$. We would have $\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 \end{bmatrix}$.

So $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$

As you can see, we only used upper rows to reduce lower rows, which is neat.

Similarly, we have $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, and this process would only involve using lower rows to reduce upper rows.

Recall our pretty pattern $LU = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$. Its inverse would be $U^{-1}L^{-1} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$.

(This last matrix is in fact related to heat distributions and wave distributions and so on. Its output coordinates are like $2x_i - (x_{i-1} + x_{i+1})$, so it is comparing each coordinate with the average of neighboring coordinates. Take heat distribution for example, if a place has higher temperature than its neighbors, then it will cool down.) ☺

Proposition 3.2.27. *A (lower or upper) trianglular matrix is invertible if and only if its diagonal entries are all non-zero.*

Proof. For upper triangular matrices, if all diagonal entries are non-zero, then we are already in REF and we see that we have full pivots. So it is invertible.

Conversely, suppose say the (i,i) entry is zero. (For example, consider $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 & 9 \\ 0 & 0 & 0 & 10 & 11 \\ 0 & 0 & 0 & 12 & 13 \\ 0 & 0 & 0 & 0 & 14 \end{bmatrix}$.)

Then the first i columns only uses $i-1$ rows, so they could produce at most $i-1$ pivots. (The first three columns in the example above contains only two pivots.) And the last $n-i$ columns obviously could produce at most $n-i$ pivots. (The last two columns in the example above contains only two pivots.) So we have at most $(i-1)+(n-i) = n-1$ pivots, which is less than full. So our matrix is NOT invertible.

For a lower triangular matrix, note that it is the transpose of an upper triangular matrix. □

Note that an invertible upper triangular matrix is already in REF. Given $U = \begin{bmatrix} a_1 & * & * \\ & \ddots & * \\ & & a_n \end{bmatrix}$, how would we further reduce it to RREF? Well, we first divide the i-th row by a_i to make all pivots into ones. Now the matrix is a unit triangular matrix.

Definition 3.2.28. *A triangular matrix (upper or lower) is **unit triangular** if all diagonal entries are 1.*

Proposition 3.2.29. *An invertible matrix is unit upper triangular iff it can be written as a series of row operations made ONLY of shearings, and ONLY use lower rows to change upper rows. (No swaps or scaling involved, and no using upper rows to change lower rows.)*

Similarly, an invertible matrix is unit lower triangular iff it can be written as a series of row operations made ONLY of shearings, and ONLY use upper rows to change lower rows. (No swaps or scaling involved, and no using lower rows to change upper rows.)

Proof. Given $U = \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 1 \end{bmatrix}$, we aim to show that $U = E_1 \dots E_k$ for shearings $E_1, \dots, E_k$ that only uses lower rows to change upper rows. This means $U = E_1 \dots E_k I$. So the question is now this: starting from the identity matrix, can we find shearings $E_1, \dots, E_k$ that only uses lower rows to change upper rows, and reach U?

Well, this is actually quite easy. Start from the second second column and work your way to the right, and you can see that this is always possible. For example, to reach $\begin{bmatrix} 1 & 2 & 3 & 4 \\ & 1 & 2 & 3 \\ & & 1 & 2 \\ & & & 1 \end{bmatrix}$, we can simply do

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \xrightarrow{r_1 \to r_1 + 2r_2} \begin{bmatrix} 1 & 2 & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \xrightarrow[r_2 \to r_2 + 2r_3]{r_1 \to r_1 + 3r_3} \begin{bmatrix} 1 & 2 & 3 & \\ & 1 & 2 & \\ & & 1 & \\ & & & 1 \end{bmatrix} \xrightarrow[r_3 \to r_3 + 2r_4]{\substack{r_1 \to r_1 + 4r_4 \\ r_2 \to r_2 + 3r_4}} \begin{bmatrix} 1 & 2 & 3 & 4 \\ & 1 & 2 & 3 \\ & & 1 & 2 \\ & & & 1 \end{bmatrix}.$$

For lower unit triangular matrices, you can do the similar thing (or simply take transpose). □

So we have the following intuitions:

1. Think of a unit upper triangular matrix as a series of shearings using lower rows to reduce upper rows.
2. Think of a unit lower triangular matrix as a series of shearings using upper rows to reduce lower rows.
3. A non-unit triangular matrix is simply a unit triangular matrix plus some row scaling.
4. Can you figure out the corresponding understanding of unit triangular matrices as column operations?

In particular, here are some nice corollaries.

Corollary 3.2.30. *Products of unit upper (lower) triangular matrices are still unit upper (lower) triangular. The inverse of a unit upper (lower) triangular matrix is still unit upper (lower) triangular.*

Proof. Products of unit upper (lower) triangular matrices means we do all the corresponding row operations, and they are all shearings using the lower rows to change upper rows. So the result is still unit upper (lower) triangular.

The inverse to $r_i \to r_i + kr_j$ is $r_i \to r_i - kr_j$. In particular, if we are using lower rows to change upper rows (i.e., $i > j$), then the inverse is also using lower rows to change upper rows. So the inverse is still unit upper (lower) triangular. □

Corollary 3.2.31. *If a matrix is upper (lower) triangular and invertible, then its inverse is also upper (lower) triangular with diagonal entries inverted.*

Proof. Just also invert the row scaling as well. □

Example 3.2.32. Sometimes the inverse of a unit triangular matrix is super easy to do. For example, $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix}$. Here the entries are simply negated. Why? Just think in terms of row operations and you shall see.

However, it is more annoying if the row operations are entangled. For example, $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$. The original matrix $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ is adding the first row to the second, then the second row to the third (and the original first row is in a sense "carried over"). To reverse, this, we are substracting the second row from the third, then substract the first row from the second (and the original first row now fail to influence the third row at all).

Basically, if all non-diagonal non-zero entries are in the same column or same row of a unit triangular matrix, then the inverse would simply negate them. Otherwise it is more complicated and you just have to do it the slow way. ☺

3.2.4 LU decomposition

We are returning to Gaussian elimination now with our new found toys and perspectives. What is Gaussian elimination?

Example 3.2.33. Suppose we have $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 4 & 6 & 8 \end{bmatrix}$. To do Gaussian elimination, first we shall attempt a top-down process where we use upper rows to reduce lower rows. In this case, we want $r_2 \to r_2 - 2r_1, r_3 \to r_3 - 4r_1$. This is the same as multiplying $L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}$ to the left of A. (Recall that the looks of L_1 is exactly what L_1 shall do to I, so you can quickly see that this is the right matrix.)

So we have $L_1 A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 2 & 4 \end{bmatrix}$. Next we want to do $r_3 \to r_3 - 2r_2$, so we are multiplying $L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}$ and get $L_2 L_1 A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & -2 \end{bmatrix}$. This is our REF, and obviously it is upper triangular. Let us call this U.

So we have $L_2 L_1 A = U$. Reorganize this, we see that $A = LU$ where $L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}^{-1} =$

$\begin{bmatrix}1&0&0\\2&1&0\\4&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\0&2&1\end{bmatrix}=\begin{bmatrix}1&0&0\\2&1&0\\4&2&1\end{bmatrix}$ is lower triangular and U is upper triangular. ☺

If you recall, our idea of elimination is to first go top down, using upper rows to reduce lower rows as much as possible, until we reach REF. This means we are using lower triangular shearing matrices, and the result we get to, an REF, is an upper triangular thing.

Suppose we got lucky, and have no need to use row swaps. Then we have lower triangular shearings $L_1, \dots, L_k$, such that $L_k \dots L_1 A = U$ for some upper triangular U. Then in particular, let $L = L_1^{-1} \dots L_k^{-1}$, we see that $A = LU$ for a lower triangular L and an upper triangular U.

All in all, this L records how to row reduce A, and U is the resulting REF.

Definition 3.2.34. *We say a square matrix A has an LU decomposition if $A = LU$ for a lower triangular L and upper triangular U.*

In essence, LU decomposition is just the matrix way of writing Gaussian elimination. Given the LU decomposition, it is very easy to tell what steps should you do. You can really just read it out.

Example 3.2.35. Suppose we have $A = LU$ as $\begin{bmatrix}1&1&1\\2&3&5\\4&6&8\end{bmatrix}=\begin{bmatrix}1&0&0\\2&1&0\\4&2&1\end{bmatrix}\begin{bmatrix}1&1&1\\0&1&3\\0&0&-2\end{bmatrix}$.

When we row reduce, the new row operations E would act as $EA = (EL)U$. So you can just focus on reducing L to I, and when you have reduced L to I successfully, A would turn into the REF matrix U automatically.

Looking at L column by column (left to right). Then we should do

$$L=\begin{bmatrix}1&0&0\\2&1&0\\4&2&1\end{bmatrix}\xrightarrow[r_3\to r_3-4r_1]{r_2\to r_2-2r_1}\begin{bmatrix}1&0&0\\0&1&0\\0&2&1\end{bmatrix}\xrightarrow{r_3\to r_3-2r_2}\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}=I.$$

Note that even without any thinking, I can literally read out the steps by reading the entries column by column (left to right).

Now we do the same thing to A, and we should reach U. Indeed, we have

$$A=\begin{bmatrix}1&1&1\\2&3&5\\4&6&8\end{bmatrix}\xrightarrow[r_3\to r_3-4r_1]{r_2\to r_2-2r_1}\begin{bmatrix}1&1&1\\0&1&3\\0&2&4\end{bmatrix}\xrightarrow{r_3\to r_3-2r_2}\begin{bmatrix}1&1&1\\0&1&3\\0&0&-2\end{bmatrix}=U.$$

In conclution, simply read out the entries of L, and we should know how to do elimination on A. Conversely, given an elimination process from an invertible matrix A to its REF, then we can read out the unit lower triangular matrix L from the elimination process. ☺

The LU decomposition is among the MOST widely used matrix decompositions. This is not just because gaussian elimination is useful, it is also because triangular matrices are super nice. For an $n \times n$ triangular matrix A, $A\boldsymbol{x} = \boldsymbol{b}$ can be solved in about $\frac{1}{2}n^2$ calculations, while in general solving $A\boldsymbol{x} = \boldsymbol{b}$ for a generic $n \times n$ matrix A takes about $\frac{1}{3}n^3$ calculations.

Remark 3.2.36. *Suppose we are going from A to its RREF. How many calculations do we need? Here if an entry changed once we count it as one calculation, and I am ignoring lower degree terms. And swapping does not count as calculations because no value is changed, only the storage location is swapped.*

If A is triangular, say lower triangular, then each entry below the diagonal would change exactly once during elimination. So we need a total of $(n-1)+(n-2)+\cdots+1$ calculations. Then we scale the diagonal entries to 1, which takes at most n calculations. So we need a total of $n+\cdots+1=\frac{1}{2}n^2+\frac{1}{2}n$ calculations.

If A is generic, since we ignored swapping, we can assume that we can just go forward elimination and then back. To use the first row to reduce all rows below, generically we need $n(n-1)$ calculations. Next, we use the second row to reduce below. Note that the first column is empty now, so we only need $(n-1)(n-2)$

calculations. And so on so forth. This takes $n(n-1)+(n-1)(n-2)+\cdots+2\times 1+1\times 0$ *calculations. Now we are left with an upper triangular matrix, which takes* $n+\cdots+1$ *calculations. So combine the two, we need* $n^2+\cdots+1^2=\frac{1}{3}n^3+\frac{1}{2}n^2+\frac{1}{6}n$ *calculations.*

So instead of solving $A\boldsymbol{x}=\boldsymbol{b}$, suppose we already know that $A=LU$. Then first we can solve $\boldsymbol{y}$ from $L\boldsymbol{y}=\boldsymbol{b}$, and then we can solve $\boldsymbol{x}$ from $U\boldsymbol{x}=\boldsymbol{y}$. This would take about n^2 steps instead of about $\frac{1}{3}n^3$ steps. When n is large, this makes a huge difference.

Remark 3.2.37. *However, keep in mind of this: even though solving* $LU\boldsymbol{x}=\boldsymbol{b}$ *is easy, finding* $A=LU$ *is not. In fact, finding* $A=LU$ *is exactly the Gaussian elimination, so it takes about* $\frac{1}{3}n^3$ *calculations. So what is the point after all?*

The point is that in practice, say we are doing a CT scan, then the A *is fixed, while the scan result* $\boldsymbol{b}$ *is unknown before hand. Suppose we have many future patients, then solving* $A\boldsymbol{x}=\boldsymbol{b}_1, A\boldsymbol{x}=\boldsymbol{b}_2,\ldots,A\boldsymbol{x}=\boldsymbol{b}_r$ *one by one would takes* $\frac{r}{3}n^3$ *calculations. On the other hand, if we do* $A=LU$ *before hand, then we only need* $\frac{1}{3}n^3+\frac{r}{2}n^2$ *calculations, which saves time. Or simply put,* LU *decomposition is basically doing most of Gaussian elimination beforehand, without even knowing what* $\boldsymbol{b}$ *is.*

Now, how about finding A^{-1} *and then simply apply* A^{-1} *to all these* $\boldsymbol{b}_1,\boldsymbol{b}_2,\ldots,\boldsymbol{b}_r$*? Well that works. But how would you find the inverse? You would use Gaussian elimination, which in essence is just* $A=LU$*. So finding* $A=LU$ *beforehand is basically the same as finding* A^{-1} *before hand, except that* LU *decomposition can be done even for non-invertible matrices.*

Now, LU decomposition might not be unique. For example, the zero matrix multiplying any upper triangular matrix would give the same answer. However, the uniqueness do exist under some special perspective.

Definition 3.2.38. *The LDU decomposition of* A *is* $A=LDU$ *where* L *is UNIT lower triangular,* U *is UNIT upper triangular, and* D *is diagonal.*

Intuitively, L is the forward elimination that gives you the REF. Then we scale pivots to 1, which is D, scaling of rows. Finally U represent the backward process that gives RREF.

Theorem 3.2.39. *If* A *is invertible and has* LU *decomposition, then the LDU decomposition of* A *is unique.*

Proof. Suppose $L_1D_1U_1=A=L_2D_2U_2$ are two LDU decompositions. I aim to show that $L_1=L_2, U_1=U_2, D_1=D_2$. Note that since A is invertible, all matrices involved here are invertible.

The key is to manipulate the equation into $L_2^{-1}L_1D_1=D_2U_2U_1^{-1}$. Now the left hand side is lower triangular, but the right hand side is upper triangular! How can a matrix be both lower triangular and upper triangular? Well it has to be diagonal, of course.

Furthermore, $L_2^{-1}L_1$ is in fact unit lower triangular, with ones on the diagonal. Now $L_2^{-1}L_1D_1$ means multiplying columns of $L_2^{-1}L_1$ with corresponding entries of D_1. So diagonal entries of $L_2^{-1}L_1D_1$ will just be the diagonal entries of D_1. But we already know that $L_2^{-1}L_1D_1$ is diagonal! So we must have $L_2^{-1}L_1D_1=D_1$. Since D_1 is invertible, we can simplify this to $L_1=L_2$.

Similarly we have $U_1=U_2$ on the right hand side. Finally, $L_2^{-1}L_1D_1=D_2U_2U_1^{-1}$ is now $D_1=D_2$. Done. □

Fun proof, yes?

Remark 3.2.40. *Why* LU *decomposition? Why not* UL *decomposition? This is largely a matter of convention.*

Suppose $A=LU$*. Then* $A^{-1}=U^{-1}L^{-1}$*, so you see that an* LU *decomposition of* A *corresponds to an* UL *decomposition for* A^{-1}*. So, if you want to do* UL *decomposition for* A*, you are essentially doing* LU *decomposition to* A^{-1}*.*

3.2.5 Diagonally Dominant Matrices

Now, is it possible to just recognize an invertible matrix by sight? Most of the time, no. However, we have seen some special cases, like diagonal matrices, triangular matrices and so on. These we can just tell by sight. Here is another kind.

Example 3.2.41. Imagine that a matrix has big entries on the diagonal, and tiny entries off the diagonal. Must it be invertible?

In 2×2 case, we have $\begin{bmatrix} \text{big} & \text{tiny} \\ \text{tiny} & \text{big} \end{bmatrix}$. It is quite obvious that the two columns are NOT parallel, hence the columns are linearly independent. So our square matrix must be invertible.

What about the 3×3 case? We have $\begin{bmatrix} \text{big} & \text{tiny} & \text{tiny} \\ \text{tiny} & \text{big} & \text{tiny} \\ \text{tiny} & \text{tiny} & \text{big} \end{bmatrix}$. To be invertible, it means it shall have full pivots (on the diagonal) after Gaussian elimination.

Imagine the first step of such an elimination. We would need to substract $\frac{\text{tiny}}{\text{big}}$ times the first row from rows below. This shall yeild something like

$$\begin{bmatrix} \text{big} & \text{tiny} & \text{tiny} \\ 0 & \text{big} - \frac{\text{tiny}}{\text{big}}\text{tiny} & \text{tiny} - \frac{\text{tiny}}{\text{big}}\text{tiny} \\ 0 & \text{tiny} - \frac{\text{tiny}}{\text{big}}\text{tiny} & \text{big} - \frac{\text{tiny}}{\text{big}}\text{tiny} \end{bmatrix}.$$

But as you can imagine, $\frac{\text{tiny}}{\text{big}}\text{tiny}$ is probably even more tiny. So it shall have little effect on the entries. We should expect big entries to remain big, and tiny entries to remain tiny. Hence we have something like

$$\begin{bmatrix} \text{big} & \text{tiny} & \text{tiny} \\ 0 & \text{big} & \text{tiny} \\ 0 & \text{tiny} & \text{big} \end{bmatrix}.$$

The next elimination step would similarly create

$$\begin{bmatrix} \text{big} & \text{tiny} & \text{tiny} \\ 0 & \text{big} & \text{tiny} \\ 0 & 0 & \text{big} \end{bmatrix}.$$

So we see that we shall have full pivots. Such a matrix must be invertible. This idea can easily be generalized to larger matrices (but the precise definition of "big" here is unfortunately quite blurry). ☺

Definition 3.2.42. *A square matrix $A = (a_{ij})_{n\times n}$ is (row) diagonally dominant if in each row, the absolute value of the diagonal entry is larger than the sum of absolute values of all other entries, i.e., $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for each i. (We can similarly define column diagonally dominant matrices.)*

Intuitively, a diagonally dominant matrix is basically a matrix that is "close to being diagonal".

Proposition 3.2.43. *A (row or column) diagonally dominant matrix is invertible.*

Before we prove it, let us first figure out when shall we see these matrices.

Example 3.2.44. Diagonally dominant matrix usually apear on things that are "mostly stable".

Let us predict the weather. Each hour, the weather could be sunny, cloudy or rainy. Furthermore, given any hour, the weather next hour is likely to be the same.

Suppose we know the following:

1. If an hour is sunny, then the next hour has 90% chance to still be sunny, and 10% chance to be cloudy.
2. If an hour is cloudy, then the next hour has 60% chance to be cloudy, and 20% chance to be rainy.

3. If an hour is rainy, then the next hour has 50% chance to be rainy, 30% chance to be cloudy, and 20% chance to be rainy.

A probability chart might look like this:

	Sunny Now	Cloudy Now	Rainy Now
Sunny Next	0.9	0.2	0.2
Cloudy Next	0.1	0.6	0.3
Rainy Next	0	0.2	0.5

.

Note that things are mostly stable because one hour is very likely to have the same weather as the next hour. We also have a natural matrix here. What linear map would this matrix represent?

Suppose a given hour has x, y, z chance to be sunny, cloudy or rainy respectively. We may write this info as a vector $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Then the next hour, we shall have $x \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0.2 \\ 0.6 \\ 0.2 \end{bmatrix} + z \begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$ chance of being sunny, cloudy or rainy.

To put this together, we see that the matrix $A = \begin{bmatrix} 0.9 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0 & 0.2 & 0.5 \end{bmatrix}$ is the transition process. It sends the probability distribution of an hour $\boldsymbol{v}$ to the probability distribution of the next hour $A\boldsymbol{v}$. In a sense, A helps us predict the future.

Note that A is (row) diagonally dominant. The diagonal entries specifically refers to the chances of a sunny hour remains sunny, a cloudy hour remains cloudy, and a rainy hour remains rainy. And these situations dominates. Essentially, this is because the weather usually likes to remain unchanged between two consecutive hours. As with most continuous process.

In situations such as heat diffusions, wave transmissions, probability evolutions and so on, you will always see such diagonally dominant matrices.

So what does it mean that these matrices can be inverted? Well, you can watch a film "Tenet" by Noland. When he creates a world where entropy is reversed, bullets go back into guns, cars drive backwards, flames freeze things, he is in a sense using the fact that these evolution matrices are invertible. ☺

How to show that something is invertible? Well, you do Gaussian elimination and see if you get full pivots, i.e., see if the RREF is the identity matrix. This can be rephrased into the following statement.

Lemma 3.2.45. *A square matrix A is invertible iff $\boldsymbol{x} = \boldsymbol{0}$ is the only solution to $A\boldsymbol{x} = \boldsymbol{0}$.*
(All outputs have unique pre-images iff the zero output has a unique pre-image.)

Proof. If A is invertible, then it is injective. So $A\boldsymbol{x} = \boldsymbol{0}$ has at most one solution, and therefore it has to be $\boldsymbol{0}$.

Now suppose $A\boldsymbol{x} = \boldsymbol{0}$ has a unique solution $\boldsymbol{x} = \boldsymbol{0}$. This means the augmented matrix $\begin{bmatrix} A & \boldsymbol{0} \end{bmatrix}$ has an RREF of $\begin{bmatrix} I & \boldsymbol{0} \end{bmatrix}$. So A has full pivots, and hence A is invertible. □

(Some textbook might WRONGLY tell you to look at the determinant of the matrix. However, finding determinant is almost always slower than doing elimination. So in practice you should NOT do that.)

Example 3.2.46. Now why are these things invertible?

Consider $\begin{bmatrix} 0.9 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0.9x + 0.2y + 0.2z \\ 0.1x + 0.6y + 0.3z \\ 0.2y + 0.5z \end{bmatrix}$. Being diagonally dominant means each output coordinate depends the MOST on the corresponding input coordinates, and less on other input coordinates.

To check for invertibility, let us consider how to make the output zero.

The first output coordinate is $0.9x + 0.2y + 0.2z$. Note that, if x has the largest non-zero absolute value among x, y, z ($|x| \geq |y|$ and $|x| \geq |z|$), then this cannot be zero. y, z are already small, and they got

multiplied with non-diagonal entries, which are also small. x is already large, and it got multiplied with the diagonal entry, which is also large.

In particular, any solution to $\begin{bmatrix} 0.9 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{0}$ must NOT have x with the largest non-zero absolute value among x, y, z.

But similarly, the solution cannot have y with the largest non-zero absolute value, and cannot have z with the largest non-zero absolute value. We are left with one choice, $x = y = z = 0$. ☺

Proof of the Propositions. Suppose A is row diagonally dominant, and $\boldsymbol{x} \neq \mathbf{0}$. Say $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ where x_i has the largest absolute value among all coordinates. Then let us consider the i-th coordinate of the output.

If $A = (a_{ij})_{n\times n}$, then the i-th coordinate of the output is

$$|\sum_j a_{ij}x_j| \geq |a_{ii}||x_i| - \sum_{j\neq i} |a_{ij}||x_j| \geq |a_{ii}||x_i| - \sum_{j\neq i} |a_{ij}||x_i| = (|a_{ii}| - \sum_{j\neq i} |a_{ij}|)|x_i| > 0.$$

So $A\boldsymbol{x} \neq 0$. We cannot combine columns of A non-trivially to get $\mathbf{0}$. So A has linearly independent column, and A is injective and hence bijective.

Finally, if A is column diagonally dominant, then A^{T} is row diagonally dominant, so we are done. □

If you check out the Chinese textbook by Liang, Tian and myself, there is also a cool example on heat distribution on wires about diagonally dominant matrix.

3.3 Block Matrices

3.3.1 Meaning of Blocks

There are two major techniques when it comes to matrices. The first one is decompositions, and we have already seen it in the form of LU decompositions or LDU decompositions. Given a complicated map, we decompose it into the composition of a chain of simpler maps.

The second one is to utilize block matrices, which is what we shall do here. We start with a mystery.

Example 3.3.1. Consider the transformation on $\mathbb{R}^3$ that rotate the xy-plane by 45 degree counterclockwise and stretch the z-axis by a factor of 2. So by looking at images of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, we see that the matrix is $\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

This is NOT a diagonal matrix. However, it is a ***block diagonal matrix***. If you divide it into four blocks as $\left[\begin{array}{cc|c} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$, you see that the non-diagonal blocks are all zero.

Now if you take inverse, we have $\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$. Hey, this is just as if we are taking inverse of each diagonal block!

If we take the inverse of a diagonal matrix, we just invert each diagonal entry. It seems that, if we take the inverse of a block diagonal matrix, we can just invert each diagonal block.

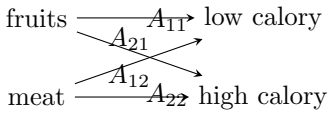
In fact, the two blocks here behaves "independently". The upper left block only uses the first two coordinates to change the first two coordinates (rotating the xy-plane), while the lower right block only uses the third coordinate to change the third coordinate (stretching the z-axis). No wonder they got inverted independently. ☺

As you can see, block matrices are NOT just a formality in grouping entries. It in fact has meanings. Each individual block is in fact a linear "submap" in some sense.

Example 3.3.2. Consider a map sending foods to nutrients. Say we have foods: apples, bananas, meat. And we have nutrients: fibers, proteins, suger. Then this map is a matrix A, such that if we have $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ apples, bananas and meat, then we have $A \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ fibers, proteins and suger. Obviously A is a 3 by 3 matrix.

Now consider the block form $A = \left[\begin{array}{cc|c} a & b & c \\ d & e & f \\ \hline g & h & i \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where A_{ij} represent the corresponding blocks.

What does A_{11} do? It sends fruits to the low calory nutrients they contain. What does A_{12} do? It send fruits to the high calory nutrients they contain. What does A_{21} do? It sends meat to the low calory nutrients it contains. What does A_{22} do? It send meat to the high calory nutrients it contains.

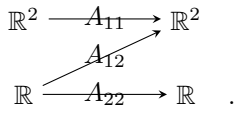


And what is A? A as a linear map is simply the collection of these four linear maps. ☺

Intuitively, when we have a block matrix, we are grouping input coordinates and output coordinates. The block A_{ij} records how the j-th group of inputing coordinates effect the i-th group of outputing coordinates.

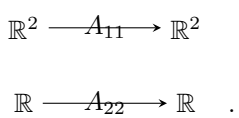
Example 3.3.3. Consider $\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 1 & 2 \\ \hline 0 & 0 & 1 \end{array}\right]$. Note that the lower left block is zero. This means the first two input coodinates does NOT effect the third output coordiante.

Indeed we have $\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 1 & 2 \\ \hline 0 & 0 & 1 \end{array}\right] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x+y+z \\ x+y+2z \\ z \end{bmatrix}$.



This is a ***block upper triangular matrix***.

In particular, block diagonal means each groups of coordinates only effect themselves. In particular, instead of one system, it is more like many separate independent systems, one for each diagonal block. Here is a picture for $\left[\begin{array}{cc|c} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$.



☺

Definition 3.3.4. *A **block diagonal matrix** is a matrix whose block form is* $\begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}$ *for square matrices* $A_1, \ldots, A_k$.

*A **block upper triangular matrix** is a matrix whose block form is* $\begin{bmatrix} A_1 & * & * \\ & \ddots & * \\ & & A_k \end{bmatrix}$ *for square matrices* $A_1, \ldots, A_k$. *One can similarly define a **block lower triangular matrix**.*

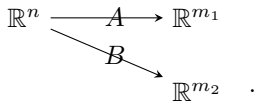
Remark 3.3.5. *Note that we usually require the diagonal blocks to be square. For example,* $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 3 & 4 \end{bmatrix}$ *is NOT considered a block diagonal matrix or a block upper triangular matrix.*

We want nice formula such as $\begin{bmatrix} A_1 & * & * \\ & \ddots & * \\ & & A_k \end{bmatrix}^{-1} = \begin{bmatrix} A_1^{-1} & * & * \\ & \ddots & * \\ & & A_k^{-1} \end{bmatrix}$, *so we want these blocks to be square.*

We are now ready to do some calculation to make these ideas rigorous.

Proposition 3.3.6. $\begin{bmatrix} A \\ B \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} A\boldsymbol{x} \\ B\boldsymbol{x} \end{bmatrix}$ *given that* A, B *has as much columns as coordinates of* $\boldsymbol{x}$.

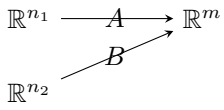
Proof. Write A, B in row vectors and this is trivial. Here is a graph if A is $m_1 \times n$ and B is $m_2 \times n$.



□

Proposition 3.3.7. $\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = A\boldsymbol{x} + B\boldsymbol{y}$ *given that* A *has as much columns as coordinates of* $\boldsymbol{x}$, *and* B *has as much columns as coordinates of* $\boldsymbol{y}$.

Proof. Write A, B in column vectors and this is trivial. Here is a graph if A is $m \times n_1$ and B is $m \times n_2$.



□

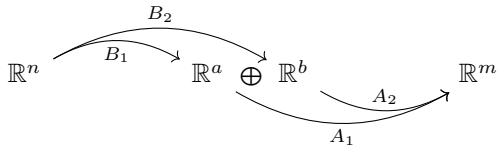
Lemma 3.3.8. *For matrices (in block form)* $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ *and* $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$, *where* A_1 *has the same number of columns as the number of rows of* B_1, *and* A_2 *has the same number of columns as the number of rows of* B_2. *Then* $AB = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1B_1 + A_2B_2$. *(Just like how a horizontal vector acts on a vertical vector.)*

Proof. For any input vector $\boldsymbol{v}$, we have

$$AB\boldsymbol{v} = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \boldsymbol{v} = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1\boldsymbol{v} \\ B_2\boldsymbol{v} \end{bmatrix} = A_1B_1\boldsymbol{v} + A_2B_2\boldsymbol{v}$$

So in short, $AB\boldsymbol{v} = (A_1B_1 + A_2B_2)\boldsymbol{v}$.

Here is a nice graph to see it. From the domain $\mathbb{R}^n$ of AB to the codomain $\mathbb{R}^m$ of AB, there are two "routes", one is via A_1B_1, and the other is via A_2B_2. So in total we just have $A_1B_1 + A_2B_2$.



□

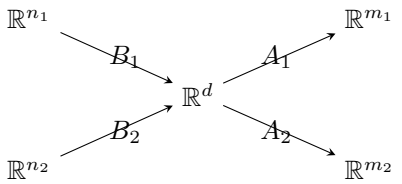
Remark 3.3.9. *Compare this with the formula of* $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \sum v_i w_i$, *and* $\begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_n^{\mathrm{T}}\end{bmatrix} = \sum \boldsymbol{a}_i\boldsymbol{b}_i^{\mathrm{T}}$.

Lemma 3.3.10. *For matrices (in block form)* $A = \begin{bmatrix}A_1 \\ A_2\end{bmatrix}$ *and* $B = \begin{bmatrix}B_1 & B_2\end{bmatrix}$, *where* A_1, A_2 *has the same number of columns as the number of rows of* B_1, B_2. *Then* $AB = \begin{bmatrix}A_1 \\ A_2\end{bmatrix}\begin{bmatrix}B_1 & B_2\end{bmatrix} = \begin{bmatrix}A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2\end{bmatrix}$. *(Just like how a column vector acts on a row vector.)*

Proof. For any input vector $\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix}$, we have

$$\begin{aligned}
&AB\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}A_1 \\ A_2\end{bmatrix}\begin{bmatrix}B_1 & B_2\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}A_1 \\ A_2\end{bmatrix}(B_1\boldsymbol{x} + B_2\boldsymbol{y}) \\
&= \begin{bmatrix}A_1B_1\boldsymbol{x} + A_1B_2\boldsymbol{y} \\ A_2B_1\boldsymbol{x} + A_2B_2\boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}\begin{bmatrix}A_1B_1 & A_1B_2\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix} \\ \begin{bmatrix}A_2B_1 & A_2B_2\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix}\end{bmatrix} \\
&= \begin{bmatrix}A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix}.
\end{aligned}$$

Here is a nice graph to see it. Say the domain of AB is $\mathbb{R}^{n_1+n_2}$, and codomain is $\mathbb{R}^{m_1+m_2}$, and the codomain of B and domain of A is $\mathbb{R}^d$. Then the only route from $\mathbb{R}^{n_i}$ to $\mathbb{R}^{m_j}$ is A_jB_i.



□

You can see the trend here. You can simply pretend that these blocks are "entries", and just do it as if you are doing matrix multiplications. You'll be just fine.

(But be careful of the order of multiplication. Whatever is originally on the left, it will end up on the left. E.g. $\begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1B_1 + A_2B_2$ is correct, while $\begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = B_1A_1 + B_2A_2$ is WRONG.)

We merely do a simple case here, since it contains all the crucial ideas for a generalization.

Proposition 3.3.11. *Sometimes we can divide a matrix into tiny blocks. Then matrix multiplications can be done block-wise, as if the blocks are just entries. For example, if* $A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$ *and* $B = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$, *where* A_1 *is* m_1 *by* n_1, A_2 *is* m_1 *by* n_2, A_3 *is* m_2 *by* n_1, A_4 *is* m_2 *by* n_2, B_1 *is* n_1 *by* r_1, B_2 *is* n_1 *by* r_2, B_3 *is* n_2 *by* r_1, B_4 *is* n_2 *by* r_2.

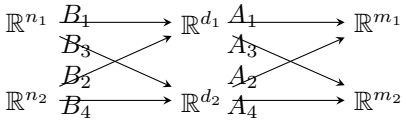
Then $AB = \begin{bmatrix} A_1B_1 + A_2B_3 & A_1B_2 + A_2B_4 \\ A_3B_1 + A_4B_3 & A_3B_2 + A_4B_4 \end{bmatrix}$. *Just as you would expect from doing multiplications of 2 by 2 matrices.*

Essentially we can divide A and B into as many blocks as we want, and multiply A and B by pretending that each block is just some number, as long as all block multiplications involved are well-defined.

Proof. Think of A as two column blocks and B as two row blocks, we have

$$\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_3 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \end{bmatrix} + \begin{bmatrix} A_2 \\ A_4 \end{bmatrix} \begin{bmatrix} B_3 & B_4 \end{bmatrix} = \begin{bmatrix} A_1B_1 + A_2B_3 & A_1B_2 + A_2B_4 \\ A_3B_1 + A_4B_3 & A_3B_2 + A_4B_4 \end{bmatrix}.$$

A picture looks like this:



As you can see from above, to move from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{m_1}$, you can do A_1B_1 or A_2B_3. So the corresponding upper left block after the multiplication is $A_1B_1 + A_2B_3$. □

As you can imagine, this is probably NOT going to make computation easier. However, it will make certain special case easier. For example you can have the following:

1. $\begin{bmatrix} A_1 & O \\ O & A_2 \end{bmatrix} \begin{bmatrix} B_1 & O \\ O & B_2 \end{bmatrix} = \begin{bmatrix} A_1B_1 & O \\ O & A_2B_2 \end{bmatrix}$. Here O means a block of all zero entries.
2. When invertible, $\begin{bmatrix} A & O \\ O & B \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & O \\ O & B^{-1} \end{bmatrix}$.
3. Products of block triangular matrices are block trianguler, and the diagonal blocks of the product is the product of corresponding diagonal blocks.

We end this section with an super mysterious special scenario, where there is no block in sight, yet the essense of a block matrix is still there.

Example 3.3.12. Sometimes there are "hidden blocks".

Consider the matrix $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 5 & 0 & 3 \end{bmatrix}$. You can check that its inverse is $\begin{bmatrix} 3 & 0 & -1 \\ 0 & \frac{1}{2} & 0 \\ -5 & 0 & 2 \end{bmatrix}$. Curiously, you may also check that $\begin{bmatrix} 2 & 1 \\ 5 & 3 \end{bmatrix}$ and $\begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$ are inverse of each other. Hey! It is as if in the original matrix

$\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 5 & 0 & 3 \end{bmatrix}$, the four corner entries form a "secret" block and the middle entry form the other block, and we are block diagonal!

How can this be? Recall the REASON behind the behavior of block diagonal matrices. It is because groups of coordinates are independent of each other. Now consider $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 5 & 0 & 3 \end{bmatrix}$, and you can see that the first coordinate and the third coordinate will NOT effect the second coordinate, and the second coordinate will NOT effect the first and third coordinate. So it indeed have the desired behavior. It is "secretly" block diagonal. ☺

3.3.2 Block Elimination and Block Inverse

Recall that elementary row operations are essentialy multiplying matrices from the left. Now sometimes we get lazy, and we will try to do MANY operations together, and do a BLOCK row operation.

Example 3.3.13. Consider $\begin{bmatrix} O & I_a & O \\ I_a & O & O \\ O & O & I_b \end{bmatrix}$. When acting on matrices with $2a + b$ rows from the left, this matrix will swap the first a rows with next a rows. In particular, $\begin{bmatrix} O & I_a & O \\ I_a & O & O \\ O & O & I_b \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} B \\ A \\ C \end{bmatrix}$ where A, B have a rows each, and C has b rows. ☺

Example 3.3.14. Consider $\begin{bmatrix} I_a & O \\ X & I_b \end{bmatrix}$ and $\begin{bmatrix} A \\ B \end{bmatrix}$ where A has a rows, B has b rows and X is any $a \times b$ matrix. We have $\begin{bmatrix} I_a & O \\ X & I_b \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A \\ B + XA \end{bmatrix}$. We are adding X times the first block row to the second block row. ☺

Example 3.3.15. Consider $\begin{bmatrix} X & O \\ O & Y \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} XA \\ YB \end{bmatrix}$. This is block row scaling. ☺

Similarly, if we apply block swapping matrices, block shearing matrices and block diagonal matrices to the right, then we are doing block column operations.

Example 3.3.16. How to find the inverse of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, provided that A is invertible?

We can first row reduce downward as we are used to. To kill C, I need to substract CA^{-1} copies of the top block rom from the bottom block row. (Note that $A^{-1}C$ would be WRONG here. The order matters.)

This gives $\begin{bmatrix} I & O \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ O & D - CA^{-1}B \end{bmatrix}$. Note that the matrix on the right hand side is invertible iff both diagonal blocks are invertible.

This yields the following theorem. ☺

Definition 3.3.17. *For a block matrix* $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *where* A, D *are square, the* ***Schur complement*** *of* A *is* $D - CA^{-1}B$ *when* A *is invertible, and the Schur complement of* D *is* $A - BD^{-1}C$ *when* D *is invertible.*

Proposition 3.3.18. *For a block matrix* $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$*, if* A *and its Schur complement are both invertible, then* M *is invertible with inverse* $\begin{bmatrix} I & -A^{-1}B \\ O & I \end{bmatrix} \begin{bmatrix} A^{-1} & O \\ O & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & O \\ -CA^{-1} & I \end{bmatrix}$.

The proof is basically finishing the block LDU decomposition and then invert each matrix. The LDU decomposition is

$$M = \begin{bmatrix} I & O \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & O \\ O & (D - CA^{-1}B) \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ O & I \end{bmatrix}.$$

Don't memorize this formula. Rather, just remember the idea of block operations, and don't be surprised when you do block operations and see a Schur complement somewhere.

3.3.3 Woodburry formula and Sherman-Morrison formula (Optional)

Block matrix multiplications are no doubt super helpful, but due to the non-commutativity, expressions could also get super long. For example, when A and its Schur complements are invertible, we have an expression for the inverse of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Multiply out the previous results, we have

$$\begin{aligned} &\begin{bmatrix} I & -A^{-1}B \\ O & I \end{bmatrix} \begin{bmatrix} A^{-1} & O \\ O & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & O \\ -CA^{-1} & I \end{bmatrix} \\ =& \begin{bmatrix} A^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ O & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & O \\ -CA^{-1} & I \end{bmatrix} \\ =& \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}. \end{aligned}$$

But when D and its Schur complements are invertible, we have another expression for the inverse of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Symetrically, this is

$$\begin{bmatrix} (A - BD^{-1}C)^{-1} & -D^{-1}C(A - BD^{-1}C)^{-1} \\ -(A - BD^{-1}C)^{-1}BD^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

Compare the upper left block in both expressions you would obtain the famous Woodbury formula, true whenever A, D are invertible and B, C have the correct number of rows and columns:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

The formula is powerful and useful in some fields. Although it looks horrible, note that we have Schur complements everywhere. In fact, the ugly portion $(A^{-1}B)(D - CA^{-1}B)^{-1}(CA^{-1})$, the three portions corresponds exactly to the LDU decomposition entries. Of course, there is no need to memorize this though. (I won't test it in the final.) Whenever you need this in the future, just google it again to get this formula.

Example 3.3.19. Let us do some special case of this. Suppose $A = I, B = \boldsymbol{e}_i, C = \boldsymbol{e}_j^{\mathrm{T}}, D = -I$ where $i \neq j$. Then we have a formula $(I - \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}})^{-1} = I + \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}$. Is this true?

Well it is. These are just shearing matrices for a single row operation, and obviously the inverse of that shearing behaves this way. Say when $i = 2, j = 1$, we have $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$.

Let us be more general. Consider the inverse $I + \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$. Forget about the precise Woodburry formula, and rather just remember this: its inverse is $I - k\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ for some number k. Now look at the equation $(I + \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})(I - k\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}) = I$, and you can solve and get $k = \frac{1}{1+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}}$.

The formula $(I + \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})^{-1} = I - \frac{\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}}{1+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}}$ is much more useful and easier to remember.

Now substitute $\boldsymbol{u}$ by $A^{-1}\boldsymbol{u}$, and multiply A^{-1} on both sides from the right, we have $(A + \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})^{-1} = A^{-1} - \frac{(A^{-1}\boldsymbol{u})(\boldsymbol{v}^{\mathrm{T}}A^{-1})}{1+\boldsymbol{v}^{\mathrm{T}}A^{-1}\boldsymbol{u}}$. This is sometimes called the Sherman-Morrison formula.

Matrices that looks like $\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ are also called rank-one matrices sometimes, because they can only have one pivot. Don't memorize the precise Sherman-Morrison. Rather, keep in mind of what it is trying to say: if we change a matrix A by a rank-one matrix, then its inverse would also change by some rank-one matrix. In fact, the Woodburry formula can be think of as a rank r version of this. ☺

Example 3.3.20. Here is another nice applicatin of the Woodburry formula. Apply this to $(I+AB)^{-1}$, and we have $(I+AB)^{-1} = I - A(I+BA)^{-1}B$. Note that A, B here does not need to be square. As long as AB is square, we are fine. Also note that AB, BA might be suqare matrices of different dimensions.

In particular, this implies that $I+AB$ is invertible iff $I+BA$ is invertible. ☺

Remark 3.3.21. *Here is an alternative proof of the fact above, that $(I+AB)^{-1} = I - A(I+BA)^{-1}B$. What is the fundamental relation between $I+AB$ and $I+BA$? It is the following fact, called a push-over identity: $A(I+BA) = (I+AB)A$ and $(I+BA)B = B(I+AB)$. Can you see the "pushing over"?*

Now inverting the $I+AB, I+BA$, and we have the push-over identity in the form of $(I+AB)^{-1}A = A(I+BA)^{-1}$.

To complete the proof, we have $I = (I+AB)^{-1}(I+AB) = (I+AB)^{-1} + (I+AB)^{-1}AB$. Now use the above identity, and we have the desired formula.

Remark 3.3.22. *This remark is entirely optional. By the summation of geometric series, we know that for real numbers $|x| < 1$, we have the formula*

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + ...$$

Apply this to AB or BA, we have

$$(I+AB)^{-1} = I - AB + ABAB - ABABAB + ...$$

$$(I+BA)^{-1} = I - BA + BABA - BABABA + ...$$

Now note that $A(I+BA)^{-1}B$ is exactly $AB - ABAB + ABABAB - ....$ This is exactly a part of the formula for $(I+AB)^{-1}$, wow! Hence we have $(I+AB)^{-1} = I - A(I+BA)^{-1}B$.

The arguments so far are not rigorous yet, and there are some convergence issues to think about. Nevertheless, it gives a very nice intuition about this fact.

3.3.4 Symmetric Matrices and LDL$^{\mathrm{T}}$ Decomposition

You surely have wondered about this. Some matrices are very pretty, in that they equal to their own transpose, say $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$.

Definition 3.3.23. *A matrix A is said to be **symmetric** if $A = A^{\mathrm{T}}$.*

So these are matrices like $\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$, where the entries are symmetric about the diagonal.

Proposition 3.3.24. *For any square matrix A, $A + A^{\mathrm{T}}$ is symmetric.*

For any $m \times n$ matrix A, AA^{T} and $A^{\mathrm{T}}A$ are symmetric.

Proof. Direct calculation. □

Now it sure looks pretty on the outside. Luckily, it is also pretty on the inside.

Proposition 3.3.25. *If $A = A^{\mathrm{T}}$ is invertible, and it has an LU decomposition, then in fact we have $A = LDL^{\mathrm{T}}$ where L is unit lower triangular, and D is diagonal.*

Proof. We know the LDU decomposition is unique. Now consider $A = LDU$ and $A = A^{\mathrm{T}} = (LDU)^{\mathrm{T}} = U^{\mathrm{T}}D^{\mathrm{T}}L^{\mathrm{T}}$, and note that these are two LDU decomposition for A. So we must have $U = L^{\mathrm{T}}$. □

We sometimes call this the LDL$^{\mathrm{T}}$ decomposition. It is in fact faster than the traditional elimination (only take half as much time). Why? Because symmetricity means we only need to work with entries below the diagonal.

Example 3.3.26. Suppose we have $\begin{bmatrix}1&2&3&4\\2&5&6&7\\3&6&8&9\\4&7&9&10\end{bmatrix}$. For the first step of the elimination, we use the first row to reduce below, and we have $\begin{bmatrix}1&2&3&4\\0&1&0&-1\\0&0&-1&-3\\0&-1&-3&-6\end{bmatrix}$. Huh. The lower right block is still symmetric!

In fact, this is no coincidence. Say our symmetric matrix is $\begin{bmatrix}a&\boldsymbol{v}^{\mathrm{T}}\\\boldsymbol{v}&S\end{bmatrix}$. Consider the elimination $\begin{bmatrix}1&\mathbf{0}^{\mathrm{T}}\\-\frac{1}{a}\boldsymbol{v}&I\end{bmatrix}\begin{bmatrix}a&\boldsymbol{v}^{\mathrm{T}}\\\boldsymbol{v}&S\end{bmatrix}=\begin{bmatrix}a&\boldsymbol{v}^{\mathrm{T}}\\\mathbf{0}&S-\frac{1}{a}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\end{bmatrix}$, we see that the lower right block must still be symmetric.

So we do not need to compute all entries during our elimination. Starting with $\begin{bmatrix}1&2&3&4\\2&5&6&7\\3&6&8&9\\4&7&9&10\end{bmatrix}$, to do the first step of the elimination, we only need to calculate the entries below the diagonal and get $\begin{bmatrix}1&2&3&4\\0&1&&\\0&0&-1&\\0&-1&-3&-6\end{bmatrix}$, and we can then fill in the blank according to the symmetricity without a thought. So for big matrices, we save about half the time. ☺

3.3.5 When do we have an LU Decomposition

Definition 3.3.27. *For a matrix A, its upper left $k\times k$ entries form its k-**th leading principal submatrix**.*

Theorem 3.3.28. *An invertible matrix A has an LU decomposition if and only if all its leading principal submatrices are invertible. (I.e., upper left square block of various sizes are all invertible.)*

Example 3.3.29. If the upper left entry is zero, then Gaussian elimination needs a swap right away. So of course there is no LU decomposition.

Consider $\begin{bmatrix}1&1&3\\2&2&3\\4&5&6\end{bmatrix}$. We now use top row to eliminate below, and we have $\begin{bmatrix}1&1&3\\0&0&-3\\0&1&-6\end{bmatrix}$. And now we need a swap, so no LU decomposition. This is precisely because the $\begin{bmatrix}2&2\end{bmatrix}$ on the second row is a multiple of $\begin{bmatrix}1&1\end{bmatrix}$ on the first row, i.e., the second leading principal submatrix is NOT invertible. (Dependent rows = NOT surjective.) ☺

Proof of Sufficiency. If $A=LU$ and A invertible, it follows that L,U are both invertible. So the diagonal entries of L,U are all non-zero.

Now consider the block form $L=\begin{bmatrix}L_{11}&O\\L_{21}&L_{22}\end{bmatrix}, U=\begin{bmatrix}U_{11}&U_{12}\\O&U_{22}\end{bmatrix}$, here $L_{11},L_{22},U_{11},U_{22}$ are all triangular with non-zero diagonal entries, and hence all invertible.

Then we have $A=LU=\begin{bmatrix}L_{11}&O\\L_{21}&L_{22}\end{bmatrix}\begin{bmatrix}U_{11}&U_{12}\\O&U_{22}\end{bmatrix}$. Ignore the rest, just focus on the upper left block of this multiplication and we see that the k-th leading principal submatrix is exactly $L_{11}U_{11}$. Hence it must be invertible. □

Proof of Necessity. Say A is $n\times n$. We do this by induction on n. The case $n=1$ is trivial because A would be a non-zero number.

Let us do the inductive step. We write block form $A=\begin{bmatrix}A_{n-1}&\boldsymbol{v}\\\boldsymbol{w}^{\mathrm{T}}&a\end{bmatrix}$. If all leading principal submatrices of A are invertible, then all leading principal submatrices of A_{n-1} are invertible. So by induction hypothesis, we have $A_{n-1}=L_{n-1}U_{n-1}$. In particular, L_{n-1}^{-1} is the elimination we would do on A_{n-1}.

Now we block eliminate A. First we have $\begin{bmatrix} I & \mathbf{0} \\ -\boldsymbol{w}^{\mathrm{T}}A_{n-1}^{-1} & 1 \end{bmatrix} A = \begin{bmatrix} A_{n-1} & \boldsymbol{v} \\ 0 & a - w^{\mathrm{T}}A_{n-1}^{-1}\boldsymbol{v} \end{bmatrix}$. Next we eliminate A_{n-1}, so we need to use L_{n-1}^{-1}. This leads to $\begin{bmatrix} L_{n-1}^{-1} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\boldsymbol{w}^{\mathrm{T}}A_{n-1}^{-1} & 1 \end{bmatrix} A = \begin{bmatrix} U_{n-1} & L_{n-1}^{-1}\boldsymbol{v} \\ 0 & a - w^{\mathrm{T}}A_{n-1}^{-1}\boldsymbol{v} \end{bmatrix}$. Now we end up with an upper triangular matrix. So reorganizing the terms, A has the following LU decomposition:

$$A = \begin{bmatrix} L_{n-1} & \mathbf{0} \\ \boldsymbol{w}^{\mathrm{T}}A_{n-1}^{-1}L_{n-1} & 1 \end{bmatrix} \begin{bmatrix} U_{n-1} & L_{n-1}^{-1}\boldsymbol{v} \\ 0 & a - w^{\mathrm{T}}A_{n-1}^{-1}\boldsymbol{v} \end{bmatrix}$$

□

3.3.6 Permutations and PLU Decomposition

Now, what if a matrix has no LU decomposition? Well, you swap the rows around, and then do the elimination.

Definition 3.3.30. *A matrix is a **permutation matrix** if as a row operations it will simply permute the rows. (As a linear map, it will simply permute the coordinates of the input vector.)*

Consider that the appearance of a matrix must exactly be what it will do to I as a row operation, we immediately have this corollary.

Corollary 3.3.31. *A matrix is a permutation matrix if and only if it has rows of I in any order, if and only if it has the column of I in any order, if and only if it has exactly a single non-zero entry of 1 in each row and in each column.*

Example 3.3.32. Here are all possible 3 by 3 permutation matrices.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The corresponds to "change nothing", $r_1 \leftrightarrow r_2$, $r_1 \leftrightarrow r_3$, $r_2 \leftrightarrow r_3$, cyclically $r_3 \to r_2 \to r_1 \to r_3$, and cyclically $r_1 \to r_2 \to r_3 \to r_1$. ☺

Here is a trivial proposition, cosider the meaning of permutation matrices as row operations.

Proposition 3.3.33. *1. There are $n!$ permutation matrices of size $n \times n$.*

2. Multiplications of permutation matrices are permutation matrices.

3. Inverse of a permutation matrix is the transpose of that permutation matrix.

4. Every permutation matrix is a product of swapping matrices.

Proof. First two are trivial.

Let us prove that the inverse of a permutation matrix is the transpose. Suppose the (i, j) entry of a permutation matrix P is 1. What does this mean? From the fact that $PI = P$, this means that the j-th row of the identity matrix is changed into the i-th row. In particular,

$$\begin{aligned} &(i, j) \text{ entry of } P^{\mathrm{T}} \text{ is } 1 \\ \Leftrightarrow &(j, i) \text{ entry of } P \text{ is } 1 \\ \Leftrightarrow &P \text{ as a row operation send the } j\text{-th row to the } i\text{-th row} \\ \Leftrightarrow &P^{-1} \text{ as a row operation send the } i\text{-th row to the } j\text{-th row} \\ \Leftrightarrow &(i, j) \text{ entry of } P^{-1} \text{ is } 1. \end{aligned}$$

Now let us prove the last one. Instead of proving it, let me give you an algorithm to do this. (The rigorous proof is just to use mathematical induction to show that this works always. I'll leave that to you.)

Suppose we want to permute $(1,2,3,4,5)$ to $(4,1,2,5,3)$. We swap to construct our desired result from left to right:

$$(1,2,3,4,5) \to (4,2,3,1,5) \to (4,1,3,2,5) \to (4,1,2,3,4) \to (4,1,2,5,3).$$

You can see how by swapping I can succeed one by one from left to right. □

Now we can make our statement about matrices without LU decomposition.

Definition 3.3.34. *A **PLU decomposition** of a matrix A means writing $A = PLU$ where P is a permutation matrix, L is a lower triangular matrix, and U is an upper triangular matrix.*

Theorem 3.3.35. *Every invertible matrix has a PLU decomposition.*

Proof. Say A is invertible $n \times n$ matrix. We again proceed by induction on n. The case $n = 1$ is again trivial. Let us look at the inductive step now.

If A is invertible, its first column cannot all be zero. So we can find a swap P_1 such that $P_1 A$ has nonzero $(1,1)$ entry. Say $A = \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \boldsymbol{w} & A_{n-1} \end{bmatrix}$ where $a \neq 0$.

Next elimination gives $P_1 A = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}} \end{bmatrix}$.

Since $a \neq 0$ and A invertible, we see that $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}$ must be invertible and $(n-1) \times (n-1)$. So it has a PLU decomposition, say $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}} = P_{n-1}L_{n-1}U_{n-1}$.

So to go on elimination, we need to permute the bottom $n-1$ rows of $\begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}} \end{bmatrix}$, and we now have

$$\begin{aligned} P_1 A &= \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix} \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & L_{n-1}U_{n-1} \end{bmatrix} \\ P_1 A &= \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & L_{n-1} \end{bmatrix} \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & U_{n-1} \end{bmatrix} \\ A &= P_1^{-1} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & L_{n-1} \end{bmatrix} \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & U_{n-1} \end{bmatrix}. \end{aligned}$$

Now we are almost done, except that $\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix}$ is in the wrong order, and they do not commute! What should we do? We multiply and decompose.

$$\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}\boldsymbol{w} & P_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}P_{n-1}^{-1}\boldsymbol{w} & I \end{bmatrix}$$

So we have our desired PLU decomposition

$$A = (P_1^{-1} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P_{n-1} \end{bmatrix})(\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \frac{1}{a}P_{n-1}^{-1}\boldsymbol{w} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & L_{n-1} \end{bmatrix}) \begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & U_{n-1} \end{bmatrix}.$$

Here the parenthesis helps you see the P,L,U portion of the decomposition. □

Don't memorize these proofs. The point is to show you how block multiplication and block elimination works, so focus on that. If you look at all the proofs, for the criteria of LU decomposition and for the existence of PLU decomposition, you see that the key is just to proceed with the elimination, and write the next step in block form.

Part II

Abstract Structures

Chapter 4

Abstract Vector Space

4.1 Motivation

We have mastered $\mathbb{R}^n$. Now we move on to abstract vector spaces. As you shall soon discover, abstract vector spaces are "pretty much" the same as $\mathbb{R}^n$, and upon picking a basis, they behave exactly like $\mathbb{R}^n$. In some sense, there is nothing to learn. However, they represent a shift in perspective which is more geometric, more spatial, and more fundamental.

In short, an (abstract) vector space refers to any set in which we can do linear combinations. We call elements of vector spaces as vectors. Also, maps between vector spaces that respect the linear structures are linear maps.

There are several reasons why we cannot do everything in $\mathbb{R}^n$.

Example 4.1.1 (Infinite dimensional vector spaces)**.** At the start of this class, we have an example where we add or substract sound waves. This example is pretty much about the same kind of phenomena: sometimes we can do linear combinations of certain objects, yet these objects cannot be written in coordinates.

Let V represent the space of all infinitely differentiable real functions. (I.e., functions $f : \mathbb{R} \to \mathbb{R}$ such that $f^{(k)}(x)$ exists everywhere for all k.) Functions such as $e^x, \sin x, \cos x$ and polynomials are all elements of V.

Now, if several functions are infinitely differentiable, then their linear combinations are infinitely differentiable. (More specifically, via the formula $(af + bg)'(x) = af'(x) + bg'(x)$.) In particular, we see that elements of V can linearly combine into some other elements of V. V is a place where you can do linear algebra.

We can also study linear maps. Let $D : V \to V$ be the map sending each f to f'. Then you can also immediately see that this is a linear map. (Again via the formula $(af + bg)'(x) = af'(x) + bg'(x)$.) One might also study the map $M : V \to V$ that sends $f(x)$ to $xf(x)$.

Here is a funny formula, which is secretly related to Heisenberg's uncertainty principle in physics. We have $DM - MD = I$, where $I : V \to V$ is the identity map sending f to itself. To see this, note that $[(DM - MD)f](x) = (DMf)(x) - (MDf)(x) = (xf(x))' - xf'(x) = f(x)$. This is very interesting though. We know that over $\mathbb{R}^n$, we could never have linear maps $A, B : \mathbb{R}^n \to \mathbb{R}^n$ such that $AB - BA = I$. (We would do this in the homework.)

So our space V is unlike $\mathbb{R}^n$ for any n. It is in fact an infinite dimensional space, and there is no "standard basis" in the conventional sense. Elements of V can be linearly combined, yet they do not have "coordinates". Linear maps can be computed, yet they do not have "matrices". ☺

Luckily for us, infinite dimensional spaces are NOT the main concern of our class. Almost everything in our class should be finite dimensional.

Example 4.1.2 (Subspaces)**.** A subspace is, loosely speaking, a vector space contained within another larger vector space.

Consider the plane $\mathbb{R}^2$. Let $V = \{\begin{bmatrix} x \\ x \end{bmatrix} \in \mathbb{R}^2 : x \in \mathbb{R}\}$, the subset of all vectors whose coordinates are the same. This is a straight line. Let $W = \{\begin{bmatrix} x \\ -x \end{bmatrix} \in \mathbb{R}^2 : x \in \mathbb{R}\}$, the subset of all vectors whose coordinates add up to zero. This is also a straight line.

You can quickly verify that any linear combinations of vectors in V should stay in V. Such things are called "subspaces" of $\mathbb{R}^2$. Similarly, W is also a subspace, as linear combinations of vectors in W should stay in W.

Sometimes, we do not care about the whole space $\mathbb{R}^2$. Rather, I just want to study a map sending elements of V to elements of W. Imagine that V is an elastic band, and I would like to "relocate" this elastic band into the position of W, and then stretch it by a scale of three.

In effect, we have a map $f : V \to W$ that sends $\begin{bmatrix} x \\ x \end{bmatrix}$ to $\begin{bmatrix} 3x \\ -3x \end{bmatrix}$. You can quickly verify that

$$f(a\begin{bmatrix} x \\ x \end{bmatrix} + b\begin{bmatrix} y \\ y \end{bmatrix}) = \begin{bmatrix} 3ax + 3by \\ -3ax - 3by \end{bmatrix} = af(\begin{bmatrix} x \\ x \end{bmatrix}) + bf(\begin{bmatrix} y \\ y \end{bmatrix}).$$

So our map f is linear.

However, note that V, W are both lines (one-dimensional). So, should f be some 1×1 matrix? But that would be absurd. It sends a vector with two coordinates to a vector with two coordinates. Maybe it is a 2×2 matrix? But then its domain and codomain are both one-dimensional. Hmm.

Now, you may say "screw this, I'm just going to use a 2×2 matrix to represent f." Fine. Note that the matrix $\begin{bmatrix} 3 & 0 \\ 0 & -3 \end{bmatrix}$ would indeed send every $\begin{bmatrix} x \\ x \end{bmatrix}$ to $\begin{bmatrix} 3x \\ -3x \end{bmatrix}$. Aha! Is this the matrix for f?

However, also note that the matrix $\begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix}$ would indeed send every $\begin{bmatrix} x \\ x \end{bmatrix}$ to $\begin{bmatrix} 3x \\ -3x \end{bmatrix}$. And so does the matrix $\begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix}$, which is not even invertible, and so does infinitely many other matrices. And all these matrices have very different properties and behaviors. Yet they somehow ALL represent the linear map f? Impossible!

The trouble is again related to the standard basis. When we write elements of V, W in coordinates, we are expressing them as linear combinations of the standard basis, i.e., $\begin{bmatrix} a \\ b \end{bmatrix}$ really stands for $a\boldsymbol{e}_1 + b\boldsymbol{e}_2$. However, the standard basis vectors are OUTSIDE of V or W.

As you can imagine, it might NOT be a good idea to express elements inside of V as linear combinations of vectors outside of V. By using outside vectors to express inside vectors, V, W are one-dimensional spaces, yet their elements have more than one coordinates. You have redundant information flying around. And everything is messed up accordingly in the ensuing calculations.

In short, you CANNOT rely on the regular "coordinate" argument anymore, and you should abandon the standard basis. In other words, you would have to think of V, W as abstract vector spaces.

Maybe we can just name vectors in V, W according to their distance to the origin. Then $\begin{bmatrix} a \end{bmatrix} \in V$ now represents a vector in V with distance a to the origin (upper right being positive), and $\begin{bmatrix} b \end{bmatrix} \in W$ now represents a vector in W with distance b to the origin (lower right being positive). In this way, we can write a matrix $L = \begin{bmatrix} 3 \end{bmatrix}$, and it just multiply this distance by 3. ☺

Example 4.1.3 (Change of basis)**.** Let us prove that the three medians of a triangle are concurrent, i.e., sharing the same intersection point. I'm sure you have a bazzilion ways to do it in highschool. However, let us try to do it with brute force computation!

Now, generically, our triangle ABC lies in $\mathbb{R}^2$, but the three vertices could be anywhere! So the coordinate should be like $\begin{bmatrix} a_x \\ a_y \end{bmatrix}$, $\begin{bmatrix} b_x \\ b_y \end{bmatrix}$ and $\begin{bmatrix} c_x \\ c_y \end{bmatrix}$, where all six numbers are unknown. Now, we might try to just go ahead and compute with this. We may find the three midpoints, and then find the line equations of the three

median, and then intersect them to see what happens. Or, we might as well kill ourselves, since this would surely be one ugly computation process. Let us give up for now.

So what can we do? Well, a better way is to "forget" about your old coordinate system. Just erase your x-axis and y-axis and such. And we shall simply re-define our coordinate system. How? Well, a coordinate system requires three things: You need to pick an origin, then you need to pick the directions for your x-axis and y-axis, i.e., you need to pick the vectors $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$.

Let us now define the origin to be A. Nice! Now A has coordinates $\begin{bmatrix}0\\0\end{bmatrix}$. Then let us define $\begin{bmatrix}1\\0\end{bmatrix}$ to be the vector AB and $\begin{bmatrix}0\\1\end{bmatrix}$ as the vector AC. Then now $B = \begin{bmatrix}1\\0\end{bmatrix}$ and $C = \begin{bmatrix}0\\1\end{bmatrix}$.

Huh, now the calculations are super easy. The three mid points of the three sides of the triangle are $D = \begin{bmatrix}1/2\\1/2\end{bmatrix}, E = \begin{bmatrix}0\\1/2\end{bmatrix}, F = \begin{bmatrix}1/2\\0\end{bmatrix}$. The three medians are $AD : x = y$, $BE : x+2y = 1$, and $CF : 2x+y = 1$. You can check that all three lines went through the point $\begin{bmatrix}1/3\\1/3\end{bmatrix}$. End of proof.

In other words, if the original coordinates (tied to the standard basis) suck, then we would like to change coordinate (by using a different set of vectors as "basis"). As a result of that, $\mathbb{R}^n$ is no longer the same $\mathbb{R}^n$, and every vector before and after the basis-change would have different coordinates. Nevertheless, the underlying ABSTRACT space and abstract vectors are the same space. The space did not change at all, and the vectors did not change at all. We simply changed the names (coordinates).

I hope this example illustrates the difference between ABSTRACT entities (independent of coordinates), and the NOMINAL entities (dependent of coordinates). ☺

So what is abstraction? My favorite explanation is one by Aristotle. He thinks that abstraction is about "forgetting intentionally". For example, say I am looking at a football, and I'm trying to reach the abstract concept of "sphere". What should I do? According to Aristotle, I should intentially "forget" its color, "forget" its smell, "forget" its texture and its location, and so on until the only thing that I still remember is its shape, the sphere.

(You see, I have a terrible memory, and I forget things all the time. That is probably why I am good at mathematics.)

But how can forgetting things help us? In the example above, we intentionally FORGET the origin and the coordinate axes, so that we can re-pick our own. Given $\mathbb{R}^n$, if you forget where the origin is, and you also forget where the standard basis vectors are, then you would reach some abstract entity.

In our case, according to the level of abstraction, we have the following concepts:

1. $\mathbb{R}^n$ is, well, $\mathbb{R}^n$. We know it already. It comes with a GIVEN origin, and a standard basis $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$. Now, the location of the origin and the direction of these standard basis elemtents might be inconvenient, but there's nothing much you can do about it, because at the level of $\mathbb{R}^n$, coordinates must be fixed, and the standard basis vectors must be fixed.

2. Now, we can choose to "forget" where the basis is. Then we arrive at an abstract vector space V, i.e., it is like $\mathbb{R}^n$, but we forget where the standard basis vectors are. Then as a result, you CANNOT write things in coordinates anymore, since you no longer remember which vectors are the standard basis vectors. Given an element of V, you may write $\boldsymbol{v} \in V$, but you do not have a coordinate expression. Nevertheless, given $\boldsymbol{v}, \boldsymbol{w}$, you can do linear combinations of them just fine. You would still have $2\boldsymbol{v} + \boldsymbol{v} = 3\boldsymbol{v}$, and you don't need coordinates to do that.

3. (Optional) Now starting from abstract vector spaces, you can further "forget" where the origin is. Oops. Then we arrive at an ***affine space***. It is just like $\mathbb{R}^n$, but not only you forget to remember where the basis vectors are, you also forget where the origin is. In particular, you CANNOT add vectors or scale vectors now. Why? For example, scaling $\boldsymbol{v}$ means the arrow from the origin to $\boldsymbol{v}$ is tripled in length, but we don't know where the origin is! So what can we do? Well, for example, we

can find "mid-points". Given two vector $\boldsymbol{v}, \boldsymbol{w}$, you can find their midpoint by doing $\frac{1}{2}\boldsymbol{v} + \frac{1}{2}\boldsymbol{w}$, and geometrically you can see that this process does NOT require the origin. It is essentially still the same set as $\mathbb{R}^n$ with the same algebraic structure, but you just forget where the origin is. (Our class does NOT study affine spaces. You can check out the subject of affine geometry if you like.)

So, why abstract? Why forget? The short answer is that we forget in order to remember, or re-choose. The more you forget, the more freedom you get to enjoy.

Hopefully now you have a vague sense of what abstract vector spaces are. On the plane, we can fix a vector $\boldsymbol{v}$. Under some basis, it will have a "name" as maybe $\begin{bmatrix}1\\2\end{bmatrix}$. Under a different basis, it will have a different "name" as maybe $\begin{bmatrix}3\\4\end{bmatrix}$. But despite the different names, the underlying vector is the same $\boldsymbol{v}$. We want to focus on the essential vector rather than to focus on the superficial name.

4.2 Axioms of Abstract Vector Spaces

Informally, we want to define abstract vector space as a set, in which you can do linear combinations. Now let us ponder the exact meaning of this, and hopefully we shall arrive at a collection of axioms that give us a good definition. You can safely skip most of this subsection and jump right to the end to see the definition. Meanwhile, if you are ever curious why we use these definitions, then here is a relatively detailed discourse of them.

Keep in mind of this key idea: All the laws serves one purpose: we want to simplify ALL expressions into linear combinations. Look at the following expressions. What calculation law would you want to have, in order to simplify them into linear combinations of the form $a\boldsymbol{v} + b\boldsymbol{w}$?

1. $(\boldsymbol{v} + \boldsymbol{w}) + \boldsymbol{v}$.
2. $2(\boldsymbol{v} + \boldsymbol{w})$.
3. $2(3\boldsymbol{v})$.
4. $2\boldsymbol{v} + 3\boldsymbol{v}$.
5. $\boldsymbol{v} + (-1)\boldsymbol{v}$.
6. ....

Example 4.2.1 (Associativity and Commutativity)**.**

Say I have $\boldsymbol{w} + \boldsymbol{v} + \boldsymbol{w}$. The obvious thing to do is to simplify it into $\boldsymbol{v} + 2\boldsymbol{w}$. But what is actually involved? Specifically, we need to do the following:

$$(\boldsymbol{w} + \boldsymbol{v}) + \boldsymbol{w} = (\boldsymbol{v} + \boldsymbol{w}) + \boldsymbol{w} = \boldsymbol{v} + (\boldsymbol{w} + \boldsymbol{w}) = \boldsymbol{v} + 2\boldsymbol{w}.$$

Here we mainly talk about the first two steps. The first is called commutativity, and the second is called associativity. It is trivially obvious that we want these properties to be part of the definition of a vector space.

But if you are curious, let us discuss them further. For new students of mathematics, a common misconception is that maybe commutativity is stronger than associativity. That is WRONG. The two are independent.

Pause and think about this. Given $(a+b)+c$, if we only have associativity, can you write out all possible equivalent re-formulations for this formula? How many are there? What if we only have commutativity?

Eitherway, without associativity or commutativity of addition, $(\boldsymbol{w} + \boldsymbol{v}) + \boldsymbol{w}$ will NEVER be able to be simplified into $\boldsymbol{v} + 2\boldsymbol{w}$. Then the very concept of linear combination will no longer work the way we want. ☺

Remark 4.2.2 ((Optional) Operation Order and Operand Order)**.**

Associativity states that $(a+b)+c=a+(b+c)$, so we may change the order of the operations. But Commutativity states that $(a+b)+c=(b+a)+c$. Here the order of operations are NOT changed, because we are always adding a and b first, then we add the result with c. What is allowed by commutativity is to change the order of OPERANDS (or summands), i.e., the objects of the operation (or the sum).

In short, if we want to change the calculation order of the "+" symbols, we need associativity. If we need to change the order of a, b, c, then we need commutativity. One deals with operations while fixing the operands, the other deals with operands while fixing the operations, so their actions are completely independent of each other.

Case in point, for our example $(\boldsymbol{w}+\boldsymbol{v})+\boldsymbol{w}$, without commutativity, association will only ever change the order in which we compute the addition symbols, but the order of summands will always remain $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{w}$. So the two w cannot be moved together to be simplified. On the other hand, without associativity, commutativity will only ever change the order of $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{w}$, but the order of computation is unchanged. So we will ALWAYS compute the sum of $\boldsymbol{v}$ and $\boldsymbol{w}$ first, and never able to compute $\boldsymbol{w}+\boldsymbol{w}$ first as we would have liked.

Of course, these discussion is only relevant if you aspire to learn more algebra after this class. For the purpose of this class, we want vector addition to be both. So no order matters any way.

Example 4.2.3 (Zero Vector)**.**

Say I have $0\boldsymbol{v}+0\boldsymbol{w}$. Surely I would like to simplify that to $\mathbf{0}$, the zero vector, yes?

Afterall, the zero vector is special. All linear maps must send the zero vector to the zero vector, and all vectors are parallel to the zero vector, and so on. You just have to have a zero vector.

Now, what does a zero vector do? The defining trait of the zero vector is the fact that $\mathbf{0}+\boldsymbol{v}=\boldsymbol{v}$ for all other $\boldsymbol{v}$. Note that this immediately implies that the zero vector is unique. If $\mathbf{0}, \mathbf{0}'$ are two zero vectors, then $\mathbf{0}=\mathbf{0}+\mathbf{0}'=\mathbf{0}'$. ☺

Example 4.2.4 (Negation of vectors)**.** Our study is called "linear algebra", and "linear" is the adjective form of "line". A line extends both ways to infinity. In particular, given any direction, there must be an opposite direction.

In short, for any vector $\boldsymbol{v}$, there must be a vector $\boldsymbol{w}$ such that $\boldsymbol{v}+\boldsymbol{w}=\mathbf{0}$. We call this the negation of $\boldsymbol{v}$, and denote it as $-\boldsymbol{v}$. The concept of negations immediately allow us to SUBSTRACT vectors. We define $\boldsymbol{x}-\boldsymbol{y}$ to be $\boldsymbol{x}+(-\boldsymbol{y})$.

Now one can immediately see that given any vector $\boldsymbol{v}$, its negation is unique. If $\boldsymbol{w}, \boldsymbol{w}'$ are both its negations, then $\boldsymbol{w}=\boldsymbol{w}+\boldsymbol{v}+\boldsymbol{w}'=\boldsymbol{w}'$.

The existence of negations of vectors also indicates that we have the law of cancellations when it comes to vector additions, i.e., if $\boldsymbol{v}+\boldsymbol{w}=\boldsymbol{u}+\boldsymbol{w}$, then $\boldsymbol{v}=\boldsymbol{u}$. ☺

All the axioms so far only concerns with vector addition. Now let us move to scalar multiplications

Example 4.2.5 (Scalar Multiplications are linear on scalars)**.** All calculational laws serves to simplify expressions into linear combinations. Say we see $2\boldsymbol{v}+3\boldsymbol{v}$? Surely we want to simplify that into $5\boldsymbol{v}$. This requires a law of distribution on scalars, i.e., $(a+b)\boldsymbol{v}=a\boldsymbol{v}+b\boldsymbol{v}$.

Finally, what about $2(3\boldsymbol{v})$? Surely we want to simplify that into $6\boldsymbol{v}$. We want a weird law of "associativity for scalar multiplications" that goes like $(ab)\boldsymbol{v}=a(b\boldsymbol{v})$. However, looking at this from another angle, and it would not be weird anymore: let $m_{\boldsymbol{v}}$ be the map that sends a real number x to the vector $x\boldsymbol{v}$, then $m_{\boldsymbol{v}}(ax+by)=(ax+by)\boldsymbol{v}=(ax)\boldsymbol{v}+(by)\boldsymbol{v}=a(x\boldsymbol{v})+b(y\boldsymbol{v})=am_{\boldsymbol{v}}(x)+bm_{\boldsymbol{v}}(y)$, i.e., $m_{\boldsymbol{v}}$ is linear.

So don't think of these as law of distributions or some weird law of associativity. The best way to think about this is that, for fixed $\boldsymbol{v}$, the output $k\boldsymbol{v}$ is linear on the input k. ☺

Example 4.2.6 (Scalar Multiplications are linear on vectors)**.** On the other hand, what if we see $2(\boldsymbol{v}+\boldsymbol{w})$? Surely we want this to be $2\boldsymbol{v}+2\boldsymbol{w}$, right? So we need a law of distribution on vectors, i.e., $k(\boldsymbol{v}+\boldsymbol{w})=k\boldsymbol{v}+k\boldsymbol{w}$.

In particular, let m_k be the map that sends vectors $\boldsymbol{v}$ to $k\boldsymbol{v}$. Then $m_k(a\boldsymbol{v}+b\boldsymbol{w})=k(a\boldsymbol{v}+b\boldsymbol{w})=k(a\boldsymbol{v})+k(b\boldsymbol{w})=(ak)\boldsymbol{v}+(bk)\boldsymbol{w}=a(k\boldsymbol{v})+b(k\boldsymbol{w})=am_k(\boldsymbol{v})+bm_k(\boldsymbol{w})$. So m_k is linear.

Again, I think the best way to think about this is that, for fixed k, the output $k\boldsymbol{v}$ is linear on the input $\boldsymbol{v}$. ☺

We see that scalar multiplication is linear on each of the two inputs. This sort of thing is called ***bilinear***.

Now there are some interesting consequences for the bilinearity of scalar multiplications.

Example 4.2.7 (One is respected)**.** Consider $1(k\boldsymbol{v}) = (1 \times k)\boldsymbol{v} = k\boldsymbol{v}$. It is reasonably obvious that when we multiply a vector by 1, we do not want to change that vector. However, what if you have $1\boldsymbol{v}$ and no k to be a mediator? Then we do not know what might happen here.

Note that this has to be a separate axiom. Nothing else implies this. If one were to define, say, $1\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}x\\0\end{bmatrix}$, then check carefully and see that all previous axioms are NOT violated. ☺

Example 4.2.8 (Zero is respected)**.** We have $0\boldsymbol{v} + \boldsymbol{v} = (0+1)\boldsymbol{v} = \boldsymbol{v}$. Substract $\boldsymbol{v}$ from both sides, we see that $0\boldsymbol{v} = \boldsymbol{0}$ for all vectors.

In fact, since $k \mapsto k\boldsymbol{v}$ is linear, 0 has to goes to the zero vector. ☺

Example 4.2.9 (Minus one is respected)**.** Note that $\boldsymbol{v}+(-1)\boldsymbol{v} = (1-1)\boldsymbol{v} = \boldsymbol{0}$. So we see that $(-1)\boldsymbol{v} = -\boldsymbol{v}$ is the negation indeed. ☺

Let us now define an abstract vector space (over $\mathbb{R}$).

Definition 4.2.10. *An abstract vector space is a set V together with two operations, a vector addition $+ : V \times V \to V$ and a scalar multiplication $m : \mathbb{R} \times V \to V$, such that the following is true:*

Vector addition gives an abelian group:

1. *(Associative and Commutative) Vector addition is associative and commutative.*
2. *(Zero Vector) There is a zero vector $\boldsymbol{0} \in V$ such that $\boldsymbol{0} + \boldsymbol{v} = \boldsymbol{v}$ for all $\boldsymbol{v} \in V$.*
3. *(Negation Vector) For any vector $\boldsymbol{v} \in V$, there is a unique negation $-\boldsymbol{v} \in V$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}$. (This also gives the law of cancellation for vector additions.)*

Scalar multiplication is bilinear:

1. *(Linear in scalars) $(ab)\boldsymbol{v} = a(b\boldsymbol{v})$ and $(a+b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$ for all $a, b \in \mathbb{R}, \boldsymbol{v} \in V$.*
2. *(Linear in vectors) $a(\boldsymbol{v} + \boldsymbol{w}) = a\boldsymbol{v} + a\boldsymbol{w}$ for all $a \in \mathbb{R}, \boldsymbol{v}, \boldsymbol{w} \in V$.*
3. *(Respect identity) $1\boldsymbol{v} = \boldsymbol{v}$.*

Remark 4.2.11. *Uniqueness of the zero vector and the negation vector are consequences of their definition and associativity. The negation axiom also allows us to define vector substractions via $\boldsymbol{v} - \boldsymbol{w} = \boldsymbol{v} + (-\boldsymbol{w})$.*

The (linear in scalars) axiom shows that $1\boldsymbol{v} + 0\boldsymbol{v} = (1+0)\boldsymbol{v} = 1\boldsymbol{v}$. Now use the (Negation vector) axiom to cancel the $1\boldsymbol{v}$ on both sides, we see that $0\boldsymbol{v} = \boldsymbol{0}$.

Now use the (respect identity) axiom, we have $(-1)\boldsymbol{v} + \boldsymbol{v} = (-1)\boldsymbol{v} + 1\boldsymbol{v} = (-1+1)\boldsymbol{v} = 0\boldsymbol{v} = \boldsymbol{0}$. Hence $(-1)\boldsymbol{v}$ is the negation of $\boldsymbol{v}$.

Theorem 4.2.12 (Optional)**.** *In a vector space V, if we have vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ and scalars $a_1, \ldots, a_t$ such that they are all combined into a single vector using vector addition and scalar multiplication, then we can simplify the expression into $b_1\boldsymbol{v}_1 + \cdots + b_k\boldsymbol{v}_k$ for some scalars $b_1, \ldots, b_k$.*

Proof. Say the starting formula is $f(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, a_1, \ldots, a_t)$ which involves s operations (i.e., scalar addition and vector multiplication). We do this by mathematical induction on the number of operations s. If the number of operation involved is zero, then we have a single vector $\boldsymbol{v}_1$ without any operation, and we have $\boldsymbol{v}_1 = 1\boldsymbol{v}_1$ by the last axiom of vector space, so we are done with $b_1 = 1$.

Suppose the statement is true when the number of operations is less than s. Now suppose we have s operations in total.

If the last operation is a scalar multiplication, WLOG say scalar multiplicatino by a_t, then the expression must be $f(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) = a_t g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t)$. Now g involves $s-1$ operations, and hence by induction hypothesis $g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) = b_1'\boldsymbol{v}_1 + \cdots + b_k'\boldsymbol{v}_k$ for some scalars $b_1', \dots, b_k'$. Therefore

$$\begin{aligned} f(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) =& a_t g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) \\ =& a_t(b_1'\boldsymbol{v}_1 + \cdots + b_k'\boldsymbol{v}_k) \\ =& a_t(b_1'\boldsymbol{v}_1) + \cdots + a_t(b_k'\boldsymbol{v}_k) \\ =& (a_t b_1')\boldsymbol{v}_1 + \cdots + (a_t b_k')\boldsymbol{v}_k. \end{aligned}$$

Note that we must involve the fact that scalar multiplication is linear on the vector, and the "weird association law" that $a(b\boldsymbol{v}) = (ab)\boldsymbol{v}$.

If the last operation is a vector addition, then we must have $f(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) = g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) + h(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t)$. Now g, h must both have less than s operations, therefore by induction hypothesis $g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) = c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k$ for some scalars $c_1, \dots, c_k$, and $b(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) = c_1'\boldsymbol{v}_1 + \cdots + c_k'\boldsymbol{v}_k$ for some scalars $c_1', \dots, c_k'$. So we have

$$\begin{aligned} f(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) =& g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) + h(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, a_1, \dots, a_t) \\ =& (c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k) + (c_1'\boldsymbol{v}_1 + \cdots + c_k'\boldsymbol{v}_k) \\ =& (c_1 + c_1')\boldsymbol{v}_1 + \cdots + (c_k + c_k')\boldsymbol{v}_k. \end{aligned}$$

Note that we must use the law of association, the law of commutativity, and the fact that scalar multiplication is linear on the scalar.

So we are done. □

Remark 4.2.13. *In the proof above, we have used ALL axioms except for the zero vector axiom and the negation vector axiom.*

The zero vector axiom is in fact the case of the theorem above when $k = 0$. If there are no vectors, the linear combination of no vector must be $\mathbf{0}$.

The negation axiom is in fact redundant. Given $\boldsymbol{v}$, its negation must be $(-1)\boldsymbol{v}$, which you can verify from other axioms.

You don't have to memorize these axioms. Rather, just remember the purpose of them. ALL of them serves the same purpose: whatever you do with vector additions and scalar multiplications, in the end it is just a linear combination. To have a nice concept of "linear combination" is the purpose of all of them.

Now this immediately prompts ups to define linear maps, which are maps that perserve vector space structures.

Definition 4.2.14. *Given two vector space V, W, a map $f : V \to W$ is linear if $f(\boldsymbol{v} + \boldsymbol{w}) = f(\boldsymbol{v}) + f(\boldsymbol{w})$ and $f(k\boldsymbol{v}) = kf(\boldsymbol{v})$. I.e., f preserves linear combinations.*

4.3 (Optional) Axioms and mathematical structures

Defining mathematical structures via a bunch of axioms is very important to mathematicians. For non-mathematicians, they are less useful. But for mathematicians, this is the best way to differentiate different structures, and identify similar structures. Here are some mathematical structures for those interested.

(These are absolutely not required. You don't need to know any of these.)

Definition 4.3.1. *A **group** is a set G with an operation $G \times G \to G$ (which may be written multiplicatively or additively, however you like), such that the following axioms hold. (Here the result of an operation of g and h is written simply as gh.)*

1. *(Associativity) $(xy)z = x(yz)$ for all $x, y, z \in G$.*

2. *(Identity) There is an element $e \in G$ such that $eg = ge = g$ for all $g \in G$. (We call e the identity element.)*

3. *(Inverse) For each $g \in G$, we can find $h \in G$ such that $gh = hg = e$.*

Here associativity is the heart of all mathematical wonders. The latter two axioms guarantee that we have the law of cancellation. Finally, the uniqueness of the identity element and the uniqueness of the inverse element can be deduced from the axioms.

Example 4.3.2.

The set of integers $\mathbb{Z}$ with the operation of addition is a group.

The set of integers $\mathbb{Z}$ with the operation of multiplication is NOT a group, because many elements have no multiplicative inverse. We see here that merely specifying the set is NOT enough. From now on, we shall write things like $(\mathbb{Z}, +)$ and $(\mathbb{Z}, \times)$ to specify which operations are in use. So $(\mathbb{Z}, +)$ is a group and $(\mathbb{Z}, \times)$ is not a group.

The set of natural numbers $(\|, \|+)$ is NOT a group, because many elements have no additive inverse.

The rational numbers $\mathbb{Q}$, real numbers $\mathbb{R}$, complex numbers $\mathbb{C}$ are all groups (under addition).

Throwing away zero, then $(\mathbb{Q} - \{0\}, \times)$ is in fact a group. The same is true for $(\mathbb{R} - \{0\}, \times)$ and $(\mathbb{C} - \{0\}, \times)$.

Any vector space V with vector addition form a group.

All invertible diagonal $n \times n$ matrices with matrix multiplication form a group. ☺

Note that for all the groups above, we have an extra law, the law of commutativity.

Definition 4.3.3. *A group is an **abelian group** if we have the law of commutativity. (Abel is the name of a famous mathematician. "Abelian" is the adjective of "Abel".)*

What about non-abelian groups? They are also super important!

Example 4.3.4.

All invertible $n \times n$ matrices with matrix multiplication form a group. If $n \neq 1$, this group is NOT abelian.

All invertible upper triangular $n \times n$ matrices with matrix multiplication form a group. If $n \neq 1$, this group is NOT abelian. You can also just look at unit upper triangular matrices, and it is still a group, and if $n \neq 1, 2$, this group is NOT abelian.

All permutations on n objects from a group. When $n \neq 1, 2$, this is NOT abelian.

In fact, for any set X, all bijective functions $f : X \to X$ form a group. (The operation is function composition.) Usually this is not abelian

We can also add extra conditions here. All continuous bijective function $f : \mathbb{R} \to \mathbb{R}$ form a group (with function composition as the group operation).

All symmetries of a triangle form a non-abelian group. There are six elements here, three reflections and three rotations (here we treat the identity symmetry as a "zero rotation'). ☺

The last example is very revealing. Ultimately, one might see the study of groups as the study of symmetries. Symmetries of geometric shapes, symmetries of the universe, symmetries among various symmetries themselves, you name it!

Now, groups are super important, and the axioms are very few. However, this makes them HARD to study. With only a few axioms, there are many possible ways a group might look like. The study of finite non-abelian groups is still pretty much a big mystery to us. Nevertheless, they already help us solve many difficult problems.

Two prominant kinds are the various "straight edge and compass" problems (e.g., can you trisect an arbitrary angle using a compass and an (unmarked) ruler?), and the statement "polynomials of degree 5 and above have no radical solutions." (I.e., no nice formula exists. Say if the polynomial has degree 2, we have $\frac{b \pm \sqrt{b^2 - 4ac}}{2a}$. But for polynomials of degree 5 and above, no such formula exists.) These are done by studying potential symmetries among solutions to an algebraic equation.

Remark 4.3.5. *The mathematician most famous for inventing most of group theory is a French super genius called Galois. He also solve the polynomial problem mentioned above.*

He died when he was 20 years old, duelling for a woman he loved. Yet in his short life, his contribution to math is more than most mathematicians' whole life's work. Wow, he did ALL THAT before turning 20 years old? This surely makes us feel bad, yes? Comparatively, what was I doing when I was 20? Well... I guess I was also chasing after a woman, who is now my wife and the mother of my two children. Hey, I guess I'd rather have my life than Galois' after all.

His life story is quite awesome. Check it out.

In short, groups are useful, but they are usually way too hard to study. To make them easier to study, we need to add more axioms.

Definition 4.3.6. *A **field** is a set $\mathbb{F}$ with two operations $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ and $\times : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, such that the following axioms hold.*

1. *$(\mathbb{F}, +)$ is an abelian group. We usually write the additive identity element as 0.*
2. *$(\mathbb{F} - \{0\}, \times)$ is an abelian group. We usually write the multiplicative identity element as 1.*
3. *We have the law of distribution.*

So technically, the statement of being an "abelian group" needs four axioms. So we have a total of nine axioms for a field. But with these many axioms, they are MUCH easier to study.

Example 4.3.7.

$\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are all very useful fields.

Let us define $\mathbb{F}_p$ for a prime number p as the set $\{0, \dots, p-1\}$, where addition and multiplication are done mod p. (E.g., $(p-2)+3 = 1$ and so on. We simply treat $p = 0$ always.) Then $\mathbb{F}_p$ is a field.

If p is not a prime, then the above construction will NOT be a field. For example, if we consider mod 6, then 2 will have no multiplicative inverse.

There are other fields, but the ones above are the most important ones. ☺

Now look at the definition of a vector space again.

Definition 4.3.8. *A vector space over a field $\mathbb{F}$ is a set V together with two operations, a vector addition $+ : V \times V \to V$ and a scalar multiplication $m : \mathbb{F} \times V \to V$, such that the following axioms hold. (We write the result of scalar multiplication of $k \in \mathbb{F}$ with $\boldsymbol{v} \in V$ as $k\boldsymbol{v}$.)*

1. *$(V, +)$ is an abelian group.*
2. *($\mathbb{F}$-linear structure) $(a+b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$ and $a(b\boldsymbol{v}) = (ab)\boldsymbol{v}$.*
3. *($\mathbb{F}$-linear structure respect V-group structure) $k(\boldsymbol{v} + \boldsymbol{w}) = k\boldsymbol{v} + k\boldsymbol{w}$, and $1\boldsymbol{v} = \boldsymbol{v}$.*

Remark 4.3.9. *Technically, to define a vector space, the group requirement means four axioms, and that $\mathbb{F}$ is a field means nine more axioms, and then we need four more for the vector space. So these 17 axioms can only mean one thing: vector spaces are super easy to learn, as we have so many axioms (tools) to help us!*

Specifically, pay attention to the second axiom here on $\mathbb{F}$-linear structures. Intuitively this says that each $\boldsymbol{v}$ is contained in a "line" made of various $k\boldsymbol{v}$, and structure-wise, this "line" looks exactly like $\mathbb{F}$. Addition on this "line" is the same as addition on $\mathbb{F}$, and multiplication on this "line" is the same as multiplication on $\mathbb{F}$. (So if $\mathbb{F} = \mathbb{R}$, this means every vector is contained in something that looks like $\mathbb{R}$, i.e., a line.)

So what is a vector space over $\mathbb{F}$? Conceptually, it is a space where elements are trapped inside various "lines". And the field $\mathbb{F}$ describes the shape and property of these "lines".

Our class focus mainly on the case of $\mathbb{R}$, so that our intuitions about lines will be spot on.

If we study vector spaces over $\mathbb{C}$ (which is geometrically a plane), then each "line" actually looks like a plane. Then we are no longer doing "linear algebra", but technically "planar algebra"....

If we study vector spaces over $\mathbb{F}_2$, then we would venture into computer science, since the all scalars are now 0 and 1. Discrete world still have geometry though. My personal favorite finite geometric object is the Fano plane, and there is a board game called "SET" which is essential investigating vector spaces over $\mathbb{F}_3$.

4.4 How to study a vector space

There is usually only one way to study a vector space: by finding a basis for it.

Definition 4.4.1. *We say a collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in V$ is a **basis** of V (plural form is "bases") if any vector of V is a UNIQUE linear combination of them.*

Definition 4.4.2. *We say a vector space V has dimension k if it has a basis made of k vectors.*

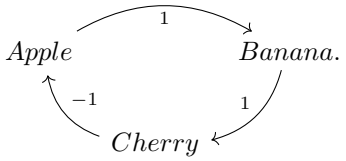
Remark 4.4.3. *In our class we require a basis to only contain finitely many elements. We say a space is finite dimensional if there is a basis. For infinite dimensional spaces, they might not have a basis at all. We may construct something called a Hammel basis, but there will be some curious situations that is very different from finite dimensional spaces.*

Remark 4.4.4. *Can a space V have two bases with different number of elements? It cannot. The proof will be given in later.*

Example 4.4.5. Our favorite example of a vector space is of course $\mathbb{R}^n$, and our favorite basis for $\mathbb{R}^n$ is the standard basis, i.e., $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$. Every vector is a unique linear combination of them.

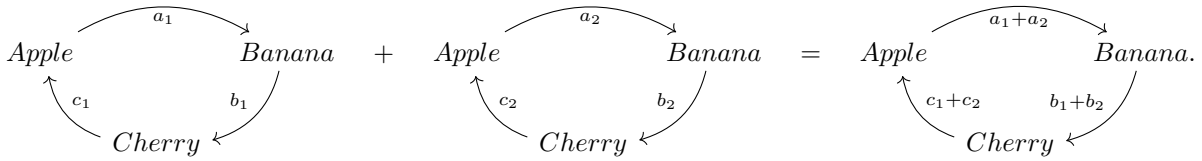
In fact, if $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$, then what are the coordinates $v_1, \ldots, v_n$? They are precisely the coefficents when we write $\boldsymbol{v}$ as a linear combination of the standard basis vectors. ☺

Example 4.4.6. Given three objects, say apple, banana and cherry, a person's preferences about them can be written as a "preference circle". For example, the following graph indicates that a person likes apple more than banana, and banana more than cherry.



Here 1 means the arrow goes from the more preferred option to the less preferred option, and -1 means the opposite. A person can have six possible preferences, see if you can write them all out. (Note that a person's preference must be contradiction-free, i.e., transitive. If I prefer A to B, and B to C, then I prefer A to C.)

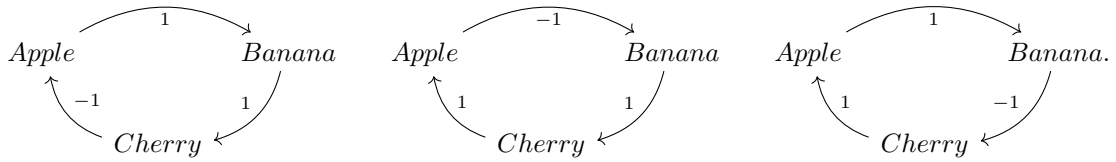
Now, given many people with their preference cycles, we can add them "coordinate-wise" like this:



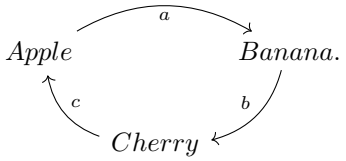
In effect, we are polling the results of peoples preference, to get the preference of the whole population. Say if we add up the preference cycles of many people, and it turns out that the arrow from apple to banana is

positive, this means more people prefer apple to banana than the oppositive. As you can see, this gives a vector space, the "population preference space". Can you find a basis made of three vectors?

A basis is in fact made of these three:



To see uniqueness, say we have an arbirary preference population cycle such as this:



We can decompose it uniquely as a linear combination of the previous basis with coefficients $\frac{a+b}{2}$, $\frac{b+c}{2}$, $\frac{c+a}{2}$.

Also, here is a fun thing to try: try to combine many preference cycles such that all three arrows are positive. So in your population, more people like apple than banana, and more people like banana than cherry, and more people like cherry than apple. This is called a voting paradox, and it can actually happen. (This shows why american elections are flawed, when you are forced to choose between two candidates rather than from all candidates. Politicians exploit this all the time.) ☺

Example 4.4.7. Basic calculus shows that the solutions to the differential equation $f''(x) = -f(x)$ must be $a\sin x + b\cos x$ for some $a, b \in \mathbb{R}$. So the solution space is the space of functions $\{a\sin x + b\cos x : a, b \in \mathbb{R}\}$, and you can see that this is a vector space. A basis is made of $\sin x, \cos x$, so this space has dimension two. ☺

The point of finding a basis is to find coordinates.

Definition 4.4.8. *Given a vector space V and a basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$, then for each vector $\boldsymbol{v} \in V$, if $\boldsymbol{v} = a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n$, then we say the **coordinates** of $\boldsymbol{v}$ under $\mathcal{B}$ is $\boldsymbol{v}_{\mathcal{B}} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$.*

Example 4.4.9. In the solution space V to the differential equation $f''(x) = -f(x)$, if we pick $\sin x$ and $\cos x$ as basis $\mathcal{B}$, then the solution $g(x) = 2\sin x + 2\cos x$ will have coordinates $g_{\mathcal{B}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.

However, you may check that $\sin(x + \frac{\pi}{4}), \cos(x + \frac{\pi}{4})$ is also a basis $\mathcal{C}$ for V. (Because $\sin(x + \frac{\pi}{4}) = \frac{\sqrt{2}}{2}\sin(x) + \frac{\sqrt{2}}{2}\cos(x)$ and $\cos(x + \frac{\pi}{4}) = \frac{\sqrt{2}}{2}\cos(x) - \frac{\sqrt{2}}{2}\sin(x)$.) Under this basis, the coordinates for $g(x)$ will now be $g_{\mathcal{C}} = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$.

As you can see, coordinates are ephemeral, i.e., they change all the time. Think of the vector g as a person, and coordinates of g as "names" of g. A person can change its name all the time, but it is still the same person. When we want to understand a solution to the differential equation $f''(x) = -f(x)$, it is usually better to deal with the abstract vector g. The coordinates are only temporary tools to help us do some calculations, but they are not innate to g. Coordinates depends on our choice of basis $\mathcal{B}$, and $\mathcal{B}$ has nothing to do with g. ☺

We now see some examples of vector spaces for some very important distinctions.

Example 4.4.10 (Subspaces of $\mathbb{R}^n$)**.** Consider $\mathbb{R}^3$ and the plane W defined by $x+y+z=0$. If $\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}$ and $\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$ lies on this plane (i.e., $x_1+y_1+z_1 = x_2+y_2+z_2 = 0$), then obviously their sum still lies on this plane, and their scalar multiples would also lie on this plane. So this is a subspace.

Any line through the origin in $\mathbb{R}^n$ is a vector space. Any plane through the origin is a vector space. Any solution set to $A\boldsymbol{x} = \boldsymbol{0}$ for any matrix A is a vector space, because $A\boldsymbol{x} = \boldsymbol{0}$ and $A\boldsymbol{y} = \boldsymbol{0}$ would imply that $A(s\boldsymbol{x} + t\boldsymbol{y}) = \boldsymbol{0}$.

In fact, these are ***subspaces*** of $\mathbb{R}^n$, because they are subsets who are also vector spaces using the vector addition and scalar multiplication inherited from $\mathbb{R}^n$.

In contrast, take the plane $x+y+z=1$ in $\mathbb{R}^3$. Then $\boldsymbol{e}_1, \boldsymbol{e}_2$ is in it, but $2\boldsymbol{e}_1, \boldsymbol{e}_1+\boldsymbol{e}_2$ are not in it. This is NOT a vector space. In particular, any subspace must contain the origin of the ambient space.

In short, only planes (and lines and hyperplanes and so on) that contains the origin can be subspaces. Those that do not contain the origin are not. (They are called affine spaces, because they are translations of vector spaces.) ☺

Example 4.4.11 (Coordinates and Nature)**.** Consider $\mathbb{R}^3$ and the plane W defined by $x+y+z=0$. This is a subspace, and as a plane it should have dimension two.

Now pick an element of W, say $\boldsymbol{w} = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$. Wait, if $\boldsymbol{w} \in W$, and W is two dimensional, then why would $\boldsymbol{w}$ have three coordinates, not two?

The correct view is to treat $\begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}$ not as coordinates, but rather as the nature of $\boldsymbol{w}$. To find the coordinates of $\boldsymbol{w}$ in W, we first need to pick a basis for W. Say the basis $\mathcal{B}$ made of $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$. Then the coordinates of $\boldsymbol{w}$ are $\boldsymbol{w}_{\mathcal{B}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, because $\boldsymbol{w}$ is the sum of the first basis vector and twice the second basis vector.

When we are dealing with vectors in $\mathbb{R}^n$, since we almost always us the standard basis, the nature of any vector $\boldsymbol{v} \in \mathbb{R}^n$ would coincide with the coordinates of $\boldsymbol{v}$ under the standard basis. However, if we are dealing with a vector $\boldsymbol{w}$ in some subspace W of $\mathbb{R}^n$, then the nature of $\boldsymbol{w}$ is a list of n real numbers, but the coordinates of $\boldsymbol{w}$ will depend on our choice of basis for W, and the coordinates will be a list of $\dim W$ real numbers. ☺

Here is a definition used above.

Definition 4.4.12. *A subset W of a vector space V is a* ***subspace*** *if it is also a vector space itself, using the same vector addition and scalar multiplication as V.*

Here let us see some more exotic examples of subspaces. In particular, a set may be both a vector space and NOT a vector space, depending on how you define your vector addition and scalar multiplication.

Example 4.4.13. Consider the plane $x+y+z=1$ in the space $\mathbb{R}^3$. We already know that this is NOT a vector space under usual vector addition and scalar multiplication. But what if we change the definition of scalar multiplication and vector addition?

Suppose we define $\begin{bmatrix} a \\ b \\ c \end{bmatrix} \boxplus \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a+x-1 \\ b+y \\ c+z \end{bmatrix}$, and $k \boxtimes \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ka-k+1 \\ kb \\ kc \end{bmatrix}$. Then you can go ahead and verify that this gives a vector space structure for the plane $x+y+z=1$.

This is not an aribtrary structure. This structure is called "declare $\begin{bmatrix}1\\0\\0\end{bmatrix}$ to be the origin", and thus transforming an affine space into a vector space. Can you see why? Check that $\boldsymbol{e}_1$ is indeed the "zero vector" for $\boxplus$.

Let V be the plane $x+y+z=1$ with $\boxplus, \boxtimes$ as its operations, then it is a vector space. Let W be the plane $x+y+z=0$, so it is also a vector space. Consider the map $f: W \to V$ that goes like $f(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{e}_1$. Then you can in fact verify that $f(a\boldsymbol{x} + b\boldsymbol{y}) = a \boxtimes f(\boldsymbol{x}) \boxplus b \boxtimes f(\boldsymbol{y})$. So this f is a linear map. In fact, this is a bijection with inverse $g(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{e}_1$.

In short, the "vector space" V is obtained by shifting W while preserving its linear structure faithfully. ☺

Remark 4.4.14. *Here the requirement is used to rule out the following weird phenomena: Sometimes V is a vector space, and under the vector addition and scalar multiplication of V, W is NOT a subspace. (Say $V = \mathbb{R}^3$ and W is the plane given by $x+y+z=1$.)*

However, we may define some weird and exotic vector addition and scalar multiplication, which shall make W into a vector space. (We shall see this construction later.) Then W by itself is now a vector space. Nevertheless, it is still NOT a subspace of V, because it does not have the same vector addition and scalar multiplication as V. In fact W will never be a subspace of V.

Proposition 4.4.15. *A subset W of V is a subspace of V if and only if it is closed under vector addition and scalar multiplication of V. (In particular, for any $\boldsymbol{v}, \boldsymbol{w} \in V$ and $k \in \mathbb{R}$, we should have $\boldsymbol{v}+\boldsymbol{w}, k\boldsymbol{v} \in V$.)*

Why would something fail to be a subspace? Here are the most common scenarios.

Example 4.4.16.
1. (Addition not closed) Take the dual cone $z^2 = x^2 + y^2$ in the space $\mathbb{R}^3$. Here $\pm\boldsymbol{e}_1 + \boldsymbol{e}_3$ are both in it, but their sum $2\boldsymbol{e}_3$ is NOT in this cone. So addition is not defined on the cone. (And can only be defined on the larger set $\mathbb{R}^3$.) (However, scalar multiplication is fine though.)
2. (Scalar multiplication not closed) Take the first quadrant of $\mathbb{R}^2$. This is NOT a vector space, because if $\boldsymbol{v} \neq \boldsymbol{0}$ is in it, then $(-1)\boldsymbol{v}$ is NOT in it. So scalar multiplication is NOT defined on this set, but rather, only on the larger set $\mathbb{R}^2$.

☺

Corollary 4.4.17. *A subspace W must contain the zero vector of V. And the **inclusion map** $\iota: W \to V$ defined as $\boldsymbol{w} \mapsto \boldsymbol{w}$ is linear. (As a side note, the inclusion map is always injective, but maybe not surjective, unlike the identity map.)*

Note that, if W uses a different definition of vector addition and scalar multiplication, then the inclusion map is not going to be linear any more.

Definition 4.4.18 (Useful Subspaces for Matrices)**.** *Given an $m \times n$ matrix A, the following four subspaces are sometimes called the **fundamental subspaces for** A.*

1. $\mathrm{Ran}(A)$*, the **range** or **column space** which contains all vectors that are images of the linear map A. (I.e., all possible linear combinations of columns of A). This is a subspace of $\mathbb{R}^m$.*
2. $\mathrm{Ker}(A)$*, the **kernel** or **zero set** or **null space**, which contains all solutions $\boldsymbol{x}$ to $A\boldsymbol{x} = \boldsymbol{0}$. This is a subspace of $\mathbb{R}^n$.*

Proposition 4.4.19. $\mathrm{Ran}(A)$ *and* $\mathrm{Ker}(A)$ *are indeed subspaces.*

Proof. Take any $\boldsymbol{v}, \boldsymbol{w} \in \mathrm{Ran}(A)$ and any $a, b \in \mathbb{R}$. Then by definition of range, $\boldsymbol{v} = A\boldsymbol{x}$ for some $\boldsymbol{x}$ and $\boldsymbol{w} = A\boldsymbol{y}$ for some $\boldsymbol{y}$. Then $a\boldsymbol{v} + b\boldsymbol{w} = A(a\boldsymbol{x} + b\boldsymbol{y}) \in \mathrm{Ran}(A)$. So this is indeed a subspace.

Take any $\boldsymbol{v}, \boldsymbol{w} \in \mathrm{Ker}(A)$ and any $a, b \in \mathbb{R}$. Then by definition of range, $A\boldsymbol{v} = A\boldsymbol{w} = \boldsymbol{0}$. Then $A(a\boldsymbol{v} + b\boldsymbol{w}) = aA\boldsymbol{v} + bA\boldsymbol{w} = a\boldsymbol{0} + b\boldsymbol{0} = \boldsymbol{0}$. □

Proposition 4.4.20. *For an $m \times n$ matrix A, the corresponding linear map is surjective if and only if* $\mathrm{Ran}(A) = \mathbb{R}^m$. *It is injective if and only if* $\mathrm{Ker}(A) = \{\mathbf{0}\}$.

Proof. The first statement is literally the definition of surjectivity.

For the second statement, if A is injective, then $A\boldsymbol{x} = \mathbf{0}$ will only have one solution. Therefore it has to be $\boldsymbol{x} = \mathbf{0}$. So $\mathrm{Ker}(A) = \{\mathbf{0}\}$. Conversely, suppose $\mathrm{Ker}(A) = \{\mathbf{0}\}$. Suppose $A\boldsymbol{v} = A\boldsymbol{w}$, then $A(\boldsymbol{v} - \boldsymbol{w}) = \mathbf{0}$, and hence $\boldsymbol{v} - \boldsymbol{w} \in \mathrm{Ker}(A)$ and it has to be the zero vector. So $\boldsymbol{v} = \boldsymbol{w}$. Hence A is injective. □

Now we move away from subspaces, and onto more abstract spaces and linear maps.

Example 4.4.21 (Spaces of Matrices)**.**

1. All $m \times n$ matrices is a vector space $M_{m\times n}$, because we can add matrices and scale matrices. This space is mn dimensional, and a basis is made of matrices of the form $\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}$, i.e., it has only a single entry 1 and all other entries are zeros.
2. In fact, given any vector spaces V, W, the set of linear maps from V to W is a vector space $\mathcal{L}(V, W)$. We can add linear maps or scale a linear map in the obvious manner. (I.e., $(f + g)(\boldsymbol{v}) = f(\boldsymbol{v}) + g(\boldsymbol{v}), (kf)(\boldsymbol{v}) = k(f(\boldsymbol{v}))$.)
3. The ***trace*** of a square matrix is the sum of its diagonal entries. This is a linear map from $M_{n\times n} \to \mathbb{R}$.
4. The transpose is a linear map from $M_{m\times n}$ to $M_{n\times m}$.
5. The space of all $n \times n$ upper triangular matrices is a vector space and in fact a subspace of $M_{n\times n}$. Can you find its dimension and find a basis?
6. Transpose can be restricted to a map from the space of upper triangular matrices to the space of lower triangular matrices. (This will be a linear bijection.)
7. The space of unit upper triangular matrices is NOT a vector space. If U_1, U_2 are unit upper triangular, then $U_1 + U_2$ is NOT. (It also does not contain the zero matrix, hence not a subspace.)
8. The space of all $n \times n$ invertible matrices is NOT a vector space. If A is invertible, then $A - A$ is not. (It also does not contain the zero matrix, hence not a subspace.)
9. The space of all $n \times n$ symmetric matrices is a vector space. Can you find its dimension and find a basis? The transpose map restricted to this domain and codomain is the identity map.
10. Matrix inversion is NOT a linear map, because in general, $(A + B)^{-1} \neq A^{-1} + B^{-1}$.
11. Let us say a 3×3 matrix is a ***magic matrix*** if entries in each row, each column and each of the two diagonals add up to the same number. For example, we have the famous $\begin{bmatrix} 2 & 9 & 4 \\ 7 & 5 & 3 \\ 6 & 1 & 8 \end{bmatrix}$. Then all such matrices form a subspace of $M_{3\times 3}$. Can you see why?

☺

So far, we have been concerning ourselves with real vector spaces. However, we sometimes might want to change the realm of coefficients, say maybe we want to allow all complex numbers to be coefficients. This will NOT affect vector addition, but when we do scalar multiplications, we can scale by more choices.

Example 4.4.22.

1. The set of complex numbers $\mathbb{C}$ is a real vector space (of dimension 2 and basis 1 and i), since we can do real linear combinations of complex numbers.
2. However, the set of complex numbers $\mathbb{C}$ is also a complex vector space. If we allow all complex numbers to be scalars, then all elements of $\mathbb{C}$ are scalar multiples of 1. So 1 is itself a basis for $\mathbb{C}$, and $\mathbb{C}$ has complex dimension one. We can use $\mathbb{C}^n$ to denote the standard n-dimensional complex vector spaces, and everything works pretty much the same as $\mathbb{R}^n$, except that now we have complex numbers in place of real numbers.

3. The set of real numbers $\mathbb{R}$ is an infinite dimensional vector space over the set of rational numbers $\mathbb{Q}$. This example is just for fun.

☺

A large part of calculus concerns linear maps on function spaces. These are infinite dimensional spaces

Example 4.4.23. 1. The space of all continuous functions from any domain D to $\mathbb{R}$, $\mathcal{C}(D)$, is a vector space. Here we add or scale functions in the obvious manner, i.e., $(f+g)(x) = f(x) + g(x)$ and $(kf)(x) = kf(x)$ for all $x \in D$. Here D can be $\mathbb{R}$, or any interval, or the set of all people with f sending each person to their age/height/salary, etc..

2. The space of all differentiable functions from $\mathbb{R}$ to $\mathbb{R}$ is a vector space. Here we add or scale functions in the obvious manner.

3. The space of all smooth functions (i.e., infinitely differentiable) from $\mathbb{R}$ to $\mathbb{R}$, $\mathcal{C}^\infty(\mathbb{R})$, is a vector space. Here we add or scale functions in the obvious manner. (Usually $\mathcal{C}^k(\mathbb{R})$ means the space of all continuously k-times differentiable real functions.)

4. The differentiation map $\frac{d}{dx}$ from the space of continuously differentiable real function $\mathcal{C}^1(\mathbb{R})$ (functions whose derivatives are continuous) to the space of continuous functions $\mathcal{C}(\mathbb{R})$ is a linear map.

5. The integration map from a to b for any $a, b \in \mathbb{R}$ is a linear map $\int_a^b$ from the space $\mathcal{C}([a,b])$ to $\mathbb{R}$. (Just an optional side note. If vectors in $\mathbb{R}^n$ are "column vectors", then linear maps $\mathbb{R}^n \to \mathbb{R}$ are "row vectors". In this perspective, elements (functions) in $\mathcal{C}([a,b])$ are "columns", then the integration $\int_a^b$ is a "row".)

6. The limit operation $\lim_{n\to\infty}$ is a linear map from the space of convergent sequences to $\mathbb{R}$.

☺

Now, the following example illustrates that it is VERY important to figure out what is this underlying vector space, and VERY important to see if linear combinations make sense. And sometimes (when linear combinations make no sense) certain things are best NOT treated as a vector space.

Example 4.4.24 (Optional)**.** The spectral colors (colors on a rainbow) depends only on their wavelength. Think about the wavelength as elements of $\mathbb{R}$, then we see that all spectral colors are subsets of the vector space $\mathbb{R}$. In this sense, color is one-dimensional.

However, what about those RGB nonsense? In a computer, a color is usually encoded with a TRIPLE of real numbers, indicating the amount of red, green and blue. Does that not make color three dimensional?

It turns out that there are NON-spectral colors, and this is peculier to human. In the eyes of a human there are three types of cone cells, and they are activated by roughly the color red, green and blue respectively. Depending on how the three types of cone cells are activated, our brain will decide its color.

For example, we have the color violet, and the color purple, and they appear similar in our eyes. Violet feels a little more blue-ish, and it is in fact a spectral color. It has objectively NO relation with red or blue, and it is simply a different wave length. Purple is essentially a mixture of red and blue. So why do they appear similar in our eyes?

It turns out that when we see violet, weirdly our "red" cone sell is activated by a tiny bit. So, even though violet has nothing to do with red, our brain is convinced that we see a hint of redness in it. For this reason, human thinks violet and purple are similar. There are many animals with different numbers and kinds of cone cells, and to them, maybe violet and purple are NOT similar at all. "Foolish humans...." They would think.

So, as far as us humans are concerned, since we have three kinds of cone cells, our color perception depends on how much each type of cone cells is activated. So we need three real numbers to express our color perception. So even though objective spectral colors form a one-dimensional structure, our human color perception form a three-dimensional structure.

Color is objective, but the perception of color (whether two colors are "close") is subjective.

Furthermore, before we declare that we are working in a vector space, we need to think about the meaning of vector addition and scalar multiplication. On the spectral color line, what does it mean to add wavelength? Probably no meaning. So it might be best that we DON'T think of it as a vector space. We need the continuous structure of $\mathbb{R}$ more than its linear structure.

On the space $\mathbb{R}^3$ of human color perceptions, we add RGB coefficients all the time in computers. It is a bit similar to mixing colored lights. So it is OK to treat this as a vector space. ☺

4.5 Basis and Dimensions

4.5.1 Linear Combination Map and Coordinate Map

In advanced mathematics, teacher would try to show you that everything is a map. In linear algebra, I endeavor to show you that everything is a linear map. In particular, a basis is a linear map.

Given a vector space V and a basis $\mathcal{B}$, we can write each vectors in V into coordinates. This gives rise to a natural map:

Definition 4.5.1. *For any ordered basis $\mathcal{B} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ in a vector space V, we define the **coordinate map** to be the map $(-)_\mathcal{B}$ from V to $\mathbb{R}^n$ such that each input vector is sent to its coordinates under $\mathcal{B}$. I.e., we send $\boldsymbol{v}$ to $\boldsymbol{v}_\mathcal{B}$.*

Note that the order of the basis vectors is IMPORTANT, because if we change the order of basis vectors, the resulting coordinates will also change order.

Let us also define the inverse process.

Definition 4.5.2. *For any ordered collection of vectors $\mathcal{C} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ in a vector space V, we define the **linear combination map** to be the map from $\mathbb{R}^k$ to V such that input coordinates are used as coefficients to do linear combinations of these vectors.*

We sometimes lazily use $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ to denote this map, so we have $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = \sum a_i \boldsymbol{v}_i$, which agrees with our matrix multiplication instincts. We also simply write $\mathcal{C} \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix}$ to denote this linear combination.

Note the similarity between this and matrix-vector multiplications. However, the map $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ is technically not a matrix, merely an abstract linear map. And $\boldsymbol{v}_i$ are abstract vectors that are usually NOT a column of numbers. Nonetheless, the i-th "column" $\boldsymbol{v}_i$ is exactly the image of $\boldsymbol{e}_i \in \mathbb{R}^k$ under the map $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$, which is why the formula looks familiar.

From now on, if we say we pick a basis $\mathcal{B}$, then this is an ordered collection of vectors $\mathcal{B} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ that form a basis. Then we can also use the notation $\mathcal{B}$ for the linear combination map $\mathcal{B} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) : \mathbb{R}^n \to V$. We no longer distinguish between a basis and its corresponding linear combination map.

Now, for any vector $\boldsymbol{v} \in V$ with basis $\mathcal{B}$, clearly $\boldsymbol{v}$ is the linear combinations of vectors in $\mathcal{B}$ with coefficents as $\boldsymbol{v}_\mathcal{B}$. In particular, we always have $\boldsymbol{v} = \mathcal{B}(\boldsymbol{v}_\mathcal{B})$ for all $\boldsymbol{v} \in V$. Therefore, we see that the coordinate map is always the inverse of the linear combination map.

Proposition 4.5.3. *For a vector space V and its basis $\mathcal{B}$, the coordinate map with respect to the basis $\mathcal{B}$ is the inverse of the linear combination map, i.e., the coordinate map is $\mathcal{B}^{-1}$.*

Proof. See discussion above. □

Corollary 4.5.4. *For a vector space V and its basis $\mathcal{B}$, the map $\mathcal{B} : \mathbb{R}^n \to V$ is a linear bijection. In particular, the coordinate map is also linear.*

Proof. $\mathcal{B}$ is linear because it is doing linear combinations. And $\mathcal{B}$ is bijective because it has an inverse. And finally, the coordinate map is the inverse of a linear map, so it is also linear. □

In short, what is a basis? A basis is a linear bijection that connectes V with $\mathbb{R}^n$. The converse is also true.

Proposition 4.5.5. *$\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ form a basis for V iff the linear combination map $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is bijective.*

Proof. $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ being surjective means all vectors in V are linear combinations of these vectors.

$(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ being injective means distinct linear combinations of these vectors give distinct results. In particular, if a vector in V is a linear combination of these vectors, then the coefficients are unique.

Now the statement looks just like a trivial repetition of the definition.... □

Extracting from this proof, we have some really interesting new concepts.

Definition 4.5.6. *For any ordered collection of vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ in a vector space V, we say they are linearly independent if distinct linear combinations of them gives distinct results. (I.e., linear combination map $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is injective.)*

Definition 4.5.7. *For any ordered collection of vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ in a vector space V, we say they are spanning if all vectors of V are linear combination of them. (I.e., linear combination map $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is surjective.) We also say that $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ span the space V.*

Corollary 4.5.8. *$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ form a basis iff it is linearly independent and spanning.*

These concepts are defined exactly the same as the cases in $\mathbb{R}^n$, and they behave the same way.

Example 4.5.9. Intuitively, linear independence means no redundancy.

Say your goal is to use vectors to span $\mathbb{R}^2$. (I.e., we want to use some vectors to express all other vectors.) You pick $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_1 + \boldsymbol{e}_2$. Oops, the last vector is redundant. So these three vectors are NOT linearly independent.

Spanning means we have enough vectors to span everything.

Together, if you have enough vectors to span everything, and none is redundant, then you have a basis. ☺

4.5.2 Existence of Basis and Uniqueness of Dimension

The idea is to grab "good" vectors one by one, until they form a basis. However, we need to make sure that the vectors we grab are linearly independent. Here are some useful lemmas for future use.

Lemma 4.5.10 (Criteria for independence)**.** *TFAE:*

1. *Vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are linearly independent.*
2. *$\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$ implies that all $a_i = 0$.*
3. *None of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ is zero and none is a linear combination of the rest.*

Proof. The proofs here are not the most efficient ones. However, it might be beneficial for you to see more flavors of proofs.

We prove the equivalence cyclically. If they are linearly independent (i.e., all vectors has unique coordinates), then $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ is injective. So if $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = \boldsymbol{0} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)\boldsymbol{0}$, we must have $\begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = \boldsymbol{0}$, i.e., all a_i are zero.

Now suppose $\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$ implies that all $a_i = 0$. Suppose for contradiction that $\boldsymbol{v}_i = \sum_{j \neq i} a_j \boldsymbol{v}_j$ (this include the case of $\boldsymbol{v}_i = \boldsymbol{0}$), then we have $\boldsymbol{v}_i - \sum_{j \neq i} a_j \boldsymbol{v}_j = \boldsymbol{0}$. And this is a linear combination of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$

and the coefficients are not all zero. In particular, $a_i = 1 \neq 0$. Contradiction, Hence none of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ is zero and none is a linear combination of the rest.

Now suppose none of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ is zero and none is a linear combination of the rest. Let us prove the injectivity of $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$. If $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$. Suppose for contradiction that some $a_i \neq b_i$, say WLOG (without loss of generality) that $a_1 \neq b_1$. Then $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$ implies that $\mathbf{0} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \begin{bmatrix} a_1 - b_1 \\ \vdots \\ a_k - b_k \end{bmatrix} = \sum (a_i - b_i) \boldsymbol{v}_i$, and therefore $\boldsymbol{v}_1 = \sum_{i \neq 1} \frac{a_i - b_i}{a_1 - b_1} \boldsymbol{v}_i$. So $\boldsymbol{v}_1 = \mathbf{0}$ or is a linear combination of the rest, which cannot happen. So we have $a_1 = b_1$, and similarly $a_i = b_i$ for all i. Hence $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ is injective, the vectors are linearly independent. □

Here the best tool is the second criterion. The third is hard to use, but it paints a nice conceptual picture, and explains the reason for the word "independence".

Let us see how the criterion is used in practice.

Lemma 4.5.11 (Independence extension lemma)**.** *If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are linearly independent, and $\boldsymbol{v}_{k+1}$ is NOT a linear combination of them, then $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k+1}$ are linearly independent.*

Proof. Suppose $\sum_{i=1}^{k+1} a_i \boldsymbol{v}_i = \mathbf{0}$. Suppose $a_{k+1} \neq 0$, then we have $v_{k+1} = \sum_{i=1}^{k} \frac{a_i}{a_{k+1}} \boldsymbol{v}_i$, impossible. So $a_{k+1} = 0$. But then we have $\sum_{i=1}^{k} a_i \boldsymbol{v}_i = \mathbf{0}$, so by linear independence of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$, we see all $a_i = 0$. □

So with this tool in mind, let us prove the existence of a basis and a dimension.

Definition 4.5.12. *A vector space is **finite dimensional** if it cannot have infinitely many linearly independent vectors. Otherwise it is **infinite dimensional**.*

Theorem 4.5.13. *A finite dimensional vector space V has a basis.*

Proof. If $V = \{0\}$, then by convension we say $\varnothing$ is its basis. This is just a convention.

If $V \neq \{0\}$, then pick any nonzero vector $\boldsymbol{v}_1$. If this single vector is spanning, then it is a basis all by itself, so we are done.

If it is not spanning, then there are vectors in V that is NOT a multiple of $\boldsymbol{v}_1$. Say let us pick any one, $\boldsymbol{v}_2$. Then by the independence extension lemma, $\boldsymbol{v}_1, \boldsymbol{v}_2$ is linearly independent. If it is now spanning, then we have found a basis, done.

If it is not spanning, then there are vectors in V that is NOT a linear combination of $\boldsymbol{v}_1, \boldsymbol{v}_2$. Say let us pick any one, $\boldsymbol{v}_3$. Then by the independence extension lemma, $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ is linearly independent. If it is now spanning, then we have found a basis, done.

If not, we just keep going. We cannot keep going forever, because by definition we cannot have infinitely many linearly independent vectors. This process must terminate somewhere, and when it terminates, we have a basis. □

Example 4.5.14. Let us see this in action. Consider the vector space $\mathcal{P}_2$, the space of polynomials of degree at most 2. How to find a basis?

First we pick any non-zero element, say the constant polynomial 1. Does this spans $\mathcal{P}_2$? Not yet. Obviously it only spans all constant polynomials.

So we pick anything that is NOT a constant polynomial, say x. Now, does $1, x$ span $\mathcal{P}_2$? Not yet. Their linear combination can reach any degree 0 or degree 1 polynomial, but they can never combine to give you a degree two polynomial.

So we pick anything that is NOT a degree 0 or degree 1 polynomial, say x^2. Now they span everything.

So we have a basis $1, x, x^2$ for our space $\mathcal{P}_2$.

Note that this gives a bijection, the linear combination map $(1, x, x^2) : \mathbb{R}^3 \to \mathcal{P}_2$ that sends $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ to $a + bx + cx^2$. Note that the "columns" of the map $(1, x, x^2)$ are not really columns, but merely abstract vectors.

Conversely, for any polynomial $a + bx + cx^2 \in \mathcal{P}_2$, we can try to find the coordinate of this polynomial under the basis $1, x, x^2$. This is obviously just $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$. Hence we have a ***coordinate map*** $C : \mathcal{P}_2 \to \mathbb{R}^3$ that sends $a + bx + cx^2$ to its coordinates $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ under the basis $1, x, x^2$. Obviously C and $(1, x, x^2)$ are inverse map of each other. We always have such a coordinate map, because by the definition of basis, $(1, x, x^2)$ is always a linear bijection. ☺

So it seems that any finite dimensional vector space V has a bijective linear map to some $\mathbb{R}^n$. We want to call this n the dimension of the space. But this needs one more theorem.

Theorem 4.5.15 (Dimension is well-defined)**.** *In a finite dimensional vector space, any two basis contain the same number of elements. And this number is called the* ***dimension*** *of this space.*

Proof. Suppose V has a basis with m vectors, then there is a bijective linear map between $\mathbb{R}^m$ and V. Now suppose V has another basis with n vectors. Then there is a bijective linear map between $\mathbb{R}^n$ and V. So there is a bijective linear map between $\mathbb{R}^n$ and $\mathbb{R}^m$, i.e., an invertible $m \times n$ matrix! But this could only happen when $m = n$. □

So we have established the following idea: An abstract vector space is either infinite dimensional, or it has a basis of n vectors where n is by definition the dimension of this space. Note that a space usually have many many basis, but all of them has the same number of vectors.

Example 4.5.16. Another basis for $\mathcal{P}_2$ is $1, x+1, (x+1)^2$. For any polynomial $p(x)$, we can express it as a unique linear combination of these three. Note that $1 = 1, x = (x+1) - 1$ and $x^2 = (x+1)^2 - 2(x+1) + 1$, so everything that is spanned by $1, x, x^2$ is spanned by $1, x+1, (x+1)^2$. Furthermore, it is obvious that for $1, x+1, (x+1)^2$, each one is NOT a linear combination of previous ones. So by independence extension lemma they are linearly independent.

To find coordinates for $p(x) \in \mathcal{P}_2$, suppose $p(x) = a+b(x+1)+c(x+1)^2$. Then $p(-1) = a$ and $p'(-1) = b$ and $p''(-1) = 2c$. So the coordinates are $\begin{bmatrix} p(-1) \\ p'(-1) \\ \frac{p''(-1)}{2} \end{bmatrix}$ or $\begin{bmatrix} \frac{p(-1)}{0!} \\ \frac{p'(-1)}{1!} \\ \frac{p''(-1)}{2!} \end{bmatrix}$ if you like to generalize this to higher degree polynomials. (The similarity to Tayler expansion in calculus is NOT a coincidence, but out of the scope of this class. We don't explore that here in our linear algebra class.)

Take P_2, the space of polynomials of degree at most 2. Take a vector $\boldsymbol{w} = 3x^2 + 4x + 1$ and basis $1, x+1, (x+1)^2$. Then the linear combination map is $L : \mathbb{R}^3 \to P_2$ that sends $\boldsymbol{e}_1, \boldsymbol{e}_2$ and $\boldsymbol{e}_3$ to $1, x+1, (x+1)^2$. we have $L^{-1}(\boldsymbol{w}) = \begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix}$ by our previous formula. You can check that we indeed have $3x^2 + 4x + 1 = -2(x+1) + 3(x+1)^2$.

The basis $1, x+1, (x+1)^2$ is preferable than $1, x, x^2$ when x is near -1 most of the time. ☺

So, it turns out that a vector space is either infinite dimensional, or essentially $\mathbb{R}^n$ because it has a linear bijection to $\mathbb{R}^n$. In linear algebra, we call a bijective linear map an ***isomorphism***. This is a very important concept in all of mathematics, it means two things are structurally the same, and they only differ in names.

Example 4.5.17. Most practically, isomorphisms allows us to "transfer" calculations. Say we have a bijective linear map $L : V \to W$. Then to do calculation in V, it is enough to do the corresponding calculation in W.

Suppose I want to add $v_1 \in V$ and $v_2 \in V$, but I don't know how to do this. I only know how to add things in W. Well, no worry. First I send then into W by applying L. Now I add $L(v_1)$ and $L(v_2)$, which I know how to do, because I'm now in W. Then I apply the inverse of L. So I see that $v_1 + v_2 = L^{-1}(L(v_1) + L(v_2))$.

Most practically, if we find a basis for V, this means we have a bijection $V \leftrightarrow \mathbb{R}^n$. Then any calculation we want to do in V, we can simply do it in $\mathbb{R}^n$ instead. Just write vectors of V in terms of coordinates, and now we are in $\mathbb{R}^n$, where everything is familiar. After we finish calculation in $\mathbb{R}^n$, just do the linear combination according to our chosen basis in V to go back to V. ☺

Remark 4.5.18 (Optional Remarks on Isomorphisms)**.** *This is the idea of* ***isomorphism****. For any two things with structures, we say they are isomorphic if, from some perspective (i.e., through some bijective function), I can almost pretend that they are the same thing (i.e., the bijective function happens to matches the structures on both sides exactly).*

For another example of an isomorphism, in Chinese we know that yi plus yi is er. Then I build a function, called a dictionary, that maps yi to one, er to two, and so forth. Then I see that one plus one is two, this is still correct. So the dictionary is a function that matches the mathematical structure of things. Then USING this isomorphism, if I prove some mathematical theorem in Chinese, then I don't really need to prove it in English again. I know the mathematical theorem will be true in English as well, I know this even without doing any translation myself. The very EXISTENCE of a dictionary already guarantees that the same mathematical theorem should be true in English as well.

The idea is this: as far as linear structures are concerned, the existence of a bijective linear map means the two things have IDENTICAL linear strucutre. So if I FIX a bijective map between two vector spaces, then I can PRETEND that they are identical. I'll be fine as long as I always go back and forth according to the same bijection.

And under this pretending, I can see that the two things are practically indistinguishable. Say there is a bijective linear map between V, W. Then if V has dimension n, then W must also have dimension n. If all triangles in V have concurrent medians, then all triangles in W have concurrent medians. If everyone living in V loves Chinese food in a linear way, then everyone living in W loves Chinese food in the same linear way. In short, whenever I prove some big theorem in V about its linear structure, then the same theorem must be true for W as well. Let me repeat: as long as I FIX this bijective linear map, then they are indistinguishable, they are identical, there are no difference between them.

However, if one day, I decide to go back and forth using ANOTHER linear bijection between them, then all of a sudden, all previous "pretending" no longer works. This is just like if I pick another basis in V, then the expressions are now all different, and angles and length of vectors are different, and so on. It is like waking up from a dream....

So this is how it works. Say you prove something about V (say V has dimension n). To transfer this property to W, you first build a bijective linear map between them. Now W have the same property from the perspective of this linear map. Then you show that this property of W is INNATE, that it is really just about W itself, independent of any choice of basis and independent of how W is connected to other spaces. Then W has this property INDEPENDENT from the bijection of our choice. (You see why being independent of choice of basis is important.)

Coordinates of a vector, entries of a matrix, angles and dot products, these things are all DEPENDENT of basis. They are all illusions. The true linear property are those independent of basis. If $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, then the $(1,1)$ entry of A will change under a different basis, but the surjectivity of A shall never change. It will always remain surjective.

All in all, the correct way to say this is the following: Not all finite dimensional spaces are $\mathbb{R}^n$, but each of them looks like some $\mathbb{R}^n$ for some n.

Here are some easy result that follows from the power of isomorphism:

Theorem 4.5.19. *For any vector space V, suppose it has a basis of n elements. Then we have the following facts:*

1. *Every basis must have exactly n vectors.*
2. *A collection of more than n vectors must be linearly DEPENDENT.*
3. *A collection of less than n vectors cannot be spanning.*
4. *A collection is a basis if and only if it is linearly independent and has n vectors.*
5. *A collection is a basis if and only if it is spanning and has n vectors.*
6. *Every linearly independend collection of vectors can be extended to a basis.*
7. *Every spanning system contains a basis.*

Proof. First statement is already done.

For the second and third statement, pretend your space is $\mathbb{R}^n$, and we have a collection $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Then the linear combination map goes from $\mathbb{R}^k$ to $\mathbb{R}^n$, hence it is an $n \times k$ matrix. So if $k > n$, it cannot be injective (and the collection cannot be linearly independent). If $k < n$, then it cannot be surjective (and the collection cannot be spanning).

For the fourth and fifth statement, this is because an $n \times k$ matrix is a bijection if and only if it is square and injective, and if and only if it is square and surjective.

For the last two statements, simply combine the results above. Say we have a linearly independent collection. Then as long as I have less than n vectors, i shall repeatedly use independence extension lemma, until I reach n vectors, and at that moment I must have a basis.

If we have a spanning collection, then as long as I have more than n vectors, i shall have redundant vectors. I throw them away, until I reach n vectors, and at that moment I must have a basis. □

Remark 4.5.20. *So say we are in an n dimensional space. To find a basis, we first pick any v_1. Then we pick any v_2 linearly independent from v_1. Then we pick any v_3 linearly independent from v_1, v_2, and so on. As soon as we hit v_n, we are done. This must be a basis. There is no need to check if it is spanning. (Interesting question to think about: When I pick my v_3, why v_2 cannot be a linear combination of v_1 and v_3?)*

Alternatively, say we have lots of vectors that would certainly span our n dimensional space. Then we can keep throwing away redundant vecotrs (vectors that are linear combination of others). When we have n vectors left, we must arrive at a basis. There is no need to check linear independency.

Here we have a special corollary for subspaces.

Definition 4.5.21. *For vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, their **span** is the subspace of all linear combinations of them.*

Corollary 4.5.22. *If vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are linearly independent, then their span has dimension k.*

Remark 4.5.23. *With the idea above, we see that the independency extension lemma is the formal rigorous statement of the following informal intuition: for an m dimensional subspace, if it moves in a new direction (new vector not contained in the original subspace), then we get an $m+1$ dimensional subspace. Point moves to make a line, and a line moves to make a plane, these are all special cases of the lemma.*

And linear dependence happen exactly like this: we add a new vector to our collection, but it fails to contribute a new dimension!

4.5.3 Dimensionality from first principles (Optional)

In the proof above, we took a huge detour. Long ago we established Gaussian elimination, and from the RREF of matrices we see that $\mathbb{R}^m$ and $\mathbb{R}^n$ cannot be isomorphic. (I.e., they have different linear structure.) This gives us that dimension for abstract vector spaces are well-defined.

However, can we do this without the detour? Can we prove this fact by simply using AXIOMS of a vector space, i.e., from first principles? Yes. Warning though, the proof is a bit abstract and hardcore. (But the idea is simple: you eat, then you poop, then you eat, then you poop, and repeat, until you eat all that is required.)

Remark 4.5.24. *The benefit of doing this in first principles is that it can be generalized to infinite dimensions. With some minor changes to the proofs below, one can show that any vector space has a "Hammal" basis, a subset H such that any vector of V is a linear combination of some (finitely many) vectors in H, and everything in H are linearly independent.*

The most powerful lemma is of course again the independence extension lemma. But let us rewrite it in a different form.

Lemma 4.5.25 (Extending Linear Independency)**.** *For any vector space V, if L is a linearly independent collection of vectors, and a vector $\boldsymbol{w}$ is not in the span of L, then $L \cup \{\boldsymbol{w}\}$ is still linearly independent. (Here L is treated as a subset of V.)*

In contrast, if you have a spanning collection, but it is NOT linearly independent, then it means you have some redundent vectors here. You may simply throw them away. However, WHAT to throw a way can be tricky.

Example 4.5.26. Say we have a collection of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_1 + \boldsymbol{e}_2, \boldsymbol{e}_3$ in $\mathbb{R}^3$, where e_i are the standard basis. This is a spanning collection but linearly dependent. You may throw away any single one of the first three, and you will then get a basis. But you CANNOT throw away $\boldsymbol{e}_3$, since that will destroy the spanning property. ☺

So WHAT to throw away can be tricky. But here comes a nice perspective: we can look at this problem from the opposite angle. What NOT to throw away?

Lemma 4.5.27 (Extending Linear Independency with a Finite Cap)**.** *For any vector space V, if L is a linearly independent collection of vectors, and S is a FINITE spanning system, and $L \subseteq S$, then there is a basis B containing L and contained in S.*

Proof. If L is already spanning, then we are done. Suppose this is not the case.

Let W be the span of L, then $W \neq V$. In particular, S cannot be contained in W. So we can find $\boldsymbol{v} \in S \setminus W$, and $L \cup \{\boldsymbol{v}\}$ would be a linearly independent set between L and S, according to the last lemma.

If $L \cup \{\boldsymbol{v}\}$ is spanning, then this is a basis and we are done. If not, we repeat above process to throw in another $\boldsymbol{w} \in S \setminus span(L \cup \{\boldsymbol{v}\})$, and now $L \cup \{\boldsymbol{v}, \boldsymbol{w}\}$ would be a linearly independent set between L and S. If this is now spanning, then we have a basis. If not, keep doing this.

Since S is finite, either we find a basis somewhere along the process above, or we keep getting larger and larger linearly independent subset of S, until we fill up all of S. But since all sets obtained this way are linearly independent, we would see that S is itself linearly independent. Then S is a basis. □

Remark 4.5.28. *We already know that, for some spanning collections, there might be some redundancy that we can throw away. The above lemma states that we actually have some freedom in choosing what NOT to throw away. You can choose to KEEP a linearly independent subset in the spanning collection, and ONLY throw away vectors among the rest in your spanning collection.*

Btw, setting $L = \varnothing$ implies that any finite spanning collection contains a basis.

As you can see, a linearly independent collection is "may or may not be enough", while a dependent spanning collection is "may or may not be redundant". This leads to the following lemma.

Lemma 4.5.29 (Any Finite Spanning System has no less elements than any Linear Independent Set)**.** *For any vector space V, if L is a linearly independent collection of vectors, and S is a FINITE spanning system, then $|L| \leq |S|$.*

Proof. Suppose $L \subseteq S$, then obviously we are done. So we proceed by induction from here. We shall do induction on how many elements of L is NOT in S. The base step is when $L \subseteq S$, and we have already said that this is trivial.

We assume that we have proven that, if L only contains at most t elements not in S, then we are done. Let us proceed to try to prove it for the case when L only contains $t+1$ elements not in S.

Pick any $\boldsymbol{v} \in L \setminus S$. Then consider $S \cup \{\boldsymbol{v}\}$. Obviously $S \cup \{\boldsymbol{v}\}$ is still spanning, and it contains a linearly independent subset $L \cap (S \cup \{\boldsymbol{v}\})$, so there is a basis B between $L \cap (S \cup \{\boldsymbol{v}\})$ and $S \cup \{\boldsymbol{v}\}$.

Now how many elements does L have that is not in B? You will see that now B contains all of $L \cap S$ and also contains $\boldsymbol{v}$, so L now have at most t elements not in B, rather than $t+1$. And B is spanning. So by induction hypothesis, $|L| \leq |B|$. Now we only need to prove that $|B| \leq |S|$.

Indeed, since S is spanning, $S \cup \{\boldsymbol{v}\}$ is linearly DEPENDENT since $\boldsymbol{v}$ can be written as a linear combination of the rest. So, it is NOT a basis, so B is a proper subset of $S \cup \{\boldsymbol{v}\}$. So $|B| < |S| + 1$, which means that $|B| \leq |S|$.

To sum up, we have inductively shown that, as long as L is finite, $|L| \leq |S|$.

What if L is infinite to begin with? Then pick any subset $L' \subseteq L$ with $|S|+1$ elements. Now L' is finite and linearly independent, so we must have $|L'| \leq |S|$. But that would be a contradiction with our choice of L! So such a case is impossible. So if some spanning collection S is finite, then any linearly independent collection L must also be finite. □

Remark 4.5.30. *So an intuitive proof goes like the following: You pick an element $\boldsymbol{v}$ in L but not in S. Consider $S \cup \{\boldsymbol{v}\}$. (S eat something new.) Now $S \cup \{\boldsymbol{v}\}$ must have redundancies, and you can throw away a redundant vector while KEEPING the linearly independent subset $L \cap (S \cup \{\boldsymbol{v}\})$. (S pooped away the redundancy.) This way we obtain a new set S' that is of the same size as S, but now S' contains one more elements of L than S does. Repeat this, and eventually we can put all of L into some spanning set of the same size as S. So we are done. (S finished eating all of L while pooping away $|L|$ irrelevant redundancies.)*

Or, to put into simple words, how to achieve anything in life? You need to go one step at a time, and keep throwing away redundant things.

There are other ways to prove this lemma. But our proof here is easily generalizable, that it also applied to mathematical areas other than linear algebra.

NOTE that, what if a vector space has NO finite spanning sets at all, that ALL spanning sets are infinite? Then in that case, this lemma proves nothing at all. However, with some twists and transfinite induction (a version of mathematical induction that goes beyond infinity), the idea will carry over.

4.5.4 Infinite dimensional spaces (Optional)

Just some examples to show you what might happen.

Example 4.5.31.

1. Consider the space of all sequences, $\mathbb{R}^{\mathbb{N}}$. Say $(1,1,1,...)$ is a sequence, and $(3,1,4,1,5,9,2,6,...)$ is a sequence, and $(\frac{2}{1}, \frac{3}{2}, \frac{4}{3}, ...)$ is also a sequence, and so on. We can add sequences by adding them term-wise, and we can multiply a real number to a sequence by mutliply the number term-wise. This makes $\mathbb{R}^{\mathbb{N}}$ a vector space. This vector space is INFINITE DIMENSIONAL. You can try to come up with a basis, and you can try $(1,0,0,...), (0,1,0,...)$ and so forth. Surely, it seems like any sequence is like some infinite linear combination of them, no? But are they really basis? Not really. Infinite dimensional spaces works differently.

2. Why are they not really a basis? First let us note that the set S_c of all converging sequences is a subspace in $\mathbb{R}^{\mathbb{N}}$. For example, sequence $(\frac{2}{1}, \frac{3}{2}, \frac{4}{3}, ...)$ is converging with limit 1. This is easy to see, because if $(a_n), (b_n)$ are converging to values a and b, then $(a_n + b_n)$ would converge to $a+b$, and $k(a_n)$ would converge to ka. So this is indeed a subspace. This is also NOT the whole space, because

there are non-convergin sequences out there. Now you see that the vectors $(1,0,0,...),(0,1,0,...)$ are all contained in S_c. So they span AT MOST S_c. Their span, by definition of a subspace, will always stay in S_c. Since there are divergent sequences like $(1,2,3,...)$ in $\mathbb{R}^{\mathbb{N}} - S_c$, there is simply no way this sequence would be in the span of vectors contained entirely in S_c. In fact, $(1,0,0,...),(0,1,0,...)$ does NOT even span S_c, because they all have limit zero, so all of their linear combinations could only have limit zero.

3. Suffice to say, there is no such thing as an "infinite" linear combination. Whenever we do a linear combination, we are always only allowed to use finitely many vectors, even if we have an infinite collection of vectors at our hands. Otherwise we may run into paradoxical phenomina as explained above.

4. Consider the set of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, denoted sometimes as $C(\mathbb{R})$. For two functions $f : x \mapsto f(x)$ and $g : x \mapsto g(x)$, we can define their sum as $f + g : x \mapsto f(x) + g(x)$. We can also multiply a function with a real number as $kf : x \mapsto k(f(x))$. This makes $C(\mathbb{R})$ a real vector space. This is NOT finite dimensional.

5. Consider the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}$ that is periodic with period 2π. Then the Fourier theorem says the following: any such function f must be the sum of some series $f(x) = c + \sum_n a_n \sin(nx) + \sum_n b_i \cos(nx)$. The sums go through all possible positive integers. This statement has a very "basis"-feel to it. Namely, f seems like a linear combination of the constant function and the sines and cosines. Furthermore, you might in fact do "dot products" for these functions, and define $f \cdot g = \int_0^{2\pi} f(x)g(x)dx$. Then you will see that the constant function, sines and cosines are mutually orthogonal! This is not just a "basis", but in fact an orthogonal "basis". This is the start of Fourier analysis. It is NOT a true basis in the traditional sense. Not every function is a linear combination of them. However, every function is a LIMIT of linear combinations of them. (We do not always have such nice things for infinite dimensional basis, btw.)

☺

4.6 Entries of a Linear Map and Changing Basis

4.6.1 Entries of a Linear Map

Matrices are linear maps between $\mathbb{R}^n$ and $\mathbb{R}^m$. But what about abstract linear maps on abstract vector spaces?

If we have a linear map $L : V \to W$, then there is no matrix right away. However, recall that picking a basis would essentially turn the vector space into $\mathbb{R}^n$. So, if we pick a basis for V and pick a basis for W, then V is now effectively $\mathbb{R}^n$ and W is now effectively $\mathbb{R}^m$. Then L is now some map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Hence it is now a matrix.

Keep in mind that any matrix expression of L will depend on our choice of basis for the domain and codomain. In particular, there is no THE matrix for L, since we may pick different basis, and then L may appear to be different matrices. Just as a vector may have different coordiantes under different basis, a linear map may have different matrices under different basis.

Definition 4.6.1. *Given a basis $\mathcal{B}$ for a space V and a basis $\mathcal{C}$ for a space W, the matrix for a linear map $L : V \to W$ is $L_{\mathcal{C}\leftarrow\mathcal{B}} = \mathcal{C}^{-1}L\mathcal{B}$.*

So if the inputs are coordinates under $\mathcal{B}$, we first combine coordinates into the actual vector via the linear combination map $\mathcal{B}$, then we map through L, and finally we express the abstract image in terms of coordinates via the coordinate map $\mathcal{C}^{-1}$. As you can see, this is basically just L, expressed as a matrix using the basis of the domain and the basis of the codomain.

To summarize the idea above, think about the following diagram:

$$\begin{array}{ccc} V & \xrightarrow{L} & W \\ {\scriptstyle\mathcal{B}}\uparrow & & \uparrow{\scriptstyle\mathcal{C}} \\ \mathbb{R}^n & \xrightarrow{L_{\mathcal{C}\leftarrow\mathcal{B}}} & \mathbb{R}^m \end{array}$$

Consider this very straight forward proposition.

Proposition 4.6.2. *Given a linear map $L : V \to W$, basis $\mathcal{B}$ for V and basis $\mathcal{C}$ for W, then for any $\boldsymbol{v} \in V$, we have $L_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} = (L\boldsymbol{v})_{\mathcal{C}}$.*

Proof. This should be intuitively trivial. For a rigorous proof, we have

$$L_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} = \mathcal{C}^{-1}L\mathcal{B}\mathcal{B}^{-1}\boldsymbol{v} = \mathcal{C}^{-1}L\boldsymbol{v} = (L\boldsymbol{v})_{\mathcal{C}}.$$

□

We also have the following nice result.

Proposition 4.6.3. *Given a linear map $L : V \to W$, and a linear map $T : W \to U$, say we have basis $\mathcal{B}$ for V, basis $\mathcal{C}$ for W, and basis $\mathcal{D}$ for U. Then we have $(TL)_{\mathcal{D}\leftarrow\mathcal{B}} = T_{\mathcal{D}\leftarrow\mathcal{C}}L_{\mathcal{C}\leftarrow\mathcal{B}}$.*

Proof. This should be intuitively trivial. For a rigorous proof, we have

$$T_{\mathcal{D}\leftarrow\mathcal{C}}L_{\mathcal{C}\leftarrow\mathcal{B}} = \mathcal{D}^{-1}T\mathcal{C}\mathcal{C}^{-1}L\mathcal{B} = \mathcal{D}^{-1}TL\mathcal{B} = (TL)_{\mathcal{D}\leftarrow\mathcal{B}}.$$

□

Now, to find the matrix $L_{\mathcal{C}\leftarrow\mathcal{B}}$ of a linear map L, we do not really use the formula $\mathcal{C}^{-1}L\mathcal{B}$. Rather, we go back to the basics in this class: we compute the matrix $L_{\mathcal{C}\leftarrow\mathcal{B}}$ column by column, by inspecting $L_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{e}_i$ for each i.

Example 4.6.4. Suppose we have a linear map $M : \mathcal{P}_2 \to \mathcal{P}_3$, that multiply each polynomial by x. How can we understand this linear map?

One way is the following: we pick a basis in $\mathcal{P}_2$ and in $\mathcal{P}_3$. Then the two vector spaces are now looking like $\mathbb{R}^3$ and $\mathbb{R}^4$. Then we see that our linear map corresponds to a 4 by 3 matrix.

To find this matrix, we again ONLY inspect the image of $\boldsymbol{e}_i$ to get its columns. We always do this to find whatever matrix.

Suppose we pick basis $1, x, x^2 \in \mathcal{P}_2$ and $1, x, x^2, x^3 \in \mathcal{P}_3$, then $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ in $\mathcal{P}_2$ are in fact $1, x, x^2$. They would go to x, x^2, x^3, which are $\boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4$ in the codomain coordinates. So the corresponding matrix for M is $\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Suppose we pick basis $2, x+1, x^2 \in P_2$ and $x+1, x-1, x^2-x, x^3+x^2$. Compute the image of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. They are in fact $2, x+1, x^2$ in the domain. So they would go to $2x, x^2+x, x^3$ in the codomain. Note that $2x = (x+1)+(x-1)$, and $x^2+x = (x^2-x)+(2x) = (x^2-x)+(x+1)+(x-1)$. Finally,

$$x^3 = (x^3+x^2)-x^2 = (x^3+x^2)-(x^2-x)-x = (x^3+x^2)-(x^2-x)-\frac{1}{2}(x+1)-\frac{1}{2}(x-1).$$

So these corresponds to vectors in the codomain with coordinates $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ -1 \\ 1 \end{bmatrix}$. Then the matrix for our linear map is now $\begin{bmatrix} 1 & 1 & -\frac{1}{2} \\ 1 & 1 & -\frac{1}{2} \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$.

Either way, since we have the matrix, we can now compute whatever we need about M. For example, if the input is $2 + (x + 1) + (x^2)$, then the output coordinates should be $\begin{bmatrix} 1 & 1 & -\frac{1}{2} \\ 1 & 1 & -\frac{1}{2} \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \\ 0 \\ 1 \end{bmatrix}$, which represents the vector $\frac{3}{2}(x+1) + \frac{3}{2}(x-1) + (x^3 + x^2)$. ☺

Example 4.6.5. Suppose we have a linear map $D : \mathcal{P}_3 \to \mathcal{P}_3$, then under the obvious easy basis $1, x, x^2, x^3$ for both the domain and the codomain, what is the matrix of D?

Well, $D(1) = 0, D(x) = 1, D(x^2) = 2x, D(x^3) = 3x^2$. So the matrix for D is $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

As you can imagine, if we simply pick the space $\mathcal{P}$ of all polynomials, then under the "obvious basis" $1, x, x^2, \ldots$, then the matrix for the derivative map $D : \mathcal{P} \to \mathcal{P}$ is an infinite matrix like $\begin{bmatrix} 0 & 1 & & \\ & 0 & 2 & \\ & & 0 & \ddots \\ & & & \ddots \end{bmatrix}$. And at the same time, the map "multiplication by x" $M : \mathcal{P} \to \mathcal{P}$ is also an infinite matrix like $\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \end{bmatrix}$.

You can directly compute and see that $DM - MD = I$. Here DM is D shifted to the left, and MD is D shifted down. ☺

In real life, we humans would simply pick basis and operates at the abstract level. Then computers would go find the matrices and coordinates and calculate for us.

4.6.2 Change of Basis

The point of an abstract vector space is to change basis, so let us see how it can be done computationally.

Example 4.6.6. Again take our example of $\mathcal{P}_2$ with $\boldsymbol{w} = 3x^2 + 4x + 1$. It is $\begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix}$ under the basis $\mathcal{B} = (1, x + 1, (x + 1)^2)$.

Now similarly, for the basis $\mathcal{C} = (1, x - 1, (x - 1)^2)$, we can compute and have $\boldsymbol{w} = \mathcal{C} \begin{bmatrix} 8 \\ 10 \\ 3 \end{bmatrix}$. How can one convert between these two basis?

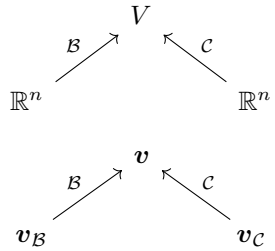
In general, suppose $\boldsymbol{v}$ in basis $\mathcal{B}$ has coordinates $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$, then what is its coordinates in basis $\mathcal{C}$? This means given $\boldsymbol{v} = \sum a_i \boldsymbol{v}_i$, how can I express $\boldsymbol{v}$ in another basis $\boldsymbol{w}_i$?

To achieve this transition, we need to know what the basis vectors $\boldsymbol{v}_i$ looks like in basis vectors $\boldsymbol{w}_i$, and substitute. Note that In our case, $1 = 1, x + 1 = 2 + (x - 1), (x + 1)^2 = 4 + 4(x - 1) + (x - 1)^2$. So if $\boldsymbol{v} = a + b(x + 1) + c(x + 1)^2$, then it will be $\boldsymbol{v} = a + 2b + b(x - 1) + 4c + 4c(x - 1) + c(x - 1)^2$ and it has new coordinates $\begin{bmatrix} a + 2b + 4c \\ b + 4c \\ c \end{bmatrix}$.

What is the relation between the two? We see that the new coordinates are in fact $\begin{bmatrix} a+2b+4c \\ b+4c \\ c \end{bmatrix} = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, a matrix times the old coordinates! Why is that? ☺

Such a matrix is called a change of coordinates matrix. But why is this a matrix?

Given a vector space V and two bases $\mathcal{B}, \mathcal{C}$, consider the linear combination maps $\mathcal{B}, \mathcal{C}$. These two maps are bijections. So we have the following graph:



So we can clearly see that the change of coordinate process is $\boldsymbol{v}_{\mathcal{C}} = \mathcal{C}^{-1}\mathcal{B}\boldsymbol{v}_{\mathcal{B}}$. Since the linear combination maps $\mathcal{B}, \mathcal{C}$ are linear bijections, we see that the change of coordinate process is $\mathcal{C}^{-1}\mathcal{B}$, which is a linear map. Furthermore, the inputs and outputs are both elements of $\mathbb{R}^n$, so it is an $n \times n$ matrix.

Let us take another approach.

Corollary 4.6.7. *Given a vector space V and bases $\mathcal{B}, \mathcal{C}$, then $I_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} = \boldsymbol{v}_{\mathcal{C}}$. (So the change of coordinate matrix is actually the matrix for the identity map.)*

Proof. $I_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} = (I\boldsymbol{v})_{\mathcal{C}} = \boldsymbol{v}_{\mathcal{C}}$.

For an alternative proof, recall that the change of coordinate matrix should be $\mathcal{C}^{-1}\mathcal{B}$. But we have $I_{\mathcal{C}\leftarrow\mathcal{B}} = \mathcal{C}^{-1}I\mathcal{B} = \mathcal{C}^{-1}\mathcal{B}$. □

This should not surprise us at all. After all, the change of basis process do NOT really change the under lying vector. It IS the identity map. But because the input vector was expressed in the old basis, and the output vector is the same vector but under the new basis, therefore the change is precisely $I_{\mathcal{B}\to\mathcal{C}}$.

Definition 4.6.8. *Given two basis $\mathcal{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ and $\mathcal{C} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ of a vector space V, then we define the **change of coordinate matrix** as a matrix $I_{\mathcal{C}\leftarrow\mathcal{B}}$, i.e., the matrix for the identity map from basis $\mathcal{B}$ to $\mathcal{C}$.*

Now, the fact that $I_{\mathcal{C}\leftarrow\mathcal{B}}$ is a matrix is MORE IMPORTANT that the fact that it is $\mathcal{C}^{-1}\mathcal{B}$ or any other expression. Because we have a very straight forward way to find a matrix: its i-th column is simple the image of $\boldsymbol{e}_i$!

So if $\mathcal{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n), \mathcal{C} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$, then the i-th column of $I_{\mathcal{C}\leftarrow\mathcal{B}}$ is $\mathcal{C}^{-1}\mathcal{B}\boldsymbol{e}_i = \mathcal{C}^{-1}\boldsymbol{v}_i$, i.e., the coordinates of $\boldsymbol{v}_i$ in the basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$.

Remark 4.6.9 (Algorithm to find change of coordinate matrix)**.** *Express each old basis vector in terms new basis vectors, and that is the corresponding column for your change of coordinate matrix. Review the example at the start of this subsection to get a better feel of this.*

Alternatively, you can also express new basis vectors in terms of old basis vectors (sometimes the problem tells you how to do this explicitly). Then you will obtain the basis transition matrix, and the inverse matrix is the change of coordinate matrix.

Example 4.6.10. Again take our example of $\mathcal{P}_2$ with basis $\mathcal{B} = (1, x+1, (x+1)^2)$ and basis $\mathcal{C} = (1, x-1, (x-1)^2)$. How can one convert coordinates in $\mathcal{B}$ into coordinates in $\mathcal{C}$?

The columns of the change of coordinate matrix should be the vectors of the old basis expressed in coodinates in the new basis. Observe that

$$\begin{array}{rcl} 1 & = & 1 \\ x+1 & = & 2+(x-1) \\ (x+1)^2 & = & 4+4(x-1)+(x-1)^2 \end{array}.$$

So the change of coordinate matrix is $I_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$, where we read the three columns from the coordinates above. ☺

Now, sometimes there is another way to do this. By taking a closer look you might see that $(1, x+1, (x+1)^2) = (1, x-1, (x-1)^2)\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$. (Just interpret the multiplication here as a "block multiplication" type of thing.) Such a matrix is called a basis transition matrix.

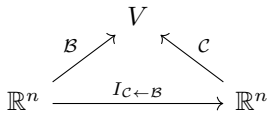
Note how the very same matrix appears both as a basis transition matrix AND a change of coordinate matrix, but in OPPOSITE directions. The matrix $\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$ is the change of coordinates matrix from the OLD coordinates to the NEW coordinates (multiplied from the left to a coordinate vector), while it is also the basis transition matrix from the NEW basis to the OLD basis (multiplied from the right to a linear combination map of the basis).

Definition 4.6.11. *Given a collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ in V and an $m \times n$ matrix $A = \begin{bmatrix} \boldsymbol{a}_1 & \ldots & \boldsymbol{a}_n \end{bmatrix}$ (so the entries of A are all real numbers rather than abstract vectors), we define $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)A = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ where $\boldsymbol{w}_i = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)\boldsymbol{a}_i$. So we simply do linear combination of these $\boldsymbol{v}_i$ according to each column of A, and get n vectors out of it.*

Definition 4.6.12. *Given two basis $\mathcal{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ and $\mathcal{C} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ of a vector space V, then we define the **basis transition matrix** as a matrix $T_{\mathcal{B}\to\mathcal{C}}$ such that $\mathcal{B}T_{\mathcal{B}\to\mathcal{C}} = \mathcal{C}$.*

Proposition 4.6.13. *If $\mathcal{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ and $\mathcal{C} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ are two bases of the same vector space V, then the $T_{\mathcal{C}\to\mathcal{B}} = I_{\mathcal{C}\leftarrow\mathcal{B}} = \mathcal{C}^{-1}\mathcal{B}$, and it is an invertible $n \times n$ matrix when $\dim V = n$.*

Proof. Look at the following diagram.



Note that a vector $\boldsymbol{v}$ lives in the V in the top, and its coordinates in $\mathcal{B}$ lives in the bottom left, and its coordinates in $\mathcal{C}$ lives in the bottom right. it is obvious that do perform a change of coordinates $I_{\mathcal{C}\leftarrow\mathcal{B}}$, it is the same as applying the linear map $\mathcal{C}^{-1}\mathcal{B}$. note that this map goes from $\mathbb{R}^n$ to $\mathbb{R}^n$, so it corresponds to an $n \times n$ matrix.

Now consider the basis transition map from $\mathcal{C}$ to $\mathcal{B}$, $T_{\mathcal{C}\to\mathcal{B}}$. It is clear that we want $\mathcal{B} = \mathcal{C}T_{\mathcal{C}\to\mathcal{B}}$. So we immediately see that $T_{\mathcal{C}\to\mathcal{B}} = \mathcal{C}^{-1}\mathcal{B}$. □

Corollary 4.6.14. *The change of coordinates matrix from $\mathcal{B}$ to $\mathcal{C}$ and the basis transition matrix from $\mathcal{B}$ to $\mathcal{C}$ are inverse of each other.*

Remark 4.6.15. *You should also note that the strange situation of $T_{\mathcal{C}\to\mathcal{B}} = I_{\mathcal{C}\leftarrow\mathcal{B}}$ is a result of associativity.*

Given $\boldsymbol{v} \in V$, we have $\mathcal{B}\boldsymbol{v}_{\mathcal{B}} = \boldsymbol{v} = \mathcal{C}\boldsymbol{v}_{\mathcal{C}}$ by definition. But the left hand side is also $(\mathcal{C}T_{\mathcal{C}\to\mathcal{B}})\boldsymbol{v}_{\mathcal{B}}$, whereas the right hand side is also $\mathcal{C}(I_{\mathcal{C}\leftarrow\mathcal{B}}\boldsymbol{v}_{\mathcal{B}})$.

Example 4.6.16. Suppose I want to buy x burgers and y cokes. Now combo A sells 2 burger and 1 coke, while combo B sells 3 burger and 4 cokes. How many combos do I need?

Well, note that $(comboA, comboB) = (burger, cokes)\begin{bmatrix}2 & 3\\ 1 & 4\end{bmatrix}$. So $\begin{bmatrix}2 & 3\\ 1 & 4\end{bmatrix}$ is the basis transition matrix from single orders to combos. Therefore, the change of coordinate map from single orders to combos would be $\begin{bmatrix}2 & 3\\ 1 & 4\end{bmatrix}^{-1} = \frac{1}{5}\begin{bmatrix}4 & -3\\ -1 & 2\end{bmatrix}$. So I need $\frac{4}{5}x - \frac{3}{5}y$ combo A and $-\frac{1}{5}x + \frac{2}{5}y$ combo B.

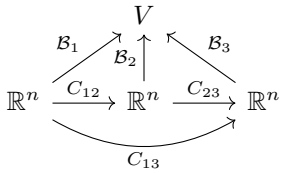
Note that a matrix inversion process is involved. Typically, a problem will look like this: we start with an old basis and old coordinates. The problem tells you how the old basis would combine into the new basis. Then the problem asks how the old coordinates would change into the new coordinates.

In effect, they give you $T_{\mathcal{B}\to\mathcal{C}}$ for free, and ask you for $I_{\mathcal{C}\leftarrow\mathcal{B}}$. These two are just inverse matrix of each other. ☺

Drawing diagrams is a very practical way to help us figure things out. Consider this:

Proposition 4.6.17. *Given three basis $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$, let C_{ij}, T_{ij} be the change of coordinates map and the transition map from $\mathcal{B}_i$ to $\mathcal{B}_j$. Then $C_{13} = C_{23}C_{12}$ (note that this will multiply things from their left, so C_{12} happens first) and $T_{13} = T_{12}T_{23}$ (note that this will multiply things from their right, so T_{12} happens first).*

Proof. Conceptually this is obvious. But we can also stare at the following diagram until we see this.



Also keep in mind that $C_{ij} = T_{ji}$.

Another alternative is to compute it directly, say $C_{23}C_{12} = \mathcal{B}_3^{-1}\mathcal{B}_2\mathcal{B}_2^{-1}\mathcal{B}_1 = \mathcal{B}_3^{-1}\mathcal{B}_1 = C_{13}$. □

Example 4.6.18 (Hen and rabbit again?)**.** The matrix $\begin{bmatrix}1 & 1\\ 2 & 4\end{bmatrix}$ is invertible. Is this a change of basis? YES!

The content of the cage (the underlying abstract entity) never changes. However, I should think of the content of the cage as a linear combination of ... what? This is precisely what the hen-rabbit problem is about.

If we describe the content of the cage using the "basis" of (head, leg), I get coordinates $\begin{bmatrix}6\\ 20\end{bmatrix}$. If we describe the content of the cage using the "basis" of (hen, rabbit), what coordinate should I get?

We have a basis transition (hen, rabbit) = (head, leg) $\begin{bmatrix}1 & 1\\ 2 & 4\end{bmatrix}$. So the answer is $\begin{bmatrix}1 & 1\\ 2 & 4\end{bmatrix}^{-1}\begin{bmatrix}6\\ 20\end{bmatrix}$. ☺

Let us take a moment and think about a change of basis in $\mathbb{R}^n$ itself. We of course always start with the standard basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$. Suppose we want to change to a new basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^n$. How would I do that?

Proposition 4.6.19. *The change of coordinate matrix to go from the standard basis in $\mathbb{R}^n$ to a new basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is $\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}^{-1}$.*

Proof. Note that for vectors in $\mathbb{R}^n$, $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ has exactly the same meaning as $\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$. So we have $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix} = I\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$. So $\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$ is the basis transition matrix from the standard basis to the new basis. □

Corollary 4.6.20. *An invertible matrix A can be thought of as a change of coordinate matrix in $\mathbb{R}^n$, to go from the standard basis to a new basis, i.e., the columns of A^{-1}.*

Here come the mind blasting moment. ALL invertible matrices might simply be seen as some change of basis process!

Proposition 4.6.21. *Say we started with an n-dimensional vector space V and a basis $\mathcal{B}$. Then for any $n \times n$ invertible matrix A, it is $C_{\mathcal{B}\to\mathcal{C}}$ for some new basis $\mathcal{C}$.*

Proof. Simply set $\mathcal{C} = \mathcal{B}A^{-1}$, and we are done. □

Remark 4.6.22. *Whenever you see an invertible matrix, then under some perspective, it will represent a change of basis.*

So an invertilbe matrix is a change of basis, but it is also a bijective linear map? How can the two be the same? Say you have a pen in your hand. You can rotate the pen (apply the linear map), or you can tilt your head (change of basis), and the pen would result in the same posture as far as your eyes can see. The two process are the same invertible matrix.

4.6.3 Change of basis for a Linear Map

Here is a thought: given a matrix for our linear map under some bases, can I convert it using the change of basis formula, to get the matrix in another basis?

Proposition 4.6.23. *Given bases $\mathcal{B}_1, \mathcal{B}_2$ for a space V and a basis $\mathcal{C}$ for a space W, then for a linear map $L : V \to W$ we have $L_{\mathcal{C}\leftarrow\mathcal{B}_2} = L_{\mathcal{C}\leftarrow\mathcal{B}_1} I_{\mathcal{B}_1\leftarrow\mathcal{B}_2}$.*

Similarly, given a basis $\mathcal{B}$ for a space V and bases $\mathcal{C}_1, \mathcal{C}_2$ for a space W, then for a linear map $L : V \to W$ we have $L_{\mathcal{C}_2\leftarrow\mathcal{B}} = I_{\mathcal{C}_2\leftarrow\mathcal{C}_1} L_{\mathcal{C}_1\leftarrow\mathcal{B}}$.

Proof. The proofs are trivial. For example, we have $L_{\mathcal{C}\leftarrow\mathcal{B}_1} I_{\mathcal{B}_1\leftarrow\mathcal{B}_2} = (LI)_{\mathcal{C}\leftarrow\mathcal{B}_2} = L_{\mathcal{C}\leftarrow\mathcal{B}_2}$.

Alternatively, look at the diagram below.

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{L_{\mathcal{C}_1\leftarrow\mathcal{B}_1}} & \mathbb{R}^m \\ \uparrow{\scriptstyle \mathcal{B}_1} & & \mathcal{C}_1\downarrow \\ V & \xrightarrow{L} & W \\ \uparrow{\scriptstyle \mathcal{B}_2} & & \mathcal{C}_2\uparrow \\ \mathbb{R}^n & \xrightarrow[L_{\mathcal{C}_2\leftarrow\mathcal{B}_2}]{} & \mathbb{R}^m \end{array}$$

(with $I_{\mathcal{B}_1\leftarrow\mathcal{B}_2}$ on the left and $I_{\mathcal{C}_1\leftarrow\mathcal{C}_2}$ on the right)

You can immediately see that for $L_{\mathcal{C}_1\leftarrow\mathcal{B}_2}$, for example, we need to go from bottom left to upper right. So $L_{\mathcal{C}_1\leftarrow\mathcal{B}_2} = L_{\mathcal{C}_1\leftarrow\mathcal{B}_1} I_{\mathcal{B}_1\leftarrow\mathcal{B}_2}$. The other one goes from upper left to bottom right, and can be proven similarly. □

Remark 4.6.24. *In particular, change of basis in the domain corresponds to multiplying an invertible matrix on the RIGHT, while change of basis in the codomain corresponds to multiplying an invertible matrix on the LEFT.*

Example 4.6.25. Note that

$$(1, x, x^2)\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = (2, x+1, x^2),$$

$$(1, x, x^2, x^3)\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = (x+1, x-1, x^2-x, x^3+x^2).$$

As a result, using the corresponding change of coordinate matrix, the matrix of our previous linear map under the ugly basis can be computed as

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -\frac{1}{2} \\ 1 & 1 & -\frac{1}{2} \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hooray, this works! ☺

In fact, the process of "diagram chase" would allow us to create many "change of basis" formula for matrices of linear maps. Let us look at this diagram:

$$\begin{array}{ccccc} \mathbb{R}^n & \xrightarrow{L_{\mathcal{C}_1 \leftarrow \mathcal{B}_1}} & \mathbb{R}^d & \xrightarrow{T_{\mathcal{D}_1 \leftarrow \mathcal{C}_1}} & \mathbb{R}^m \\ \mathcal{B}_1 \downarrow & & \downarrow \mathcal{C}_1 & & \downarrow \mathcal{D}_1 \\ U & \xrightarrow{L} & V & \xrightarrow{T} & W \\ \mathcal{B}_2 \uparrow & & \uparrow \mathcal{C}_2 & & \uparrow \mathcal{D}_2 \\ \mathbb{R}^n & \xrightarrow{L_{\mathcal{C}_2 \leftarrow \mathcal{B}_2}} & \mathbb{R}^d & \xrightarrow{T_{\mathcal{D}_2 \leftarrow \mathcal{C}_2}} & \mathbb{R}^m \end{array}$$

By snaking around the diagram, you can get results like

$$(T \circ L)_{\mathcal{D}_1 \leftarrow \mathcal{B}_2} = I_{\mathcal{D}_1 \leftarrow \mathcal{D}_2} T_{\mathcal{D}_2 \leftarrow \mathcal{C}_2} I_{\mathcal{C}_2 \leftarrow \mathcal{C}_1} L_{\mathcal{C}_1 \leftarrow \mathcal{B}_1} I_{\mathcal{B}_1 \leftarrow \mathcal{B}_2}.$$

I'm not even going to write this as a proposition or to prove it. Just stare at these diagrams and you should be able to see this.

4.6.4 Row and Column operations and the Rank Normal Form

Now let us not forget the point of doing change of basis. The point is to gain more interpretations and to understand linear algebra better.

Suppose we have a matrix $A : \mathbb{R}^n \to \mathbb{R}^m$. Now maybe A looks ugly. If I perform some change of basis, hopefully A will look pretty, and we can solve problems involving A much better, yes? So how to do this? And how simple can we simplify A?

Now, change of basis in $\mathbb{R}^n$ are simply multiplication by invertible matrices. So we can reformulate our problem into this:

Question 4.6.26. *Given an $m \times n$ matrix A, can you find invertible $m \times m$ matrix R and invertible $n \times n$ matrix C such that BAC is nice?*

Now, an invertible matrix on the left is simply a series of row operations, while an invertible matrix on the right is simply a series of column operations! In fact, combining our intuitions on row/column operations and change of basis, we have the following:

1. A row operation is a change of basis in the codomain.
2. A column operation is a change of basis in the domain.
3. RREF means doing a change of basis to the codomain, so that your system looks as simple as possible.

So how would we simplify a matrix A via change of basis? Say $A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}$. First we row reduce

to get RREF $\begin{bmatrix}1&0&0&0\\0&1&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}$. Next we column reduce, using pivots to kill the rest of each row. We now have $\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&0&0\end{bmatrix}$. Finally, let us throw zero columns to the right, and we have $\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}$. Huh.

Proposition 4.6.27. *Given any $m \times n$ matrix A with rank r, we can find invertible $m \times m$ matrix R and invertible $n \times n$ matrix C such that $RAC = \begin{bmatrix}I_{r\times r} & O\\ O & O\end{bmatrix}$. We say this is the **rank normal form** or **rank revealing form** for our matrix A.*

Proof. What we just did. Starting from A, we first change basis in codomain, i.e., row reduce A, until we reach RREF. Then change basis in domain, i.e., column reduce by using pivots to kill the rest of each row. Now only pivots are left. And then they are column-swapped to stay in the upper left block. □

Note that we write $\begin{bmatrix}I_{r\times r} & O\\ O & O\end{bmatrix}$ for simplicity, but technically it could be $\begin{bmatrix}I & O\end{bmatrix}$ or $\begin{bmatrix}I\\O\end{bmatrix}$ or I, depending on whether we have zero columns or zero rows at all.

Corollary 4.6.28. *Given any linear map, by finding the right basis for the domain and for the codomain, the map will have matrix $\begin{bmatrix}I & O\\ O & O\end{bmatrix}$ or $\begin{bmatrix}I & O\end{bmatrix}$ or $\begin{bmatrix}I\\O\end{bmatrix}$ or I. We say this is the **rank normal form** for your linear map L.*

Think about the meaning of these statements. We realized that ANY linear map, by choosing the correct basis, will look like the matrix $\begin{bmatrix}I_{r\times r} & O\\ O & O\end{bmatrix}$ or $\begin{bmatrix}I & O\end{bmatrix}$ or $\begin{bmatrix}I\\O\end{bmatrix}$ or I.

So it turns out that, as far as finite dimensions goes, if we always pick good bases, then the abstract world is extremely beautiful and simplistic. We have ONLY FOUR kinds of linear maps:

1. Each finite dimensional vector space looks like $\mathbb{R}^n$ for some n.
2. A linear injection must look like $\begin{bmatrix}I\\O\end{bmatrix}$ under the right basis for the domain and the codomain. Geometrically, this is an inclusion map. For example $\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}$ would simply send $\begin{bmatrix}x\\y\end{bmatrix}$ to $\begin{bmatrix}x\\y\\0\end{bmatrix}$. It represent the process of including the xy-plane into $\mathbb{R}^3$.
3. A linear surjection must look like $\begin{bmatrix}I & O\end{bmatrix}$ under the right basis for the domain and the codomain.Geometrically, this is a projection map. For example $\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}$ would simply send $\begin{bmatrix}x\\y\\z\end{bmatrix}$ to $\begin{bmatrix}x\\y\end{bmatrix}$. It represent the process of projecting $\mathbb{R}^3$ onto the xy-plane.
4. A linear bijection must look like the identity matrix under the right basis for the domain and the right basis for the codomain.
5. A linear map that is neither injective nor surjective looks like $\begin{bmatrix}I & O\\ O & O\end{bmatrix}$ under the right basis for the domain and the codomain.

Furthermore, here are some interesting ramifications.

Theorem 4.6.29 (Full rank decomposition)**.** *For any $m \times n$ matrix A with rank r, we can find an $m \times r$ matrix B with rank r (so it is injective), and an $r \times n$ matrix C with rank r (so it is surjective), such that $A = BC$.*

Proof. Pick the right basis, then $A = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$.

However, we have

$$\begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix} = \begin{bmatrix} I_{r\times r} \\ O \end{bmatrix} \begin{bmatrix} I_{r\times r} & O \end{bmatrix}.$$

So set $B = \begin{bmatrix} I_{r\times r} \\ O \end{bmatrix}$ and $C = \begin{bmatrix} I_{r\times r} & O \end{bmatrix}$, and we are done. □

Remark 4.6.30. *This is part of a larger phenomenon. For any map f between sets, then we can always find an injective map g and a surjective map h such that $f = g \circ h$.*

What does this mean geometrically? It means that ALL linear maps A behave like this:

1. First, we do a surjective linear map C, i.e., we collapse a few dimensions of the domain, until only r dimensions left.
2. Then, we do an injective linear map B, i.e., we use an inclusion map to put the leftover from the last step into a larger space, the codomain.

Computationally, if all basis we choose are nice, then ALL linear maps $A : \mathbb{R}^n \to \mathbb{R}^m$ of rank r behave like this:

1. Given an input $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, we drop a few coordinates and get $\begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix}$.
2. Then, we add a few zero coordinates, and get $\begin{bmatrix} x_1 \\ \vdots \\ x_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^m$.

So, all linear maps are simply "collapse a few dimensions, and then put into a larger space." From this intuition, the following corollary is completely trivial.

Corollary 4.6.31. *The rank of a matrix A is the same as* $\dim(\mathrm{Ran}(A))$.

Proof. We don't even need the full rank decomposition. Say A has rank r. Simply pick the right basis, then $A = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$. Now look, its range (column space) is spanned by $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r$ and therefore has dimension r.

Obviously changing basis could only change the names of vectors, and could not change the dimension of $\dim(\mathrm{Ran}(A))$. □

Our previous proofs show that, given a matrix A with rank r, we can find a rank normal form also with rank r. But is it possible to find some rank normal form of A with a different rank? This is NOT possible. In particular, the rank normal form is unique, as shown below.

Lemma 4.6.32. *For any $m \times n$ matrix A, pick any invertible $m \times m$ matrix R and invertible $n \times n$ matrix C, then A and RAC must have the same rank.*

Proof. No change of basis could possibly change $\dim(\mathrm{Ran}(A))$ in any way. □

Corollary 4.6.33. *The rank normal form of a matrix is unique. And two matrices A, B have the same rank if and only if we can find invertible $m \times m$ matrix R and invertible $n \times n$ matrix C, such that $B = RAC$.*

In particular, the rank of a matrix COMPLETELY characterizes the linear map. If two matrices are both $m \times n$ and have the same rank, then they are the SAME linear map, differ only by a change of basis in the domain and a change of basis in the codomain.

We end this section by a useful decompositon in practice.

Proposition 4.6.34 (Rank-one decomposition)**.** *A matrix of rank r is the sum of r rank-one matrices.*

Proof. Given a matrix A with rank r, we first do the full rank decomposition $A = BC$. Note that B has r columns, and C has r rows. Say $B = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_r \end{bmatrix}$ and $C = \begin{bmatrix} \boldsymbol{c}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{c}_r \end{bmatrix}$. Then we have

$$A = BC = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{c}_r \end{bmatrix} = \sum_{i=1}^{r} \boldsymbol{b}_i \boldsymbol{c}_i^{\mathrm{T}}.$$

Each $\boldsymbol{b}_i \boldsymbol{c}_i^{\mathrm{T}}$ is clearly a rank-one matrix. So we are done. □

As a side note, let us briefly talk about the rank of an abstract linear map. This is mostly trivial, but we establish to be technically more rigorous.

Definition 4.6.35. *Given a linear map L, we define its rank* $\mathrm{rank}(L)$ *as* $\dim \mathrm{Ran}(L)$.

Recall that $\mathrm{Ran}(L)$ is the collection of all images of L, and they form a natural subspace of the codomain. Its dimension is what we call the rank of L.

Proposition 4.6.36. *The rank of A as a matrix (i.e., number of pivots) is the same as the rank of A as a linear map.*

Proof. Notice that A and $rref(A)$ can be thought of as the same linear map, differ only by a change of basis in the codomain. So their range have the same dimension. Now the dimension of $rref(A)$ is obviously the number of pivots. □

Example 4.6.37. Let A be a matrix such that each column is an arithmetic sequence. Then each column is of the form $a \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + d \begin{bmatrix} 0 \\ \vdots \\ n-1 \end{bmatrix}$. So the range of A is contained in the span of $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ n-1 \end{bmatrix}$. So it is at most 2 dimensional.

So A has rank at most two. ☺

Let us review the full rank decomposition process. If $L : V \to W$, how can I decompose this into a surjective linear map followed by an injective linear map?

Well, first of all, we can restricte the codomain to $\mathrm{Ran}(L)$. Then the map $L' : V \to \mathrm{Ran}(L)$ would be surjective, where L' is essentially the same as L, and they only differ in the choice of codomain.

Next, we have an inclusion map $\iota : \mathrm{Ran}(L) \to W$, since $\mathrm{Ran}(L)$ must be a subspace of W.

Finally, see that $L = \iota \circ L'$ as desired.

4.7 (Optional) Alternative proof of Woodbury formula

Let us give an application of rank normal form by proving the Woodbury formula. The idea is very straight forward: we change basis to make the maps look nice.

Theorem 4.7.1. *For any $m \times n$ matrix A and $n \times m$ matrix B, we have $I_m + AB$ invertible if and only if $I_n + BA$ invertible, and we have a formula $(I_m + AB)^{-1} = I_m - A(I_n + BA)^{-1}B$.*

Proof. Note that as a linear map, A goes from $\mathbb{R}^n$ to $\mathbb{R}^m$ and B goes from $\mathbb{R}^m$ to $\mathbb{R}^n$.

Let us change basis in $\mathbb{R}^m$ and $\mathbb{R}^n$ so that A looks nice. This means we have $A = P\begin{bmatrix} I_r & O \\ O & O \end{bmatrix} Q$ for some invertible matrices P, Q. Now, under these change of bases, $B = Q^{-1}\begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} P^{-1}$ for some B_1, B_2, B_3, B_4.

So $I_m + AB = I_m + P\begin{bmatrix} B_1 & B_2 \\ O & O \end{bmatrix} P^{-1} = P\begin{bmatrix} B_1 + I_r & B_2 \\ O & I_{m-r} \end{bmatrix} P^{-1}$. So this is invertible if and only if $B_1 + I_r$ is invertible.

Similarly, $I_n + BA = Q^{-1}\begin{bmatrix} B_1 + I_r & O \\ B_3 & I_{n-r} \end{bmatrix} Q$. So this is invertible if and only if $B_1 + I_r$ is invertible. So we have $I_m + AB$ invertible if and only if $I_n + BA$ invertible.

Now for the formula, note that $(I_n + BA)^{-1} = Q^{-1}\begin{bmatrix} (B_1 + I_r)^{-1} & O \\ -B_3(B_1 + I_r)^{-1} & I_{n-r} \end{bmatrix} Q$. So $A(I_n + BA)^{-1}B = P\begin{bmatrix} (B_1 + I_r)^{-1}B_1 & (B_1 + I_r)^{-1}B_2 \\ O & O \end{bmatrix} P^{-1}$.

In contrast, $(I_m + AB)^{-1} = P\begin{bmatrix} (B_1 + I_r)^{-1} & -(B_1 + I_r)^{-1}B_2 \\ O & I_{m-r} \end{bmatrix} P^{-1}$. So we see that $A(I_n + BA)^{-1}B + (I_m + AB)^{-1} = P\begin{bmatrix} (B_1 + I_r)^{-1}(B_1 + I_r) & O \\ O & I_{m-r} \end{bmatrix} P^{-1} = I_m$. So we are done. □

4.8 Rank-Nullity Theorem and Subspace Algebra

For a linear map L, recall that $\mathrm{Ker}(L)$ is the solution set to $L\boldsymbol{x} = \boldsymbol{0}$, while $\mathrm{Ran}(L)$ is the subspace of all its images. Both of these are subspaces.

Now under the right basis, L looks like its rank normal form. So it is very easy to see the following:

Theorem 4.8.1 (Rank-Nullity Theorem)**.** $\dim(\mathit{domain}(L)) = \mathrm{rank}(L) + \dim \mathrm{Ker}(L) = \dim \mathrm{Ran}(L) + \dim \mathrm{Ker}(L)$

Note that if the domain is $\mathbb{R}^n$, then for any system $L\boldsymbol{x} = \boldsymbol{b}$, we see that $\dim(\mathit{domain}(L))$ is the total number of variables. The rank of L is the number of dependent variables. Finally, $\dim \mathrm{Ker}(L)$ means all possible free directions to move inside the solution set of $L\boldsymbol{x} = \boldsymbol{0}$, i.e., the number of free variables.

Hence, This is the abstract statement of our previously established fact that "the number of dependent variable + the number of free variable = the number of all variables". Nevertheless, here is another formal proof.

Proof. Pick nice basis, and poof! We can now assume WLOG (without loss of generality) that L is the matrix $\begin{bmatrix} I_r & O \\ O & O \end{bmatrix}$.

Its range is spanned by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_r$, and its kernel is spanned by $\boldsymbol{e}_{r+1}, \dots, \boldsymbol{e}_n$. So we are done by counting. □

Remark 4.8.2. *Intuitively, if L starts with n dimensions, and kills $n - r$ dimensions, its range is left with r dimension.*

Here are some extreme cases.

Corollary 4.8.3. *L is injective if and only if $\mathrm{Ker}(L) = \{\mathbf{0}\}$, if and only if $\mathrm{Ran}(L)$ has the same dimension as the domain. L is surjective if and only if $\mathrm{Ran}(L)$ is the same as the codomain, if and only if $\dim \mathrm{Ker}(A)$ is the dimension of the domain minus the dimension of the codomain.*

Proof. Obviously L is surjective if and only if $\mathrm{Ran}(L)$ is the same as the codomain. Using rank-nullity, the statement on $\dim \mathrm{Ker}(A)$ is immediate.

For the injective case, note that $L\boldsymbol{v} = L\boldsymbol{w}$ iff $L(\boldsymbol{v} - \boldsymbol{w}) = \mathbf{0}$ iff $\boldsymbol{v} - \boldsymbol{w} \in \mathrm{Ker}(L)$. If the kernel is trivial, we have $\boldsymbol{v} = \boldsymbol{w}$, so L is injective. Conversly if L is injective and $L\boldsymbol{v} = \mathbf{0} = L\mathbf{0}$, we have $\boldsymbol{v} = \mathbf{0}$.

So L is injective if and only if $\mathrm{Ker}(L)$ is trivial, if and only if $\dim \mathrm{Ker}(L) = 0$, if and only if $\dim \mathrm{Ran}(L)$ is the same as the dimension of the domain. □

The idea here actually immediately gives us a picture of how solution sets are like.

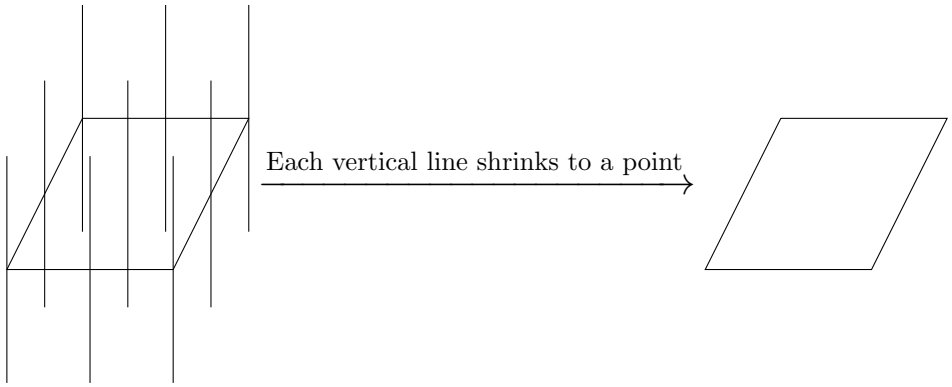
Corollary 4.8.4 (Description to all solutions to a linear system)**.** *For any linear map $L : V \to W$ and any $\boldsymbol{w} \in W$, the solution set for the equation $L\boldsymbol{x} = \boldsymbol{w}$ is a translation of $\mathrm{Ker}(L)$. More rigorously, if $\boldsymbol{x}_0$ is a solution, then the solution set is $\boldsymbol{x}_0 + \mathrm{Ker}(L) = \{\boldsymbol{x}_0 + \boldsymbol{v} \mid \boldsymbol{v} \in \mathrm{Ker}(L)\}$.*

Proof. If $L\boldsymbol{y} = \boldsymbol{w}$ too, then $L\boldsymbol{x}_0 = L\boldsymbol{y}$, so $\boldsymbol{y} - \boldsymbol{x}_0 \in \mathrm{Ker}(L)$. So $\boldsymbol{y} \in \boldsymbol{x}_0 + \mathrm{Ker}(L)$. □

Example 4.8.5. Consider $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. It sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} x \\ y \end{bmatrix}$, so it corresponds to the projection of $\mathbb{R}^3$ to the xy-plane. Its kernel is simply the z-axis.

So for any $\boldsymbol{b} \in \mathbb{R}^2$, what is the solution set to $A\boldsymbol{x} = \boldsymbol{b}$? It must be a parallel translation of the z-axis. It is simply some vertical line in $\mathbb{R}^3$. In particular, if we move $\boldsymbol{b}$ around, the solution set would also move around in a parallel manner.

In particular, the geometric behavior of A is this: in the domain $\mathbb{R}^3$, look at all the lines parallel to the z-axis. A would shrink each line to a point. Hence the range of A is just the xy-plane.

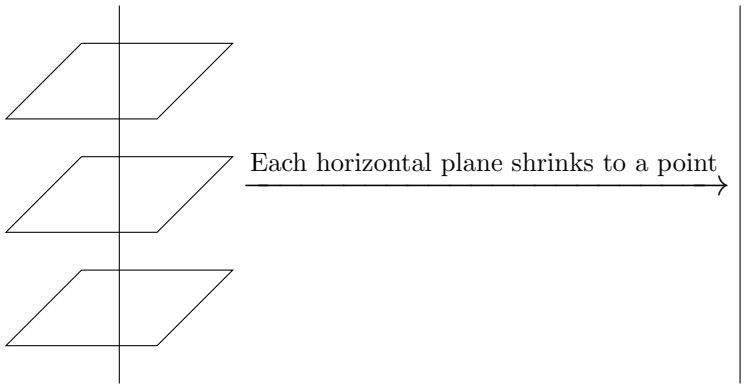


You can also see that the kernel has dimension one, the domain has dimension three. By shrinking everything parallel to the kernel, we end up with a range with dimension two.

Similarly, consider $A = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$. It sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to z, so it corresponds to the projection of $\mathbb{R}^3$ to the z-axis. Its kernel is simply the xy-plane.

So for any $b \in \mathbb{R}$, what is the solution set to $A\boldsymbol{x} = b$? It must be a parallel translation of the xy-plane. These are just various horizontal planes in $\mathbb{R}^3$. In particular, if we move b around, the solution set would also move around in a parallel manner.

In particular, the geometric behavior of A is this: in the domain $\mathbb{R}^3$, look at all the planes parallel to the xy-plane. A would shrink each plane to a point. Hence the range of A is just the z-axis.



You can also see that the kernel has dimension two, the domain has dimension three. By shrinking everything parallel to the kernel, we end up with a range with dimension one. ☺

Remark 4.8.6. *Now we have a clearer intuition on an arbitrary linear map.*

Say L has kernel dimension d. Then in fact, it is collapsing all d-dimensional affine subspaces parallel to $\mathrm{Ker}(L)$ *to points. Each d-dimensional object is now turned into a* 0*-dimensional object. So our n dimensional domain is turned in to the $n-d$ dimensional range.*

With our most powerful tool established, we are now free to study subspaces. The first important part is to study containment relation.

Proposition 4.8.7. *Let V be a finite dimensional space. If $W \subseteq V$, then* $\dim W \leq \dim V$*. If $W \subseteq V$ and* $\dim W = \dim V$*, then in fact $W = V$.*

Proof. Say $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are linearly independent vectors inside W. Note that, the definition of linear independence does not really need W. It simply tries to combine these vectors and see if we can get $\mathbf{0}$. Therefore, they are also linearly independent inside V. In particular, if V is finite dimensional, then W must also be finite dimensional.

If we pick a basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ for W, then they are linearly independent and they are in V. So we can extend it to a basis of V, and thus $\dim W \leq \dim V$.

In particular, if V is also k-dimensional, then $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ is a linearly independent collection of k vectors in V, and thus also a basis for V. So $V = W$.

For an alternative proof, consider the inclusion map $inc : W \to V$, which is linear. So say A is a matrix for it under some basis. Then the rank of A is at most the number of rows of A. But the former is just $\dim(W)$ while the latter is $\dim(V)$. So $\dim(W) \leq \dim(V)$. Furthermore, if the two are equal, then A is a square matrix. So it is injective (since it is inclusion) and square, hence bijective. But if the inclusion map is bijective, then it is in fact the identity map. So $V = W$. □

We already uses this from time to time without pointing it out. Now let us specifically consider subspaces of $\mathbb{R}^n$.

Corollary 4.8.8. *Any subspace W of $\mathbb{R}^n$ is* $\mathrm{Ran}(A)$ *for some A and* $\mathrm{Ker}(B)$ *for some B.*

Proof. Pick a basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in W \subseteq \mathbb{R}^n$ for W. Now $W = \mathrm{Ran}\begin{bmatrix}\boldsymbol{w}_1 & \ldots & \boldsymbol{w}_k\end{bmatrix}$. Done.

Now extend $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ to a basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ for V (via the independence extension lemma, essentially). Let $C = \begin{bmatrix}\boldsymbol{w}_1 & \ldots & \boldsymbol{w}_n\end{bmatrix}^{-1}$ be the change of coordimate map for this new basis, and $P = \begin{bmatrix}O & I_{n-k}\end{bmatrix}$ be the projection map sending $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ to $\begin{bmatrix} x_{k+1} \\ \vdots \\ x_n \end{bmatrix}$.

Then $\boldsymbol{v} \in W$ iff it only needs vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ among the basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, iff its last $n-k$ coordinates under the new basis are all zero, iff $PC\boldsymbol{v} = \mathbf{0}$. So $W = \mathrm{Ker}(PC)$. □

Now, given some subspaces, we are interested in how to make more subspaces.

Definition 4.8.9. *Given two subspaces U, W in V, their **intersection** $U \cap W$ is, well, um, their intersection. Their **sum** (or their **span**) $U + W$ is the set $\{\boldsymbol{u} + \boldsymbol{w} \mid \boldsymbol{u} \in U, \boldsymbol{w} \in W\}$.*

Proposition 4.8.10. *Given two subspaces, their intersection and their sum are subspaces.*

Proof. Suppose $\boldsymbol{u}, \boldsymbol{w} \in U \cap W$, and take any $a, b \in \mathbb{R}$.

Since $\boldsymbol{u}, \boldsymbol{w} \in U$ and U is a subspace, therefore $a\boldsymbol{u} + b\boldsymbol{w} \in U$. Similarly, since $\boldsymbol{u}, \boldsymbol{w} \in W$ and W is a subspace, therefore $a\boldsymbol{u} + b\boldsymbol{w} \in W$. Hence $a\boldsymbol{u} + b\boldsymbol{w} \in U \cap W$.

Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2 \in U + W$, and take any $a, b \in \mathbb{R}$.

For $i = 1, 2$, since $\boldsymbol{v}_i \in U + W$, we must have $\boldsymbol{v}_i = \boldsymbol{u}_i + \boldsymbol{w}_i$ for some $\boldsymbol{u}_i \in U, \boldsymbol{w}_i \in W$. Then $a\boldsymbol{v}_1 + b\boldsymbol{v}_2 = a(\boldsymbol{u}_1 + \boldsymbol{w}_1) + b(\boldsymbol{u}_2 + \boldsymbol{w}_2) = (a\boldsymbol{u}_1 + b\boldsymbol{u}_2) + (a\boldsymbol{w}_1 + b\boldsymbol{w}_2)$. Since the first summand is in U and the second summand is in W, the result is in $U + W$. □

In the specific case of subspaces of $\mathbb{R}^n$, we have the following nice block matrx interpretations of sum and intersections.

Proposition 4.8.11. $\operatorname{Ran}(A) + \operatorname{Ran}(B) = \operatorname{Ran}\begin{bmatrix} A & B \end{bmatrix}$, *and* $\operatorname{Ker}(A) \cap \operatorname{Ker}(B) = \operatorname{Ker}\begin{bmatrix} A \\ B \end{bmatrix}$.

Proof. $\operatorname{Ran}(A) + \operatorname{Ran}(B)$ means the all possible linear combinations of (all possible linear combinations of columns of A) and (all possible linear combinations of columns of B). $\operatorname{Ran}\begin{bmatrix} A & B \end{bmatrix}$ means the all possible linear combinations of columns of A and columns of B. Wait. Why do we even need to prove this? This is totally trivial!

$\boldsymbol{x} \in \operatorname{Ker}(A) \cap \operatorname{Ker}(B)$ means $A\boldsymbol{x} = B\boldsymbol{x} = \boldsymbol{0}$. On the other hand, $\boldsymbol{x} \in \operatorname{Ker}\begin{bmatrix} A \\ B \end{bmatrix}$ means $\begin{bmatrix} A\boldsymbol{x} \\ B\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}$. Wait. Why do we even need to prove this? Again, this is totally trivial! □

Generically, the following also provide some nice intuitions. In particular, they are analogous to the situations of subsets.

Proposition 4.8.12. *$U \cap W$ is the largest subspace contained in both, and $U + W$ is the smallest subspace containing both.*

Proof. Pretty straightforward by definition.

If V is a subspace containing both U and W, surely it contains all $\boldsymbol{u} + \boldsymbol{w}$ where $\boldsymbol{u} \in U$ and $\boldsymbol{w} \in W$. So $V \supseteq U + W$.

If V is a subspace inside both U and W, surely it is inside $U \cap W$. So $V \subseteq V \cap W$. □

Now, it might feel like sums and intersections of subspaces bear some similarity to unions (smallest subset containing both subsets) and intersections (largest subset inside both subsets) of subsets. However, there are some crucial distinctions.

Example 4.8.13. Subspace algebra can be very tricky, and very dangerous. Recall that, normally, for ordinary subsets, then we have distrubutativity of $\cup$ and $\cap$ over each other, namely, $A \cup (B \cap C) = (A \cup B) \cup (A \cup C)$ and $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$. Draw some Vayne diagram to see this.

However, suppose we look at $\mathbb{R}^2$, let U be the line $x = 0$, let V be the line $y = 0$ and let W be the line $x = y$. Note that any two of them can intersect only at the origin, and any two of them would span the whole space.

$U + (V \cap W)$ may NOT equal to $(U + V) \cap (U + W)$. The first one is U, while the second one is the whole $\mathbb{R}^2$.

$U \cap (V + W)$ may NOT equal to $U \cap V + U \cap W$. The first one is U, while the second one is the origin. ☺

However, sometimes the analogy is true. Recall that for two subsets, we have $|A \cup B| = |A| + |B| - |A \cap B|$. (This is called the ***inclusion-exclusion principle*** sometimes.) In particular, if A, B are both big but $A \cup B$ is not big enough, then $A \cup B$ would be "too crowded", and thus A, B must have big intersections. The same goes for subspaces

Theorem 4.8.14 (Inclusion-Exclusion Principle for Subspaces)**.** $\dim(V+W) = \dim V + \dim W - \dim(V \cap W)$.

Proof. The first proof is done by basis-extension and counting. Suppose $\dim(V \cap W) = k, \dim V = a, \dim W = b$.

Pick a basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ for $V \cap W$. Extend this to a basis of V as $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{a-k}$. Similarly, extend the linearly independent collection $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ to a basis of W as $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b-k}$. I claim that $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{a-k}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b-k}$ form a basis for $V+W$. (This would then imply that $\dim(V+W) = k + (a-k) + (b-k) = a + b - k$, establishing the desired result.)

Spanning is obvious, so we only prove linear independence here. Suppose $\sum a_i \boldsymbol{u}_i + \sum b_i \boldsymbol{v}_i + \sum c_i \boldsymbol{w}_i = \boldsymbol{0}$. Then $\sum a_i \boldsymbol{u}_i + \sum b_i \boldsymbol{v}_i = -\sum c_i \boldsymbol{w}_i$. Note that the left hand side is in V, while the right hand side is in W, so both sides are in $V \cap W$!

So $\sum c_i \boldsymbol{w}_i \in V \cap W \subseteq W$. However, every vector in W must be a unique linear combination of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b-k}$, and if the vector is actually in $V \cap W$, then the coordinates for $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b-k}$ must all be zero. Hence, since $\sum c_i \boldsymbol{w}_i \in V \cap W \subseteq W$, we see that all c_i are zero.

Then $\sum a_i \boldsymbol{u}_i + \sum b_i \boldsymbol{v}_i = \boldsymbol{0}$. But since $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{a-k}$ is a basis for V, this implies that all a_i and all b_i are zero.

So $\sum a_i \boldsymbol{u}_i + \sum b_i \boldsymbol{v}_i + \sum c_i \boldsymbol{w}_i = \boldsymbol{0}$ implies that all coefficients are zero. Hence $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{a-k}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{b-k}$ is a linearly independent collection of vectors. So it is indeed a basis.

Now let us do an alternative (hardcore but optional) proof. This proof is arguably more abstract, but it is what advanced mathemagicians (ha!) typically would do: we solve a problem by building TONS of structures around the problem, and see what would happen.

Consider the following chain

$$V \cap W \xrightarrow{D} V \times W \xrightarrow{S} V + W$$

Let us explain the space $V \times W$ and the maps D and S.

Here the space $V \times W$ refers to the vector space $\{\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} : \boldsymbol{v} \in V, \boldsymbol{w} \in W\}$ where addition and scalar multiplication are done in the obvious manner. Clearly if $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_a$ is a basis for V and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_b$ is a basis for W, then a basis for $V \times W$ is $\begin{bmatrix} \boldsymbol{v}_i \\ \boldsymbol{0} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{w}_j \end{bmatrix}$ for all i, j. So $\dim(V \times W) = \dim V + \dim W$.

The map D sends $\boldsymbol{x}$ to $\begin{bmatrix} \boldsymbol{x} \\ -\boldsymbol{x} \end{bmatrix}$. And the map S sends $\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix}$ to $\boldsymbol{v} + \boldsymbol{w}$. These are all linear maps, as you can verify. Furthermore, it is obvious by definition that D is injective and S is surjective.

On extra remark, we also have $\mathrm{Ran}(D) = \mathrm{Ker}(S)$. (This is called ***exact***. The spaces $V \cap W, V \times W, V + W$ form a classical short exact sequence, which is an important phenomenon in all algebra studies.) To see this, first of all obviously $S \circ D$ is the zero map (it sends everything to $\boldsymbol{0}$). Hence $\mathrm{Ran}(D) \subseteq \mathrm{Ker}(S)$. Conversely, if $\boldsymbol{v} \in \mathrm{Ker}(S)$, then $S(\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix}) = \boldsymbol{0}$, then $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{0}$, hence $\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v} \\ -\boldsymbol{v} \end{bmatrix} = D(\boldsymbol{v}) \in \mathrm{Ran}(D)$. Hence $\mathrm{Ker}(S) \subseteq \mathrm{Ran}(D)$. So we have the desired "exactness".

We have finished building. Now to prove the statement, just use rank-nullity twice, once on D and once on S. The rank-nullity on D gives

$$\dim(V \cap W) = \dim \mathrm{Ran}(D) + \dim \mathrm{Ker}(D) = \dim \mathrm{Ran}(D) = \dim \mathrm{Ker}(S).$$

Here $\dim \mathrm{Ker}(D) = 0$ because D is injective.

The rank-nullity on S gives

$$\dim(V) + \dim(W) = \dim(V \times W) = \dim \mathrm{Ker}(S) + \dim \mathrm{Ran}(S) = \dim(V \cap W) + \dim(V + W).$$

So we are done.

Intuitively, short exact sequences are used to decompose algebraic structures. We are pointing out that $V \times W$ is in some sense composed of $V \cap W$ and $V + W$. □

Remark 4.8.15. *(Entirely optional remark)*

The second proof above is actually the same as the popular proof for the inclusion exclusion principle for subsets. For subsets, a proof is typically like this: we "pretent" that there is no intersection and count elements of X and elements of Y. Now when we put these elements into $X \cup Y$, we would have overlaps. The overlaps are $X \cap Y$. Hence $|X \cup Y| = |X| + |Y| - |X \cap Y|$.

For our proof here, the space $V \times W$ is the linear algebra way of "pretending" that the two spaces have zero intersection. In particular, $\dim(V \times W) = \dim V + \dim W$.

For subsets, when we put X, Y into $X \cup Y$, we are performing inclusion maps. Here the map $S : V \times W \to V + W$ is actually the natural linear extension of both inclusion maps $V \to V + W$ and $W \to V + W$.

Finally, for subsets, we would try to look at overlaps of X, Y in $X \cup Y$. Correspondingly, overlaps for the map $S : V \times W \to V + W$ can be studied as the kernel of S. (Which is the collection of inputs with common image $\mathbf{0}$, i.e., they "overlap" at $\mathbf{0}$.)

Here is an immediate application.

Corollary 4.8.16. *If $\dim(W) = n - 1$ is a subspaces of n-dimensional space V, then for any subspace U of V, either $U \subseteq W$ or $\dim(U \cap W) = \dim(U) - 1$.*

This essentially reads that any hyperplane would always cut your dimension by one. Say you are in $\mathbb{R}^n$ and have k hyperplanes in generic position. ("Generic position" means their equations have no redundancy or contradiction.) What is the dimension of their intersection? It would be $n - k$, as each plane cut the dimension by 1.

In particular, n variables with k effective equations means your solution set is an $n - k$ dimensional (affine) space.

Here is another application of the inclusion-exclusion formula.

Corollary 4.8.17. *If $V \cap W = \{\mathbf{0}\}$, then $\dim(V + W) = \dim V + \dim W$.*

A special case of above is the concept of complement subspaces.

Definition 4.8.18. *We say two subspaces U, W of V are complements if $U + W = V$ and $U \cap W = \{\mathbf{0}\}$.*

This is just the analogy of complements for subsets, where we require $S \cup T$ to be the whole set and $S \cap T = \varnothing$.

Corollary 4.8.19. *If U, W are complements of each other in V, then $\dim V = \dim U + \dim W$.*

Another useful phenomenon is that they provides unique decomposition of vectors, similar to $\begin{bmatrix}1\\2\\3\end{bmatrix} = \begin{bmatrix}1\\2\\0\end{bmatrix} + \begin{bmatrix}0\\0\\3\end{bmatrix}$. (This is the unique decomposition where we think of the xy-plane and the z-axis as complementary subspaces.)

Corollary 4.8.20. *If U, W are complementary subspaces of V, then for any $\boldsymbol{v} \in V$, we can find unique $\boldsymbol{u} \in U$ and unique $\boldsymbol{w} \in W$ such that $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$.*

Proof. Since $U + W = V$, such decomposition must exist.

Now suppose $\boldsymbol{u}_1 + \boldsymbol{w}_1 = \boldsymbol{v} = \boldsymbol{u}_2 + \boldsymbol{w}_2$. Then rearrange this gives $\boldsymbol{u}_1 - \boldsymbol{u}_2 = \boldsymbol{w}_2 - \boldsymbol{w}_1$. Now the left hand side is in U and the right hand side is in W, so both are in $U \cap W = \{0\}$. Hence both sides are zero, $\boldsymbol{u}_1 = \boldsymbol{u}_2, \boldsymbol{w}_1 = \boldsymbol{w}_2$. □

Now again, the analogy holds at the amount of dimensions, but fail in terms of actual subspaces. In particular complements of subspaces are NOT unique. For example, in $\mathbb{R}^2$, the y-axis and the line $x = y$ are both complements to the x-axis. (This is the same example that fails the law of distribution among subspaces. This is NOT a coincidence, but we don't need to dwell on it too much at the moment.)

4.9 Rank Inequalities (Optional)

Here are somethings that are very useful sometimes, if you want to do rank estimates for some generic linear maps. Nevertheless, they are not an essential part of this class, and is not used for other contents in this lecture notes. We present them here, along with an application: to decide when a square matrix (not necessarily invertible) can have LU decompositions.

Definition 4.9.1. *Given a linear map $L : V \to W$ and a subspace $U \subseteq W$, the pullback of U via L is the subspace is $L^{-1}(U) = \{\boldsymbol{v} \mid L(\boldsymbol{v}) \in U\}$.*

Dually, given a subspace $U \subseteq V$, the pushforward of U via L is the subspace $L(U) = \{L(\boldsymbol{v}) \mid \boldsymbol{v} \in U\}$.

These can be thought of as generalizations of range and kernel. The range is the pushforward of the domain. The kernel is the pullback of $\{\boldsymbol{0}\}$. In particular, here is a very interesting example.

Proposition 4.9.2. *For any matrices A, B, we have* $\operatorname{Ran}(AB) = A(\operatorname{Ran}(B))$ *and* $\operatorname{Ker}(AB) = B^{-1}(\operatorname{Ker}(A))$.

Proof. These are very clear just from the definition. For example, we have $\operatorname{Ran}(AB) = \{AB\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^n\} = A\{B\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^n\} = A\operatorname{Ran}(B)$. □

Let us establish some dimension results that is enough for all our future computations.

Lemma 4.9.3 (Range and Kernal of restrictions of maps)**.** *Given a linear map $L : V \to W$ and a subspace U of W, let L' be the restriction of L with domain $L^{-1}(U)$ and codomain U. Then* $\operatorname{Ran}(L') = \operatorname{Ran}(L) \cap U$ *and* $\operatorname{Ker}(L') = \operatorname{Ker}(L)$.

Given a linear map $L : V \to W$ and a subspace U of V, let L' be the restriction of L with domain U and codomain $L(U)$. Then $\operatorname{Ran}(L') = L(U)$ *and* $\operatorname{Ker}(L') = \operatorname{Ker}(L) \cap U$.

Proof. The proofs are mostly trivial, but lengthy (because every equality between sets $X = Y$ need to be proven twice as $X \subseteq Y$ and $Y \subseteq X$, hence there are 8 things to prove...). Feel free to skip these proofs.

For both statements, beware of the key relation between L and L'. If $\boldsymbol{x}$ is in the domain of L', then we must have $L'\boldsymbol{x} = L\boldsymbol{x}$ by definition. If $\boldsymbol{x}$ is not in the domain of L', then $L'\boldsymbol{x}$ is not defined while $L\boldsymbol{x}$ could be defined.

Let us prove the first paragraph.

Therefore, pick any $\boldsymbol{x} \in \operatorname{Ker}(L')$. Then $L'\boldsymbol{x} = \boldsymbol{0}$. Hence $L\boldsymbol{x} = \boldsymbol{0}$ and $\boldsymbol{x} \in \operatorname{Ker}(L)$. Conversely, if $\boldsymbol{x} \in \operatorname{Ker}(L)$, then $\boldsymbol{x} \in L^{-1}(\{\boldsymbol{0}\}) \subseteq L^{-1}(U)$, hence $L'\boldsymbol{x}$ is defined. So $L'\boldsymbol{x} = L\boldsymbol{x} = \boldsymbol{0}$ and thus $\boldsymbol{x} \in \operatorname{Ker}(L')$.

Pick any $\boldsymbol{b} \in \operatorname{Ran}(L')$, then $\boldsymbol{b} = L'\boldsymbol{x}$ for some $\boldsymbol{x}$ in the domain of L'. Hence $\boldsymbol{b} = L\boldsymbol{x} \in \operatorname{Ran}(L)$. And since $\boldsymbol{x} \in L^{-1}(U)$, we see that $\boldsymbol{b} = L\boldsymbol{x} \in U$. So $\boldsymbol{b} \in \operatorname{Ran}(L) \cap U$.

Conversely, say $\boldsymbol{b} \in \operatorname{Ran}(L) \cap U$. Since $\boldsymbol{b} \in \operatorname{Ran}(L)$, by definition this means $\boldsymbol{b} = L\boldsymbol{x}$ for some $\boldsymbol{x}$. But then since $\boldsymbol{b} \in U$, this means that $\boldsymbol{x} \in L^{-1}(U)$. Hence L' is defined on $\boldsymbol{x}$, and $L'\boldsymbol{x} = L\boldsymbol{x} = \boldsymbol{b}$. So $\boldsymbol{b} \in \operatorname{Ran}(L')$.

Now let us prove the second paragraph.

$L(U)$ is literally defined as the set of images of $L\boldsymbol{u}$ for $\boldsymbol{u} \in U$, hence L' is surjective by construction. So we have $\operatorname{Ran}(L') = L(U)$.

If $\boldsymbol{x} \in \operatorname{Ker}(L')$, then $L'\boldsymbol{x} = \boldsymbol{0}$, hence $\boldsymbol{x}$ is defined (i.e., $\boldsymbol{x} \in U$), and hence $L\boldsymbol{x} = L'\boldsymbol{x} = \boldsymbol{0}$. So we also have $\boldsymbol{x} \in \operatorname{Ker}(L)$. So $\boldsymbol{x} \in \operatorname{Ker}(L) \cap U$.

Conversely, say $\boldsymbol{x} \in U \cap \operatorname{Ker}(L)$. Then $L\boldsymbol{x} = \boldsymbol{0}$, and L' is defined on $\boldsymbol{x}$. Therefore $L'\boldsymbol{x} = L\boldsymbol{x} = \boldsymbol{0}$. So $\boldsymbol{x} \in \operatorname{Ker}(L')$. □

Proposition 4.9.4 (Dimensions of pullbacks and pushforwards of subspaces)**.**

$$\dim L^{-1}(U) = \dim(U \cap \operatorname{Ran}(L)) + \dim \operatorname{Ker}(L).$$

$$\dim L(U) = \dim U - \dim(U \cap \operatorname{Ker}(L)).$$

Proof. Just apply rank-nullity on L' in the previous lemma. □

You don't need to remember the formula above. Personally I simply keep in mind that all such things can be done via rank-nullity, and I search for these formulas online (or re-deduce them myself) when I need them.

Corollary 4.9.5. *We have many rank inequalities among matrices:*

1. $\operatorname{rank}(A+B) \le \operatorname{rank}(A) + \operatorname{rank}(B)$.
2. $\operatorname{rank}(A) + \operatorname{rank}(B) - n \le \operatorname{rank}(AB) \le \min(\operatorname{rank}(A), \operatorname{rank}(B))$. *(Here n is the dimension of the domain of A which is also the codomain of B.)*
3. *(Optional)* $\max(\operatorname{rank}(A), \operatorname{rank}(B)) \le \operatorname{rank}(\begin{bmatrix} A & B \end{bmatrix}) \le \operatorname{rank}(A) + \operatorname{rank}(B)$.
4. *(Optional)* $\max(\operatorname{rank}(A), \operatorname{rank}(B)) \le \operatorname{rank}(\begin{bmatrix} A \\ B \end{bmatrix}) \le \operatorname{rank}(A) + \operatorname{rank}(B)$.
5. *(Optional)* $\operatorname{rank}(\begin{bmatrix} A & O \\ O & B \end{bmatrix}) = \operatorname{rank}(A) + \operatorname{rank}(B)$.
6. *(Optional)* $\operatorname{rank}(\begin{bmatrix} A & O \\ C & B \end{bmatrix}) \ge \operatorname{rank}(A) + \operatorname{rank}(B)$.

Proof. The goal is NOT to prove these things. Rather, the goal is to practice using rank-nullity on subspaces.

1. Obvious because all columns of $A+B$ are linear combinations of columns of A and columns of B, so $\operatorname{Ran}(A+B) \subseteq \operatorname{Ran}(A) + \operatorname{Ran}(B)$. Take dimensions and $\dim \operatorname{Ran}(A+B) \le \dim(\operatorname{Ran}(A) + \operatorname{Ran}(B)) \le \dim \operatorname{Ran}(A) + \dim \operatorname{Ran}(B)$. The last inequality is inclusion-exclusion principle.
2. Obviously any image via AB is in particular some image of A, so $\operatorname{Ran}(AB) \subseteq \operatorname{Ran}(A)$, which gives $\operatorname{rank}(AB) \le \operatorname{rank}(A)$.

   Now $\operatorname{Ran}(AB) = \{AB\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^n\} = A\{B\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^n\} = A \operatorname{Ran}(B)$. By the pushforward formula (equivalently, by rank-nullity on the map from $\operatorname{Ran}(B)$ to $A\operatorname{Ran}(B)$), we have $\dim A\operatorname{Ran}(B) = \dim \operatorname{Ran}(B) - \dim(\operatorname{Ran}(B) \cap \operatorname{Ker}(A))$. This immediatley established $\operatorname{rank}(AB) \le \operatorname{rank}(B)$.

   By the same equation, we can also have $\dim A\operatorname{Ran}(B) \ge \dim \operatorname{Ran}(B) - \dim \operatorname{Ker}(A) = \dim \operatorname{Ran}(B) + \dim \operatorname{Ran}(A) - n$. So we are done.
3. For $\begin{bmatrix} A & B \end{bmatrix}$ both sides are obvious, because $\operatorname{Ran}\begin{bmatrix} A & B \end{bmatrix} = \operatorname{Ran}(A) + \operatorname{Ran}(B)$.
4. Let n be the dimension of the common domain. Now $\operatorname{Ker}(\begin{bmatrix} A \\ B \end{bmatrix}) = \operatorname{Ker}(A) \cap \operatorname{Ker}(B)$. Take dimensions we have $\min(\dim \operatorname{Ker}(A), \dim \operatorname{Ker}(B)) \ge \dim \operatorname{Ker}(\begin{bmatrix} A \\ B \end{bmatrix}) = \dim \operatorname{Ker}(A) + \dim \operatorname{Ker}(B) - \dim(\operatorname{Ker}(A) + \operatorname{Ker}(B)) \ge \dim \operatorname{Ker}(A) + \dim \operatorname{Ker}(B) - n$. Now use the fact that rank is n minus the dimension of the kernel, and these converts to the desried equation.
5. It looks quite obvious that $\operatorname{Ran}(\begin{bmatrix} A \\ O \end{bmatrix})$ and $\operatorname{Ran}(\begin{bmatrix} O \\ B \end{bmatrix})$ have trivial intersection, while the former has the same dimension as $\operatorname{Ran}(A)$, and the latter has the same dimension as $\operatorname{Ran}(B)$.
6. Suppose A has RREF A' and rank a, and B has RREF B' and rank b. Then by row operations, $\operatorname{rank}(\begin{bmatrix} A & O \\ C & B \end{bmatrix}) = \operatorname{rank}(\begin{bmatrix} A' & O \\ * & B' \end{bmatrix})$. Now, deleting columns would obviously only reduce the dimension of the column space, so it would only decrease rank. By dropping all non-pivotal columns, we have

$\operatorname{rank}(\begin{bmatrix} A' & O \\ * & B' \end{bmatrix}) \geq \operatorname{rank}(\begin{bmatrix} I_a & O \\ O & O \\ * & I_b \\ * & O \end{bmatrix})$. By further block row operations, we have $\operatorname{rank}(\begin{bmatrix} I_a & O \\ O & O \\ * & I_b \\ * & O \end{bmatrix}) = \operatorname{rank}(\begin{bmatrix} I_a & O \\ O & O \\ O & I_b \\ O & O \end{bmatrix}) = a + b$. So we are done.

□

Remark 4.9.6. *Note that* $\begin{bmatrix} A & B \end{bmatrix}$ *and* $\begin{bmatrix} A \\ B \end{bmatrix}$*, although similar in terms of rank inequalities, might NOT have the same rank. For example, take* $\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 \end{bmatrix}$.

The meaning behind the proof is sometimes more important than the actual results. For example, consider the statement $\operatorname{rank}(A) + \operatorname{rank}(B) - n \leq \operatorname{rank}(AB)$. In the proof, we essentially established that $\dim \operatorname{Ran}(AB) \geq \dim \operatorname{Ran}(B) - \dim \operatorname{Ker}(A)$. So if A killed k dimensions, then performing the linear map A after B, A would kill at most k dimensions in $\operatorname{Ran}(B)$. Hence the resulting space $\operatorname{Ran}(AB)$ has a dimension at most $\dim \operatorname{Ran}(B) - \dim \operatorname{Ker}(A)$.

The inequality $\dim \operatorname{Ran}(AB) \geq \dim \operatorname{Ran}(B) - \dim \operatorname{Ker}(A)$ is sometimes more important than the actual rank inequality. For example, one can easily see the following fun statement, by using this inequality repeatedly. (Useful next semester for the establishement of Jordan canonical form.)

Proposition 4.9.7. *If* A *is* $n \times n$ *and* $A^k = n$*, then* $\dim \operatorname{Ker}(A) \geq \frac{n}{k}$.

Proof. Intuitively, if $\dim \operatorname{Ker}(A) < \frac{n}{k}$, then for $\operatorname{Ran}(A^k) = A^k \operatorname{Ran}(I_n)$, each application of A to the subspace would kill less than $\frac{n}{k}$ dimensions, and a total of k steps would kill less than n dimensions. So $\dim \operatorname{Ran}(A^k) > 0$, there must be surviving dimensions.

Rigorous proof goes like this: If $A^k = 0$, then we have

$$\begin{aligned} 0 &= \dim \operatorname{Ran}(A^k) \\ &\geq \dim \operatorname{Ran}(A^{k-1}) - \dim \operatorname{Ker}(A) \\ &\geq \dim \operatorname{Ran}(A^{k-2}) - 2 \dim \operatorname{Ker}(A) \\ &\geq \dots \\ &\geq \dim \operatorname{Ran}(A^0) - k \dim \operatorname{Ker}(A) = n - k \operatorname{Ker}(A). \end{aligned}$$

Hence $\operatorname{Ker}(A) \geq \frac{n}{k}$. □

Here is one last inequality between ranks that is sometimes useful.

Proposition 4.9.8 (The Inclusion-Exclusion Principle for maps)**.** *For any* $m \times d$ *matrix* A*,* $d \times k$ *matrix* B *and* $k \times n$ *matrix* C*, we have* $\operatorname{rank}(AB) + \operatorname{rank}(BC) \leq \operatorname{rank}(ABC) + \operatorname{rank}(B)$.

Intuitively, ABC is the union of AB and BC, while B is the intersection of AB and BC. Hence this is a version of "Inclusion-Exclusion Principle" for maps. This intuition also inspires a very beautiful proof. For subspaces, we have $\dim(V) + \dim(W) = \dim(V + W) + \dim(V \cap W)$, and the proof relies on the construction of an intermediate space $V \times W$ with dimension $\dim(V) + \dim(W)$. This idea can be used for all variant versions of "inclusion-exclusion principle". In this case, how to build an intermediate map for AB and BC? We want a way to combine the two maps independently, and one way to do this is via the diagram below, which represents the two submaps of a big map.

$$\mathrm{Dom}(B) \xrightarrow{\quad AB \quad} \mathrm{Cod}(A)$$

$$\mathrm{Dom}(C) \xrightarrow{\quad BC \quad} \mathrm{Cod}(B) \quad .$$

So we pick the domain as the space $\mathrm{Dom}(B) \times \mathrm{Dom}(C)$, and the codomain as $\mathrm{Cod}(A) \times \mathrm{Cod}(B)$, then we can treat AB and BC as the diagonal blocks of a big map, i.e., the "independent union" of both maps. However, how to handle the intersection map B? From the diagram above, there is an obvious place to put it, and it gives us a new diagram

$$\begin{array}{ccc} \mathrm{Dom}(B) & \xrightarrow{\quad AB \quad} & \mathrm{Cod}(A) \\ & \searrow & \\ \mathrm{Dom}(C) & \xrightarrow{\quad BC \quad} & \mathrm{Cod}(B) \end{array} \quad .$$

So we now have a map $M : \mathrm{Dom}(B) \times \mathrm{Dom}(C) \to \mathrm{Cod}(A) \times \mathrm{Cod}(B)$ with block structure $M = \begin{bmatrix} AB & \\ B & BC \end{bmatrix}$.

Proof of IEP for maps. Consider $\begin{bmatrix} AB & \\ B & BC \end{bmatrix}$. (The combination of AB and BC is represented as the diagonal blocks, while the intersection of AB and BC is represented by the lower left block.) We already know that $\mathrm{rank}(AB) + \mathrm{rank}(BC) \leq \mathrm{rank}(\begin{bmatrix} AB & \\ B & BC \end{bmatrix})$.

On the other hand, by row and column block operations we have $\begin{bmatrix} AB & \\ B & BC \end{bmatrix} \to \begin{bmatrix} & -ABC \\ B & BC \end{bmatrix} \to \begin{bmatrix} & -ABC \\ B & \end{bmatrix}$, so $\mathrm{rank}(\begin{bmatrix} AB & \\ B & BC \end{bmatrix}) = \mathrm{rank}(\begin{bmatrix} & -ABC \\ B & \end{bmatrix}) = \mathrm{rank}(B) + \mathrm{rank}(-ABC) = \mathrm{rank}(B) + \mathrm{rank}(ABC)$. So we are done. □

If you think the proof above is too "magical", then here is another proof using subspaces.

Alternative Proof. All these maps have many different domains and codomains. For example, it makes little sense to add $\dim \mathrm{Ran}(AB)$ and $\dim \mathrm{Ran}(BC)$, because $\mathrm{Ran}(AB)$ and $\mathrm{Ran}(BC)$ are not in the same underlying vector space. To make meaningful progress, we need to fix a specific underlying vector space. Say we pick the domain of A.

We rearrange the inequality as $\mathrm{rank}(BC) - \mathrm{rank}(ABC) \leq \mathrm{rank}(B) - \mathrm{rank}(AB)$. (We immediatey see that the Inclusion-Exclusion Principle for maps is a generalization of the fact that $\mathrm{rank}(BC) \leq \mathrm{rank}(B)$.)

Now $\mathrm{rank}(AB) = \dim \mathrm{Ran}(AB) = \dim A(\mathrm{Ran}(B)) = \dim \mathrm{Ran}(B) - \dim(\mathrm{Ran}(A) \cap \mathrm{Ker}(B))$. So the right hand side is simply $\dim(\mathrm{Ran}(A) \cap \mathrm{Ker}(B))$. Similarly, the left hand side is simply $\dim(\mathrm{Ran}(A) \cap \mathrm{Ker}(BC))$.

So we now want to prove that $\dim(\mathrm{Ker}(A) \cap \mathrm{Ran}(BC)) \leq \dim(\mathrm{Ker}(A) \cap \mathrm{Ran}(B))$. Since we know $\mathrm{Ran}(BC) \subseteq \mathrm{Ran}(B)$, we have $\mathrm{Ker}(A) \cap \mathrm{Ran}(BC) \subseteq \mathrm{Ker}(A) \cap \mathrm{Ran}(B)$, and hence the dimension inequality is also true. □

The alternative proof here is more revealing. Ultimately, the Inclusion-Exclusion Principle for maps is about comparing the subspaces $\mathrm{Ker}(A) \cap \mathrm{Ran}(BC)$ with $\mathrm{Ker}(A) \cap \mathrm{Ran}(B)$.

Also note that taking $B = I$, from the inequality $\mathrm{rank}(AB) + \mathrm{rank}(BC) \leq \mathrm{rank}(ABC) + \mathrm{rank}(B)$, we can also deduce that $\mathrm{rank}(A) + \mathrm{rank}(C) \leq \mathrm{rank}(AC) + n$, which we know before.

4.10 When can we LU (Optional)

We know that, for an invertible matrix A, an LU decomposition is possible if and only if all principal submatrices are invertible. But what if A is not invertible? An LU decomposition might still be possible. For examlpe, the zero matrix is the zero matrix multiplied by the zero matrix. Here is a more exotic example.

Example 4.10.1. Consider $\begin{bmatrix}0&0&0\\0&0&1\\0&1&1\end{bmatrix}$. If you attempt to do Gaussian elimination, surely you need to row swap right away! So can there still be LU decompositions?

There IS. We have $\begin{bmatrix}0&0&0\\0&0&1\\0&1&1\end{bmatrix} = \begin{bmatrix}0&0&0\\1&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}0&0&1\\0&1&1\\0&0&0\end{bmatrix}$. The non-zero portion of $\begin{bmatrix}0&0&0\\0&0&1\\0&1&1\end{bmatrix}$ can "shift-up" to satisfy the upper triangular requirement, and the "shift-down" operator is lower triangular. ☺

For a generic (maybe not invertible) square matrix A, when can we perform an LU decomposition?

Theorem 4.10.2. *A square matrix A has an LU decomposition if and only if the following is true: for all k, if $A = \begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}$ where A_{11} is $k\times k$, then we have* $\operatorname{rank}(A_{11}) + k \geq \operatorname{rank}(\begin{bmatrix}A_{11}&A_{12}\end{bmatrix}) + \operatorname{rank}(\begin{bmatrix}A_{11}\\A_{21}\end{bmatrix})$.

Proof. Let us only prove the forward direction. Suppose $A = LU$ for lower triangular L and upper triangular U, and we perform block decompositions $L = \begin{bmatrix}L_{11}&\\L_{21}&L_{22}\end{bmatrix}$ and $U = \begin{bmatrix}U_{11}&U_{12}\\&U_{22}\end{bmatrix}$, where L_{11}, U_{11} are $k\times k$.

Now $\operatorname{rank}(\begin{bmatrix}A_{11}&A_{12}\end{bmatrix}) = \operatorname{rank}(\begin{bmatrix}L_{11}U_{11}&L_{11}U_{12}\end{bmatrix}) = \operatorname{rank}(L_{11}\begin{bmatrix}U_{11}&U_{12}\end{bmatrix}) \leq \operatorname{rank}(L_{11})$, and similarly $\operatorname{rank}(\begin{bmatrix}A_{11}\\A_{21}\end{bmatrix}) \leq \operatorname{rank}(U_{11})$. So $\operatorname{rank}(\begin{bmatrix}A_{11}&A_{12}\end{bmatrix}) + \operatorname{rank}(\begin{bmatrix}A_{11}\\A_{21}\end{bmatrix}) \leq \operatorname{rank}(L_{11}) + \operatorname{rank}(U_1 1) \leq \operatorname{rank}(L_{11}U_{11}) + k = \operatorname{rank}(A_{11}) + k$. So we are done.

For the other direction, the proof is basically by mathematical induction on n. When $n = 1$, again the case is trivial.

Suppose $n > 1$. We write $A = \begin{bmatrix}a&\boldsymbol{v}^{\mathrm{T}}\\\boldsymbol{w}&A_{n-1}\end{bmatrix}$, and suppose that our conditions on rank of submatrices are all true.

Case 1:

Suppose $a \neq 0$. Then $A = \begin{bmatrix}1&\\\frac{1}{a}\boldsymbol{w}&I_{n-1}\end{bmatrix}\begin{bmatrix}a&\boldsymbol{v}^{\mathrm{T}}\\&A_{n-1}-\frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}\end{bmatrix}$. We now aim to show that $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}$ satisfy the induction hypothesis and thus has an LU decomposition.

For any positive integer $k < n-1$, let $A_{n-1} = \begin{bmatrix}A_k&B_k\\C_k&D_k\end{bmatrix}$, and let $\boldsymbol{v}^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{v}_k^{\mathrm{T}}&(\boldsymbol{v}')^{\mathrm{T}}\end{bmatrix}$, and let $\boldsymbol{w} = \begin{bmatrix}\boldsymbol{w}_k\\\boldsymbol{w}'\end{bmatrix}$. Here A_k is $k\times k$, and $\boldsymbol{v}_k, \boldsymbol{w}_k \in \mathbb{R}^k$. Then our condition on A states that $\operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}\\\boldsymbol{w}_k&A_k\end{bmatrix}) + k + 1$ $\geq \operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}\\\boldsymbol{w}_k&A_k\\\boldsymbol{w}'&C_k\end{bmatrix}) + \operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}&(\boldsymbol{v}')^{\mathrm{T}}\\\boldsymbol{w}_k&A_k&B_k\end{bmatrix})$. So we have the deduction

$$\begin{aligned}
&\operatorname{rank}(A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}}) + k\\
=&\operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}\\&A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}}\end{bmatrix}) + k - 1\\
=&\operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}\\\boldsymbol{w}_k&A_k\end{bmatrix}) + k - 1\\
\geq&\operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}\\\boldsymbol{w}_k&A_k\\\boldsymbol{w}'&C_k\end{bmatrix}) + \operatorname{rank}(\begin{bmatrix}a&\boldsymbol{v}_k^{\mathrm{T}}&(\boldsymbol{v}')^{\mathrm{T}}\\\boldsymbol{w}_k&A_k&B_k\end{bmatrix}) - 2
\end{aligned}$$

$$= \text{rank}(\begin{bmatrix} a & \boldsymbol{v}_k^{\mathrm{T}} \\ & A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}} \\ & C_k - \frac{1}{a}\boldsymbol{w}'\boldsymbol{v}_k^{\mathrm{T}} \end{bmatrix}) + \text{rank}(\begin{bmatrix} a & \boldsymbol{v}_k^{\mathrm{T}} & (\boldsymbol{v}')^{\mathrm{T}} \\ & A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}} & B_k - \frac{1}{a}\boldsymbol{w}_k(\boldsymbol{v}')^{\mathrm{T}} \end{bmatrix}) - 2$$
$$= \text{rank}(\begin{bmatrix} A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}} \\ C_k - \frac{1}{a}\boldsymbol{w}'\boldsymbol{v}_k^{\mathrm{T}} \end{bmatrix}) + \text{rank}(\begin{bmatrix} A_k - \frac{1}{a}\boldsymbol{w}_k\boldsymbol{v}_k^{\mathrm{T}} & B_k - \frac{1}{a}\boldsymbol{w}_k(\boldsymbol{v}')^{\mathrm{T}} \end{bmatrix}).$$

So our conditions on rank of submatrices are also true for $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}$. By induction hypothesis, $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}$ has LU decompositions, say $A_{n-1} - \frac{1}{a}\boldsymbol{w}\boldsymbol{v}^{\mathrm{T}} = L_{n-1}U_{n-1}$. Then $A = \begin{bmatrix} 1 & \\ \frac{1}{a}\boldsymbol{w} & I_{n-1} \end{bmatrix}\begin{bmatrix} 1 & \\ & L_{n-1} \end{bmatrix}\begin{bmatrix} a & \boldsymbol{v}^{\mathrm{T}} \\ & U_{n-1} \end{bmatrix}$. So we are done.

Case 2:

Now suppose $a = 0$. Then our condition on A states that $1 = \text{rank}(a) + 1 \geq \text{rank}(\boldsymbol{v}^{\mathrm{T}}) + \text{rank}(\boldsymbol{w})$. So at least one of $\boldsymbol{v}, \boldsymbol{w}$ is the zero vector. So A must have a zero first column or a zero first row.

Suppose that $\boldsymbol{v} \neq \boldsymbol{0}$, then $\boldsymbol{w} = \boldsymbol{0}$. Say the i-th coordinate of $\boldsymbol{v}$ is the first non-zero coordinate of $\boldsymbol{v}$. Then by row operations, we can find a lower triangular matrix L such that $LA = \begin{bmatrix} 0 & \boldsymbol{v}^{\mathrm{T}} \\ \boldsymbol{0} & A' \end{bmatrix}$ where the i-th column of A' is entirely zero. Again we aim to show that A' satisfy the induction hypothesis and thus has an LU decomposition.

For any positive integer $k < n-1$, let $A' = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}$, and let $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_k & \boldsymbol{v}' \end{bmatrix}$. Here A_k is $k \times k$, and $\boldsymbol{v}_k \in \mathbb{R}^k$. Let $t = 1$ if $k \geq i$, and $t = 0$ if $k < i$. So we have the deduction

$$\begin{aligned} &\text{rank}(A_k) + k \\ &= \text{rank}(\begin{bmatrix} 0 & \boldsymbol{v}_k^{\mathrm{T}} \\ & A_k \end{bmatrix}) + k - t \\ &\geq \text{rank}(\begin{bmatrix} 0 & \boldsymbol{v}_k^{\mathrm{T}} \\ & A_k \\ & C_k \end{bmatrix}) + \text{rank}(\begin{bmatrix} 0 & \boldsymbol{v}_k^{\mathrm{T}} & (\boldsymbol{v}')^{\mathrm{T}} \\ & A_k & B_k \end{bmatrix}) - t - 1 \\ &= \text{rank}(\begin{bmatrix} A_k \\ C_k \end{bmatrix}) + \text{rank}(\begin{bmatrix} A_k & B_k \end{bmatrix}). \end{aligned}$$

So we are done.

Case 3:

Suppose $a = 0$ and $\boldsymbol{w} \neq \boldsymbol{0}$, then $\boldsymbol{v} = \boldsymbol{0}$. This case is the same as above, where we simply take transpose of everything.

Case 4:

Finally, suppose $a = 0$ and $\boldsymbol{v} = \boldsymbol{w} = \boldsymbol{0}$. Say the i-th row is the first row of A that is not entirely zero. Then let A' be obtained from A by swapping the i-th row and first row of A. Clearly $A = LA'$ where $L = I + \boldsymbol{e}_i\boldsymbol{e}_1^{\mathrm{T}} - \boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$, which is lower triangular. So A has LU decomposition if A' has LU decomposition. We aim to show that A' is in case 2 above, so we are done. Obviously A' has zero first column and non-zero first row, so all we need is to show that the submatrices of A' satisfies the rank condition.

For any positive integer $k \leq i$, the first k rows of A' has rank exactly 1, and the first k columns of A' has rank at most $k-1$ since the first column of A' is zero. So if $A' = \begin{bmatrix} A'_{11} & A'_{12} \\ A'_{21} & A'_{22} \end{bmatrix}$ where A'_{11} is $k \times k$, we have $\text{rank}(A'_{11}) + k \geq k \geq \text{rank}(\begin{bmatrix} A'_{11} & A'_{12} \end{bmatrix}) + \text{rank}(\begin{bmatrix} A'_{11} \\ A'_{21} \end{bmatrix})$.

For any positive integer $k > i$, then the first k columns of A' must have the same rank as the first k columns of A (since they differ by a row swap), and the first k rows of A' must also have the same rank as the first k rows of A, because they are made of the same rows. Finally, the upper left $k \times k$ block of A' would also have the same rank as the upper left $k \times k$ block of A, again because they are

made of the same rows. So if $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and $A' = \begin{bmatrix} A'_{11} & A'_{12} \\ A'_{21} & A'_{22} \end{bmatrix}$ where A_{11}, A'_{11} are $k \times k$, then we have $\text{rank}(A_{11}) = \text{rank}(A'_{11})$, $\text{rank}(\begin{bmatrix} A_{11} & A_{12} \end{bmatrix}) = \text{rank}(\begin{bmatrix} A'_{11} & A'_{12} \end{bmatrix})$, and $\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} A'_{11} \\ A'_{21} \end{bmatrix}$. So by the rank condition on A we have $\text{rank}(A_{11}) + k \geq \text{rank}(\begin{bmatrix} A_{11} & A_{12} \end{bmatrix}) + \text{rank}(\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix})$, which implies $\text{rank}(A'_{11}) + k \geq \text{rank}(\begin{bmatrix} A'_{11} & A'_{12} \end{bmatrix}) + \text{rank}(\begin{bmatrix} A'_{11} \\ A'_{21} \end{bmatrix})$. So A' is indeed in case 2, and it has an LU decomposition. □

4.11 How to find Kernel (Optional)

In this section we introduce a method to find the kernel of a matrix. Given a matrix A, we aim to find a basis for $\text{Ker}(A)$.

First of all, you should already know how to do this. $\text{Ker}(A)$ is exactly the solution set to $A\boldsymbol{x} = \mathbf{0}$. So this is basically Gaussian elimination.

Example 4.11.1. Suppose A, we first reduce it to RREF, say $\begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 4 \end{bmatrix}$. Then we see the second and fourth columns are free. So the solution set must be $\begin{bmatrix} * \\ x \\ * \\ z \end{bmatrix}$, where the star portion depends on x, z, and the variables x, z are free. Using these equations we see that the star portion must be $\begin{bmatrix} -2x - 3z \\ x \\ -4z \\ z \end{bmatrix}$. So the solution set is $\text{Ker}(A) = \{ \begin{bmatrix} -2x - 3z \\ x \\ -4z \\ z \end{bmatrix} : x, z \in \mathbb{R} \}$.

But what is a basis for this space? Note that $\begin{bmatrix} -2x - 3z \\ x \\ -4z \\ z \end{bmatrix} = x \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + z \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}$. So $\text{Ker}(A)$ is spanned by $\begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}$. We also see that these two vectors are linearly independent. Therefore they form a basis for $\text{Ker}(A)$.

So we are done. As you can see, the Gaussian elimination process is enough to find a basis for the kernel, where each basis vector corresponds to a free variable. ☺

Now look at the example above. The basis vectors are columns of $\begin{bmatrix} 2 & 3 \\ -1 & 0 \\ 0 & 4 \\ 0 & -1 \end{bmatrix}$, while the RREF of the matrix is $\begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 4 \end{bmatrix}$. You can easily see that the entries $2, 3, 4$ are in common, and in somewhat corresponding locations. Is there a way to utilize this correspondence directly?

There is indeed a way. First we force the RREF into a "square" matrix such that all pivots are on

the diagonal, i.e., we change $\begin{bmatrix}1&2&0&3\\0&0&1&4\end{bmatrix}$ into $\begin{bmatrix}1&2&0&3\\0&0&0&0\\0&0&1&4\\0&0&0&0\end{bmatrix}$. Then we substract by identity, and we get $\begin{bmatrix}0&2&0&3\\0&-1&0&0\\0&0&0&4\\0&0&0&-1\end{bmatrix}$. Now the non-zero columns are exactly the basis vectors for the kernel.

The rest of the section aims to prove that this is a valid method.

Definition 4.11.2. *Given a matrix A, we get its RREF X, delete all zero rows, then add in zero-rows so that all pivots of X are on the diagonal. We call the resulting matrix A' the square-RREF of A.*

Proposition 4.11.3. *If A' is the square-RREF of A, then* $\mathrm{Ker}(A) = \mathrm{Ker}(A')$.

Proof. Note that elementary row operations, deleting zero rows, or adding zero rows do not change the solution set. □

Lemma 4.11.4. *If A' is the square-RREF of A, then it is an upper triangular matrix. And for each index i, either the i-th column of A' is $\boldsymbol{e}_i$, or the i-th row of A' is zero.*

Proof. Before the proof, you may stare at $\begin{bmatrix}1&2&0&3\\0&0&0&0\\0&0&1&4\\0&0&0&0\end{bmatrix}$ for some intuition about this lemma.

Now we begin the proof. By construction of A', each row of A' is either entirely zero, or has the first non-zero entry on the diagonal. Therefore it is upper triangular.

First, note that elementary row operations, deleting zero rows, or adding zero rows do not change whether a column is pivotal or free. Therefore, A has i-th column pivotal iff $\mathrm{RREF}(A)$ has i-th column pivotal, iff A' has i-th column pivotal.

Now, since A' is upper triangular, if its (i,i)-entry is non-zero, then its i-th column cannot be a linear combination of previous columns. So it is pivotal. So the i-th column of A is a pivotal column. By definition of RREF, if the i-th column of A is pivotal, then the pivot entry in the i-th column of $\mathrm{RREF}(A)$ is 1, and all non-pivot entries of the i-th column of $\mathrm{RREF}(A)$ must be zero. Since we obtained A' from $\mathrm{RREF}(A)$ by adding zero rows, therefore the i-th column of A' is $\boldsymbol{e}_i$.

On the other hand, pivots in the i-th column of $\mathrm{RREF}(A)$ is moved down to the (i,i)-entry of A' by construction. So if the (i,i)-entry of A' is zero, then $\mathrm{RREF}(A)$ has no pivot in the i-th column. So when we create A' from $\mathrm{RREF}(A)$, the entire i-th row of A' is a newly inserted zero row. □

Lemma 4.11.5. *If A' is the square-RREF of A, we take its diagonal entries and make a diagonal matrix D. Then $A'D = D$ and $DA' = A'$.*

Proof. Note that by the last lemma, for each index i, either $A'\boldsymbol{e}_i = \boldsymbol{e}_i$ (which implies that the (i,i)-entry of D is 1), or $\boldsymbol{e}_i^{\mathrm{T}} A' = \boldsymbol{0}^{\mathrm{T}}$ (which implies that the (i,i)-entry of D is 0).

Suppose for an index i, $A'\boldsymbol{e}_i = \boldsymbol{e}_i$. Then the (i,i)-entry of D is 1. Since D is diagonal, this means $D\boldsymbol{e}_i = \boldsymbol{e}_i$ and $\boldsymbol{e}_i^{\mathrm{T}} D = \boldsymbol{e}_i^{\mathrm{T}}$. So $A'D\boldsymbol{e}_i = A'\boldsymbol{e}_i = \boldsymbol{e}_i = D\boldsymbol{e}_i$ and $\boldsymbol{e}_i^{\mathrm{T}} DA' = \boldsymbol{e}_i^{\mathrm{T}} A'$. In particular, D and $A'D$ have the same i-th column, and A' and DA' have the same i-th row.

Now suppose for an index i, $\boldsymbol{e}_i^{\mathrm{T}} A' = \boldsymbol{0}^{\mathrm{T}}$. Then the (i,i)-entry of D is 0. Since D is diagonal, this means $D\boldsymbol{e}_i = \boldsymbol{0}$ and $\boldsymbol{e}_i^{\mathrm{T}} D = \boldsymbol{0}^{\mathrm{T}}$. So $A'D\boldsymbol{e}_i = A'\boldsymbol{0} = \boldsymbol{0} = D\boldsymbol{e}_i$ and $\boldsymbol{e}_i^{\mathrm{T}} DA' = \boldsymbol{0}^{\mathrm{T}} A' = \boldsymbol{0}^{\mathrm{T}} = \boldsymbol{e}_i^{\mathrm{T}} A'$. In particular, D and $A'D$ have the same i-th column, and A' and DA' have the same i-th row.

Either way, all columns of D and $A'D$ are identical, so $D = A'D$. And all rows of A' and DA' are identical, so $A' = DA'$. □

Lemma 4.11.6. *If A' is the square-RREF of A, then $(A')^2 = A'$.*

Proof. $(A')^2 = A'A' = A'(DA') = (A'D)A' = DA' = A'$. □

Lemma 4.11.7. *If A' is the square-RREF of A, then* $\mathrm{Ran}(A' - I) \subseteq \mathrm{Ker}(A)$.

Proof. Since $(A')^2 = A'$, we see that $A'(A' - I) = 0$. So $\mathrm{Ran}(A' - I) \subseteq \mathrm{Ker}(A') = \mathrm{Ker}(A)$. □

Theorem 4.11.8. *If A' is the square-RREF of A, then* $\mathrm{Ran}(A' - I) = \mathrm{Ker}(A)$. *Furthermore, non-zero columns of* $\mathrm{Ran}(A' - I)$ *are linearly independent and hence form a basis for* $\mathrm{Ker}(A)$.

Proof. We already have $\mathrm{Ran}(A' - I) \subseteq \mathrm{Ker}(A)$. To see $\mathrm{Ran}(A' - I) = \mathrm{Ker}(A)$, it is enough to show that $\dim \mathrm{Ran}(A' - I) \geq \dim \mathrm{Ker}(A)$.

Suppose the domain of A is n-dimensional, and A has rank r. Then A' is $n \times n$, and it has r diagonal entires with value 1 and $n - r$ diagonal entries of value 0. So $A' - I$ has $n - r$ diagonal entries of value -1. So by the lemma below, $\mathrm{rank}(A' - I) \geq n - r = \dim \mathrm{Ker}(A)$.

Now for each i, $A' - I$ has zero i-th column iff A' has i-th column $\boldsymbol{e}_i$, iff the (i, i)-entry of A' is 1, iff the i-th column of A is pivotal. So there are a total of r such columns. So $A' - I$ has exactly $n - r$ non-zero columns, and their span has dimension $\dim \mathrm{Ran}(A' - I) \geq n - r$. Hence all these columns are linearly independent. □

Lemma 4.11.9. *If U is an upper triangular matrix with k non-zero diagonal entries, then* $\mathrm{rank}(U) \geq k$.

Proof. This is basically $\mathrm{rank}(\begin{bmatrix} A & B \\ & C \end{bmatrix}) \geq \mathrm{rank}(A) + \mathrm{rank}(C)$ plus mathematical induction. □

Chapter 5

Inner Product Space

Abstraction is fun and nice and useful, but they have missing pieces. In particular, we do not have lengths or angles.

Example 5.0.1. Consider the space of polynomials of degree at most one, $\mathcal{P}_1$. Then under the basis $(1, x)$, the elements 2 and $3x$ in $\mathcal{P}_2$ have coordinates $\begin{bmatrix}2\\0\end{bmatrix}, \begin{bmatrix}0\\3\end{bmatrix}$. In particular, they seem to have length 2 and 3, and it seems like they are perpendicular.

However, under the basis $(2+3x, 3x)$, then the elements 2 and $3x$ in $\mathcal{P}_2$ now have coordinates $\begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}$. Now they have length $\sqrt{2}$ and 1, and they are no longer perpendicular.

As you can see, "length" and "angles" are NOT defined for an abstract vector spaces.

Similarly, you also lose transpose. Consider $M : \mathcal{P}_1 \to \mathcal{P}_1$ which is $a + bx \mapsto -b + ax$. Uder the basis $(1, x)$, its matrix is $\begin{bmatrix}0 & -1\\1 & 0\end{bmatrix}$, whose inverse equal to its transpose. However, under the basis $(1 + x, x)$, then its matrix is $\begin{bmatrix}-1 & -1\\2 & 1\end{bmatrix}$, whose inverse is $\begin{bmatrix}1 & 1\\-2 & -1\end{bmatrix}$, which is not the transpose. ☺

In practice, we would often want to use length and angles. However, we might still want to reserve an option to change basis. This gives rise to the definition of an inner product space.

In the sense of abstractions, an inner product space is pretty much $\mathbb{R}^n$ where you forget about your basis, but you remembered the lengths and angles. If you then also forget about the lengths and angles, then you have reached an abstract vector space.

Conversely, you can also think of an inner product space as an abstract vector space with an extra structure, the inner product (e.g. dot product), which allows you to do lengths and angles. And you can think of $\mathbb{R}^n$ as an inner product space with an extra structure, i.e., a fixed basis.

In practice, we shall see that (finite dimensional) inner product spaces are pretty much just $\mathbb{R}^n$, but you are allowed to change basis as long as the length and angle structures are preserved in the process, i.e., orthonormal change of basis.

5.1 Fundamental Theorem of Linear Algebra

Here let us plug in the missing pieces in the case of $\mathbb{R}^n$, utilizing the dot product structure. Keep in mind that these statements will NOT have easy analogues for abstract linear maps. (The analogues exist, but they require much deeper knowledge to describe.)

We now go back to the world of $\mathbb{R}^n$. Let us refrain from all change of basis at the moment, and always use the standard basis. By doing so, we have lengths, angles, dot products, and transpose, and all that, i.e., the missing pieces are now all back.

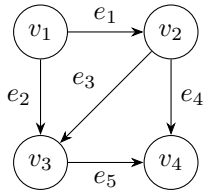
Given an $m \times n$ matrix A, there are four ***fundamental subspaces*** related to it.

1. The range of A, $\mathrm{Ran}(A)$;
2. The kernel of A, $\mathrm{Ker}(A)$;
3. The ***left range*** or ***row space*** of A, which is in fact $\mathrm{Ran}(A^{\mathrm{T}})$;
4. The ***left null space*** or ***left kernel*** of A, which is in fact $\mathrm{Ker}(A^{\mathrm{T}})$.

The names are somewhat revealing. For example, what is the subspace spanned by all rows of A? Well, these rows are just columns of A^{T}, so the space is simply $\mathrm{Ran}(A^{\mathrm{T}})$. You can also verify this: $\boldsymbol{b} \in \mathrm{Ran}(A^{\mathrm{T}})$ if and only if $\boldsymbol{b}^{\mathrm{T}} = \boldsymbol{x}^{\mathrm{T}}A$ for some $\boldsymbol{x}$. This explains the name of "left range". Similarly, for the left kernel, you can verify that $\boldsymbol{x} \in \mathrm{Ker}(A^{\mathrm{T}})$ if and only if $\boldsymbol{x}^{\mathrm{T}}A = \mathbf{0}^{\mathrm{T}}$.

These things have practical meanings in real life applications.

Example 5.1.1. Consider an electric circuit in Figure 5.1.1.



$$M = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{matrix}$$
$$\begin{matrix} v_1 & v_2 & v_3 & v_4 \end{matrix}$$

Figure 5.1.1: Graph of electric circuit and its incidence matrix M

Such a structure is mathematically called a directed graph, where we have some vertices, and some arrows (electrical wires) that goes from some vertex to another vertex. Here we think of the arrows as actual electrical wires. Given electrical potentials on the vertices, the difference between electrical potentials would induce voltages on wires. Then given the resistance of each wire, we can find out the electrical current on each wire. This is a classical situation in the study of electrical circuit.

(Why do we have arrows on the wires? Well, when we measure electrical currents, if the electricity flow in the arrow direction, we can declare this to be a "positive current", and if it flows in the opposite direction, we can declare this to be a "negative current".)

Whenever we have a graph, we can draw its incidence matrix M, where each column corresponds to a vertex, and each row corresponds to an arrow. The (i, j) entry of M is -1 if the i-th arrow starts at the j-th vertex, 1 if the i-th arrow ends at the j-th vertex, and 0 otherwise.

Now, what linear map does M represents? First of all, we see that M would send a generic input $\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$ to $\begin{bmatrix} u_2 - u_1 \\ u_3 - u_1 \\ u_3 - u_2 \\ u_4 - u_2 \\ u_4 - u_3 \end{bmatrix}$. In particular, if the inputs are electrical potentials at each vertex, then the output are their differences (i.e., voltages) on each wire!

$$\text{Electrical potentials on vertices} \xrightarrow{M} \text{Voltages on wires}$$

In particular, if we assign electrical potentials on vertices arbitrarily, then $\mathrm{Ran}(M)$ is the space of all possible voltages on wires in this circuit.

On the other hand, $\mathrm{Ker}(M)$ is the set of electrical potentials that gives zero voltages everywhere. Huh. Without any computation, just by physics intuition, you can see that $\mathrm{Ker}(M) = \{ \begin{bmatrix} u \\ u \\ u \\ u \end{bmatrix} : u \in \mathbb{R} \}$, i.e., when

all electrical potentials are identical. This should be the case as long as my graph for electrical circuit is a connected graph.

Now, what about the range and kernel of M^T?

(You might notice this formula in physics: voltage times current is power. We have many wires here. Let $\boldsymbol{v}$ be the vector recording the voltages on these wires, and let $\boldsymbol{c}$ be the vector recording the currents on these wires, then $\boldsymbol{v} \cdot \boldsymbol{c}$ is the total power of the circuit, a scalar! Hence if M acts on the voltage side, then M^{T} acts on the current side.)

Let $\boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}$ be a vector recording the electrical currents on each wire. Then note that $M^{\mathrm{T}}\boldsymbol{c} = \begin{bmatrix} -c_1 - c_2 \\ c_1 - c_3 - c_4 \\ c_2 + c_3 - c_5 \\ c_4 + c_5 \end{bmatrix}$. What are these output coordinates? Well, look at the second one. The vertex v_2 has wires e_1 flowing into it, and wires e_3, e_4 flowing out of it, and the second coordinate here is exactly recording that! In particular, the coordinates record the net inflow of electrical currents at each vertex. (Keep in mind that each c_i could be positive or negative, depending on the direction of the current.) Feel free to verify this yourself.

Electrical currents on wires $\xrightarrow{M^{\mathrm{T}}}$ Net current inflow on vertices

If you think about this, WITHOUT ouside influence, when the electrical flows are stable, we would expect each vertex to have zero net inflow, i.e., whatever flows into a vertex, it must then flow out. This is ***Kirchhoff's current law***, which says that stable currents must be in $\mathrm{Ker}(M^{\mathrm{T}})$.

If you compute $\mathrm{Ker}(M^{\mathrm{T}})$ (say via Gaussian elimination), this space is in fact spanned by $\begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 1 \end{bmatrix}$. You can see that the both corresponds to closed loops in the graph (where -1 means we are going against an arrow). So stable currents simply means currents travels in loops. For example, the vector $\begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix}$ is also in the kernel of M^{T}, and it also corresponds to a loop. (Note that it is the sum of the two loop vectors listed previously. Geometrically, this also makes sense. The two triangle loops would add up to the square loop, where the diagonal edge got cancelled, because the two triangle loops travel in opposite directions on this edge.)

There is also ***Kirchhoff's voltage law***, which states that possible voltages (i.e. elements of $\mathrm{Ran}(M)$)

must add up to zero on closed loops. Indeed, if $\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \in \mathrm{Ran}(M)$, then Gaussian elimination gives

$$\left[\begin{array}{cccc|c} -1 & 1 & 0 & 0 & u_1 \\ -1 & 0 & 1 & 0 & u_2 \\ 0 & -1 & 1 & 0 & u_3 \\ 0 & -1 & 0 & 1 & u_4 \\ 0 & 0 & -1 & 1 & u_5 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & -1 & 0 & 0 & -u_1 \\ 0 & -1 & 1 & 0 & u_2 - u_1 \\ 0 & 0 & -1 & 1 & u_1 - u_2 + u_4 \\ 0 & 0 & 0 & 0 & u_1 - u_2 + u_3 \\ 0 & 0 & 0 & 0 & u_2 - u_1 + u_5 - u_4 \end{array}\right].$$

So $\boldsymbol{u} \in \mathrm{Ran}(M)$ if and only if $u_1 - u_2 + u_3 = 0$ and $u_2 - u_1 + u_5 - u_4 = 0$, i.e., it is perpendicular to the two loop vectors $\begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$. Note that these two loop vectors would actually span all loops, i.e., they span the space $\mathrm{Ker}(M^{\mathrm{T}})$.

Wait, we have obtained a funny result here. Elements of $\mathrm{Ran}(M)$ are exactly those that are perpendicular to $\mathrm{Ker}(M^{\mathrm{T}})$.

The same relation is also present for $\mathrm{Ran}(M^{\mathrm{T}})$ and $\mathrm{Ker}(M)$. Given $\mathrm{Ran}(M^T)$, the elements of $\mathrm{Ran}(M^{\mathrm{T}})$ are possible net inflows at vertices. However, if currents flow out of a vertex, then they must flow into some other vertex. So if $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \in \mathrm{Ran}(M^{\mathrm{T}})$, then we should have $w_1 + w_2 + w_3 + w_4 = 0$, i.e., they should be perpendicular to $\mathrm{Ker}(M)$ (which is spanned by $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$). ☺

Our goal for this section, the fundamental theorem, is basically a theorem about relations between these fundamental subspaces. First of all, by rank-nullity, we already know that $\dim \mathrm{Ran}(A) + \dim \mathrm{Ker}(A) = n$ and $\dim \mathrm{Ran}(A^{\mathrm{T}}) + \dim \mathrm{Ker}(A^{\mathrm{T}}) = m$. What else do we know?

Proposition 5.1.2. *A and A^{T} have the same rank.*

Proof. Suppose the rank of A is r. Suppose the rank normal form of A is $RAC = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$ for invertible matrices R, C. Taking transpose, we have $C^{\mathrm{T}}A^{\mathrm{T}}R^{\mathrm{T}} = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$. And since $C^{\mathrm{T}}, R^{\mathrm{T}}$ are still invertible, we immediately see that the rank normal form of A^{T} is $\begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$, and hence its rank is r as well. □

(Note that this is essentially the old statement that "the number of effective equations = the number of dependent variables".)

(The porposition above also says the following fact: for a matrix A, if its columns span a space with dimension r, then its rows span a space with dimension r.)

Remark 5.1.3. *Be careful, you cannot say "pick basis and assume that $A = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$. This is becasue if you change basis arbitrarily, then A^{T} might not corresponds to the transpose anymore.*

Indeed, to go to the range normal form, A and A^{T} must actually use DIFFERENT change of basis. One needs R and C, while the other uses C^{T} and R^{T}.

So, given an $m \times n$ matrix A with rank r, then $r = \dim \mathrm{Ran}(A) = \dim \mathrm{Ran}(A^{\mathrm{T}})$, $\dim \mathrm{Ker}(A) = n - r$ and $\dim \mathrm{Ker}(A^{\mathrm{T}}) = m - r$. So we have all the dimensions.

But hey, note that $\mathrm{Ran}(A)$ and $\mathrm{Ker}(A)$ are NOT in the same space. The former is in the codomain, while the latter the domain. So in fact $\mathrm{Ran}(A)$ and $\mathrm{Ker}(A^{\mathrm{T}})$ are both in the codomain of A, while $\mathrm{Ran}(A^{\mathrm{T}})$ and $\mathrm{Ker}(A)$ are both in the domain of A. Do the subspaces in the same space have some special relation?

Yes they do. We need to introduce the concept of an orthogonal complement.

Definition 5.1.4. *We say subsapces V, W of $\mathbb{R}^n$ are orthogonal complements if they are complement as subspaces, and all vectors in V are orthogonal to all vectors in W.*

Proposition 5.1.5. $\mathrm{Ran}(A)$ *and* $\mathrm{Ker}(A^{\mathrm{T}})$ *are orthogonal complements of each other, and* $\mathrm{Ran}(A^{\mathrm{T}})$ *and* $\mathrm{Ker}(A)$ *are orthogonal complements of each other.*

Proof. It is enough to prove the first statement, because applying the first statement to A^{T} gives the second one.

Let us first show that they are orthogonal. Intuitively this is obvious. $\mathrm{Ker}(A^{\mathrm{T}})$ is made of vectors perpendicular to all rows of A^{T}, i.e., all columns of A. But $\mathrm{Ran}(A)$ is made of vectors that are linear combinations of these columns. So of course they are orthogonal subspaces.

To be more formal, pick any $\boldsymbol{v} \in \mathrm{Ran}(A)$, say $\boldsymbol{v} = A\boldsymbol{u}$ for some $\boldsymbol{u}$, and pick any $\boldsymbol{w} \in \mathrm{Ker}(A^{\mathrm{T}})$. Then $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = (A\boldsymbol{u})^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{u}^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{0} = 0$. So they are orthogonal subspaces.

Now let us show that they are complements. They are orthogonal, so obviously they have trivial intersection. Or to be more rigorous, suppose $\boldsymbol{v} \in \mathrm{Ran}(A) \cap \mathrm{Ker}(A^{\mathrm{T}})$, then $\boldsymbol{v}$ must be orthogonal to itself. Then $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 0$, and thus $\boldsymbol{v}$ has zero length. So it must be $\boldsymbol{0}$.

But the dimensions of the two subspaces add up to the dimension of the ambient space, so by inclusion-exclusion principle, their sum IS the ambient space. So they are complements.

More rigorously, if A is $m \times n$ rank r, then $\dim \mathrm{Ran}(A) = r$ while $\dim \mathrm{Ker}(A^{\mathrm{T}}) = m - \dim \mathrm{Ran}(A^{\mathrm{T}}) = m - r$. With trivial intersection, this means $\dim(\mathrm{Ran}(A) + \mathrm{Ker}(A^{\mathrm{T}})) = m = \dim \mathbb{R}^m$. Since the subspace has the same dimension with the whole space, we see that the two are the same. □

This immediately opens up a lot of interesting properties for orthogonal complements. Unlike regular complements, the orthogonal complement is unique.

Proposition 5.1.6. *For any subspace V of $\mathbb{R}^n$, then $V^\perp = \{\boldsymbol{w} \mid \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} = 0 \forall \boldsymbol{v} \in V\}$ is the only orthogonal complement of V. In particular, the orthogonal complement always exists and is unique.*

Proof. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ be a basis for V, and let $A = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_k \end{bmatrix}$ (note that A must be $n \times k$ with rank k). Now $\mathrm{Ran}(A) = V$.

But also, $\boldsymbol{v} \in \mathrm{Ker}(A^{\mathrm{T}})$ iff $A^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$ iff $\boldsymbol{v}$ is orthogonal to all $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ iff $\boldsymbol{v} \in V^\perp$. So $\mathrm{Ker}(A^{\mathrm{T}}) = V^\perp$. We see that $V, V^\perp$ are indeed orthogonal complements.

Now we go for uniqueness. Suppose W is also an orthogonal complement of V. Then on one hand, all vectors of W are orthogonal to all vectors of V, so $W \subseteq V^\perp$. On the other hand, since $W, V^\perp$ are both complement subspaces to V, they both have dimension $n - \dim(V)$. So they are the same. □

Corollary 5.1.7. $(V^\perp)^\perp = V$.

Proof. $V, (V^\perp)^\perp$ are both orthogonal complements of $V^\perp$, so by uniqueness we have equality. □

For example, the xy-plane and the z-axis in $\mathbb{R}^3$ are the unique orthogonal complements of each other. The orthogonal complement is a MUCH better analogy to the complements of subsets. In particular, we have the de Morgan law:

Proposition 5.1.8. $(V + W)^\perp = V^\perp \cap W^\perp, (V \cap W)^\perp = V^\perp + W^\perp$

Proof. Pick basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_a$ for V and basis $\boldsymbol{w}_1, \dots, \boldsymbol{w}_b$ for W. Let $A = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_a \end{bmatrix}$ and $B = \begin{bmatrix} \boldsymbol{w}_1 & \dots & \boldsymbol{w}_b \end{bmatrix}$. So $\operatorname{Ran}(A) = V, \operatorname{Ran}(B) = W, \operatorname{Ker}(A^{\mathrm{T}}) = V^{\perp}, \operatorname{Ker}(B^{\mathrm{T}}) = W^{\perp}$.

Now we already know sum and intersections are related to block matrices. We have $V+W = \operatorname{Ran}\begin{bmatrix} A & B \end{bmatrix}$ and $V^{\perp} \cap W^{\perp} = \operatorname{Ker}\begin{bmatrix} A^{\mathrm{T}} \\ B^{\mathrm{T}} \end{bmatrix}$. Then we immediately see that they are orthogonal complements because $\begin{bmatrix} A & B \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} A^{\mathrm{T}} \\ B^{\mathrm{T}} \end{bmatrix}$.

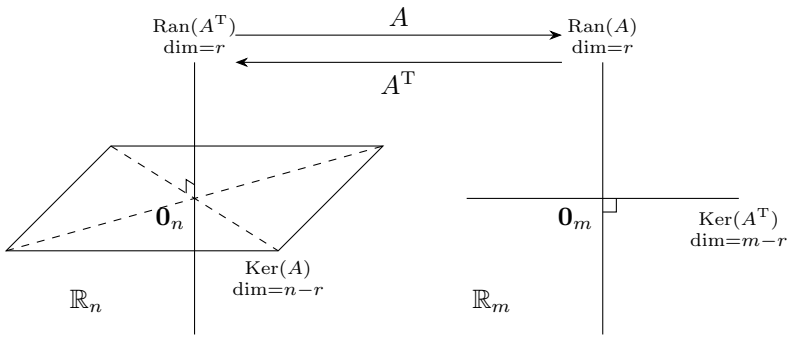
To see the second formula, just take orthogonal complement of the first formula on both sides. □

Hope you have fun. You can definitely see how the abstract results on rank-nullity, inclusion-exclusion principle, and subspace algebra mixes really well with concrete things in $\mathbb{R}^n$ such as block matrices and transpose.

So now we have the complete fundamental theorem of linear algebra (FTLA for short).

Theorem 5.1.9. *Given any $m \times n$ matrix A, then $\operatorname{Ran}(A)^{\perp} = \operatorname{Ker}(A^{\mathrm{T}})$ and $\operatorname{Ker}(A)^{\perp} = \operatorname{Ran}(A^{\mathrm{T}})$, and $\dim \operatorname{Ran}(A) = \dim \operatorname{Ran}(A^{\mathrm{T}})$.*

This is simply a collection of all previous results. We can see graphically here for the matrix $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ as the following



Here are some examples of how FTLA relates to real problems.

Example 5.1.10. Again consider Figure 5.1.1. Here recall that $\operatorname{Ker}(M^{\mathrm{T}})$ is the space of currents that gives zero net inflow at each vertex, i.e., currency that satisfy Kirchhoff's current law, while $\operatorname{Ran}(M)$ is the space of voltages that sums to zero on closed loops, i.e., voltages that satisfy Kirchhoff's voltage law. In a stable case, both laws should be satisfied. The total power of our electrical circuit is the dot product of the voltage vector and the current vector, which now must be zero, because the two subspaces are orthogonal complements. Huh.

Why is that? This is because we have no battery in our system. With no battery, the electrical circuit is essentially "dead", so there can be no power output.

Suppose on each edge their is a batter providing b_i extra voltage to flow in the positive direction. Then we have a battery vector $\boldsymbol{b}$. Assume that $\boldsymbol{x}$ is the electrical potential vector and $\boldsymbol{y}$ is the current vector, then we have $\boldsymbol{b} - M\boldsymbol{x} = R\boldsymbol{y}$, where R is the diagonal matrix whose diagonal entries are the resistence on each wire. Furtheremore we know that $M^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$ given Kirchhoff's current law. So we can solve $\begin{bmatrix} R & M \\ M^{\mathrm{T}} & O \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \mathbf{0} \end{bmatrix}$ to find stable solutions to our system. ☺

(The rest of this subsection is entirely optional.)

Let us insert a corollary before we proceed with the last example.

Corollary 5.1.11. $\operatorname{Ker}(A^{\mathrm{T}}A) = \operatorname{Ker}(A)$.

Proof. Suppose $\boldsymbol{v} \in \mathrm{Ker}(A)$. Then $A^{\mathrm{T}}A\boldsymbol{v} = A^{\mathrm{T}}\mathbf{0} = \mathbf{0}$, so $\boldsymbol{v} \in \mathrm{Ker}(A^{\mathrm{T}}A)$.

Conversely, $\boldsymbol{v} \in \mathrm{Ker}(A^{\mathrm{T}}A)$ means $A^{\mathrm{T}}A\boldsymbol{v} = \mathbf{0}$, which means $A\boldsymbol{v} \in \mathrm{Ker}(A^T)$. So in particular, $A\boldsymbol{v} \in \mathrm{Ker}(A^{\mathrm{T}}) \cap \mathrm{Ran}(A) = \{\mathbf{0}\}$. So $A\boldsymbol{v} = \mathbf{0}$, $\boldsymbol{v} \in \mathrm{Ker}(A)$.

Intuitively, the image coming from A would perfectly dodge the kernel of A^{T}. So $A^{\mathrm{T}}A$ kills exactly thos killed by A in the first step, and the second step A^{T} would fail to kill anything new. □

Example 5.1.12. (Optional)

Now, block row operation gives $\begin{bmatrix} R & M_G \\ O & -M_G^{\mathrm{T}}R^{-1}M_G \end{bmatrix}$, and then block column operation gives $\begin{bmatrix} R & O \\ O & -M_G^{\mathrm{T}}R^{-1}M_G \end{bmatrix}$. Note that since resistance are usually all non-zero, R has full rank, i.e., $\mathrm{rank}(R) = 5$. Since diagonal entries of R^{-1} are all positive, we have $R^{-1} = D^2$ for some diagonal D. Then $\mathrm{Ker}(M_G^{\mathrm{T}}R^{-1}M_G) = \mathrm{Ker}((DM_G)^{\mathrm{T}}(DM_G)) = \mathrm{Ker}(DM_G) = \mathrm{Ker}(M_G)$ because D is bijective. This is spanned by $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ and hence 1-dimensional, so $M_G^{\mathrm{T}}R^{-1}M_G$ has rank $4 - 1 = 3$. So $\begin{bmatrix} R & M_G \\ O & -M_G^{\mathrm{T}}R^{-1}M_G \end{bmatrix}$ is a 9×9 matrix of rank 8.

Note that $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ is in the kernel, which has dimension 1, so it in fact spans the kernel. This means our system, given battery inputs, will have unique solution up to shifting all electrical potential by the same constant. ☺

Example 5.1.13. (Optional)

Let us consider a surprise example from light out puzzle. This also serves as an example of how things work over the field $\mathbb{F}_2$, where we only have coeficient $0, 1$ and we think $1 + 1 = 0$.

Say we have a single row of five tiles. Each tile is either lit up or light out. Whenever we press a tile, then this tile and its adjacent tiles change status. At any given time, if we use 1 to represent a lit-up tile and 0 to represent a tile whose light is off, then the status of tiles is a vector in $(\mathbb{F}_2)^5$. (So we have $2^5 = 32$ possible initial configurations.) For example, $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ means the first, second and forth tile are llit up.

We can press some of the tiles, and this input translates to a vector $\boldsymbol{p} \in (\mathbb{F}_2)^5$. For example, $\boldsymbol{p} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ means we are pressing the first, second and forth tile. Then since each tile when pressed would change the status of itself and all adjacent tiles, we have a matrix $M = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$, and the total change of status would be $M\boldsymbol{p}$. If the lights started with status $\boldsymbol{v}$, then after pressing $\boldsymbol{p}$, we would end up with status $M\boldsymbol{p} + \boldsymbol{v}$. We would turn off the light iff $M\boldsymbol{p} = -\boldsymbol{v}$. Note that over $\mathbb{F}_2$, we have $1 = -1$ for scalars, so we are looking to solve $\boldsymbol{p}$ from $M\boldsymbol{p} = \boldsymbol{v}$.

Now, M is not invertible. You may check that $M\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix} = \mathbf{0}$. In fact, gaussian elimination gives $rref(M) = \begin{bmatrix}1&0&0&0&1\\0&1&0&0&1\\0&0&1&0&0\\0&0&0&1&1\\0&0&0&0&0\end{bmatrix}$. So $\operatorname{Ker}(M)$ is in fact spanned by $\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix}$.

Now by rank nullity theorem, this means we have $\dim \operatorname{Ran}(M) = 5 - \dim \operatorname{Ker}(M) = 4$. So not all initial status can be solved. In fact, since our matrix is symmetric, we have $\operatorname{Ran}(M) = \operatorname{Ran}(M^{\mathrm{T}}) = \operatorname{Ker}(M)^{\perp}$ by FTLA. So a status $\begin{bmatrix}a\\b\\c\\d\\e\end{bmatrix}$ can be solved iff it is "orthogonal" to $\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix}$, i.e., $a+b+d+e=0$. This means among the first, second, forth, fifth lights, only even number of them are lit up. Note that since $\operatorname{Ran}(M)$ is 4 dimensional, only $2^4 = 16$ initial status out of $|(\mathbb{F}_2)^5| = 2^5 = 32$ possible initial status are solvable.

Furthermore, given any solution $\boldsymbol{p}$ to an initial status, then $\boldsymbol{p} + \begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix}$ is also an solution. We have exactly two solutions to any initial status.

Check out the english wikipedia page of "Light Out (Game)", where they listed the two vectors the span the kernel of M when there are 5×5 array of tiles. (So the space is $(\mathbb{F}_2)^{25}$ instead of $(\mathbb{F}_2)^5$. For cosmetic reasons, they write the vectors N_1, N_2 like matrices, when they are in fact a column vector in $(\mathbb{F}_2)^{25}$.) So any initial status can have 4 solutions, and only 2^{23} initial cases are solvable out of a total of 2^{25}. ☺

5.2 Inner Product Space

Now abstract vector spaces are sometimes annoying. They have no angle, no dot product, no transpose. We want these things sometimes. If abstraction is forgetting, then sometimes it seems that we have forgotten too much.

So we need some intermediate structure. This is inner product space, i.e., an abstract vector space where you can do "dot product", yet there is no standard basis.

Definition 5.2.1. *An inner product space is an abstract vector space V over $\mathbb{R}$ equipped with an **inner product**, which is an operation $\langle -,-\rangle$ that sends two vectors $\boldsymbol{v},\boldsymbol{w}$ of V into a scalar $\langle \boldsymbol{v},\boldsymbol{w}\rangle$ in $\mathbb{R}$, such that*

1. *(Symmetric) $\langle \boldsymbol{v},\boldsymbol{w}\rangle = \langle \boldsymbol{w},\boldsymbol{v}\rangle$ for all $\boldsymbol{v},\boldsymbol{w}\in V$.*
2. *(Bilinear) $\langle a\boldsymbol{u}+b\boldsymbol{v},\boldsymbol{w}\rangle = a\langle \boldsymbol{u},\boldsymbol{w}\rangle + b\langle \boldsymbol{v},\boldsymbol{w}\rangle$ for all $a,b\in\mathbb{R}, \boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in V$. And same thing for the other side.*
3. *(Positive Definite) $\langle \boldsymbol{v},\boldsymbol{v}\rangle \geq 0$ with equality iff $\boldsymbol{v}=\mathbf{0}$.*

The inner product here essentially serves as an abstract analogue to the "dot product". If you think dot product, then all these properties are obviously needed. The last one, in particular, allows us to define the ***length*** of an abstract vector $\boldsymbol{v}$ in an inner product space as $\|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v},\boldsymbol{v}\rangle}$.

So, why is this enough? Remember, a major goal of this is to have angles, because orthogonality is very useful in many situations (and also required for building FTLA in abstract spaces). Traditionally in $\mathbb{R}^n$, if

two vectors have angle θ, then we can work out θ from the fact that $\cos\theta = \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$. (Surely you've seen this formula, at least for $\mathbb{R}^2$?) So we would to define angles via the formula $\cos\theta = \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$. But for this to work at all, we need the following famous inequality.

Theorem 5.2.2 (Cauchy-Schwarz Inequality)**.** *For any $\boldsymbol{v}, \boldsymbol{w}$ in an inner product space, we have $\langle \boldsymbol{v}, \boldsymbol{w}\rangle^2 \leq \langle \boldsymbol{v}, \boldsymbol{v}\rangle\langle \boldsymbol{w}, \boldsymbol{w}\rangle$, with equality iff $\boldsymbol{v}, \boldsymbol{w}$ are colinear.*

There are a million proofs of this inequality, so you are very welcome to track them down online. The english wikipedia has two proofs at least. Here I present my favorite, not because it is simple, but because it is intuitive.

Definition 5.2.3. *We write $\boldsymbol{v} /\!\!/ \boldsymbol{w}$ if one is a multiple of the other. We write $\boldsymbol{v} \perp \boldsymbol{w}$ if $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0$ in our inner product space.*

Lemma 5.2.4. *Given any $\boldsymbol{v}, \boldsymbol{w}$, let $P_{\boldsymbol{v}}(\boldsymbol{w})$ be $\frac{\langle \boldsymbol{v},\boldsymbol{w}\rangle}{\langle \boldsymbol{v},\boldsymbol{v}\rangle}\boldsymbol{v}$. (We call $P_{\boldsymbol{v}}$ the **projection to $\boldsymbol{v}$**.) Then $\boldsymbol{v} /\!\!/ P_{\boldsymbol{v}}(\boldsymbol{w})$ and $\boldsymbol{v} \perp \boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w})$.*

Proof. $\boldsymbol{v} /\!\!/ P_{\boldsymbol{v}}(\boldsymbol{w})$ is obvious. By definition $P_{\boldsymbol{v}}(\boldsymbol{w}) = \frac{\langle \boldsymbol{v},\boldsymbol{w}\rangle}{\langle \boldsymbol{v},\boldsymbol{v}\rangle}\boldsymbol{v}$ is a multiple of $\boldsymbol{v}$.

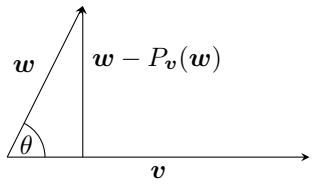
Now let us move on to perpendicularity.

$$\langle \boldsymbol{v}, \boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w})\rangle = \langle \boldsymbol{v}, \boldsymbol{w} - \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\boldsymbol{v}\rangle = \langle \boldsymbol{v}, \boldsymbol{w}\rangle - \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\langle \boldsymbol{v}, \boldsymbol{v}\rangle = \langle \boldsymbol{v}, \boldsymbol{w}\rangle - \langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0.$$

□

Where did we get the formula $P_{\boldsymbol{v}}(\boldsymbol{w}) = \frac{\langle \boldsymbol{v},\boldsymbol{w}\rangle}{\langle \boldsymbol{v},\boldsymbol{v}\rangle}\boldsymbol{v}$? Well, we literally borrowed this from the case of $\mathbb{R}^n$, and we simply swapped the dot products with inner products. And (unsurprisingly) it works in exactly the same way.

Proof of Cauchy-Schwarz. Intuitively, the angle between $\boldsymbol{v}, \boldsymbol{w}$ can be found by looking at the right triangle made by $\boldsymbol{w}, P_{\boldsymbol{v}}(\boldsymbol{w}), \boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w})$. We need all three edges to be "well-defined vectors", in which case I must have my well-defined angle. Then Cauchy-Schwarz would be true.



In this setting, "well-defined vectors" just mean they don't violate the positive definiteness. It turns out that we only need to check the last one. (This also the edges most related to the measurement of the angle.) We have

$$\begin{aligned}
0 \leq & \langle \boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w}), \boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w})\rangle \\
= & \langle \boldsymbol{w}, \boldsymbol{w}\rangle - 2\langle \boldsymbol{w}, \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\boldsymbol{v}\rangle + \langle \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\boldsymbol{v}, \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\boldsymbol{v}\rangle \\
= & \langle \boldsymbol{w}, \boldsymbol{w}\rangle - 2\frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}\langle \boldsymbol{w}, \boldsymbol{v}\rangle + (\frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle})^2\langle \boldsymbol{v}, \boldsymbol{v}\rangle \\
= & \langle \boldsymbol{w}, \boldsymbol{w}\rangle - \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle^2}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}.
\end{aligned}$$

Rearrange terms and we are done. It is also easy to see that we have equality iff $\boldsymbol{w} = P_{\boldsymbol{v}}(\boldsymbol{w})$ iff $\boldsymbol{w} /\!\!/ \boldsymbol{v}$. □

This establishes angles in an inner product space. We say two vectors $\boldsymbol{v}, \boldsymbol{w}$ has ***angle*** $\arccos(\frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|})$. Here the range of arccos function is $[0, \pi]$. It is easy to check that two non-zero vectors have angle zero iff they are in the same direction, angle π iff they are in the opposite direction, and angel $\frac{\pi}{2}$ iff they are orthogonal. All is as expected.

If you play some more, you can get more things that you would expect. Here are some optional things you might want to try or skip.

Proposition 5.2.5 (Triangle inequality)**.** $\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\|$

Proof. Square both sides, then compare both sides. You will see that this is just Cauchy-Schwarz in disguise.

$$\|\boldsymbol{v} + \boldsymbol{w}\|^2 - (\|\boldsymbol{v}\| + \|\boldsymbol{w}\|)^2 = \langle \boldsymbol{v} + \boldsymbol{w}, \boldsymbol{v} + \boldsymbol{w}\rangle - \langle \boldsymbol{v}, \boldsymbol{v}\rangle - \langle \boldsymbol{w}, \boldsymbol{w}\rangle - 2\|\boldsymbol{v}\|\|\boldsymbol{w}\| = 2\langle \boldsymbol{v}, \boldsymbol{w}\rangle - 2\|\boldsymbol{v}\|\|\boldsymbol{w}\| \leq 0.$$

The last inequaltiy is due to Cauchy-Schwarz. □

Proposition 5.2.6 (Pythagorean Theorem (Gou Gu Ding Li))**.** $\|\boldsymbol{v} + \boldsymbol{w}\|^2 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2$ *if and only if* $\boldsymbol{v} \perp \boldsymbol{w}$*.*

Proof. In general, we have

$$\|\boldsymbol{v} + \boldsymbol{w}\|^2 = \langle \boldsymbol{v} + \boldsymbol{w}, \boldsymbol{v} + \boldsymbol{w}\rangle = \langle \boldsymbol{v}, \boldsymbol{v}\rangle + \langle \boldsymbol{w}, \boldsymbol{w}\rangle + 2\langle \boldsymbol{v}, \boldsymbol{w}\rangle.$$

Now orthogonality gives $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0$, so we are done. □

Here is a very important corollary of Pythagorean Theorem.

Proposition 5.2.7. *If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are mutually orthogonal (i.e. pairwise orthogonal) non-zero vectors in an inner product space V, then they are linearly independent.*

Proof. Suppose $\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$. Then $0 = \|\sum a_i \boldsymbol{v}_i\|^2 = \sum \|a_i \boldsymbol{v}_i\|^2$ because the vectors involved are mutually orthogonal. So $0 = \sum a_i^2 \|\boldsymbol{v}_i\|^2$. Now since all $\|\boldsymbol{v}_i\|^2$ are positive, we must have all $a_i = 0$.

Here is an alternative proof NOT using Pythagorean theorem. If $\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$, we simply take inner product on both sides with $\boldsymbol{v}_j$. Then we would get $a_j \|\boldsymbol{v}_j\|^2 = 0$, hence $a_j = 0$. Since this is true for all j, we are done again. □

Remark 5.2.8. *Recall that for a collection of vectors, pairwise independent does NOT imply collective independence.*

However, pairwise orthogonal would indeed imply collective independence. This is amazingly useful.

THe next is not an inequality, but an equality.

Proposition 5.2.9 (Polarization identity)**.** $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \frac{1}{2}(\|\boldsymbol{v} + \boldsymbol{w}\|^2 - \|\boldsymbol{v}\|^2 - \|\boldsymbol{w}\|^2)$

Proof. Straightforward to verify. □

This identity has an important meaning: it means that any length structure would in fact induce an angle structure. This has some interesting ramifications for many study of geometry, and also the study of infinite dimensional spaces, but sadly we do not explore them in this class.

Nevertheless, applying this polariztion identity in the context of dot product, we have some very interesting results.

Lemma 5.2.10. *For $m \times n$ matrices A, B, if $\boldsymbol{v}^{\mathrm{T}} A \boldsymbol{w} = \boldsymbol{v}^{\mathrm{T}} B \boldsymbol{w}$ for all $\boldsymbol{v} \in \mathbb{R}^m, \boldsymbol{w} \in \mathbb{R}^n$, then $A = B$.*

Proof. By assumption we have $\boldsymbol{e}_i^{\mathrm{T}} A \boldsymbol{e}_j = \boldsymbol{e}_i^{\mathrm{T}} B \boldsymbol{e}_j$. So A, B have identical entries. □

Lemma 5.2.11. *For $n \times n$ symmetric matrices A, B, if $\boldsymbol{v}^{\mathrm{T}} A \boldsymbol{v} = \boldsymbol{v}^{\mathrm{T}} B \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{R}^n$, then $A = B$.*

Proof. Look at this adaptation of the famous polarization identity on inner products.

$$\boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w} = \frac{1}{2}[(\boldsymbol{v}+\boldsymbol{w})^{\mathrm{T}}A(\boldsymbol{v}+\boldsymbol{w}) - \boldsymbol{v}^{\mathrm{T}}A\boldsymbol{v} - \boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w}].$$

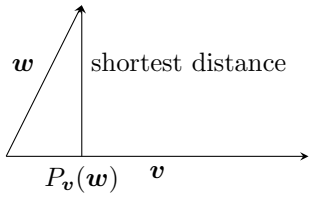
Hence if $\boldsymbol{v}^{\mathrm{T}}A\boldsymbol{v} = \boldsymbol{v}^{\mathrm{T}}B\boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{R}^n$, then $\boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w} = \boldsymbol{v}^{\mathrm{T}}B\boldsymbol{w}$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$. So $A = B$. □

As you have seen here, we have a new way to think about a symmetric matrix A. In the past, we have been treating matrices as linear maps. However, given a symmetric matrix A, you can also think of it as a ***bilinear map*** that sends a pair of vectors $\boldsymbol{v}, \boldsymbol{w}$ to the scalar $\boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w}$. The polarization identity is true for all symmetric bilinear maps, and hence we have the results above.

The last result is not very useful in this class, but very useful in real life applications.

Proposition 5.2.12 (Projection minimizes distance)**.** *$\|\boldsymbol{w} - x\boldsymbol{v}\| \geq \|\boldsymbol{w} - P_{\boldsymbol{v}}(\boldsymbol{w})\|$ for all $x \in \mathbb{R}$, with equality if and only if $x\boldsymbol{v} = P_{\boldsymbol{v}}(\boldsymbol{w})$. In short, on the line spanned by $\boldsymbol{v}$, $P_{\boldsymbol{v}}(\boldsymbol{w})$ is the closest to $\boldsymbol{w}$.*

Proof. What the proposition is saying is very intuitive. On the line spanned by $\boldsymbol{v}$, the closest point to the point $\boldsymbol{w}$ is the projection of $\boldsymbol{w}$ to the line.



Consider all possible distance from $\boldsymbol{w}$ to the line spanned by $\boldsymbol{v}$. The square of this distance is, for some unknown x, $\|\boldsymbol{w} - x\boldsymbol{v}\|^2 = \langle \boldsymbol{w}, \boldsymbol{w}\rangle - 2x\langle \boldsymbol{w}, \boldsymbol{v}\rangle + x^2\langle \boldsymbol{v}, \boldsymbol{v}\rangle$. Take derivative and find minimum, and we see that the minimum is reached exactly when $x = \frac{\langle \boldsymbol{w}, \boldsymbol{v}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}$. So we are done.

In fact, plug this in and find minimum, and it would be $\langle \boldsymbol{w}, \boldsymbol{w}\rangle - \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle^2}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}$. Now this minimum is still a square length, so it is non-negative. Then this would give a proof of Cauchy-Schwartz. (Most textbook uses this proof.) □

Let us see some example of abstract inner product space.

Example 5.2.13.

1. $\mathbb{R}^n$ with dot product. Obvious. We also call this inner product space the ***Euclidean space***.
2. $\mathbb{R}^n$ where we define $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^{\mathrm{T}}D\boldsymbol{w}$, and D is a fixed diagonal matrix with positive diagonal entries.
3. In general, if we define $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w}$ on $\mathbb{R}^n$ for some fixed matrix A, then symmetricity of $\langle -, -\rangle$ is just the same as the symmetricity of A. Bilinearity is immediate. We say a symmetric matrix is ***positive definite*** if $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w}$ is positive definite, and therefore an inner product. Some example is $A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$. You can check that $\begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = x^2 + 4xy + 5y^2 = (x+2y)^2 + y^2 \geq 0$, and with equality iff $x = y = 0$.
4. Consider $\mathbb{R}^4$ with $D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$. This matrix is NOT positive definite, because $\boldsymbol{e}_4^{\mathrm{T}}D\boldsymbol{e}_4 = -1 < 0$. So $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^{\mathrm{T}}D\boldsymbol{w}$ is NOT an inner product, merely a ***symmetric bilinear form***. This structure is still important though. For example, we see that $\left\| \begin{bmatrix} x \\ y \\ z \\ t \end{bmatrix} \right\| = \sqrt{x^2 + y^2 + z^2 - t^2}$, and this is the length

structure used for special relativity. In general, spaces with bilinear forms of all kinds are useful and important. We only study inner product space because it is easier and intuitive.

☺

Let us see some more abstract example.

Example 5.2.14. 1. Let V be the space of continuous real functions defined on the interval $[0, 2\pi]$. Define $\langle f, g\rangle = \int_0^{2\pi} f(x)g(x)\,\mathrm{d}x$. Check and see that now V is an inner product space. Also check and see that $\sin(x), \cos(x), \sin(2x), \cos(2x), \dots$ are all orthogonal to each other. So we are giving V a structure where different frequencies are orthogonal to each other. This is the starting point of Fourier analysis.

2. Let V be the space of random variables whose value is a random number in the interval $[0, 1]$. For any random variable X, we can define its expected value (or average value) as $\mathbb{E}(X)$. Then we can define its variance as $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$. This measures how much X vary from being constant. Finally, we can (using polarization identity) define covariance as $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$. This being positive means X, Y are usually large together or small together, i.e., they are most likely to change in the same direction. This being negative means they are most likely to change in opposite directions. Hence the name "covariance". Now, V with covariance is NOT an inner product space. You can check that Cov is bilinear and symmetric, and $\mathrm{Cov}(X, X) \geq 0$, but $\mathrm{Cov}(X, X) = 0$ does NOT imply $X = 0$. Rather, X could simply be any constant. So close. We say that Cov is NOT positive definite, but ***positive semi-definite***. It is a ***semi-inner product***, and V with Cov is merely a ***semi-inner product space***.

3. Nonetheless, people define correlation between two random variables as $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$. What is this? This is the cosine of the angle! Note that $\rho = 1$ if and only if $X = kY + b$ for some $k > 0$, and $\rho = -1$ if and only if $X = kY + b$ for some $k < 0$. So basically, in a semi-inner product space, things behave pretty much the same as an inner product space, as long as you ignore vectors whose length is zero (constant variables, in this case).

4. Let W be the space of random variables whose value is a random number in the interval $[0, 1]$, with expected value zero. NOW W and covariance is a genuin inner product space. In statistics, when we get a bunch of data, often people would like to "center" the data, i.e., substract each data by the average. Why do this? Because a data is in V which is only a semi-inner product space, but after we center the data, it now lives in W which is a genuin inner product space.

5. This is just an observation here. Suppose two random variables are INDEPENDENT, then of course their covariance is zero, i.e., they are orthogonal in the (semi-)inner product structure. Be careful though, the converse is not true. Zero covariance variables might still be dependent, when the dependence is non-linear. Nevertheless, when you see data that are orthogonal, it does not hurt to pause and think if they are independent or not.

☺

Here is a more interesting example, in lieu of previous discussions on infinite dimenisional spaces in the last chapter.

Example 5.2.15. This is again an optional example. It aims to further explain why $\boldsymbol{e}_i$ in the sequence space fail to be a basis.

Let us consider space S of all real sequences. Consider $\boldsymbol{e}_i \in V$ where $\boldsymbol{e}_i$ is the sequence $(0, \dots, 0, 1, 0 \dots)$ where the 1 is at the i-th location. We already know that these vectors cannot span $\mathbb{R}^{\mathbb{N}}$. What can they actually span?

Algebraically, only finite linear combinations are allowed. And you can quickly see that finite linear combinations of these $\boldsymbol{e}_i$ must only have finitely many non-zero terms, i.e., they are actually finite sequences. Let us call this S_{fin}. This is also an infinite dimensional vector space, and it is strictly smaller than the

space S. In particular, sequences such as $(1, \frac{1}{2}, \frac{1}{3}, \dots)$ are NOT spanned by these $\boldsymbol{e}_i$, as they have infinitely many terms.

However, let us consider adding some geometry, by giving a length/angle structure to sequences. For two sequences $(a_0, a_1, \dots)$ and $(b_0, b_1, \dots)$, how would you do a "dot product"? Well, one obvious way to do this is to define $\langle (a_0, a_1, \dots), (b_0, b_1, \dots) \rangle = \sum_{n \in \mathbb{N}} a_n b_n$, simply the infinite analogue of a dot product.

However, consider the sequence $(1, 1, \dots)$. Then what is the length of this sequence? It is $\sqrt{\langle (1, 1, \dots), (1, 1, \dots) \rangle} = \sqrt{\infty} = \infty$. Hey! That is bad.

As a result, consider the space ℓ^2 of all sequences $(a_0, a_1, \dots)$, such that $\sum a_n^2 < \infty$. You can verify yourself that this is indeed a subspace of S, and we can make it an inner product space by using the dot product above.

Now consider $(1, \frac{1}{2}, \frac{1}{3}, \dots)$ again. It is not a linear combination of those $\boldsymbol{e}_i$. However, we can "approximate" this sequence using elements of S_{fin}, by looking as the sequences $(1, 0, \dots), (1, \frac{1}{2}, 0, \dots), \dots$. In particular, $(1, \frac{1}{2}, \frac{1}{3}, \dots)$ is a limit of linear combinations of these $\boldsymbol{e}_i$ under our new length structure! Indeed, we can compute and see that

$$\lim_{n \to \infty} \| (1, \frac{1}{2}, \frac{1}{3}, \dots) - \sum_{k=0}^{n} \frac{1}{k} \boldsymbol{e}_k \| = \lim_{n \to \infty} \sum_{k=n+1}^{\infty} \frac{1}{k^2} = 0.$$

So even though we cannot obtain all vectors using merely linear combinations, we can obtain all vectors in ℓ^2 using linear combinations and geometry (limits)!

You may also see that sequences such as $(3, 1, 4, 1, 5, 9, 2, 6, \dots)$ are still NOT obtainable using these $\boldsymbol{e}_i$. The above limit will fail to converge to zero for this sequence still. ℓ^2 is the "geometric span" of these $\boldsymbol{e}_i$, under our given length/angle structure.

You need to be careful though. Geometry depends on your definition of inner products. For example, let us now define a new inner product on sequence spaces $\langle (a_0, a_1, \dots), (b_0, b_1, \dots) \rangle = \sum_{n \in \mathbb{N}} \frac{1}{2^n} a_n b_n$. Now you can verify that sequences such as $(3, 1, 4, 1, 5, 9, 2, 6, \dots)$ are now also limits of linear combinations of these $\boldsymbol{e}_i$! The "geometric span" of these $\boldsymbol{e}_i$ will be much larger than before. On the other hand, if you define a new inner product as $\langle (a_0, a_1, \dots), (b_0, b_1, \dots) \rangle = \sum_{n \in \mathbb{N}} n a_n b_n$, then even the sequence $(1, \frac{1}{2}, \frac{1}{3}, \dots)$ will fail to be a limit of linear combinations of these $\boldsymbol{e}_i$.

The point is this: for infinite linear combinations, convergence depends on geometry, which usually depends on your definition of inner products. ☺

5.3 (Optional) Adjoint: Abstract "transpose"

To conclude this section, let us recall that we have another goal with the introduction of an inner product structure. I.e., transpose. How to relate transpose to inner products?

Proposition 5.3.1. *Given a linear map $L : V \to W$ between inner product spaces, there is a UNIQUE linear map $L^* : W \to V$ such that $\langle L\boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{v}, L^* \boldsymbol{w} \rangle$. (We call this the **adjoint** of A.)*

Before we prove this, think about what would happen in a Euclidean space ($\mathbb{R}^n$ with dot product). There we see that $\boldsymbol{v}^{\mathrm{T}} A^{\mathrm{T}} \boldsymbol{w} = (A\boldsymbol{v})^{\mathrm{T}} \boldsymbol{w}$. So L^* is just basically the transpose of L, but generalized into an abstract setting.

Let us first establish this "transpose" on vectors.

Lemma 5.3.2 (Vector Transpose)**.** *Given any linear map α from an inner product space V to $\mathbb{R}$, there is a UNIQUE vector $\boldsymbol{w}$ such that $\alpha(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v} \rangle$ for all $\boldsymbol{v}$. We write $\alpha = \langle \boldsymbol{w}, - \rangle$ (or $\alpha = \langle \boldsymbol{w} |$, as physicists prefer).*

(Physisists also like to call $\langle -, - \rangle$ as a "braket". So they call this evaluation thing $\langle \boldsymbol{w} |$ a "bra", and a regular vector $\boldsymbol{v}$ a "ket". They also sometimes write $| \boldsymbol{v} \rangle$ for regular vectors.)

In short, transpose on a vector means "this guy is no longer a vector. Instead, it is now waiting to EAT another vector." This duality between linear evaluations and vectors are also sometimes called the (finite dimensional) Riesz representation theorem, i.e., any linear evaluation can be "represented" by a vector.

Proof of Lemma. Let V^* be the space of linear maps from V to $\mathbb{R}$. (Check yourself that this is indeed a vector space.) Define the map $\langle -| : V \to V^*$ by sending $\boldsymbol{w}$ to $\langle \boldsymbol{w}|$, the corresponding linear evaluations. Our goal is to show that this is bijective.

To see injectivity, suppose $\langle \boldsymbol{v}|$ is the zero map. Then $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0$ for all $\boldsymbol{w} \in V$. In particular, $\langle \boldsymbol{v}, \boldsymbol{v}\rangle = 0$, so we have $\boldsymbol{v} = \boldsymbol{0}$. (Intuitively, only the zero vector can be perpendicular to all vectors.) This shows that $\langle -|$ is injective.

Now, note that if $n = \dim V$, then V^* is the space of linear maps from a n-dimensional space to a 1-dimensional space. By picking basis, this is the space of $1 \times n$ matrices. So $\dim V^* = n$. But $\mathrm{Ran}(\langle -|)$ is also n dimensional by rank-nullity, so $\langle -|$ is surjective. □

Proof of proposition. For any $\boldsymbol{w} \in W$, consider $\langle L(-), \boldsymbol{w}\rangle : V \to \mathbb{R}$. This is a linear evaluation! So in fact, it is $\langle \boldsymbol{x}|$ for a unique $\boldsymbol{x} \in V$. We define $L^*(\boldsymbol{w}) = \boldsymbol{x}$ in this mannor, so we now have a well-defined map $L^* : W \to V$. Also we immediately have $\langle L\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{v}, L^*\boldsymbol{w}\rangle$ by construction.

We only need to show that L^* is linear. Note that $\langle L^*(a\boldsymbol{u} + b\boldsymbol{w}), \boldsymbol{v}\rangle = \langle a\boldsymbol{u} + b\boldsymbol{w}, L\boldsymbol{v}\rangle = a\langle \boldsymbol{u}, L\boldsymbol{v}\rangle + b\langle \boldsymbol{w}, L\boldsymbol{v}\rangle = a\langle L^*\boldsymbol{u}, \boldsymbol{v}\rangle + b\langle L^*\boldsymbol{w}, \boldsymbol{v}\rangle = \langle aL^*\boldsymbol{u} + bL^*\boldsymbol{w}, \boldsymbol{v}\rangle$, so we see that as linear evaluations we have $\langle L^*(a\boldsymbol{u} + b\boldsymbol{w})| = \langle aL^*\boldsymbol{u} + bL^*\boldsymbol{w}|$. By injectivity of $\langle -|$, we have $L^*(a\boldsymbol{u} + b\boldsymbol{w}) = aL^*\boldsymbol{u} + bL^*\boldsymbol{w}$ as desired.

Finally, let us show that such L^* is unique. Suppose T is another linear map suc that $\langle L\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{v}, T\boldsymbol{w}\rangle$, then $\langle T\boldsymbol{w}, \boldsymbol{v}\rangle = \langle L^*\boldsymbol{w}, \boldsymbol{v}\rangle$ for all $\boldsymbol{v}, \boldsymbol{w}$. So $T\boldsymbol{w} = L^*\boldsymbol{w}$ again by injectivity of $\langle -|$. □

So we have a well-defined concept of "transpose" for maps between inner product spaces. Here are some very nice corollaries.

Corollary 5.3.3. $(A^*)^* = A$.

Proof. Note that $\langle \boldsymbol{v}, (A^*)^*\boldsymbol{w}\rangle = \langle A^*\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{v}, A\boldsymbol{w}\rangle$. So $A, (A^*)^*$ are both adjoint of A^*. By the uniqueness of adjoint, we have $(A^*)^* = A$. □

Corollary 5.3.4. *If $\langle \boldsymbol{v}, A\boldsymbol{w}\rangle = \langle \boldsymbol{v}, B\boldsymbol{w}\rangle$ for all $\boldsymbol{v}, \boldsymbol{w}$, then we have $A = B$.*

Proof. Again they are both adjoint of A^*. □

5.4 Gram Matrices and Cholesky decomposition

We make a bold statement here. We claim that, as a matter of fact, all finite dimensional inner product spaces are Euclidean. (After picking the right basis, your inner product is actually the same as the dot product.)

But first let us see a matrix phenomenon.

Example 5.4.1. Consider $\begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. After calculation, you shall see that this is $a_1b_1 + 2a_1b_2 + 3a_2b_1 + 4a_2b_2$. Hey, the (i,j)-entry of $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ is exactly the coefficient for a_ib_j. This is NOT a coincidence.

Think about this. When we multiply $\begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, a_1 will multiply entries in the first row, while a_2 will multiply entries in the second row. And when we multiply $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, b_1 will multiply entries in the first column, and b_2 will multiply entries in the second column. All in all, who will multiply the (i,j)-entry of the matrix? Well, it is in the i-th row and j-th column, so it is multiplied by a_i and b_j. This is why the (i,j)-entry of the matrix ends up as the coefficient for a_ib_j. ☺

Proposition 5.4.2. $\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} c_{11} & \dots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \dots & c_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_2 \end{bmatrix} = \sum_{i,j} c_{ij}a_ib_j$.

Proof.

$$\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} C \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = (\sum_i a_i \boldsymbol{e}_i^{\mathrm{T}}) C (\sum_j b_j \boldsymbol{e}_j) = \sum_{i,j} a_i b_j (\boldsymbol{e}_i^{\mathrm{T}} C \boldsymbol{e}_j).$$

Here $\boldsymbol{e}_i^{\mathrm{T}} C \boldsymbol{e}_j$ is the (i,j) entry of C. □

Now let us again look at an inner product space. Given a (finite dimensional) inner product space V, it is first and foremost an abstract vector space. So after picking some basis, it is just $\mathbb{R}^n$. How would the inner product interact with this?

Say $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is a basis. Then given any vector in V, say $\boldsymbol{a} = \sum a_i \boldsymbol{v}_i$ and $\boldsymbol{b} = \sum b_j \boldsymbol{a}_j$, their inner product is $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \sum_i a_i \boldsymbol{v}_i, \sum_j b_j \boldsymbol{v}_i \rangle = \sum_{i,j} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. It is clear that, in fact, by knowing the values of $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ for all i, j, we will know the whole inner product!

Let us go further. If you look closely, you might see that

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i,j} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \dots & \langle \boldsymbol{v}_1, \boldsymbol{v}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{v}_n, \boldsymbol{v}_1 \rangle & \dots & \langle \boldsymbol{v}_n, \boldsymbol{v}_n \rangle \end{bmatrix} \begin{bmatrix} b_1 & \dots & b_n \end{bmatrix}.$$

So if we pick basis and turn all vectors into coordinates, then the inner product is represented by some symmetric matrix $G = \begin{bmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \dots & \langle \boldsymbol{v}_1, \boldsymbol{v}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{v}_n, \boldsymbol{v}_1 \rangle & \dots & \langle \boldsymbol{v}_n, \boldsymbol{v}_n \rangle \end{bmatrix}$, called the ***Gram matrix*** for this basis. And given vectors with coordinates $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$, the inner product is just $\boldsymbol{v}^{\mathrm{T}} G \boldsymbol{w}$.

Definition 5.4.3. *For any finite dimensional inner product space and any basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, the corresponding **Gram matrix** is the symmetric matrix whose (i,j) entry is $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$.*

We have caught a glimps of the important fact: any finite dimensional inner product space, after picking a basis, becomes simply $\mathbb{R}^n$ with exotic inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}^{\mathrm{T}} G \boldsymbol{w}$ for some nice symmetric matrix G. Or we can simply write $(\mathbb{R}^n, G)$ for short, indicating that the abstract vector space is $\mathbb{R}^n$, but instead of the dot product, we use the exotic inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}^{\mathrm{T}} G \boldsymbol{w}$.

Remark 5.4.4. *In case you did not notice this, the Euclidean space ($\mathbb{R}^n$ with dot product) is essentially the case when G is the identity matrix.*

Definition 5.4.5. *A symmetric matrix S is positive definite if $\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x} \geq 0$ with equality if and only if $\boldsymbol{x} = \boldsymbol{0}$.*

It is obvious that a symmetric matrix G is positive definite if and only if $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}^{\mathrm{T}} G \boldsymbol{w}$ is indeed an inner product.

Proposition 5.4.6. *For any finite dimensional inner product space and any basis, the corresponding Gram matrix is positive-definite.*

Proof. Trivial by definition of positive definite matrices. □

In conclusion, we see that any finite dimensional inner product space, upon picking a basis, is the same as $(\mathbb{R}^n, G)$ for some positive-definite matrix G. The study of inner product spaces is now the study of positive-definite matrices.

Example 5.4.7. If A is invertible, then $AA^{\mathrm{T}}, A^{\mathrm{T}}A$ are both positive definite. They are obviously symmetric. Furthermore, if $\boldsymbol{x}^{\mathrm{T}} A^{\mathrm{T}} A \boldsymbol{x} = \|A\boldsymbol{x}\| \geq 0$ with equality if and only if $A\boldsymbol{x} = \boldsymbol{0}$ if and only if $\boldsymbol{x} = \boldsymbol{0}$ (A is invertible).

For a non-zero real number x, we know x^2 is positive. The fact that $AA^{\mathrm{T}}, A^{\mathrm{T}}A$ are both positive-definite are natural generalizations of this. (In contrast, A^2 is not guaranteed to be anything. Pick rotation matrix by $\frac{\pi}{2}$ on $\mathbb{R}^2$, and you shall see that $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^2 = -I$ is thoroughly negative.) ☺

Now, suppose we have a symmetric matrix S. How would I know if it is positive definite or not? Let us see an example first.

Example 5.4.8. Consider $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & 8 \\ 3 & 8 & 14 \end{bmatrix}$. I claim that A is positive definite, i.e., $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$, with equality if and only if $\boldsymbol{x} = \mathbf{0}$.

Suppose $\boldsymbol{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Then first we want to show that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} \geq 0$. We have $\begin{bmatrix} x & y & z \end{bmatrix} A \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x^2 + 6y^2 + 14z^2 + 4xy + 6xz + 16yz$. To show that this is always non-negative, we may try to complete the squares. First let us group up everything that is related to x into a square, and we have $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = (x+2y+3z)^2 + 2y^2 + 5z^2 + 4yz$. Next, for the rest, let us group everything related to y and have $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = (x+2y+3z)^2 + 2(y+z)^2 + 3z^2$. Then the only thing left is z^2, which is by itself a square. So we see that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} \geq 0$, with equality only if $x + 2y + 3z = y + z = z = 0$. And this condition can be solved as $x = y = z = 0$. So indeed, A is positive definite.

But let us rethink this process one more time. What are we doing when we complete the square? Let us think about this. Originally, we have variables x, y, z. Now I want to make a linear change of variables, where $x' = x + 2y + 3z$, and $y' = y + z$ and $z' = z$. Then I realize that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = (x')^2 + 2(y')^2 + 3(z')^2 = \begin{bmatrix} x' & y' & z' \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$. So we indeed have a sum of squares.

Observe that $\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. In another words, we have $U = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ and $D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$, such that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = (U\boldsymbol{x})^{\mathrm{T}} D(U\boldsymbol{x})$. So in fact $A = U^{\mathrm{T}} DU$. The process of completing squares on the polynomial $x^2 + 6y^2 + 14z^2 + 4xy + 6xz + 16yz$ corresponds to the process of writing A as $U^{\mathrm{T}} DU$. This is the LDU decomposition of A!

Imagine this. Suppose we want to complete the square. Then we would perform some change of variables where x, y, z becomes x', y', z' so that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x}$ becomes a linear combination of squares of x', y', z' (with positive coefficients). This change of variable thing can be achieved as $\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = E \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ for some INVERTIBLE E. If completion of square can be achieved, then we have $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = (E\boldsymbol{x})^{\mathrm{T}} D(E\boldsymbol{x})$ for some diagonal matrix D with positive diagonal entries. In particular, we have $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} (E^{\mathrm{T}} DE)\boldsymbol{x}$. So in fact $A = E^{\mathrm{T}} DE$.

But what is this E? Any invertible matrix can be thought of as a series of row operations or column operations. So $(E^{-1})^{\mathrm{T}} AE^{-1} = D$, i.e., we are reducing A to a diagonal matrices via "symmetric" row and column operations! So starting with A, if we add the first column to the second column, then we immediately add the first row to the second row. So on so forth, until we get a diagonal matrix. If you always do your reduction top to bottom and left to right, then you end up with $A = LDL^{\mathrm{T}}$.

Think about the following. We start with $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & 8 \\ 3 & 8 & 14 \end{bmatrix}$. The corresponding polynomial for $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x}$ is $x^2 + 6y^2 + 14z^2 + 4xy + 6xz + 16yz$. Now we substract the first row from the second row and the first column from the second column (note that the order does not matter, because $(BA)C = B(AC)$). Then we end up with $A' = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & 5 \\ 3 & 5 & 14 \end{bmatrix}$. This corresponds to the transformation $x^2 + 6y^2 + 14z^2 + 4xy + 6xz + 16yz = (x+y)^2 + 3y^2 + 14z^2 + 2(x+y)y + 6(x+y)z + 10yz$. So it works! These "symmetric row+column operations" are exactly corresponding to the change of variables in the polynomial $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x}$! ☺

First of all, there is a change of basis formula here.

Proposition 5.4.9. *On an abstract inner product space V, suppose the Gram matrix for the basis $\mathcal{B}$ is $G_{\mathcal{B}}$, and the Gram matrix under the basis $\mathcal{C}$ is $G_{\mathcal{C}}$, then we have $I_{\mathcal{B}\leftarrow\mathcal{C}}^{\mathrm{T}} G_{\mathcal{B}} I_{\mathcal{B}\leftarrow\mathcal{C}} = G_{\mathcal{C}}$.*

Proof. For any $\boldsymbol{v}, \boldsymbol{w} \in V$, we have

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}_{\mathcal{B}}^{\mathrm{T}} G_{\mathcal{B}} \boldsymbol{w}_{\mathcal{B}} = \boldsymbol{v}_{\mathcal{C}}^{\mathrm{T}} I_{\mathcal{B}\leftarrow\mathcal{C}}^{\mathrm{T}} G_{\mathcal{B}} I_{\mathcal{B}\leftarrow\mathcal{C}} \boldsymbol{w}_{\mathcal{C}}.$$

We also have

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}_{\mathcal{C}}^{\mathrm{T}} G_{\mathcal{C}} \boldsymbol{w}_{\mathcal{C}}.$$

Since $\boldsymbol{v}_{\mathcal{C}}, \boldsymbol{w}_{\mathcal{C}}$ are arbitrary, we have the result $I_{\mathcal{B}\leftarrow\mathcal{C}}^{\mathrm{T}} G_{\mathcal{B}} I_{\mathcal{B}\leftarrow\mathcal{C}} = G_{\mathcal{C}}$. □

Now let us example the LDU property of a positive definite matrix. To start, we at least want a positive definite matrix S to be invertible.

Proposition 5.4.10. *Positive definite matrices are invertible.*

Proof. If S is not invertible, say $\boldsymbol{x}$ is a nonzero vector in the kernel, then $\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x} = \boldsymbol{0}$, so S is not positive definite. □

But we also have another nice property of positive definiteness.

Proposition 5.4.11. *If A is positive definite, then all leading principal submatrices are positive definite.*

Proof. Let A_k be the k-th LPS. Then $\boldsymbol{x}^{\mathrm{T}} A_k \boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}} & \boldsymbol{0}^{\mathrm{T}} \end{bmatrix} A \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{0} \end{bmatrix} \geq 0$, with equality if and only if $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{0}$, if and only if $\boldsymbol{x} = \boldsymbol{0}$. □

Remark 5.4.12. *In fact all "principal submatrices" are positive definite. We have not defined this term yet.... But for example, all diagonal entries are positive.*

Corollary 5.4.13. *If A is positive definite, then it has LU decomposition (in fact LDL^{T} decomposition, since A is symmetric).*

Corollary 5.4.14. *A is positive definite if and only if it is invertible with an LDL^{T} decomposition $A = LDL^{\mathrm{T}}$, and all diagonal entries of D are positive.*

Proof. ***⇒:***

We only need to show that D has positive diagonal entries. Note that the i-th diagonal entry is $\boldsymbol{e}_i^{\mathrm{T}} D \boldsymbol{e}_i = ((L^{\mathrm{T}})^{-1}\boldsymbol{e}_i)^{\mathrm{T}} A ((L^{\mathrm{T}})^{-1}\boldsymbol{e}_i) \geq 0$, and equality cannot hold because $(L^{\mathrm{T}})^{-1}\boldsymbol{e}_i \neq \boldsymbol{0}$ (L is unit triangular and thus invertible).

⇐:

Note that obviously D is positive definite, because $\boldsymbol{x}^{\mathrm{T}} D \boldsymbol{x}$ is a positive linear combination of squares of all coordiates. Now $\boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} = (L\boldsymbol{x})^{\mathrm{T}} D (L\boldsymbol{x}) \geq 0$, and with equality if and only if $L\boldsymbol{x} = \boldsymbol{0}$. But L is unit triangular and thus invertible. So $L\boldsymbol{x} = \boldsymbol{0}$ if and only if $\boldsymbol{x} = \boldsymbol{0}$. □

So how to see if a matrix is positive definite? Well first it needs to be symmetric, which is easy to check. Then we do Gaussian elimination (which corresponds to an attempt to "complete the squares") to get the $\mathrm{LDL}^{\mathrm{T}}$ decomposition. If we fail to get an LU decompositon (have to swap at some point), then A is NOT positive definite. If we get the $\mathrm{LDL}^{\mathrm{T}}$ decomposition, then look at the diagonal entries of D. If they are all positive, then $\boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x}$ as a polynomial is a positive sum of squares, so A is positive definite.

This has an interesting consequence:

Theorem 5.4.15 (Cholesky decomposition)**.** *If A is positive definite, then $A = LL^{\mathrm{T}}$ for an invertible lower triangular L. If we require the diagonal entries of L to be positive, then the decomposition is unique.*

Proof. We have $A = LDL^{\mathrm{T}}$. Since diagonal entries of D are positve, taking their square roots, we have diagonal D' whose square is D. Now $A = (LD')(LD')^{\mathrm{T}}$.

For uniqueness, note that $A = LDL^{\mathrm{T}}$ is unique. Suppose $A = BB^T$ for some invertible lower triangular matrix B with positive diagonal entries, then we can write $B = B'D'$ where B' is unit lower triangular, and D' is diagonal with positive diagonal entries. Then $A = B'D'D'(B')^{\mathrm{T}}$. By uniqueness of the LDU decomposition, we have $B' = L$ and $(D')^2 = D$. Since D' is diagonal with positive diagonal entries, each entry must be the unique square root of the corresponding entry of D. So D' is uniquely determined by D. So $B = LD'$ is also uniquely determined,. □

Remark 5.4.16. *We previously have seen that if A is invertible, then AA^{T} is positive definite. Now we see that if G is positive definite, then it is in fact AA^{T} for some invertible A. This is the generalization of the fact that all positive numbers are squares.*

Corollary 5.4.17. *Any finite dimensional inner product space is isomorphic to the Euclidean space.*

Proof. We know any finite dimensional space is isomorphic to $(\mathbb{R}^n, G)$ for some positive definite G. Then $G = LL^{\mathrm{T}}$ for an invertible upper triangular L.

Now we do the change of basis to get a Gram matrix $L^{-1}G(L^{-1})^{\mathrm{T}} = I$, and we are now isomorphic to $(\mathbb{R}^n, I)$. □

Well, as it turned out, finite dimensional inner product spaces are all the same as Euclidean spaces!

Remark 5.4.18. *It might now seem like a waste of time to introduce inner product spaces at all. They are all Euclidean!*

However, the point lies in two things. First of all, for infinite dimensional spaces, you have no choice but to do it abstractly. This involves super important spaces like function spaces and random variable spaces. And in many cases, your choice of inner product will endow the infinite dimenisonal space with different geometry.

Secondly, just like abstract vector spaces allow us to focus on invariant things under a change of basis, inner product space allows us to focus on invariant things under a change of orthonormal basis. We shall now study orthonormal basis in the next section.

5.5 Orthonormal Basis

Now, given an inner product space, we know it is isomorphic to a Euclidean space. This means by picking the right basis, the inner product will BE the dot product. Say $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ is this basis. Then if the identity matrix is our Gram matrix, we must conclude that $\|\boldsymbol{q}_i\| = 1$ for all i, and $\boldsymbol{q}_i \perp \boldsymbol{q}_j$ whenever $i \neq j$.

Definition 5.5.1. *An **orthogonal** basis (OGB for short) in an inner product space is $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ where all vectors are orthogonal to each other.*

*An **orthonormal** basis (ONB for short) is an orthogonal basis where, in addition, all vectors are unit vectors. (I.e., vectors with length one.)*

Let us take Euclidean space as an example. The standard basis is obviously a good orthonormal space. But sometimes we do need other orthonormal basis.

Example 5.5.2. Say I am pushing a box down a slope. Then we have pushing force, friction, normal force and gravity acting on the box. It is best NOT to choose the horizontal direction and vertical direction as orthonormal basis. Rather, it is better to choose the direction parallel to the slope and orthogonal to the slope as orthonormal basis. This way, three out of four forces lies on the coordinate axis, so computations are easier. ☺

A really nice property of ONB is that their coordinates are super easy to find.

Proposition 5.5.3. *Given an orthonormal basis $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ of V, then for any $\boldsymbol{v} \in V$, we have $\boldsymbol{v} = \sum \langle \boldsymbol{v}, \boldsymbol{q}_i \rangle \boldsymbol{q}_i$. In particular, the coordinate map for the basis is $(\boldsymbol{q}_1, \dots, \boldsymbol{q}_n)^{-1} = \begin{pmatrix} \langle \boldsymbol{q}_1 | \\ \vdots \\ \langle \boldsymbol{q}_n | \end{pmatrix}$ that sends each $\boldsymbol{v} \in V$ to $\begin{bmatrix} \langle \boldsymbol{q}_1, \boldsymbol{v} \rangle \\ \vdots \\ \langle \boldsymbol{q}_n, \boldsymbol{v} \rangle \end{bmatrix}$.*

Proof. Suppose $\boldsymbol{v} = \sum a_i \boldsymbol{q}_i$. Now for each i, we take inner product with $\boldsymbol{q}_i$ on both sides, we have $\langle \boldsymbol{v}, \boldsymbol{q}_i \rangle = a_i$ as desired. □

Corollary 5.5.4. *If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is an OGB, then for any $\boldsymbol{v} \in V$, the coordinates of $\boldsymbol{v}$ for $\boldsymbol{v}_i$ is $\frac{\langle \boldsymbol{v}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle}$.*

Proof. Let $\boldsymbol{q}_i = \frac{1}{\|\boldsymbol{v}_i\|} \boldsymbol{v}_i$, then $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ form an orthonormal basis. So $\boldsymbol{v}_i = \sum \langle \boldsymbol{v}, \boldsymbol{q}_i \rangle \boldsymbol{q}_i = \sum \frac{\langle \boldsymbol{v}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i$. □

Again, keep in minds that $\frac{\langle \boldsymbol{v}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i$ is simply the projection of $\boldsymbol{v}$ to the direction of $\boldsymbol{v}_i$. So essentially, we are saying that any vector $\boldsymbol{v}$ equals to the sum of its projections into orthogonal directions.

Remark 5.5.5. *Recall that physicists sometimes like to write $\langle \boldsymbol{v}|$ for the linear map that sends each input $\boldsymbol{w}$ to the output $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$.*

Given ONB $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ of V, then for any $\boldsymbol{v} \in V$, we have

$$\boldsymbol{v} = \sum \langle \boldsymbol{v}, \boldsymbol{q}_i \rangle \boldsymbol{q}_i = \sum |\boldsymbol{q}_i\rangle\langle \boldsymbol{q}_i|(\boldsymbol{v}) = (\sum |\boldsymbol{q}_i\rangle\langle \boldsymbol{q}_i|)(\boldsymbol{v}).$$

Since this is true for all $\boldsymbol{v}$, we can conclude that $\sum |\boldsymbol{q}_i\rangle\langle \boldsymbol{q}_i|) = I$ the identity map. This is actually not something new. Note that under dot product, $\langle \boldsymbol{q}_i|$ is simply $\boldsymbol{q}_i^{\mathrm{T}}$. Recall that ages ago, we have an example

$$\begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \end{bmatrix} \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \end{bmatrix} + \begin{bmatrix} -\frac{4}{5} \\ \frac{3}{5} \end{bmatrix} \begin{bmatrix} -\frac{4}{5} & \frac{3}{5} \end{bmatrix} = I.$$

This is exactly an example of the formula $\sum |\boldsymbol{q}_i\rangle\langle \boldsymbol{q}_i|) = I$. This represents the process of the decomposition of a vector $\boldsymbol{v}$ into orthogonal components.

Example 5.5.6. Consider the OGB $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$. (To make this an ONB, we would need to divide them all by 2.)

Given a vector, say $\boldsymbol{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$. To find coordinates with respect to the new basis, we simply take the inner product (which in this case is simply the dot product). For example, $\frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ gives the coefficient for the basis vector $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. It appears that we simply get the average of $1, 2, 3, 4$.

In general, we can see that

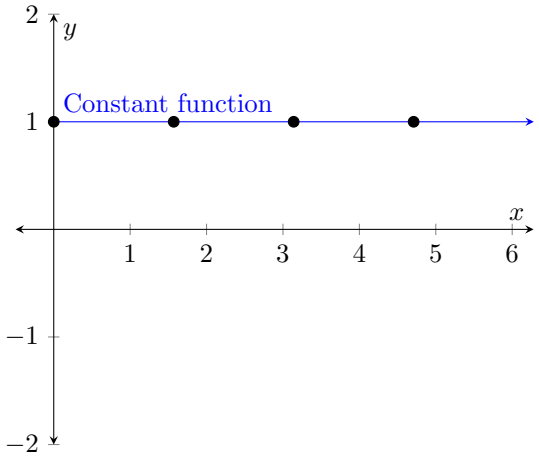
$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \frac{a+b+c+d}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{a-b+c-d}{4} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} + \frac{a+b-c-d}{4} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} + \frac{a-b-c+d}{4} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}.$$

☺

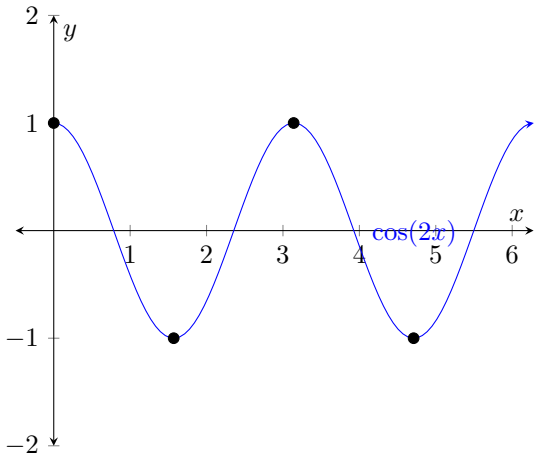
Example 5.5.7. Now is the time to introduce a famous orthogonal basis, the Haar wavelet basis. The basis $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$ is an example of Haar wavelet basis in dimension four. In general, this is an OGB where each vector only uses ± 1 as its coordinates.

Why the name "wavelet"? Well, we have the following correspondence.

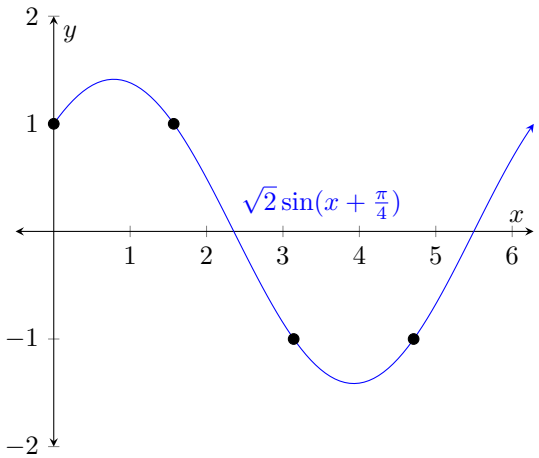
1. The vector $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ corresponds to the constant wave, i.e., the function $f(x) = 1$ for all x.



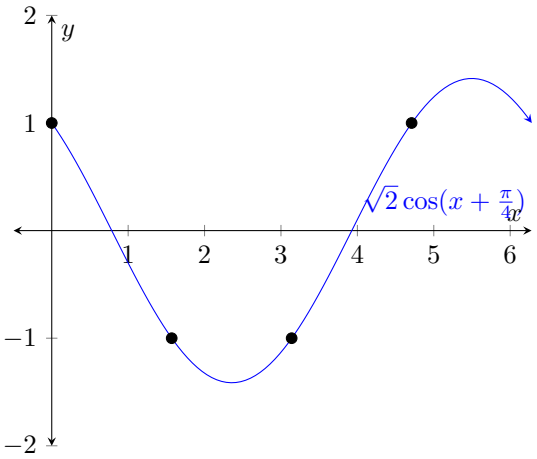
2. The vector $\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$ corresponds to the wave $\cos(2x)$, and we simply sample its value at $x = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$.



3. The vector $\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$ corresponds to the wave $\sqrt{2}\sin(x + \frac{\pi}{4})$, and again we simply sample its value at $x = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$.



4. The vector $\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$ corresponds to the wave $\sqrt{2}\cos(x + \frac{\pi}{4})$, and again we simply sample its value at $x = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$.



As you can see, they are essentially discrete version of waves, where we have some extra constants to make sure that coordinates are ± 1 for these vectors.

Haar wavelet basis is widely used for jpeg image processing. Here we merely take $\mathbb{R}^{16}$ as an example, but keep in mind that in practice it is usually a much larger space.

Suppose we have 16 pixels in greyscale, so we can represent each pixel by a number, the "grey-ness". Then we could use the vector $\begin{bmatrix} a_1 \\ \vdots \\ a_{16} \end{bmatrix}$ to represent the following picture.

a_1	a_2	a_5	a_6
a_3	a_4	a_7	a_8
a_9	a_{10}	a_{13}	a_{14}
a_{11}	a_{12}	a_{15}	a_{16}

Now suppose the picture takes up too much space! So we want to store less numbers, and maybe the picture is a little blurry, but we can still recognize the general shapes. Say we "blend" two adjacent pixels by takining their average. Let $b_i = \frac{a_{2i-1}+a_{2i}}{2}$, then we are replacing the picture with $\begin{bmatrix} a_1 \\ \vdots \\ a_{16} \end{bmatrix}$ with $\begin{bmatrix} b_1 \\ b_1 \\ b_2 \\ b_2 \\ \vdots \\ b_8 \\ b_8 \end{bmatrix}$. The picture now looks like

b_1	b_1	b_3	b_3
b_2	b_2	b_4	b_4
b_5	b_5	b_7	b_7
b_6	b_6	b_8	b_8

You can imagine that our picture is a little blurry now, but when there are many many pixels, then averaging each pair would have neglegible effect on your picture in general. But what if we still want to compress the size of the data? Maybe I can again take the average of adjacent b_i's. Let $c_i = \frac{b_{2i-1}+b_{2i}}{2}$, then

we are replacing the picture with $\begin{bmatrix} a_1 \\ \vdots \\ a_{16} \end{bmatrix}$ with $\begin{bmatrix} c_1 \\ c_1 \\ c_1 \\ c_1 \\ c_2 \\ \vdots \\ c_4 \end{bmatrix}$. The picture now looks like

c_1	c_1	c_2	c_2
c_1	c_1	c_2	c_2
c_3	c_3	c_4	c_4
c_3	c_3	c_4	c_4

Now the picutre is a lot more blurry, but we saved a lot of data storage. We only need to store four numbers. If we have many many pixels, you can continue this process for ever to get more and more blurry pictures with less and less storage need.

With this compression process in mind, let us look at the Haar wavelet basis. In the case of $\mathbb{R}^4$, they are columns $\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \boldsymbol{c}_4$ of the matrix $H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$.

Now given a picture with 4 pixels, say

a_1	a_2
a_3	a_4

Then we have decomposition

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \frac{a_1+a_2+a_3+a_4}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{a_1+a_2-a_3-a_4}{4} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} + \frac{a_1-a_2+a_3-a_4}{4} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} + \frac{a_1-a_2-a_3+a_4}{4} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}.$$

Look at the first term $\frac{a_1+a_2+a_3+a_4}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. This is exactly the picture when I averaged ALL pixels! Now we only need to store one number, and the picture is too blurry to see anything.

Now look at the first two terms.

$$\frac{a_1+a_2+a_3+a_4}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{a_1+a_2-a_3-a_4}{4} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{a_1+a_2}{2} \\ \frac{a_1+a_2}{2} \\ \frac{a_3+a_4}{2} \\ \frac{a_3+a_4}{2} \end{bmatrix}.$$

This is exactly the picture where I averaged the top two pixels and the bottom two pixels!

Now look at the first term and the third term. We have

$$\frac{a_1+a_2+a_3+a_4}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{a_1-a_2+a_3-a_4}{4} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{a_1+a_3}{2} \\ \frac{a_2+a_4}{2} \\ \frac{a_1+a_3}{2} \\ \frac{a_2+a_4}{2} \end{bmatrix}.$$

This is exactly the picture where I averaged the left two pixels and the right two pixels!

As you can see, given a picture, we can first write it under the Haar wavelet basis. Then "blur adjacent pixels" can be done simply by dropping coordiates. The more coordinate you drop, the more blurry your picture can be.

So how can one construct this Haar wavelet basis in general? One can iterate the process of $H_{2n} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$. Feel free to check that H_{2n} would still be symmetric and has orthogonal columns. In particular, in the 16 pixel case, we have the Haar wavelet basis:

$$\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1
\end{bmatrix}.$$

Note that the Haar wavelet basis is NOT ONB, merely OGB. To make it orthonormal, one needs to use columns of $\frac{1}{\sqrt{n}} H_n$ instead. ☺

Here we see an interesting matrix H_n whose columns are orthogonal and unit vectors. In general, the following types of matrices are very important.

5.6 Orthogonal Matrices

From now on, let us focus on the Euclidean space!

Let us do an alternative characterization of these matrices. If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ form an ONB, then the inverse of $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ is the coordinate map, which is $\begin{bmatrix} \langle \boldsymbol{v}_1 | \\ \vdots \\ \langle \boldsymbol{v}_n | \end{bmatrix}$ as we have seen. Under dot product, the "bra" simply means transpose. So we have an interesting discovery. In $\mathbb{R}^n$ with the dot product, if the columns of $U = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix}$ form an ONB, then $U^{-1} = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_n^{\mathrm{T}} \end{bmatrix} = U^{\mathrm{T}}$.

Definition 5.6.1. *We say an $n \times n$ invertible matrix A is an **orthogonal matrix** if $A^{-1} = A^{\mathrm{T}}$.*

Now, what does it mean to have $A^{\mathrm{T}}A = I$? Suppose $A = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix}$, then the (i, j) entry of $A^{\mathrm{T}}A$ would be $\boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_j$. So $A^{\mathrm{T}}A = I$ implies that columns of A are mutually orthogonal normal vectors. Similarly, $AA^{\mathrm{T}} = I$ implies that rows of A are mutually orthogonal normal vectors.

Proposition 5.6.2. *For an $n \times n$ invertible matrix A, TFAE:*

1. *A is orthogonal.*
2. *Rows of A form an orthonormal basis to the Euclidean space $\mathbb{R}^n$.*
3. *Columns of A form an orthonormal basis to the Euclidean space $\mathbb{R}^n$.*
4. *$\langle A\boldsymbol{v}, A\boldsymbol{w} \rangle = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$. (Note that the inner product here is just dot product.)*
5. *$\|A\boldsymbol{v}\| = \|\boldsymbol{v}\|$ for all $\boldsymbol{v} \in \mathbb{R}^n$.*

Proof. We are given that A is square. So $A^{-1} = A^{\mathrm{T}}$ iff $AA^{\mathrm{T}} = I$ iff $A^{\mathrm{T}}A = I$. By writing A in columns, we see that $A^{\mathrm{T}}A = I$ iff columns of A form ONB. Similarly, by writing A in rows, we see that $AA^{\mathrm{T}} = I$ iff rows of A form ONB. So it is already obvious tha the first three statements are equivalent.

Now, if A is orthogonal, then $\langle A\boldsymbol{v}, A\boldsymbol{w}\rangle = \boldsymbol{v}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{w} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{w}\rangle$. So the first one implies the forth one. Conversely, if $\langle A\boldsymbol{v}, A\boldsymbol{w}\rangle = \langle \boldsymbol{v}, \boldsymbol{w}\rangle$, then $\boldsymbol{v}^{\mathrm{T}}(A^{\mathrm{T}}A)\boldsymbol{w} = \boldsymbol{v}^{\mathrm{T}}I\boldsymbol{w}$ for all $\boldsymbol{v}, \boldsymbol{w}$. This means $A^{\mathrm{T}}A = I$.

Finally, the last statement is also equivalent because due to polarization identity, length determines angle. □

In general, orthogonal matrices would map the standard basis to some orthonormal basis. In particular, above proposition shows that orthogonal matrices as a linear map is exactly a "rigid motion" that preserves length and angles. (This means we have some rotation+reflection going on.) In particular, we have the following:

Lemma 5.6.3. *If A, B are orthogonal, then AB, A^{-1} are orthogonal.*

Proof. Obviously $A^{-1}(A^{-1})^{\mathrm{T}} = A^{-1}(A^{\mathrm{T}})^{-1} = (A^{\mathrm{T}}A)^{-1} = I$. And we have $(AB)(AB)^{\mathrm{T}} = A(BB^{\mathrm{T}})A^{\mathrm{T}} = AA^{\mathrm{T}} = I$.

There are alternative proofs. For example, $\|AB\boldsymbol{v}\| = \|B\boldsymbol{v}\| = \|\boldsymbol{v}\|$ for all $\boldsymbol{v}$ because A, B are orthogonal. But then this means AB is orthogonal. □

Another interpretation of an orthogonal matrix is the following: they corresponds to change of basis between orthonormal basis.

Proposition 5.6.4. *In an inner product space V, the change of coordinate matrix from an orthonormal basis to another orthonormal basis is an orthogonal matrix. Conversely, given any orthonormal basis $\mathcal{B}$, then $\mathcal{B}U$ is an orthonormal basis if U is orthogonal.*

Proof. WLOG suppose that V is the Euclidean space. Then any orthonormal basis forms an orthogonal matrix. Say the two orthonormal basis are columns of A, B. Note that both A, B are orthogonal and therefore invertible. Then because $A(A^{-1}B) = B$, we see that the basis transition matrix from A to B is $A^{-1}B$, and therefore the change of coordinate matrix is just $B^{-1}A$, which is still orthogonal.

For the "converse" part, say an orthonormal basis are columns of A, then for orthogonal U, AU is still orthogonal. So the columns of AU still form an orthonormal basis. □

Here is a preview of a future important theorem, that we do not proof at the moment.

Theorem 5.6.5 (Singular Value Decomposition)**.** *For any matrix A, we can find orthogonal matrices U, V such that $UAV = \begin{bmatrix} \Sigma & O \\ O & O \end{bmatrix}$ where Σ is diagonal.*

This is a super upgrade from rank normal form, which says there must be some bases to make our map pretty. Now we see that, in fact, there must be some orthogonal bases to make our map pretty. We can make our map pretty in a way that is also compatible with the inner product structure of the domain and codomain.

Remark 5.6.6. *In practice, people only perform change of basis using orthogonal matrices, because such change of basis process would preserve the length of error term.*

For example, suppose we have a vector $\boldsymbol{v}$. Under some basis, we measured it and get its coordinate $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. However, measurements are never precise. So technically, we have $\boldsymbol{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \boldsymbol{e}$ for some vector $\boldsymbol{e}$ recording the tiny error.

Suppose now I want to use a new basis. Let A be the change of coordinate matrix, and say $\begin{bmatrix}1\\2\\3\end{bmatrix}$ *is changed into* $\begin{bmatrix}4\\5\\6\end{bmatrix}$*. However, due to the error term, the ACTUAL coordinates of* $\boldsymbol{v}$ *in the new basis should be* $A(\begin{bmatrix}1\\2\\3\end{bmatrix}+\boldsymbol{e})=\begin{bmatrix}4\\5\\6\end{bmatrix}+A\boldsymbol{e}$.

If A is an orthogonal matrix, then $\|A\boldsymbol{e}\|=\|\boldsymbol{e}\|$. In particular, if the initial error is tiny, then after the change of coordinates, the resulting error is still tiny.

This is not true for non-orthogonal matrices. For example, if $A=\begin{bmatrix}1&2\\0&1\end{bmatrix}$ *and the error term is* $\boldsymbol{e}=\begin{bmatrix}0\\0.1\end{bmatrix}$*, then* $A\boldsymbol{e}=\begin{bmatrix}0.2\\0.1\end{bmatrix}$*. The size of error is now $\|A\boldsymbol{e}\|=\frac{\sqrt{5}}{10}$, whereas the original error has size $\frac{1}{10}$. The size of error is more than doubled.*

For this reason, people NEVER perform non-orthonormal change of basis. We do not want the error term to be magnified.

Let us conclude this section by some more examples of orthogonal matrices.

Example 5.6.7. Rotation matrices $R_\theta=\begin{bmatrix}\cos\theta&-\sin\theta\\\sin\theta&\cos\theta\end{bmatrix}$ are orthogonal matrices, because it preserves length.

What about 3D rotations? Any rotation in $\mathbb{R}^3$ must be rotating around a line. Suppose we pick an orthonormal basis $\boldsymbol{q}_1,\boldsymbol{q}_2,\boldsymbol{q}_3$ such that $\boldsymbol{q}_1$ is the axis of rotation, and we rotate around this by rotating the plane $\operatorname{span}(\boldsymbol{q}_2,\boldsymbol{q}_3)$ via R_θ. Then under this orthonormal basis, our rotation looks like $\begin{bmatrix}1&\mathbf{0}^{\mathrm{T}}\\\mathbf{0}&R_\theta\end{bmatrix}$.

So, what does this looks like under the original basis, i.e., the standard basis? Since change of coordinate maps are simply arbitrary orthogonal matrices, we are looking at $U\begin{bmatrix}1&\mathbf{0}^{\mathrm{T}}\\\mathbf{0}&R_\theta\end{bmatrix}U^{-1}$ for some arbitrary orthogonal matrix U, where $U=\begin{bmatrix}\boldsymbol{q}_1&\boldsymbol{q}_2&\boldsymbol{q}_3\end{bmatrix}$. Here we first use $U^{-1}=U^{\mathrm{T}}=\begin{bmatrix}\boldsymbol{q}_1^{\mathrm{T}}\\\boldsymbol{q}_2^{\mathrm{T}}\\\boldsymbol{q}_3^{\mathrm{T}}\end{bmatrix}$ to change coordinates from the standard basis to our new orthonormal basis, then use the matrix under our new basis, and finally use U to change the coordinates back to the standard basis.

All in all, we see that 3D rotations are ALL $U\begin{bmatrix}1&\mathbf{0}^{\mathrm{T}}\\\mathbf{0}&R_\theta\end{bmatrix}U^{-1}$ for some orthogonal matrix U.

We are going to see structures like BAB^{-1} a lot. For linear transformations where the domain is the same as the codomain, and the domain and codomain are required to change basis simultaneously, then we would always go from A to BAB^{-1} for some invertible B, i.e., the change of coordinate matrix. ☺

So we have figured out all 2D and 3D rotation matrices. What about higher dimensions?

Definition 5.6.8. *A **Givens rotation** is a matrix G_{ij}^θ whose $(i,i),(i,j),(j,i),(j,j)$ entries form R_θ, and the rest looks exactly like the identity matrix.*

Obviously the Givens rotations are rotations. They rotate the x_ix_j-plane while fixing all other coordinate axis. As a result, their compositions are all rotations. It turns out that this is it. We don't have enough time to do this in class though, so we shall leave it at that.

Example 5.6.9. What about reflections, which always preserve the length as well? For example, we know that $\begin{bmatrix}0&1\\1&0\end{bmatrix}$ is an orthogonal matrix and in fact a reflection about the line $x=y$ on $\mathbb{R}^2$.

Suppose we want to reflect about a hyperplane with unit normal vector $\boldsymbol{q}_1$. Then we pick any orthonormal basis $\boldsymbol{q}_2, \dots, \boldsymbol{q}_n$ for the hyper plane, and then $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ would form an orthonormal basis for $\mathbb{R}^n$. Under this basis, the reflection would send $\boldsymbol{q}_1$ to $-\boldsymbol{q}_1$, and preserve the rest. So the matrix under this basis is $\begin{bmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$.

As a result, a generic reflection must be $U \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} U^{-1}$ for any orthogonal matrix U. Note that the matrix in the middle is in fact $I - 2\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$. So a reflection must be $U(I - 2\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}})U^{-1} = U(I - 2\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}})U^{\mathrm{T}} = UU^{\mathrm{T}} - 2U\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}U^{\mathrm{T}} = I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$, where $\boldsymbol{u}$ is the first column of U, i.e., $\boldsymbol{q}_1$, the unit normla vector to our hyperplane.

In fact we have already seen this before. The reflection to a hyper plane is exactly $I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ where $\boldsymbol{u}$ is the unit normal vector. ☺

Definition 5.6.10. *A **Householder transformation** is $I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ for some unit vector $\boldsymbol{n}$. Or abstractly on an inner product space, it is $I - 2|\boldsymbol{n}\rangle\langle\boldsymbol{n}|$ for some unit vector $\boldsymbol{n}$.*

It is easy to verify that this is indeed an orthogonal matrix. In fact its inverse and transpose are both itself.

Here is an interesting result that provide some nice intuitions. The proof is optional and not required.

Theorem 5.6.11 (Geometric meaning of an orthogonal matrix)**.** *An orthogonal matrix is either a product of Givens rotations (and thus a rotation it self), or is the product of a series of Givens rotations and a Householder transformation (and thus a reflection then rotation).*

5.7 (Optional) Geometrix meaning of an orthogonal matrix

To see this proof, first we shall need a few lemmas. In the following, by rotation we shall always means a product of a series of Givens rotations.

Lemma 5.7.1. *If R is a rotation, then R^{-1}, $\begin{bmatrix} R & O \\ O & I \end{bmatrix}$, $\begin{bmatrix} I & O \\ O & R \end{bmatrix}$ are still rotations.*

Proof. Just verify for all Givens rotations. Then use the fact that inverse matrices of a product is the product with the inverses with reversed order. And for a product of block diagonal matrices, we can can simply compute the product on each diagonal block. □

Lemma 5.7.2. *Given any two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ of the same length and $n \geq 2$, then there is a rotation R such that $R\boldsymbol{v} = \boldsymbol{w}$.*

Proof. We proceed by induction. When $n = 2$, this is obvious.

For generic n, suppose $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \boldsymbol{v}' \end{bmatrix}$. Then by induction hypothesis, I can find R_1 a rotation on $\mathbb{R}^{n-1}$ such that $R_1\boldsymbol{v}' = \begin{bmatrix} \|\boldsymbol{v}'\| \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. Then we see that $\begin{bmatrix} 1 & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{0} & R_1 \end{bmatrix} \boldsymbol{v} = \begin{bmatrix} v_1 \\ \|\boldsymbol{v}'\| \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

Now I find a rotation R_2 such that $R_2 \begin{bmatrix} v_1 \\ \|\boldsymbol{v}'\| \end{bmatrix} = \begin{bmatrix} \|\boldsymbol{v}\| \\ 0 \end{bmatrix}$. Then $\begin{bmatrix} R_2 & O \\ O & I \end{bmatrix} \begin{bmatrix} v_1 \\ \|\boldsymbol{v}'\| \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \|\boldsymbol{v}\|\boldsymbol{e}_1$.

So composing the two, we have found a rotation R such that $R\boldsymbol{v} = \|\boldsymbol{v}\|\boldsymbol{e}_1$. Similarly, we can find a rotation R' such that $R'\boldsymbol{w} = \|\boldsymbol{w}\|\boldsymbol{e}_1$. Since the two vectors have the same length, we see that $R\boldsymbol{v} = R'\boldsymbol{w}$, so $(R')^{-1}R$ is the desired rotation. □

Lemma 5.7.3. *If H is a Householder transformation, then $\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & H \end{bmatrix}$ is a householder transformation.*

Proof. Suppose $H = I_n - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ for some unit vector H. We now look at $\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & H \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & I_n - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} \end{bmatrix} = I_{n+1} - \begin{bmatrix} 0 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} \end{bmatrix} = I_{n+1} - 2\begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix}\begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix}^{\mathrm{T}}$. And it is easy to verify that $\begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix}$ is still a unit vector. □

Now we can proceed to prove the theorem. It is of course a proof by induction. Let me write the theorem again for clarity.

Theorem 5.7.4 (Geometric meaning of an orthogonal matrix)**.** *An orthogonal matrix is either a product of Givens rotations (and thus a rotation it self), or is the product of a series of Givens rotations and a Householder transformation (and thus a reflection then rotation).*

Proof. When $n = 1$ or $n = 2$ this is trivial. We proceed by induction.

For generic n, consider any orthogonal matrix Q, and let its first column be $\boldsymbol{q}$. This is a unit vector. So we can find a rotation R such that $R\boldsymbol{q} = \boldsymbol{e}_1$. So $RQ = \begin{bmatrix} 1 & \star \\ \mathbf{0} & \star \end{bmatrix}$.

Since RQ is still an orthogonal matrix, its columns must still be orthogonal to each other. In particular, the upper right block must actually be zero. So we have $RQ = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & Q' \end{bmatrix}$ for some matrix Q'. Further more, again because RQ is orthogonal, $\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & Q' \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & Q' \end{bmatrix}^{\mathrm{T}}$, and therefore we see that $\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & (Q')^{-1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & (Q')^{\mathrm{T}} \end{bmatrix}$. So we see that Q' is orthognal and with smaller dimension.

So by induction, Q' is either a rotation (in which case Q is a rotation), or a rotation times a Householder transformation (in which case Q is a rotation times a Householder transformation). □

Here is a nice generalization to the affine case. We can now understand ALL distance preserving transformations on $\mathbb{R}^n$, even the non-linear ones. We define ***distance*** between two vectors $\boldsymbol{v}, \boldsymbol{w}$ to simply be $\|\boldsymbol{v} - \boldsymbol{w}\|$.

Lemma 5.7.5 (Distance ratio and convex linear combinations)**.** *In $\mathbb{R}^n$, suppose $\|\boldsymbol{v} - \boldsymbol{w}\| = r$. Then for any $0 \le t \le 1$, the point $t\boldsymbol{v} + (1-t)\boldsymbol{w}$ has distance $(1-t)r$ to the point $\boldsymbol{v}$, and has distance tr to the point $\boldsymbol{w}$.*

Conversely, if $\boldsymbol{x}$ has distance $(1-t)r$ to the point $\boldsymbol{v}$ and has distance tr to the point $\boldsymbol{w}$ for some $0 \le t \le 1$, then $\boldsymbol{x} = t\boldsymbol{v} + (1-t)\boldsymbol{w}$.

Proof. Note that $\boldsymbol{v} - (t\boldsymbol{v} + (1-t)\boldsymbol{w}) = (1-t)(\boldsymbol{v} - \boldsymbol{w})$, and $\boldsymbol{w} - (t\boldsymbol{v} + (1-t)\boldsymbol{w}) = t(\boldsymbol{w} - \boldsymbol{v})$. Hence the distance ratio is exactly $(1-t)$ to t. Given total distance r between $\boldsymbol{v}, \boldsymbol{w}$, the rest is obvious.

For the second portion, suppose $\boldsymbol{x}$ has distance $(1-t)r$ to the point $\boldsymbol{v}$ and has distance tr to the point $\boldsymbol{w}$ for some $0 \le t \le 1$. Intuitively, this means that $\boldsymbol{x}$ is in the ball around $\boldsymbol{v}$ with radius tr, and also in the ball around $\boldsymbol{w}$ with radius $(1-t)r$. But since $\boldsymbol{v}, \boldsymbol{w}$ have a distance of $tr + (1-t)r$, the two balls have unique intersection, and then we are done.

Let us do this more rigorously. The triangle inequality $\|\boldsymbol{v}-\boldsymbol{w}\| \leq \|\boldsymbol{v}-\boldsymbol{x}\| + \|\boldsymbol{w}-\boldsymbol{x}\|$ achieved equality. This means the corresponding Cauchy-Schwarz inequality (by squaring both sides and simplify) between $\boldsymbol{v}-\boldsymbol{x}$ and $\boldsymbol{w}-\boldsymbol{x}$ achieves equality. Hence these two are parallel. In particular, $\boldsymbol{x}$ lies on the line through the two points $\boldsymbol{v}, \boldsymbol{w}$. Since we have distance requirement with $tr, (1-t)r \leq r$, $\boldsymbol{x}$ must lie on the line segment between $\boldsymbol{v}, \boldsymbol{w}$.

If either $\boldsymbol{v}-\boldsymbol{x}$ or $\boldsymbol{w}-\boldsymbol{x}$ is the zero vector, then the statement is trivial. Suppose they are both non-zero. Let $-k(\boldsymbol{v}-\boldsymbol{x}) = \boldsymbol{w}-\boldsymbol{x}$ for some $k \geq 0$. Rearrange and we have $\boldsymbol{x} = \frac{k}{k+1}\boldsymbol{v} + \frac{1}{k+1}\boldsymbol{w}$. By the first portion of our lemma, we see that the distance between $\boldsymbol{x}$ and $\boldsymbol{v}$ should be $\frac{1}{k+1}r$. Hence $1-t = \frac{1}{k+1}$ and $t = \frac{k}{k+1}$. So we see that $\boldsymbol{x} = t\boldsymbol{v} + (1-t)\boldsymbol{w}$. □

Theorem 5.7.6 (Classification of isometries on $\mathbb{R}^n$)**.** *If $f : \mathbb{R}^n \to \mathbb{R}^n$ is any map that preserves distance, i.e., $\|\boldsymbol{v}-\boldsymbol{w}\| = \|f(\boldsymbol{v})-f(\boldsymbol{w})\|$, then we must have $f(\boldsymbol{v}) = A\boldsymbol{v}+\boldsymbol{b}$ for some constant vector $\boldsymbol{b}$ and orthogonal matrix A.*

Proof. For any points $\boldsymbol{v}, \boldsymbol{w}$ and any $0 \leq t \leq 1$, let us first show that $f(t\boldsymbol{v}+(1-t)\boldsymbol{w}) = tf(\boldsymbol{v}) + (1-t)f(\boldsymbol{w})$.

Set $r = \|\boldsymbol{v}-\boldsymbol{w}\|$. Then $t\boldsymbol{v}+(1-t)\boldsymbol{w}$ has distance $(1-t)r$ to $\boldsymbol{v}$, and distance tr to $\boldsymbol{w}$.

Now we apply the map f, then $f(\boldsymbol{v}), f(\boldsymbol{w})$ shall also have distance r. The point $f(t\boldsymbol{v}+(1-t)\boldsymbol{w})$ must still have distance $(1-t)r$ to $f(\boldsymbol{v})$, and distance tr to $f(\boldsymbol{w})$. But this in turn implies that $f(t\boldsymbol{v}+(1-t)\boldsymbol{w}) = tf(\boldsymbol{v}) + (1-t)f(\boldsymbol{w})$.

The lemma below shows that this implies that $f(\boldsymbol{v}) = A\boldsymbol{v}+\boldsymbol{b}$ for some constant vector $\boldsymbol{b}$ and some matrix A. Note that we must have $f(\boldsymbol{0}) = \boldsymbol{b}$. In particular, we have

$$\|A\boldsymbol{x}\| = \|f(\boldsymbol{x})-\boldsymbol{b}\| = \|f(\boldsymbol{x})-f(\boldsymbol{0})\| = \|\boldsymbol{x}-\boldsymbol{0}\| = \|\boldsymbol{x}\|.$$

So A corresponds to a linear map that preserves distance. Hence it is an orthogonal matrix. □

Lemma 5.7.7 (Classification of all affine maps)**.** *If $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfies $f(t\boldsymbol{v}+(1-t)\boldsymbol{w}) = tf(\boldsymbol{v}) + (1-t)f(\boldsymbol{w})$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ and all $0 \leq t \leq 1$, then $f(\boldsymbol{x}) = A\boldsymbol{x}+\boldsymbol{b}$.*

Proof. (Intuitively, an n-dimensional affine space is basically $\mathbb{R}^n$, but you are NOT allowed to do linear combinations in general. You can only perform "convex linear combinations", which means $t\boldsymbol{v}+(1-t)\boldsymbol{w}$ with $0 \leq t \leq 1$. An affine map is therefore a map that preserves convex linear combinations.)

Set $\boldsymbol{b} = f(\boldsymbol{0})$, and set $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{b}$. Let us show that g is linear, then we are done. To start, we have $g(\boldsymbol{0}) = \boldsymbol{0}$. So far so good. Now let us check scalar multiplications.

For $0 \leq k \leq 1$, we have $k\boldsymbol{v} = k\boldsymbol{v} + (1-k)\boldsymbol{0}$. Hence $f(k\boldsymbol{v}) = kf(\boldsymbol{v}) + (1-k)f(\boldsymbol{0})$. Rearrange terms and use the fact that $\boldsymbol{b} = f(\boldsymbol{0})$, we have $f(k\boldsymbol{v}) - \boldsymbol{b} = k(f(\boldsymbol{v}) - \boldsymbol{b})$. Hence $g(k\boldsymbol{v}) = kg(\boldsymbol{v})$.

For $k > 1$, we have $\boldsymbol{v} = \frac{1}{k}(k\boldsymbol{v}) + (1-\frac{1}{k})\boldsymbol{0}$. Using similar idea to above, we would obtain $g(\boldsymbol{v}) = \frac{1}{k}g(k\boldsymbol{v})$.

For $k < 0$, we have $\boldsymbol{0} = (\frac{1}{1-k})(k\boldsymbol{v}) + (1-\frac{1}{1-k})\boldsymbol{v}$. Using similar idea to above, we would obtain $g(\boldsymbol{0}) = (\frac{1}{1-k})g(k\boldsymbol{v}) + (1-\frac{1}{1-k})g(\boldsymbol{v})$. Rearrange and use the fact that $g(\boldsymbol{0}) = \boldsymbol{0}$, we would get $g(k\boldsymbol{v}) = kg(\boldsymbol{v})$ again.

All in all, we have $g(k\boldsymbol{v}) = kg(\boldsymbol{v})$ for all $k \in \mathbb{R}$.

Now let us check vector addition. Consider $g(\boldsymbol{v}+\boldsymbol{w})$. Note that $\boldsymbol{v}+\boldsymbol{w} = \frac{1}{2}(2\boldsymbol{v}) + \frac{1}{2}\boldsymbol{w}$. Hence $f(\boldsymbol{v}+\boldsymbol{w}) = \frac{1}{2}f(2\boldsymbol{v}) + \frac{1}{2}f(2\boldsymbol{w})$. Rearrange this, and we get $g(\boldsymbol{v}+\boldsymbol{w}) = \frac{1}{2}g(2\boldsymbol{v}) + \frac{1}{2}g(2\boldsymbol{w}) = g(\boldsymbol{v}) + g(\boldsymbol{w})$. So we are done. □

In short, any distance-preserving transformation on $\mathbb{R}^n$ must be a composition of translations, reflections and rotations.

5.8 (Optional) Rotations and Skew-Symmetric Matrices

These two kinds of matrices are closely related to each other. Let us see how. This is among the best ways to understand and represent rotations.

Definition 5.8.1. *We say a matrix A is **skew-symmetric** if $A^{\mathrm{T}} = -A$.*

Example 5.8.2. If $A^{\mathrm{T}} = -A$, then look at the diagonal entries on the left hand side and the right hand side. We must conclude that all diagonal entries for A are zero.

So if A is 2×2 and skew-symmetric, then it can only be something like $A = \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$. Note that this is invertible as long as $a \neq 0$.

If A is 2×2 and skew-symmetric, then it can only be something like $A = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}$. Note that this NEVER invertible. If $a = b = c = 0$, then this is the zero matrix. If not, then $A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{0}$, so the kernel is non-zero. ☺

Definition 5.8.3. *Given any matrix A, we define the matrix e^A to be the series $I + A + \frac{A^2}{2!} + \ldots$.*

This series always converge. However, proving it is beyond the scope of the class. Let us just take it for granted for now.

Example 5.8.4. Recall that $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$, and $\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots$

Now consider a matrix $A = \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$. Note that this is a skew-symmetric matrix. We now have

$$\begin{aligned}
\mathrm{e}^A &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} + \frac{1}{2!}\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}^2 + \frac{1}{3!}\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}^3 + \ldots \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} + \begin{bmatrix} -\frac{a^2}{2!} & 0 \\ 0 & -\frac{a^2}{2!} \end{bmatrix} + \begin{bmatrix} 0 & \frac{a^3}{3!} \\ -\frac{a^3}{3!} & 0 \end{bmatrix}^3 + \ldots \\
&= \begin{bmatrix} 1 - \frac{a^2}{2!} + \ldots & -a + \frac{a^3}{3!} - \ldots \\ a - \frac{a^3}{3!} + \ldots & 1 - \frac{a^2}{2!} + \ldots \end{bmatrix} \\
&= \begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}.
\end{aligned}$$

In particular, the exponential matrix for any 2 by 2 skew symmetric matrix is a rotation matrix. ☺

Proposition 5.8.5. *If $AB = BA$, then $\mathrm{e}^{A+B} = \mathrm{e}^A \mathrm{e}^B$.*

Proof. At an abstract level, the exponential map and its properties are derived from the basic properties of addition and multiplication. So if we have commutativity, then matrix addition and matrix multiplication now satisfies all the properties of regular addition and regular multiplication of real numbers. So the exponential map would exhibit exactly the same behavior as the one for regular real numbers.

To be more rigorous, here is the computational proof.

$$\begin{aligned}
\mathrm{e}^A \mathrm{e}^B =& (\sum_{i=0}^{\infty} \frac{1}{i!} A^i)(\sum_{j=0}^{\infty} \frac{1}{j!} A^j) \\
=& \sum_{i,j=0}^{\infty} \frac{1}{i!j!} A^i B^j \\
=& \sum_{i,j=0}^{\infty} \frac{C_{i+j}^i}{(i+j)!} A^i B^j \\
=& \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i=0}^{k} C_k^i A^i B^{k-i}
\end{aligned}$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!}(A+B)^k$$
$$= \mathrm{e}^{A+B}.$$

Can you spot the step that uses $AB = BA$? (In particular, if $AB \neq BA$, this computation would be false at that step.) □

Just be careful that in general, $\mathrm{e}^{A+B} \neq \mathrm{e}^A \mathrm{e}^B$ when $AB \neq BA$.

Theorem 5.8.6. *If A is skew-symmetric, then e^A is orthogonal.*

Proof. Note that $A^{\mathrm{T}} = -A$, so in particular we have $AA^{\mathrm{T}} = -A^2 = A^{\mathrm{T}}A$. So $\mathrm{e}^A \mathrm{e}^{A^{\mathrm{T}}} = \mathrm{e}^{A+A^{\mathrm{T}}} = \mathrm{e}^O = I$, the identity map.

All I need to do now is to show that $(\mathrm{e}^A)^{\mathrm{T}} = \mathrm{e}^{A^{\mathrm{T}}}$. But this is obvious.

$$(\mathrm{e}^A)^{\mathrm{T}} = (I + A + \frac{A^2}{2!} + \dots)^{\mathrm{T}} = I + A^{\mathrm{T}} + \frac{(A^{\mathrm{T}})^2}{2!} + \cdots = \mathrm{e}^{A^{\mathrm{T}}}.$$

So e^A is orthogonal. □

Note that e^A is in fact always a rotation. And in fact, this goes both ways. If Q is a rotation (no reflection involved), then $Q = \mathrm{e}^A$ for some skew-symmetric A. However, these are trickier to prove and we don't have the tools to do it yet.

Example 5.8.7. Consider any generic 3 by 3 skew symmetric matrix $A = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}$. Note that $A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{0}$.

Now consider the rotation e^A. Applying this to $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$, we see that

$$\mathrm{e}^A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = (I + A + \dots) \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

All the A portion of the series would NOT contribute at all, because they multiply $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ to zero. In particular, e^A is a rotation around the line containing $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$. (In fact, it is a rotation with radian degree $\| \begin{bmatrix} a \\ b \\ c \end{bmatrix} \|$. So if $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ has length 2π, then $\mathrm{e}^A = I$. However, this is much easier to do after we have eigenvalues, so we don't prove it here.)

As you can see, in practice, the orthogonal matrices are usually ugly and hard to compute. Those cosines and sines would give ugly numbers everywhere. However, their "logarithm", the skew symmetric matrices are much prettier and intuitive, and the entries would have immediate geometric meanings behind them.

Also note that here we in fact have $A\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \times \boldsymbol{v}$. So we see that cross product is the logarithm of rotations. This is why in physics and multivariable calculus, things involving rotations are usually related to the cross product. ☺

5.9 Gram-Schmidt Orthogonalization and QR decomposition

5.9.1 First Perspective: Algorithm to find an orthonormal basis

So far, we are using the existence of an ONB for free. We know they exist, because any inner product space is isomorphic (as inner product spaces) to the Euclidean space. But how to specifically find one?

In practice, people usually start with a random (non-orthonormal) basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, and tries to transform them into an ONB. This is the famous Gram-Schmidt Orthogonalization process.

Example 5.9.1. How can I make a bunch of vectors orthogonal to each other?

Suppose I have a stack of papers, but they are NOT stacked up-right. Rather, they are stacked in a tiled way, forming a parallelepiped. The three edges are $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, where the first two are edges of a paper, and the third vector is the direction of how they are stacked.

In particular, although $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthogonal to each other, $\boldsymbol{v}_3$ is currently NOT orthogonal to $\boldsymbol{v}_1, \boldsymbol{v}_2$, and my parallelepiped is NOT a rectangular box.

Now this stack is annoying to carry. So I take the stack, and hit the desk with it on its side. Then I hit the desk with it on the other side. Now I will have a rectangular box stack, and it is now easy to carry.

When I hit the stack on its $\boldsymbol{v}_2$ side, I am doing a SHEARING (preserving the base and height) to make sure that $\boldsymbol{v}_3$ is now orthogonal to $\boldsymbol{v}_1$. And then when I hit the stack on the other side, now $\boldsymbol{v}_3$ is also sheared to be orthogonal to $\boldsymbol{v}_2$. Then I am done. The staking direction is finally orthogonal to the other two vectors.

This is the Gram-Schmidt orthogonalization process. ☺

Let us look at another example for more computational detail.

Example 5.9.2. Suppose we only have two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ spanning a two dimensional inner product space. To make them orthogonal, we need to shear $\boldsymbol{v}_2$ along the direction of $\boldsymbol{v}_1$ (i.e., adding multiples of $\boldsymbol{v}_1$ to $\boldsymbol{v}_2$), until it is orthogonal to $\boldsymbol{v}_1$. How much shearing do we need?

Suppose the shearing we need is $(\boldsymbol{v}_1, \boldsymbol{v}_2)\begin{bmatrix}1 & k\\ 0 & 1\end{bmatrix} = (\boldsymbol{v}_1, \boldsymbol{v}_2 + k\boldsymbol{v}_1)$. Then we must have $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 + k\boldsymbol{v}_1\rangle = 0$. Solving for k, we see that we need $k = -\frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1\rangle}$.

So the Gram Schmidt orthogonalization process would transform the basis $\boldsymbol{v}_1, \boldsymbol{v}_2$ to the OGB $\boldsymbol{v}_1, \boldsymbol{v}_2 - \frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1\rangle}\boldsymbol{v}_1$. To make it ONB, just then divide each vector by its lengh, and we are done. ☺

Proposition 5.9.3. *Given any linearly independent $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ in an abstract vector space, let*

$$\begin{aligned}
\boldsymbol{w}_1 =& \boldsymbol{v}_1\\
\boldsymbol{w}_2 =& \boldsymbol{v}_2 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_2\rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1\rangle}\boldsymbol{w}_1\\
\boldsymbol{w}_3 =& \boldsymbol{v}_3 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_3\rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1\rangle}\boldsymbol{w}_1 - \frac{\langle \boldsymbol{w}_2, \boldsymbol{v}_3\rangle}{\langle \boldsymbol{w}_2, \boldsymbol{w}_2\rangle}\boldsymbol{w}_2\\
&\vdots\\
\boldsymbol{w}_k =& \boldsymbol{v}_k - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_k\rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1\rangle}\boldsymbol{w}_1 - \frac{\langle \boldsymbol{w}_2, \boldsymbol{v}_k\rangle}{\langle \boldsymbol{w}_2, \boldsymbol{w}_2\rangle}\boldsymbol{w}_2 - \cdots - \frac{\langle \boldsymbol{w}_{n-1}, \boldsymbol{v}_k\rangle}{\langle \boldsymbol{w}_{n-1}, \boldsymbol{w}_{n-1}\rangle}\boldsymbol{w}_{n-1}.
\end{aligned}$$

Then $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ are all non-zero and mutually orthogonal. Let $\boldsymbol{q}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$, then $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$ is an orthonormal collection of vectors.

*If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ is a basis, then $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ is OGB and $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$ is ONB. We call this ONB the **Gram-Schmidt orthogonalization** of our original basis.*

Proof. Let us prove orthogonality by induction. Suppose $\boldsymbol{w}_1, \dots, \boldsymbol{w}_i$ are pairwise orthogonal. (The initial case $i = 1$ is trivial.)

Then consider $\boldsymbol{w}_{i+1}$. For any $j \leq i$, we have

$$\langle \boldsymbol{w}_{i+1}, \boldsymbol{w}_j \rangle = \langle \boldsymbol{v}_{i+1} - \sum_{k=1}^{i} \frac{\langle \boldsymbol{w}_k, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{w}_k, \boldsymbol{w}_k \rangle} \boldsymbol{w}_k, \boldsymbol{w}_j \rangle = \langle \boldsymbol{v}_{i+1}, \boldsymbol{w}_j \rangle - \sum_{k=1}^{i} \frac{\langle \boldsymbol{w}_k, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{w}_k, \boldsymbol{w}_k \rangle} \langle \boldsymbol{w}_k, \boldsymbol{w}_j \rangle = \langle \boldsymbol{v}_{i+1}, \boldsymbol{w}_j \rangle - \frac{\langle \boldsymbol{w}_j, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{w}_j, \boldsymbol{w}_j \rangle} \langle \boldsymbol{w}_j, \boldsymbol{w}_j \rangle = 0.$$

So $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i+1}$ are also mutually orthogonal. So by induction, we see that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are mutually orthogonal.

Finally, since we went from the basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ of the space $\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ to the collection $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ using only shearings (invertible operations), the result must still be a basis of the space $\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. So $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are all non-zero. The rest of the proposition is obvious. □

Remark 5.9.4. *Note that we started with vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. To proceed, we first make sure that all later vectors are orthogonal to the first vector. So we have*

$$\begin{aligned}
\boldsymbol{w}_1 =& \boldsymbol{v}_1 \\
& \boldsymbol{v}_2 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1 \\
& \boldsymbol{v}_3 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_3 \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1 \\
& \vdots \\
& \boldsymbol{v}_n - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_n \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1.
\end{aligned}$$

Now all later vectors are orthogonal to the first vector. Then we do the shearing to make sure that all later vectors are orthogonal to the second vector. So we have

$$\begin{aligned}
\boldsymbol{w}_1 =& \boldsymbol{v}_1 \\
\boldsymbol{w}_2 =& \boldsymbol{v}_2 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1 \\
& \boldsymbol{v}_3 - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_3 \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1 - \frac{\langle \boldsymbol{w}_2, \boldsymbol{v}_3 \rangle}{\langle \boldsymbol{w}_2, \boldsymbol{w}_2 \rangle} \boldsymbol{w}_2 \\
& \vdots \\
& \boldsymbol{v}_n - \frac{\langle \boldsymbol{w}_1, \boldsymbol{v}_n \rangle}{\langle \boldsymbol{w}_1, \boldsymbol{w}_1 \rangle} \boldsymbol{w}_1 - \frac{\langle \boldsymbol{w}_2, \boldsymbol{v}_n \rangle}{\langle \boldsymbol{w}_2, \boldsymbol{w}_2 \rangle} \boldsymbol{w}_2.
\end{aligned}$$

Now all later vectors are orthogonal to the second vector. So on so forth. This is the process of Gram Schmidt orthogonalization.

5.9.2 Second Perspective: QR decomposition

Now, take a closer look. When we shear, we are ALWAYS substracting multiples of earlier vectors from later vectors! For example, in $\mathbb{R}^2$, to go from $\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}$ to $\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 - \frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle} \boldsymbol{v}_1 \end{bmatrix}$, we only need to use the first column to reduce the second column. This corresponds to a column operation that is upper triangular! In particular, $\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \begin{bmatrix} 1 & -\frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 - \frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle} \boldsymbol{v}_1 \end{bmatrix}$.

Think of Gram-Schmidt as an series of operations on columns of the invertible matrix $A = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix}$, then we are multiplying a series of upper triangular matrices to the right of it, and we get OGB $\begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix} U = \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_n \end{bmatrix}$. Now we divide each column by its length, and we get $\begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix} UD = \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_n \end{bmatrix} D =$

$\begin{bmatrix}\boldsymbol{q}_1 & \dots & \boldsymbol{q}_n\end{bmatrix}$ where D is the diagonal matrix represending the division of each column by its length. And the end result here should be an ONB. In particular, the matrix $Q = \begin{bmatrix}\boldsymbol{q}_1 & \dots & \boldsymbol{q}_n\end{bmatrix}$ is an orthogonal matrix. So $AUD = Q$. We can rearrange this as $A = QR$ where $R = D^{-1}U^{-1}$ is upper triangular.

So now we get the matrix-decomposition perspective of Gram-Schmidt. (Just like Gaussian elimination and LU decomposition are essentially the same, Gram-Schmidt and QR below are essentially the same.)

Theorem 5.9.5 (QR decomposition)**.** *Let A be any $n \times k$ matrix with linearly independent columns (injection). Then we can find an $n \times k$ matrix Q with orthonormal columns and an upper triangular $k \times k$ matrix R such that $A = QR$. If we require all diagonal entries of R to be positive, then this decomposition is unique.*

In particular, if A is invertible, then $A = QR$ for some orthogonal matrix Q and upper triangular matrix R.

Proof. If A has linearly independent columns, then the Gram-Schmidt process means we do AU for some upper triangular U (which represents the shearings and the final scaling of each column), and now the columns of AU are orthonormal. So $Q = AU$ is a matrix with orthonormal columns. So $A = QU^{-1}$ is the desired QR decomposition.

We delay the proof of uniqueness for now. Later perspectives will make the proof of uniqueness trivial. □

Remark 5.9.6. *The QR decomposition IS the Gram-Schmidt orthogonalization. How to find the QR decomposition? One way is to first work out Q by Gram Schmidt orthogonalization, then R follows from the shearing process.*

Example 5.9.7. Suppose we have $\boldsymbol{v}_1 = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}, \boldsymbol{v}_2 = \begin{bmatrix}1\\1\\0\\0\end{bmatrix}, \boldsymbol{v}_3 = \begin{bmatrix}2\\1\\1\\0\end{bmatrix}, \boldsymbol{v}_4 = \begin{bmatrix}2\\0\\0\\0\end{bmatrix}$. So we started with $A = \begin{bmatrix}1&2&4&4\\1&2&2&0\\1&0&2&0\\1&0&0&0\end{bmatrix}$.

We start our shearing. To shear everything in A to be orthogonal to $\boldsymbol{v}_1$, we end up with

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

Here the upper triangular matrix is recording my column operations using the first column to reduce all the columns to the right.

Next we shear the last two columns to be orthogonal to the second column. We end up with

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

Finally, we shear the last column to be orthogonal to the third. We end up with

$$\begin{bmatrix}1&1&1&2\\1&1&-1&-2\\1&-1&1&-0\\1&-1&-1&-0\end{bmatrix} = \begin{bmatrix}1&1&1&1\\1&1&-1&-1\\1&-1&1&-1\\1&-1&-1&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&1\\0&0&0&1\end{bmatrix}.$$

So we have

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

Now normalize these columns, we get the Haar wavelet basis.

All in all, we get the following QR decomposition:

$$\begin{bmatrix}1&1&2&2\\1&1&1&0\\1&0&1&0\\1&0&0&0\end{bmatrix} = (\frac{1}{2}\begin{bmatrix}1&1&1&1\\1&1&-1&-1\\1&-1&1&-1\\1&-1&-1&1\end{bmatrix})\begin{bmatrix}2&2&4&2\\0&2&2&2\\0&0&2&2\\0&0&0&2\end{bmatrix}.$$

☺

5.9.3 Third Perspective: Cholesky decomposition

Let us spend a little time here to discuss what it means to have orthonormal columns.

Example 5.9.8. Consider $Q = \begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}$. This is a matrix with orthonormal columns. Calculation shows that $Q^{\mathrm{T}}Q = I$, yet Q is not an orthogonal matrix. Why? Because Q is not a square matrix, and $QQ^{\mathrm{T}} = \begin{bmatrix}1&&\\&1&\\&&0\end{bmatrix} \neq I$.

For square matrices, $Q^{\mathrm{T}}Q = I$ implies that Q is orthogonal. For non-square matrix, $Q^{\mathrm{T}}Q = I$ only means that Q has orthonormal columns, but may have redundant rows, as shown in the lemma below. ☺

Lemma 5.9.9. *A matrix A has orthonormal columns iff $A^{\mathrm{T}}A = I$. A matrix A has orthonormal rows iff $AA^{\mathrm{T}} = I$.*

Proof. Write $A = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}$, then the (i,j)-entry of $A^{\mathrm{T}}A$ is exactly $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_j$. So $A^{\mathrm{T}}A = I$ means $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_j = 0$ when $i \neq j$, and $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_j = 1$ when $i = j$. So we are done.

The other statement about orthonormal rows is simply the same thing applied to A^{T}. □

The previous proof of QR decomposition and description of Gram-Schmidt orthogonalization is very "algorithmic" in nature. They basically goes along the say line of reason and logic of Gaussian elimination. Instead of using upper rows to kill lower rows, to achieve something triangular, we now use left columns to shear right columns into orthogonal position. (And thus R is upper triangular.)

However, you are no longer a high school student anymore. The hall mark of a mature learner is to learn the SAME thing through DIFFERENT persepctives. Everything we do in this class, you can understand them as an algorithm (Gram Schmidt), or as a matrix decomposition (QR), or as a change of basis process between bases, or as a geometric relation of subspaces, or as an induction process on block matrices, or sometimes even as a mathematical version of physics statements or facts from other sciences.

So now we turn to more perspectives and descriptions and proofs of the Gram-Schmidt process and QR decomposition.

Let me deliver a fatal blow to your sanity: We have actually proven QR decomposition a long long time ago!

Alternative proof of QR decomposition. Given any matrix A with independent columns, then $A^{\mathrm{T}}A$ is positive definite. So by Cholesky decomposition, there is a unique lower triangular matrix L with positive diagonal entries such that $A^{\mathrm{T}}A = LL^{\mathrm{T}}$.

This immediately implies that $L^{-1}A^{\mathrm{T}}A(L^{-1})^{\mathrm{T}} = I$. So we see that $A(L^{-1})^{\mathrm{T}}$ has orthonormal columns, say $Q = A(L^{-1})^{\mathrm{T}}$. Then $A = QL^{\mathrm{T}}$. We are done. □

This proof makes so much sense, because previously, we go from a Gram matrix G to the dot product via some change of basis. If columns of A are the basis, then $A^{\mathrm{T}}A$ is precisely the gram matrix, and the change of basis implemented by L^{T} gives the orthnormal basis Q whose Gram matrix is now I. The following three process are the same:

1. The process of going from some arbitrary invertible matrix to an orthogonal matrix. ($A = QR$)
2. The process of going from some arbitrary basis to an orthonormal basis. (Column view of $A = QR$.)
3. The process of going from some arbitrary Gram matrix to the identity matrix. ($A^{\mathrm{T}}A = R^{\mathrm{T}}Q^{\mathrm{T}}QR = R^{\mathrm{T}}R$, the Cholesky decomposition.)

Let us also prove uniqueness of the QR decomposition here.

Proof of the uniqueness of QR decomposition. Given any matrix A with independent columns, let $A = Q_1R_1 = Q_2R_2$ are two QR decompositions. Then we have $A^{\mathrm{T}}A = R_1^{\mathrm{T}}Q_1^{\mathrm{T}}Q_1R_1 = R_1^{\mathrm{T}}R_1$, and similarly $A^{\mathrm{T}}A = R_2^{\mathrm{T}}R_2$. But then these are two Cholesky decompositions of the same positive definite matrix. So by uniqueness of the Cholesky decomposition, $R_1 = R_2$. And hence we also have $Q_1 = AR_1^{-1} = AR_2^{-1} = Q_2$. □

5.9.4 Fourth Perspective: Geometry of the subspace chain

Suppose we started with A, and we perform Gram-Schmidt to get Q. However, we are too lazy to record the shearing process, so we forget to have R. What to do?

Well, recall that the Gram-Schmidt process $A = QR$ is essentially a change of basis process. The i-th column of R should exactly be the coordinates of the i-th column of A under the basis Q.

Example 5.9.10. If you go from A to Q, but forget to record your column operations. How to find R now?

Well, one can also retroactively solve R from $A = QR$, which implies that $R = Q^{-1}A = Q^{\mathrm{T}}A$.

Writing A, Q in columns $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ and $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$, we see that the (i, j) entry of R is simply $\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{v}_j$. For example, when $n = 3$, the QR decomposition looks like:

$$\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3\end{bmatrix} = \begin{bmatrix}\boldsymbol{q}_1 & \boldsymbol{q}_2 & \boldsymbol{q}_3\end{bmatrix} \begin{bmatrix}\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_3 \\ \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_3 \\ \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_3\end{bmatrix}.$$

This is also obvious since $\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{v}_j$ is the i-th coordinate of $\boldsymbol{v}_j$ under the orthonormal basis. The i-th column of R should exactly be the coordinates of the i-th column of A under the basis Q.

Using previous example, you can compute and verify that indeed,

$$(\frac{1}{2}\begin{bmatrix}1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1\end{bmatrix}) \begin{bmatrix}1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}2 & 2 & 4 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2\end{bmatrix}.$$

(Here the matrix $\frac{1}{2}\begin{bmatrix}1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1\end{bmatrix}$ is symmetric and orthogonal, so it is its own inverse.)

Here is a mystery though. Can you see why must the matrix $\begin{bmatrix}\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{v}_3 \\ \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{v}_3 \\ \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_2 & \boldsymbol{q}_3^{\mathrm{T}}\boldsymbol{v}_3\end{bmatrix}$ always be upper triangular? It is not obvious right away. ☺

Let us unravel this mystery.

Proposition 5.9.11. *Suppose Gram Schmidt brings the basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ to the orthonormal basis $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$. Then* $\operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i) = \operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_i)$.

Proof. This comes from the Gram Schmidt shearing process. We created $\boldsymbol{q}_i$ only using $\boldsymbol{v}_1, \dots, \boldsymbol{v}_i$, so obviously $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_i) \subseteq \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$. But since both collections are linearly independent, they have the same dimension. So we have it. □

This immediatly shows that, conversely, $\boldsymbol{v}_j$ is a linear combination of $\boldsymbol{q}_1, \dots, \boldsymbol{q}_j$, and do not use $\boldsymbol{q}_{j+1}, \dots, \boldsymbol{q}_n$ at all. So its coordinates $\boldsymbol{q}_i^{\mathrm{T}} \boldsymbol{v}_j$ when $i > 0$ are all zero. So $\begin{bmatrix} \boldsymbol{q}_1^{\mathrm{T}} \boldsymbol{v}_1 & \boldsymbol{q}_1^{\mathrm{T}} \boldsymbol{v}_2 & \boldsymbol{q}_1^{\mathrm{T}} \boldsymbol{v}_3 \\ \boldsymbol{q}_2^{\mathrm{T}} \boldsymbol{v}_1 & \boldsymbol{q}_2^{\mathrm{T}} \boldsymbol{v}_2 & \boldsymbol{q}_2^{\mathrm{T}} \boldsymbol{v}_3 \\ \boldsymbol{q}_3^{\mathrm{T}} \boldsymbol{v}_1 & \boldsymbol{q}_3^{\mathrm{T}} \boldsymbol{v}_2 & \boldsymbol{q}_3^{\mathrm{T}} \boldsymbol{v}_3 \end{bmatrix}$ is upper triangular.

Remark 5.9.12. *If you think carefully, buried beneath this is the fact that the inverse of an upper triangular matrix is upper triangular. If $\boldsymbol{q}_i$ only uses $\boldsymbol{v}_1, \dots, \boldsymbol{v}_i$ for each i, then $\boldsymbol{v}_i$ only uses $\boldsymbol{q}_1, \dots, \boldsymbol{q}_i$ for each i.*

The observation here also gives us a geometric way to achieve Gram-Schmidt.

Example 5.9.13. Again consider any basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ in $\mathbb{R}^3$. We can actually find $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$ geometrically, without any calculation!

Here is how. First of all, obviouosly $\boldsymbol{q}_1 = \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}$. This is easy.

Now $\boldsymbol{q}_2$ must be a linear combination of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. So it lies in the plane $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)$. Inside this plane, we should also pick $\boldsymbol{q}_2$ that is perpendicular to the line $\operatorname{span}(\boldsymbol{v}_1)$.

Hey! Given a plane and a line inside it, there are ONLY TWO unit vectors in the plane normal to the line. One is in the same "half-plane" as $\boldsymbol{v}_2$, the other is in the opposite "half-plane". So $\boldsymbol{q}_2$ must be the one in the same "half-plane" as $\boldsymbol{v}_2$. This choice is unique.

Now consider $\boldsymbol{q}_3$. It has to lie in the space $\mathbb{R}^3 = \operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$, and it must be orthogonal to the plane $\operatorname{span}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)$. Now, in the space, there are ONLY TWO unit vectors orthogonal to the plane, one in each "half-space". We let $\boldsymbol{q}_3$ be the one in the same "half-space" as $\boldsymbol{v}_3$.

Then $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$ must be the result of Gram-Schmidt. We simply have no other choice in each step. ☺

So now we proceed with the subspace version of Gram-Schmidt. As you shall see, the Gram-Schmidt does NOT depends on the specific vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ at all. It ONLY depends on the CHAIN of SUBSPACES they generate.

Proposition 5.9.14. *Given a chain of subspaces $\{\boldsymbol{0}\} \subset V_1 \subset V_2 \subset \cdots \subset V_n$ where $\dim V_k = k$, then there is an orthonormal basis $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ such that $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k) = V_k$. This basis is unique up to sign. (I.e., all such orthonormal basis must be $\pm\boldsymbol{q}_1, \dots, \pm\boldsymbol{q}_n$.)*

If it helps, just imagine that V_i is $\operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$ if you like.

Proof. To start, V_1 is one dimensional. So pick any unit vector in it and we are done. Note that in an one-dimensional space, there are only TWO possible choice of unit vectors, and they are negation of each other. So we get $\pm\boldsymbol{q}_1$.

Next, we want to find $\boldsymbol{q}_2$. To do this, we need to find a vector in V_2 that is perpendicular to V_1. However, since V_2 is a plane and V_1 is a line, there are ONLY two unit vectors perpendicular to V_1. So we get $\pm\boldsymbol{q}_2$.

This goes on until the end. At each step, when we pick $\boldsymbol{q}_k$, we are trying to find a vector in V_k perpendicular to V_{k-1}. But since V_{k-1} is only one dimension less than V_k, it is a hyperplane in V_k, and has only two unit normal vectors. So we get $\pm\boldsymbol{q}_k$. □

One technical lemma is needed here. We prove it below.

Lemma 5.9.15. *Let V be an inner product space with dimension n, and let W be a subspace with dimension $n-1$. Then there are only two unit vectors in V orthogonal to W.*

Proof. All vectors perpendicular to W are in $W^\perp$, which must be one dimensional. Hence $W^\perp$ contains only two unit vectors. □

In short, according to this view, we started with $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$. And each $\boldsymbol{q}_k$ is basically the unit normal vector of the hyperplane $\operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1})$ in $\operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$, and there are only two such vectors. But since we require $A = QR$ to have R with positive diagonal entries, we are requiring that $\langle \boldsymbol{q}_k, \boldsymbol{v}_k \rangle > 0$. So we see that only one choice of unit normal vector could make this positive, and the other one makes this negative (the two choices of unit normal vectors are negations of each other). So we choose the positive one.

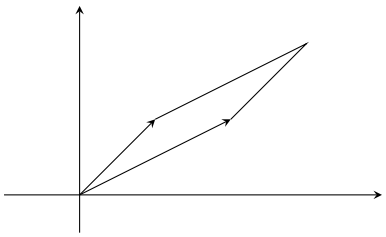
5.9.5 Fifth Perspective: Parallelotope

What is a matrix? So far, we have several interpretations.

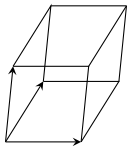
1. We can think of an invertibla matrix A as a basis, by treating its columns as basis vectors.
2. We can also think of any matrix A as a linear map, sending $\boldsymbol{v}$ to $A\boldsymbol{v}$.
3. We can also think of any matrix A as a bilinear map, sending a pair $\boldsymbol{v}, \boldsymbol{w}$ to $\boldsymbol{v}^{\mathrm{T}} A\boldsymbol{w}$. (If A is furthermore symmetric and positive definite, then this bilinear map is an inner product.)
4. We can think of an invertibla matrix A as a change of coordinate matrix or basis transition matrix. So in this sense, A is doing nothing, merely changing the names of objects.

Now let us use a new interpretation. A could represent a parallelotope.

Example 5.9.16. Consider $A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$. We may think of the two column vectors as two edges of a parallelogram.



What if $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}$? Then A represents a parallelopiped in the space $\mathbb{R}^3$.



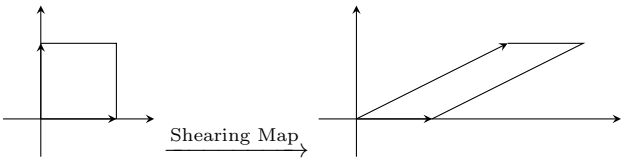
What about $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$? Well, we have two edges, and they live in $\mathbb{R}^3$. Therefore, this is a parallelogram in space!

The term ***parallelotope*** is the generalization of parallelograms and parallelopipeds. In general, an $m \times n$ matrix A can represent an n-dimensional parallelotopes in $\mathbb{R}^m$. ☺

Example 5.9.17. Consider the matrix equation

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}.$$

Let us interpret the first $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ as a linear map, and the other two matrices in the equation as parallelograms. Then we are saying that the shearing map $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ is sending the unit square $\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$ to the parallelogram $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$. Indeed we have



☺

Now, so far we have been focused on the relation between A and Q in the decomposition $A = QR$, and think of R as the transformation process (basis transition matrix). What if we think of Q as the transformation process, and consider A and R as parallelotopes? These two views make a very interesting comparison.

Example 5.9.18. Suppose $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ and $R = \begin{bmatrix} \sqrt{2} & \frac{3}{2}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{2} \end{bmatrix}$. Here you can check that $A = QR$, and R is obviously upper triangular with positive diagonal entries, and Q is orthogonal because it is a rotation by 45 degree counterclockwise.

Now let us treat A as a parallelogram in $\mathbb{R}^2$ whose edges are columns of A (and one vertex is at the origin). Draw this yourself. We see that $Q^{-1}A = R$. This means, if we apply Q^{-1} to the columns of A (edges of your parallelogram), you get columns of R (another parallelogram). In short, if you rotate the parallelogram by A via Q^{-1}, i.e., clockwise by 45 degree, you get a parallelogram whose first edge is on the positive x-axis (because R is upper triangular and the first diagonal entry of R is positive), and its also on the upper half plane (because the second diagonal entry of R is also positive).

In general, you may think of $A = QR$ as such a process. Say $n = 3$. Imagine that A is a parallelepiped. We are rotation/reflecting the A parallelepiped to get the R parallelepiped. First we rotate the parallelepiped so that its first edge is in the positive x-axis. Now we keep this first edge fixed, but we rotate it around the x-axis so that its second edge is now on the positive xy-plane. Now the "base" of our parallelepiped is fixed, so we may choose to reflect it or not, so that the third edge is in the positive half space. All these rotations and reflections are Q^{-1}, and they transform A into a parallelepiped whose three edges are like $\begin{bmatrix} positive & * & * \\ 0 & positive & * \\ 0 & 0 & positive \end{bmatrix}$, i.e., R.

Since Q is a rigid motion, the resulting parallelepiped by R has the same shape as A, but the location is now unique.

This process generalizes to higher dimensions very easily. (The higher dimension versions of a parallelogram or parallelepiped is called a parallelotope.) ☺

The idea that Q is a rigid motion transforming the parallelotope A to a parallelotope R turns out to be VERY IMPORTANT. It turns out that it gives rise to a BETTER way to do Gram-Schmidt than Gram-Schmidt. (Provide greater numerical stability, i.e., the calculations would NOT magnify error term.) The idea is to use a series of Householder transformations, i.e., reflections, to go from A to R. Furthermore, since we are working on parallelotopes, why restrict ourselves to n-dim parallelotopes? One can in fact look at n-dimensional parallelotopes in $\mathbb{R}^m$ and start doing rigid motions.

The following is a more general version of QR decomposition. However, pay special attention to the proof, because that is what most computers would actually do to FIND the QR decomposition.

Theorem 5.9.19. *Given any $m \times n$ matrix A with $m \geq n$, we have QR decomposition $A = QR$ where Q is $m \times m$ and orthogonal, and $R = \begin{bmatrix} R_1 \\ O \end{bmatrix}$ is an $m \times n$ block matrix where R_1 is $n \times n$ and upper triangular with NON-NEGATIVE diagonal entries. If A has full rank (i.e., injective), then R_1 has positive diagonal entries.*

Proof. Suppose $n = 1$. Then A is a single vector $\boldsymbol{v}$, and Q is any orthogonal matrix with first column $\frac{1}{\|\boldsymbol{v}\|}\boldsymbol{v}$, and $R = \|\boldsymbol{v}\|$. All is trivial.

By induction, suppose $n > 1$. Let $\boldsymbol{a}$ be the first column of A, and find a Householder transformation H such that $H\boldsymbol{a} = \|\boldsymbol{a}\|\boldsymbol{e}_1$. This is possible due to the lemma below this proof.

Then $HA = \begin{bmatrix} \|\boldsymbol{a}_1\| & b^{\mathrm{T}} \\ \boldsymbol{0} & A_{n-1} \end{bmatrix}$. (Geometrically, we perfomed a relfection so that the first edge of the parallelotope is now on the positive x_1-axis.)

Here A_{n-1} had $n-1$ columns, so by induction $A_{n-1} = Q_{n-1}R_{n-1}$ as desired, where $R_{n-1} = \begin{bmatrix} R_1 \\ O \end{bmatrix}$ is an $(m-1)\times(n-1)$ block matrix where R_1 is $(n-1\times(n-1)$ and upper triangular with non-negative diagonal entries.

Now consider $\begin{bmatrix} 1 & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{0} & Q_{n-1}^{-1} \end{bmatrix} HA = \begin{bmatrix} \|\boldsymbol{a}_1\| & b^{\mathrm{T}} \\ \boldsymbol{0} & R_1 \\ \boldsymbol{0} & O \end{bmatrix}$. Now the upper portion $\begin{bmatrix} \|\boldsymbol{a}_1\| & b^{\mathrm{T}} \\ \boldsymbol{0} & R_1 \end{bmatrix}$ is $n \times n$ and upper triangular, with non-negative diagonal entries. So we are done.

Finally, if A has full rank and $A = QR$, then since Q is invertible, $R = \begin{bmatrix} R_1 \\ O \end{bmatrix}$ must also have full rank. So R_1 can only have positive diagonal entries. □

Lemma 5.9.20. *If $\|\boldsymbol{v}\| = \|\boldsymbol{w}\|$ in $\mathbb{R}^n$, then there is a Householder matrix $H = I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ for a unit vector $\boldsymbol{n}$ such that $H\boldsymbol{v} = \boldsymbol{w}$.*

Proof. If $\boldsymbol{v} = \boldsymbol{w} = \boldsymbol{0}$, pick $H = I$ and we are done. From now on, assume that $\boldsymbol{v}, \boldsymbol{w} \neq \boldsymbol{0}$.

Let us first get some intuition. Draw the arrow $\boldsymbol{v}$ and $\boldsymbol{w}$, and imagine this reflection process. It is obvious that $\boldsymbol{v} - \boldsymbol{w}$ must be perpendicular to the hyperplane of reflection.

Now we start our proof. Let $\boldsymbol{n} = \frac{\boldsymbol{v}-\boldsymbol{w}}{\|\boldsymbol{v}-\boldsymbol{w}\|}$, and set $H = I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$. Then direct computation would yeild $H\boldsymbol{v} = \boldsymbol{w}$. Yay. □

Remark 5.9.21. *In the end, if you use the formula $H = I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ and the fact that $\boldsymbol{n} = \frac{\boldsymbol{v}-\boldsymbol{w}}{\|\boldsymbol{v}-\boldsymbol{w}\|}$, we have $H = I - 2\frac{(\boldsymbol{v}-\boldsymbol{w})(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}}{(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{w})}$. This is not surprising at all, since we already know that $\frac{(\boldsymbol{v}-\boldsymbol{w})(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}}{(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{w})}$ is the projection to the direction $\boldsymbol{v} - \boldsymbol{w}$.*

Note that the proof above is NOT just a proof, but in fact a computationally viable and excellent way to compute the QR decomposition. We started with a big parallelotope. First we reflect it so that the first edge is on the positive x_1-axis. Then induction means we again perform a reflection, so that the first two edges are on the positive half of the x_1x_2-plane. So on so forth, until we are done and get R.

(I would really love to make an animiation showing an example of such process.... But I do not know how....)

For many cases, when trying to find the QR decomposition, a computer would usually just reflect away. Orthogonal matrices usually tends to "keep small errors small". Whereas shearings R might magnify errors. So applying Q^{-1} to A to find R is a MUCH better idea, compared with applying R^{-1} and find Q. By doing a series of reflections, we can find the QR decomposition of A without enlarge any initial errors.

Remark 5.9.22. *In practice, the traditional Gram-Schmidt would require about $2mn^2$ addition / substraction / multiplication / divisions, and n square root calculations. On the other hand, using a series of Householders takes about $2mn^2 - \frac{2}{3}n^3$ addition / substraction / multiplication / divisions, and n square root calculations. It is both faster and more stable.*

The stuff below is super optional! I would strongly advise you to skip it. I add it here for completeness only.

Let us see how many calculations are involved.

Suppose we are doing regular Gram schmidt. We start with an $m \times n$ matrix $\begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}$. First let us use the first column to change later columns. First we need to calculate $\langle \boldsymbol{v}_1, \boldsymbol{v}_i \rangle$ for each i, and this takes m multiplications and $m-1$ additions for each i, hence a total of mn multiplications and $(m-1)n$ additions. Then we compute $\frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle}$ for all $i \neq 1$, which takes $n-1$ multiplications (we treat division as a multiplication). Next we calculate $\frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle}\boldsymbol{v}_1$ for each i and substract this from the i-th column. This takes

m multiplications and m additions for each column, giving a total of $m(n-1)$ multiplications and $m(n-1)$ additions. We used a total of $mn+n-1+m(n-1) = 2mn+n-m-1$ multiplications/divisions, and $(m-1)n+m(n-1) = 2mn-n-m$ additions/substractions. Let us just say $4mn$ calculations for short.

Next we use the second columns to reduce later columns. So we are essentially dealing with a matrix with m rows and $n-1$ columns, so it takes about $4m(n-1)$ calculations. By this pattern, we see that Gram-Schmidt takes a total of $4m(n+\cdots+1)$ calculations, which is about $2mn^2$. Now we have OGB, but to get ONB, we need to calculate the length, which requires $(2m+1)n$ operations and n square root calculations. Note that $(2m+1)n << 2mn^2$, so we can ignore that for simplicity.

Now let us look at the Householder approach. Given first column $\boldsymbol{v}_1$, first we calculate $\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1$, which takes $2m-1$ calculations. Next we take square root to get $\|\boldsymbol{v}_1\|$, and we compute $\boldsymbol{n} = \boldsymbol{v}_1 - \|\boldsymbol{v}_1\|\boldsymbol{e}_1$, which is a single substraction. (Note that $\boldsymbol{n}$ is not a unit vector, but we shall deal with it later.) So far we have $2m$ calculations and one square root operations.

Next, we want to apply $H = I - 2\frac{\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}}$ to each column. We can calculate $\frac{2}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}}$ with $2m$ calculations, and multiply this to $\boldsymbol{n}$ takes another m calculations.

So the i-th column $\boldsymbol{v}_i$ should turn into $\boldsymbol{v}_i - 2\boldsymbol{n}\frac{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{v}_i}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}}$, and the quantity $2\boldsymbol{n}\frac{1}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}}$ is already pre-calculated. The dot product $\boldsymbol{n}^{\mathrm{T}}\boldsymbol{v}_i$ takes $2m-1$ calculations, and multiply this to $2\boldsymbol{n}\frac{1}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}}$ takes m calculations, and substract this from $\boldsymbol{v}_i$ takes another m. This takes a total of $4m-1$ calculations for each $i \neq 1$. Hence this is a total of $(4m-1)(n-1)$ calculations.

The whole process so far takes $5m+(4m-1)(n-1)$ calculations, which is about $4mn$, and one square root. Now we have completed our induction step, and we no longer need to touch the first row or the first column. So we do this to the $(m-1)\times(n-1)$ lower right matrix left, and this takes about $4(m-1)(n-1)$ calculations. By induction, we need $4mn+4(m-1)(n-1)+\cdots+4(m-n)(n-n)$. Compare this with the previous Gram-Schmidt calculation of $4mn+4m(n-1)+\cdots+4m(n-n)$, and you see where the saving happened.

A more specific calculation shows that $4mn+4(m-1)(n-1)+\cdots+4(m-n)(n-n) = 4(m-n)(n+\cdots+1)+4(n^2+\cdots+1) = 2(m-n)n^2+\frac{4}{3}n^3 = 2mn^2-\frac{2}{3}n^3$.

Now, note that for $A = Q\begin{bmatrix}R_1\\O\end{bmatrix}$, the lower block of zeros means many rows on the right of Q are NOT used at all. If we write $Q = \begin{bmatrix}Q_1 & Q_2\end{bmatrix}$ accordingly, we see that $A = Q_1R_1$. So we have the following corollary:

Corollary 5.9.23. *Given any $m\times n$ matrix A with $m \geq n$, we have QR decomposition $A = QR$ where Q is $m\times n$ matrix with orthonormal columns, and R is $n\times n$ and upper triangular with positive diagonal entries.*

What if you have an $m\times n$ matrix A with $m < n$? Then A^{T} has a QR decomposition $A^{\mathrm{T}} = QR$, and we see that $A = R^{\mathrm{T}}Q^{\mathrm{T}}$ where R^{T} is lower triangular, and Q^{T} is still orthogonal. So we have this so-called RQ decomposition instead.

Furthermore, recall that for Gram-Schmidt, we are working on columns of A from left to right. What if you do this from right to left? Then you have $A = QR$ where R is lower triangular. In conclusion, you can have QR (if $m \geq n$) or RQ (if $m \leq n$), and R maybe upper or lower triangular. In our class, for simplicity, we usually just do QR where R is upper triangular.

5.10 Projections and Applications

5.10.1 Algebraic Projections

The study of orthogonality cannot be complete without the study of orthogonal projections, which is our goal here. However, we take a slight detour and investigate a special property of matrices. This property will end up surprisingly connected to projections.

Definition 5.10.1. *A linear map $P: V \to V$ (i.e., a square matrix) is **idempotent** if $P^2 = P$.*

Note that $P^2 = P$ immediately implies that $P^k = P$ for all positive integer k. ("idem" means "itself", and "potent" means "power". So "idempotent" literally means that its powers are all itself.) However, P might not be invertible. In fact, most of the time it is not.

Example 5.10.2. If P is invertible, we see that $P^2 = P$ implies $P = I$. So the only invertible idempotent matrix is I.

Apart from I, all other idempotent matrices are NOT invertible. Another easy example is the zero matrix O. We clearly have $O^2 = O$. ☺

As will be evident in the next examples, trace is an important aspect of the study of idempotent matrices.

Definition 5.10.3. *We define the* ***trace*** *of a square matrix A to be* $\text{trace}(A) = \sum a_{ii}$*, the sum of diagonal entries of A.*

Proposition 5.10.4. $\text{trace}(AB) = \text{trace}(BA)$ *for any $m \times n$ matrix A and $n \times m$ matrix B.*

Proof. See homework. □

Remark 5.10.5. $\text{trace}(AB) = \text{trace}(BA)$ *means we can "cyclically permute" matrix multiplications while preserving the trace. For example* $\text{trace}(ABC) = \text{trace}(A(BC)) = \text{trace}((BC)A) = \text{trace}(BCA)$*. (Note that non-cyclic permutations might change the trace. We do not have* $\text{trace}(ABC) = \text{trace}(ACB)$ *in general. Take* $A = C^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ *for example.)*

Let us see some examples of idempotent matrices. They are indeed projections!

Example 5.10.6. Consider a diagonal matrix D whose diagonal entries are all 0 or 1. You can also immediately see that it has $D^2 = D$. In fact, these are the only idempotent diagonal matrices. (Can you see why? This is essentially due to the fact that $x^2 = x$ has only two solutions, $x = 0$ and $x = 1$.)

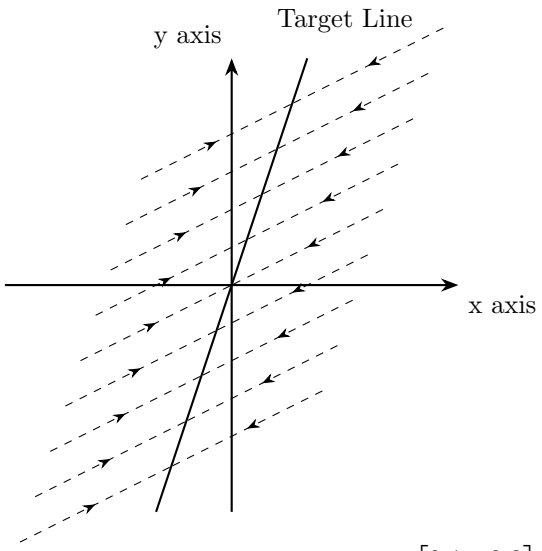
Consider such a matrix, say $D = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix}$. What is this? It will send a generic vector $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} x \\ y \\ 0 \end{bmatrix}$. So this is a projection to the xy-plane. In fact, consider any diagonal matrix D whose diagonal entries are 0 or 1, then you see that it is a projection to some coordiate-subspaces (i.e., subspaces spanned by some coordinate axis).

It is obvious that in this case, $\text{trace}(D) = \text{rank}(D)$ is the dimension of the target space of your projection. ☺

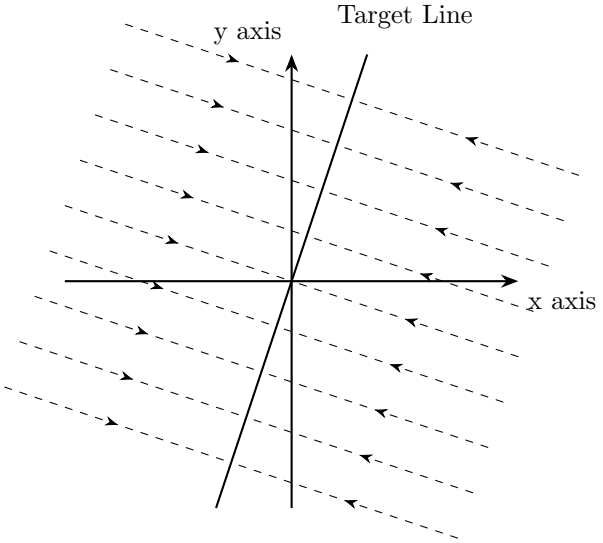
Example 5.10.7. Here let us see a non-diagonal example. Consider $P = \begin{bmatrix} -0.2 & 0.4 \\ -0.6 & 1.2 \end{bmatrix}$. You can easily verify that $P^2 = P$. What does this matrix do?

In this case, one can easily see that $\text{Ran}(P)$ is spanned by $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$. It will map ALL vectors to the direction of $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$. Furthermore, you can check that $P \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. In fact, $P^2 = P$ literally means that P fixes its own range. This has a very "projection" feel to it.

This is what we call an ***oblique projection***. It will project everything to the line $y = 3x$ on the plane $\mathbb{R}^2$, but NOT orthogonally. Rather, every point goes to this line along the direction of $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$. You can check in particular that $\text{Ker}(P)$ is spanned by $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$.



Compare this with the matrix $Q = \begin{bmatrix} 0.1 & 0.3 \\ 0.3 & 0.9 \end{bmatrix}$. You will again find that $Q^2 = Q$, and $\text{Ran}(Q)$ is spanned by $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ as well. However, now the kernel is orthogonal to this range, and we have an ***orthogonal projection***.



Note that we have $\text{trace}(P) = \text{trace}(Q) = 1$, which is the dimension of the target space of our projections. This is exactly the case of diagonal idempotent matrices!

Finally, for the matrix P above, consider $I - P = \begin{bmatrix} 1.2 & -0.4 \\ 0.6 & -0.2 \end{bmatrix}$. You can easily verify that $(I-P)^2 = I-P$ as well. What does this matrix do?

In this case, one can easily see that $\text{Ran}(I - P)$ is spanned by $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, so $\text{Ran}(I - P) = \text{Ker}(P)$. Similarly, you can verify that $\text{Ker}(I - P) = \text{Ran}(P)$. The matrix $I - P$ projects to exactly what P kills, and it kills exactly what P projects! it is the "complement projection" of P.

Again, you can verify that $\text{trace}(I - P) = 1$, which is again the dimension of the target space of this projections. ☺

We are now ready to understand idempotent matrices. On one hand, obviously all (oblique or orthogonal) projections must be idempotent, because projecting twice is the same as projecting once. On the other

hand, the following proposition states that, as linear maps, idempotent matrices are (oblique or orthogonal) projections. So the two are the same.

Proposition 5.10.8. *If $P : V \to V$ is idempotent, we have the following conclusions:*

1. $\mathrm{Ran}(P)$ *is the collection of vectors* $\boldsymbol{v}$ *such that* $P\boldsymbol{v} = \boldsymbol{v}$.
2. $I - P$ *is also idempotent.*
3. $\mathrm{Ran}(I - P) = \mathrm{Ker}(P)$ *and* $\mathrm{Ran}(P) = \mathrm{Ker}(I - P)$.
4. $\mathrm{Ran}(P)$ *and* $\mathrm{Ran}(I - P) = \mathrm{Ker}(P)$ *are complement subspaces. (So P and $I - P$ are projections onto complement subspaces.)*
5. *For any $\boldsymbol{v} \in V$, then $\boldsymbol{v} = P\boldsymbol{v} + (I - P)\boldsymbol{v}$ is the unique decomposition such that the first summand is in* $\mathrm{Ran}(P)$ *and the second summand is in* $\mathrm{Ran}(I - P)$.

Proof. Suppose $\boldsymbol{v} \in \mathrm{Ran}(P)$, say $\boldsymbol{v} = P\boldsymbol{w}$. Then $P\boldsymbol{v} = P^2\boldsymbol{w} = P\boldsymbol{w} = \boldsymbol{v}$. Conversely, if $\boldsymbol{v} = P\boldsymbol{v}$, then obviously $\boldsymbol{v} \in \mathrm{Ran}(P)$. So $\mathrm{Ran}(P)$ is exactly the collection of fixed points of P.

Now $(I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P$, so $I - P$ is also idempotent.

Note that $\boldsymbol{v} \in \mathrm{Ran}(P)$ iff $P\boldsymbol{v} = \boldsymbol{v}$ iff $(I - P)\boldsymbol{v} = \boldsymbol{0}$, so we see that $\mathrm{Ran}(I - P) = \mathrm{Ker}(P)$. Apply this fact to the idempotent $I - P$, we get $\mathrm{Ker}(I - P) = \mathrm{Ran}(I - (I - P)) = \mathrm{Ran}(P)$.

Now let us show that $\mathrm{Ran}(P)$ and $\mathrm{Ker}(P)$ are complements. If $\boldsymbol{v} \in \mathrm{Ran}(P) \cap \mathrm{Ker}(P)$, then $\boldsymbol{v} = P\boldsymbol{v} = \boldsymbol{0}$. So the intersection is zero. On the other hand, the decomposition $\boldsymbol{v} = P\boldsymbol{v} + (I - P)\boldsymbol{v}$ holds for all $\boldsymbol{v} \in V$, so $\mathrm{Ran}(P) + \mathrm{Ran}(I - P) = V$. So $\mathrm{Ran}(P)$ and $\mathrm{Ker}(P) = \mathrm{Ran}(I - P)$ are complement subspaces.

Any decomposition according to complement subspaces must be unique. Hence the decomposition $\boldsymbol{v} = P\boldsymbol{v} + (I - P)\boldsymbol{v}$ is unique. □

The important takeaway here is that whenever you have an idempotent P, you should always think about the pair P and $I - P$. They give you two complement projections. Note that so far, we have NOT used any inner product structure in our definition of idempotent matrices and the derivation of properties. These are true for all abstract vector spaces (finite or infinite dimensional), and without an inner product structure, we cannot differentiate between oblique and orthogonal projections. We simply call them all projections, and none is preferred over others.

We are missing one last piece of the puzzle. We have seen in previous examples that $\mathrm{trace}(P) = \dim \mathrm{Ran}(P)$, it gives the dimension of the subspace we are projecting onto. This is NOT a coincidence.

Proposition 5.10.9. *If P is an idempotent square matrix, then* $\mathrm{trace}(P) = \mathrm{rank}(P)$.

Proof. Here we shall use the fact that $\mathrm{trace}(AB) = \mathrm{trace}(BA)$, whose proof is in the homework.

Suppose P is acting on $\mathbb{R}^n$, so it is a square matrix.

If $\dim \mathrm{Ran}(P) = 1$, then P is rank one, so $P = \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ for some vectors $\boldsymbol{u}, \boldsymbol{v}$. Furthermore, $\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} = P = P^2 = \boldsymbol{u}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u})\boldsymbol{v}^{\mathrm{T}} = (\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u})P$, so we have $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u} = 1$. It is easy to see now $\mathrm{trace}(P) = \sum v_i u_i = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u} = 1$.

Now suppose $\dim \mathrm{Ran}(P) = r$. Then $P = UV$ where U is $n \times r$ and injective, and V is $r \times n$ and surjective. Then $UV = P = P^2 = UVUV$.

But since U is injective, we have the law of left cancellation for U. So $UV = UVUV$ implies that $V = VUV$. Also, since V is surjective, we have the law of right cancellation for V. So $V = VUV$ implies that $I_{r\times r} = VU$.

So $\mathrm{trace}(P) = \mathrm{trace}(UV) = \mathrm{trace}(VU) = \mathrm{trace}(I_{r\times r}) = r$. Note that this is essentially the same proof as above. □

Here it might seem a bit surprising that the trace is the rank. Nominally trace seems to be computed from entries, which are dependent on our choice of basis. But $\dim \mathrm{Ran}(P)$ is the rank, which is independent of choice of basis (and also independent of any inner product structure). So could it be that trace is in fact independent of basis?

Definition 5.10.10. *We say two square matrices A, B are **similar** if we can find invertible C such that $A = CBC^{-1}$.*

Note that the domain and codomain of a square matrix are the same. If we perform a change of basis simultaneously, we would result in similar matrices.

For example, consider $L : V \to V$, and V has basis $\mathcal{B}, \mathcal{C}$. Then $L_{\mathcal{C}\leftarrow\mathcal{C}} = I_{\mathcal{C}\leftarrow\mathcal{B}} L_{\mathcal{B}\leftarrow\mathcal{B}} I_{\mathcal{B}\leftarrow\mathcal{C}}$. Note that on the right side of the equation, the left-most and right-most matrices are inverse of each other. So $L_{\mathcal{C}\leftarrow\mathcal{C}}$ and $L_{\mathcal{B}\leftarrow\mathcal{B}}$ are similar matrices.

Now, trace for linear transformations is indeed independent of any choice of basis (and also independent from inner product structure).

Proposition 5.10.11. *If A, B are similar matrices, then they have the same trace.*

Proof. We have $\operatorname{trace}(CAC^{-1}) = \operatorname{trace}(C^{-1}CA) = \operatorname{trace}(A)$. □

Interpreting C as such a change of basis, then A, CAC^{-1} are referring to the same linear transformation, differ only by a change of basis. (Recall that a linear transformation is a linear map whose domain and codomain are the same.) In this sense, trace is really about the underlying linear map, and not about the nominal matrix.

Let us conclude this section now. A projection is P with $P^2 = P$, and the subspace it is projecting to is $\operatorname{Ran}(P)$, and the dimension is $\operatorname{trace}(P)$.

5.10.2 Orthogonal Projection

Previously we study projections without any reference to the inner product structure. Now, with an inner product structure, we see that most projections are probably oblique projections. However, we are more interested in orthogonal projections. These are projections that kills things orthogonal to $\operatorname{Ran}(P)$.

Definition 5.10.12. *A projection $P : V \to V$ is an **orthogonal projection** if $\operatorname{Ran}(P) \perp \operatorname{Ran}(I-P)$. (I.e., $\operatorname{Ran}(P) \perp \operatorname{Ker}(P)$.)*

The nice thing about orthogonal projections is that they give "best approximation in W" of any vector $\boldsymbol{v} \in V$. This makes them extremely useful.

Proposition 5.10.13. *If $P : V \to V$ is an orthogonal projection to $W \subseteq V$, then $P\boldsymbol{v}$ is the unique vector in W closest to $\boldsymbol{v}$. I.e., we have $\|\boldsymbol{v} - \boldsymbol{w}\| \geq \|\boldsymbol{v} - P\boldsymbol{v}\|$ for all $\boldsymbol{w} \in W$, with equality iff $\boldsymbol{w} = P\boldsymbol{v}$.*

Proof. Consider the decomposition $\boldsymbol{v} = P\boldsymbol{v} + (I-P)\boldsymbol{v}$. From the geometric meaning, we see that $P\boldsymbol{v}$ should be inside W, while $(I-P)\boldsymbol{v} \in \operatorname{Ran}(I-P) = \operatorname{Ker}(P) = \operatorname{Ran}(P)^\perp = W^\perp$.

Now our goal is to compare $\boldsymbol{v} - \boldsymbol{w}$ and $\boldsymbol{v} - P\boldsymbol{v}$. However, note that the vectors $\boldsymbol{v} - \boldsymbol{w}, \boldsymbol{v} - P\boldsymbol{v}, \boldsymbol{w} - P\boldsymbol{v}$ form a triangle. Furthremore, $\boldsymbol{w} - P\boldsymbol{v}$ is a linear combination of vectors in W, hence it is still in W. On the other hand, $\boldsymbol{v} - P\boldsymbol{v} = (I-P)\boldsymbol{v}$ is perpendicular to everything in W. Hence the triangle made of $\boldsymbol{v} - \boldsymbol{w}, \boldsymbol{v} - P\boldsymbol{v}, \boldsymbol{w} - P\boldsymbol{v}$ is a right triangle.

So by Pythagorean theorem, $\|\boldsymbol{v} - \boldsymbol{w}\|^2 = \|P\boldsymbol{v} - \boldsymbol{w}\|^2 + \|(I-P)\boldsymbol{v}\|^2 \geq \|(I-P)\boldsymbol{v}\|^2 = \|(\boldsymbol{v} - P\boldsymbol{v}\|^2$, and equality holds iff $\|P\boldsymbol{v} - \boldsymbol{w}\|^2 = 0$ iff $\boldsymbol{w} = P\boldsymbol{v}$.

(This proof is highly geometric. See if you can draw the picture. This is hilariously a geometric proof to an analytic property of an algebraic operator.) □

We have previously seen that for a subspace, it might have many oblique projections. But this is not the case any more. Another nicest thing about an orthogonal projection is that it is unique for the subspace.

Proposition 5.10.14. *Given a subspace $W \subseteq V$, then there is a unique orthogonal projection $P : V \to V$ with $\operatorname{Ran}(P) = W$.*

Proof. Uniqueness is guaranteed by the last proposition, since the image of $\boldsymbol{v}$ after orthogonal projection to W must be the unique vector in W closest to $\boldsymbol{v}$. Let us show existence.

One idea is to simply define $P\boldsymbol{v}$ as the unique vector in W closest to $\boldsymbol{v}$. Then we can manually and painstakingly verify that P is linear, and $P^2 = P$, and $W = \mathrm{Ran}(P) \perp \mathrm{Ker}(P)$.

Here is another approach. Suppose WLOG that $V = \mathbb{R}^n$. Find any ONB for W, say $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$. Then extend this to a full ONB of the whole space $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$.

For any $\boldsymbol{v} \in V$, we have an orthogonal decomposition $\boldsymbol{v} = \boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{v} + \cdots + \boldsymbol{w}_n\boldsymbol{w}_n^{\mathrm{T}}\boldsymbol{v}$. However, in this decomposition, the portion $\boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{v} + \cdots + \boldsymbol{w}_k\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{v}$ is inside W, while the rest $\boldsymbol{w}_{k+1}\boldsymbol{w}_{k+1}^{\mathrm{T}}\boldsymbol{v} + \cdots + \boldsymbol{w}_n\boldsymbol{w}_n^{\mathrm{T}}\boldsymbol{v}$ is perpendicular to W.

In particular, let $P = \sum_{i=1}^{k} \boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}$, then this is the desired projection.

Let us reformulate the approach above. (Consider this the third proof if you like.) Pick any ONB $Q = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_k)$ for W, and set $P = QQ^{\mathrm{T}}$. We aim to show that P is the desired orthogonal projection to W.

To start, $\mathrm{Ker}(P) = \mathrm{Ker}(QQ^{\mathrm{T}}) = \mathrm{Ker}(Q^{\mathrm{T}}) = \mathrm{Ran}(Q)^{\perp} = W^{\perp}$.

Since P is symmetric, we also have $\mathrm{Ran}(P) = \mathrm{Ran}(P^{\mathrm{T}}) = \mathrm{Ker}(P)^{\perp} = (W^{\perp})^{\perp} = W$. So far so good.

Finally, note that the (i, j) entry of $Q^{\mathrm{T}}Q$ is $\boldsymbol{e}_i^{\mathrm{T}}Q^{\mathrm{T}}Q\boldsymbol{e}_j = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j$. Since columns of Q form ONB, this is 1 if $i = j$ and 0 if $i \neq j$. Hence $Q^{\mathrm{T}}Q = I_{k\times k}$. So we have $P^2 = QQ^{\mathrm{T}}QQ^{\mathrm{T}} = QI_{k\times k}Q^{\mathrm{T}} = QQ^{\mathrm{T}} = P$. So P is the desired orthogonal projection. □

Here is an important remark. Compare the following two formula:

1. For a unit vector $\boldsymbol{u}$, the orthogonal projection to the line spanned by $\boldsymbol{u}$ is the matrix $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$.
2. For an ONB Q for a subspace W, the orthogonal projection to $\mathrm{Ran}(Q)$ is the matrix QQ^{T}.

It is quite obvious that the projection formula QQ^{T} is a generalization of our old formula $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$.

Furthermore, here is something else to be careful. Note that columns of Q do NOT form a basis for the whole space, merely for a subspace. So Q is not square. If $\dim W = k$, then Q is $n \times k$ and it is injective. So $P = QQ^{\mathrm{T}}$ is an $n \times n$ matrix. We alsk know that $\mathrm{Ran}(P) = W$, hence this matrix has rank exactly k. So P is NOT invertible when $k < n$.

However, $Q^{\mathrm{T}}Q$ would be a $k \times k$ matrix. And in fact, since columns of Q form a basis for W, we have $\mathrm{Ker}(Q) = \{\boldsymbol{0}\}$, and hence $\mathrm{Ker}(Q^{\mathrm{T}}Q) = \mathrm{Ker}(Q) = \{\boldsymbol{0}\}$. So $Q^{\mathrm{T}}Q$ is always invertible. In fact, as we have seen in the calculations in the proposition above, we always have $Q^{\mathrm{T}}Q = I_{k\times k}$.

Let me again stress these two formula. If columns of Q represent an ONB for a subspace W of dimension k, then

1. QQ^{T} is the orthogonal projection to W.
2. $Q^{\mathrm{T}}Q = I_{k\times k}$.

Now, if $P^2 = P$, then it is a (possibly oblique) projection. Can we tell which projections are orthogonal? We can indeed. This turns out to be surprisingly easy.

Proposition 5.10.15. *A square matrix P is an orthogonal projection iff $P^2 = P$ and $P^{\mathrm{T}} = P$. (I.e., orthogonal projection = symmetric projection.)*

Note that we implicitly used the inner product structure because we used transpose. And here we are refering to the dot product as the inner product. For abstract inner product spaces, pick orthonormal basis and then do this.

Proof. Suppose $P^2 = P$ and $P^{\mathrm{T}} = P$. Then $\mathrm{Ran}(P) = \mathrm{Ran}(P^{\mathrm{T}}) = \mathrm{Ker}(P)^{\perp}$. Done.

Suppose P is an orthogonal projection, then it is QQ^{T} as above. Then we can easily verify that $P^2 = P$ and $P = P^{\mathrm{T}}$. □

So orthogonal projections are uniquely defined, easy to identify, and very useful. To make it even better, we have another formula.

Proposition 5.10.16. *Given any subspace $W \subseteq \mathbb{R}^n$, let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ be a basis, and let $A = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_k \end{bmatrix}$. Then the orthogonal projection to W is $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$.*

In general, if A is injective, then $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ is the orthogonal projection to $\mathrm{Ran}(A)$.

Proof. Since A is injective, $A^{\mathrm{T}}A$ is also injective. (See previous homework.) But $A^{\mathrm{T}}A$ is also square, so it is bijective.

Columns of A is a basis for W. To find ONB for W, we perform Gram-Schmidt on A, which means we do QR decomposition $A = QR$ where Q has orthonormal columns and R is upper triangular with positive diagonal entries. So columns of Q form ONB for W, and R is invertible.

Now

$$\begin{aligned} A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} =& QR(R^{\mathrm{T}}Q^{\mathrm{T}}QR)^{-1}R^{\mathrm{T}}Q^{\mathrm{T}} \\ =& QR(R^{\mathrm{T}}R)^{-1}R^{\mathrm{T}}Q^{\mathrm{T}} \\ =& QRR^{-1}(R^{\mathrm{T}})^{-1}R^{\mathrm{T}}Q^{\mathrm{T}} \\ =& QQ^{\mathrm{T}}. \end{aligned}$$

So this is indeed the desired orthogonal projection. □

Now the proof above may seem like magic. Let us now try to demystify it by doing the following comparision of formula.

1. Given a (non-unit) vector $\boldsymbol{v}$, the orthogonal projection to the line spanned by $\boldsymbol{v}$ is $\frac{\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}$.
2. Given a matrix A with independent columns, the orthogonal projection to $\mathrm{Ran}(A)$ is $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$.

Look at the denominator of $\frac{\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}$ and the inversed portion of $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$, you shall see that they are pretty much the same formula, with the latter generalizing the former.

Remark 5.10.17. *In the expression $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$, why don't we expand the parenthesis?*

If you try, you see that $A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} = AA^{-1}(A^{\mathrm{T}})^{-1}A^{\mathrm{T}} = I$. Huh, this CANNOT be! What went wrong?

The failure of this logic lies in the fact that the identity $(AB) = B^{-1}A^{-1}$ has an ASSUMPTION! Only for square matrices this is true.

In our case, A is $n \times k$, so you CANNOT use this formula.

In fact, if $n = k$, then A being injective means it is bijective, $\mathrm{Ran}(A) = \mathbb{R}^n$ *the whole space. Then projection to the whole space obviously must be I. This is why expanding the parenthesis gives the identity matrix as our projection.*

You can prove that in general, for $n \times k$ matrix A, AA^{T} is $n \times n$ with rank the same as A, and $A^{\mathrm{T}}A$ is $k \times k$ with rank the same as A. In our case, A has rank k, hence $A^{\mathrm{T}}A$ must be invertible, and AA^{T} is not invertible when $k < n$.

5.10.3 Applications of orthogonal projections

As we have discussed before, orthogonal projections are highly useful because they gives approximations. Consider the following example.

Example 5.10.18. Suppose we want to solve the linear system:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \\ 3.001 \end{bmatrix}.$$

Oops, we have no solution. But do we declare failure? No we do not. Rather, we declare this to be FAKE NEWS and we firmly believe that there must be a solution. So what could happen? Well, it DOES

look suspicious. We ALMOST have a solution of $\boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, if only that third coordinates of $\boldsymbol{b}$ changes a tiny little bit. In fact, everything else are integers, why the hell is this coordinate 3.001? Probably due to some error.

Sometimes we collect data, and try to solve $A\boldsymbol{x} = \boldsymbol{b}$. We know there must be a solution, because this is a real life problem where the solution must exist. However, possibly due to unaccounted noise or minor errors, it appears our system has no solution. What to do now?

In this case, one should try to find the BEST $\boldsymbol{x}$, such that $A\boldsymbol{x}$ is as close as $\boldsymbol{b}$ as possible!

So what could $A\boldsymbol{x}$ be? Well, it is obviously always in $\mathrm{Ran}(A)$. We want the vector in $\mathrm{Ran}(A)$ that is closest to $\boldsymbol{b}$, i.e., we want to project $\boldsymbol{b}$ orthogonally to the subspace $\mathrm{Ran}(A)$.

Luckily we already know how to do this! If A is injective (it usually is in practice), we just need $\boldsymbol{b}' = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\boldsymbol{b}$ and solve $A\boldsymbol{x} = \boldsymbol{b}'$ instead.

Let us simplify a bit. We have $\boldsymbol{b}' = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\boldsymbol{b}$ and $A\boldsymbol{x} = \boldsymbol{b}'$. This means $A\boldsymbol{x} = \boldsymbol{b}' = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\boldsymbol{b}$. Since we assumed that A is injective, by left cancellation we have $\boldsymbol{x} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\boldsymbol{b}$. Yay! We now have a formula to find this $\boldsymbol{x}$, so that $A\boldsymbol{x}$ is as close to $\boldsymbol{b}$ as possible. This is the "closest" to a solution of $A\boldsymbol{x} = \boldsymbol{b}$.

Now the formula is actually not that good. To actually calculate $\boldsymbol{x}$ using this formula, you would need to calculate the inverse of some matrix, which takes a long time to do. Therefore, we can further simplify this expression to $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ instead, and solve this using Gaussian elimination. Finally, no matrix inversion! Finding $\boldsymbol{x}$ using $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ is in fact faster than finding $\boldsymbol{x}$ using $\boldsymbol{x} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\boldsymbol{b}$. (Also, finding the inverse of a matrix would usually results in magnifying the error term, if any initial error is involved.)

To sum up, how to find the best approximated solution? Given a system $A\boldsymbol{x} = \boldsymbol{b}$ which might not have a solution, we first apply A^{T} to both sides, and then solve $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ instead. ☺

Definition 5.10.19. *Given A, a least square solution to a system $A\boldsymbol{x} = \boldsymbol{b}$ is the solution to the system $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$.*

Why is it called the least square solution? Because that is what orthogonal projection would do.

Proposition 5.10.20. *For any linear system $A\boldsymbol{x} = \boldsymbol{b}$ which may or may not have a solution, then $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ always have a solution. Furthremore, if $\boldsymbol{x}_0$ is a solution, then $A\boldsymbol{x}_0$ is as close to $\boldsymbol{b}$ as possible.*

Proof. First of all, let us show that there is a solution. Note that the right hand side is

$$A^{\mathrm{T}}\boldsymbol{b} \in \mathrm{Ran}(A^{\mathrm{T}}) = \mathrm{Ker}(A)^{\perp} = \mathrm{Ker}(A^{\mathrm{T}}A)^{\perp} = \mathrm{Ran}(A^{\mathrm{T}}A).$$

Hence we can always find some $\boldsymbol{x}$ such that $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$. Done.

Now suppose $\boldsymbol{x}_0$ satisfies $A^{\mathrm{T}}A\boldsymbol{x}_0 = A^{\mathrm{T}}\boldsymbol{b}$. Let us show that $A\boldsymbol{x}_0$ is the unique vector in $\mathrm{Ran}(A)$ closest to $\boldsymbol{b}$.

First we perform full rank decomposition $A = BC$, where B is injective and C is surjective. Next we perform QR decomposition $B = QR$. So Q has orthonormal columns and R is invertible. By the lemma below, we have $\mathrm{Ran}(A) = \mathrm{Ran}(Q)$, hence the projection to $\mathrm{Ran}(A)$ is just QQ^{T}.

Now $A = QRC$, where RC is surjective. Let $S = RC$ for simplicity. Plug in $A = QS$ to $A^{\mathrm{T}}A\boldsymbol{x}_0 = A^{\mathrm{T}}\boldsymbol{b}$, we have

$$S^{\mathrm{T}}Q^{\mathrm{T}}QS\boldsymbol{x}_0 = S^{\mathrm{T}}Q^{\mathrm{T}}\boldsymbol{b}.$$

This simplifies to

$$S^{\mathrm{T}}S\boldsymbol{x}_0 = S^{\mathrm{T}}Q^{\mathrm{T}}\boldsymbol{b}.$$

Since S is surjective, S^{T} is injective and has left cancellation. So $S\boldsymbol{x}_0 = Q^{\mathrm{T}}\boldsymbol{b}$. Now apply Q on both sides, we see that $A\boldsymbol{x}_0 = QQ^{\mathrm{T}}\boldsymbol{b}$. So our porposition is correct. □

Lemma 5.10.21. *For any $m \times n$ matrix A, we do the full rank decomposition $A = BC$ and then QR decomposition $B = QR$. Then $\mathrm{Ran}(A) = \mathrm{Ran}(Q)$.*

Proof. Homework. Use definition of surjectivity of C. □

The conclusion is this. If you want to solve $A\boldsymbol{x} = \boldsymbol{b}$ but there is no solution, then simply solve $A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ instead. This gives you as close to a solution as possible.

Let me give you a practical example here.

Example 5.10.22. We uses CT scans to find tumor in a person. To simplify the problem, suppose this person only consists of four pixels, each pixel is a square with unit side length. Each pixel could be bones, flesh, blood or tumors. See the picture.

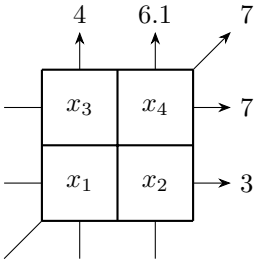


Figure 5.10.1: CT scan decay

Suppose we uses X-rays to go through these pixels, and the strength of the X-ray would decay by $1, 2, 3, 4$ units per unit length through bones, flesh, blood and tumors respectively. For example, if the lower left pixel is a bone and the lower right cell is flesh, then a ray going horizontally through the lower cells would decay by a total of $1 + 2 = 3$ units. And if the upper left cell is blood and the lower right cell is flesh, then a ray going diagonally from upper left to lower right would decay by a total of $\sqrt{2}(1 + 3) = 4\sqrt{2}$ units.

Now suppose the rate of decay in each pixel per unit length is x_1, x_2, x_3, x_4, and we uses five X-rays as shown. The total decay of each ray after measurement, is also shown. Can you figure out what is the content of each cell?

Our linear system is

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \\ 7 \\ 4 \\ 6.1 \end{bmatrix}.$$

If you attempt to solve this directly, you shall see that there is no solution. This is in fact always the case in practice. Dust in the air, trumbling patients and so on, there are always some noises that will give rise to inconsistencies.

However, we can try to find the least square solution. We consider the new system

$$\begin{bmatrix} 1 & 0 & \sqrt{2} & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & \sqrt{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \sqrt{2} & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & \sqrt{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 7 \\ 7 \\ 4 \\ 6.1 \end{bmatrix}.$$

This simplifies to

$$\begin{bmatrix} 4 & 1 & 1 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 2 & 1 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 7 + 7\sqrt{2} \\ 9.1 \\ 11 \\ 13.1 + 7\sqrt{2} \end{bmatrix}.$$

Now the matrix on the left is invertible. So we solve for the least square solution, and find out that $x_1 \approx 0.950, x_2 \approx 2.075, x_3 \approx 3.025, x_4 \approx 4.000$. So we see that the four pixels x_1, x_2, x_3, x_4 corresponds to bones, flesh, blood and tumors. The tumor is located in pixel four, the upper right pixel. Problem solved.

Note that the matrix $B = \begin{bmatrix} 4 & 1 & 1 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 2 & 1 & 1 & 4 \end{bmatrix}$ depends only on the way we positioned our machine. So we are going to solve $B\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ for many different $\boldsymbol{b}$ again and again. So in practice, it is better to first perform the LU decomposition of B beforehand, and then solve for the solution according to each patient's $\boldsymbol{b}$. ☺

Example 5.10.23. Suppose we want to determine how our education effect your income. We collect data from n people, $(x_1, y_1), ..., (x_n, y_n)$, where x_i is the number of years of education received by the i-th person, and y_i is the eventual income of this person.

Suppose we believe in a linear model, that education and income should vaguely lies around some line $y = kx + b$. In this case, we aim to find the line that BEST FIT our data. Then I can claim that on average, an extra year of education shall increase your income by k.

So our model is $Y = kX + b + E$ where k, b are unknown constants and E represent factors other than education. X, Y here are random variables that represent the years of educations and income of a random person. How to find the best k, b to fit our data?

Let $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $\boldsymbol{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, then ideally, IF the data lies on a line perfectly, we should have $\boldsymbol{y} = k\boldsymbol{x} + b\boldsymbol{u}$. Here $\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}$ are all known, and we aim to solve for the unknown k and b.

Rearrange this, we are trying to solve $\begin{bmatrix} \boldsymbol{x} & \boldsymbol{u} \end{bmatrix} \begin{bmatrix} k \\ b \end{bmatrix} = \boldsymbol{y}$, which is a linear system. Now, of course there are other factors that influences one's income, so our data CANNOT lies on a line perfectly. So this system has no solution. What should we do? We go for the LEAST SQUARE solution instead.

So we are looking at the new system

$$\begin{bmatrix} \boldsymbol{x}^{\mathrm{T}} \\ \boldsymbol{u}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} & \boldsymbol{u} \end{bmatrix} \begin{bmatrix} k \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}} \\ \boldsymbol{u}^{\mathrm{T}} \end{bmatrix} \boldsymbol{y}.$$

This simplifies to

$$\begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix} \begin{bmatrix} k \\ b \end{bmatrix} = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}.$$

The solution is $k = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$ and $b = \overline{y} - k\overline{x}$, where $\overline{x} = \frac{1}{n}\sum x_i, \overline{y} = \frac{1}{n}\sum y_i$.

Now note that $k = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$, and covariance is an "inner product" for random variables. So this is the projection formula of the random variable Y to the random variable X, wow! (Compare with the formula $\frac{\langle \boldsymbol{v}, \boldsymbol{w} \rangle}{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$ of projecting $\boldsymbol{w}$ to $\boldsymbol{v}$.) This surprisingly makes a lot of sense. If we are trying to predict Y via X, the best we can approximate is to get the projection of Y to X.

You can also check that in this case, $\mathbb{E}(E) = 0$ and $\mathrm{Var}(E)$ is minimized under this condition. So the "error term" is as small as possible.

This is a very simple case of ***linear regression***. ☺

Example 5.10.24. Does death penalty discourage murder? It seems logical. But theory must be tested by data. So let us see a way to test this.

Let $X = 1$ if a country or a state has death penalty, and $X = 0$ if a country or a state has no death penalty. Let Y be the murder rate in a state or a country. Then if you pick many many countries, you get many many data $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$. Here all x_i are 0 or 1, while all y_i are some real numbers.

Let us fit these data into a line $Y = kX + b + E$ with the best possible k, b and $\mathbb{E}(E) = 0$. Then what does this mean? Well, if a country has no death penalty, then $X = 0$, and then we see that the average murder rate is b. If a country has death penalty, then $X = 1$, and then we see that the average murder rate is $b + k$. In particular, k is EXACTLY the effect of death penalty on murder rate.

After collecting data, we can compute $k = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$, and $b = \bar{y} - \frac{\text{Cov}(x,y)}{\text{Var}(x)}\bar{x}$. It turned out that, on almost all the studies conducted, k is positive!!! Oops. It turned out that death penalty is actually associated to a higher murder rate. That is very counter-intuitive. What happened?

Attempted explanation 1: Correlation is not causality. Maybe it is the other way around: Countries with higher murder rate is more likely to punish them severely, so maybe high Y value caused $X = 1$, instead of $X = 1$ causing high Y value. Unfortunately, this causality effect can be addressed with better statistic approach and better ways to collect data. (E.g., looking at countries that changed their death penalty laws, and compare the murder rate before and after.) To my knowledge, it turned out that even if we strictly investigate the causality of X on Y, k is STILL positive. So this explanation failed. Death penalty in fact lead to a higher murder rate.

Attempted Explanation 2: One argument is that death penalty encourage murder because death penalty is murder in itself. You are killing criminals, sure, but you planned and premeditated and then killed them. The message you are sending with a death penalty is that it is OK to kill when you have a justified reason. So people might think it is OK to kill when they feel justified for some reason.

Attempted Explanation 3: There is a minute but non-zero chance that all previous studies are all collectively unlucky, and we simply get the unlucky data. Maybe death penalty do deter murder. Who knows?

Also keep in mind that this is NOT an endorsement to the total revokation of death penalty. There are other philosophical concerns to think about. (E.g., Immanuel Kant.) ☺

Part III

Coordinate Invariants

Chapter 6

Determinants

6.1 Introduction

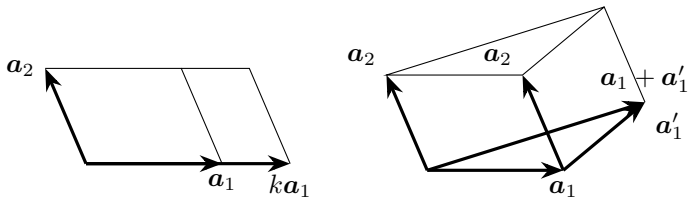
6.1.1 Oriented Area and Oriented Volume

Given a parallelogram in $\mathbb{R}^2$, suppose two of its adjacent edges are $\boldsymbol{v} = \begin{bmatrix} a \\ c \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} b \\ d \end{bmatrix}$. Then instead we shall focus on the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Our goal here is to find the area of the parallelogram, i.e., $Area(\boldsymbol{v}, \boldsymbol{w})$ or $Area(A)$. How can we do this?

First let us list some obvious properties of area.

Example 6.1.1. Consider the parallelograms $\begin{bmatrix} \boldsymbol{v} & \boldsymbol{w} \end{bmatrix}$ and $\begin{bmatrix} k\boldsymbol{v} & \boldsymbol{w} \end{bmatrix}$ in the graph below. Obviously if I multiply an edge by k, then the area is multiplied by k. So $Area(k\boldsymbol{v}, \boldsymbol{w}) = Area(\boldsymbol{v}, k\boldsymbol{w}) = kArea(\boldsymbol{v}, \boldsymbol{w})$.



(a) Scaling (b) Vector Addition

Figure 6.1.1: Vectors and Area

Similarly, consider the case of $Area(\boldsymbol{a}_1 + \boldsymbol{a}_1', \boldsymbol{a}_2)$ above. Treating the direction of $\boldsymbol{a}_2$ as "base" and the perpendicular direction as "height", you can see that $Area(\boldsymbol{a}_1 + \boldsymbol{a}_1', \boldsymbol{a}_2) = Area(\boldsymbol{a}_1, \boldsymbol{a}_2) + Area(\boldsymbol{a}_1', \boldsymbol{a}_2)$. It seems to suggect that area is bilinear!

However, that is not the case. The CORRECT formula is $Area(k\boldsymbol{v}, \boldsymbol{w}) = Area(\boldsymbol{v}, k\boldsymbol{w}) = |k|Area(\boldsymbol{v}, \boldsymbol{w})$, because area is always positive. And we in fact have $Area(\boldsymbol{a}_1 + \boldsymbol{a}_1', \boldsymbol{a}_2) = Area(\boldsymbol{a}_1, \boldsymbol{a}_2) \pm Area(\boldsymbol{a}_1', \boldsymbol{a}_2)$, where the sign depends on whether $\boldsymbol{a}_1, \boldsymbol{a}_1'$ are on the same sides of $\boldsymbol{a}_2$ or not. (Can you draws these out?)

So area is NOT bilinear. ☺

This is an annoying situation. Area of a parallelogram is almost bilinear, but not truly. This is largely due to the need for area to be positive, as is very evident in the formula $Area(k\boldsymbol{v}, \boldsymbol{w}) = Area(\boldsymbol{v}, k\boldsymbol{w}) = |k|Area(\boldsymbol{v}, \boldsymbol{w})$.

So, if we allow negative area, then problem solved. The "signed" area, or ***oriented area*** will be bilinear.

What is oriented area? Loosely speaking, for each shape, we can "orient" the shape to be clockwise (CW) or counter-clockwise (CCW). If we give it a CCW orientation, then it will have positive area. And if we give it a CW orientation, then it will have negative area.

Example 6.1.2. You might have already encountered the concept of oriented area. For example, in calculus, a definite integral is the area below the curve. But sometimes the curve goes below the x-axis, resuting in a negative area. See the graph below.

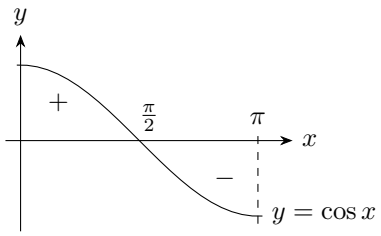


Figure 6.1.2: Oriented Area

In some sense, you can use the "clockwise" and "counterclockwise" to tell orientation. For example, if we go along the positive x-axis, you will see that we are going counterclockwise around the positive region, and clockwise around the negative region. A very informal intuition is to assume that the boundary of your region is oriented clockwise (CW) or counterclockwise (CCW), where a CCW oriented region has positive area, and a CW oriented region has a negative area. ☺

So now let us see how this oriented area could be calculated. Given two $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^2$, we use $\det(\boldsymbol{v}, \boldsymbol{w})$ or $\det\begin{bmatrix}\boldsymbol{v} & \boldsymbol{w}\end{bmatrix}$ to denote the oriented area of the corresponding parallelogram, where we go along the edge vector $\boldsymbol{v}$ first, and then go along $\boldsymbol{w}$ from the end point of $\boldsymbol{v}$ to the end point of $\boldsymbol{v}+\boldsymbol{w}$. This will give us a clear idea about whether the orientation is CW or CCW.

In particular, $\begin{bmatrix}\boldsymbol{v} & \boldsymbol{w}\end{bmatrix}$ and $\begin{bmatrix}\boldsymbol{w} & \boldsymbol{v}\end{bmatrix}$ both represent the same parallelogram, but with opposite orientation.

Now here are some properties of oriented area.

1. (Normalized) First of all, the unit square going CCW should have area 1. I.e., we want $\det(\boldsymbol{e}_1, \boldsymbol{e}_2) = \det(I) = 1$.
2. (Bilinear) We do this because we want the oriented area to be bilinear. I.e., we want $\det(a\boldsymbol{u}+b\boldsymbol{v}, \boldsymbol{w}) = a\det(\boldsymbol{u}, \boldsymbol{w}) + b\det(\boldsymbol{v}, \boldsymbol{w})$.
3. (Anti-symmetry) If we swap the two edges, it is the same parallelogram but with reversed orientation. I.e., we want $\det(\boldsymbol{v}, \boldsymbol{w}) = -\det(\boldsymbol{w}, \boldsymbol{v})$.

Note that these properties immediately implies two interesting results:

1. (Flat parallelogram has no area) Swapping the two vectors gives $\det(\boldsymbol{v}, \boldsymbol{v}) = -\det(\boldsymbol{v}, \boldsymbol{v})$, which implies that $\det(\boldsymbol{v}, \boldsymbol{v}) = 0$. Indeed, if the two edges of the parallelogram coincide, then the area must be zero. In fact, conversly, if we have bilinearity and the fact that $\det(\boldsymbol{v}, \boldsymbol{v}) = 0$, then we can also deduce that $\det(\boldsymbol{v}, \boldsymbol{w}) + \det(\boldsymbol{w}, \boldsymbol{v}) = \det(\boldsymbol{v}+\boldsymbol{w}, \boldsymbol{v}+\boldsymbol{w}) - \det(\boldsymbol{v}, \boldsymbol{v}) - \det(\boldsymbol{w}, \boldsymbol{w}) = 0$, so we have antisymmetry. Note that we essentially used the polarization identity here.
2. (Shearing preserves area) We have $\det(\boldsymbol{v}+\boldsymbol{w}, \boldsymbol{w}) = \det(\boldsymbol{v}, \boldsymbol{w}) + \det(\boldsymbol{w}, \boldsymbol{w}) = \det(\boldsymbol{v}, \boldsymbol{w})$.

Now, to calculate $\det(A)$, we can try to swap/scale/shear columns of A, and $\det(A)$ should change accordingly. This gives us an idea to calculate this. For $A = \begin{bmatrix}a & b\\ c & d\end{bmatrix}$, assuming we do not see zero in the denominators below (I.e., $d \neq 0$ and $ad - bc \neq 0$), then we have

$$\det(A) = \det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \det\begin{bmatrix} a - \frac{bc}{d} & b \\ 0 & d \end{bmatrix} = \det\begin{bmatrix} a - \frac{bc}{d} & 0 \\ 0 & d \end{bmatrix} = (a - \frac{bc}{d})d \det\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = ad - bc.$$

This process corresponds to shearing a parallelogram until the two edges are on the coordinate axis. See the graph below, where $A = \left[\boldsymbol{a}_1, \boldsymbol{a}_2\right], \left[\boldsymbol{b}_1, \boldsymbol{b}_2\right], \left[\boldsymbol{c}_1, \boldsymbol{c}_2\right]$ are the matrices in the first three steps of the above calculation.

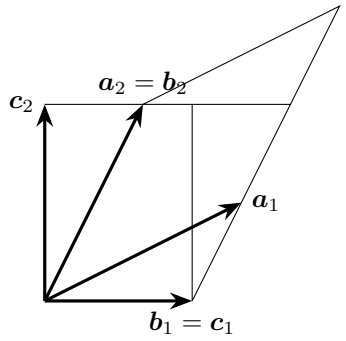


Figure 6.1.3: Column Shearing and Area

Definition 6.1.3. *Given a* 2×2 *matrix* $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, *we define* $\det(A) = ad - bc$.

You can now easily verify that this satisfy all the desired relations. I.e., this is INDEED a well-defined oriented area.

Now we have tackled the two-dimensional case of parallelograms, what about parallelepipeds? Can we define an oriented volume?

Example 6.1.4. My heart is on my left side. You can rotate me, shear me, stretch me, walk me around, throw me over the moon, and my heart will always on my left side. If a person has his/her heart on the right side, then this person must not be me, right?

But one day, I lookd into a mirror. The person in the mirror will have his/her heart on the right side. As we can see, the orientation is reversed. ☺

Now, there are arguments that one can think of 3D orientations as "inward" and "outward". Also, you can imagine the orientation refers to "positive mass" and "negative mass", which is sometimes used to solve some problems in physics. I myself think of the positive volume as something in the REAL world, and negative volume as something in a MIRROR world. Feel free to do these things if it helps. However, as we move on into higher and higher dimensions, we will lose sight of our ability to do this. At this point, it is better to revert back to the basic of WHAT IS AN ORIENTED VOLUME.

Again, we want oriented volume to to satisfy analogous properties as oriented area. Given three vectors in $\mathbb{R}^3$, we can imagine that they form a parallelepiped. Then we want the following properties:

1. (Normalized) We want $\det(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3) = \det(I) = 1$.
2. (Multilinear) We want $\det(a\boldsymbol{u} + b\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = a\det(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) + b\det(\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$, and the same for the second column and for the third column.
3. (Alternating) If we swap the two edges, it is the same parallelepiped but with reversed orientation. I.e., we want $\det(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = -\det(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v})$, and the same thing if we swap any pair of edges in general.

The first one is trivial. The second one can be seen geometrically. Treating the parallelogram $(\boldsymbol{x}, \boldsymbol{y})$ as the base, and the orthogonal direction as height, you shall see that $\det(\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = \det(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) + \det(\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$

and $\det(k\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) = k\det(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y})$. For the third one, it is easier to understand with an example. Consider the parallelepiped $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$. Reflect this about the plane $x = y$ in $\mathbb{R}^3$, and you will get the parallelepiped $(\boldsymbol{e}_2, \boldsymbol{e}_1, \boldsymbol{e}_3)$. In general, swapping a pair of edges means doing some reflection.

Again, note that these properties immediately implies two interesting results:

1. (Flat parallelepiped has no volume) If two vectors among $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ are the same, then there is no volume and $\det(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = 0$.
2. (Shearing preserves volumn) We have $\det(\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{w}) = \det(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$.

Using these ideas, we can establish the formula. Note that the formula is a lot uglier and not nice at all.

Definition 6.1.5. *Given a* 3×3 *matrix* $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$, *we define* $\det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$.

There are many smart ways to memorize this formula, but honestly there is no need. Such methods are usually for 3×3 matrices only, so using them has the danger of doing higher dimensional determinants wrong.

We can go on similarly, and in general, given vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$, then the ***determinant*** of them $\det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ is the oriented n-dimensional volumn of the corresponding n-dimensional parallelotope. We may also think of the input vectors as columns, forming an $n \times n$ square matrix A, and talk about $\det(A)$.

But the four dimension determinant formula will have 24 terms in the determinant. In general, the determinant of $n \times n$ matrices will have $n!$ terms. Instead, this is how we define determinants.

Definition 6.1.6. *We define an* $n \times n$ *determinant to be a function* $\det : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}$ *with* n *input vectors from* $\mathbb{R}^n$*, such that*

1. *(Identity)* $\det(I) = 1$.
2. *(Multiplinear)* $\det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ *is linear on each input* $\boldsymbol{v}_i$.
3. *(Alternating) If we swap the* i*-th input* $\boldsymbol{v}_i$ *and the* j*-th input* $\boldsymbol{v}_j$ *for any* $i \neq j$*, then the determinant is negated.*

Theorem 6.1.7. *For each* n*, determinant exists and is unique.*

We do the proof of this theorem later. For now, let us focus on some more intuitions.

Example 6.1.8. Suppose A is not invertible. Then columns of A are linearly dependent, i.e., they would fail to span $\mathbb{R}^n$ as expected, and would only span some smaller dimensional thing. As a result, the n-dimensional oriented volume of A would be zero. So $\det(A) = 0$. ☺

Example 6.1.9. Suppose we have a diagonal matrix D. Then as a parallelotope, all edges are orthogonal to each other! So the area is very easy: it is simply the product of all side lengths, i.e., all the diagonal entries of D.

We can generalize this a bit more. Consider an upper triangular matirx, say $U = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$. Taking the first edge as "base" which has length a, then the height is exactly c. So the area is exactly ac. Here b does not matter at all, because it effects neither the base nor the height. So $\det(U) = ac$.

What about $U = \begin{bmatrix} a & b & d \\ 0 & c & e \\ 0 & 0 & f \end{bmatrix}$? Taking the first two edges as "base", we see that the base area is ac, while the height is f. So the volume is $\det(U) = acf$.

You can probably see immediately that for any triangular matrix T, then $\det(T)$ is the product of diagonal entries.

Similarly, consider a block diagonal matrix $M = \begin{bmatrix} A & O \\ O & B \end{bmatrix}$ where A, B are square. Then the A portion and the B portion are orthogonal to each other. It is not hard to see that we have $\det(M) = \det(A)\det(B)$. In fact, it is also not hard to see that $\det \begin{bmatrix} A & C \\ O & B \end{bmatrix} = \det(A)\det(B)$ as well. Taking A portion as "base", then the corresponding height is exactly $\det(B)$, and C is irrelavant. ☺

6.1.2 Volume Scaling Factor

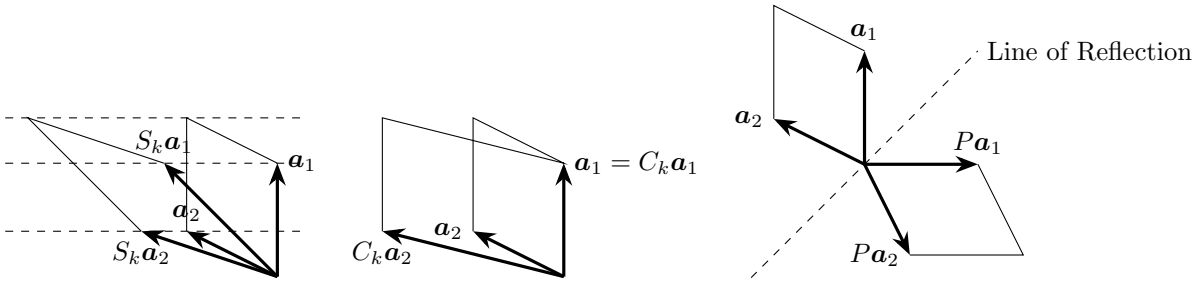
So far, by thinking of a square matrix A as an n-dimensional parallelotope in $\mathbb{R}^n$, we realized that if we do column operations, AE, then $\det(A)$ and $\det(AE)$ are related. Swapping columns would negate the determinant, shearing columns will preserve the determinant, and finally, scaling columns will scale the determinants. This subsection is devoted to row operations, $\det(EA)$.

But again, let us try to give it a meaning. Given a square matrix A, you may think of it as a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^n$. Given a parallelotope $B = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}$, then $AB = \begin{bmatrix} A\boldsymbol{v}_1 & \dots & A\boldsymbol{v}_n \end{bmatrix}$ is the image of the original parallelotope after the transformation A.

So here is the big question: in general, how would A changes volumes of parallelotope? What is $\frac{\det(AB)}{\det(B)}$? Let me spoil the answer first: It is always a constant, independent of the input parallelotope B. A would simply change all parallelotope volume by the same factor. This is the volume scaling factor of A, and let us write it as $\delta(A)$. For examlpe, rotations will always have a volume scaling factor of 1, and reflections will always have a volume scaling factor of -1.

Let us see some examples here.

Example 6.1.10. Suppose we have a parallelogram, say $\begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} = \begin{bmatrix} 0 & -0.8 \\ 1 & 0.4 \end{bmatrix}$. We can try to apply shearings $S_k = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$, scalings $\begin{bmatrix} k & 0 \\ 0 & 1 \end{bmatrix}$ or swapping $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ to it. As you can see from the graph below, when we apply S_{-1}, C_2, P to it, the area is changed by a fixed factor. Shearing again preserves the oriented area, swapping preserves the absolute area but changed the orientation, and finally scaling just scale the area.



(a) Shearing (b) Scaling (c) Swap

Figure 6.1.4: Row Operations and Area

Note that these row operations give an interesting contrast to column operations. For example, if we compare $\det(A)$ with $\det(AE)$ where E is a shearing, then we are shearing one edge of the parallelogram along the direction of another. However, if we are comparing $\det(A)$ with $\det(EA)$, then we are shearing the whole coordinate charts. In our case, the entire y-axis is sheared in the direction of the x-axis by our linear transformation E, and the parallelogram simply changes with it.

In these sense, the E of AE are applying changes to the edges of the parallelogram A, where as the E of EA are applying changes to the whole $\mathbb{R}^2$, and A simply changes along. If we use four matches to enclose

a parallelogram A on a paper, then AE is trying to move/stetch/rotate the matches, while EA is trying to move/stetch/rotate the whole paper.

Alternatively, if you think of E of EA as a change of coordinate process, then the underlying parallelogram is in fact unchanged, and we are simply looking at A through a different perspective, e.g., squint our eyes, tilt our heads, apply a distorting mirror, these sort of things. And A will end up appearing to have a different area from before. The change from A to EA is nominal but not essential. In contrast, AE means we are essentially changing the parallelogram we are studying.

If these discussions only serves to confuses you, then you are also welcome to just understand AE and EA in terms of column/row operations.

But either way, the pattern is preserved. Shearing fixes the oriented area, scaling scales the area, and swapping flips the orientation. These are true for both EA and AE. And we have $\delta(S_k) = 1, \delta(C_k) = k, \delta(P) = -1$. ☺

So in the case of elementary matrices, they indeed have a corresponding volume scaling factor, independent of the input parallelotope. What about matrices in general?

Now recall that, given any invertible matrix A, it is the product of many elementary matrices, $A = E_1 \dots E_k$. As a result, we see that

$$\det(AB) = \det(E_1 \dots E_k B) = \delta(E_1) \det(E_2 \dots E_k B) = \cdots = \delta(E_1) \dots \delta(E_k) \det(B).$$

In particular, we see that $\delta(A) = \prod \delta(E_i)$. As you can see, this depends ONLY on A, and NOT on B at all.

Example 6.1.11. Suppose $A = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$. Let us calculate $\delta(A)$.

Note that $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. The right matrix preserves the area, while the left matrix scale the area by 2. So $\delta(A) = 2$. ☺

What if A is not invertible? Then $\mathrm{Ran}(A)$ is less than n-dimensional. So any input B, AB will have zero n-dimensional volume. So $\delta(A) = 0$. Either way, we see that each linear map has a unique fixed volume scaling factor.

Definition 6.1.12. *For each $n \times n$ matrix A, we define its volumn scaling factor $\delta(A)$ to be the number such that, for any input parallelotope B, $\det(AB) = \delta(A) \det(B)$. (I.e., the linear map A scales the volumn by $\delta(A)$ always.)*

Now what is this $\delta(A)$? Note that, since AB as a linear map means we do B and then do A, obviously $\delta(AB) = \delta(A)\delta(B)$.

Corollary 6.1.13. $\delta(AB) = \delta(A)\delta(B)$.

Proof. For any parallelotope C, we have $\det(ABC) = \delta(A) \det(BC) = \delta(A)\delta(B) \det(C)$. Since this is true for all C, by definition $\delta(AB)$ must be the constant $\delta(A)\delta(B)$. □

For any elementary matrix E, $\delta(AE) = \delta(A)\delta(E)$ by definition. This immediatly implies the following:

1. Taking E as a swapping, we see that swapping columns of A will negate $\delta(A)$.
2. Taking E as a scaling, we see that scaling columns of A will scale $\delta(A)$.
3. Taking E as a shearing, we see that shearing columns of A will scale $\delta(A)$.
4. Obviously $\delta(I) = 1$, as the identity map would not change anything.

These looks oddly suspicious! They are exactly the defining properties of determinant!

Proposition 6.1.14. $\delta(A) = \det(A)$.

Proof. We have $\delta(AB) = \delta(A)\det(B)$ by definition of the volume scaling factor. Taking $B = I$, we see that $\delta(A) = \det(A)$. □

So as a function on square matrices, the determinant is the same as the volume scaling factor. (Intuitively this is also obvious. To find the volume scaling factor, simply check out what would happen to a unit cube. The map A send the unit cube I to the parallelotope A, so $\delta(A) = \det(A)$.)

Finally, we also have an interesting property.

Proposition 6.1.15. $\det(A) = \det(A^{\mathrm{T}})$.

Proof. Suppose A is not invertible. Then both A, A^{T} are not invertible, so $\det(A) = 0 = \det(A^{\mathrm{T}})$.

If A is invertible, then we can reduce A to I through a series of column operations. But then we can do the corresponding row operations to A^{T}, and it will be reduced to $I^{\mathrm{T}} = I$.

Since these operations determines the determinant, we see that $\det(A) = \det(A^{\mathrm{T}})$.

To be more detailed, if $A = E_1 \dots E_k$ for elementary matrices $E_1, \dots, E_k$, then $A^{\mathrm{T}} = E_k^{\mathrm{T}} \dots E_1^{\mathrm{T}}$. Since $\det(E_i) = \det(E_i^{\mathrm{T}})$ for all elementary matrices, we have $\det(A) = \prod \det(E_i) = \prod \det(E_i^{\mathrm{T}}) = \det(A^{\mathrm{T}})$. □

So now we have an alternative understanding of the determinant. We can think of $\det(A)$ as the n-volume of the n-parallelotope A, or as the volume scaling factor of the linear map A. But the latter perspective gives us the following two very interesting additional properties of the determinant.

1. $\det(AB) = \det(A)\det(B)$. This is obvious and natural from the volume scaling factor perspective.
2. $\det(A) = \det(A^{\mathrm{T}})$. This is because row operations and column operations do the same thing to $\det(A)$.

It is also very interesting if we combine this with some of our previous knowledges.

Example 6.1.16. Consider the LDU decomposition $A = LDU$. Here L and U are unit triangular, so their determinants are both one. So $\det(A) = \det(D)$. If you think about this, U, L means we are shearing the edges and also shearing the space, until we end up with a parallelotope D whose edges are on the coordinate axes. All these shearings preserves volume, so $\det(A) = \det(D)$ is simply the product of all edge lengths of D, i.e., all diagonal entries of D. (In some textbooks, i.e., Gilbert Strang, diagonal entries of D in the LDU decomposition is called the pivot values for A. This is because $L^{-1}A$ gives the row echelon form DU where pivots will be the corresponding diagonal entries of D.)

Now consider the QR decomposition $A = QR$. Q here is a rotation/reflection, so its volume scaling factor must be $\det(Q) = \pm 1$. R is an upper triangular matrix with positive diagonal entries, so $\det(R) > 0$. So in this sense, we have $\det(A) = \det(Q)\det(R)$, where $\det(R)$ is the absolute volumn of the parallelotope A, and $\det(Q)$ is the orientation of A. ☺

Remark 6.1.17. *Note that, A not only scales the volume of any input parallelotope by $\det(A)$. In fact ANY shape in $\mathbb{R}^n$, after transformation by A, will have its n-dimensional volume scaled by $\det(A)$. (To see this, chop up the shape into infinitesimal tiny parallelotopes, and take limit. This is a very standard argument in calculus.)*

This fact is crutial in multivariable calculus. In regular calculus, we integrate $\int f(x)dx$, and here dx intuitively refers to a infinitesimal tiny arrow in the direction of positive x-axis. We add up all the tiny vectors $f(x)dx$, and the result is $\int f(x)dx$.

In multivariable calculus, we integrate $\int f(x,y)dxdy$, and here $dxdy$ intuitively refers to an oriented parallelogram made by tiny vectors dx and dy. So we add up all the tiny oriented area $f(x,y)dxdy$, and the result is $\int f(x,y)dxdy$

Suppose we are doing $\int f(2x+y,y)dxdy$, and we want to do a change fo variable into $dzdy$ where $z = 2x + y$. Then $dz = 2dx + dy$ is a tiny arrow in the $\begin{bmatrix}2\\1\end{bmatrix}$ direction, and the resulting $dzdy$ will have a different oriented area from $dxdy$. Note that the map $A = \begin{bmatrix}2 & 0\\1 & 1\end{bmatrix}$ is exactly the linear transformation that changes the parallelogram $dxdy$ to $dzdy$. This means $dzdy = \det(A)dxdy = 2dxdy$. So we have $\int f(2x+y,y)dxdy = \int f(z,y)(\frac{1}{2}dzdy)$. If you miss the factor $\frac{1}{2}$, then your calculation is wrong.

6.2 Permutation Issue

6.2.1 Parity of a Permutation

Now all discussions so far depends on the existance and uniqueness of the concept of higher dimensional oriented volume. In lower dimensions, we can work it out and simply produce a formula for $\det(A)$. But for higher dimensions, we need to do this more rigorously. We want to rigorously define "oriented volume" in any dimension.

Example 6.2.1. Say we have an oriented 4-parallelotope $\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4\end{bmatrix}$. Now $\begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_1 & \boldsymbol{v}_3 & \boldsymbol{v}_2\end{bmatrix}$ describes the same parallelotope, but does it have the same orientation as before, or the opposite orientation?

Well, we need to check the number of swaps. One might do

$$\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_1\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_1 & \boldsymbol{v}_3 & \boldsymbol{v}_2\end{bmatrix}.$$

So we see that two swaps are needed. So they should have the same orientation.

But hold on a second! What if I do permutations differently? What if by some other way of swapping things, I go from $\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4\end{bmatrix}$ to $\begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_1 & \boldsymbol{v}_3 & \boldsymbol{v}_2\end{bmatrix}$ wiht odd number of swaps, hence they would have different orientation? Then we would have a contradiction at hour hand. The whole concept of oriented volume would collapse and become total nonesense.

Luckily, that may never happen. For example, let us do a different series of swaps. One might have

$$\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_3 & \boldsymbol{v}_2 & \boldsymbol{v}_1 & \boldsymbol{v}_4\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_3 & \boldsymbol{v}_2 & \boldsymbol{v}_4 & \boldsymbol{v}_1\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_1\end{bmatrix} \to \begin{bmatrix}\boldsymbol{v}_4 & \boldsymbol{v}_1 & \boldsymbol{v}_3 & \boldsymbol{v}_2\end{bmatrix}.$$

Then we have four swaps, still even. ☺

So, how is this "orientation" thing defined? It seems to be someting induced by reflections. Consider a swapping matrix P, which is supposed to change orientations. Given a permutation matrix P, if it is done via an even number of swaps, then it should preserve orientation. If it is done via an odd number of swaps, then it should negate the orientation. This motivates us to do the following definition:

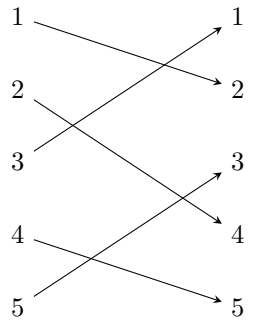
Definition 6.2.2. *A permutation is called an* ***even permutation*** *if it is the composition of even number of swaps. A permutation is called an* ***odd permutation*** *if it is the composition of odd number of swaps.*

Our goal is to rigorously show that, each permutation must be either even, or odd, but not both. This is the very fundation of orientation. Without this fact, then orientation could not ever be consistently defined. We now do this via some super fun diagrams.

Lemma 6.2.3. *Any permutation is the composition of swaps.*

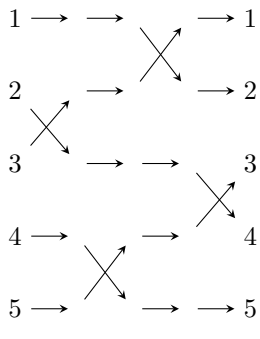
Proof. There are many many ways to do this. However, here we give a proof based on diagrams, and we merely provide an example here. You should be able to generalize this yourself.

Suppose a permutation sends $1, 2, 3, 4, 5$ to $2, 4, 1, 5, 3$. We can use the diagram below to denote this:



.

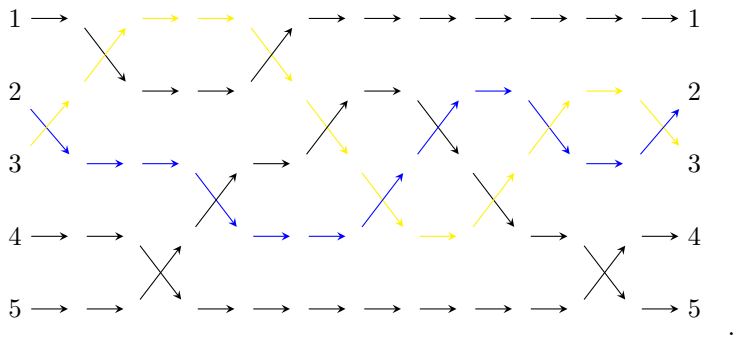
Did you see the crossings? Now, we repreduce these crossings via swaps, in the following way:



As you can see, we have decomposed a permutation into a series of swaps. It is easy to generalize this to any permutation. □

Lemma 6.2.4. *The identity permutation can only be the composition of even number of swaps, but not odd number of swaps.*

Proof. Suppose we decompose the identity permutation into a series of swaps. For illustration purpose, you can look at the following diagram:



Note that the composition is identity means each path goes ends up at the same number as where they started. I claim that for any such diagram, we must have an even number of "crossings" in total.

For example, consider the blue path from 2 to 2 and the yellow path from 3 to 3. Since the yellow path started below the blue path and ends up below the blue path, whenever it "crosses above", it must eventually "crosses back down". So the number of crossings between the blue path and the yellow path is even. (If you like, this is like a discrete version of intermediate value theorem in calculus.)

But the same is true for any pair of paths here. So the total number of crossings must be even. So the identity can only be the composition of even numbers of swaps. □

Theorem 6.2.5. *A permutation must be either even or odd, but not both.*

Proof. Suppose $P = P_1 \dots P_s = Q_1 \dots Q_t$ is the decomposition of the permutation P into swaps $P_1, \dots, P_s$ or swaps $Q_1, \dots, Q_t$. Then $I = PP^{-1} = P_1 \dots P_s Q_t \dots Q_1$. (Here note that the inverse of a swap is itself.) So the identity permutation is the composition of $s + t$ numbers of swaps. But then $s + t$ must be even, so s, t must be both even or both odd. □

As a result, orientation is well-defined. For the purpose of later use, let us do one last definition.

Definition 6.2.6. *For a permutation P, we define its sign* $\operatorname{sign}(P)$ *to be* 1 *if P is even, and* -1 *if P is odd. So if P is the composition of k swaps, then* $\operatorname{sign}(P) = (-1)^k$.

Note that many traditional Chinese textbook often refers to a concept called "inversion number", which refers to the number of inersections in our diagrams.

You should NOT focus on the number of intersections (i.e., inversion numbers). The numbers themselves are utterly useless. Only the parity (i.e., even or odd) matters.

6.2.2 Cycle Decomposition (Optional)

The methods in the last subsection is in fact NOT how people study permutations most of the time. A more useful concept is cycle decomposition. This is hard to establish for beginners, but in practis it is both easy to compute and easy to use.

First let us see an example.

Example 6.2.7. Suppose our permutation sends $1, 2, 3, 4, 5$ to $2, 3, 1, 5, 4$ respectively. Then this means we have a loop $1 \to 2 \to 3 \to 1$ and a loop $4 \leftrightarrow 5$, or "cycles". Cycle decomposition means we think of our permutation as disjoint loops (cycles) like these. It is a fact that the cycle decomposition of a permutation must exist and be unique. ☺

Theorem 6.2.8 (Cycle Decomposition)**.** *Given a permutation σ on the indices $1, 2, ..., n$, then we can break down the set $\{1, ..., n\}$ into a union of disjoint subsets, such that P acts as a single "loop" on each subset.*

Proof. We proceed by mathematical induction. If $n = 1$, this is trivial.

For generic $n > 1$, consider the subset $S = \{1, \sigma(1), \sigma(\sigma(1)), ...\}$. Let k be the smallest positive integer such that $\sigma^k(1) = 1$. (Here σ^k means we iterate σ by k-times). Let us first see that $1, \sigma(1), ..., \sigma^{k-1}(1)$ are all distinct. If for some $a < b < k$ we have $\sigma^a(1) = \sigma^b(1)$, then since permutations are bijections, we have $\sigma^{b-a}(1) = 1$ and $b - a < k$, contradiction. So indeed $S = \{1, \sigma(1), \dots, \sigma^{k-1}(1)\}$ and it has k elements. Clearly σ acts as a "loop" on S. Furthermore, if $\sigma(i) \in S$, then $i \in \sigma^{-1}(S) = S$.

But then for the rest of the indices, $\{1, ..., n\} - S$ has only $n - k$ elements. And if $i \notin S$, we already know that $\sigma(i) \notin S$. So σ also permutes indices in $\{1, ..., n\} - S$. So by induction hypothesis, σ acts on $\{1, ..., n\} - S$ as a union of disjoint "loops". So we are done. □

Finding a cycle decomposition is extremely fast. You simply follow the idea of the proof above, and figure it out one cycle at a time. For example, if σ sends $1, 2, 3, 4, 5$ to $4, 5, 1, 3, 2$, then we simply compute $1 \mapsto \sigma(1) = 4 \mapsto \sigma(4) = 3 \mapsto \sigma(3) = 1$ and we closed up the first loop. Then we move on to the first un-used index, and compute $2 \mapsto \sigma(2) = 5 \mapsto \sigma(5) = 2$ and we closed up the second loop. Done. So we can perform a cycle decomposition almost as fast as simply reading the entire permutation.

Lemma 6.2.9. *If a permutation is a single big cycle on n indices, then it is the product of $n - 1$ swaps.*

Proof. We proceed by mathematical induction. If $n = 1$, this is trivial.

Say the permutation σ is $1 \mapsto 2 \mapsto ... \mapsto n \mapsto 1$. Let $\sigma' = 1 \mapsto 2 \mapsto ... \mapsto n - 1 \mapsto 1$ and $\sigma'(n) = n$, and let $\tau = n - 1 \mapsto n \mapsto n - 1$ be a swap that fixes $1, ..., n - 2$. Then we can verify that σ is the composition $\sigma' \circ \tau$. So σ needs one more swap than σ', which is a product of $n - 2$ swaps by induction hypothesis. So σ is the product of $n - 1$ swaps. □

Corollary 6.2.10. *If a permutation has a cycle decomposition into $c_1, ..., c_k$ cycles, then its parity is $c_1 + \cdots + c_k - k$.*

Proof. Each cycle needs $c_i - 1$ swaps. □

This is the best way to figure out the parity of a permutation. For super super large permutations, the diagrams in the last subsection will be an unreadable mess. Cycle decomposition is much more superior.

6.3 Uniqueness and Existence of Determinants

Definition 6.3.1. *We define a function* $\det : M_{n\times n} \to \mathbb{R}$ *to be a* ***determinant function*** *if it is multilinear (linear in each column), alternating (swapping two columns will negate value), and normalized (*$\det(I) = 1$*).*

Note that we are going to write $\det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ *sometimes, and this simply means* $\det(A)$ *where* $A = \begin{bmatrix}\boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n\end{bmatrix}$

You can interpret the determinant function as n-dimensional oriented volume, or as volume scaling factor, it does not matter. Our goal is to show that this exist and is unique. So we can simply say "the determinant" instead of "a determinant function".

But before we show that it exists, let us see some examples of what is NOT a determinant.

Example 6.3.2. A multilinear function is NOT linear (unless there is only one input). For example, in $\mathbb{R}^2$, $\det(3A) \neq 3\det(A)$. Rather, it is $9\det(A)$. We CANNOT just pull out the common factor 3 from the MATRIX. Rather, we have to pull out the common factor 3 from EACH COLUMN, and thus we pulled out two 3's instead.

So linear functions like trace are NOT determinants.

Nevertheless, there are some behavioral analogies. For example, a linear combination of multilinear functions are still multilinear. For example, If f, g are multilinear, then $(f+g)(\boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w}) + g(\boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{u}, \boldsymbol{w}) + f(\boldsymbol{v}, \boldsymbol{w}) + g(\boldsymbol{u}, \boldsymbol{w}) + g(\boldsymbol{v}, \boldsymbol{w}) = (f+g)(\boldsymbol{u}, \boldsymbol{w}) + (f+g)(\boldsymbol{v}, \boldsymbol{w})$. ☺

Example 6.3.3. Consider the function $f : M_{n\times n} \to \mathbb{R}$ with f such that $f(A) = a_{11}a_{22}\ldots a_{nn}$, i.e., it sends any matrix to the product of diagonal entries. This is multilinear. Let us just prove that it is linear in the first column. We have

$$\begin{aligned}
&f(\begin{bmatrix} b_{11}+c_{11} & a_{12} & \ldots & a_{1n} \\ b_{21}+c_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}+c_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}) \\
=&(b_{11}+c_{11})a_{22}\ldots a_{nn} \\
=&b_{11}a_{22}\ldots a_{nn} + c_{11}a_{22}\ldots a_{nn} \\
=&f(\begin{bmatrix} b_{11} & a_{12} & \ldots & a_{1n} \\ b_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}) + f(\begin{bmatrix} c_{11} & a_{12} & \ldots & a_{1n} \\ c_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}).
\end{aligned}$$

You can esily verify the rest. However, this is NOT alternating. Indeed, $f(I) = 1$, but for any non-identity permutation matrix P, since it will fail to preserve all index, some diagonal entry will be zero. So $f(P) = 0$, whereas we should have $\det(P) = 1$ for even permutations and -1 for odd permutations.

Note that this is a start though. Note that if we set $f_P(A) = f(AP)$ for a fixed permutation P, then f_P is also multilinear because P merely permute the columns.

Suppose we set $g(A) = \sum_P f_P(A)$. For example, over $\mathbb{R}^3$ we have

$$g(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) + f(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}) + f(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u}) + f(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}) + f(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}) + f(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{u}).$$

This is a linear combination of multilinear functions f_P for all P, and hence multilinear. However, it is NOT alternating. Rather, it is symmetric. As you can see, permuting columns in g would NOT change the value. What if we want something alternating? We would have to define g differently. Suppose we define

$$g(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) - f(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}) + f(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u}) - f(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}) + f(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}) - f(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{u}).$$

Then this shall be alternating as desired. So the desired function is $g(A) = \sum_P \pm f_P(A)$, where we add this for even permutations, and substract this for odd permutations. ☺

Theorem 6.3.4. *The determinant function exists.*

Proof. Let us start with any non-zero multilinear function $f : M_{n\times n} \to \mathbb{R}$. Say pick f such that $f(A) = a_{11}a_{22}\dots a_{nn}$, i.e., it sends any matrix to the product of diagonal entries.

Now, obviously this f is NOT a determinant function, because it is not alternating. What should we do then? WE FORCE IT!

We define a new function f_P as $f_P(A) = f(AP)$ for each permutation P. Let $\text{sign}(P)$ be 1 if P is an even permutation, and -1 if P is an odd permutation. Set $g(A) = \sum_P \text{sign}(P) f_P(A)$.

What have I done? Well, since each f_P is still multilinear, their linear combination is also multilinear. Furthermore, if S is a swap, then $g(AS) = \sum_P \text{sign}(P) f_P(AS) = \sum_P \text{sign}(P) f_{PS}(A) = -\sum_P \text{sign}(PS) f_{PS}(A) = -\sum_{P'} \text{sign}(P') f_{P'}(A) = -g(A)$. Here we set $P' = PS$ and used a change of index for the sum. So g is multilinear and alternating.

Now what about normalized? Well, check out $g(I)$. Note that $f(I) = 1$ and $f_P(I) = 0$ as long as P is not the identity permutation. So $g(I) = f(I) = 1$. So we are done. □

Remark 6.3.5. *A philosophy remark here. The initial multilinear function f does not actually matter much. Pick any f, and define $g = \sum \text{sign}(P) f_P$. Then g must be multilinear and alternating. If $g(I) = 1$, then it is the determinant and we are done. If $g(I) \neq 1, 0$, then define $h = \frac{1}{g(I)} g$, and it will be the determinant. The only potential trouble is if $g(I) = 0$. If that happens, then you need to pick a different f.*

Corollary 6.3.6 (Leibniz formula of determinants, i.e., the "big formula")**.** *A formula for a determinant function is* $\det(A) = \sum_\sigma \text{sign}(\sigma) a_{\sigma(1),1} \dots a_{\sigma(n),n}$. *Here σ ranges over all permutations on the index set $\{1, 2, ..., n\}$, and $a_{i,j}$ is the (i, j) entry of A.*

Proof. Just pick $f(A) = \prod a_{ii}$ and construct g as before. □

So we have established existence. Note how unwieldly this formula is. For $n \times n$ matrices, we have $n!$ terms. Oof! Don't memorize this, and don't ever use this formula, unless you have to. Also, don't let your computer compute this formula, if you don't want your computer to burn itself out. This takes forever to compute.

The only point of this formula is to show that determinants DO exist. Now a more practical approach of determinant lies in the fact that it is unique, which we now prove.

Theorem 6.3.7. *The determinant function is unique.*

Proof. Suppose f, g are two determinant functions. Suppose $f(A) = g(A)$ for some A. Then since AE is simply doing corresponding column operations on A, and since f, g are both alternating and multilinear, therefore we have $f(AE) = g(AE)$.

But we have $f(I) = 1 = g(I)$. By applying a series of column operations, I can get to any invertible matrix from I. So we have $f(A) = g(A)$ for all invertible A.

If A is not invertible, then AE will have first column zero for some E. (One column is the linear combination of others. Shear this column into zero using other columns, then swap it to the left.)

By multilinearity, we have $f(AE) = 0 = g(AE)$. So $f(A) = g(A)$ as well. □

Here is an alternative way to get the big formula. It is less conceptual and more computational. But the idea is simple: the big formula is simply this: write all inputs as linear combinations of standard basis vectors, and use multilinearity to expand everything!

Example 6.3.8. Consider $\det(A)$ where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. We have

$$\begin{aligned}
\det(A) =& \det(\begin{bmatrix} a \\ c \end{bmatrix}, \begin{bmatrix} b \\ d \end{bmatrix}) \\
=& \det(a\boldsymbol{e}_1 + c\boldsymbol{e}_2, b\boldsymbol{e}_1 + d\boldsymbol{e}_2) \\
=& ab\det(\boldsymbol{e}_1, \boldsymbol{e}_1) + ad\det(\boldsymbol{e}_1, \boldsymbol{e}_2) + bc\det(\boldsymbol{e}_2, \boldsymbol{e}_1) + cd\det(\boldsymbol{e}_2, \boldsymbol{e}_2)
\end{aligned}$$

$$= ad - bc.$$

Again, as you can see, we get the big formula by simply expanding everything using multilinearity. ☺

Let us now do an alternative proof of the Leibniz formula.

Alternative proof of Leibniz formula. Let $\boldsymbol{a}_k$ be the k-th column of A, and let $a_{i,j}$ be the (i,j) entry of A. Then $\boldsymbol{a}_k = \sum_{i_k=1}^{n} a_{i_k,k}\boldsymbol{e}_{i_k}$ for each k.

Now

$$\begin{aligned}\det(A) &= \det(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)\\ &= \det(\sum_{i_1=1}^{n} a_{i_1,1}\boldsymbol{e}_{i_1}, \dots, \sum_{i_n=1}^{n} a_{i_n,n}\boldsymbol{e}_{i_n})\\ &= \sum_{(i_1,\dots,i_n)} a_{i_1,1}\dots a_{i_n,n}\det(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_n}).\end{aligned}$$

Now, if a tuple $(i_1, \dots, i_n)$ have repeated indices, then $\det(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_n}) = 0$. If there is no repeated indices, then $(i_1, \dots, i_n)$ is a permutation of $(1, \dots, n)$, and thus it corresponds to some permutaiton P, and $\det(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_n}) = \det(P) = \operatorname{sign}(P)$.

Hence we have

$$\begin{aligned}\det(A) &= \sum_{(i_1,\dots,i_n)} a_{i_1,1}\dots a_{i_n,n}\det(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_n})\\ &= \sum_{P} \operatorname{sign}(P) a_{P(1),1}\dots a_{P(n),n}.\end{aligned}$$

Here P ranges over all possible permutations, and $P(k)$ is the number that P would permute k to. □

So here is an important conclusion: whatever the big formula can do, you can also do by simply using multilinearity. The big formula is nothing more than the ultimate expression after using multilinearity on everything. Multilinearity is enough. We shall almost NEVER use the big formula!

Example 6.3.9. We have the 3×3 determinant formula

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

Can you see which term corresponds to which permutation? ☺

Example 6.3.10. We have the 2×2 determinant formula $\det\begin{bmatrix} a & b\\ c & d\end{bmatrix} = ad - bc$. So the area of the parallelogram is the difference between the area of two squares. Can you prove this statement using high school planar geometry? (Hint: geometrically speaking, multilinearity means, for example, we should do $\begin{bmatrix} a\\ c\end{bmatrix} = \begin{bmatrix} a\\ 0\end{bmatrix} + \begin{bmatrix} 0\\ c\end{bmatrix}$. As a result, the parallelogram would break down into (the sum or difference of) two parallelograms. Now shear these parallelograms ultil they are rectangles.) ☺

6.4 Base, Height, Cofactor Expansion

Example 6.4.1. Given a 3×3 matrix $A = \begin{bmatrix} a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}$, we have

$$\det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}.$$

Boy this is ugly. However, let us try to establish some pattern here. For example, let us take all the terms involving a_{11}. Then we see $a_{11}(a_{22}a_{33} - a_{23}a_{32})$. This is exactly a_{11} times the determinant of the lower right block! We shall in this section generalize this to all matrices and all entries.

Basically, for a_{ij}, the corresponding coefficient for a_{ij} is $(-1)^{i+j}\det(A_{ij})$, where A_{ij} is the submatrix of A by removing the i-th row and j-th column. ☺

Our goal is to achieve the following: we want to arrange the big formula for $\det(A)$ as

$$\det(A) = a_{ij}(\text{Blahblah}) + (\text{Other stuff that does not involve } a_{ij}).$$

In short, we want to figure out what terms in the big formula would involve the entry a_{ij}. For this purpose, let us also use a concept from calculus.

Definition 6.4.2. *Given a function of many variables, say $f(x, y, z)$, the partial derivative $\frac{\partial}{\partial x}f$ means holding the other variables constant, and take derivative with respect to x. I.e., $\frac{\partial}{\partial x}f(x, y, z) = \lim_{\mathrm{d}x\to 0}\frac{f(x+\mathrm{d}x,y,z)-f(x,y,z)}{\mathrm{d}x}$.*

Example 6.4.3. Say $f(x, y) = x^2y$. Then $\frac{\partial}{\partial x}f = 2xy$ and $\frac{\partial}{\partial y}f = x^2$.

For determinant, we have n^2 variables $a_{11}, a_{12}, \ldots, a_{nn}$. Above we showed that for 3×3 matrix A, $\frac{\partial}{\partial a_{11}}\det(A) = a_{22}a_{33} - a_{23}a_{32}$. Basically, suppose we have

$$\det(A) = a_{ij}(\text{Blahblah}) + (\text{Other stuff that does not involve } a_{ij}).$$

Then $\frac{\partial}{\partial a_{ij}}\det(A)$ is exactly the "Blahblah" portion.

In terms of calculus, we shall later show that $\frac{\partial}{\partial a_{ij}}\det(A) = (-1)^{i+j}\det(A_{ij})$. Calculus aside, the intuition is the answer to the following question: How would a particular entry a_{ij} influence $\det(A)$? The answer is via all the entriess NOT in the same column and NOT in the same row.

This makes sense, considering how the big furmula works. Each term of the big formula uses exactly one entry in each row and each column. So all the entries in the same row would NEVER be multiplied to produce a term, and the same for all entries in the same column. Their contributions to the determinant are "disjoint". ☺

Definition 6.4.4. *For a matrix A, we use A_{ij} to denote the submatrix of A by removing the i-th row and j-th column. Then we call $c_{ij} = (-1)^{i+j}\det(A_{ij})$ the (i, j) **cofactor** of A.*

*In particular, the matrix C whose (i, j) entry is c_{ij} is the **cofactor matrix** of A.*

Remark 6.4.5. *The name "cofactor" is like this: in the big formula $\det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$, what terms have factor a_{11}? Then these terms make up $a_{11}(a_{22}a_{33} - a_{23}a_{32})$ or $a_{11}c_{11}$. Hence c_{11} is the "cofactor" to the factor a_{11}.*

Lemma 6.4.6. $\frac{\partial}{\partial a_{11}}(\det(A)) = \det(A_{11}) = c_{11}$.

Proof. Suppose B is the matrix obtained by increasing the $(1, 1)$ entry of A by a tiny bit $\mathrm{d}x$. We want to see how $\det(A)$ changes into $\det(B)$.

Note that A, B differ only in the first column. If A has first column $\boldsymbol{a}$, then B has first column $\boldsymbol{a} + \mathrm{d}x\boldsymbol{e}_1$. So

$$\det(B) = \det(\boldsymbol{a} + \mathrm{d}x\boldsymbol{e}_1, rest) = \det(\boldsymbol{a}, rest) + \det(\mathrm{d}x\boldsymbol{e}_1, rest) = \det(A) + \det(\mathrm{d}x\boldsymbol{e}_1, rest).$$

In particular, we see that $\det(B) - \det(A) = \det(\mathrm{d}x\boldsymbol{e}_1, rest) = \det\begin{bmatrix}\mathrm{d}x & \star\\ \boldsymbol{0} & A_{11}\end{bmatrix}$. This is block upper triangular, so this is $\mathrm{d}x\det(A_{11}) = c_{11}\,\mathrm{d}x$.

So $\frac{\partial}{\partial a_{11}}(\det(A)) = \lim_{\mathrm{d}x\to 0}\frac{\det(B)-\det(A)}{\mathrm{d}x} = c_{11}$. □

Proposition 6.4.7 (How entries effect determinant)**.** $\frac{\partial}{\partial a_{ij}}(\det(A)) = (-1)^{i+j}\det(A_{ij}) = c_{ij}$.

Proof. Let P_i be the permutation matrix that (as row operation) send indices 1 to 2, 2 to 3, ..., $i-1$ to i, and then i to 1. All other indices are fixed. Let P_j be defined similarly but as a column operation. Let $B = P_iAP_j$.

Then by interpreting these permutation matrices as row permutations and column permutations, one can directly verify that the $(1,1)$ entry of B, b_{11}, is exactly a_{ij}, and the submatrix B_{11} is exactly A_{ij}.

Now $\det(A_{ij}) = \det(B_{11}) = \frac{\partial}{\partial b_{11}}(\det(B)) = \frac{\partial}{\partial a_{ij}}(\det(P_iAP_j)) = \det(P_i)\frac{\partial}{\partial a_{ij}}(\det(A))\det(P_j)$. Here the last equality is because $\det(P_i), \det(P_j)$ are constants, so we can simply take them out of the derivative.

Now P_i has $i-1$ crosses in its diagram, and P_j has $j-1$ crosses in its diagram, so we see that $\det(P_i)\det(P_j) = (-1)^{i+j-2} = (-1)^{i+j}$. Here we use $i+j$ instead of $i+j-2$ because it looks prettier. Hence, $\det(A_{ij}) = (-1)^{i+j}\frac{\partial}{\partial a_{ij}}(\det(A))$. □

You should loosely interpret above proposition as this: the formula for determinant is $\det(A) = a_{ij}c_{ij} +$ other stuff, where "other stuff" do not use the entry a_{ij} at all. Putting these together, we would have a new formla for determinant.

Example 6.4.8.

$$\begin{aligned}\det(A) =& a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}\\ =& a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{21}(-a_{12}a_{33} + a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22})\\ =& a_{11}c_{11} + a_{21}c_{21} + a_{31}c_{31}.\end{aligned}$$

Basically, each term in the big formula for $\det(A)$ must uses one entry in the first column. And if a term uses a_{i1}, then it is a term contained inside $a_{i1}c_{i1}$. ☺

Theorem 6.4.9 (Laplace expansion)**.** *Fix an index k. Then* $\det(A) = a_{1k}c_{1k} + \cdots + a_{nk}c_{nk}$ *(expansion via the k-th column), and similarly* $\det(A) = a_{k1}c_{k1} + \cdots + a_{kn}c_{kn}$ *(expansion via the k-th row).*

Proof. Look at the Lebniz formula. Each term must be has exactly ONE of $a_{1k}, \ldots, a_{nk}$ as a factor. (Since each term only contains a single entry from the k-th column). In particular, by taking common factors, we can write $\det(A) = a_{1k}(\text{stuff}) + a_{2k}(\text{stuff}) + \cdots + a_{nk}(\text{stuff})$, and all the "stuff" here only uses entries NOT on the k-th column (since the one on the k-th column is already picked).

Now by taking partial derivatives, we see that $\frac{\partial}{\partial a_{ik}}(\det(A))$ is exactly the "stuff" after a_{ik}, and we have $\frac{\partial}{\partial a_{ik}}(\det(A)) = c_{ik}$. So we are done.

The expansion via rows is the same as the expansion via columns. □

Example 6.4.10. In general, the Laplace expansion would reduce the calculation of a single $n \times n$ determinant to the calculation of n determinants of $(n-1)\times(n-1)$ determinants. For examlpe, we have

$$\det\begin{bmatrix}1&2&3\\4&5&6\\7&8&9\end{bmatrix} = 1\times\det\begin{bmatrix}5&6\\8&9\end{bmatrix} - 4\times\det\begin{bmatrix}2&3\\8&9\end{bmatrix} + 7\times\det\begin{bmatrix}2&3\\5&6\end{bmatrix}.$$

Note the sign here are alternating. Be careful not to mess it up.

Now this in general did not save too many time. It is merely doing the big formula in a more organized way. However, sometimes it will save you some time.

For examlpe, consider $\det(A) = \det\begin{bmatrix}0&1&2&3\\0&4&0&0\\5&6&7&8\\0&9&10&11\end{bmatrix}$. By doing Laplace expansion in the first column, we have $\det\begin{bmatrix}0&1&2&3\\0&4&0&0\\5&6&7&8\\0&9&10&11\end{bmatrix} = 5\det\begin{bmatrix}1&2&3\\4&0&0\\9&10&11\end{bmatrix}$. This is nice because the first column has many zeros. Keep picking rows and columns with many zeros, and we can expand along the second row and get $-20\det\begin{bmatrix}2&3\\10&11\end{bmatrix} = 160$.

But let us do this again by using the big formula directly. Then you shall immediately see that this again. Recall that, in the big formula, each term only uses, and must use exactly ONE entry from each column and each row. Now loot at $\det \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 4 & 0 & 0 \\ 5 & 6 & 7 & 8 \\ 0 & 9 & 10 & 11 \end{bmatrix}$. If a term in the big formula wants to be non-zero, then it MUST use the 5. There is no other non-zero alternatives. Similarly, the term MUST also use 4 in the second row. There is no other alternative.

Now since a nonzero term must use the 5 in the first column and the 4 in the second row, the rest is like $\det \begin{bmatrix} - & - & 2 & 3 \\ - & 4 & - & - \\ 5 & - & - & - \\ - & - & 10 & 11 \end{bmatrix}$. The dashed out portion does not matter, because NO NONZERO TERM will use them ever. They might as well all be zero. Now, since a non-zero term must already picked 5 and 4 here, there are only two possibilities left: pick 2 and 11, or pick 3 and 10.

So the two non-zero terms are $5\times4\times2\times11$ and $5\times4\times3\times10$. Be careful of the sign, which comes from the position of these entries. So we have $\det(A) = 5\times4\times2\times11 \det \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + 5\times4\times3\times10 \det \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$. The first permutation matrix can be done in 1 swap, while the second can be done in 2 swaps. So we have $\det(A) = -5 \times 4 \times 2 \times 11 + 5 \times 4 \times 3 \times 10 = 160$.

As you can see, there is a clear equivalence between Laplace expansion and the big formula. Laplace expansion is simply grouping terms in the big formula by common factors. Intuitively, Laplace expansion is nothing more than doing the big formula in a more organized fashion.

Here is something else to think about, if you like. At least how many zeros do you need for a 4×4 matrix, to guarantee that all terms in the big formula are zero? (Answer is 4, and they must all lie on the same column or same row.)

Let us do this YET AGAIN. When we say "the non-zero terms must pick 5 in the first column", and thereby simplifying the determinant $\det \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 4 & 0 & 0 \\ 5 & 6 & 7 & 8 \\ 0 & 9 & 10 & 11 \end{bmatrix}$ into $\det \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 4 & 0 & 0 \\ 5 & 0 & 0 & 0 \\ 0 & 9 & 10 & 11 \end{bmatrix}$. Hey, this is just a column operation! We can then do a row operation using the second row to shear others, and get $\det \begin{bmatrix} 0 & 0 & 2 & 3 \\ 0 & 4 & 0 & 0 \\ 5 & 0 & 0 & 0 \\ 0 & 0 & 10 & 11 \end{bmatrix}$. And we get the results again.

See? Ultimately, the big formula is just multilinearity. So any fancy argument using the big formula must ultimately be the same as some column/row operations.

Let us do this YET YET AGAIN. Consider some row/column swaps as

$$\det \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 4 & 0 & 0 \\ 5 & 6 & 7 & 8 \\ 0 & 9 & 10 & 11 \end{bmatrix} = (-1)^3 \det \begin{bmatrix} 5 & 6 & 7 & 8 \\ 0 & 1 & 2 & 3 \\ 0 & 9 & 10 & 11 \\ 0 & 4 & 0 & 0 \end{bmatrix} = (-1)^4 \det \begin{bmatrix} 5 & 8 & 7 & 6 \\ 0 & 3 & 2 & 1 \\ 0 & 11 & 10 & 9 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

Now note that this is block upper triangular, with blocks $\begin{bmatrix} 5 \end{bmatrix}, \begin{bmatrix} 3 & 2 \\ 11 & 10 \end{bmatrix}, \begin{bmatrix} 4 \end{bmatrix}$. So the answer is $(-1)^4 \times 5 \times 4 \times \det \begin{bmatrix} 3 & 2 \\ 11 & 10 \end{bmatrix} = 160$. As you can see, why were we able to utilize the zero entries so effectively? It is precisely because these zeros allow us to make things block upper triangular, which reduced the calculation of a big determinant to be simplified into a smaller determinant.

If the zeros are smartly arranged, so that you cannot get block triangular things no matter how you permute, then none of the above methods would have worked. ☺

Recall that the Leibniz formula (big formula) is basically just writing all columns in the standard basis, and expand everything using multilinearity. And afterwards, by grouping terms in the big formula via common factors along a column or a row, we have the Laplace expansion.

So multilinearity = Leibniz formula = Laplace expansion. They are simply describing the same structure in different perspectives. The first one is the perspective of linear algebra. The Leibniz formula is from the perspective of permutation theory. And the Laplace expansion is in fact a geometric perspective.

Proposition 6.4.11 (Laplace expansion is base times height)**.** *Consider a square matrix* $A = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \end{bmatrix}$ *and its cofactor matrix* $C = \begin{bmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_n \end{bmatrix}$*. Then* $\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{c}_j$ *is* 0 *if* $i \neq j$ *and* $\det(A)$ *if* $i = j$*. The same is true for rows.*

In particular, $\boldsymbol{c}_i$ *is perpendicular to* $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_{i+1}, \dots, \boldsymbol{a}_n$*. And the length* $\|\boldsymbol{c}_i\|$ *is exactly the* $(n-1)$*-dimensional absolute volume of the* $(n-1)$*-dimensional parallelotope* $(\boldsymbol{a}_1, \dots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_{i+1}, \dots, \boldsymbol{a}_n)$ *in* $\mathbb{R}^n$*.*

Proof. We already know that $\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{c}_i = \det(A)$. Now, suppose we have $i \neq j$, and we aim to show that $\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{c}_j = 0$.

Let us prove the case when $j = 1$ and $i \neq 1$. Imagine the Laplace expansion for the following two matrices.

$$A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}.$$

$$B = \begin{bmatrix} \boldsymbol{a}_i & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}.$$

To both matrices, if we perform the Laplace expansion along the first column, the cofactors are actually identical! Therefore we have $\det(A) = \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{c}_1$ and $\det(B) = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{c}_1$.

On the other hand, clearly B has repeated column. So $\det(B) = 0$. Hence we see that $\boldsymbol{c}_1 \perp \boldsymbol{a}_i$.

So we are done. We see that $\boldsymbol{c}_j$ is perpendicular to $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{j-1}, \boldsymbol{a}_{j+1}, \dots, \boldsymbol{a}_n$, i.e., perpendicular to the "base". Then Volumn $= |\det(A)| = |\boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{c}_j| = \|\text{Projection of } \boldsymbol{a}_j \text{ to } \boldsymbol{c}_j\| \|\boldsymbol{c}_j\| = \text{height} \times \|\boldsymbol{c}_j\|$. So we see that $\|\boldsymbol{c}_j\| =$ "Base Area". □

Corollary 6.4.12 (Big Formula for Inverse Matrix)**.** $C^{\mathrm{T}} A = A^{\mathrm{T}} C = \det(A) I$*. In particular, if* A *is invertible, then its inverse is* $A^{-1} = \frac{1}{\det(A)} C^{\mathrm{T}}$*. Note that this gives a formula for the inverse.*

Proof. The (i, j) entry of $C^T A$ is $\boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{a}_j$. So we are done. □

Example 6.4.13. Suppose $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then the cofactor matrix is $C = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$. And the inverse is $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Intuitively, the cofactors are the "entries of inverse transpose, but without denominator". ☺

The formula here may sounds nice, but it is actually too slow to compute. Each entry of C is a determinant, and you would have to calculate forever. Our old approach, i.e., Gaussian elimination on $\begin{bmatrix} A & I \end{bmatrix}$, would find A^{-1} much faster.

Corollary 6.4.14 (Cramer's rule)**.** *Suppose* A *is invertible. Then the unique solution to* $A\boldsymbol{x} = \boldsymbol{b}$ *is a vector whose* i*-th coordinate is* $x_i = \frac{\det(A_i)}{\det(A)}$*, where* A_i *is a matrix obtained by replacing the* i*-th row of* A *by* $\boldsymbol{b}$*. Note that this gives a formula for the solution of the linear system. (Too ugly to be useful though. Gaussian elimination is much faster.)*

Proof. Let us simply prove the formula for x_1. The rest are the same.

Consider $A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}$ and $A_1 = \begin{bmatrix} \boldsymbol{b} & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}$. Then $\det(A_1) = \det \begin{bmatrix} x_1 \boldsymbol{a}_1 + \cdots + x_n \boldsymbol{a}_n & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix} = \det \begin{bmatrix} x_1 \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}$ by shearing columns. Then we have $\det(A_1) = x_1 \det \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix} = x_1 \det(A)$. So we are done.

There is also a more geometric proof (with essentially the same idea). Note that as parallelotopes, A, A_1 share the same "base" made of $\boldsymbol{a}_2, \dots, \boldsymbol{a}_n$. So the ratio of determinant is simlpy the ratio of height. So the key is to compare how $\boldsymbol{a}_1$ contribute to the height, and how $\boldsymbol{b}$ contribute to the height.

Now height is the portion perpendicular to the base, i.e., perpendicular to $\boldsymbol{a}_2, \dots, \boldsymbol{a}_n$. If $\boldsymbol{b} = x_1\boldsymbol{a}_1 + \cdots + x_n\boldsymbol{a}_n$, then $x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ does NOT contribute to the height at all, because this is parallel to the base. Only $x_1\boldsymbol{a}_1$ component of $\boldsymbol{b}$ would contribute. So the height of A_1 is x_1 times the height of A. So we are done. □

Remark 6.4.15. *The big formula for determinant, inverse matrix, and solutions for linear system are ALL useless for computational purpose. There are too many terms, and too many additions and multiplications to do. It is usually MUCH faster to use Gaussian elimination for all of them.*

However, the importance lies in perspectives. How would an entry contribute to the determinant? How would zero entries help simplify the determinant? How does parallelotopes has to do with solving linear systems? THESE are the main takeaways.

6.5 (Optional) Generalized Pythagorean theorem and Cauchy-Binet formula

Note that in the discussions above, we have essentially solved a geometric problem.

Corollary 6.5.1 ($(n-1)$-dimensional parallelotope in $\mathbb{R}^n$)**.** *Suppose we have an $(n-1)$-dimensional parallelotope in $\mathbb{R}^n$, whose edges form an $n \times (n-1)$ matrix A. Let A_i be the square matrix obtained by removing the i-th row of A. Then the $(n-1)$-dimensional (absolute) volumn is $\sqrt{\sum \det(A_i)^2}$.*

Proof. Consider adding an arbitrary column $\boldsymbol{x}$ to the left of A to get an $n \times n$ square matrix B. Then the $(n-1)$-parallelotope is the "base" of B using all but the first columns. Let $\boldsymbol{c}_1$ be the first column of the cofactor matrix of B, then $\|\boldsymbol{c}_1\|$ is our desired results.

Now $\boldsymbol{c}_1$ has coordinates $\pm\det(A_i)$. So we are done. □

Note that A_i is obtained by deleting the i-th coordinate of all edges of the parallelotope. So they are projection image of A. For example, if $n = 3$ and A represent some parallelogram, then A_1 is the projection to yz-plane.

Corollary 6.5.2 (Generalized Pythagorean Theorem)**.** *Given a right tetrahedron (a "corner of the walls", i.e., three adjacent edges of it are mutually orthogonal to each other), the squared area of the oblique face is the sum of the squared area of the right-triangle faces.*

Any flat shape in $\mathbb{R}^3$, say it has area S, and its "shadows", i.e., projections to the three coordinate planes, has are S_{xy}, S_{yz}, S_{zx}, then $S^2 = S_{xy}^2 + S_{yz}^2 + S_{zx}^2$. (However, this is only true for flat shapes, i.e., entirely contained in an affine 2-dimensional space. Surfaces with a curvature will fail to have this.)

(You can easily see generalizations of this in higher dimensions. But the proof relies on Cauchy-Binet formula, which is poven below.)

However, in comparison, here is an interesting result.

Proposition 6.5.3 (k-dimensional parallelotope in $\mathbb{R}^n$)**.** *Consider a k-dimensional parallelotope in $\mathbb{R}^n$ (inner product is dot product), with edges forming a matrix $A = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_k\end{bmatrix}$. Then the k-dimensional (absolute) volumn is $\sqrt{\det(A^{\mathrm{T}}A)}$.*

Proof. Let $W = \mathrm{Ran}(A)$ and pick an orthonormal basis for $W^\perp$, $Q = \begin{bmatrix}\boldsymbol{q}_{k+1} & \dots & \boldsymbol{q}_n\end{bmatrix}$. Now all these vectors are unit vectors, mutually orthogonal, and all orthogonal to the parallelotope represented by A. As a result, $\begin{bmatrix}A & Q\end{bmatrix}$ is an n-dimensional parallelotope whose n-dimensional absolute volumn is the same as the k-dimensional absolute volume of A.

So this is $|\det\begin{bmatrix}A & Q\end{bmatrix}| = \sqrt{\det\begin{bmatrix}A & Q\end{bmatrix}^{\mathrm{T}}\det\begin{bmatrix}A & Q\end{bmatrix}} = \sqrt{\det(\begin{bmatrix}A & Q\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}A & Q\end{bmatrix})} = \sqrt{\det\begin{bmatrix}A^{\mathrm{T}}A & A^{\mathrm{T}}Q\\ Q^{\mathrm{T}}A & Q^{\mathrm{T}}Q\end{bmatrix}}$.

Now $Q^{\mathrm{T}}Q = I$ because Q has orthonormal columns. Furthermore, $A^{\mathrm{T}}Q = (Q^{\mathrm{T}}A)^{\mathrm{T}} = O$, because all columns of Q are orthogonal to the parallelotope A. So our expression simplifies to $\sqrt{\det \begin{bmatrix} A^{\mathrm{T}}A & O \\ O & I \end{bmatrix}} = \sqrt{\det(A^{\mathrm{T}}A)}$. □

Putting the two together, this hinted at a much stronger formula.

Theorem 6.5.4 (Cauchy-Binet Formula)**.** *Say A is an $n \times k$ matrix with $n \geq k$. Then $\det(A^{\mathrm{T}}A) = \sum \det(A')^2$, where A' is any $k \times k$ square submatrix of A, and we are summing all such possible suqare submatrices.*

Taking polarization identity, we in fact has $\det(A^{\mathrm{T}}B) = \sum \det(A')\det(B')$ where the sum is over all CORRESPONDING $k \times k$ square submatrices A', B' of A, B.

Proof. It turned out that it is easier to prove $\det(A^{\mathrm{T}}B) = \sum \det(A')\det(B')$. Think of both sides as taking the input $\begin{bmatrix} A & B \end{bmatrix}$. Then both sides are multilinear in columns. So, by expanding everything in terms of standard basis vectors, we only need to prove the case when all columns are some standard basis vector.

In this case, if A has repeated columns, it is easy to verify that both sides are zero. So suppose A has distinct columns. (Since columns of A are standard basis vectors, this means A is a permutation matrix.) Now note that swapping columns of A gives a minus sign on both sides, so it does not change the validity of the formula. On the other hand, simultaneously changing rows of A and B, then both sides are preserved as well.

So WLOG by swapping columns of A, and simultaneously swapping rows of A and B, we can assume that $A = \begin{bmatrix} I_k \\ O \end{bmatrix}$. Say $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$. Then both sides are $\det(B_1)$, so the statement is true. □

This concludes the generalizations of Pythagorean theorem into all dimensions. Given a k-dimensional flat shape in $\mathbb{R}^n$, then (k-dimensional volumn of this shape)^2 is the sum of all (k-dimensional volumns of its shadows on a k-dimer over all possible k-dimensional coordinate subspaces. Here, coordinate subspaces refers to subspaces spanned by the cooridnate axes.

6.6 (Optional) Determinant Tricks

6.6.1 Block eliminations

So, how to actually compute determinants? The answer is as always: Gaussian elimination. I mean, this is the source of all computations. How to solve $A\boldsymbol{x} = \boldsymbol{b}$? Gaussian elimination. How to find LU decomposition? Gaussian eliminations. How to find QR decomposition? Do Gaussian elimination on $A^{\mathrm{T}}A$.

Given a matrix A, we do Gaussian elimination, which is equivalent to $PA = LDU$. Then $\det(A) = \operatorname{sign}(P)\det(D)$, because L, U are unit triangular. $\det(D)$ is just the product of diagonal entries, because it is a rectangular parallelotope. Note that you don't have to restrict yourself to row opertions though. If needed, column operations are just fine.

Just keep in mind: shearings preserve determinant, scalings scale determinants, and swappings NEGATE the determinant. The last one is easy to forget sometimes. You can do rows or columns as you see fit.

Example 6.6.1. Consider the matrix $\begin{bmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 0 & 2 & 0 \end{bmatrix}$. We cab swap the first row and the third row, then swap the first column and the secod colume. So we have

$$\det \begin{bmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 0 & 2 & 0 \end{bmatrix} = \det \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} = -\det \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix}.$$

This is now lower triangular, so the determinant is -2. ☺

In general, if you see no obvious clue about what to do, then Gaussian elimination is (as always) your best bet. Do NOT apply the big formula, because it would then take forever (unless we are facing some super special matrices). And in practice, no one ever use the big formula.

Now, recall that the funcation of column/row operations to simplify determinant is the fact that $\det(AB) = \det(A)\det(B)$. As a result, we can in fact use block operations to simply determinants. We have the following:

1. Block shearings preserve determinants as $\det \begin{bmatrix} I & A \\ O & I \end{bmatrix} = 1$.
2. Block scalings scale determinants as $\det \begin{bmatrix} A & \\ & I \end{bmatrix} = \det(A)$.
3. Block swaps might or might NOT change the sign. We have $\det \begin{bmatrix} & I_m \\ I_n & \end{bmatrix} = (-1)^{mn}$.

The last one might needs some justification. We want to move the m last columns to the left. The first of the m last columns, swap one by one to the left, needs a total of n swaps. Then the next one needs a total of n swaps as well. So on so forth for all m of them. So we performed a total of mn swaps. Alternatively, draw the diagram for $\begin{bmatrix} & I_m \\ I_n & \end{bmatrix}$ and see that the m parallel lines going up right would intersect ALL the n parallel lines going down right so we have a total of mn intersections.

Here is a nice application.

Proposition 6.6.2. *For any $m \times n$ matrix A and an $n \times m$ matrix B, $\det(I_m + AB) = \det(I_n + BA)$. Note that the suqare matrix on the left and on the right might have DIFFERENT sizes.*

Note that there are some deep relations between AB and BA in terms of eigenvalues, so one way is to go there. Here let me give you two other proofs.

Proof. This proof is straight forward. Change basis to simplify, then compute.

Say $A = P \begin{bmatrix} I_r & O \\ O & O \end{bmatrix} Q$ by rank normal form. This means we have changed basis in the domain and codomain of $A : \mathbb{R}^n \to \mathbb{R}^m$. So the corresponding transformation to B will be $B = Q^{-1}B'P^{-1}$ for some matrix B'. Say $B' = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$.

Now

$$\begin{aligned} \det(I_m + AB) &= \det(I_m + P \begin{bmatrix} B_1 & B_2 \\ O & O \end{bmatrix} P^{-1}) \\ &= \det(P(I_m + \begin{bmatrix} B_1 & B_2 \\ O & O \end{bmatrix})P^{-1}) \\ &= \det(I_m + \begin{bmatrix} B_1 & B_2 \\ O & O \end{bmatrix}) \\ &= \det \begin{bmatrix} B_1 + I_r & B_2 \\ O & I_{m-r} \end{bmatrix} \\ &= \det(B_1 + I_r). \end{aligned}$$

Similarly, $\det(I_n + BA) = \det(I_m + Q^{-1} \begin{bmatrix} B_1 & O \\ B_3 & O \end{bmatrix} Q) = \det(B_1 + I_r)$. So we are done. □

Proof. This is a traditional proof using the magic of block matrices.

When you have an $m \times n$ matrix and an $n \times m$ matrix, how would you put them into a single block matrix? You don't have much option but to do $\begin{bmatrix} I_n & B \\ A & I_m \end{bmatrix}$.

Now by block row-shearing, we have $\det\begin{bmatrix} I_n & B \\ A & I_m \end{bmatrix} = \det\begin{bmatrix} I_n & B \\ O & I_m - AB \end{bmatrix} = \det(I_m - AB)$, and by column-shearing we have $\det\begin{bmatrix} I_n & B \\ A & I_m \end{bmatrix} = \det\begin{bmatrix} I_n - BA & B \\ O & I_m \end{bmatrix} = \det(I_n - BA)$. So we have $\det(I_m - AB) = \det(I_n - BA)$. Replace A by $-A$ and we are done. □

Corollary 6.6.3. *$I_n + AB$ is invertible iff $I_m + BA$ is invertible. In particular, $I + \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ is invertible iff $1 + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u} \neq 0$.*

As cool as this corollary is, it suffers from the fact that it is redundant. The core of the argument is the LDU block decomposition. However, the LDU block decomposition of $\begin{bmatrix} I_n & B \\ A & I_m \end{bmatrix}$ already tells you that $I_n + AB$ is invertible iff $I_m + BA$ is invertible, in fact it gives you the FORMULA for finding the inverses, i.e., the Sherman-Morrison formula.

Nevertheless, both $\det(I_m + AB) = \det(I_n + BA)$ and the Sherman-Morrison formula is used mainly for ONE thing: low rank perturbations. If you realize that the $n \times n$ determinant you are working on is "almost" some nice matrix, where the difference has rank k which is small, then formula $\det(I + AB) = \det(I + BA)$ might reduce the computation of an $n \times n$ determinant to a $k \times k$ determinant. Let us see some examples.

Example 6.6.4. Consider $A = \begin{bmatrix} 1 + a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1}b_n \\ a_nb_1 & \dots & a_nb_{n-1} & 1 + a_nb_n \end{bmatrix}$. Find the determinant.

(Note that there is a tedious method of doing fancy row/column operations, and then set up inductions, and then find the generic formula. Try yourself if you want some mental exercise/torture.)

When we do mathematics, or science in general, it is VITAL to NOT ignore your first instinct and intuitions. Capture the fleeting glimps of genius perspective in your mind, and identify the related tools from that perspective, and the solution will be natural.

Think about this. If only all the ones on the diagonals are GONE, then our matrix would be $\begin{bmatrix} a_1b_1 & \dots & a_1b_n \\ \vdots & \ddots & \vdots \\ a_nb_1 & \dots & a_nb_n \end{bmatrix}$. This determinant is easy: it is zero (when $n > 1$). Because it has rank 1: all its columns are parallel! In fact, let $\boldsymbol{a}, \boldsymbol{b}$ be the obvious vector recording those a_i, b_i, then this matrix is $\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$.

In particular, $A = I + \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$. It is close to identity, where the difference has rank one. Time to use the formula $\det(I + AB) = \det(I + BA)$! We have $\det(A) = \det(I + \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}) = 1 + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{a} = 1 + \sum a_ib_i$. Done. ☺

Example 6.6.5. Consider $A = \begin{bmatrix} 1 + a_1 + b_1 & a_1 + b_2 & \dots & a_1 + b_n \\ a_2 + b_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1} + b_n \\ a_n + b_1 & \dots & a_n + b_{n-1} & 1 + a_n + b_n \end{bmatrix}$. Find the determinant.

(Note that there is a tedious method of doing fancy row/column operations, and then set up inductions, and then find the generic formula. This is even worse than last one.)

Think about this. If only all the ones on the diagonals are GONE, then our matrix would be $\begin{bmatrix} a_1 + b_1 & \dots & a_1 + b_n \\ \vdots & \ddots & \vdots \\ a_n + b_1 & \dots & a_n + b_n \end{bmatrix}$. This determinant is easy: it is zero (when $n > 2$). Because it has rank 2: all its columns are spanned by $\boldsymbol{a}, \boldsymbol{u}$, where $\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ and $\boldsymbol{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

In fact, let $\boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$, then this matrix is $\begin{bmatrix} \boldsymbol{a} & \boldsymbol{u} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix}$.

In particular, $A = I + \begin{bmatrix} \boldsymbol{a} & \boldsymbol{u} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix}$. It is close to identity, where the difference has rank two. Time to use the formula $\det(I + AB) = \det(I + BA)$!

We have $\det(A) = \det(I + \begin{bmatrix} \boldsymbol{a} & \boldsymbol{u} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix}) = \det(I_2 + \begin{bmatrix} \boldsymbol{u}^{\mathrm{T}} & \boldsymbol{b}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{u} \end{bmatrix}) = \det \begin{bmatrix} 1 + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{a} & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} \\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{a} & 1 + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u} \end{bmatrix} = (1 + \sum a_i)(1 + \sum b_i) - n \sum a_i b_i$. Done.

If you would like some statistical interpretation, note that we can further expand via $(1 + \sum a_i)(1 + \sum b_i) - n \sum a_i b_i = 1 + n\mathbb{E}(a) + n\mathbb{E}(b) - n^2(\mathbb{E}(ab) - \mathbb{E}(a)\mathbb{E}(b)) = 1 + n\mathbb{E}(a) + n\mathbb{E}(b) - n^2\mathrm{Cov}(a, b)$. So the degree one terms record the expected value, while the degree two term record the covariance. ☺

Example 6.6.6. This is a tough example. Many beginners will bang their head against this using row and column operations, hoping to find some magical operations or magical Laplace expansion, or hoping to set up induction, and all to no avail.

Consider $A = \begin{bmatrix} a_1 & b & \dots & b \\ b & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ b & \dots & b & a_n \end{bmatrix}$. You may try to do some operations and inductions, but the inductive formula would end up as $\det(A_n) = (a_n - b)\det(A_{n-1}) + b\prod_{i=1}^{n-1}(a_i - b)$. The generic solution is NOT easy to work out from this (albeit possible).

Instead, think about this. Wouldn't it be nice if we have a matrix where all entries are b? Then the answer is probably just zero, due to all the repeating columns. Similarly, if there are no b and all we had are the diagonal entries, then this is also nice, since determinant of a diagonal matrix is easy.

Capture this intuition and capitalize on this intuition. Our observation is essentially that our matrix A is rank one away from diagonal. Let $D = \begin{bmatrix} a_1 - b & & \\ & \ddots & \\ & & a_n - b \end{bmatrix}$ and $\boldsymbol{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Then $A = D + b\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$. Since A and D differ by a small rank, this means it is time for the formula $\det(I_m - AB) = \det(I_n - BA)$.

Suppose $a_i \neq b$ for all i, i.e., D is invertible. Then $\det(A) = \det(D + b\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}) = \det(D)\det(I + (bD^{-1}\boldsymbol{u})\boldsymbol{u}^{\mathrm{T}}) = \det(D)(1 + \boldsymbol{u}^{\mathrm{T}}(bD^{-1}\boldsymbol{u})) = [\prod(a_i - b)](1 + b\sum \frac{1}{a_i - b})$. Done. In particular, we reduced our calculation of $n \times n$ determinants to 1×1 determinants.

What if some $a_i = b$ happens? Note that in the formula $[\prod(a_i - b)](1 + b\sum \frac{1}{a_i - b})$, the denominators are "fake" , since once multiplied by $\prod(a_i - b)$, each summands of $1 + b\sum \frac{1}{a_i - b}$ would be clear of denominators. So it is in fact the same formula. Say $a_k = b$. Then the only summand without factor $(a_i - k)$ would be $b\prod_{i \neq k}(a_i - b)$, and this is the answer.

Of course, if some $a_k = b$, in fact it is easy to use row operations to make it upper triangular where the diagonals are simply the factors of $b\prod_{i \neq k}(a_i - b)$. I'll leave that method to you. ☺

Let me iterate again: the esssense of structural relation between $I + AB$ and $I + BA$ is a rank perturbation. Suppose we want to study an $n \times n$ matrix X, and we don't know how. However, it looks close to some nice matrix Y, say $X - Y$ has rank k. Then $X - Y = UV$ for some $n \times k$ matrix U and $k \times n$ matrix V. Then $X = Y + UV = Y(I_n + (Y^{-1}U)V)$. And we can utilize the relation between $I_n + (Y^{-1}U)V$ and $I_k + VY^{-1}U$. And now we have reduced the study of X, some $n \times n$ matrix, to the study of $I_k + VY^{-1}U$, some $k \times k$ matrix.

Example 6.6.7. Let us prove again that $A^{-1} = \frac{1}{\det(A)}C^{\mathrm{T}}$, without talking about the geometry of parallelopipes at all.

Consider $\frac{\partial}{\partial a_{ij}}\det(A)$. In essense, we are interested in $\det(A + \mathrm{d}x\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}})$, i.e., what would happen to the determinant if we increase the (i, j) entry a bit. Note that this is a rank one difference from A.

So we have $\det(A + \mathrm{d}x\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}) = \det(A)\det(I + \mathrm{d}xA^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}) = \det(A)(1 + \mathrm{d}x\boldsymbol{e}_j^{\mathrm{T}}A^{-1}\boldsymbol{e}_i)$. In particular, $\frac{\partial}{\partial a_{ij}}\det(A) = \lim_{\mathrm{d}x\to 0}\frac{\det(A+\mathrm{d}x\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}})-\det(A)}{\mathrm{d}x} = \lim_{\mathrm{d}x\to 0}\frac{\det(A)\,\mathrm{d}x\boldsymbol{e}_j^{\mathrm{T}}A^{-1}\boldsymbol{e}_i)}{\mathrm{d}x} = \det(A)\boldsymbol{e}_j^{\mathrm{T}}A^{-1}\boldsymbol{e}_i$.

But we also know that this is c_{ij}. So we see that $c_{ij} = \det(A)\boldsymbol{e}_j^{\mathrm{T}}A^{-1}\boldsymbol{e}_i$, i.e., the (i,j) entry of $\det(A)(A^{-1})^{\mathrm{T}}$. So $C = \det(A)(A^{-1})^{\mathrm{T}}$, from which we see that $A^{-1} = \frac{1}{\det(A)}C^{\mathrm{T}}$. ☺

Here is one final super interesting application.

Theorem 6.6.8 (Derivative of the determinant near identity is trace)**.** *Let* $f_A(t) = \det(I + tA)$. *Then* $f'_A(0) = \mathrm{trace}(A)$.

Proof. We go from I gradually towards $I + tA$. How to do this gradually? We do this one rank at a time. (One column at a time.)

Say $A = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_n\end{bmatrix}$. Then $A = \sum \boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}}$, this way we write A as a sum of n matrices of rank 1. (Can you see why?)

Now when t is infinitesimally small, we see that $\prod(I+t\boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}}) = I+t\sum \boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}}+t^2(\text{stuff}) = I+tA+t^2(\text{stuff})$, where t^2 would be too small to be relevant. So $\det(I + tA) = \det(\prod(I + t\boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}})) + t^2(\text{stuff})$ when t is infinitesimally small. (Note that the determinant is just a big polynomial in the entries, so the extra terms using t^2 entries would ALL have factor t^2.)

Note that $\det(\prod(I+t\boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}})) = \prod(\det(I+t\boldsymbol{a}_i\boldsymbol{e}_i^{\mathrm{T}})) = \prod(1+t\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{a}_i) = \prod(1+ta_{ii}) = 1+t\,\mathrm{trace}(A)+t^2(\text{stuff})$.

So all in all, we have $\det(I + tA) = 1 + t\,\mathrm{trace}(A) + t^2(\text{stuff})$. Taking derivative at $t = 0$, our result is obvious. □

Note that we did NOT use ANY property of trace other than its definition as the sum of diagonals. So in fact this gives a super cool proof that $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.

Proposition 6.6.9. $\mathrm{trace}(AB) = \mathrm{trace}(BA)$. *In particular, this implies that* $\mathrm{trace}(XAX^{-1}) = \mathrm{trace}(A)$, *i.e., trace is invariange under change of basis.*

Proof. Consider $\det(I + tAB) = \det(I + tBA)$. Take derivative and done. This is my personal favorite proof. (And in fact this explains what trace and determinants are, from the perspective of Lie group and Lie algebra, though they are outside the scope of this class.) □

6.6.2 Shear and Expand and Induction

This is the most traditional strategy to tackle any hard and tricky determinant problem. You do row and column operations to generate as much zero entries as possible. Hopefully these zeroes will concentrate, and when they concentrate in some row or in some column, you expand along that row or that column.

Sometime, this expansion allows you to reduce some $n \times n$ determinant into some $(n-1)\times(n-1)$ determinant, and then you can use induction.

Example 6.6.10. Consider an $n \times n$ matrix $A_n = \begin{bmatrix} 1 & -1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & 1 & 1 \end{bmatrix}$. What is $\det(A_n)$?

Let us do Laplace expansion alongthe first column. Then we see that

$$\det(A_n) = 1\det(A_{n-1}) + 1(-\det\begin{bmatrix}-1 & O \\ & A_{n-2}\end{bmatrix}) = \det(A_{n-1}) + \det(A_{n-2}).$$

In particular, if we think of the $\det(A_n)$ as a sequence for increasing n, then each term is the sum of previous two terms. Sounds familiar? This is the same induction scheme for the famous Fibbonacci sequence.

Furthermore, $\det(A_1) = 1$ and $\det(A_2) = 2$. Hence this sequence goes like $1, 2, 3, 5, 8, 13, \dots$. It is indeed the Fibbonacci sequence.

If you have hardcore skills in working out sequences using mathematical induction, the you can find a formula. We have $\det(A_n) = \frac{\phi^{n+1}-(1-\phi)^{n+1}}{\sqrt{5}}$ where $\phi = \frac{1+\sqrt{5}}{2}$ is the golden ratio. ☺

Example 6.6.11. Consider $\det A$ where $A = \begin{bmatrix} a_1 & b & \dots & b \\ c & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ c & \dots & c & a_n \end{bmatrix}$, where $b \neq c$. Note that the case when $b = c$ is solved previously using the formula $\det(I + AB) = \det(I + BA)$.

This determinant is notoriously tricky for beginners. Here I did NOT present the fastest solution. However, this is a faithful implement of the strategy lined out above. The purpose of doing this problem is NOT the result or speed, but rather on the strategy itself. I will intentionally try to AVOID magical row/column operations that simplify this at various places.

You see, the obvious direction to go is to, say, take the second to last column and substract this from the last column. This gives $\begin{bmatrix} a_1 & b & \dots & b & 0 \\ c & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & b & 0 \\ c & \dots & c & a_{n-1} & b - a_{n-1} \\ c & \dots & c & c & a_n - c \end{bmatrix}$.

Now we have a concentration of zeros on the last column, so we expand. This gives us:

$$\det A = (a_n - c) \det \begin{bmatrix} a_1 & b & \dots & b \\ c & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ c & \dots & c & a_{n-1} \end{bmatrix} + (-1)(b - a_{n-1}) \det \begin{bmatrix} a_1 & b & \dots & b & b \\ c & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & b & b \\ c & \dots & c & a_{n-2} & b \\ c & \dots & c & c & c \end{bmatrix}.$$

Now the first term here is clearly a good set up for induction. What about the second term here?

Let $A_k = \begin{bmatrix} a_1 & b & \dots & b \\ c & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ c & \dots & c & a_k \end{bmatrix}$ and $B_k = \begin{bmatrix} a_1 & b & \dots & b & b \\ c & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & b & b \\ c & \dots & c & a_k & b \\ c & \dots & c & c & c \end{bmatrix}$. We have $\det(A_n) = (a_n - c) \det(A_{n-1}) + (a_{n-1} - b) \det(B_{n-2})$. Let us first figure out the B_k portion of this formula.

Again use the second to last column of B_k and substract this from the last column. Then we see that

$$\det(B_k) = \det \begin{bmatrix} a_1 & b & \dots & b & 0 \\ c & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & b & 0 \\ c & \dots & c & a_k & b - a_k \\ c & \dots & c & c & 0 \end{bmatrix} = (a_k - b) \det(B_{k-1}).$$

This set up the induction nicely. So we have $\det(B_k) = (a_k - b) \det(B_{k-1}) = (a_k - b)(a_{k-1} - b) \det(B_{k-2}) = \cdots = [\prod_{i=1}^{k} (a_i - b)] \det(B_0)$. Note that $B_0 = \begin{bmatrix} c \end{bmatrix}$. So $\det(B_k) = c \prod_{i=1}^{k} (a_i - b)$.

Go back to A_n, we have $\det(A_n) = (a_n - c) \det(A_{n-1}) + (a_{n-1} - b) \det(B_{n-2}) = (a_n - c) \det(A_{n-1}) + c \prod_{i=1}^{n-1} (a_i - b)$. Now ideally, one must be able to deduce the generic formula from this inductive formula. In parctice, while doable, this is a bit annoying and lengthy. So instead, we cheat by using symmetry.

Note that $\det(A_n) = \det(A_n^{\mathrm{T}})$, yet the difference between A_n and A_n^{T} is simply swapping b, c. As a result of this symmetry, if we have $\det(A_n) = (a_n - c) \det(A_{n-1}) + c \prod_{i=1}^{n-1} (a_i - b)$, we must symmetrically have $\det(A_n) = (a_n - b) \det(A_{n-1}) + b \prod_{i=1}^{n-1} (a_i - c)$. Now use $(a_n - b)$ times the first equation minus $(a_n - c)$ times the second equation, we obtain the equation $(c - b) \det(A_n) = c \prod_{i=1}^{n} (a_i - b) - b \prod_{i=1}^{n} (a_i - c)$.

In particular, if $b \neq c$, then we are done. We have $\det(A_n) = \frac{c}{c-b}\prod_{i=1}^{n}(a_i - b) + \frac{b}{b-c}\prod_{i=1}^{n}(a_i - c)$. ☺

Note that this example here cannot be done via low rank perturbations, since the b portion and c portion are both triangular with almost full rank.

Here is another very tough one.

Example 6.6.12. Consider the matrix $A = \begin{bmatrix} a & 1 & & \\ n & \ddots & \ddots & \\ & \ddots & \ddots & n \\ & & 1 & a \end{bmatrix}$. Find the determinant. (This is the notorious Kac-Sylvester matrix.)

This is seriously annoying, even with all the zero entries. Say if you expand along the last row, then you shall see that you FAIL to set up an induction, because the 1 through n and n through 1 anti-symmetric pattern here.

The magic trick is this. Consider say $n = 5$ for illustration purppose. We have $\begin{bmatrix} a & 1 & & & & \\ 5 & a & 2 & & & \\ & 4 & a & 3 & & \\ & & 3 & a & 4 & \\ & & & 2 & a & 5 \\ & & & & 1 & a \end{bmatrix}$. Now, we add the i-th row to the $(i-2)$-th row, $(i-4)$-th row and so on, for all i. We do these row operations in the order of increasing i. Effectively, it is as if we are going from A to EA for the matrix $E = \begin{bmatrix} 1 & 0 & 1 & \dots & \cdot \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & 1 \\ & & & \ddots & 0 \\ & & & & 1 \end{bmatrix}$, where the values alternate between 1 and 0 for this upper triangular matrix E.

After all these row shearings, we end up with $EA = \begin{bmatrix} a & 5 & a & 5 & a & 5 \\ 5 & a & 5 & a & 5 & a \\ & 4 & a & 5 & a & 5 \\ & & 3 & a & 5 & a \\ & & & 2 & a & 5 \\ & & & & 1 & a \end{bmatrix}$.

This looks neat, yes?

Next, we do E^{-1} as a column operation, and go from EA to EAE^{-1}. In terms of operations, it means we substract the i-th column from the $(i+2)$-th column, $(i+4)$-th column, and so on, for all i. We do these column operations in the order of increasing i.

This way all the extra a's introduced by shearing will cancel out, we end up with $\begin{bmatrix} a & 5 & & & & \\ 5 & a & & & & \\ & 4 & a & 1 & & \\ & & 3 & a & 2 & \\ & & & 2 & a & 3 \\ & & & & 1 & a \end{bmatrix}$.

Note that this is block lower triangular. So $\det(A_5) = (a^2 - 5^2)\det(A_3) = (a+5)(a-5)\det(A_3)$.

By induction, you can see that $\det(A_n) = (a+n)(a-n)\det(A_{n-2}) = (a+n)(a+n-2)\dots(a-n+2)(a-n)$. For example, for A_5 the determinant is $(a+5)(a+3)(a+1)(a-1)(a-3)(a-5)$.

Note that, essentially, we doing $\det(EAE^{-1})$, and as a linear transformation, EAE^{-1} is merely A after a change of basis! Indeed, for any invertible X, then $\det(XAX^{-1}) = \det(X)\det(A)\det(X)^{-1} = \det(A)$. So just like trace is invariant under a change of basis, determinant is also invariant under a change of basis. And similar matrices must always have the same determinant.

In essence, E^{-1} is a basis on which the linear map A has slightly better behavior. We change basis, and induction appear. ☺

6.6.3 Polynomial Interpretation, Interpolation and the Vandermonde Matrix

One niche use of the determinant is to show that a matrix is invertible. However, bear in mind that MOST of the time it is easier to simply look at the kernel. Nevertheless, the following case is both interesting and useful later.

Example 6.6.13. Two points determines a line. Three points determines a parabola $p(x) = a + bx + cx^2$. In general, you can imagine that $n + 1$ points can determine a polynomial of degree n. Can we prove this?

Consider an unknown parabola $p(x) = a + bx + cx^2$. Suppose we know that $p(1) = 2, p(2) = 3, p(3) = 4$, let us find the unique parabola through them. Then we have a linear system

$$\begin{aligned} a + b + c &= 2 \\ a + 2b + 4c &= 3 \\ a + 3b + 9c &= 4. \end{aligned}$$

Write this in matrix form, and we have

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}.$$

So, to find a, b, c, it is enough to solve this system using Gaussian elimination. Note that the matrix here is really nice. As a linear map, it sends "coefficients" of a polynomial to "evaluations" of the polynomial at the given points. So, given evaluations $p(1) = 2, p(2) = 3, p(3) = 4$, to solve the unknown coefficients is the same as doing the inverse of this linear map. ☺

Definition 6.6.14. *The Vandermonde matrix is* $V_{n+1} = \begin{bmatrix} 1 & a_1 & \dots & a_1^n \\ 1 & a_2 & \dots & a_2^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{n+1} & \dots & a_{n+1}^n \end{bmatrix}$ *for some chosen constants* $a_1, \dots, a_{n+1} \in \mathbb{R}$.

Example 6.6.15. Let V be the space of polynomials of degree at most n. This is an $(n+1)$ dimensional space. Pick any $x_1, \dots, x_{n+1} \in \mathbb{R}$, then we can build a linear map $E : V \to \mathbb{R}^{n+1}$ such that $E(p) = \begin{bmatrix} p(x_1) \\ \vdots \\ p(x_{n+1}) \end{bmatrix}$.

This process is secretly linear. Even though $p(x)$ is probably NOT linear in x, but the expression $p(x)$ is linear in p. What I mean is, given two polynomials p, q, then $(ap + bq)(x) = ap(x) + bq(x)$.

Now, if E is a linear map, what is the matrix for E? Well, pick basis $1, x, x^2, \dots, x^n$ for V, and pick the standard basis for $\mathbb{R}^{n+1}$.

The image for $E(1)$ is $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. The image $E(x^k)$ is $\begin{bmatrix} x_1^k \\ \vdots \\ x_{n+1}^k \end{bmatrix}$ for each k. So all in all, the matrix for E is $\begin{bmatrix} 1 & x_1 & \dots & x_1^n \\ 1 & x_2 & \dots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n+1} & \dots & x_{n+1}^n \end{bmatrix}$.

As you can see, the Vandermonde matrix is precisely the linear map E. ☺

Theorem 6.6.16 (Polynomial Interpolation Theorem)**.** *Given any distinct inputs $x_1, \ldots, x_{n+1}$ and corresponding outputs $y_1, \ldots, y_{n+1}$, there is a UNIQUE polynomial $p(x)$ of degree at most n such that $p(x_i) = y_i$ for all i.*

Proof. Suppose the unknown polynomial is $p(x) = a_0 + a_1 x + \cdots + a_n x^n$. Then let V be the Vandermonde matrix for $x_1, \ldots, x_{n+1}$, we have

$$V \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n+1} \end{bmatrix}.$$

So we have unique solution if and only if V is invertible.

We do this by showing that the determinant is non-zero. Look at the lemma below. □

Proposition 6.6.17. $\det(V_n) = \prod_{i<j}(a_j - a_i)$*. In particular, if all a_i are distinct, then V_n is invertible.*

Proof. We give two proofs.

The first proof is traditional. To find $\det V_{n+1}$, going from RIGHT to LEFT, we smartly replace each column $\boldsymbol{c}_i$ by $\boldsymbol{c}_i - a_{n+1}\boldsymbol{c}_{i-1}$. (Yeah, this is NOT a typo. We meant to write a_{n+1} here.) These are all shearings and they preserve the determinant. (The total column action here is equivalent to multiplying $\begin{bmatrix} 1 & -a_{n+1} & & \\ & \ddots & \ddots & \\ & & \ddots & -a_{n+1} \\ & & & 1 \end{bmatrix}$ to the right of V_{n+1}.)

Then magically, we now only need to calculate $\det \begin{bmatrix} 1 & (a_1 - a_{n+1}) & a_1(a_1 - a_n) & \ldots & a_1^{n-1}(a_1 - a_n) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & (a_n - a_{n+1}) & a_n(a_n - a_{n+1}) & \ldots & a_n^{n-1}(a_n - a_{n+1}) \\ 1 & 0 & 0 & \vdots & 0 \end{bmatrix}$.

Now expand along the last row (or consider non-zero terms in the big formula), we see that this is $(-1)^{n+1} \det \begin{bmatrix} (a_1 - a_{n+1}) & a_1(a_1 - a_n) & \ldots & a_1^{n-1}(a_1 - a_n) \\ \vdots & \ddots & \ddots & \vdots \\ (a_n - a_{n+1}) & a_n(a_n - a_{n+1}) & \ldots & a_n^{n-1}(a_n - a_{n+1}) \end{bmatrix}$.

Take out common factors, we have $(-1)^{n+1}(a_1 - a_{n+1}) \ldots (a_n - a_{n+1}) \det \begin{bmatrix} 1 & a_1 & \ldots & a_1^{n-1} \\ 1 & a_2 & \ldots & a_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & \ldots & a_n^{n-1} \end{bmatrix} = \prod_{1 \le i \le n}(a_{n+1} - a_i) \det V_n$. Hey! By induction (and the base case is very trivial), we have

$$\det V_{n+1} = [\prod_{1 \le i \le n} (a_{n+1} - a_i)][\prod_{1 \le i < j \le n} (a_j - a_i)] = \prod_{1 \le i < j \le n+1} (a_j - a_i).$$

Well, this is not easy to think of, but it gets the job done. Now let me show you my favorite proof.

Recall that, according to the big formula, $\det(V_n)$ is simply a big polynomial on the variables $a_1, \ldots, a_n$. What are its factors?

Note that, if $a_i = a_j$, then the i-th row and j-th row of V_n would coincide, so we would have $\det(V_n) = 0$. This means that the polynomial $a_j - a_i$ is a factor of the polynomial $\det(V_n)$ for all $i \neq j$. (In fact, since we already know the answer, these are ALL the factors. HAHAHA!)

In particular, we see that $\prod_{i<j}(a_j - a_i)$ is a factor of $\det(V_n)$. Now each term in the big formula of $\det(V_n)$ has a total degree of $1 + 2 + \cdots + (n-1) = \frac{1}{2}n(n-1)$, which is EXACTLY the degree of $\prod_{i<j}(a_j - a_i)$. So we have $\det(V_n) = k \prod_{i<j}(a_j - a_i)$.

Now check the coefficient of the term $a_2a_3^3\dots a_n^{n-1}$ in both sides. $\prod_{i<j}(a_j-a_i)$ can only reach this term by always pick a_j in each factor. $\det(V_n)$ can only reach this term by picking the diagonal entries. So we see that both coefficients are one. Hence $k=1$ and we are done. (On the product side, this is obvious. On the determinant side, this term can only come from the diagonal.) □

Remark 6.6.18. *The second proof above is useful for many similar things. For example, consider the Cauchy matrix* $\begin{bmatrix} \frac{1}{x_1+y_1} & \cdots & \frac{1}{x_1+y_n} \\ \vdots & \ddots & \vdots \\ \frac{1}{x_n+y_1} & \cdots & \frac{1}{x_n+y_n} \end{bmatrix}$. *What is this determinant? Try it yourself. People has worked out the LDU decompositions and inverse of Cauchy matrices as well (see wikipedia and related papers).*

Here's some optional fun fact. A special case of the Cauchy matrix is the Hilbert matrix, where $x_i=y_i=i$ for all i. If we let V be the space of polynomials of degree at nost $n-1$, with basis $1,x,\dots,x^{n-1}$, and inner product $\langle p(x),q(x)\rangle=\int_0^1 p(x)q(x)\,\mathrm{d}x$, then the $n\times n$ Hilbert matrix is actually the corresponding Gram matrix.

Polynomial interpolation is the major reason we bother with the Vandermonde matrix. (Another reason is the study of eigenspaces in later chapters.) Now that we have its determinant, it is time to find its inverse. How to find $V_n^{-1}\boldsymbol{e}_1$, for example? (I.e., how to find the first column of V_n^{-1}.)

Given any distinct inputs $a_1,\dots,a_n$, we want to find the UNIQUE polynomial $p(x)$ of degree at most $n-1$ such that $p(a_i)=0$ for all $i\neq 1$ and $p(a_1)=1$. Then $x-a_i$ for all $i\neq 1$ should be a factor of this polynomial.

Consider the polynomial $\prod_{i\neq 1}(x-a_i)$. It has degree $n-1$, and takes value 0 on $a_2,\dots,a_n$, and it takes value $\prod_{i\neq 1}(a_1-a_i)$ at a_1. So let $p(x)=\frac{\prod_{i\neq 1}(x-a_i)}{\prod_{i\neq 1}(a_1-a_i)}$, and we are done. (Note the similarity in ideas with the second proof of the Vandermonde determinant! Finding factor polynomials is the essense of this problem. Column operations work, but not as elegant, because it is not the essense of the problem.)

Proposition 6.6.19 (Lagrange Interpolation)**.** *Given any distinct inputs $a_1,\dots,a_{n+1}$ and corresponding outputs $b_1,\dots,b_{n+1}$, then the UNIQUE polynomial $p(x)$ of degree at most n is $p=\sum b_kp_k$, where $p_k(x)=\frac{\prod_{i\neq k}(x-a_i)}{\prod_{i\neq k}(a_1-a_i)}$.*

The coeffients of p_k are the entries of V_{n+1}^{-1}. They are not pretty, and it is usually easier to do abstractly via polynomials.

One can go even further to get the more general Hermite interpolation, though that will require Jordan canonical form, and thus outside of current ability yet.

6.6.4 (Optional) A determinant game

This game is proposed by a Putnam mathematical contest. Suppose Alice and Bob are playing a game. We have an $n\times n$ matrix with unfilled entries. Alice and Bob take turns to fill out the entries. Say if Alice goes first, she can fill out a number in an entry. Then Bob can go, and fill out another number in another entry. The game ends when the matrix is filled.

Now if the game ends with an invertible matrix, then Alice wins. If the game ends with determinant zero, the Bob wins. Suppose Alice goes first, and $n=2020$, who has the winning strategy?

Proposition 6.6.20. *If n is even and Alice goes first, then Bob always win.*

Proof. If Alice write a number in the first or second column, then Bob write the same number in the same row but in the second or first column, respectively. If Alice write a number outside of the first and second columns, then Bob also fill out an arbitrary entry outside of the first and second columns.

This way, Bob can guarantee that the first two column of the resulting matrix is always the same. So Bob wins. □

Now, this will NOT work if n is odd. In that case, there would be an ODD number of entries outside of the first and second columns. So if Alice keep playing outside of the first two columns, then Bob would eventually be forced to go first in the first and second column as there is nowhere else to play. Thus this strategy would not work.

However, Bob might still win with some other strategies.

Proposition 6.6.21. *If $n = 3$ and Alice goes first, then Bob always win.*

Proof. Note that we can permute the rows and columns at all time without changeing the invertibility of the matrix. So WLOG suppose Alice played in the upper left corner. If Alice fill it by zero, then Bob can fill a zero below. Then Alice must fill the last entry in the first column by a non-zero number, else she would loose by Bob filling out the first column with all zeros. Next Bob can fill out the zero at the center, and Alice has no choice but to block at the center-right entry, else lose by a central row of zeros. Then Bob can fill out the last of the upper 2×2 block, so that this entire block is zero. This 2×2 block of zeros would force the matrix to have rank one, and hence determinant zero. Bob wins.

Obviously Alice is better off playing a non-zero number to start with. Suppose the first step of Alice is non-zero in the upper left corner. Now, by wisdom granted to us via the tic-tac-toe game, Bob's winning strategy is to fill out the center entry by zero, thus we have $\begin{bmatrix} a_1 & & \\ & 0 & \\ & & \end{bmatrix}$. Note that, from here on out, Alice CANNOT allow a row of zeros or a column of zeros. The tic-tac-toe feeling is strong now.

Now Alice play next. Afterwards Bob can keep filling out zeros to force a 2×2 block of zeros, and Alice has no choice but to block rows or columns on the way. This 2×2 block of zeros would force the matrix to have rank one, and hence determinant zero. Specifically, we have the following four possibilities, where the subscript indicates the order of steps. Alice will play $a_1, a_2, \ldots, a_5$ while Bob will only play zeros, $0_1, \ldots, 0_5$.

The four possibilities are (up to taking transpose)

$$\begin{bmatrix} a_1 & a_3 & a_5 \\ a_2 & 0_1 & 0_4 \\ a_4 & 0_2 & 0_3 \end{bmatrix}, \begin{bmatrix} a_1 & a_3 & a_5 \\ a_4 & 0_1 & 0_3 \\ a_2 & 0_2 & 0_4 \end{bmatrix}, \begin{bmatrix} a_1 & 0_4 & 0_3 \\ a_3 & 0_1 & 0_2 \\ a_5 & a_2 & a_4 \end{bmatrix}, \begin{bmatrix} a_1 & 0_3 & 0_4 \\ a_3 & 0_1 & 0_2 \\ a_5 & a_4 & a_2 \end{bmatrix}.$$

□

However, Bob do not always win. For example, if $n = 1$, obviously Alice would win. The generic case when n is odd is still unknown. I have no idea about who might win the $n = 5$ game. (But I think playing only zeros is no longer a viable strategy for Bob.)

What if Bob plays first? Then if n is odd, Bob can just fill out some entry outside of the first and second columns, and then resume the copy-cat strategy to win.

What if n is even and Bob goes first? I have no clue yet. You might want to try the easy case when $n = 2, 4$.

Chapter 7

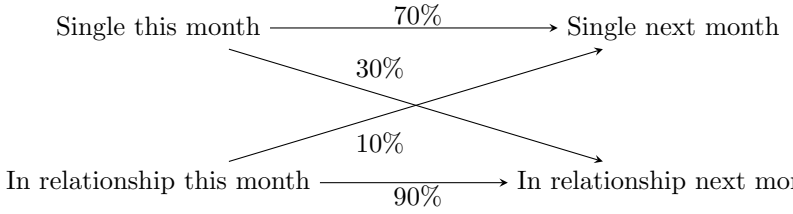
Eigenstuff

7.1 Introduction

Eigenvalues first started as a vital piece of dynamics. Think about the following example.

Example 7.1.1. Suppose we are studying in the famous Terribly Happy University (THU for short). Students are generally in two romantic status: single, or in a relationship.

If a student is single, since everyone is so happy, suppose 30% will be in a relationship next month. However, 70% shall remain single. If a student is in a relationship, then suppose 90% shall remain in the relationship next month, while 10% will break up and become single.



.

Now as months go by, what would happen in this dynimical system?

Say we started with x students single and y students in a relationship. Write this as a vector $\boldsymbol{x}_0 = \begin{bmatrix} x \\ y \end{bmatrix}$. Then next month, we shall have a distribution of $\boldsymbol{x}_1 = A\boldsymbol{x}_0$, where $A = \begin{bmatrix} 0.7 & 0.1 \\ 0.3 & 0.9 \end{bmatrix}$. Can you see why?

And then the next month, we shall have a distribution of $\boldsymbol{x}_2 = A\boldsymbol{x}_1$. As months go by, we are basically applying A again and again iteratively. So our question is now this: what is the limiting behavior of the sequence $\boldsymbol{x}_0, A\boldsymbol{x}_0, A^2\boldsymbol{x}_0, \ldots$?

Say $\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, i.e., 100% of the students are single initially. Then calculation yields the sequence as (rounding to the forth digit after the decimal point):

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}, \begin{bmatrix} 0.52 \\ 0.48 \end{bmatrix}, \begin{bmatrix} 0.412 \\ 0.588 \end{bmatrix}, \begin{bmatrix} 0.3472 \\ 0.6528 \end{bmatrix}, \begin{bmatrix} 0.3083 \\ 0.6917 \end{bmatrix}, \begin{bmatrix} 0.2850 \\ 0.7150 \end{bmatrix}, \begin{bmatrix} 0.2710 \\ 0.7290 \end{bmatrix}, \begin{bmatrix} 0.2626 \\ 0.7374 \end{bmatrix}, \begin{bmatrix} 0.2576 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 0.2545 \\ 0.7455 \end{bmatrix}, \begin{bmatrix} 0.2527 \\ 0.7473 \end{bmatrix} \cdots$$

As you can see, the sequence seems to be converging to $\boldsymbol{x}_\infty = \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix}$. Eventually, 75% of the students are in a relationship, and only 25% of the students will remain single.

Is there a better and more rigorous way to find $\boldsymbol{x}_\infty$? Yes there is! If $\boldsymbol{x}_\infty$ is indeed the stable equilibrium, then we should have $A\boldsymbol{x}_\infty = \boldsymbol{x}_\infty$. In particular, $\boldsymbol{x}_\infty \in \mathrm{Ker}(A - I)$. Now $A - I = \begin{bmatrix} -0.3 & 0.1 \\ 0.3 & -0.1 \end{bmatrix}$, and its

kernel is spanned by $\begin{bmatrix}1\\3\end{bmatrix}$, which shows that 25% of the students are single in the equilibrium. We find our limiting behavior right away!

Not only so, we can also find the speed of convergence. Suppose our current distribution is $\begin{bmatrix}0.25+x\\0.75-x\end{bmatrix}$ for some x. Then $A\begin{bmatrix}0.25+x\\0.75-x\end{bmatrix} = \begin{bmatrix}0.25+0.6x\\0.75-0.6x\end{bmatrix}$. As you can see, the deviation is shrunk by a factor of 0.6, wow!

In particular, if we started with $\begin{bmatrix}0.25+x\\0.75-x\end{bmatrix}$, then next month we would have $\begin{bmatrix}0.25+0.6x\\0.75-0.6x\end{bmatrix}$, and next month we would have $\begin{bmatrix}0.25+0.6^2x\\0.75-0.6^2x\end{bmatrix}$. The deviation would shrunk exponentially towards zero.

Let us sum up the phenomina here. First of all, we can observe that

$$A\begin{bmatrix}1\\3\end{bmatrix} = \begin{bmatrix}1\\3\end{bmatrix},$$
$$A\begin{bmatrix}1\\-1\end{bmatrix} = 0.6\begin{bmatrix}1\\-1\end{bmatrix}.$$

Both of these have significant meanings. The first one describes the eventual equilibrium must be in this direction. The second one tells us how we would converge to this equilibrium: in the direction of $\begin{bmatrix}1\\-1\end{bmatrix}$, and the speed is governed by the value 0.6.

To sum up, the solutions to $A\boldsymbol{x} = \lambda\boldsymbol{x}$ completely describe ALL info we need for this dynamical system. ☺

We'd better give this beautiful info names. They are called eigenstuff.

Definition 7.1.2. *For a matrix A, if $A\boldsymbol{x} = \lambda\boldsymbol{x}$ for some $\lambda \in \mathbb{R}$ and some nonzero vector $\boldsymbol{x}$, then we call λ an **eigenvalue** of A, and $\boldsymbol{x}$ an **eigenvector** of A for the eigenvalue λ.*

Why do we require $\boldsymbol{x}$ here to be non-zero? Mainly because if we allow the zero vector, then $A\mathbf{0} = \lambda\mathbf{0}$ for all λ. These solutions are entirely trivial and useless.

Now, a major point of eigenstuff is to help us understand the behavior of $A, A^2, A^3, \ldots$. In some very nice cases, this is very easy to do. For example, if A is diagonal say $\begin{bmatrix}a & \\ & b\end{bmatrix}$, then obviously A^k is done by simply raising the diagonal entries to the corresponding power, i.e., $\begin{bmatrix}a^k & \\ & b^k\end{bmatrix}$. But for a generic A, say $A = \begin{bmatrix}0.7 & 0.1\\0.3 & 0.9\end{bmatrix}$, how can we obtain a formula for calculating A^k? This is where eigenstuff come in to help. They help describe the "dynamics" behind the linear transformation A.

Example 7.1.3. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a basis of $\mathbb{R}^n$ and they are also eigenvectors of A, say $A\boldsymbol{x}_i = \lambda_i\boldsymbol{x}_i$. One might call this an ***eigenbasis*** of $\mathbb{R}^n$ for A.

Then in particular we have $A\begin{bmatrix}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n\end{bmatrix} = \begin{bmatrix}\lambda_1\boldsymbol{x}_1 & \ldots & \lambda_n\boldsymbol{x}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}$.

Let $X = \begin{bmatrix}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n\end{bmatrix}, \Lambda = \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}$, then we see that $AX = X\Lambda$, and $A = X\Lambda X^{-1}$. Hey!

If you remember, structures such as $X\Lambda X^{-1}$ refers to a change of basis for a linear transformation. In particular, it means that if we change basis from the standard basis to the eigenbasis, then A will change into a diagonal matrix Λ.

This is indeed the case. If we pick $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ as the new basis, then $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$ says exactly that A is diagonal in this new basis.

So, how to study A? We first go into this new basis. Now our linear map is represented by a diagonal matrix Λ, and its k-th power is simply $\Lambda^k = \begin{bmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{bmatrix}$. Powers of a diagonal matrix is super easy to do! Now we change back to the standard basis, and we see that $A^k = X\Lambda^k X^{-1} = X \begin{bmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{bmatrix} X^{-1}$.

Alternatively, one might also compute directly via $A^k = X\Lambda X^{-1} X\Lambda X^{-1} \dots X\Lambda X^{-1}$. Note that all the X and X^{-1} in the middle will cancel out, and we have $A^k = X\Lambda^k X^{-1}$.

Alternatively, I also enjoy this lovely diagram below, which make things quite clear.

$$\begin{array}{ccccccc} \mathbb{R}^n & \xrightarrow{\Lambda} & \mathbb{R}^n & \xrightarrow{\Lambda} & \dots & \xrightarrow{\Lambda} & \mathbb{R}^n \\ {\scriptstyle X}\downarrow & & {\scriptstyle X}\downarrow & & & & {\scriptstyle X}\downarrow \\ \mathbb{R}^n & \xrightarrow{A} & \mathbb{R}^n & \xrightarrow{A} & \dots & \xrightarrow{A} & \mathbb{R}^n \end{array}$$

For example, in our previous case of $A = \begin{bmatrix} 0.7 & 0.1 \\ 0.3 & 0.9 \end{bmatrix}$, we see that $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ is an eigenvector for the eigenvalue 1, and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ is an eigenvector for the eigenvalue 0.6. Since the two linearly independent vectors span $\mathbb{R}^2$, this immediately imply that $\begin{bmatrix} 0.7 & 0.1 \\ 0.3 & 0.9 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 0.6 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix}^{-1}$. (Calculate and see that this is indeed correct.)

We also immediately obtain the formula

$$\begin{bmatrix} 0.7 & 0.1 \\ 0.3 & 0.9 \end{bmatrix}^k = \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 0.6 \end{bmatrix}^k \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 0.6^k \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & -\frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} + \frac{3}{4}(0.6^k) & \frac{1}{4} - \frac{1}{4}(0.6^k) \\ \frac{3}{4} - \frac{3}{4}(0.6^k) & \frac{3}{4} + \frac{1}{4}(0.6^k) \end{bmatrix}.$$

Thus we have obtained a direct formula for A^k.

Finally, let us calculate $A^\infty = \lim_{k\to\infty} A^k = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{3}{4} \end{bmatrix}$. What is this? This is the oblique projection to the line in the direction $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ along a kernel direction of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. ☺

This technique is called the similarity diagonalization of A.

Definition 7.1.4. *We say square matrices A, B are **similar** if we can find an invertible X such that $A = XBX^{-1}$. (I.e., A, B only differ by a change of basis.)*

*If A is similar to a diagonal matrix, then we say that A is **diagonalizable**.*

Proposition 7.1.5. *A square matrix A is diagonalizable if and only if eigenvectors of A span the whole domain.*

Proof. If eigenvectors of A span the whole domain, then we can pick some of them to form a basis, and thus A is diagonalizable as shown in the example above. We only need to prove the other direction now.

Suppose $A = XDX^{-1}$ for some invertible $X = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix}$ and diagonal $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$. Then $AX = XD$, which means $A\boldsymbol{x}_i = d_i \boldsymbol{x}_i$. Since X is invertible, all these $\boldsymbol{x}_i$ are nonzero, so they are all eigenvectors. Again since X is invertible, they span the whole space. □

Now, most of the matrices are diagonalizable. (Exceptions exist but rare.) Then our goal is simple: how to find these eigenvalues and eigenvectors?

Proposition 7.1.6. *λ is an eigenvalue of A if and only if $A-\lambda I$ is not invertible, if and only if* $\det(A-\lambda I) = 0$*, if and only if* $\det(\lambda I - A) = 0$.

And eigenvectors for the eigenvalue λ are exactly non-zero vectors of $\mathrm{Ker}(A - \lambda I)$.

Proof. Just look at definition. □

Example 7.1.7. Again consider the matrix $A = \begin{bmatrix} 0.7 & 0.1 \\ 0.3 & 0.9 \end{bmatrix}$. If x is an eigenvalue, then we should have $\det(xI - A) = 0$.

Now $\det(xI - A) = \det \begin{bmatrix} x - 0.7 & -0.1 \\ -0.3 & x - 0.9 \end{bmatrix} = x^2 - 1.6x + 0.6$. So as you can see, x is an eigenvalue if and only if $x^2 - 1.6x + 0.6 = 0$. Solving this gives $x = 1$ or $x = 0.6$. So the only eigenvalues for A are 1 and 0.6. ☺

So the standard procedure goes like this: for our $n \times n$ matrix A, consider $\det(A - xI)$. This is some polynomial in x with degree n. Solve $\det(A - xI) = 0$, then solutions to this polynomial equation are EXACTLY the eigenvalues.

Once we have the eigenvalues, for each eigenvalue λ, we find eigenvectors by solving $\mathrm{Ker}(A - \lambda I)$ using Gaussian elimination.

Definition 7.1.8. *The **characteristic polynomial** of an $n \times n$ square matrix A is* $\det(xI - A)$*, a polynomial in x of degree n. We usually write this polynomial as $p_A(x)$.*

Note that $\det(xI - A) = (-1)^n \det(A - xI)$. Our ultimate goal is to check if $A - xI$ is invertible or not, so this sign does not matter. We use $\det(xI - A)$ so that x^n will have a positive coefficient. If you choose $\det(A - xI)$ and n is odd, then your x^n will have coefficient -1, which is a bit ugly, but ultimatelty does not matter.

Proposition 7.1.9. *The roots of characteristic polynomial of A are exactly the eigenvalues of A.*

Proof. $\det(xI - A) = 0$ if and only if $A - xI$ is not invertible if and only if $A\boldsymbol{v} = x\boldsymbol{v}$ has a non-zero solution if and only if x is an eigenvalue. □

To illustrate this strategy, let us look at an example of non-diagonalizable matrices.

Example 7.1.10. Consider shearing $E = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. The characteristic polynomial is very simple, it is $(x-1)^2$. So the only eigenvalue is 1.

However, $\mathrm{Ker}(E - I)$ is spanned by $\boldsymbol{e}_1$. So the ONLY eigenvectors of E are multiples of $\boldsymbol{e}_1$.

In this case, eigenvectors of E fails to span $\mathbb{R}^2$. E is NOT diagonalizable. In general, all shearings fail to be diagonalizable. ☺

Now, this strategy works. And almost ALL textbook will teach you to use this strategy. However, this strategy is bogus when $n \geq 5$. Why? Because there is no algebraic formula to solve a polynomial of degree 5 or more. (In practice, a 3×3 matrix will already yields a degree three polynomial, and finding those roots will be a headache already.)

In fact, the reality is the opposite. How would a computer nowaday find roots of a polynomial $p(x)$? The computer would build a matrix whose characteristic polynomial is $p(x)$, then use other methods (variants or improvements of the ***iterated QR algorithm***) to approximate eigenvalues, thus finding approximated roots of $p(x)$. Oh the irony.

Nevertheless, as long as we are dealing with square matrices of dimension at most 4, we can use the traditional strategy as desired. Luckily, most problems we face in this class will have $n \leq 4$.

7.2 Intuitions on Eigenstuff

Sometimes, we can tell the eigenstuff by sight. Here are some examples, which shall further help us understand the intuitions behind the eigenstuff.

Example 7.2.1 (Eigenstuff by sight)**.** Sometimes there are some obvious properties among entries, and they in fact indicates eigenstuff right away.

Consider $A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$. Note how each row adds up to 5? If you apply A to $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, you will exactly add up each row of A and get $\begin{bmatrix} 5 \\ 5 \end{bmatrix}$, so this is an eigenvector for the eigenvalue 5. ☺

Proposition 7.2.2. *If each row of a matrix A adds up to the same number λ, then λ is an eigenvalue of A, and $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ is an eigenvector for this eigenvalue.*

What if each column of A adds up to the same number, say λ? Then we can actually also conclude that A has eigenvalue λ.

Proposition 7.2.3. *A and A^{T} share the same eigenvalues. In fact, they share the same characteristic polynomial.*

Proof. Note that $p_A(x) = \det(xI - A) = \det(xI - A^{\mathrm{T}}) = p_{A^{\mathrm{T}}}(x)$, where the middle equality is true since determinant is invariant under transpose. So A and A^{T} share the same characteristic polynomial, and thus they also share the same eigenvalues. □

Corollary 7.2.4. *If each column of a matrix A adds up to the same number λ, then λ is an eigenvalue of A. (But the eigenvectors are harder to see now.)*

In many applications, such as our dynamical system before, each column of the matrix represents some probability distribution. Therefore, entries in each column will add up to 1. So 1 will always be an eigenvalue for such a matrix. But if the system is evolving via a matrix A and A has eigenvalue 1, then A has a nonzero eigenvector $\boldsymbol{v}$. So $A\boldsymbol{v} = \boldsymbol{v}$, our system must have an equilibrium!

Definition 7.2.5. *A square matrix is a Markov matrix if all entries are positve, and each column adds up to* 1*. (These are exactly the matrices describing some probabilistic evolutions.)*

Proposition 7.2.6. *A Markov matrix must have an eigenvalue 1.*

Proof. Duh. □

For Markov matrices, finding eigenvectors for the eigenvalue 1 is very important. This corresponds to finding the equilibrium of the system. Here is another application of this.

Example 7.2.7. Do you know how google works?

Given a key word, there are many many webpages containing that key word. How can we rank these pages in order, so that the "best" or "most relevant" page comes up on top?

In the pre-google time, search engines do very stupid things, like rank them alphabetically.... And there are also certain search engines that rank pages according to how much these pages pay them, like certain search engine that shall remain unnamed, but you probably have a candidate in mind.

Another way to rank them is to rank according to the number of visits. Surely the most visited website is the "best" result for your search, yes? However, there are some pitfalls. For example, the number of visits are highly volatile. They change every second! It takes a tremendous amount of work and money to keep track of this, and it might also slow down the search process. We want something that is more stable, yet also indicative of the quality of the webpages.

And one day, two young guys named Larry Page and Sergey Brin (the founders of google) came up with a brilliant idea. This idea started google, and its efficiencies, stabilities and speed was far more superior to all other search engines. It immediately beat all competitors and grew into a super company.

So what is this billion dollar idea? This is called the Page rank algorithm. The internet is connected, with webpages linked to one another. So the basic idea is this:

1. If a website is linked to by many many other website, then this website must be important.
2. If some important website links to your website, then this makes your website important.
3. However, if a website is linked to many many many other websites, then it will fail to transfer much importance to each of these individual websites.

Consider the following cases, where we have four websites v_1, v_2, v_3, v_4, v_5, and they are linked to each other as in the graph below. Suppose these websites have importance x_1, x_2, x_3, x_4, x_5. Then v_3 lends weights to both v_1, v_5, so each of v_1, v_5 would get $\frac{1}{2}x_3$ importance from v_3. So on so forth, so we have this graph.

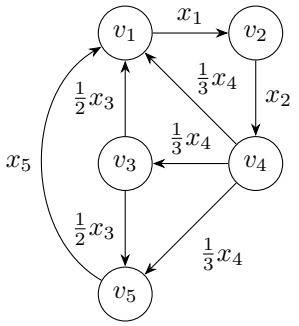


Figure 7.2.1: Graph G of websites

The idea is that your importance equal to the total importance received. For example, the importance of v_1 is x_1, and the total importance received is $\frac{1}{2}x_3 + \frac{1}{3}x_4 + x_5$, so we want $x_1 = \frac{1}{2}x_3 + \frac{1}{3}x_4$. Then collectively, we see that our importance $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_5 \end{bmatrix}$ must be solved by this:

$$\begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{3} & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 \end{bmatrix} \boldsymbol{x} = \boldsymbol{x}.$$

So we are looking for an eigenvector $\boldsymbol{x}$ for the eigenvalue 1 of our matrix. Note that this is a Markov matrix, so such an eigenvector must exist. Then the page rank algorithm would simply list the websites from the most important (largest x_i) to the least important (smallest x_i).

In this cases, one solution is $\boldsymbol{x} = \begin{bmatrix} 6 \\ 6 \\ 2 \\ 6 \\ 3 \end{bmatrix}$. So we put v_1, v_2, v_4 up top in the search result, as they are the most important, then we list webpage v_5, and finally we list webpage v_3, which is the least important.

(As you can probably see from the graph, v_4 is probably some "search engine" website, which is important and has link to almost everyone, but lends little weight to each. v_1 is probably some authority website that

every one links to. And v_2 is some remote website that is hard to reach, but yet very valuable since v_1 only links to v_2. Here v_3, v_5 are not important. v_5 basically just says "yeah just go check out v_1". v_3 is even less useful, as it links to an unhelpful site v_5 other than v_1.) ☺

Above we have focused on the algebra side of the intuition. Now let us turn to geometry. What does it mean geometrically that $A\boldsymbol{x} = \lambda\boldsymbol{x}$? This means the linear transformation of A would FIX this line spanned by $\boldsymbol{x}$.

Example 7.2.8. Suppose all entries of a 2×2 matrix A are positive. I claim that A has a positive eigenvalue whose eigenvector spans a line through the first-third quadrants.

(And A has another eigenvector which spans a line through the second-forth quadrand, with positive eigenvalue if $\det(A) > 0$, and negative eigenvalue if $\det(A) < 0$, and zero eigenvalue if $\det(A) = 0$. I shall not prove these things in the parenthesis though, which is similar but more tedius. Determinant matters here because they indicate whether your map preserves or reverse orientation.)

How can I see this? Let us say $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$.

Image in the domain, we have a ray shooting from the origin in the direction of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Now this ray rotates clockwise, sweeps through the entire first quadrant, and ends up in the direction of $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Now these are happening in the domain of A. If we map these to the codomain, then in the codomain my ray will start in the direction of $A\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, then it will sweep counter-clockwise and end up in the direction of $A\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

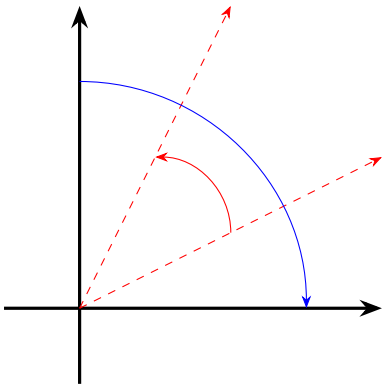


Figure 7.2.2: Sweeping Rays, input = blue and output = red.

Now look at both of these sweeping process in $\mathbb{R}^2$ simultaneously, and we see that somewhere the input ray and the output ray must meet. (This is a variant of the intermediate value theorem in calculus.) This is a line fixed by A, and hence vectors on this line are eigenvectors of A. Since the fixed ray stays in the first quadrant, the eigenvalue must be positive.

Note that $\det(A)$ is negative here. In particular, it is an orientation reversing linear map. As a result, the input clockwise movement becomes a counter-clockwise movement in the output. ☺

I shall mention the following theorem without proof. (We don't use it in this class.) But surely you can imagine how it is true. The proof is basically just a high dimensional version of intermediate value theorem.

Theorem 7.2.9 (Perron-Frobenius Theorem)**.** *If all entries of an $n\times n$ matrix A are positive, then A must*

have a positive eigenvalue $\lambda > 0$*, and an eigenvector* $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ *for* λ *such that all* $x_i > 0$*. (So it is a "positive eigenvector".)*

Finally, let us see some geometric examples.

Example 7.2.10. Consider a reflection in $\mathbb{R}^3$, say about the plane $x+y+z=0$. Note that a normal vector is $\boldsymbol{n} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, and thus this reflection is $A = I - 2\frac{\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}}{\boldsymbol{n}^{\mathrm{T}}\boldsymbol{n}} = I - \frac{2}{3}\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{bmatrix}$.

Now, we can in fact find all eigenvectors using only the geometric intuition, without any computation. Obviously our map A will reflect $\boldsymbol{n}$, so $\boldsymbol{n}$ is an eigenvector for the eigenvalue -1. Furthermore, any non-zero vector on the plane $x+y+z=0$ are preserved, so they are all eigenvectors for the eigenvalue 1. Finally, it is geometrically obvious to see that there are no more fixed lines. All other lines will be reflected to a different line. So we are done.

In particular, pick any two vector $\boldsymbol{v}, \boldsymbol{w}$ on the plane $x+y+z=0$, then $\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{w}$ form a basis of $\mathbb{R}^3$ made of eigenvectors of A. As a result, A is diagonalized under this basis into $\begin{bmatrix} -1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$. ☺

Example 7.2.11. Let us now think about rotations R_θ on $\mathbb{R}^2$. Suppose θ is NOT a multiple of π. Then since every line is rotated, there is no fixed line. In particular, R_θ will have NO eigenvalue.

However, there is no need to despair. Consider the matrix $R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. The characteristic polynomial is x^2+1. So even though it has no REAL root, it has two COMPLEX roots, $\pm\mathrm{i}$.

Consider $R - \mathrm{i}I = \begin{bmatrix} -\mathrm{i} & -1 \\ 1 & -\mathrm{i} \end{bmatrix}$, you can see that the kernel is spanned by $\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}$. Indeed, you can check that $R\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix} = \mathrm{i}\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}$.

Similarly, eigenvalues for $-\mathrm{i}$ is $\begin{bmatrix} 1 \\ \mathrm{i} \end{bmatrix}$. (You can get this by imitating the strategy above, or you can simply take complex conjugates on $R\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix} = \mathrm{i}\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}$.)

Eitherway, $\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}, \begin{bmatrix} 1 \\ \mathrm{i} \end{bmatrix}$ form a basis for $\mathbb{R}^2$. So we have $R = XDX^{-1}$ where $X = \begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix}, D = \begin{bmatrix} \mathrm{i} & \\ & -\mathrm{i} \end{bmatrix}$. We say that R is NOT diagonalizable over $\mathbb{R}$, but IS diagonalizable over $\mathbb{C}$.

It seems that in certain cases, complex matrices are unavoidable. Previously, we have been completely focused on real matrices. But now, we have to deal with complex matrices. ☺

As such, now is a very good time to take a detour to look at complex linear algebra.

7.3 Complex Numbers

As we can see above, even if we only want to study real matrices, complex nubmers would still pop up as eigenvalues and eigenvectors. Here let us quickly consider what is a complex number, and how would complex linear algebra look like. We assume that the readers have a basic familiarty for complex numbers. Nevertheless, we list the basic definition and calculation rules here.

Definition 7.3.1. *The set of complex numbers is the set* $\mathbb{C} = \{a + b\mathrm{i} : a, b \in \mathbb{R}\}$ *where we define complex addition and complex multiplication as*

$$(a+b\mathrm{i}) + (c+d\mathrm{i}) = (a+c) + (b+d)\mathrm{i},$$

$$(a+b\mathrm{i})(c+d\mathrm{i}) = (ac-bd)+(ad+bc)\mathrm{i}.$$

Here $a+c, b+d, ac-bd, ad+bc$ are all calculations of real numbers.

So in short, a complex number is a REAL linear combination of 1 and i, and we require $\mathrm{i}^2 = -1$. You can easily verify that complex addition and multiplication are commutative and associative, and we have law of distribution, and so on. We also have the formula $-(a+b\mathrm{i}) = -a - b\mathrm{i}$, and $(a+b\mathrm{i})^{-1} = \frac{a}{\sqrt{a^2+b^2}} - \frac{b}{\sqrt{a^2+b^2}}\mathrm{i}$. These allow us to do substractions between complex numbers, and divisions when the divider is non-zero. Everything is verifiably nice.

Here are some final bits of definitions.

Definition 7.3.2. *Given a complex number $a+b\mathrm{i}$, we define its real part as $\mathrm{Re}(a+b\mathrm{i}) = a$, and its imaginary part as $\mathrm{Im}(a+b\mathrm{i}) = b$. (Note that traditionally we use b, not $b\mathrm{i}$.) We also define its absolute value (or modulus or norm or length or magnitude.... it's alll the same to me) as $|a+b\mathrm{i}| = \sqrt{a^2+b^2}$, and its complex conjugate is $\overline{a+b\mathrm{i}} = a - b\mathrm{i}$.*

Now, this traditional definition of complex numbers seems artificial and a bit surreal. What does i even mean? How would such a system exist? Let us observe some interesting comparisons.

	Complex Numbers	Special 2×2 matrices
Elements	$a+b\mathrm{i}$	$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$
Addition	$(a+b\mathrm{i})+(c+d\mathrm{i}) = (a+c)+(b+d)\mathrm{i}$	$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} + \begin{bmatrix} c & -d \\ d & c \end{bmatrix} = \begin{bmatrix} a+c & -b-d \\ b+d & a+c \end{bmatrix}$
Multiplication	$(a+b\mathrm{i})(c+d\mathrm{i}) = (ac-bd)+(ad+bc)\mathrm{i}$	$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}\begin{bmatrix} c & -d \\ d & c \end{bmatrix} = \begin{bmatrix} ac-bd & -ad-bc \\ ad+bc & ac-bd \end{bmatrix}$
Real part	$\mathrm{Re}(a+b\mathrm{i}) = a$	$\frac{1}{2}\operatorname{trace}\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = a$
Absolute value	$\lvert a+b\mathrm{i}\rvert^2 = a^2+b^2$	$\det\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = a^2+b^2$
Complex conjugate	$\overline{a+b\mathrm{i}} = a-b\mathrm{i}$	$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$
Motivation	$\mathrm{i}^2 = -1$	$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^2 = -I.$

The point is this: even though complex numbers might look "unreal", they are actually part of our everyday life. You can simply think of 1 as the identity map I, and think of i as rotation counter-clockwise by 90 degree, then a complex number is simply a linear combination of the two. As far as calculations go, $a+b\mathrm{i}$ is usually easier to work with. But as far as interpretations go, $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ tells you the meaning behind this complex number.

So as we can see, a complex number is basically a linear combination between the identity map and the "rotation by 90" map. In fact, all rotations of $\mathbb{R}^2$ are closely related to complex numbers.

Example 7.3.3. Consider $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. Clearly this corresponds to $\cos\theta + \mathrm{i}\sin\theta$. You can also see that these are precisely the ***unit complex numbers***, i.e., complex numbers with absolute value one. It is in general a very good intuition to think of unit complex numbers as representations of rotations.

Consider $kI = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$. This corresponds to the complex number k, which is purely real. So a purely real number as a complex number would represent the "stretch everything" map, where we stretch everything by a factor of k.

Now consider a generic complex number $a+b\mathrm{i}$. The corresponding linear map is $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Now both columns have the same length $\sqrt{a^2+b^2}$. If we divide by this length, then $\begin{bmatrix} \frac{a}{\sqrt{a^2+b^2}} \\ \frac{b}{\sqrt{a^2+b^2}} \end{bmatrix}$ would be a unit vector. Set $\frac{a}{\sqrt{a^2+b^2}} = \cos\theta$ and $\frac{b}{\sqrt{a^2+b^2}} = \sin\theta$ for some theta, we see that $\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} \sqrt{a^2+b^2} & \\ & \sqrt{a^2+b^2} \end{bmatrix} R_\theta$.

In short, a complex number $a+b\mathrm{i}$ as a linear map would mean a rotation map then a "stretch all" map. We have $a+b\mathrm{i} = \sqrt{a^2+b^2}(\cos\theta + \mathrm{i}\sin\theta)$, where the rotation amount θ is determined by the direction of $\begin{bmatrix} a \\ b \end{bmatrix}$, and the stretching factor is determined by the absolute value. ☺

The decomposition $a+b\mathrm{i} = \sqrt{a^2+b^2}(\cos\theta+\mathrm{i}\sin\theta)$ is the famous polar decomposition of complex numbers. It corresponds to decomposing the linear map $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ into a stretching and a rotation. To better write this, we need some extra notational convention.

Definition 7.3.4. *For a complex number z, we define* $\mathrm{e}^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$.

In general, for a square matrix A, we define $\mathrm{e}^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots$.

We shall use the fact that e^z and e^A always converge, without any proof. The proof is pure analysis and has nothing to do with linear algebra.

Lemma 7.3.5. $\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, *and* $\mathrm{e}^{\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}} = R_\theta$. *For any two complex nubmers $z, w \in \mathbb{C}$, we have* $\mathrm{e}^{z+w} = \mathrm{e}^z\mathrm{e}^w$.

Proof. Direct calculations. Recall that $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$, and $\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$

Now consider a matrix $A = \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$. Note that this is a skew-symmetric matrix. We now have

$$\begin{aligned}
\mathrm{e}^A &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} + \frac{1}{2!}\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}^2 + \frac{1}{3!}\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}^3 + \dots \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} + \begin{bmatrix} -\frac{a^2}{2!} & 0 \\ 0 & -\frac{a^2}{2!} \end{bmatrix} + \begin{bmatrix} 0 & \frac{a^3}{3!} \\ -\frac{a^3}{3!} & 0 \end{bmatrix}^3 + \dots \\
&= \begin{bmatrix} 1 - \frac{a^2}{2!} + \dots & -a + \frac{a^3}{3!} - \dots \\ a - \frac{a^3}{3!} + \dots & 1 - \frac{a^2}{2!} + \dots \end{bmatrix} \\
&= \begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}.
\end{aligned}$$

□

Corollary 7.3.6 (Euler identity)**.** $\mathrm{e}^{\mathrm{i}\pi} = -1$.

As far as I'm concerned, I think the Euler identity is neither surprising nor mysterious. It simply means that on $\mathbb{R}^2$, rotating things by 180 degree is the same as negating everything.

Corollary 7.3.7 (Polar decomposition)**.** *For any complex number z, there are unique real numbers $r > 0$ and $0 \le \theta < 2\pi$, such that $z = r\mathrm{e}^{\mathrm{i}\theta}$.*

Remark 7.3.8. *If we think of complex numbers as 2 by 2 matrices as before, then $z = \mathrm{e}^{\mathrm{i}\theta}r$ in fact corresponds to a QR decomposition.*

Now we have understood almost everything about complex numbers. The next property, however, is the most important. It is why we bother.

Theorem 7.3.9 (Fundamental Theorem of Algebra)**.** *Any complex polynomial $p(x)$ of degree n has n roots counting multiplicities. (E.g., we say $(x-a)^3(x-b)^2$ has a root a with multiplicity 3, and a root b with multiplicity 2. So counting multiplicity, the roots are a, a, a, b, b, five in total.)*

Given an $n \times n$ matrix, its characteristic polynomial is some degree n polynomial. The roots of this polynomial are exactly the eigenvalues of your matrix. However, the fundamental theorem of algebra would then reveal that there are exactly n roots (counting multiplicities). Hence an $n \times n$ matrix would always have exactly n eigenvalues (counting algebraic multiplicities, which we shall define later).

Let us see some examples of this. For $n = 2$, we shall study some matrices and see that they all have two (potentially complex) eigenvalues. Further, the complex eigenvalues are indicative of what your matrix is trying to do.

Example 7.3.10. Consider $A = \begin{bmatrix} 3 & 3 \\ 1 & 5 \end{bmatrix}$. Then the characteristic polynomial is $x^2 - 8x + 12$. The roots are exactly 2 and 6, so these are your eigenvalues. $\operatorname{Ker}(A - 2I) = \operatorname{Ker} \begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix} = \operatorname{span}(\begin{bmatrix} 3 \\ -1 \end{bmatrix})$, and $\operatorname{Ker}(A - 6I)$ is spanned by $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. (Note that each row adds up to six.)

In short, we have a diagonalization $A = \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & \\ & 6 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix}^{-1}$.

Note that both eigenvalues are purely real. What does real complex numbers do? They stretch. Our matrix indeed stretch in the $\begin{bmatrix} 3 \\ -1 \end{bmatrix}$ direction by a factor of 2, and stretch in the $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ direction by a factor of 6. At this stage, you should be able to visualize the geometric action of this linear map on $\mathbb{R}^2$. ☺

Example 7.3.11. Consider $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. Then the characteristic polynomial is $(x - \cos\theta)^2 + \sin^2\theta$. The roots are exactly $\cos\theta \pm \mathrm{i}\sin\theta$, or $\mathrm{e}^{\pm \mathrm{i}\theta}$. As you can see, the matrix is doing rotation, and its eigenvalues are unit complex numbers, which also represent rotations.

Diagonalization here is $R_\theta = \begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\mathrm{i}\theta} & \\ & \mathrm{e}^{-\mathrm{i}\theta} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix}^{-1}$. ☺

7.4 Complex Linear Algebra

Now moving forward, we shall need to do a lot of complex matrix stuff. In general, most of the things are exactly like before. The most important one is matrix multiplication.

Example 7.4.1. As a simple example, we have $\begin{bmatrix} 1 & \mathrm{i} \\ 1+\mathrm{i} & 1-\mathrm{i} \end{bmatrix} \begin{bmatrix} \mathrm{i} \\ 2+\mathrm{i} \end{bmatrix} = \begin{bmatrix} 1 \times \mathrm{i} + \mathrm{i} \times (2+\mathrm{i}) \\ (1+\mathrm{i}) \times \mathrm{i} + (1-\mathrm{i}) \times (2+\mathrm{i}) \end{bmatrix} = \begin{bmatrix} -1+3\mathrm{i} \\ 2 \end{bmatrix}$. As you can see, such things are exactly as before.

As an optional side note, just like we can think of $a + b\mathrm{i}$ as the real matrix $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, we can similarly think of complex vectors or matrices as "block matrices" with real entries. For example, the same equation above may be written as

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

And you can see that everything works just the same.

In particular, if you seek geometric intuitions on complex vectors, you can think of each complex vector with n coordinates as "a pair of real vectors with $2n$ coordinates". ☺

While doing these, keep in mind of the following: if someting requires only linear structure, but NO inner product strucutre, then all formulas are exactly the same as the real case. These includes:

1. Matrix multiplications.
2. Matrix inversions.
3. Gaussian Elimination.
4. Rank-Nullity Theorem. ($\dim \operatorname{Ran}(L) + \dim \operatorname{Ker}(L) = \dim \operatorname{dom}(L)$.)
5. Trace and determinant formula. (And anything that is similarity-invariant, i.e., invariant under an arbitrary change of basis. Here for example we have $\operatorname{trace}(XAX^{-1}) = \operatorname{trace}(A), \det(XAX^{-1}) = \det(A)$.)
6. Eigenstuff, characteristic polynomial, diagonalizability, etc. (All are similarity-invariant.)

Example 7.4.2. Keep in mind that for trace and determinant, the block matrix analogy does not work, even though all the formula are the same.

For example, consider $A = \begin{bmatrix} 1 & \mathrm{i} \\ 1+\mathrm{i} & 1-\mathrm{i} \end{bmatrix}$. Then the trace is $2-\mathrm{i}$. However, the corresponding real block matrix $\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}$ cannot have a complex trace, since all entries are real. The trace of this real block matrix is in fact always $2\operatorname{Re}(\operatorname{trace}(A))$.

Similarly, $\det(A) = 1(1-\mathrm{i}) - \mathrm{i}(1+\mathrm{i}) = 2 - 2\mathrm{i}$, while the corresponding real block matrix can only have real determinant since all entries are real. In fact, the real block matrix should always have a determinant of $|\det(A)|^2$.

I shall leave the verification of these correspondences to you. ☺

However, the inner product structure (and therefore transpose) is now vastly different. For the matrix $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, taking transpose would send it to $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$. This corresponds to sending the complex number $a + b\mathrm{i}$ to its complex conjugate $a - b\mathrm{i}$.

Example 7.4.3. Consider the complex matrix $A = \begin{bmatrix} 1 & \mathrm{i} \\ 1+\mathrm{i} & 1-\mathrm{i} \end{bmatrix}$. Its corresponding real block matrix is $\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}$.

Now if we take transpose on the real matrix, we would get $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 \\ 0 & 1 & 1 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix}$. This corresponds to the complex matrix $\begin{bmatrix} 1 & 1-\mathrm{i} \\ -\mathrm{i} & 1+\mathrm{i} \end{bmatrix}$, which is the transpose then complex conjugate of A.

In summary, when we do transpose to real matrices, the analogous operation on complex matrices should be transpose conjugate. ☺

Definition 7.4.4. *Given a complex matrix A, we define its **adjoint** to be $A^* = \overline{A^{\mathrm{T}}}$, where the bar indicates that we are taking complex conjugates on each entry.*

Definition 7.4.5. *Given $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^n$, we define the dot product between them to be $\boldsymbol{v}^* \boldsymbol{w}$.*

Example 7.4.6. Consider $\boldsymbol{v} = \begin{bmatrix} 1 \\ \mathrm{i} \end{bmatrix}$. Then $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{v} = 1 + (-1) = 0$. Clearly this should NOT be the length squared of the vector!

If we have $\boldsymbol{v} = \begin{bmatrix} z \\ w \end{bmatrix}$, then in fact $\boldsymbol{v}^* \boldsymbol{v} = \begin{bmatrix} \overline{z} & \overline{w} \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix} = \overline{z} z + \overline{w} w = |z|^2 + |w|^2$. This now makes much more sense to be the length squared of the vector. ☺

Example 7.4.7. In general, what does it mean for two complex vectors to be perpendicular?

Suppose $\boldsymbol{v} = \begin{bmatrix} 1 \\ \mathrm{i} \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}$. You can easily calculate and see that $\boldsymbol{v}^* \boldsymbol{w} = 0$. So these two complex vectors are perpendicular.

In real block form, $\boldsymbol{v}$ corresponds to $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 1 & 0 \end{bmatrix}$ while $\boldsymbol{w}$ corresponds to $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}$. As you can see, ALL FOUR columns are mutually orthogonal. This is what it means to be complex orthogonal. ☺

There are other analogous thing to do. Consider these definitions.

Definition 7.4.8.

1. *A complex matrix A is unitary if $A^* = A^{-1}$. (This is analogous to real orthogonal matrices.)*
2. *A complex matrix A is Hermitian (or self-adjoint as physicists prefer) if $A^* = A$. (This is analogous to real symmetric matrices.)*
3. *A complex matrix A is skew-Hermitian (or skew-adjoint) if $A^* = -A$. (This is analogous to real skew-symmetric matrices.)*

We also have this analogous theorem.

Theorem 7.4.9. *If A is an $m \times n$ complex matrix. Then* $\mathrm{Ran}(A)^{\perp} = \mathrm{Ker}(A^*)$ *and* $\mathrm{Ker}(A)^{\perp} = \mathrm{Ran}(A^*)$. *Here orthogonality is in the sense of complex dot product.*

7.5 (Optional) Fundamental Theorem of Algebra

Here I present my personal favorite proof, using topology.

Consider the polynomial $p(z) = z^n$. Then $p : \mathbb{C} \to \mathbb{C}$ is a (non-linear) map from the plane to the plane. What is its behavior?

Intuitively, if z represents some rotation, then z^n simply rotate n times the original amount. This is the guiding intuition for all the analysis below.

Suppose in the domain, we are looking at a big circle of radius R, centered at the origin. And we let a point z walk around this circle ONCE, counter-clockwise. What would happen to $p(z)$ in the codomain?

Let $z = R\mathrm{e}^{\mathrm{i}\theta}$, and we are increaing θ from 0 to 2π. Then $p(z) = R^n \mathrm{e}^{\mathrm{i}(n\theta)}$, where θ goes from 0 to 2π. It is not hard to see that $p(z)$ in the codomain would walk around a big circle of radius R^n, centered at the origin, n times! $p(z)$ is moving much faster than z, with n times the angular speed of z.

Now consider a generic polynomial $p(z) = z^n +$lower degree terms. Here we assume the leading coefficient is 1 for simplicity. Now, if I pick a super super large R, then for any z on the big circle of radius R, centered at the origin, it will have a super super large absolute value $|z| = R$. In particular, $|z^n| = R^n$ will be much much larger than any lower degree terms. We would effectively have $p(z) \approx z^n$ for all these z on the big circle of radius R, centered at the origin.

So this big circle of radius R in the domain is approximately mapped to some big circle of radius R^n around the origin n-times. Since this is just an approximation, the actual image of the circle will have some minor perturbations caused by the lower degree terms. But they should be minor if R is super super large.

All in all, if in the domain we pick the big circle of radius R, then its image is some curve winding around the origin n-times, approximately a big circle of radius R^n.

Now, let us gradually shrink R towards the origin. As R is shrinked to the origin, its image in the codomain would shrink towards a single point $p(0)$. Since its image is a curve winding around the origin n-times, when it shrinks to a single point, it shall SWEEP THROUGH the origin n-times.

So, inside the disc of radius R around the origin in the domain, we have n inputs z with $p(z) = 0$.

7.6 Algebraic multiplicity and Schur Decomposition

Important: Without specification, everything in this section is over $\mathbb{C}$. Entries are allowed to be complex, coordinates are allowed to be complex, and so on, unless specifically mentioned.

Just like polynomials can have repeated roots, matrices can have repeated eigenvalues. We use the name multiplicity to refer to the number of times they are repeated. Let us first study algebraic multiplicity.

Given an eigenvalue λ of a matrix A, whose characteristic polynomial is $p_A(x)$, then we know λ is a root of p_A. However, roots of a polynomial have multiplicities. For example, $(x-1)^3(x-2)^4$ has seven roots, 1 with multiplicity 3 and 2 with multiplicity 4.

Definition 7.6.1. *The algebraic multiplicity of an eigenvalue λ of A is the multiplicity of λ as roots of the characteristic polynomial $p_A(x)$. (Later we shall have another concept called geometric multiplicity, which shall be different.)*

Example 7.6.2. Consider $A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 2 \end{bmatrix}$. Clearly $\det(xI - A) = (x-1)^3(x-2)$.

We say A has eigenvalues $1, 1, 1, 2$ counting algebraic multiplicity. We also say that A has eigenvalues $1, 2$ NOT counting algebraic multiplicity. ☺

Proposition 7.6.3. *If A is $n \times n$, then it has n eigenvalues (counting algebraic multiplicities). Equivalently, if $\lambda_1, \ldots, \lambda_k$ are eigenvalues of A NOT counting algebraic multiplicity, then the sum of algebraic multiplicities is $\sum m_a(\lambda_i) = n$.*

Furthermore, if these eigenvalues are $\lambda_1, \ldots, \lambda_n$ counting algebraic multiplicity, then $p_A(x) = \prod(x - \lambda_i)$.

Proof. This is simply the fundamental theorem of algebra. Each polynomial has at most n roots. □

The concept of multiplicities allows us to look at the eigenstuff of a matrix globally, and in fact would produce some amazing results.

Proposition 7.6.4. *Given an $n \times n$ matrix A, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues (counting algebraic multiplicity). Then $\operatorname{trace}(A) = \sum \lambda_i$ and $\det(A) = \prod \lambda_i$.*

Let us see the 2×2 case before the formal proof.

Example 7.6.5. Consider $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then $xI - A = \begin{bmatrix} x-a & -b \\ -c & -d \end{bmatrix}$. So the characteristic polynomial is $(x-a)(x-d) - bc = x^2 - (a+d)x + ad - bc = x^2 - \operatorname{trace}(A)x + \det(A)$. On the other hand, if the eigenvalues are λ_1, λ_2, then the characteristic polynomial should also be $(x - \lambda_1)(x - \lambda_2) = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2$. Compare the two, and we are done.

Note that this is essentially Vieta's formula, which relates the roots of a polynomial (eigenvalues of your matrix) with the coefficients of a polynomial (calculated from entries of your matrix). ☺

Proof. Say $A = \begin{bmatrix} \boldsymbol{a}_1 & \ldots & \boldsymbol{a}_n \end{bmatrix}$. Then $\det(xI - A) = \det(x\boldsymbol{e}_1 - \boldsymbol{a}_1, \ldots, x\boldsymbol{e}_n - \boldsymbol{a}_n)$. Now we expand everything using multilinearity.

Note that to get contribution to the term x^{n-1}, we must pick all but one of the $x\boldsymbol{e}_i$. As a result, in the expansion, the coefficient for x^{n-1} is $\det(-\boldsymbol{a}_1, x\boldsymbol{e}_2, \ldots, x\boldsymbol{e}_n) + \cdots + \det(x\boldsymbol{e}_1, \ldots, x\boldsymbol{e}_{n-1}, -\boldsymbol{a}_n) = -a_{11} - \cdots - a_{nn} = -\operatorname{trace}(A)$. However, this is also $-\sum \lambda_i$ by Vieta's formula. So we are done.

Similarly, consider the constant term in the expansion of $\det(xI - A) = \det(x\boldsymbol{e}_1 - \boldsymbol{a}_1, \ldots, x\boldsymbol{e}_n - \boldsymbol{a}_n)$. Then we must avoid ALL $x\boldsymbol{e}_i$. So we have constant term $\det(-\boldsymbol{a}_1, \ldots, -\boldsymbol{a}_n) = (-1)^n \det(A)$. However, this is also $(-1)^n \prod \lambda_i$ by Vieta's formula. So we are done. □

Alternative proof. These proofs are just for fun. Suppose $p_A(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$.

Since $p_A(x) = \det(xI - A)$, therefore $p_A(0) = \det(-A) = (-1)^n \det(A)$. On the other hand, $p_A(0) = a_0 = (-1)^n \prod \lambda_i$ by Vieta's formula. Hence we are done.

Now consider $f(t) = \det(I + tA)$. Then $f(-\frac{1}{x}) = \det(I - \frac{1}{x}A) = \frac{1}{x^n}\det(xI - A) = \frac{1}{x^n}p_A(x) = 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}$. Therefore, by setting $t = -\frac{1}{x}$, we see that $f(t) = 1 - a_{n-1}t + a_{n-2}t^2 - \cdots + (-1)^n a_0 t^n$. So $f'(0) = -a_{n-1} = \sum \lambda_i$ by Vieta's formula. On the other hand, we know $f'(0) = \operatorname{trace}(A)$. So we are done. □

Example 7.6.6. Consider $\begin{bmatrix} 2 & -1 \\ -2 & 3 \end{bmatrix}$. Since each row adds up to 1, we see that 1 is an eigenvalue. Since the trace is 5, we see that the other eigenvalue is 4. Done. Such techniques are very helpful to speed calculate the eigenvalues of small matrices. ☺

Here is also a handy corollary.

Corollary 7.6.7. *A is invertible iff it has no zero eigenvalue.*

Proof. A is invertible iff $\det(A) \neq 0$ iff $\prod \lambda_i \neq 0$ iff all $\lambda_i \neq 0$. □

In fact, by similar methodology (but more complicated computations), we can prove the following. I obmit the proof here since the computations are a bit too messy.

Definition 7.6.8. *A k principal submatrix of A is any $k \times k$ submatrix whose diagonal is on the diagonal of A. (Note that a submatrix does NOT have to be a block. For example, $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ has three principal submatrices $\begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$, $\begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}$ and $\begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}$, whose diagonal is indeed on the diagonal of the original matrix.)*

Theorem 7.6.9. *The coefficient for x^{n-k} in $p_A(x)$ is exactly $(-1)^k S_k$, where S_k is the sum of determinants of all k principal submatrices of A.*

By generalized Vieta's formula, this is also $(-1)^k \sum_{1 \leq i_1 < \cdots < i_k \leq n} \lambda_{i_1} \ldots \lambda_{i_k}$, sum of all possible products of k of the roots.

So in particular, we have $S_k = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \lambda_{i_1} \ldots \lambda_{i_k}$.

Remark 7.6.10. *Given an $n \times n$ matrix A, then its only $n \times n$ principal submatrix is A itself. So $S_n = \det(A)$.*

Its 1×1 principal submatrices are the diagonal entries. So $S_1 = \operatorname{trace}(A)$.

You can see that the above theorem is a generalization of the result we have proven.

Let us see some examples both utilizing $\operatorname{trace}(A) = \sum \lambda_i$ and $\det(A) = \prod \lambda_i$, and some examples of the weird theorem above.

Example 7.6.11. Consider $A = \begin{bmatrix} -1 & 1 & 1 \\ -4 & 3 & 2 \\ -4 & 1 & 4 \end{bmatrix}$. What are the eigenvalues? How to diagonalize this matrix?

Well, note that each row adds up to one. So one eigenvalue is 1. We also have $\operatorname{trace}(A) = 6$ and $\det(A) = 6$.

Suppose the eigenvalues are $1, \lambda_2, \lambda_3$, then $\lambda_2 + \lambda_3 = 5$ and $\lambda_2\lambda_3 = 6$. We see that the two unknown eigenvalues must be 2 and 3.

Calculate $\mathrm{Ker}(A-\lambda I)$ with $\lambda = 1,2,3$ via Guassian elimination, we have eigenvector $\begin{bmatrix}1\\1\\1\end{bmatrix}$ for the eigenvalue 1, eigenvector $\begin{bmatrix}1\\2\\1\end{bmatrix}$ for the eigenvalue 2, and eigenvector $\begin{bmatrix}1\\2\\2\end{bmatrix}$ for the eigenvalue 3. These three vectors form a basis, so we have found our diagonalization:

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

Now, let us consider the 2×2 principal submatrices. We have

$$S_2 = \det\begin{bmatrix}-1&1\\-4&3\end{bmatrix} + \det\begin{bmatrix}-1&1\\-4&4\end{bmatrix} + \det\begin{bmatrix}3&2\\1&4\end{bmatrix} = 1+0+10 = 11.$$

On the other hand, we have

$$\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 = 1\times 2 + 2\times 3 + 1\times 3 = 11.$$

Indeed, the two are the same. In fact, you can verify that the characteristic polynomial is indeed $p_A(x) = x^3 - 6x^2 + 11x - 6$. The value of S_1 (trace), S_2, S_3 (determinant) corresponds to the three coefficients. ☺

Example 7.6.12. For triangular matrices, this correspondence is even better.

Consider $A = \begin{bmatrix}1&1&1\\0&2&1\\0&0&3\end{bmatrix}$. What are the eigenvalues? How to diagonalize this matrix?

Well, note that each row adds up to three. So one eigenvalue is 3. The first column indicates that $A\boldsymbol{e}_1 = \boldsymbol{e}_1$, so one eigenvalue must be 1. Finally, the trace is 6, so the last eigenvalue must be $6-1-3=2$.

Calculate $\mathrm{Ker}(A-\lambda I)$ with $\lambda = 1,2,3$ via Guassian elimination, we have eigenvector $\begin{bmatrix}1\\0\\0\end{bmatrix}$ for the eigenvalue 1, eigenvector $\begin{bmatrix}1\\1\\0\end{bmatrix}$ for the eigenvalue 2, and eigenvector $\begin{bmatrix}1\\1\\1\end{bmatrix}$ for the eigenvalue 3. These three vectors form a basis, so we have found our diagonalization:

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

Now, let us consider the 2×2 principal submatrices. We have

$$S_2 = \det\begin{bmatrix}1&1\\&2\end{bmatrix} + \det\begin{bmatrix}2&1\\&3\end{bmatrix} + \det\begin{bmatrix}1&1\\&3\end{bmatrix} = 1\times 2 + 2\times 3 + 1\times 3 = 11.$$

On the other hand, we have

$$\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 = 1\times 2 + 2\times 3 + 1\times 3 = 11.$$

Indeed, the two are the same. You can even see that the three principal submatrices in this case corresponds perfectly with pairs of eigenvalues! This is an amazing thing that should always happen to triangular matrices. ☺

Note that the key here lies on the structure of triangular matrices. Let us extrapolate the result here:

Proposition 7.6.13. *If* $A = \begin{bmatrix} B & \star \\ O & C \end{bmatrix}$, *then the characteristic polynomials have the relation* $p_A(x) = p_B(x)p_C(x)$.

Proof. This is a direct corollary of the similar results on the determinants of block triangular matrices. □

Corollary 7.6.14. *If* A *is triangular, then its eigenvalues (counting algebraic multiplicity) are exactly the diagonal entries.*

As a result, for a triangular matrix such as $A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}$, it is perfectly clear now why its 2×2 principal submatrices are in perfect correspondence with pairs of eigenvalues: because each 2×2 principal submatrix takes a pair of diagonal entries!

At this stage, it is very clear that triangular matrices are super amazing. Their determinant is easy, their eigenvalues are easy, everything is easy. Don't you wish that all matrices are triangular? Well.... Wish granted!

Theorem 7.6.15 (Schur Decomposition)**.** *For any square matrix* A*, let its eigenvalues be* $\lambda_1, \dots, \lambda_n$ *counting algebraic multiplicity, in any prescribed order. Then one can find an invertible matrix* X *such that* $A = XTX^{-1}$ *where* T *is upper triangular, and the diagonal entries of* T *are exactly* $\lambda_1, \dots, \lambda_n$ *in the desired order.*

Furthermore, we can take X *to be a unitary matrix.*

In short, for any matrix, you can change basis and make it triangular. Furthermore, if needed, you can restrict it to an orthonormal change of basis!

Proof. First pick any eigenvector for the eigenvalue λ_1, say $\boldsymbol{x}_1 \neq \mathbf{0}$. WLOG, we can pick $\boldsymbol{x}_1$ to be a unit vector.

Extend $\boldsymbol{x}_1$ into a basis (or orthonormal basis), then these vectors as columns would form a matrix X_1, where the first column of X_1 is $\boldsymbol{x}_1$. (Note that X_1 will be a unitary matrix as well.)

Now consider a change of basis to this new basis X_1, resulting in $X_1^{-1}AX_1$. Note that since $A\boldsymbol{x}_1 = \lambda_1 \boldsymbol{x}_1$, the first basis vector is sent to λ_1 times itself. Therefore, after this change of basis, $X_1^{-1}AX_1 = \begin{bmatrix} \lambda_1 & \star \\ \mathbf{0} & A_{n-1} \end{bmatrix}$ for some $(n-1) \times (n-1)$ matrix A_{n-1}.

(Equivalently, observe that $X_1^{-1}AX_1 = X_1^{-1}A \begin{bmatrix} \boldsymbol{x}_1 & \star \end{bmatrix} = X_1^{-1} \begin{bmatrix} \lambda_1 \boldsymbol{x}_1 & \star \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{e}_1 & \star \end{bmatrix}$.)

Either way, now apply mathematical induction on the dimension n of A. The $n = 1$ case is trivial. For generic n, we can assume that the $n-1$ case is already done. Therefore, we can assume that $A_{n-1} = X_{n-1}T_{n-1}X_{n-1}^{-1}$ for some unitary X_{n-1}. Furthermore, since the eigenvalues of A_{n-1} are exactly $\lambda_2, \dots, \lambda_n$, we can make sure that the diagonal entries of T_{n-1} are exactly these values in the desired order by induction hypothesis.

So we have $A = X_1 \begin{bmatrix} 1 & \\ & X_{n-1} \end{bmatrix} \begin{bmatrix} \lambda_1 & \star \\ \mathbf{0} & T_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \\ & X_{n-1}^{-1} \end{bmatrix} X_1^{-1}$. So we are done. □

Remark 7.6.16. *Note that for most of our purpose, it is enough to have* $A = XTX^{-1}$ *for any invertible* X*. The fact that we can furthermore choose* X *to be unitary is not important for our class at the moment. However, in terms of computation stability, this is HUGE. It means this factorization* $A = XTX^{-1}$ *will NOT magnify errors.*

Currently, if you ask a computer to find roots of a polynomial $p(x)$*, it will attempt to first build a matrix* A *whose characteristic polynomial is* $p(x)$*, and then perform the Schur decomposition* $A = XTX^{-1}$*, and then read the diagonal entries of* T*.*

Both diagonalization and triangularization tries to do the same thing: we attempt to find a basis, so that after the change of basis, our matrix looks prettier (and therefore easier to compute with). In this sense, diagonalizations are the better version. Triangular matrices are of course much uglier than diagonal matrices.

However, the advantage is two-fold. First, for diagonalization $A = XDX^{-1}$, the matrix X might not be unitary, and error terms might be magnified severly. For triangularization, this is not a problem since X is unitary. Second, some matrices do NOT have diagonalizations. But ALL matrices have triangularizations. The universality is in itself a huge advantage already.

Example 7.6.17. Suppose A has eigenvalues $\lambda_1, \ldots, \lambda_n$ counting algebraic multiplicity, then $A = XTX^{-1}$ where $T = \begin{bmatrix} \lambda_1 & * & * \\ & \ddots & * \\ & & \lambda_n \end{bmatrix}$.

Then $A^2 = XTX^{-1}XTX^{-1} = XT^2X^{-1}$, where $T^2 = \begin{bmatrix} \lambda_1^2 & * & * \\ & \ddots & * \\ & & \lambda_n^2 \end{bmatrix}$ where the star portion might change in ugly manner, but the diagonal portion is still nice. In particular, we see that A^2 has eigenvalues $\lambda_1^2, \ldots, \lambda_n^2$ counting algebraic multiplicity. You can imagine that, for A^k, the eigenvalues would be $\lambda_1^k, \ldots, \lambda_n^k$ counting algebraic multiplicity.

Now consider $A^2 + A$, which is a polynomial in A. Since $A = XTX^{-1}$ and $A^2 = XT^2X^{-1}$, we can see that $A^2 + A = X(T^2 + T)X^{-1}$ where $T^2 + T = \begin{bmatrix} \lambda_1^2 + \lambda_1 & * & * \\ & \ddots & * \\ & & \lambda_n^2 + \lambda_n \end{bmatrix}$. So the eigenvalues for $A^2 + A$ are $\lambda_1^2 + \lambda_1, \ldots, \lambda_n^2 + \lambda_n$ counting algebraic multiplicity. ☺

Proposition 7.6.18. *If A has eigenvalues $\lambda_1, \ldots, \lambda_n$, then for any polynomial $p(x)$, the matrix $p(A)$ has eigenvalues $p(\lambda_1), \ldots, p(\lambda_n)$. (For example, $A+kI$ would have eigenvalues $\lambda_1+k, \ldots, \lambda_n+k$. And A^2 would have eigenvalues $\lambda_1^2, \ldots, \lambda_n^2$.)*

Proof. Write $A = XTX^{-1}$. Then $p(A) = Xp(T)X^{-1}$. Since T is triangular with diagonal entries $\lambda_1, \ldots, \lambda_n$, we see that $p(T)$ is still triangular with diagonal entries $p(\lambda_1), \ldots, p(\lambda_n)$. So we are done. □

Here is an optional alternative proof without use Schur decomposition. It is a purely algebraic manipulation of the characteristic polynomial.

Optional Proof. Warning: This proof is ugly and not that illuminating. But serious math lovers should know that this is possible.

It is very easy to see that $p(\lambda_i)$ is indeed an eigenvalue for $p(A)$. If $A\boldsymbol{v} = \lambda_i\boldsymbol{v}$, then it is easy to verify that $A^k\boldsymbol{v} = \lambda_i^k\boldsymbol{v}$. Since $p(A)$ is a linear combination of powers of A, we see that $p(A)\boldsymbol{v} = p(\lambda_i)\boldsymbol{v}$. The difficulty here lies in the algebraic multiplicity. To show that, one has to attack the characteristic polynomial.

Consider $\det(\lambda I - p(A))$ for some fixed λ. What is $\lambda I - p(A)$? This is simply another polynomial of A! This is the most vital observation. Let $q(t) = \lambda - p(t)$. By the fundamental theorem of linear algebra, we have $q(t) = k\prod(t - t_i)$ for some complex values $k, t_1, \ldots, t_n$. So we have

$$\det(\lambda I - p(A)) = \det(q(A)) = k^n \prod \det(A - t_i I) = (-k)^n \prod_i p_A(t_i).$$

Now use the fact that $p_A(x) = \prod(x - \lambda_j)$, we see that

$$(-k)^n \prod_i p_A(t_i) = (-k)^n \prod_{i,j}(t_i - \lambda_j) = \prod_j (k \prod_i (\lambda_j - t_i)) = \prod_j q(\lambda_j) = \prod_j (\lambda - p(\lambda_j)).$$

This is true for all λ. So $\det(\lambda I - p(A)) = \prod_j(\lambda - p(\lambda_j))$. In particular, the characteristic polynomial has roots $p(\lambda_j)$ as desired. □

By similar arguments (either by Schur decomposition or by characteristic polynomial manipulation), one can see the following:

Proposition 7.6.19. *If A has eigenvalues $\lambda_1, \dots, \lambda_n$, then A^{T} has the same set of eigenvalues. And if A is invertible, then A^{-1} has eigenvalues $\lambda_1^{-1}, \dots, \lambda_n^{-1}$.*

Example 7.6.20. The fact that $p(A)$ has eigenvalues $p(\lambda)$ is very convenient when analyzing eigenvalues of A.

For example, suppose $A^2 = A$, i.e., the (oblique) projection matrices. What are the possible eigenvalues of such a matrix?

We have $A^2 - A = 0$. Therefore, if λ is an eigenvalue of A, then $\lambda^2 - \lambda$ must be an eigenvalue for $A^2 - A = O$, so we have $\lambda^2 - \lambda = 0$. This implies that $\lambda = 0$ or 1.

So all eigenvalues of A are either 0 or 1. Indeed, a projection does two things: kill $\mathrm{Ker}(A)$ and preserve $\mathrm{Ran}(A)$. $\mathrm{Ker}(A)$ is exactly the eigenspace for 0, while $\mathrm{Ran}(A)$ is exactly the eigenspace for 1. ☺

Example 7.6.21. Consider a Householder reflection H. As a reflection, we must have $H^2 = I$. In particular, eigenvalues of H must have $\lambda^2 = 1$. As a result, $\lambda = \pm 1$. Indeed, the eigenspace for 1 is the hyperplane of reflection, while the eigenspace for -1 is the normal direction of the hyperplane, which shall be reflected by H via $H\boldsymbol{n} = -\boldsymbol{n}$. ☺

There is a common theme behind both examples. Any polynomial that kills A would determine the eigenvalues of A. These eigenvalues then in turn determines the behavior of A. So in many cases, if you want to study linear maps with certain behavior, you can just specify some polynomial $p(x)$ and consider all matrices that satisfy this polynomial via $p(A) = 0$. As we have see here, $A^2 = A$ gives (oblique) projections, while $A^2 = I$ would actually give you all (oblique) reflections. If you want the orthogonal versions, you can further require some condition between the matrix and its transpose (or adjoint in case of complex matrices). For example, by requiring $A = A^{\mathrm{T}}$ for real matrices, then $A^2 = A$ would give you orthogonal projection, and $A^2 = I$ would give you orthogonal reflection.

Remark 7.6.22. *This remark is completely optional.*

How to perform a Schur decomposition? If you know all the eigenvalues already, then this is super easy. Just look at the proof and figure it out from there. However, if you do not know the eigenvalues of A, how would you do this?

One famous algorithm is the iterated QR-algorithm. Today computers usually use some variants or improved version of it. The idea is this:

1. *Set $A_0 = A$, and perform QR decomposition $A_0 = Q_0 R_0$.*
2. *Set $A_1 = R_0 Q_0$, and perform QR decomposition $A_1 = Q_1 R_1$.*
3. *Set $A_2 = R_1 Q_1$, and perform QR decomposition $A_2 = Q_2 R_2$.*
4. *....*
5. *R_∞ would probably converge to T in the Schur decomposition $A = XTX^{-1}$.*

(This does not always converge. The precise condition is annoying to state clearly. However, in practice, most of the time it should work.)

What is the idea of this? Well, consider these calculations.

1. *$AQ_0 = Q_0 R_0 Q_0 = Q_0 Q_1 R_1$.*
2. *$AQ_0 Q_1 = Q_0 Q_1 R_1 Q_1 = Q_0 Q_1 Q_2 R_2$.*
3. *....*
4. *$A(Q_0 \dots Q_k) = (Q_0 \dots Q_{k+1}) R_{k+1}$.*

Let $X_k = Q_0 \dots Q_k$. Then you see that $AX_k = X_{k+1} R_{k+1}$ where X_k is orthogonal (or unitary in the complex case). Taking limit $k \to \infty$, we see that $AX_\infty = X_\infty R_\infty$, and hence $A = X_\infty R_\infty X_\infty^{-1}$ where X_∞ is unitary and R_∞ is upper triangular. Done.

7.7 Geometric multiplicity and diagonalization

Last section is mostly designed towards eigenvalues. However, to achieve diagonalization, we also need eigenvectors, which are the goal for this section.

Given an eigenvalue λ for a matrix A, then by definition there must be eigenvectors for this eigenvalue, i.e., some non-zero $\boldsymbol{v}$ such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$. In particular, $\operatorname{Ker}(A - \lambda I) \neq \{\mathbf{0}\}$. We can generally think of $\operatorname{Ker}(A - \lambda I)$ as the space of "λ eigenvectors" (with the exception of $\mathbf{0}$ in it).

Definition 7.7.1. *Given an eigenvalue λ for a matrix A, its **eigenspace** is $\operatorname{Ker}(A - \lambda I)$. (This is also exactly made of the zero vector and all eigenvalues for λ.)*

*We say the **geometric multiplicity** of λ is $\dim \operatorname{Ker}(A - \lambda I)$.*

Intuitively, geometric multiplicity measures how many eigenvectors you have for this eigenvalue.

Now, if we want A to be diagonalizable, we want eigenvectors of A to span the whole domain. This means we want the sum space $\sum \operatorname{Ker}(A - \lambda I)$ to be the whole space. In particular, we hope that $\dim(\sum \operatorname{Ker}(A - \lambda I)) = n$.

This sum space is not that easy to study. However, let us show that in fact $\dim(\sum \operatorname{Ker}(A - \lambda I)) = \sum \dim \operatorname{Ker}(A - \lambda I)$. In this sense, we only need to check if the geometric multiplicity adds up to n. The phenomenon here is independent subspaces.

Definition 7.7.2. *For subspaces $W_1, \dots, W_k$ of V, we say they are linearly independent if, for any $\boldsymbol{w}_1 \in W_1, \dots, \boldsymbol{w}_k \in W_k$, then $\sum \boldsymbol{w}_i = \mathbf{0}$ implies all $\boldsymbol{w}_i$ are $\mathbf{0}$.*

Example 7.7.3. Recall that on $\mathbb{R}^2$, the x-axis, the y-axis and the line $x = y$ are pairwise independent, but collectively NOT independent. In particular,

$$\begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}-1\\-1\end{bmatrix} = \mathbf{0},$$

this violates the linear independence condition. Indeed, in this case $\dim \sum W_i = 2$ while $\sum \dim W_i = 3$.

In contrast, in $\mathbb{R}^3$, the three coordinate axes are independent subspaces. If we have

$$\begin{bmatrix}a\\0\\0\end{bmatrix} + \begin{bmatrix}0\\b\\0\end{bmatrix} + \begin{bmatrix}0\\0\\c\end{bmatrix} = \mathbf{0},$$

then we must have $a = b = c = 0$. Indeed, in this case $\dim \sum W_i = \sum \dim W_i = 3$. ☺

Let us first establish linear independence among subspaces.

Proposition 7.7.4. *For subspaces $W_1, \dots, W_k$ of a vector space V, the following are equivalent:*

1. *The subspaces are linearly independent.*
2. *If $\mathcal{B}_1, \dots, \mathcal{B}_k$ are bases for $W_1, \dots, W_k$ respectively, then $\bigcup \mathcal{B}_i$ is a basis for $\sum W_i$.*
3. $\dim(\sum W_i) = \sum \dim W_i$.

Proof.

Downward implications:

Suppose the subspaces are linearly independent, and we have bases $\mathcal{B}_1, \dots, \mathcal{B}_k$ for $W_1, \dots, W_k$ respectively. Let us first show that $\bigcup \mathcal{B}_i$ is linearly independent. Suppose some linear combination of vectors in $\bigcup \mathcal{B}_i$ gives the zero vector. Then we have

$$\text{Linear Combination}(\bigcup \mathcal{B}_i) = \mathbf{0}$$
$$\text{Linear Combination}(\mathcal{B}_1) + \cdots + \text{Linear Combination}(\mathcal{B}_k) = \mathbf{0}.$$

Now, note that each Linear Combination($\mathcal{B}_i$) is some vector in W_i. So by linear independence of subspaces, this implies that each Linear Combination($\mathcal{B}_i$) is zero. However, $\mathcal{B}_i$ is also linearly independent. Hence all coefficients are zero.

As a result, we see that $\bigcup \mathcal{B}_i$ is linearly independent. Since it also trivially span $\sum W_i$, it must be a basis for $\sum W_i$.

Now $\dim \sum W_i = |\bigcup \mathcal{B}_i| = \sum |\mathcal{B}_\rangle| = \sum \dim W_i$. Here absolute value symbol means "the number of elements".

Upward implications:

Suppose $\dim \sum W_i = \sum \dim W_i$, and we have bases $\mathcal{B}_1, \dots, \mathcal{B}_k$ for $W_1, \dots, W_k$ respectively. Then we have $\dim \sum W_i = \sum \dim W_i = \sum |\mathcal{B}_\rangle| \geq |\bigcup \mathcal{B}_i|$. Yet we obviously have $\bigcup \mathcal{B}_i$ spanning $\sum W_i$. Hence $\dim \sum W_i = |\bigcup \mathcal{B}_i|$ and $\bigcup \mathcal{B}_i$ is a basis for $\sum W_i$.

We can also see that $\sum |\mathcal{B}_\rangle| = |\bigcup \mathcal{B}_i|$, hence these sets $\mathcal{B}_1, \dots, \mathcal{B}_k$ are disjoint.

Now suppose we have $\boldsymbol{w}_1 \in W_1, \dots, \boldsymbol{w}_k \in W_k$ such that $\sum \boldsymbol{w}_i = \boldsymbol{0}$. Now for the basis $\bigcup \mathcal{B}_i$, each $\boldsymbol{w}_i$ is a linear combination of vectors in $\mathcal{B}_i$, so $\sum \boldsymbol{w}_i$ is a linear combination of vectors in $\bigcup \mathcal{B}_i$. But this linear combination is equal to $\boldsymbol{0}$, hence all coefficients are zero. In particular, $\boldsymbol{w}_i$ must be the zero linear combination of vectors in $\mathcal{B}_i$. So $\boldsymbol{w}_i = \boldsymbol{0}$. □

Remark 7.7.5. *There are other equivalent conditions as well. For example, for independent subspaces, the intersection of W_i with the sum of the other subspaces must be zero. This is another necessary and equivalent condition, even though we do not need this.*

Next, we shall see that eigenspaces are all independent.

Remark 7.7.6. *Eigenspaces for different eigenvalues clearly have zero intersection. If $\boldsymbol{v} \in \operatorname{Ker}(A - \lambda I) \cap \operatorname{Ker}(A - \mu I)$, then $A\boldsymbol{v} = \lambda \boldsymbol{v}$ and $A\boldsymbol{v} = \mu \boldsymbol{v}$. Therefore $\lambda \boldsymbol{v} = \mu \boldsymbol{v}$. Hence either $\boldsymbol{v} = \boldsymbol{0}$, or $\lambda = \mu$.*

However, this is merely "pairwise independence", which is not enough for collective independence.

Proposition 7.7.7. *Suppose A has eigenvalues (NOT counting algebraic multiplicity) $\lambda_1, \dots, \lambda_k$, and let $V_1, \dots, V_k$ be corresponding eigenspaces. Then these spaces are linearly independent.*

Proof. Pick non-zero $\boldsymbol{v}_i \in V_i$ for each i. Let us show that $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are linearly independent.

Suppose $\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$. What can we do? Since $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, we can try to repeatedly applying A to this equation. This would yeilds the following system of equations:

$$\begin{aligned} a_1 \boldsymbol{v}_1 + \dots \quad & + a_k \boldsymbol{v}_k = 0 \\ a_1 \lambda_1 \boldsymbol{v}_1 + \dots \quad & + a_k \lambda_k \boldsymbol{v}_k = 0 \\ a_1 \lambda_1^2 \boldsymbol{v}_1 + \dots \quad & + a_k \lambda_k^2 \boldsymbol{v}_k = 0 \\ \vdots & \\ a_1 \lambda_1^{k-1} \boldsymbol{v}_1 + \dots & + a_k \lambda_k^{k-1} \boldsymbol{v}_k = 0. \end{aligned}$$

Writing in terms of matrices we have

$$\begin{bmatrix} a_1 \boldsymbol{v}_1 & \dots & a_k \boldsymbol{v}_k \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \dots & \lambda_1^{k-1} \\ 1 & \lambda_2 & \dots & \lambda_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_k & \dots & \lambda_k^{k-1} \end{bmatrix} = \boldsymbol{0}.$$

Note that the right matrix is a Vandermonde matrix. Since $\lambda_1, \dots, \lambda_k$ are eigenvalues WIHTOUT counting algebraic multiplicities, they are all distinct. Therefore this Vandermonde matrix is invertible. So $\begin{bmatrix} a_1 \boldsymbol{v}_1 & \dots & a_k \boldsymbol{v}_k \end{bmatrix} = O$. As a result, since all $\boldsymbol{v}_i$ are non-zero, we have $a_i = 0$ for all i. So these vectors are linearly independent. □

Remark 7.7.8. *The idea behind this proof is called the power method. Recall that we started our study of eigenvalues EXACTLY by trying to study sequences like $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \dots$. Therefore, in many computational applications, people usually compute the sequence $A, A^2, A^3, \dots$ and then extrapolate eigenvalues or eigenvalue behaviors.*

This is also the second major use of Vandermonde matrix.

In some sense, the very motivation of eigensutff come from the study of $A, A^2, A^3, \dots$. It is not surprising that iterations are crucial for many parts of eigenstuff.

We now come to the meat of this section.

Corollary 7.7.9. *An $n \times n$ matrix A is diagonalizable iff its geometric multiplicities $m_g(\lambda)$ add up to n.*

Proof. Note that $\dim(\sum V_i) = \sum \dim V_i$ is the sum of geometric multiplicity. So if the geometric multiplicities add up to n, then $\sum V_i$ is the whole domain. So the whole domain is spanned by eigenvectors, and we have A diagonalizable.

The other direction is simply the reverse of the deductions above. □

Now compare the two multiplicities. $\sum m_a(\lambda_i) = n$ always, while $\sum m_g(\lambda_i)$ might or might not be n. If the eigenvectors fail to span the whole domain, then they shall span some subspace, so we in fact have $\sum m_g(\lambda_i) < n$ in that case. Geometric multiplicities seems to be smaller or equal to algebraic multiplicities. This is indeed the case.

Theorem 7.7.10. *Let λ be an eigenvalue of A. Then $m_g(\lambda) \le m_a(\lambda)$. I.e., for each eigenvalue, geometric multiplicities are less than or equal to the algebraic multiplicities.*

Proof. Let $V \subseteq \mathbb{R}^n$ be the eigenspace for λ with dimension $m = m_g(\lambda)$. Pick a basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_m$, and extend this into a basis for the whole space. Then in this new basis, our matrix A changes into $B = X^{-1}AX$ for some invertible X.

Now consider the first m columns of B. Since $A\boldsymbol{v}_i = \lambda \boldsymbol{v}_i$ for all $1 \le i \le m$, we see that $B\boldsymbol{e}_i = \lambda \boldsymbol{e}_i$ for all $1 \le i \le m$. So $B = \begin{bmatrix} \lambda I_{m\times m} & \star \\ O & B_1 \end{bmatrix}$.

Consider the characteristic polynomial, and use the upper triangular block structure, we see that $\det(xI - B) = \det \begin{bmatrix} (x-\lambda) I_{m\times m} & \star \\ O & xI - B_1 \end{bmatrix} = (x-\lambda)^m p_{B_1}(x)$. Here $p_{B_1}(x)$ is the characteristic polynomial of B_1.

In particular, we see that λ is a root of the characteristic polynomial of B at least m times. Finally, since A, B differs only via a change of basis, they have the same characteristic polynomial, same eigenvalues and same algebraic multiplicity.

As a more computational write-up of the same proof, we have

$$\begin{aligned} AX =& A \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_m & X_1 \end{bmatrix} \\ =& \begin{bmatrix} \lambda \boldsymbol{v}_1 & \dots & \lambda \boldsymbol{v}_m & Y_1 \end{bmatrix} \\ =& \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_m & X_1 \end{bmatrix} \begin{bmatrix} \lambda I_{m\times m} & \star \\ O & B_1 \end{bmatrix} \\ =& X \begin{bmatrix} \lambda I_{m\times m} & \star \\ O & B_1 \end{bmatrix}. \end{aligned}$$

So $X^{-1}AX = \begin{bmatrix} \lambda I_{m\times m} & \star \\ O & B_1 \end{bmatrix}$, and hence the characteristic polynomial of A contains factor $(x-\lambda)^m$. □

In some sense, algebraic multiplicity cares only about the result of a triangularization (Schur decomposition). In contrast, geometric multiplicity cares only about the result of a potential diagonalization. In particular, let us now conclude this venture into a criteria for diagonalizability.

Theorem 7.7.11 (Criteria for diagonalizability)**.** *A matrix A is diagonalizable if and only if for each eigenvalue λ, its algebraic multiplicity equals to its geometric multiplicity.*

Proof. Note that $\sum m_g(\lambda) \leq \sum m_a(\lambda) = n$. So the left hand side equal to n iff equality holds iff $m_a = m_g$ for all λ. □

In particular, why does diagonalizability fail? It is because of DEFECTIVE eigenvalues.

Definition 7.7.12. *An eigenvalue is **defective** if its geometric multiplicity is STRICTLY smaller than its algebraic multiplicity.*

Example 7.7.13. Consider an upper triangular matrix. Then the algebraic multiplicity only cares about the diagonal entries. The rest could be whatever. However, if the entries above the diagonal are too bad, your matrix might NOT be diagonalizable. Then you would fail to have enough geometric multiplicity.

The most important example to keep in mind is $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Its only eigenvalue is 1, with algebraic multiplicity 2 and geometric multiplicity 1. (Can you compute this geometric multiplicity yourself?)

In general, we can look at a ***Jordan block***, which is $J_\lambda = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$. This is the MOST DEFECTIVE case possible, where the algebraic multiplicity of the only eigenvalue λ is n, while the geometric multiplicity is 1. ☺

Remark 7.7.14. *For matrix that cannot be diagonalized, we can in fact always "Jordanize", which is as close to diagonal as possible. For any square matrix A, we can always find decomposition $A = XJX^{-1}$ where J is block diagonal, and each diagonal block is a Jordan block. This matrix J (also called the Jordan canonical form of A) is unique up to the permutation of these diagonal blocks. In particular, A is diagonalizable if and only if its Jordan canonical form is diagonal.*

We do not prove it here though.

Now that the most defective case is dealt with, let us look at the opposite case. This is a very useful scenario.

Definition 7.7.15. *An eigenvalue is **simple** if its algebraic multiplicity is* 1.

Proposition 7.7.16. *A simple eigenvalue cannot be defective. In particular, if all eigenvaleus of A are simple, then A is diagonalizable.*

In other words, if all n eigenvalues of A are distinct, then A is diagonalizable.

Proof. Note that by definition, an eigenvalue λ must have corresponding non-zero eigenvectors. Its eigenspace has dimension at least 1. So $m_g(\lambda)$ is at least 1.

If the eigenvalue λ is simple, then $1 \leq m_g(\lambda) \leq m_a(\lambda) = 1$. So we must have $m_g = m_a$. □

Corollary 7.7.17. *If a triangular matrix has distinct diagonal entries, then it is diagonalizable.*

Let us see some examples to get a better computational idea on diagonalization. In practice, one computational way to get diagonalization for computers is to first find the Schur triangularization, and then diagonalize using row/column operations.

Example 7.7.18. Consider $\begin{bmatrix} 1 & 4 & 5 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix}$. How can we diagonalize it?

Let us try the shearing $\begin{bmatrix} 1 & k & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 4 & 5 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & -k & \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4+k & 5+6k \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix}$. (Think in terms of row and column operations to compute faster.) So by setting $k = -4$, we see that our original matrix is similar to $\begin{bmatrix} 1 & 0 & -19 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix}$.

Next we try $\begin{bmatrix} 1 & & \\ & 1 & k \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -19 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & -k \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -19 \\ 0 & 2 & 6+k \\ 0 & 0 & 3 \end{bmatrix}$. By setting $k = -6$, we are now at $\begin{bmatrix} 1 & 0 & -19 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$.

Finally, we try $\begin{bmatrix} 1 & & k \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -19 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & & -k \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -19+2k \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. Set $k = \frac{19}{2}$, and we are done. Indeed A can be diagonalized.

Set $X = \begin{bmatrix} 1 & 4 & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & 6 \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & -\frac{19}{2} \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 & \frac{29}{2} \\ & 1 & 6 \\ & & 1 \end{bmatrix}$. Then the process above shows that $X^{-1}AX = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 3 \end{bmatrix}$, so $A = X \begin{bmatrix} 1 & & \\ & 2 & \\ & & 3 \end{bmatrix} X^{-1}$. In particular, A has eigenvalues $1, 2, 3$, and the corresponding eigenspaces are spanned by eigenvectors that are corresponding columns of X, respectively. (Note that since $m_g = m_a = 1$, all eigenspaces are one-dimensional.)

In contrast, recall that $\begin{bmatrix} 1 & 1 & \\ & 1 & 1 \\ & & 1 \end{bmatrix}$ cannot be diagonalized. If we try shearing, we would typically have $\begin{bmatrix} 1 & k & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \\ & 1 & 1 \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & -k & \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & k \\ & 1 & 1 \\ & & 1 \end{bmatrix}$. There is no way to kill the $(1,2)$ entry, and the cancellation is precise due to the (algebraically) repeated eigenvalues.

This gives a computational intuition about why algebraically repeated eigenvalues are crucial for defectiveness. Intuitively, for a triangular matrix, to kill the (i,j) entry using row/column operations, we usually need the (i,i) eigenvalue and (j,j) eigenvalue to be different. ☺

Now, let us combine the two multiplicities and have some applications.

Example 7.7.19. Consider again $\begin{bmatrix} 1+a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & 1+a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \dots & 1+a_nb_n \end{bmatrix}$. What is its determinant?

Let us do this by finding the eigenvalues. Then it is enough to find the eigenvalues of $A = \begin{bmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \dots & a_nb_n \end{bmatrix}$.

Note that this matrix has rank at most 1, so $\dim \operatorname{Ker}(A)$ is at least $n-1$. In particular, 0 is an eigenvalue of A with geometric multiplicity at least $n-1$. This in turn implies that 0 also has algebraic multiplicity at least $n-1$.

So, since A has n eigenvalues in total (counting algebraic multiplicity), and we already know that $n-1$ of them are all zero. We just need to find the last one. But since $\operatorname{trace}(A) = \sum a_ib_i$, we see that the last one is simply $\sum a_ib_i$. So A has eigenvalue $0, \dots, 0, \sum a_ib_i$.

Now our goal is to find $\det(I+A)$. Note that $A+I$ has eigenvalues $1, \dots, 1, 1+\sum a_ib_i$, so $\det(I+A) = 1+\sum a_ib_i$. Done. ☺

Example 7.7.20. Consider again $\begin{bmatrix} 1+a_1+b_1 & a_1+b_2 & \dots & a_1+b_n \\ a_2+b_1 & 1+a_2+b_2 & \dots & a_2+b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n+b_1 & a_n+b_2 & \dots & 1+a_n+b_n \end{bmatrix}$. What is its determinant?

Let us do this by finding the eigenvalues. Then it is enough to find the eigenvalues of $A = \begin{bmatrix} a_1+b_1 & a_1+b_2 & \dots & a_1+b_n \\ a_2+b_1 & a_2+b_2 & \dots & a_2+b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n+b_1 & a_n+b_2 & \dots & a_n+b_n \end{bmatrix}$. Note that this matrix has rank at most 2, so $\dim \mathrm{Ker}(A)$ is at least $n-2$. In particular, 0 is an eigenvalue of A with algebraic multiplicity at least $n-2$.

So, since A has n eigenvalues in total (counting algebraic multiplicity), and we already know that $n-2$ of them are all zero. We just need to find the last two. Let them be λ_1, λ_2. How to find them?

From trace we easily have $\lambda_1 + \lambda_2 = \mathrm{trace}(A) = \sum a_i + \sum b_i$. This corresponds to the coefficient of the x^{n-1} term in the characteristic polynomial. However, let us look at the coefficient of the x^{n-2} terms. This gives $\sum_{i<j} \lambda_i \lambda_j = S_2$, where S_2 is the sum of all 2×2 principal minors of A. Note that since all but two eigenvalues are zero, the left hand side is in fact simply $\lambda_1\lambda_2$, the product of the only two non-zero eigenvalues. And the right hand side is $\sum_{i<j} \det \begin{bmatrix} a_i+b_i & a_i+b_j \\ a_j+b_i & a_j+b_j \end{bmatrix} = \sum_{i<j} [(a_i+b_i)(a_j+b_j) - (a_i+b_j)(a_j+b_i)] = \sum_{i<j}(a_ib_j + b_ia_j - a_ib_i - a_jb_j)$.

Now our goal is to find $\det(I+A)$. Note that $A+I$ has eigenvalues $1, \dots, 1, 1+\lambda_1, 1+\lambda_2$, so $\det(I+A) = (1+\lambda_1)(1+\lambda_2) = 1 + (\lambda_1+\lambda_2) + \lambda_1\lambda_2 = 1 + \sum a_i + \sum b_i + \sum_{i<j}(a_ib_j + b_ia_j - a_ib_i - a_jb_j)$. Done.

For aethetic purpose, one might want to further simplify by $\sum_{i<j}(a_ib_j + b_ia_j - a_ib_i - a_jb_j) = \frac{1}{2}\sum_{i,j}(a_ib_j + b_ia_j - a_ib_i - a_jb_j)$, since the cases of $i<j$ and $i>j$ are symmetric, and the terms when $i=j$ are all zero. Then this further simplify to $\frac{1}{2}\sum_{i,j}(a_ib_j + b_ia_j - a_ib_i - a_jb_j) = \frac{1}{2}((\sum a_i)(\sum b_i) + (\sum a_i)(\sum b_i) - n(\sum a_ib_i) - n(\sum a_ib_i)) = (\sum a_i)(\sum b_i) - n\sum a_ib_i$.

Then we have $\det(I+A) = 1 + \sum a_i + \sum b_i + (\sum a_i)(\sum b_i) - n\sum a_ib_i$. ☺

7.8 Limit and Conquer

Suppose A is diagonalizable, say $A = XDX^{-1}$. Then we have a very nice formula $A^k = XD^kX^{-1}$ and more generally $p(A) = Xp(D)X^{-1}$, which are all easy to calculate. What if A is NOT diagonalizable? It turned out that we can still calculate $p(A)$ with some formula. Except that derivatives are now involved.

Example 7.8.1. For a diagonal matrix $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$, then we have $D^k = \begin{bmatrix} d_1^k & & \\ & \ddots & \\ & & d_n^k \end{bmatrix}$. Therefore, for any polynomial $p(x)$, note that $p(D)$ is a linear combination of powers of D. Therefore we can easily verify that in fact $p(D) = \begin{bmatrix} p(d_1) & & \\ & \ddots & \\ & & p(d_n) \end{bmatrix}$.

Now consider a non-diagonalizable matrix, a Jordan block $J = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$. Even though this matrix is NOT diagonalizable, consider $J_t = \begin{bmatrix} 2 & 1 \\ 0 & 2+t \end{bmatrix}$. Now for $t \neq 0$, then all eigenvalues of J_t would be simple. So J_t is diagonalizable for $t \neq 0$. In particular, even though J is NOT diagonalizable itself, it is a limit of diagonal matrices.

Now for $t \neq 0$, we have $J_t = \begin{bmatrix} 1 & \frac{1}{t} \\ & 1 \end{bmatrix} \begin{bmatrix} 2 & \\ & 2+t \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{t} \\ & 1 \end{bmatrix}$. As a result, we see that

$$p(J_t) = \begin{bmatrix} 1 & \frac{1}{t} \\ & 1 \end{bmatrix} \begin{bmatrix} p(2) & \\ & p(2+t) \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{t} \\ & 1 \end{bmatrix} = \begin{bmatrix} p(2) & \frac{p(2+t)-p(t)}{t} \\ & p(2+t) \end{bmatrix}.$$

Hey! That looks like a derivative formula.

Now take limit $t \to 0$, we see that $p(J) = \begin{bmatrix} p(2) & p'(2) \\ & p(2) \end{bmatrix}$. This is true for all polynomial $p(x)$. (Note that $\lim p(J_t) = p(\lim J_t) = p(J)$ because p is continuous. Algebraically, continuity means "it commute with the limit operator".) ☺

Remark 7.8.2. *In fact, the argument above extends to all analytic functions. (An analytic function is a function that equal to its own Taylor expansion.)*

Suppose $f(x)$ is a function with Taylor expansion. Then $f(x)$ is a limit of polynomial functions. Hence $f(D)$ for $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$ is $f(D) = \begin{bmatrix} f(d_1) & & \\ & \ddots & \\ & & f(d_n) \end{bmatrix}$. This gives us a way to define $f(A)$ in general. If $A = XDX^{-1}$, we simply define $f(A) = Xf(D)X^{-1}$, since the X, X^{-1} portion are simple changing the basis.

One can then use the same proof to see that $f(\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}) = \begin{bmatrix} f(2) & f'(2) \\ 0 & f(2) \end{bmatrix}$. For example, $\mathrm{e}^{\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}} = \begin{bmatrix} e^2 & e^2 \\ 0 & e^2 \end{bmatrix}$.

This idea of proving easy cases, and then extend by continuity is a very standard strategy, and in fact very useful everywhere. We are doing linear algebra, so everything is linear or at worst polynomial. So everything is continuous. This makes taking limit very easy to impliment. Here are two optional lemmas you might be interested in. The results is more important than the proof. The proof is a standard mathematical technique about tiny perturbations.

Lemma 7.8.3. *For any square matrix A, it is the limit of a sequence of matrices A_n with distinct eigenvalues. (In particular, all A_n are diagonalizable since all eigenvalues are simple.) We may also require that all A_n are invertible.*

Proof. Consider the Schur decomposition $A = QUQ^{-1}$. Say $U = \begin{bmatrix} \lambda_1 & * & * \\ & \ddots & * \\ & & \lambda_n \end{bmatrix}$. We aim to perturb the diagonal entries of U a bit, to get matrices with distinct eigenvalues. Let $A_t = QU_tQ^{-1}$ where $U_t = \begin{bmatrix} \lambda_1 + t & * & * \\ & \ddots & * \\ & & \lambda_n + nt \end{bmatrix}$. Obviously $\lim_{t\to 0} A_t = A$. We only need to show that for small enough non-zero t, all A_t have distinct eigenvalues.

Let g be the smallest non-zero gap between eigenvalues of A. (E.g., if A has eigenvalues $1, 1, 4, 6, 6, 6$, then $g = 2$.) I claim that for all $0 < |t| < \frac{g}{2n}$, A_t has distinct eigenvalues.

Indeed, eigenvalues of A_t must be $\lambda_1 + t, \ldots, \lambda_n + nt$. For any $i \neq j$, if $\lambda_i = \lambda_j$, then $\lambda_i + it \neq \lambda_j + jt$ due to $|t| > 0$. If $\lambda_i \neq \lambda_j$, then the gap $|\lambda_i - \lambda_j|$ is non-zero and therefore at least g. So we have

$$|(\lambda_i + it) - (\lambda_j + jt)| \geq |\lambda_i - \lambda_j| - i|t| - j|t| \geq g - i|t| - j|t| > 2nt - nt - nt = 0.$$

Hence $\lambda_i + it \neq \lambda_j + jt$ as well. Either way, A_t has distinct eigenvalues and thus is diagonalizable.

(Intuitively, if g is smallest gap, and $2nt < g$, then for all λ_i, λ_j with a gap, then the total perturbation is $it + jt \leq 2nt < g$, which is not enough to fill in this gap.)

If you want all A_t to be invertible as well, then you can again let g be the smallest non-zero gap, both between eigenvalues of A and between non-zero eigenvalues of A and zero. Then the same proof works in the same way. □

So, sometimes to prove some generic theorem about continuous things, we may simply assume that our matrices are diagonalizable or invertible or both, and prove the special case. Then we take limit, and we would have obtained all cases.

As an application, consider the following proof.

Theorem 7.8.4. *Say $n > k$. For any $n \times k$ matrix A and $k \times n$ matrix B, then $p_{AB}(x) = x^{n-k} p_{BA}(x)$. In particular, AB and BA have the same eigenvalues and the same algebraic multiplicity, with the exception of the eigenvalue 0. AB has $n-k$ more zeros as eigenvalues than BA.*

Note that $\det(I + AB) = \det(I + BA)$ is a very simple corrolary of the fact here. If we add identity, then we see that $I + AB$ has $n - k$ more ones as eigenvalues than $I + BA$, and otherwise they have identical eigenvalues. Taking product, and we are done.

Proof. We prove the theorem by first proving it for invertible A, then proving it for square A, B, and then prove the general case when A, B are rectangular.

*A **is invertible:*** If A is invertible, then $A(BA)A^{-1} = AB$. So AB and BA are similar, and we are done.

*A **is NOT invertible:*** Consider $A + tI$. Since A has finitely many eigenvalues, by choosing really tiny t, we can make sure that eigenvalues of $A + tI$ are ALL non-zero. So A is the limit of invertible matrices A_t. Now since $A_t B$ and BA_t have identical characteristic polynomial, by taking limit, AB and BA will have the same characteristic polynomial.

*A, B **are NOT square:*** We add zero columns to A and obtain $A' = \begin{bmatrix} A & O \end{bmatrix}$, a square matrix, and add zero rows to B and obtain B', a square matrix. Note that $A'B' = AB$ while $B'A' = \begin{bmatrix} BA & \\ & O \end{bmatrix}$.

Since $A'B'$ and $B'A'$ have the same characteristic polynomial, we see that $p_{AB}(x) = x^{n-k} p_{BA}(x)$. □

As you can see, each step is trivial. This is a very fundamental relation between AB and BA, and in turn, between $I + AB$ and $I + BA$. Everything that we've done previously, like $\operatorname{trace}(AB) = \operatorname{trace}(BA)$ or $\det(I + AB) = \det(I + BA)$ are all corollaries of this.

We now end this section with the all powerful Cayley-Hamilton theorem.

Proposition 7.8.5. *The process of sending $n \times n$ matrices A to the complex number $p_A(A)$ is continuous.*

Proof. The coefficients of $p_A(x)$ are $(-1)^k S_k$, which is some polynomials in entries of A. So all coefficients of $p_A(x)$ depends on A continuously. So in the calculation of $p_A(A)$, we are adding or multiplying things that all depends on A continuously. Therefore this is a continuous process. □

Theorem 7.8.6 (Cayley-Hamilton)**.** *We have $p_A(A) = O$ for any square matrix A.*

Proof. If A is diagonalizable, then $A = XDX^{-1}$ where the diagonal entries of D are the eigenvalues. Then $p_A(A) = X p_A(D) X^{-1}$. However, $p_A(D)$ simply applies p_A to the eigenvalues, which are roots of p_A. So $p_A(D) = O$. So $p_A(A) = O$.

Now suppose A is NOT diagonalizable. Pick a sequence A_n of diagonalizable matrices A_n such that $A = \lim A_t$.

So $p_A(A) = \lim p_{A_n}(A_n) = \lim O = O$. □

There are many theoretical consequence of Cayley-Hamilton. Here is one:

Corollary 7.8.7. *If A is invertible, then there is a polynomial $p(x)$ such that $A^{-1} = p(A)$.*

Proof. Consider $p_A(x)$. Since $\det(A) \neq 0$, the constant term of $p_A(x)$ is non-zero. So $p_A(x) = xq(A) + c$ for some $c \neq 0$ and some polynomial $q(x)$.

Now since $p_A(A) = O$, we see that $Aq(A) + cI = O$. Rearrange terms, and we have $A^{-1} = -\frac{1}{c} q(A)$. Done. □

Here is an even more powerful version.

Corollary 7.8.8. *If $p(A) = O$ for some polynomial $p(x)$ with degree d, then for any polynomial $q(x)$, $q(A) = r(A)$ for some polynomial $r(x)$ with degree strictly less than d.*

Proof. For any polynomial $q(x)$, perform polynomial division, and we have $q(x) = a(x)p(x) + r(x)$ for some polynomial $a(x)$ and remainder $r(x)$. Hence $r(x)$ has degree strictly less than the degree of $p(x)$, and $q(A) = a(A)p(A) + r(A) = r(A)$. □

Since we always have $p_A(A) = O$, we see that for any $n \times n$ matrix A, we only need to consider polynomials of A with degree at most $n - 1$. No higher degree polynomials of A is ever needed.

Example 7.8.9. Suppose $A^2 = A$, show that $I + A$ is invertible and find its inverse.

How to do this? Well, since eigenvalues of A must satisfy $\lambda^2 = \lambda$, we see that $\lambda = 0, 1$. So eigenvalues of A are 0 or 1. Therefore, eigenvalues of $I + A$ are 1 and 2, and in particular $I + A$ is invertible.

Now inverse of $I + A$ must be some polynomial of $I + A$, which is in term some polynomial of A. However, $A^2 = A$, so any polynomial of A must be $a_0 I + a_1 A$.

If $a_0 I + a_1 A$ is the inverse, then we have $I = (I + A)(a_0 I + a_1 A) = a_0 I + (a_0 + 2a_1)A$. So we want $a_0 = 1$ and $a_0 + 2a_1 = 0$, which gives $a_0 = 1$ and $a_1 = -\frac{1}{2}$. So we have $(I + A)^{-1} = I - \frac{1}{2}A$. ☺

Here is a slightly more complicated example.

Example 7.8.10. Suppose $A^3 + A^2 + A + I = O$, show that $2A + I$ is invertible and find its inverse.

How to do this? Well, since eigenvalues of A must satisfy $\lambda^3 + \lambda^2 + \lambda + 1 = 0$, we see that $\lambda = -1, \mathrm{i}, -\mathrm{i}$. So eigenvalues of A are $-1, \mathrm{i}, -\mathrm{i}$. Therefore, eigenvalues of $2A + I$ are $-1, 1 + 2\mathrm{i}, 1 - 2\mathrm{i}$, and in particular $2A + I$ is invertible.

Now inverse of $2A + I$ must be some polynomial of $2A + I$, which is in term some polynomial of A. However, $A^3 + A^2 + A + I = O$, so any polynomial of A must be $a_0 I + a_1 A + a_2 A^2$.

If $a_0 I + a_1 A + a_2 A^2$ is the inverse, then we have

$$I = (2A+I)(a_0 I + a_1 A + a_2 A^2) = a_0 I + (2a_0 + a_1)A + (2a_1 + a_2)A^2 + (2a_2)A^3 = (a_0 - 2a_2)I + (2a_0 + a_1 - 2a_2)A + (2a_1 - a_2)A^2.$$

So we want $a_0 - 2a_2 = 1$, $2a_0 + a_1 - 2a_2 = 0$ and $2a_1 - a_2 = 0$, which gives $a_0 = -\frac{3}{5}$, $a_1 = -\frac{2}{5}$ and $a_2 = -\frac{4}{5}$. So we have $(2A + I)^{-1} = -\frac{3}{5}I - \frac{2}{5}A - \frac{4}{5}A^2$. ☺

So you always have only finitely many coefficients to consider, and can always attempt at something like the example above. In fact, for any function f with converging Taylor expansion, then $f(A) = r(A)$ for some polynomial $r(x)$ with degree strictly less than d.

Example 7.8.11. For $A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$, we have $\mathrm{e}^A = \begin{bmatrix} e^2 & e^2 \\ 0 & e^2 \end{bmatrix} = \mathrm{e}^2 A - \mathrm{e}^2 I$. ☺

7.9 (Optional) Classification of 2×2 real matrices

Since we were studying real matrices, here are the most important properties of them.

Proposition 7.9.1. *If A is a real square matrix, then any complex eigenvalues would come in conjugate pairs, with same algebraic and geometric multiplicities, and their corresponding eigenvectors are also in conjugate pairs.*

Proof. If $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then $\overline{A\boldsymbol{v}} = \overline{\lambda\boldsymbol{v}}$, and hence $A\overline{\boldsymbol{v}} = \overline{\lambda}\overline{\boldsymbol{v}}$. The rest is easy verifications along similar lines. □

Proposition 7.9.2. *If a real square matrix A has a real eigenvalue, then we can find real eigenvectors for the eigenvalue. Furthermore, real eigenvectors would span the (complex) eigenspace of λ.*

Proof. Note that if $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then $\overline{A\boldsymbol{v}} = \overline{\lambda\boldsymbol{v}}$. Since both A and λ are real, we see that $A\overline{\boldsymbol{v}} = \lambda\overline{\boldsymbol{v}}$. As a result, both $\boldsymbol{v}, \overline{\boldsymbol{v}}$ are eigenvectors for λ. Then $\frac{1}{2}(\boldsymbol{v} + \overline{\boldsymbol{v}})$, $\frac{1}{2\mathrm{i}}(\boldsymbol{v} - \overline{\boldsymbol{v}})$ are real eignevectors for λ and their span would contain $\boldsymbol{v}$.

In particular, not only λ has real eigenvectors, we also see that real eigenvectors are enough to span the eigenspace. □

Another interesting but less important fact happen when n is odd.

Proposition 7.9.3. *If A is an $n \times n$ real square matrix, and n is odd, then A must have a real eigenvalue (and hence with corresponding real eigenvectors).*

Proof. Note that $p_A(x)$ is a polynomial of odd degree. In particular, it is easy to see that $\lim_{x\to\infty} p_A(x) = \infty$ while $\lim_{x\to-\infty} p_A(x) = -\infty$. By intermediate value theorem, somewhere we must be able to find a real λ with $p_A(\lambda) = 0$. □

Corollary 7.9.4. *Any 3-dimensional rotation must have an axis of rotation. (In fact, this is true for all odd-dimensional rotations.)*

Proof. Let this rotation be R, which would be an orthogonal matrix with determinant 1. Then suppose it has an eigenvalue λ. Say $R\boldsymbol{v} = \lambda\boldsymbol{v}$. However, since R is an orthogonal matrix, we have $\|\boldsymbol{v}\| = \|R\boldsymbol{v}\| = |\lambda|\|\boldsymbol{v}\|$, so we see that all eigenvalues must be complex numbers with absolute value 1.

In particular, eigenvalues of R are conjugate pairs of non-real numbers with absolute value 1, and some -1 and some 1. Now since R is a rotation, it preserves orientation, so the product of all its eigenvalues are 1. But the conjugate pairs of complex eigenvalues must multiply to 1, so we see that R must have even numbers of eigenvalue -1.

Now let us count. R has n eigenvalues in total, and n is odd. R must have even numbers of non-real eigenvalues, and even numbers of -1, so it seems that it must have an odd number (hence non-zero number) of eigenvalues 1. In particular, R has eigenvalue 1. And the corresponding real eigenvector must be an axis of rotation. □

Here let us attempt to classify the behavior of all 2×2 real matrices. We finally have enough tools to do so. Throughout this subsection, suppose A is a 2×2.

Example 7.9.5. Suppose A is NOT invertible. If A has rank 0, then obviously the only case is $A = O$. Now suppose A has rank 1.

Then the eigenvalues of A are $0, \lambda$. Suppose $\lambda \neq 0$. Then A has distinct eigenvalues and thus diagonalizable. So $A = XDX^{-1}$ where $D = \begin{bmatrix} \lambda & \\ & 0 \end{bmatrix}$. In particular, $(\frac{1}{\lambda}A)^2 = \frac{1}{\lambda}A$. So A is a multiple of some oblique or orthogonal projection.

As a side note, $\operatorname{trace}(A) = 0 + \lambda = \lambda \neq 0$ in this case.

If $\lambda = 0$, yet $A \neq O$, so we see that A is NOT diagonalizable. By Schur decomposition, we see that $A = QTQ^{-1}$ for some orthogonal matrix Q, where $T = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} = a\boldsymbol{e}_1\boldsymbol{e}_2^{\mathrm{T}}$ for some non-zero a. Also note that $\begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} a & \\ & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{a} & \\ & 1 \end{bmatrix}$. So all these matrices are similar to $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. The matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ does this: it sends the y-axis to the x-axis, and it sends the x-axis to the origin. It is also funny that $A \neq O$ but $A^2 = O$.

As a side note, $\operatorname{trace}(A) = 0$ in this case.

In summary we have the following classification:

1. If A has rank zero, then $A = O$.
2. If A has rank one and $\operatorname{trace}(A) \neq 0$, then A is a multiple of oblique or orthogonal projection.
3. If A has rank one and $\operatorname{trace}(A) = 0$, then A is similar to the Jordan block $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

☺

Now we are ready to study the invertible matrices. Note that in this case, A will have two non-zero eigenvectors λ_1, λ_2. And furthermore, we have $\operatorname{trace}(A) = \lambda_1 + \lambda_2$ and $\det(A) = \lambda_1\lambda_2$. In particular, the trace and determinant of A completely determines the eigenvalues!

For the moment, let us assume that $\det(A) = 1$. So A would preserve area. First we study rotation like behavior, which would obviously preserve area.

Example 7.9.6 (Elliptic rotations)**.** If A has non-real eigenvalues, then note that these eigenvalues come in conjugate pairs. So the two eigenvalues must be $\lambda, \overline{\lambda}$. Also note that, by premises, $\lambda\overline{\lambda} = 1$, so λ is a unit complex number.

Now by polar decomposition, we must have $\lambda = \mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$ for some θ. Furthermore, since this should NOT be purely real, we have $|\cos\theta| < 1$. For the record, we can see that in this case $|\mathrm{trace}(A)| = |2\cos\theta| < 2$

Now since our eigenvalues are non-real, they are conjugate pairs, and thus they are distinct, so A is diagonalizable. So A is similar to $\begin{bmatrix} \mathrm{e}^{\mathrm{i}\theta} & \\ & \mathrm{e}^{-\mathrm{i}\theta} \end{bmatrix}$. However, the matrix $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is also similar to $\begin{bmatrix} \mathrm{e}^{\mathrm{i}\theta} & \\ & \mathrm{e}^{-\mathrm{i}\theta} \end{bmatrix}$. Hence A is similar to R_θ.

Note that R_θ is a simple rotation. You can think of the plane $\mathbb{R}^2$ as made of concentric circles around the origin, and R_θ simple rotates all these circles by the same amount.

If we change basis and get A, then A would be an "elliptic rotation". You can think of the plane $\mathbb{R}^2$ as made of concentric ellipses around the origin, and A simple rotates all these ellipses simultaneously.

What is the speed of this rotation? It turns out that different points on the ellipse will be rotated with different speed, depending on how close they are to the origin. It is actually a version of Kepler's second law of planetary motion, where the area swept in unit time should be the same. Note the fact that $\det(A) = 1$, so A should preserve area. In particular, the triangle made by $\boldsymbol{v}, A\boldsymbol{v}$ and the triangle made by $A\boldsymbol{v}, A^2\boldsymbol{v}$ should have the same area, and so on. This can help us determine the speed of rotation induced by A. ☺

We are now left with the cases where the eigenvalues are both real. Suppose they are distinct.

Example 7.9.7. Suppose the eigenvalues are distinct. Since their product is 1, they are $\lambda, \frac{1}{\lambda}$ where $\lambda \neq \pm 1$. Note that in this case, we necessarily have $|\mathrm{trace}(A)| > 2$. Let us assume in this case that $\lambda > 0$, so both eigenvalues are positive.

Since the two eigenvalues are distinct, $A = XDX^{-1}$ with $D = \begin{bmatrix} \lambda & \\ & \frac{1}{\lambda} \end{bmatrix}$. So A after a change of basis becomes D. What is the geometric behavior of this D?

Note that $D\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \lambda x \\ \frac{1}{\lambda} y \end{bmatrix}$. In particular, it would fix the product of the coordinates! As you can imagine, it will slide points along hyperbolas such as curves $xy = k$ in $\mathbb{R}^2$. These hyperbolas are "orbits" of D.

After a change of basis, A would become a "hyperbolic rotation". Let $\boldsymbol{v}, \boldsymbol{w}$ be a basis of $\mathbb{R}^2$ made of eigenvectors of A, and consider all hyperbolas with these two lines as asymptotes. Then A simply slide each point along the hyperbola it lies on.

What is the speed of this rotation? Again, since $\det(A) = 1$, Kepler's second law of planetary motion applies. The triangle made by $\boldsymbol{v}, A\boldsymbol{v}$ and the triangle made by $A\boldsymbol{v}, A^2\boldsymbol{v}$ should have the same area, and so on. This can help us determine the speed of rotation induced by A.

What if we started with $\lambda < 0$? Then we can apply the analysis above to $-A$. Hence A is the negation of a hyperbolic rotation. You may think of it as $A = -B$ where B is the hyperbolic rotation. So A would perform a hyperbolic rotation and then negate the outcome vector. ☺

Now we are left with the case where the two eigenvalues are real and identical. We have obvious candidates like $\pm I$. However, there is also another special linear map of this case that preserve area, i.e., shearing.

Example 7.9.8. The two eigenvalues must be $\lambda_1 = \lambda_2 = 1$ or $\lambda_1 = \lambda_2 = -1$. Note that in this case, we have $|\mathrm{trace}(A)| = 2$.

If A is diagonalizable with identical eigenvalues, then $A = XDX^{-1}$ where $D = \pm I$. But then we would have $A = \pm I$. So these are the only possibility in this case.

Suppose A is NOT diagonalizable. Then $A = QTQ^{-1}$ for some triangular T. Suppose $\lambda_1 = \lambda_2 = 1$, then we have $T = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$, so A after a change of basis is a shearing transformation.

What if $\lambda_1 = \lambda_2 = -1$? Then again, we simply apply the arguments above to $-A$. So $A = -B$ where B is a shearing transformation. ☺

We have now covered all cases. Let us make a summary.

Proposition 7.9.9. *If* $\det(A) = 1$*, then the geometric behavior of* A *is exactly one of the following, characterized by trace.*

1. *(*$\text{trace}(A) > 2$*) Sliding points along hyperbolas with common asymptotes.*
2. *(*$\text{trace}(A) = 2$*) Either* $A = I$ *or* A *is a shearing along some direction.*
3. *(*$|\text{trace}(A)| < 2$*) Rotations around concentric ellipses of the same eccentricity.*
4. *(*$\text{trace}(A) = -2$*) Either* $A = -I$ *or* A *is a shearing along some direction then negation.*
5. *(*$\text{trace}(A) < -2$*) Sliding points along hyperbolas with common asymptotes, and then negation.*

Note that the case $|\text{trace}(A)| < 2$ is sometimes called elliptic, and the case $|\text{trace}(A)| > 2$ is sometimes called hyperbolic, for obvious reason. A less intuitive name is that the case of $\text{trace}(A) = \pm 2$ is sometimes called parabolic. (This is because the parabola is an intermediate state between the ellipses and the hyperbolas.)

Now what if $\det(A) \neq 1$? We have two scenarios. Let us focus now on the case of positive determinant.

Suppose $\det(A) > 0$. Then let $t = \sqrt{\det(A)}$, we have $\det(\frac{1}{t}A) = 1$. So $\frac{1}{t}A$ is one of the cases above. In particular, $A = tB$ for some B in one of the cases above. A is basically B and then a uniform scaling.

Example 7.9.10. If B is an elliptic rotation and $t > 1$, then A will "spiral out". For each application of A, you shall rotate, and then uniformly scale outward. Repeatedly apply A, and you will see all points spiraling away from the origin. Reversely, if $0 < t < 1$ in this case, then you shall see a spiral in process instead. Note that this is the case of non-real eigenvalues with $\text{trace}(A)^2 < 4\det(A)$ ☺

Example 7.9.11. Suppose B is a hyperbolic sliding (i.e., with positive eigenvalues). Note that this corresponds to the case where A is diagonalizable with distinct eigenvalues. Suppose that both eigenvalues are positive.

Then we have $A = XDX^{-1}$ with $D = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}$. If we have $0 < \lambda_1, \lambda_2 < 1$, then A will shrink everything towards the origin. The origin is like a "sink" where everything would flows towards. If $\lambda_1, \lambda_2 > 1$, then A will now push everything away from the origin. The origin is like a "source" where everything pours outwards.

If one eigenvalue is larger than 1, while the other is less than 1, then the picture is similar to a hyperbolic sliding, even though the orbits are not necessarily hyperbolas. The origin is called a "saddle" point, where points would flow in through one eigen-direction, and yet pushed away along another eigen-direction.

The intermediate state, where one eigenvalue is exactly 1, means we have a line of fixed points, and we stretch or shrink along the other direction, depending on whether the other eigenvalue is larger than or less than 1. The geometric behaviors are a bit like the process of opening doors or closing doors.

Finally, the case when A has negative eigenvalues are simply above process and then negation. ☺

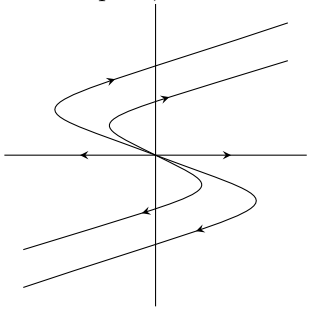
We are again left with the cases of identical real eigenvalues. If $B = \pm I$, then A is also a multiple of identity, and this case is trivial. So let us now look at the case when B is a shearing, or alternatively, when A is similar to some $\begin{bmatrix} \lambda & 1 \\ & \lambda \end{bmatrix}$

Example 7.9.12. Consider the case $A = \begin{bmatrix} \lambda & 1 \\ & \lambda \end{bmatrix}$ where $\lambda = 100$. Then we can think of A as two process: first we perform a tiny shearing $\begin{bmatrix} 1 & \frac{1}{100} \\ & 1 \end{bmatrix}$, and then we scale everything by 100.

First imagine that the imput is a vector in a direction above but close to the negative x-axis. Then since the y-component is tiny, shearing should effect very little of it. Yet the scaling should effect it greatly and move it towards the upper left. As a result, the vector would move towards its upper left.

However, also because $A\begin{bmatrix}x\\y\end{bmatrix}$ is parallel to $\begin{bmatrix}x+\frac{1}{100}y\\y\end{bmatrix}$, the vector is slowly but surely turning clockwise. Eventually, once the vector is turned past the y-axis, then both the shearing and the scaling would push it to the right, and the vector would just go faster and faster right ever since.

This phenomena is called a degenerate source. All points are pushed away, but the direction of push is somewhat spiral, somewhat asymptotic.



In general, suppose A has identical positive eigenvalues but is NOT diagonalizable. When $\det(A) > 1$, its behavior would be some degenerate source as shown. When $\det(A) < 1$, then it would be some degenerate sink, which spirals in. Finally, when $\det(A)$, then this is just a regular shearing.

If A has negative eigenvalues, then A is the above process and then negation. ☺

We have one last, case, the case when $\det(A) < 0$. Note that in this case, $p_A(x)$ is a degree two polynomial with negative constant coefficient, and hence it always have distinct real roots, one positive and one negative. In particular, A MUST be diagonalizable.

Consider $\begin{bmatrix}\lambda_1 & \\ & \lambda_2\end{bmatrix}$ where $\lambda_1 > 0, \lambda_2 < 0$. This is essentially $\begin{bmatrix}1 & \\ & -1\end{bmatrix}$ and then we stretch or shrink a bit in each axis. So all such matrices must be (maybe oblique) reflection and then some stretches.

7.10 Linear Differential Equations

7.10.1 Differential Equations with only One Function

Example 7.10.1. What is the solution to $f'(x) = f(x)$? Consider the differential operator D, which is linear. We are trying to find $Df = f$, which would be eigenvectors of D for eigenvalue 1. ☺

As you can see, taking derivative is a linear operator. To study this operator D, obviously we need to understand its eigenvalues and eigenvectors. Well, guess what? We do.

Proposition 7.10.2. *Let D be the derivative operator. Then any complex number $\lambda \in \mathbb{C}$ is an eigenvalue of D with geometric multiplicity one. Its eigenspace is spanned by $\mathrm{e}^{\lambda x}$.*

Proof. This is a basic calculus class result. □

Before moving on, recall that if $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then it is very easy to verify that $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v}$ for all polynomial $p(x)$.

Example 7.10.3. Consider the harmonic osciliator. Say your position is $f(t)$, then we know $mf''(t) = -kf(t)$. However, suppose that we also have frictions proportional to your speed. Then now we have $mf''(t) = -bf'(t) - kf(t)$. Then the differential equation is $mf'' + bf' + kf = 0$ for some positive constants m, b, k. What is the solution?

Let us assume $m = 1, b = 3, k = 2$. Then $f'' + 3f' + 2f = 0$. Let $p(x) = x^2 + 3x + 2$, then naturally $p(D)f = 0$. So we are trying to find eigenvectors for the eigenvalue 0.

Now $p(-1) = p(-2) = 0$. So since e^{-x} and e^{-2x} are eigenvectors of D for the eigenvalues -1 and -2 respectively, it turned out that they are both eigenvectors of $p(D)$ for the eigenvalue $p(-1) = 0 = p(-2)$.

I state without proof here that for any polynomial p of degree d, then any eigenspace of $p(D)$ is complex d dimensional. So the solution to our differential equation is the span of e^{-x} and e^{-2x}.

So our harmonic osciliator with friction in this case behaves like $a\mathrm{e}^{-t} + b\mathrm{e}^{-2t}$ and the limit as $t \to \infty$ is 0. So eventually we stopped moving at the origin. ☺

Proposition 7.10.4. *Let D be the derivative operator. For any polynomial p of degree d, then any eigenspace of $p(D)$ is d dimensional.*

Proof. Check out some odinary differential equation class for this proof. □

With this idea, you should be able to almost solve all differential equations that looks like $p(D)f = 0$. If you see $f''' + 6f'' + 11f' + 6f = 0$? NO PROBLEM. The roots for p are 1,2,3, so all solutions are linear combintions of $\mathrm{e}^x, \mathrm{e}^{2x}, \mathrm{e}^{3x}$. As long as all the roots of p are distinct, we can find all solutions to differential equations this way.

Let us check out another case. This allow us to handle complex eigenvalues with better grace.

Example 7.10.5. Now suppose we have no friction, so we have $f'' + f = 0$. Then $p(D)f = 0$ for $p(x) = x^2 + 1$, and the roots for p are $\pm\mathrm{i}$. The two eigenvectors are $\mathrm{e}^{\mathrm{i}t}$ and $\mathrm{e}^{-\mathrm{i}t}$, and they are both eigenvectors for $p(D)$ for the eigenvalue 0. Any solution must look like $z\mathrm{e}^{\mathrm{i}t} + w\mathrm{e}^{-\mathrm{i}t}$ for some complex numbers z, w. But how can I find all real solutions?

Well, note that the (complex) span of $\mathrm{e}^{\mathrm{i}t}, \mathrm{e}^{-\mathrm{i}t}$ is the same as the (complex) span of $\cos t, \sin t$. So the answer is simply all real linear combinations of $\cos t, \sin t$.

Alternatively, note that $p(D)e^{\mathrm{i}t} = 0$ means that $p(D)\cos t + \mathrm{i}p(D)\sin t = 0$, so $p(D)\cos t = 0$ and $p(D)\sin t = 0$.

Finally, note that the solutions are all periodic. Indeed, if there is no friction, then the osciliator would simply bounce forever, in a periodic manner. ☺

Techniques here are vaid, as long as $p(D)f = 0$ and the polynomial $p(x)$ have distinct roots. What if $p(x)$ has repeated roots?

Example 7.10.6. Consider any potential solution to $(D^2 - 2D + I)f = 0$. Since the roots of $p(x)$ are both 1, we see that e^x is an eigenvector with eigenvalue 0. However, since our polynomial has degree two, the solution space is suppose to be two dimensional. What could another basis vector be?

This could be xe^x. You can feel free to verify that indeed $D^2 - 2D + I$ would kill this function. This is NOT an eigenvector, but a ***generalized eigenvector***, which we shall discuss next semester. ☺

7.10.2 (Optional) Eigenspace of $p(A)$ and $p(\frac{\mathrm{d}}{\mathrm{d}x})$

In last subsection, we claim that if p has degree n, then $\mathrm{Ker}(p(D))$ would have dimension n for the differential operator D. Why is that?

Lemma 7.10.7. $\dim \mathrm{Ker}(A^k) \leq k \dim \mathrm{Ker}(A)$. *Here A can be any abstract linear transformation.*

Keep in mind of the example where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Note that A first kill the entire x-axis, while sending everyone else to the x-axis, setting up to be killed in the next application of A. So $A^2 = 0$, which is also easy to see computationally.

Note that, informally, $\dim \mathrm{Ker}(A)$ is how many things A would kill in one step. If this is known, then how many thing would A kill in k steps? Well, intuitively, A^k would kill $k \dim \mathrm{Ker}(A)$ stuff. However, since there might be "repetitive kill", we see that $\dim \mathrm{Ker}(A^k) \leq k \dim \mathrm{Ker}(A)$ in general.

Proof. Note that $A : \mathrm{Ker}(A^{n+1}) \to \mathrm{Ker}(A^n)$ is a pullback relation, i.e., if $A\boldsymbol{v} \in \mathrm{Ker}(A^n)$, then A will kill $A\boldsymbol{v}$ in n steps, and hence A will kill $\boldsymbol{v}$ in $n + 1$ steps.

Now use the pull back dimension formula, we have $\dim \mathrm{Ker}(A^{n+1}) = \dim(\mathrm{Ker}(A^n) \cap \mathrm{Ran}(A)) + \dim \mathrm{Ker}(A) \leq \dim \mathrm{Ker}(A^n) + \dim \mathrm{Ker}(A)$.

Then by induction, it is easy to see that $\dim \mathrm{Ker}(A^k) \leq k \dim \mathrm{Ker}(A)$. □

Lemma 7.10.8. *If $p(x)$ and $q(x)$ has greatest common factor $g(x)$, then we can find polynomials $a(x)$ and $b(x)$ such that $a(x)p(x) + b(x)q(x) = g(x)$.*

Proof. Search for the name "Euclidean algorithm". There is a highly similar statements for integers. □

Proposition 7.10.9. *If $p(x)$ and $q(x)$ has no common root, then* $\operatorname{Ker}(p(A)q(A)) = \operatorname{Ker}(p(A)) + \operatorname{Ker}(q(A))$ *and* $\operatorname{Ker}(p(A)) \cap \operatorname{Ker}(q(A)) = \{\mathbf{0}\}$. *In particular, we have* $\dim \operatorname{Ker}(p(A)q(A)) = \dim \operatorname{Ker}(p(A)) + \dim \operatorname{Ker}(q(A))$.

Proof. Obviously $q(A)\boldsymbol{v} = \mathbf{0}$ or $p(A)\boldsymbol{v} = \mathbf{0}$ would both imply $p(A)q(A)\boldsymbol{v} = \mathbf{0}$. So we have $\operatorname{Ker}(p(A)) + \operatorname{Ker}(q(A)) \subseteq \operatorname{Ker}(p(A)q(A))$.

Now since $p(x)$ and $q(x)$ has no common root, therefore we can find $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$. In particular, $\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v}$. So if $p(A)q(A)\boldsymbol{v} = \mathbf{0}$, we see that $a(A)p(A)\boldsymbol{v} \in \operatorname{Ker}(q(A))$ and $b(A)q(A)\boldsymbol{v} \in \operatorname{Ker}(p(A))$. So $\boldsymbol{v} \in \operatorname{Ker}(p(A)) + \operatorname{Ker}(q(A))$. We have established that $\operatorname{Ker}(p(A)q(A)) = \operatorname{Ker}(p(A)) + \operatorname{Ker}(q(A))$.

Now let us show that they have trivial intersection. Suppose $\boldsymbol{v} \in \operatorname{Ker}(p(A)) \cap \operatorname{Ker}(q(A))$. Note that again we have $\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} = \mathbf{0} + \mathbf{0} = \mathbf{0}$. So we are done. □

Theorem 7.10.10. *Suppose $p(x) = \prod (x - \lambda_i)^{m_i}$ is a factorization into distinct roots λ_i with multiplicity m_i. Then* $\dim \operatorname{Ker}(p(A)) = \sum \operatorname{Ker}(A - \lambda_i I)^{m_i} \leq \sum m_i \operatorname{Ker}(A - \lambda_i I)$.

Proof. Just combine previous results. □

Corollary 7.10.11. $\dim \operatorname{Ker}(p(\frac{\mathrm{d}}{\mathrm{d}x}))$ *is exactly the degree of p.*

Proof. Since all eigenspaces of $\frac{\mathrm{d}}{\mathrm{d}x}$ are one-dimensional, we see that $\dim \operatorname{Ker}(p(\frac{\mathrm{d}}{\mathrm{d}x})) = \sum \dim \operatorname{Ker}(\frac{\mathrm{d}}{\mathrm{d}x} - \lambda_i I)^{m_i} \leq \sum m_i$, which is the degree of p.

We now just need to show that $\dim \operatorname{Ker}(\frac{\mathrm{d}}{\mathrm{d}x} - \lambda_i I)^{m_i}$ is indeed m_i. We already know that it is at most m_i, so we just need to find m_i linearly independent vectors in it. The answer is $\mathrm{e}^{\lambda_i x}, x\mathrm{e}^{\lambda_i x}, \ldots, x^{m_i - 1}\mathrm{e}^{\lambda_i x}$. So we are done. □

7.10.3 Linear Systems of Differential Equations

Example 7.10.12. I take a bottle of milk out of the fridge, and put it into a bowl of hot water to warm it up. Let us say that at time t, the temperature of the milk is $M(t)$ and the temperature of the water is $W(t)$ Then we have $M'(t) = a(W(t) - M(t))$ and $W'(t) = b(M(t) - W(t))$, where a, b are some positive constants to be determined. How to solve this system?

We have two functions, $M(t), W(t)$. We also have the following descriptions: $\begin{cases} M'(t) = -aM(t) + aW(t) \\ W'(t) = bM(t) - bW(t) \end{cases}$

So in fact we have $\begin{bmatrix} M'(t) \\ W'(t) \end{bmatrix} = \begin{bmatrix} -a & a \\ b & -b \end{bmatrix} \begin{bmatrix} M(t) \\ W(t) \end{bmatrix}$. ☺

Example 7.10.13. Say we want to solve $f''' + 6f'' + 11f' + 6f = 0$. Let $g = f'$, and $h = f''$. Then we have $\begin{bmatrix} f' \\ g' \\ h' \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} \begin{bmatrix} f \\ g \\ h \end{bmatrix}$. So a high order linear differential equation becomes a first order linear system of differential equations.

(Also be mindful of the fact that the matrix $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix}$ is EXACTLY thetranspose of the companion matrix for the polynomial $x^3 + 6x^2 + 11x + 6$.) ☺

How to solve these systems? We are trying to solve $\boldsymbol{v}' = A\boldsymbol{v}$. If $f' = kf$, then the solution is $\mathrm{e}^{kx} f(0)$. So I can make the following guess: If $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$, then the solution is probably related to e^{At}, which is a matrix. I claim that the solution should be $\mathrm{e}^{At}\boldsymbol{v}(0)$.

In particular, if the initial conditions are arbitrary, then $\mathrm{e}^{At}\boldsymbol{v}(0)$ could be anything in $\operatorname{Ran}(\mathrm{e}^{At})$, i.e., some linear combinations of columns of e^{At}. So the solution space to $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$ is simply $\operatorname{Ran}(\mathrm{e}^{At})$.

Proposition 7.10.14. *For any diagonal matrix D, the solution space to $\frac{\mathrm{d}}{\mathrm{d}x}\boldsymbol{v} = D\boldsymbol{v}$ is $\mathrm{Ran}(\mathrm{e}^{Dt})$. In fact, the solution is $\mathrm{e}^{Dt}\boldsymbol{v}(0)$ where $\boldsymbol{v}(0)$ is some arbitrary initial value.*

Proof. Say $D = diag(a, b)$. Then the system reads $f' = af$ and $g' = bg$. So we know the solution is that f is a multiple of e^{at} and g is a multiple of e^{bt}. So $\begin{bmatrix} f \\ g \end{bmatrix}$ is a linear combination of columns of $\begin{bmatrix} \mathrm{e}^{at} & 0 \\ 0 & \mathrm{e}^{bt} \end{bmatrix} = \mathrm{e}^{Dt}$.

Now plut in $t = 0$ to the equation $\boldsymbol{v}(t) = \begin{bmatrix} \mathrm{e}^{at} & 0 \\ 0 & \mathrm{e}^{bt} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$, we see that $\boldsymbol{v}(t) = \begin{bmatrix} \mathrm{e}^{at} & 0 \\ 0 & \mathrm{e}^{bt} \end{bmatrix} \boldsymbol{v}(0)$. □

Proposition 7.10.15. *For a diagonalizable matrix $A = BDB^{-1}$, the solution space to $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$is $\mathrm{Ran}(\mathrm{e}^{At})$. In fact, the solution is $\mathrm{e}^{At}\boldsymbol{v}(0)$ where $\boldsymbol{v}(0)$ is some arbitrary initial value.*

Proof. Set $\boldsymbol{w}(t) = B^{-1}\boldsymbol{v}(t)$. Then $BD\boldsymbol{w}(t) = (B\boldsymbol{w}(t))' = B\boldsymbol{w}'(t)$. Here B comes out of the derivative because it is constant. Since B is also invertible, we have $D\boldsymbol{w} = \boldsymbol{w}'$. Hence the solution space to $\boldsymbol{w}$ is $\mathrm{Ran}(\mathrm{e}^{Dt})$.

Now $\boldsymbol{v} = B\boldsymbol{w}$. Hence the solution space to $\boldsymbol{v}$ is $\mathrm{Ran}(B\mathrm{e}^{Dt}) = \mathrm{Ran}(B\mathrm{e}^{Dt}B^{-1}) = \mathrm{Ran}(\mathrm{e}^{At})$. Here the first equality is true because B^{-1} is invertible and does not effect the codomain at all. □

The statement is still true for non-diagonalizable A, but we don't prove it here.

Example 7.10.16. In real life, many couples behave in a periodic way. They are very sweet with each other for a while, and they argue and fight for a while, and then they are sweet again, and then they fight again. In short, their romantic relation exhibit a periodic behavior. What are some possible explanations?

Say two person A and B are romantically involved. The love of A for B is $f(t)$, a function of time, and the love of B for A is $g(t)$, a function of time. Now assume that A is a normal person. To a normal person, the more you are loved, the more you love the other. So $f'(t)$ is proportional of $g(t)$. Say $f'(t) = g(t)$. Assume that, unfortunately, B is a very unappreciative person. The more you love B, then more B takes you for granted. Then more you ignore B, the more B is obssessed with you. So $g'(t)$ is proportional to $-f(t)$, say $g'(t) = -f(t)$.

Then $\begin{bmatrix} f' \\ g' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix}$. So the solution space is $\mathrm{Ran}(\mathrm{e}^{\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} t})$.

Now $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = (\frac{1}{-2\mathrm{i}} \begin{bmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{bmatrix}) \begin{bmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{bmatrix} \begin{bmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{bmatrix}$. So $\mathrm{e}^{\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} t}$ is in fact $(\frac{1}{-2\mathrm{i}} \begin{bmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{bmatrix}) \begin{bmatrix} \mathrm{e}^{\mathrm{i}t} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}t} \end{bmatrix} \begin{bmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{bmatrix}$, which is $\begin{bmatrix} \frac{\mathrm{e}^{\mathrm{i}t}+\mathrm{e}^{-\mathrm{i}t}}{2} & \frac{\mathrm{e}^{\mathrm{i}t}-\mathrm{e}^{-\mathrm{i}t}}{2\mathrm{i}} \\ -\frac{\mathrm{e}^{\mathrm{i}t}-\mathrm{e}^{-\mathrm{i}t}}{2\mathrm{i}} & \frac{\mathrm{e}^{\mathrm{i}t}+\mathrm{e}^{-\mathrm{i}t}}{2} \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$.

So there are constants a, b such that $f(t) = a\cos t + b\sin t$ and $g(t) = -a\sin t + b\cos t$. As you can see, they are periodic. ☺

7.10.4 (Optional) Non-linear Romantic Dynamics

Even for non-linear differential equations, linear algebra is still a very powerful tool. To see this, let us first try to have a slightly more realistic dynamic model for romantic relations.

Example 7.10.17. Again let $f(t), g(t)$ be the love the two person A, B have for each other. What is $f'(t)$? Well, we know it should be proportional to $g(t)$, where the proportion is positive if A is appreciative, and negative if A is non-appreciative, and 0 if A love B but does not really care if B loves back or not. So let us say A has an appreciativeness of k_A and B has an appreciativeness of k_B.

On the other hand, our energy is finite. Your love for another person cannot be infinity, because you don't have that much energy. For each person, there might be an ideal amount of love L, and if your love for another person is more than L, you will start to feel tired and emotionally drained. Let us say the ideal amount of love for A and for B are L_A and L_B. Then $f'(t)$ should be proportional to $L_A - f(t)$ and $g'(t)$ should be proportional to $L_B - g(t)$.

So our model is this: $\begin{bmatrix} f'(t) \\ g'(t) \end{bmatrix} = \begin{bmatrix} k_A g(t)(L_A - f(t)) \\ k_B f(t)(L_B - g(t)) \end{bmatrix}$. In a healthy relation, we hope that $\lim f(t) = L_A$ while $\lim g(t) = L_B$. In an indifferent relation, we have $\lim f(t) = \lim g(t) = 0$. In an unhealthy relation, $\lim f(t)$ and $\lim g(t)$ fail to exist because they are both periodic. ☺

In general, a non-linear system is like this: $\begin{bmatrix} f'(t) \\ g'(t) \end{bmatrix} = \begin{bmatrix} h_1(f(t), g(t)) \\ h_2(f(t), g(t)) \end{bmatrix}$. Then in particuylar, given a pairs of values of f and g, we can calculate f' and g'. In short, IF WE ALEADY KNOW where we are, then we know the direction we are moving to. This gives us a vector field.

Example 7.10.18. (Draw pictures of some vector fields. Electromagnetic fields and so on.) ☺

A generic vector field may have many behaviors, but what is important are the ***fixed points*** of this vector field. It could be a sink, a source, or a saddle, or neither. How can we study this behavior?

A fixed point, or an equilibrium, is a point on the fg-plane where if you start there, you never move. So you are looking for (f, g) values that makes $f' = g' = 0$. So this is like solving the values for f and g from $h_1(f, g) = 0$ and $h_2(f, g) = 0$.

Recall that in one-variable calculus, to find a local max or local min of a function $f(x)$, we first find critical points by looking at the information in the first derivative, to solve $f'(x) = 0$. Then AT THESE CRITICAL POINTS, we try to look at their second derivative to classify their behavior. It turned out that we can do the same thing here. This is the Hartman-Grobman Theorem.

We try to solve $f' = g' = 0$ to find all "critical points", or equilibriums. Then at these points, we try to look at their second order derivative to classify their behavior. Recall that $f' = h_1$ and $g' = h_2$, so the second order derivatives for f and g are the derivatives for h_1 and h_2. You will have to look at the matrix $\begin{bmatrix} \frac{\partial h_1}{\partial f} & \frac{\partial h_1}{\partial g} \\ \frac{\partial h_2}{\partial f} & \frac{\partial h_2}{\partial g} \end{bmatrix}$.

Example 7.10.19. Consider a highly simplified case. Say $\boldsymbol{v}'(t) = \begin{bmatrix} h_1(\boldsymbol{v}) \\ h_2(\boldsymbol{v}) \end{bmatrix}$, and for each point $\boldsymbol{v} = \begin{bmatrix} f(t) \\ g(t) \end{bmatrix}$, we have constant $\begin{bmatrix} \frac{\partial h_1}{\partial f} & \frac{\partial h_1}{\partial g} \\ \frac{\partial h_2}{\partial f} & \frac{\partial h_2}{\partial g} \end{bmatrix} = A$. What do we have?

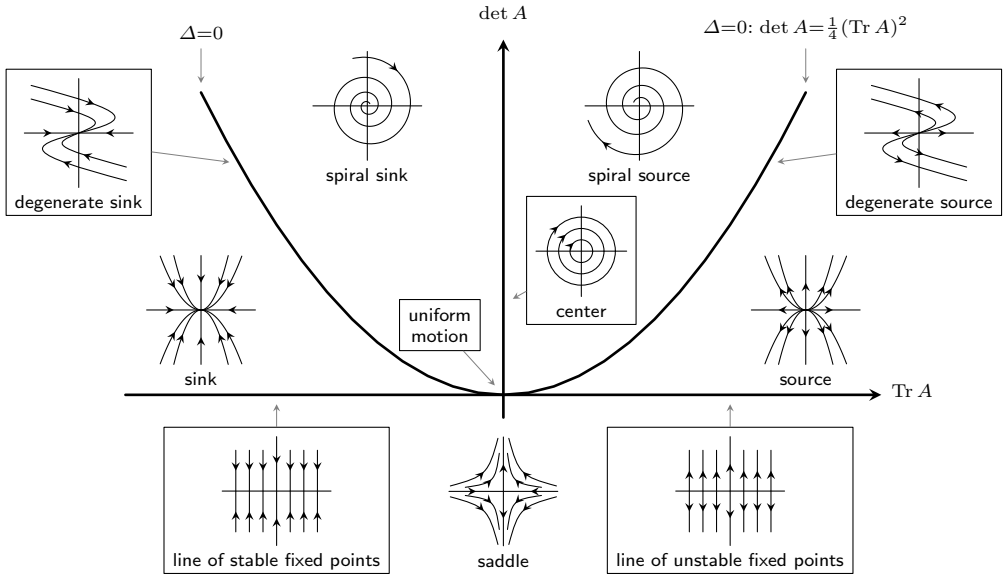
Note that if we have constant $\begin{bmatrix} \frac{\partial h_1}{\partial f} & \frac{\partial h_1}{\partial g} \\ \frac{\partial h_2}{\partial f} & \frac{\partial h_2}{\partial g} \end{bmatrix} = A$, then the vector function $\boldsymbol{h}(\boldsymbol{v}) = \begin{bmatrix} h_1(\boldsymbol{v}) \\ h_2(\boldsymbol{v}) \end{bmatrix}$ will have constant derivative. In particular, it is easy to guess and verify that $\boldsymbol{h}(\boldsymbol{v}) = A\boldsymbol{v}$.

As a result, we see that $\boldsymbol{v}'(t) = \boldsymbol{h}(\boldsymbol{v}) = A\boldsymbol{v}(t)$. So the solution is $\mathrm{e}^{At}\boldsymbol{v}(0)$.

In particular, if the second derivative of your evolution is A, then locally, vectors would flow around your critical point in approximately the same way vectors flow according to the linear map e^A.

Now note that e^A is a real 2×2 matrix with positive eigenvalues. By our previous classification, we can obtain the following diagram.

Poincaré Diagram: Classification of Phase Portraits in the $(\det A, \operatorname{Tr} A)$-plane



☺

Example 7.10.20. Critical points for our romantic relation system are $(f, g) = (0, 0)$ and $(f, g) = (L_A, L_B)$. The second derivative matrix is $\begin{bmatrix} -k_A g & k_A(L_A - f) \\ k_B(L_B - g) & -k_B f \end{bmatrix}$. So at (L_A, L_B), this second derivative matrix is $\begin{bmatrix} -k_A L_B & 0 \\ 0 & -k_B L_A \end{bmatrix}$. So the two eigenvalues are exactly $-k_A L_B$ and $-k_B L_A$, both are real, and the eigendirctions are horizontal and vertical. Also note that L_A and L_B are by assumption positive.

So if both k_A, k_B are positive, then this is a sink. It attracts everyone around. So two appreciative person are likely to be attracted to their ideal love limits. If k_A, k_B are both negative, this is negative. So two unappreciative person are repelled by their love limits, and they may never work out. Finally, if k_A, k_B have different sign, then this is a saddle. It will attract from the eigendirection for positive k-value and repel from the eigendirection for the negative k-value. It seems to promote periodic behavior.

And at $(0, 0)$, this second derivative matrix is $\begin{bmatrix} 0 & k_A L_A \\ k_B L_B & 0 \end{bmatrix}$. So if k_A, k_B have the same sign, then we have two real eigenvalues of distinct signs, and we have a saddle at the origin. If k_A, k_B have distinct signs, then we have two imaginary eigenvalues and trace zero, and so we tend to rotate around the origin. ☺

Example 7.10.21. Let us specifically analyze the case when $k_A > 0$ and $k_B > 0$. Hopefully, we are always in a relation that is mutually appreciative. Then (L_A, L_B) is a sink, while $(0, 0)$ is a saddle, with an attracting eigendirection in the 1st-3rd quadrant and a repelling eigendirection in the 2nd-4th quadrant. Draw the picture to see some flows. As you can see, there seems to be some curve through the origin, above which we will always be attracted to our ideal love limit, and below which we will be repelled to mutual hatred.

This curve has an equation of $\frac{L_A}{k_A}(\frac{f(t)}{L_A} - \ln|1 - \frac{f(t)}{L_A}|) = \frac{L_B}{k_B}(\frac{g(t)}{L_B} - \ln|1 - \frac{g(t)}{L_B}|)$. I call this the confession line. Suppose you are an appreciative person, and you are secretly in love with another appreciative person. Then should you confess your love and start your romance? Well, if you two are above the confession line, then even if the other person does not love you back, you should still go ahead and confess, because EVERYTHING above the confession line will eventually be attracted to the mutual ideal love limits. However, if you two are below the confession line, then hold it for the moment, because your confession would only result in eventual mutual hatred.

Just as a side note, the curve along which your love will evolve is always $\frac{L_A}{k_A}(\frac{f(t)}{L_A} - \ln|1 - \frac{f(t)}{L_A}|) = \frac{L_B}{k_B}(\frac{g(t)}{L_B} - \ln|1 - \frac{g(t)}{L_B}|) + C$ for some constant C. The direction of your evolution depends on the critical point situation.

Take a look at any curve above the confession line. There is a valuable lesson here. Suppose you are trying to persue someone, and you know you two should be above the confession line, then don't give up! Look at the place where your curve intersect with the x or y axis. This is the moment where your love to the other person dropped to the bottum. This is your darkest moment, then moment of despair. But LOOK, this is also exactly the moment where the other person finally starts to like you. This is not chicken soup for the soul. This is mathematics. ☺

Example 7.10.22. What if both k_A and k_B are negative? Then (L_A, L_B) is repelling, and $(0,0)$ is a saddle, attracting along some direction in the 1st-3rd quadrant, and repelling along some direction of the 2nd-4th quadrant. The solution curves are still $\frac{L_A}{k_A}(\frac{f(t)}{L_A} - \ln|1 - \frac{f(t)}{L_A}|) = \frac{L_B}{k_B}(\frac{g(t)}{L_B} - \ln|1 - \frac{g(t)}{L_B}|) + C$.

Drawing some solution curves, you see that the curves are almost IDENTICAL with the two positive k-value case, except that all directions are negated. And invariabley, it will end up with one person, say A, extremely hate the other, and the other, say B, wierdly reached the ideal love limit, since B loves to be hated. I think it is safe to conclude that two unappreciative person would never be together. ☺

Example 7.10.23. What if we have an appreicative person A and an unappreciative person B? Then (L_A, L_B) is a saddle attracting in the horizontal direction and repelling in the vertical direction. And $(0,0)$ is a swirling point. The relation would behave in a clockwise and periodic manner. The solution curves still have the same equations though, $\frac{L_A}{k_A}(\frac{f(t)}{L_A} - \ln|1 - \frac{f(t)}{L_A}|) = \frac{L_B}{k_B}(\frac{g(t)}{L_B} - \ln|1 - \frac{g(t)}{L_B}|) + C$.

Let us start from the point where your periodic curve intersect with the negative f-axis. At this moment, the jerk B start to have an interest in A, while A hates B. Then B started to persue A, and after a long persuing process, B eventually mananged to get A's love back. Very shortly after, there will be a sweet spot where the two are almost close to their ideal love limit. They will probably get married at that point. However, immediately after marriage, the jerk B, being the jeriest jerk, starts to lose interest in A. Very shortly after, our curve intersect with the f-axis agiain, marking the point where A completely lose interest in B. This is probably the point where A started cheating on B. Now A is still in love with B, and in fact the love of A for B is at maximum at this moment. So A will desperately try to cling on to B, to win B's love back. But this just annoys B extremely and speed up the process of B hating A. After a very long and mutually painful process, our curve eventually hit negative g-axis, marking the spot where the two no longer love each other. In most cases, they get a divorce and never see each other again.

In some cases some couples manage to stay married. Oh boy, then they will only get to torture themselves over and over agian with a very tiny sweep period and very very long period of mutual hatred. I mean, why bother? And the sweeter the sweet time is, the longer the mutual hatred will be, as can be seen from these orbiting curves. ☺

To sum up, if $f' = h_1(f,g)$ and $g' = h_2(f,g)$, you can find critical points by solving $h_1 = h_2 = 0$, and you can classify your critical points as sources or sinks or stuff by looking at the eigenstuff of $\begin{bmatrix} \frac{\partial h_1}{\partial f} & \frac{\partial h_1}{\partial g} \\ \frac{\partial h_2}{\partial f} & \frac{\partial h_2}{\partial g} \end{bmatrix}$.

7.11 Spectral Theorem

7.11.1 Spectral Theorem for Normal Matrices

Remark 7.11.1. *(Optional) This is just to explain the name of "spectral theorem".*

Given any system, say a single hydrogen atom, physicists would use a linear operator H to describe its evolution (usually the Hamiltonian operator). Each state of the system is described by some function, and H is a linear map acting on these functions.

If one intends to study potential orbits of the electrons, then we are seeking states that are stable under the evolution, i.e., eigenvectors (or eigenfunctions) of H. If $Hf = \lambda f$, then f is an orbit, and the eigenvalue λ is the corresponding energy state of this orbit.

So when electrons change orbit, it will change energy and therefore emit a light with certain frequence (i.e., certain color). These are called the spectrum of the hydrogen atom.

As you can see, this ultimately depends on the eigenvalue. As a result, people usually use the name "spectrum" to refer to the set of possible eigenvalues of a matrix or operator, and people usually use the name "spectral theorem" to refer to a big theorem governing the structures of eigenvalues and eigenvectors. The diagonalization $A = XDX^{-1}$ is also sometimes called a ***spectral decomposition****.*

In this section we shall study the fundamental eigen structure of normal matrices. First let us see some examples.

Example 7.11.2. Consider $P = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$. It has trace one and determinant zero, so the eigenvalues are 0 and 1 (distinct) and therefore the matrix can be diagonalized as $D = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix}$. Since $D^2 = D$, we see that $P^2 = P$. This is an oblique projection. Indeed, the eigenspace for the eigenvalue 1 is $\mathrm{Ker}(P-I) = \mathrm{Ker}(I-P) = \mathrm{Ran}(P)$ according to properties of projections, and this is spanned by $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. The the eigenspace for the eigenvalue 0 is $\mathrm{Ker}(P)$ obviously spanned by $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. So the two eigenspaces are NOT orthogonal.

But if we have $P = \begin{bmatrix} 0.1 & 0.3 \\ 0.3 & 0.9 \end{bmatrix}$, then since $P^2 = P = P^{\mathrm{T}}$, this is an orthogonal projection. The eigenspace for eigenvalue 1 is $\mathrm{Ker}(P - I) = \mathrm{Ker}(I - P) = \mathrm{Ran}(P)$, which is orthogonal to $\mathrm{Ker}(P)$ the eigenspace for the eigenvalue 0.

This is not a singular occurence. You can also verify this: if $H^2 = I$, it is an (oblique) reflection, and it is an orthogonal projection if and only if $H = H^{\mathrm{T}}$. Note that the "mirror" or the reflection is $\mathrm{Ker}(H - I)$, the set of fixed points, while the direction of reflection is $\mathrm{Ker}(H + I)$, the set of vectors $\boldsymbol{v}$ such that $H\boldsymbol{v} = -\boldsymbol{v}$. So the two eigenspaces are orthogonal if and only if H is symmetric.

But this is not a phenomenon to symmetric matrices alone. Consider $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. We have

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix} \begin{bmatrix} \mathrm{i} & \\ & -\mathrm{i} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix}^{-1}.$$

As you can see, the eigenspace for i is spanned by $\begin{bmatrix} 1 \\ -\mathrm{i} \end{bmatrix}$ and the eigenspace for $-\mathrm{i}$ is spanned by $\begin{bmatrix} 1 \\ \mathrm{i} \end{bmatrix}$. These two complex vectors are orthogonal as well. ☺

So, what kinds of matrices would have orthogonal eigenspaces? It seems like we need A to have some relation to A^{T}, or in the case of complex matrices, we need A to have some relation to A^*. Now we can have our definition.

As a side note, since we are interested in orthogonal eigenspaces, throughout this section we need inner product structure. We shall always assume that our space is $\mathbb{R}^n$ or $\mathbb{C}^n$ with the usual dot product as the inner product structure.

Definition 7.11.3. *A matrix is normal if $AA^* = A^*A$.*

This includes real symmetric matrices, real skew-symmetric matrices, real orthogonal matrices, complex Hermitian matrices, complex skew-Hermitian matrices and complex unitary matrices. There are also normal matrices that are none of these special cases, such as $\begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix}$.

Now, the following lemma is the key reason why these matrices are so nice: its eigenstructure respect the complex conjugate.

Lemma 7.11.4. *If A is normal and $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then $A^*\boldsymbol{v} = \overline{\lambda}\boldsymbol{v}$.*

Proof. If $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then $(A - \lambda I)\boldsymbol{v} = \boldsymbol{0}$. Note that we must utilize the inner product structure here. So we rewrite the equation as $\|(A - \lambda I)\boldsymbol{v}\| = 0$, and then $\boldsymbol{v}^*(A - \lambda I)^*(A - \lambda I)\boldsymbol{v} = 0$.

Now comes the use of normality. Since $AA^* = A^*A$, we have a commutativity $(A - \lambda I)^*(A - \lambda I) = A^*A - \overline{\lambda}A - \lambda A^* + \lambda\overline{\lambda}I = AA^* - \overline{\lambda}A - \lambda A^* + \lambda\overline{\lambda}I = (A - \lambda I)(A - \lambda I)^*$.

As a result, $0 = \boldsymbol{v}^*(A - \lambda I)^*(A - \lambda I)\boldsymbol{v} = \boldsymbol{v}^*(A - \lambda I)(A - \lambda I)^*\boldsymbol{v} = \|(A - \lambda I)^*\boldsymbol{v}\| = \|(A^* - \overline{\lambda}I)\boldsymbol{v}\|$. So we see that $(A^* - \overline{\lambda}I)\boldsymbol{v} = \boldsymbol{0}$, and $A^*\boldsymbol{v} = \overline{\lambda}\boldsymbol{v}$. □

Lemma 7.11.5. *If T is a (complex) upper triangular matrix and T is normal, then T is diagonal.*

Proof. Suppose the upper left entry of T is λ, so we have $T = \begin{bmatrix} \lambda & * \\ \boldsymbol{0} & T' \end{bmatrix}$. Then since T is upper triangular, we have $T\boldsymbol{e}_1 = \lambda\boldsymbol{e}_1$. Then $T^*\boldsymbol{e}_1 = \overline{\lambda}\boldsymbol{e}_1$, and therefore taking adjoint we have $\boldsymbol{e}_1^*T = \lambda\boldsymbol{e}_1^* = \lambda\boldsymbol{e}_1^{\mathrm{T}}$. This implies that the first row of T is $\lambda\boldsymbol{e}_1^{\mathrm{T}}$.

In particular, we have $T = \begin{bmatrix} \lambda & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{0} & T' \end{bmatrix}$.

Now T' is an upper triangular normal matrix of smaller size. So we are done by induction. □

We are now ready for the big theorem.

Theorem 7.11.6 (Spectral Theorem for Normal Matrices)**.** *If A is normal, then we can find unitary U and diagonal D such that $A = UDU^* = UDU^{-1}$.*

In particular, A is always diagonalizable, there is an orthonormal basis made of eigenvectors of A, and all geometric multiplicities are equal to corresponding algebraic multiplicities.

Proof. Consider a Schur decomposition $A = UTU^*$. Then $AA^* = A^*A$ implies that $TT^* = T^*T$. But then this means T is triangular and normal, and therefore diagonal. So we are done. □

Since A is diagonalizable, the domain $\mathbb{C}^n$ of A can be decomposed as the direct sum of the eigenspaces of A. In fact, these eigenspaces must be mutually orthogonal! This is already evident from the spectral theorem above, but you can also enjoy the following independents proof.

Corollary 7.11.7. *If A is normal, then its eigenspaces are mutually orthogonal.*

Proof. Let us show that eigenspaces are mutually orthogonal. Suppose $A\boldsymbol{v} = \lambda\boldsymbol{v}$ and $A\boldsymbol{w} = \mu\boldsymbol{w}$ where $\lambda \neq \mu$.

Now $\boldsymbol{v}^*(A\boldsymbol{w}) = \mu\boldsymbol{v}^*\boldsymbol{w}$. On the other hand, $(\boldsymbol{v}^*A)\boldsymbol{w} = (A^*\boldsymbol{v})^*\boldsymbol{w} = (\overline{\lambda}\boldsymbol{v})^*\boldsymbol{w} = \lambda\boldsymbol{v}^*\boldsymbol{w}$. Since $\mu \neq \lambda$, we have no choice but to conclude that $\boldsymbol{v} \perp \boldsymbol{w}$. □

Remark 7.11.8. *Note that the requirement of being normal is necessary and sufficient for the spectral theorem here. Suppose $A = UDU^*$ for some unitary U. Then $A^* = UD^*U^*$, and $AA^* = UDU^*UDU^* = UDD^*U^*$. Since diagonal matrices commutes with each other, we have $DD^* = D^*D$. So $AA^* = UDD^*U^* = UD^*DU^* = A^*A$. So A is normal.*

Example 7.11.9. Let V be the space of periodic smooth real functions with period 2π. We make it an inner product space via the inner product $\langle f, g\rangle = \int_0^{2\pi} f(x)g(x)\,\mathrm{d}x$, i.e., we integrate the product over a single period.

Now, consider the derivative operator $D = \frac{\mathrm{d}}{\mathrm{d}x}$. (If you are familiar with integration by parts, skip this paragraph.) One of the most important property of the derivative is the product rule or Leibniz rule, i.e., $D(fg) = (Df)g + f(Dg)$. By integration, we would obtain $fg|_a^b = \int_a^b Df(x)g(x)\,\mathrm{d}x + \int_a^b f(x)Dg(x)\,\mathrm{d}x$. This technique is called integration by parts. Namely, if one attempt to do integration of a product $\int_a^b Df(x)g(x)\,\mathrm{d}x$, then one can integrate one factor function while differentiating the other factor function, and have $\int_a^b Df(x)g(x)\,\mathrm{d}x = (fg)|_a^b - \int_a^b f(x)Dg(x)\,\mathrm{d}x$.

Now take integration from 0 to 2π, and note that fg is periodic and thus $(fg)|_0^{2\pi} = 0$. So we have $\langle Df, g\rangle = -\langle f, Dg\rangle$. If you think of $\langle Df, g\rangle$ as $(Df)^{\mathrm{T}}g = f^{\mathrm{T}}D^{\mathrm{T}}g$, and think of the right hand side $-\langle f, Dg\rangle$ as $-f\mathrm{T}Dg$, we see that $D^{\mathrm{T}} = -D$. So the differentiation operator is skew-symmetric.

In practice, many physical phenomena (heat, wave, etc.) are related to the second derivative. Then we have $(D^2)^{\mathrm{T}} = (D^{\mathrm{T}})^2 = (-D)^2 = D^2$, so this would be a symmetrix operator.

In particular, eigenspaces of D^2 are orthogonal. So you immediately see that, for example, $\sin x, \sin(2x), \dots$ are all orthogonal because they are eigenvectors of D^2 for different eigenvalues. (I.e., for integers $m \neq n$, we must have $\int_0^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x = 0$.)

(Optional) It is also fun to do this directly but with the same idea as above, i.e., different eigenvalues lead to $\lambda \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = \mu \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$, which means $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} = 0$. For integers $m \neq n$, we have

$$\int_0^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x = -\frac{1}{n}\cos(nx)\sin(mx)|_0^{2\pi} - m\int_0^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x = -m\int_0^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x.$$

But similarly we have

$$\int_0^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x = -\frac{1}{m}\sin(nx)\cos(mx)|_0^{2\pi} - n\int_0^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x = -n\int_0^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x.$$

So since $m \neq n$, all must be zero. ☺

Now let us move on to specific matrices. First we consider real symmetric or complex Hermitian matrices.

Proposition 7.11.10. *If $A = A^*$, then all eigenvalues are real. If A is a real matrix and $A = A^{\mathrm{T}}$, then furthermore, the spectral decomposition $A = QDQ^{-1}$ can be made so that Q, D are both real matrices. (So Q is real orthogonal and D is real diagonal.)*

Proof. For all normal matrices, we know if $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then $A^*\boldsymbol{v} = \overline{\lambda}\boldsymbol{v}$. However, if furthermore $A = A^*$, then in fact we also have $A^*\boldsymbol{v} = A\boldsymbol{v} = \lambda\boldsymbol{v}$. Since an eigenvector $\boldsymbol{v}$ must be non-zero, we must have $\lambda = \overline{\lambda}$.

Now if A is real and symmetric, then immediately we see that A is diagonalizable and all eigenvalues are real. So all eigenspaces are spanned by real vectors, and they are all orthogonal to each other. By picking an orthonormal basis for each eigenspace, together they form an orthonormal basis for $\mathbb{R}^n$ made of eigenvectors. Say this basis is the real orthogonal matrix Q. Then $A = QDQ^{-1}$. □

Remark 7.11.11. *Note that this is a necessary and sufficient condition. Conversely, if $A = QDQ^{-1}$ for unitary Q and real diagonal D, then $A^* = QD^*Q^{-1} = QDQ^{-1} = A$, so A is Hermitian. And if Q is real orthogonal, then A is also real, and thus symmetric.*

Let us consider the geometric implication of this. Most of the matrices are diagonalizable. Therefore, they are scalings along eigendirections.

Example 7.11.12. Consider $A = XDX^{-1}$ where $D = diag(2, 3)$ and $X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Then A would stretch in the x-axis direction by a factor of 2, and stretch in the $x = y$ direction by a factor of 3. (If you are curious, it will repel points along curves $\ln|y| = \frac{\ln 3}{\ln 2}\ln|x - y| + k$, or via parametrization $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} a2^t + b3^t \\ b3^t \end{bmatrix}$ for constants a, b.)

Note that the two direction of stretching are NOT perpendicular. We are stretching along an oblique frame of reference. However, in the case of normal matrices, X would be unitary or in the real case orthogonal. This means we are stretching along some orthogonal frame of reference.

In particular, consider the effect of A on a unit circle. We are stretching the circle in the x-axis direction by a factor of 2, and stretch in the $x = y$ direction by a factor of 3. Since the vectors $\boldsymbol{e}_1, \boldsymbol{u} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ are on the original circle, after the action of A, the vectors $2\boldsymbol{e}_1, 3\boldsymbol{u}$ are on the resulting ellipse. You can draw the unique ellipse centered around the origin through points $\pm 2\boldsymbol{e}_1, \pm 3\boldsymbol{u}$. Note that $\boldsymbol{e}_1, \boldsymbol{u}$ are NOT the direction of the major-axis and the minor-axis of the resulting ellipse (because they are oblique).

Consider a symmetric operation $A = QDQ^{-1}$ where Q is orthogonal, and $D = diag(2, 3)$. Then we are stretching along an orthogonal fram of reference. In particular, the two eigendirections are EXACTLY the directions of the major-axis and the minor-axis, and the lengths of the resulting ellipse is EXACTLY the eigenvalues of A. (Because that is how much A stretches.)

In short, $A = QDQ^{-1}$ means the geometric action of A is rotation–coordinate stretch–rotation back. ☺

7.11.2 (Optional) Other special cases of spectral theorem

We are mainly concerned with unitary matrices, skew-Hermitian matrices, real orthogonal matrices and real skew-symmetric matrices.

Proposition 7.11.13. *If U is a unitary matrix, then all its eigenvalues are unit complex numbers. In particular, $U = QDQ^{-1}$ where Q is unitary and D is diagonal with diagonal entries $d_k = \mathrm{e}^{\mathrm{i}\theta_k}$.*

Proof. Again note that U is normal, so $U\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $U^*\boldsymbol{v} = \overline{\lambda}\boldsymbol{v}$. However, $U^* = U^{-1}$. As a result, we see that $\lambda^{-1} = \overline{\lambda}$, so λ is a unit complex number. □

Remark 7.11.14. *Note that this is also necessary and sufficient. If $D = \begin{bmatrix} \mathrm{e}^{\mathrm{i}\theta_1} & & \\ & \ddots & \\ & & \mathrm{e}^{\mathrm{i}\theta_n} \end{bmatrix}$ and $U = QDQ^*$ for some unitary Q, then $U^{-1} = QD^{-1}Q^* = QD^*Q^* = U^*$. So U is unitary.*

Geometrically, a unitary matrix means up to an orthogonal change of variable, we simply rotate the complex number in each coordinate in their respective complex plane.

For the real case, we have an even better result.

Lemma 7.11.15. *If A is real and normal, and $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$ for real vectors $\boldsymbol{v}, \boldsymbol{w}$ and non-real eigenvalue λ, then we have $\|\boldsymbol{v}\| = \|\boldsymbol{w}\|, \boldsymbol{v} \perp \boldsymbol{w}$.*

Proof. Say $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$ for real vectors $\boldsymbol{v}, \boldsymbol{w}$ and non-real eigenvalue λ. Then since A is real, by taking complex conjugate, we have $A(\boldsymbol{v} - \mathrm{i}\boldsymbol{w}) = \overline{\lambda}(\boldsymbol{v} - \mathrm{i}\boldsymbol{w})$.

However, since eigenspaces for distinct eigenvalues are mutually orthogonal, and $\lambda \neq \overline{\lambda}$, hence $0 = (\boldsymbol{v} + \mathrm{i}\boldsymbol{w})^*(\boldsymbol{v} - \mathrm{i}\boldsymbol{w}) = \|\boldsymbol{v}\|^2 - \|\boldsymbol{w}\|^2 - 2\mathrm{i}(\boldsymbol{w}^*\boldsymbol{v})$. Then the real part and imaginary parts must both be zero, and we see that $\|\boldsymbol{v}\| = \|\boldsymbol{w}\|, \boldsymbol{v} \perp \boldsymbol{w}$. □

Theorem 7.11.16. *If A is real orthogonal, then we have $A = QDQ^{-1}$ where Q is real orthogonal, and D is real block diagonal where each diagonal block is either 1×1 with value ± 1, or it is some 2×2 rotation matrix R_θ.*

Proof. If all eigenvalues of A are ± 1, then since all eigenvalues are real, $A = UDU^*$ for unitary U and real diagonal D. So, taking adjoint, we have $A^* = UD^*U^* = UDU^* = A$. But A is also real, so it is real symmetric. So, we can perform $A = QDQ^{-1}$ with real orthogonal Q and real diagonal D where the diagonal entries are ± 1. So we are done.

Suppose this is not the case. Then A has an eigenvalue $\lambda \neq \pm 1$. Since A is unitary, $|\lambda| = 1$, and hence λ cannot be real. But A is also a real matrix, so complex eigenstuff come in pairs!

Say $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$ for real vectors $\boldsymbol{v}, \boldsymbol{w}$. By previous lemma, we have $\|\boldsymbol{v}\| = \|\boldsymbol{w}\|, \boldsymbol{v} \perp \boldsymbol{w}$. So let us scale $\boldsymbol{v}, \boldsymbol{w}$ simultaneously so that they are now unit vectors.

Since $\boldsymbol{v}, \boldsymbol{w}$ are unit mutually orthogonal vectors, we can complete them into an orthogonal basis for $\mathbb{R}^n$, and as columns they form an orthogonal matrix X. Also since $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$, by setting $\lambda = \cos\theta + \mathrm{i}\sin\theta$, we can deduce that $A\begin{bmatrix}\boldsymbol{v} & \boldsymbol{w}\end{bmatrix} = \begin{bmatrix}\boldsymbol{v} & \boldsymbol{w}\end{bmatrix} R$ for some rotation matrix R. As a result. $AX = X\begin{bmatrix} R & O \\ O & A_1 \end{bmatrix}$. The rest is standard induction. □

Consider the implication of this. It gives a complete description of the geometric behavior of an orthogonal matrix A. It means we can decompose the whole space into mutually orthogonal planes which A would rotate independently, and then a subspace of points fixed by A, and finally a subspace of points reflected by A.

However, note that $\begin{bmatrix} -1 & \\ & -1 \end{bmatrix}$ can also be thought of as a rotation. So we have the following result:

Corollary 7.11.17. *If A is orthogonal and $\det(A) = 1$, then A simply rotates mutually orthogonal planes independently. If $\det(A) = -1$, then A would first rotates mutually orthogonal planes independently, and then reflect along a direction orthogonal to all the planes of rotation.*

In particular, an orthogonal matrix is either a rotation, or a rotation plus reflection.

Now let us move on to the skew-Hermitian case. They are intuitively the opposite of being Hermitian.

Proposition 7.11.18. *If A is skew-Hermitian, then all eigenvalues are purely imaginary.*

Proof. Same old same old. Since A is normal, $A\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $A^*\boldsymbol{v} = \overline{\lambda}\boldsymbol{v}$. But since $A^* = -A$, we see that $\overline{\lambda} = -\lambda$, so λ has no real part. □

Remark 7.11.19. *This is again necessary and sufficient, and the proof is trivial.*

Also note that if A is skew-Hermitian, then all diagonal entries of A must be purely imaginary because $A^* = -A$. In particular, consider $\begin{bmatrix} a\mathrm{i} & \\ & a\mathrm{i} \end{bmatrix}$ for real a. This is NOT a Hermitian matrix, because the diagonal is not real. This is in fact Skew-Hermitian, depsite the apparent look of symmetricity.

If A is a skew-symmetric matrix, then its eigenvalues are precisely pairs of conjugate purely imaginary numbers, so we can have the following fact.

Corollary 7.11.20. *If A is an $n \times n$ real symmetric matrix, and n is odd, then A is NOT invertible.*

Proof. A has an odd number of eigenvalues, but all the non-real eigenvalues come in pairs. So A must have a real eigenvalue. However, the only number that is both real and purely imaginary is 0. So A has eigenvalue 0. □

Note that the skew-Hermitian matrix $\begin{bmatrix} a\mathrm{i} & \\ & -a\mathrm{i} \end{bmatrix}$ is similar to the real matrix $\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$.

Proposition 7.11.21 (Darboux basis)**.** *If A is real skew-symmetric and invertible, then we have $A = Q\begin{bmatrix} O & -D \\ D & O \end{bmatrix}Q^{-1}$ for some real orthogonal Q and invertible diagonal D.*

Proof. Suppose all eigenvalues of A are zero. Then since A is also diagonalizable, we must have $A = O$ and the whole statement is trivial.

Suppose we have some non-zero eigenvalues. Suppose $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$ for real vectors $\boldsymbol{v}, \boldsymbol{w}$ and non-zero λ. Then λ cannot be real. So $\|\boldsymbol{v}\| = \|\boldsymbol{w}\|, \boldsymbol{v} \perp \boldsymbol{w}$. So let us scale $\boldsymbol{v}, \boldsymbol{w}$ simultaneously so that they are now unit vectors.

Since $\boldsymbol{v}, \boldsymbol{w}$ are unit mutually orthogonal vectors, we can complete them into an orthogonal basis for $\mathbb{R}^n$, and as columns they form a real orthogonal matrix X. Also since $A(\boldsymbol{v} + \mathrm{i}\boldsymbol{w}) = \lambda(\boldsymbol{v} + \mathrm{i}\boldsymbol{w})$, by setting $\lambda = a\mathrm{i}$ for a real a, we can deduce that $A\begin{bmatrix} \boldsymbol{v} & \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v} & \boldsymbol{w} \end{bmatrix} R$ for $R = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}$. As a result. $AX = X\begin{bmatrix} R & O \\ O & A_1 \end{bmatrix}$.

Now by induction, $X^{-1}AX$ is now $\begin{bmatrix} & a & & \\ -a & & & \\ & & & -D \\ & & D & \end{bmatrix}$. Permute the second block row and third block row, while permuting the second block column with the third block column, we see that $P^{-1}X^{-1}AXP = \begin{bmatrix} & & a & \\ & & & -D \\ -a & & & \\ & D & & \end{bmatrix}$, and we are done. Note that a permutation matrix P is real orthogonal. □

The most fascinating aspect of a skew symmetric matrix, however, lies in its connection to rotations.

Proposition 7.11.22. *If A is skew-Hermitian, then e^A is unitary. Conversely, if B is unitary, then $B = \mathrm{e}^A$ for some skew-Hermitian A.*

Proof. Say A is skew-Hermitian. Then $A = Q\begin{bmatrix} a_1\mathrm{i} & & \\ & \ddots & \\ & & a_n\mathrm{i} \end{bmatrix}Q^*$. Then $\mathrm{e}^A = Q\begin{bmatrix} \mathrm{e}^{a_1\mathrm{i}} & & \\ & \ddots & \\ & & \mathrm{e}^{a_n\mathrm{i}} \end{bmatrix}Q^*$ will have unit complex eigenvalues, and hence this is unitary.

Conversely, if B is unitary, then $B = Q \begin{bmatrix} \mathrm{e}^{a_1\mathrm{i}} & & \\ & \ddots & \\ & & \mathrm{e}^{a_n\mathrm{i}} \end{bmatrix} Q^*$, and thus the skew-Hermitian $A = Q \begin{bmatrix} a_1\mathrm{i} & & \\ & \ddots & \\ & & a_n\mathrm{i} \end{bmatrix} Q^*$ would yield $\mathrm{e}^A = B$. □

The case is a bit trickier for the real matrices.

Lemma 7.11.23. $\det(\mathrm{e}^A) = \mathrm{e}^{\mathrm{trace}(A)}$

Proof. Let $x_1, ..., x_n$ be eigenvalues for A. Then $\mathrm{e}^{x_1}, \ldots, \mathrm{e}^{x_n}$ are the eigenvalues of e^A.
So, $\det(\mathrm{e}^A) = \prod \mathrm{e}^{x_i} = \mathrm{e}^{\sum x_i} = \mathrm{e}^{\mathrm{trace}(A)}$. □

Proposition 7.11.24. *If A is real skew-symmetric, then e^A is orthogonal with determinant 1, i.e., a rotation. Conversely, if B is a real orthogonal matrix with $\det(B) = 1$, i.e., a rotation, then $B = \mathrm{e}^A$ for some skew symmetric A.*

Proof. If A is real, by the formula of exponentiation we see that e^A is also real. Since A is skew-Hermitian, e^A is real and unitary, i.e., real orthogonal. Finally, $\det(\mathrm{e}^A) = \mathrm{e}^{\mathrm{trace}(A)} = \mathrm{e}^0 = 1$, this is because all diagonal entries of A must be zero, courtesy of $A^{\mathrm{T}} = -A$.

Conversely, if B is real orthogonal with $\det(B) = 1$, then QBQ^{T} is block diagonal where each block is either 1×1 with value 1, or some 2×2 rotation matrix. (Note that pairs of -1 on the diagonal is also a rotation matrix, and there is no left out -1 eigenvalues, because $\det(B) = 1$.)

Now each rotation matrix $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is the exponentiation $\mathrm{e}^{\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}}$. So we are done. □

Example 7.11.25. Sometimes it is much more geometric to consider the logarithm of a rotation, i.e., the skew-symmetric matrices, rather than the rotation matrix. For a quick example, $\mathrm{e}^{\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, so the skew-symmetric matrix tells you immediately the angle of rotation, whereas the actual rotation matrix looks like a big mess with ugly values coming from sines and cosines.

Consider the skew-symmetric matrix $A = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}$. What is the rotation e^A? Well, let $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. Then first of all you should notice that $A\boldsymbol{v} = \boldsymbol{0}$. As a result, $\mathrm{e}^A\boldsymbol{v} = \mathrm{e}^0\boldsymbol{v} = \boldsymbol{v}$. So $\boldsymbol{v}$ is the axis of rotation!

Now how much is this rotation? Consider the eigenvalues of A, which should be $\theta\mathrm{i}, -\theta\mathrm{i}, 0$ for some real θ. Then the eigenvalues of e^A are $\mathrm{e}^{\theta\mathrm{i}}, \mathrm{e}^{-\theta\mathrm{i}}, 1$, so e^A is a rotation by θ. So we just need to find out what θ is.

Now consider the sum of 2×2 principal minors of A, which gives $(\theta\mathrm{i})(-\theta\mathrm{i}) = \det\begin{bmatrix} 0 & -c \\ c & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} = a^2 + b^2 + c^2 = \|\boldsymbol{v}\|^2$. In particular, $\theta = \pm\|\boldsymbol{v}\|$.

So the direction of $\boldsymbol{v}$ gives the axis of rotation, while the length of $\boldsymbol{v}$ gives the amount of rotation. ☺

7.11.3 Definiteness

Here is also an important application.

Definition 7.11.26. *Given a Hermitian matrix A, we say it is **positive definite** if $\boldsymbol{v}^*A\boldsymbol{v} \geq 0$ with equality if and only if $\boldsymbol{v} = \boldsymbol{0}$. (Can you see why $\boldsymbol{v}^*A\boldsymbol{v}$ must be real?)*

We can have a bunch of similar definitions.

Definition 7.11.27. *Given a Hermitian matrix A,*

1. *we say it is* ***negative definite*** *if $\boldsymbol{v}^* A\boldsymbol{v} \leq 0$ with equality if and only if $\boldsymbol{v} = \boldsymbol{0}$.*
2. *we say it is* ***positive semidefinite*** *if $\boldsymbol{v}^* A\boldsymbol{v} \geq 0$.*
3. *we say it is* ***negative semidefinite*** *if $\boldsymbol{v}^* A\boldsymbol{v} \leq 0$.*
4. *we say it is* ***indefinite*** *if $\boldsymbol{v}^* A\boldsymbol{v}$ could be positive for some $\boldsymbol{v}$ and negative for some other $\boldsymbol{v}$.*

The definition is not new in itself. We have already studied positive definite symmetric matrices when we study inner product spaces. However, we now have a new criterion.

Corollary 7.11.28. *A Hermitian matrix is positive definite if and only if all eigenvalues are positive.*

Proof. Suppose A is Hermitian. Note that $\boldsymbol{v}^* A\boldsymbol{v} = \boldsymbol{v}^* QDQ^*\boldsymbol{v} = \boldsymbol{w}^* D\boldsymbol{w}$, here $\boldsymbol{w} = Q^*\boldsymbol{v}$. Now if $\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$, then $\boldsymbol{w}^* D\boldsymbol{w} = \sum \lambda_i |w_i|^2$.

So if all eigenvalues of A are positive, then $\boldsymbol{v}^* A\boldsymbol{v} = \boldsymbol{w}^* D\boldsymbol{w} = \sum \lambda_i |w_i|^2 \geq 0$, with equality if and only if all $w_i = 0$, if and only if $\boldsymbol{w} = \boldsymbol{0}$, if and only if $Q^*\boldsymbol{v} = \boldsymbol{0}$. Note that Q is unitary, and hence invertible, so equality happens if and only if $\boldsymbol{v} = \boldsymbol{0}$. So A is positive definite.

Conversely, suppose some eigenvalues of A are NOT positive. Say $A\boldsymbol{v} = \lambda\boldsymbol{v}$ for some $\boldsymbol{v} \neq \boldsymbol{0}$ and $\lambda \leq 0$. Then $\boldsymbol{v}^* A\boldsymbol{v} = \lambda\|\boldsymbol{v}\|^2 \leq 0$, so A is NOT positive definite. □

To put this side to side with previous result on positive definiteness, we see that the followings are all equivalent:

1. A is positive definite.
2. All eigenvalues of A are positive.
3. All pivots are positive, i.e., A has LDU decomposition $A = LDL^*$ and the diagonal entries in D are all positive.
4. All leading principal submatrices are positive definite.
5. All leading principal submatrices have positive determinant.
6. All principal submatrices are positive definite.
7. All principal submatrices have positive determinant.
8. We have a lower triangular invertible L with $A = LL^*$.
9. We have $A = BB^*$ for invertible B.

The equivalence between the first three are the most important. The others are not really important. (Except for maybe the last two, as they provide a nice analogy between the positiveness of XX^* and of x^2 for real numbers.)

Similar statements is true for all other (semi)definiteness. For example, the followings are all equivalent:

1. A is positive semi-definite.
2. All eigenvalues of A are non-negative.
3. All pivots are non-negative, i.e., we have LDU decomposition $A = LDL^*$ and the diagonal entries in D are all non-negative.

4. All leading principal submatrices are positive semi-definite.
5. All principal submatrices are positive semi-definite.
6. (Determinants of leading principal matrices are no longer useful. Say $\begin{bmatrix} 0 & \\ & -1 \end{bmatrix}$, then the leading principal minors are all zero (non-negative), yet it is NOT positive semi-definite. It is in fact negative semidefinite, because all pivots are ≤ 0.)
7. All principal matrices have non-negative determinants.
8. We have a lower triangular L with $A = LL^*$. (But we allow L to be non-invertible.)
9. We have $A = BB^*$ for B. (But we allow B to be non-invertible.)

Beware that determinants of leading principal submatrices are no longer useful.

Remark 7.11.29. *(Optional)*
Note that the LDU decompositions for positive semi-definite matrices are a bit tricky. We previously know that when A is INVERTIBLE, then LDU decomposition exists if and only if all leading principal matrices are invertible. Here A could have non-invertible leading principal submatrices. What do we do? We have to start over.

The proof goes like this. Suppose the upper left entry of A is non-zero. Then $A = \begin{bmatrix} a & \boldsymbol{v}^* \\ \boldsymbol{v} & A_1 \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^* \\ \frac{1}{a}\boldsymbol{v} & I \end{bmatrix} \begin{bmatrix} a & \mathbf{0}^* \\ \mathbf{0} & A_1 - \frac{1}{a}\boldsymbol{v}\boldsymbol{v}^* \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{a}\boldsymbol{v}^* \\ \mathbf{0} & I \end{bmatrix}$. Since A is positive semi-definite, therefore $\begin{bmatrix} a & \mathbf{0}^* \\ \mathbf{0} & A_1 - \frac{1}{a}\boldsymbol{v}\boldsymbol{v}^* \end{bmatrix}$ is positive semi-definite, and therefore if it has LDU decomposition, then we can then deduce a corresponding LDU decomposition for A.

If $a = 0$, then I claim that $\boldsymbol{v} = \mathbf{0}$. Suppose not, let us show that A is NOT positive semi-definite. Pick any $\boldsymbol{w} \in \mathbb{C}^{n-1}$ such that $\boldsymbol{v}^*\boldsymbol{w} = -1$. Then consider $\begin{bmatrix} 1 & t\boldsymbol{w}^* \end{bmatrix} A \begin{bmatrix} 1 \\ t\boldsymbol{w} \end{bmatrix} = -2t + t^2\boldsymbol{w}^*A_1\boldsymbol{w}$. Let t be closer and closer to zero from the positive side, then the t^2 term would be ignorable and this would be negative.

So if $a = 0$ and A is positive semi definite, we in fact have $A = \begin{bmatrix} 0 & \mathbf{0}^* \\ \mathbf{0} & A_1 \end{bmatrix}$ and A_1 is positive semi-definite. So again by induction we are done.

For the negative cases, the short answer is that A is negative definite if and only if $-A$ is positive definite, and the same for semi-cases. For example, the followings are all equivalent:

1. A is negative definite.
2. All eigenvalues of A are negative.
3. All pivots are negative. I.e., we have LDU decomposition $A = LDL^*$ and the diagonal entries in D are all negative.
4. The $k \times k$ leading principal submatrix is negative definite if k is odd, and positive definite if k is even.
5. All $k \times k$ principal submatrices are negative definite if k is odd, and positive definite if k is even.
6. The leading principal matrices have alternating determinant signs, i.e., the $k \times k$ leading principal matrix has determinant with sign $(-1)^k$.
7. All $k \times k$ principal matrices have determinant with sign $(-1)^k$.
8. We have a lower triangular invertible L with $A = -LL^*$.
9. We have $A = -BB^*$ for invertible B.

For the negative semidefinite cases, the followings are all equivalent:

1. A is negative semi-definite.
2. All eigenvalues of A are non-positive.
3. All pivots are non-positive. I.e., we have LDU decomposition $A = -LDL^*$ and the diagonal entries in D are all non-positive.
4. All leading principal submatrices are negative semi-definite.
5. All principal submatrices are negative semi-definite.
6. (The determinant of leading principal matrices are no longer useful.)
7. All $k \times k$ principal matrices have non-positive determinant for odd n, and non-negative determinant for even n.
8. We have a lower triangular L with $A = -LL^*$. (But we allow L to be non-invertible.)
9. We have $A = -BB^*$ for B. (But we allow B to be non-invertible.)

Finally, for the indefinite case, the followings are all equivalent:

1. A is indefinite.
2. Some eigenvalues of A are positive and some are negative.
3. Either A has no LDU decomposition, or $A = LDL^*$ where the diagonal entries of D contains both positive and negative values.

In particular, if A is Hermitian but has NO LDU decomposition, then it is indefinite.
Let us see some interesting applications.

Example 7.11.30. Given a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$, how to find a local maximum or minimum? We have the following traditional approach.

1. Solve $f'(x) = 0$ to find critical points, say we find out that $f'(x_0) = 0$.
2. If $f''(x_0) > 0$, then $x = x_0$ is a local minimum.
3. If $f''(x_0) < 0$, then $x = x_0$ is a local maximum.
4. If $f''(x_0) = 0$, then we failed and have no clue what would happen.

Now consider a twice differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$, say $f(x, y) = (x + y + 2)^4 + x^2 + y^2$. Note that its graph would be some surface in $\mathbb{R}^3$. How to find a local maximum or minimum? Note that f has two partial derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} : \mathbb{R}^2 \to \mathbb{R}$, both are functions with two variables. In our example, we have $\frac{\partial f}{\partial x}(x, y) = 4(x + y)^3 + 2x$ and $\frac{\partial f}{\partial y}(x, y) = 4(x + y)^3 + 2y$.

Then we see that f has four second derivatives, because $\frac{\partial f}{\partial x}$ has two derivatives and $\frac{\partial f}{\partial y}$ has two derivatives. We write these as $\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial y \partial x}, \frac{\partial^2 f}{\partial y^2}$. Here the notation $\frac{\partial^2 f}{\partial x \partial y}$ means we take the y-derivative first, and then we take the x-derivative. And the notation $\frac{\partial^2 f}{\partial x^2}$ means we take the x-derivative twice.

Then they form a matrix, the Hessian of f, $H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$. In our case, we have $H(f)(x, y) = \begin{bmatrix} 12(x + y)^2 + 2 & 12(x + y)^2 \\ 12(x + y)^2 & 12(x + y)^2 + 2 \end{bmatrix}$. Hey, this is a real symmetric matrix!

Indeed, your future calculus class would show you that for all real continuously twice-differentiable f, we always have $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$, so $H(f)$ is always symmetric.

Now to find local minimum or local maximum, we do the following approach. (The proof is higher dimensional Taylor expansion, which should be in your future calculus class.)

1. Solve $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ to find critical points, say we find out that one critical point is $x = x_0, y = y_0$.
2. If $H(f)$ at $x = x_0, y = y_0$ is positive definite, then $x = x_0$ is a local minimum.
3. If $H(f)$ at $x = x_0, y = y_0$ is negative definite, then $x = x_0$ is a local maximum.
4. If $H(f)$ at $x = x_0, y = y_0$ is other cases, then we failed and have no clue what would happen.

In our case, solving $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ gives $x = y = 0$, so this is the only critical point. At this point, $H(f) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ is positive definite. So we have a local minimum. ☺

7.11.4 (Optional) Some multivariable calculus

These stuff should be covered in any multivariable calculus. However, since they are related to our class, I will put the proof here to be more self-contained.

Theorem 7.11.31. *For any continuously twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, let $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$, so it is a continuously differentiable function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$. Let Hf be the Hessian of f, so it is a continuous function $Hf : \mathbb{R}^n \to M_{n\times n}$.*

If for some $\boldsymbol{x}_0 \in \mathbb{R}^n$, we have $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $Hf(\boldsymbol{x}_0)$ positive/negative definite, then $\boldsymbol{x}_0$ is a local minimum/maximum of f.

Proof. In lack of calculus materials, this proof is somewhat conceptual and non-rigorous.

Say $n = 2$ and we are looking at a function $f(x, y)$. Then first we increase x by a tiny bit, and our one-variable calculus knowledge tell us that $f(x + \mathrm{d}x, y) = f(x, y) + \mathrm{d}x \frac{\partial f}{\partial x}(x, y) + \frac{1}{2}\,\mathrm{d}x^2 \frac{\partial^2 f}{\partial x^2}(x, y)$ for $\mathrm{d}x \to 0$. (Obviously we are ignoring infinitesimals of degree 3 and above, so we treat things such as $\mathrm{d}x^3$ as zero.)

Now we increase y by a tiny bit. The left hand side is $f(x + \mathrm{d}x, y + \mathrm{d}y)$. On the right hand side, we have three terms.

The first term is $f(x, y + \mathrm{d}y)$ and it can be simplified to $f(x, y) + \mathrm{d}y \frac{\partial f}{\partial y}(x, y) + \frac{1}{2}\,\mathrm{d}y^2 \frac{\partial^2 f}{\partial y^2}(x, y)$.

The second term is $\mathrm{d}x \frac{\partial f}{\partial x}(x, y + \mathrm{d}y)$, and it can be simplified to $\mathrm{d}x \frac{\partial f}{\partial x}(x, y) + \mathrm{d}x\,\mathrm{d}y \frac{\partial^2 f}{\partial y \partial x}(x, y)$.

Finally, the last term is $\frac{1}{2}\,\mathrm{d}x^2 \frac{\partial^2 f}{\partial x^2}(x, y+\mathrm{d}t)$, and it can be simplified to $\frac{1}{2}\,\mathrm{d}x^2 \frac{\partial^2 f}{\partial x^2}(x, y) + \frac{1}{2}\,\mathrm{d}x^2\,\mathrm{d}y \frac{\partial^3 f}{\partial y \partial x^2}(x, y)$. But the new extra term here has infinitesimal degree 3, so we ignore it. Hence this term is actually unchanged.

All in all, if the input changes from (x, y) to $(x, y) + (\mathrm{d}x, \mathrm{d}y)$, the out put is

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

Now use the fact that $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$, we can simplify above into the following version of higher dimensional Taylor expansion approximation, assuming $\mathrm{d}\boldsymbol{x} \to \mathbf{0}$.

$$f(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) = f(\boldsymbol{x}) + (\nabla f(\boldsymbol{x}))^{\mathrm{T}}\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\,\mathrm{d}\boldsymbol{x}^{\mathrm{T}}(Hf(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}.$$

So if $\nabla f(\boldsymbol{x}) = 0$, this simplifies to $f(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\,\mathrm{d}\boldsymbol{x}^{\mathrm{T}}(Hf(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$. This will always be larger than $f(\boldsymbol{x})$ if $\frac{1}{2}Hf(\boldsymbol{x})$ is positive definite, and this will always be smaller than $f(\boldsymbol{x})$ if $\frac{1}{2}Hf(\boldsymbol{x})$ is negative definite. □

As an extra remark, the idea here also proved $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$ by itself. On one hand, we have

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

But on the other hand, by symmetry we also have

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

Hence we indeed have $\frac{\partial^2 f}{\partial y\partial x} = \frac{\partial^2 f}{\partial x\partial y}$.

Alternatively, consider the following exposition for some more intuitions on this fact.

Example 7.11.32. Given a function on two variables $f(x,y)$, it does not has a single derivative function. Rather, it has partial derivatives. Its first derivatives are $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$. Sometimes in calculus people would write this as a vector $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$, called the ***gradient*** of f.

What about the second derivatives? Each of the partitial derivative has two further partial derivatives. So we have four second derivatives, i.e., $\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x\partial y}, \frac{\partial^2 f}{\partial y\partial x}$. Here $\frac{\partial^2 f}{\partial x\partial y} = \frac{\partial}{\partial x}(\frac{\partial f}{\partial y})$. Sometimes in calculus people would write this into a matrix $H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y\partial x} \\ \frac{\partial^2 f}{\partial x\partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$, called the ***Hessian*** of f. This matrix would record all second order differential information about f.

However, an amazing phenomenon here is the fact that, when the function is continuously twice differentiable, then $H(f)$ is ALWAYS a symmetric matrix!

For example, consider $f(x,y) = x^2\mathrm{e}^{xy}$. Then $\frac{\partial^2 f}{\partial y\partial x} = \frac{\partial}{\partial y}(2x\mathrm{e}^{xy} + x^2y\mathrm{e}^{xy}) = 2x^2\mathrm{e}^{xy} + (x^2\mathrm{e}^{xy} + x^3y\mathrm{e}^{xy}) = 3x^2\mathrm{e}^{xy} + x^3y\mathrm{e}^{xy}$, while $\frac{\partial^2 f}{\partial x\partial y} = \frac{\partial}{\partial x}(x^3\mathrm{e}^{xy}) = 3x^2\mathrm{e}^{xy} + x^3y\mathrm{e}^{xy}$. Indeed, the mixed derivatives are the same.

We do not provide a full proof here, but merely gives an intuition as to why this happens. Consider the graph of $f(x,y)$, which is some surface in $\mathbb{R}^3$. Suppose I am standing on this surface. Say I extend my arms on the x direction and lay my arm on the surface, then the slope of my arm records the value of $\frac{\partial f}{\partial x}$ at my location.

Now, I keep my arm in the x direction, and I walk in the y direction. Since my arm is laid on the surface, my arm's slope will change up and down depending on how the surface flows. In particular, if I walk in the y direction, the change in my x-directional arm's slope would be $\frac{\partial^2 f}{\partial y\partial x}$.

Say $\frac{\partial^2 f}{\partial y\partial x} > 0$. Then as I walk in the y-direction, my arm would raise, which indicates that in some positive x and positive y direction, there is a bump, which would raise my arm.

Now suppose I lay my arm in the y direction and walk in the x direction. Then the SAME bump in the positive x and positive y direction would again raise my arm. So we have $\frac{\partial^2 f}{\partial x\partial y} = \frac{\partial^2 f}{\partial y\partial x} > 0$ because it is caused by the very same bump.

(Note that the assumptions that f is continuously twice differentiable is crutial. It means locally the function can be approximated by degree 2 polynomials, i.e., approximately some paraboloids where there is only one bump or pit. Multiple bumps or pits in the positive x and positive y direction would cause this symmetricity to fail.)

Here is also an alternative intuition. Suppose we start with $f(\boldsymbol{v})$. Consider a super tiny square with vertices $\boldsymbol{a} = \boldsymbol{v}, \boldsymbol{b} = \boldsymbol{v} + \mathrm{d}s\boldsymbol{e}_1, \boldsymbol{c} = \boldsymbol{v} + \mathrm{d}t\boldsymbol{e}_2, \boldsymbol{d} = \boldsymbol{v} + \mathrm{d}s\boldsymbol{e}_1 + \mathrm{d}t\boldsymbol{e}_2$. Now you can also see that one should have $\mathrm{d}s\,\mathrm{d}t\frac{\partial^2 f}{\partial x\partial y} = (f(\boldsymbol{d}) - f(\boldsymbol{b})) - (f(\boldsymbol{c}) - f(\boldsymbol{a})) = f(\boldsymbol{a}) + f(\boldsymbol{d}) - f(\boldsymbol{b}) - f(\boldsymbol{c})$. However, we can also see that $\mathrm{d}s\,\mathrm{d}t\frac{\partial^2 f}{\partial y\partial x} = (f(\boldsymbol{d}) - f(\boldsymbol{c})) - (f(\boldsymbol{b}) - f(\boldsymbol{a})) = f(\boldsymbol{a}) + f(\boldsymbol{d}) - f(\boldsymbol{b}) - f(\boldsymbol{c})$. So the two are the same.

Eitherway, we see that the Hessian of a matrix is always symmetric. From the analysis above, you might also notice that $H(f)$ is crucial in analysing the curvature of surfaces, which is very important in geometry or general relativity, which states that gravity is just curvature in space. ☺

7.12 Singular Value Decomposition

We now go back to the world of the real numbers. No more complex numbers. Futhermore, everything in this section should be about linear maps, not linear transformations. I.e., we think of the domain and codomain as DIFFERENT spaces. So we don't have to change basis simultaneously.

7.12.1 Introduction

Recall that for any linear map $A : \mathbb{R}^n \to \mathbb{R}^m$, we can always change basis so that $RAC = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix}$, where r is the rank of A. Now, suppose we only want to perform orthonormal change of basis. Can we still do this?

The Singular Value Decomposition (SVD) is doing exactly this.

Theorem 7.12.1. *For any $m \times n$ real matrix A, we can find $m \times m$ real orthogonal matrix U and $n \times n$ real orthogonal matrix V such that $A = U\Sigma V^{\mathrm{T}}$, where $\Sigma = \begin{bmatrix} D & O \\ O & O \end{bmatrix}$ and D is a real $r \times r$ diagonal matrix with positive diagonal entries, and r is the rank of A.*

In the introduction section, we do not prove this theorem. But first of all, let us get some perspective about why this is useful and important, even for square matrices.

Example 7.12.2. Given a 2×2 matrix A, we know A would map circles in $\mathbb{R}^2$ into ellipses. However, what is the resulting eccentricity fo these ellipses?

We have already seen that when $A = XDX^{-1}$ with $X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, D = \begin{bmatrix} 2 & \\ & 3 \end{bmatrix}$, then the unit circle would be mapped to an ellipse whose major half-axis has length > 3 and minor half-axis has length < 2. Where is the direction and length of the major half-axis and the minor half-axis?

Suppose now we have a decomposition $A = U\Sigma V^{\mathrm{T}}$, note that in this case, Σ is a genuin diagonal matrix. Now the U, V are both rotations or reflections, and Σ is a stretching along the coordinate axis. So V^{T} would do nothing to the unit circle. Then Σ would stretch the unit circle along coordinate axis to get an ellipse. Finally U rotate/reflect the ellipse. As a result, it is very easy to see that the half-axis lengths are exactly the diagonal entries of Σ. These values are called the singular values of A. As you can see, eigenvalues do NOT define the shape of the resulting ellipse. Only singular values could define their shapes.

A related question is this: suppose I squeeze a ping pong ball in several directions, where will the ping pong ball most likely crack? This is slightly more complicated, but the idea is similar: the forces are not necessarily orthogonal, but their combined action will push the ping pong ball to deform into elliptic shape (ellipsoid). Singluar value decomposition will help us identify the shape of this ellipsoid, and the shape of deformation would tell us where the ping pong ball would crack. ☺

Example 7.12.3. The singular value decomposition is also related to other decompositions.

Suppose we have $A = U\Sigma V^{\mathrm{T}}$ for a square matrix A. Writing this decomposition as $A = (U\Sigma U^{\mathrm{T}})(UV^{\mathrm{T}})$. This is called a polar decomposition, since $U\Sigma U^{\mathrm{T}}$ must have real (and in fact non-negative) eigenvalues, and UV^{T} must have eigenvalues that are unit complex numbers. This is also a matrix analogue of the complex number polar decomposition $z = r\mathrm{e}^{\mathrm{i}\theta}$. ☺

Now let us finally consider this decomposition from a rank perspective. In many applicaitions of linear algebra, it is veru beneficial to write a rank r matrix as the sum of r matrices of rank one.

Example 7.12.4. Consider the France map. If we use number 1 for blue, 2 for red, and 0 for white, then the France map is something like $M = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 & 2 & \dots & 2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & 1 & 0 & \dots & 0 & 2 & \dots & 2 \end{bmatrix}$. If such a matrix is $m \times n$, then we would need to store mn numbers in the computer.

However, note that such a matrix must have rank one. We must have $M = \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ where $\boldsymbol{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ and $\boldsymbol{v} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 & 2 & \dots & 2 \end{bmatrix}$. And to store $\boldsymbol{u}, \boldsymbol{v}$, we only need $m + n$ numbers. This way, we saved a lot of memory space in our computer!

Consider the Benin flag. If 3 is green and 4 is yellow, then the 2×3 Benin flag would look like $\begin{bmatrix} 3 & 4 & 4 \\ 3 & 2 & 2 \end{bmatrix}$. In general, the $m \times n$ version of this flag always have rank 2, and we have $\begin{bmatrix} 3 & 4 & 4 \\ 3 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 4 & 4 \\ 0 & 2 & 2 \end{bmatrix}$,

the sum of two rank one matrices. Since each rank one matrix must be $\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ for some $\boldsymbol{u}, \boldsymbol{v}$, we see that we only need to store $2(m+n)$ numbers instead of mn numbers.

In general, for an $m \times n$ matrix with rank r, we do not store mn numbers in the computer. Rather, we would only store $r(m+n)$ numbers.

You can also check out the Greece flag, which has rank three. I have not found any rank four flag yet, so please let me know if you find any. Finally, the Chinese flag and the American flag both have infinite rank. ☺

The example above does not involve SVD yet. However, you can think of SVD as the best rank one decomposition.

Consider the nature of SVD for an $m \times n$ matrix $A = U\Sigma V^{\mathrm{T}}$. Suppose $U = \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_m \end{bmatrix}, V = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}$, and $\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & O \end{bmatrix}$ for nonzero $\sigma_1, \dots, \sigma_r$. Since U, V are invertible, it is easy to see that A has rank exactly r. Furthermore, $A = \sum \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. So essentially, we are decomposing A into a linear combination of rank one matrices.

Furthremore, consider any two such rank one matrices, say $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}, \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^{\mathrm{T}}$ for any $i \neq j$. Then we would have $\boldsymbol{u}_i \perp \boldsymbol{u}_j$ and $\boldsymbol{v}_i \perp \boldsymbol{v}_j$ by construction. So the two rank one matrices are "as orthogonal to each other as possible". So SVD is decomposing A into a linear combination of "mutually orthogonal" rank one matrices!

7.12.2 The fundation of SVD

How can we obtain the singular value decomposition? Again, note that A is NOT necessarily square.

Let us do some backward observation. Suppose $A = U\Sigma V^{\mathrm{T}}$ is possible, with $\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & O \end{bmatrix}$.

We usually assume that $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

A very important observation is that $A^{\mathrm{T}}A = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$. Note that $\Sigma^{\mathrm{T}}\Sigma$ is diagonal and square, and V is orthogonal, so this is in fact the spectral decomposition for the symmetric matrix $A^{\mathrm{T}}A$. So to get V, you can simply find all eigenvectors for $A^{\mathrm{T}}A$ and you are done. Furthermore, we have $\Sigma^{\mathrm{T}}\Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ & \ddots & & \\ & & \sigma_r^2 & \\ & & & O \end{bmatrix}$.

So these $\sigma_1, \dots, \sigma_r$ in Σ are simple the square roots of the non-zero eigenvalues of $A^{\mathrm{T}}A$. Fortunately, we know $A^{\mathrm{T}}A$ must always be positive semidefinite, so the square roots are all real.

In short, the spectral decomposition of $A^{\mathrm{T}}A$ will give you both Σ and V. Similarly, the spectral decomposition of AA^{T} will give you both U and Σ. Is this enough to find the whole SVD? Well, there is one more missing ingredient.

Example 7.12.5. Consider $A = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$ and $B = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}$. Then $AA^{\mathrm{T}} = BB^{\mathrm{T}}$ and $A^{\mathrm{T}}A = B^{\mathrm{T}}B$. However, A, B are different matrices with different SVD.

In particular, in terms of rank decomposition, the SVD of A is $A = \boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + \boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}}$, while the SVD for B is $B = \boldsymbol{e}_2\boldsymbol{e}_1^{\mathrm{T}} + \boldsymbol{e}_1\boldsymbol{e}_2^{\mathrm{T}}$. As you can see, the U portion and V portion of the SVD for A and B uses the same columns, but their MATCHING is different.

AA^{T} tells you what columns U should have, and $A^{\mathrm{T}}A$ tells you what columns V should have, but they give no information on how to order these things to match each other. It turns out that the order of columns for U, V are very important. Recall the decomposition $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, where the $\boldsymbol{u}_i, \boldsymbol{v}_i$ are in fact paired up. So the order of columns of U and of V must MATCH each other. ☺

A very interesting observation is that we also see that $A^{\mathrm{T}}A$ and AA^{T} has the same positive eigenvalues $\sigma_1^2, \ldots, \sigma_r^2$. Can you see why without SVD? (Hint: How are the eigenvalues of AB and BA related? Note that one matrix is $m \times m$ and the other is $n \times n$, they do not have the same amounts of eigenvalues.)

Definition 7.12.6. *A **singluar value** $\sigma > 0$ for A is the square root of a non-zero eigenvalue of $A^{\mathrm{T}}A$ or AA^{T}.*

*Given an $m \times n$ matrix A, we say $\boldsymbol{v} \in \mathbb{R}^n$ is a **right singular vector** if it is an eigenvector for $A^{\mathrm{T}}A$. (This is related to the V in $A = U\Sigma V^{\mathrm{T}}$, hence it is "right".)*

We say it is a singular vector of A for the singluar value $\sigma > 0$ if it is an eigenvector of $A^{\mathrm{T}}A$ for the eigenvalue σ^2.

*We say $\boldsymbol{u} \in \mathbb{R}^m$ is a **left singular vector** if it is an eigenvector of AA^{T}. (This is related to the U in $A = U\Sigma V^{\mathrm{T}}$, hence it is "left".)*

We say it is a singular vector of A for the singluar value $\sigma > 0$ if it is an eigenvector of AA^{T} for the eigenvalue σ^2.

How to work out the correspondence between the left and right singular vectors?

Again let us do some backward deduction. Consider $A\boldsymbol{v}_i$. Since all other $\boldsymbol{v}_j$ are orthogonal to $\boldsymbol{v}_j$, via the decomposition $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$ we see that in fact $A\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$ for $i \leq r$. So actually $\boldsymbol{v}_i$ gives us $\boldsymbol{u}_i$ right away when $i \leq r$. Similarly, we can also see that $\boldsymbol{u}_i^{\mathrm{T}} A = \sigma_i \boldsymbol{v}_i^{\mathrm{T}}$.

Lemma 7.12.7 (Ping Pong of left/right singular vectors)**.** *If $\boldsymbol{v}$ is a right singular vector of A for the singular value σ, then $A\boldsymbol{v}$ is a left singular vector of A for the singular value σ.*

If $\boldsymbol{u}$ is a left singular vector of A for the singular value σ, then $A^{\mathrm{T}}\boldsymbol{u}$ is a right singular vector of A for the singular value σ.

Proof. Suppose $\boldsymbol{v}$ is a right singular vector for the singular value σ. Now $(AA^{\mathrm{T}})(A\boldsymbol{v}) = A(A^{\mathrm{T}}A\boldsymbol{v}) = A(\sigma^2\boldsymbol{v}) = \sigma^2 A\boldsymbol{v}$. Hence $A\boldsymbol{v}$ is a left singular vector of A for the singular value σ. The other statement is proven similarly. □

Lemma 7.12.8. *For a right singular vector $\boldsymbol{v}$ for the singular value σ, $\|A\boldsymbol{v}\| = \sigma\|\boldsymbol{v}\|$. If $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthogonal right singular vectors, then $A\boldsymbol{v}_1, A\boldsymbol{v}_2$ are orthogonal left singular vectors.*

Proof. For a right singular vector $\boldsymbol{v}$ for the singular value σ, $\|A\boldsymbol{v}\|^2 = \boldsymbol{v}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{v} = \sigma^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$.

If $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthogonal right singular vectors, then $\langle A\boldsymbol{v}_1, A\boldsymbol{v}_2\rangle = \boldsymbol{v}_1^{\mathrm{T}}(A^{\mathrm{T}}A\boldsymbol{v}_2) = \sigma^2\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2$ for some $\sigma > 0$. Since $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthogonal, the calculation above gives zero. □

Theorem 7.12.9 (Singular Value Decomposition)**.** *For any real $m \times n$ matrix A of rank r, we can find $m \times m$ orthogonal matrix $U = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_m \end{bmatrix}$, and $n \times n$ orthogonal matrix $V = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix}$, and "diagonal" $m \times n$ matrix $\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & O \end{bmatrix}$ with positive $\sigma_1, \ldots, \sigma_r$, such that $A = U\Sigma V^{\mathrm{T}}$.*

Here $\boldsymbol{u}_i$ is a left singular vector for singular value σ_i for $i \leq r$, and $\boldsymbol{u}_i$ is in $\mathrm{Ker}(A^{\mathrm{T}})$ if $i > r$. Similarly, $\boldsymbol{v}_i$ is a right singular vector for singular value σ_i for $i \leq r$, and $\boldsymbol{v}_i$ is in $\mathrm{Ker}(A)$ if $i > r$.

Proof. Since eigenvalues of $A^{\mathrm{T}}A$ which are all non-negative, and $A^{\mathrm{T}}A$ have the same rank as A, let us order the eigenvalues from large to small as $\sigma_1^2 \geq \cdots \geq \sigma_r^2 > 0$. Then we have spectral decomposition $A^{\mathrm{T}}A = VDV^{\mathrm{T}}$ where V is orthogonal and D is diagonal, with diagonal entries $\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0$. And columns of V are right singular vectors by definition.

Let $\boldsymbol{u}_i = \frac{1}{\sigma_i} A\boldsymbol{v}_i$ for all $i \leq r$. Now we have mutually orthogonal unit vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$, we extend this arbitrarily to an orthonormal basis for the codomain $\mathbb{R}^m$. So now we have $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, and $U = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_m \end{bmatrix}$ is an orthogonal matrix.

Now $A\begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_r & \boldsymbol{v}_{r+1} & \ldots & \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} \sigma_1\boldsymbol{u}_1 & \ldots & \sigma_r\boldsymbol{u}_r & O \end{bmatrix} = U\Sigma$. So $AV = U\Sigma$. So we are done. □

Alternative proof. This proof is a bit more elegant, but lacks some finesse.

Again start by performing any spectral decomposition $A^{\mathrm{T}}A = VDV^{\mathrm{T}}$. Then we see that $(AV)^{\mathrm{T}}(AV) = D$. In particular, columns of AV are mutually orthogonal. Then by a generalized version of QR decomposition, we have $AV = U\Sigma$ where U is orthogonal and Σ is "diagonal" as desired. (We obmit this proof of the generalized QR decomposition. It is strikingly similar to the last proof....) □

Remark 7.12.10. *Note that $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ are all in* $\mathrm{Ran}(A)$ *since $\boldsymbol{u}_i = \frac{1}{\sigma_i} A\boldsymbol{v}_i$, and $\boldsymbol{u}_{r+1}, \dots, \boldsymbol{u}_m$ are all in* $\mathrm{Ker}(A^{\mathrm{T}})$.

Given that the two subspaces are orthogonal complements, and given their dimensions, it is not hard to see that $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ in fact form an orthonormal basis for $\mathrm{Ran}(A)$*, while $\boldsymbol{u}_{r+1}, \dots, \boldsymbol{u}_m$ form an orthonormal basis for* $\mathrm{Ker}(A^{\mathrm{T}})$.

The same thing is true for $\boldsymbol{v}_i$ and the subspaces $\mathrm{Ran}(A^{\mathrm{T}}), \mathrm{Ker}(A)$*. In particular, SVD means simultaneously finding the BEST orthonormal basis for ALL subspaces related to A.*

A common geometric interpretation of SVD is this: $A = U\Sigma V^{\mathrm{T}}$ means A is rotation, then stretch, and the rotation. However, here is another one.

Corollary 7.12.11 (Polar decomposition)**.** *For any square matrix A, we have $A = SQ$ where S is symmetric positive semi-definite and Q is orthogonal.*

Proof. $A = U\Sigma V^{\mathrm{T}} = (U\Sigma U^{\mathrm{T}})(UV^{\mathrm{T}})$. □

This is akin to $z = re^{\mathrm{i}\theta}$ for complex numbers, where $r \geq 0$ and $e^{\mathrm{i}\theta}$ is unit complex. Recall that a symmetric matrix is just a stretch along orthogonal frames. So the geometric action of A is simply a rotation, and the stretch along orthogonal frams (where the stretching factors are the singular values, and the stretching directions are columns of U). In particular, we have the following.

Corollary 7.12.12. *If an $n \times n$ matrix A acts on a unit n-dim sphere, the result is an n-dim ellipsoid whose half-axes have length the same as singular values of A, and the direction of these axes are the same as columns of U for the SVD $A = U\Sigma V^{\mathrm{T}}$.*

In this sence, if we are thinking of A as a linear transformation, then usually U is more prominent in its interpretation.

Example 7.12.13. Consider a Ping Pong ball squeezed by 10 different forces. Where on the ball would crack first?

Well, under these 10 forces, the Ping Pong ball would be "deformed", and the location with the most deformation would be the most likely to crack. The deformed Ping Pong ball could imaginably be some ellipsoid, and the location with most deformation could be determined by the direction of the half-axes of the ellipsoid. As you can imagine, this must be related to SVD some how!

Skipping some details on mechanical analysis involving stress tensor, if the 10 forces are $\boldsymbol{F}_1, \dots, \boldsymbol{F}_{10}$, then we can set $A = \begin{bmatrix} \frac{\boldsymbol{F}_1}{\sqrt{\|\boldsymbol{F}_1\|}} & \cdots & \frac{\boldsymbol{F}_{10}}{\sqrt{\|\boldsymbol{F}_{10}\|}} \end{bmatrix}$. The SVD for A would yeild these half-axes and the amount of corresponding deformation.

Now instead of a Ping Pong ball under pressure, imagine your head, which is under pressure due to all the classes you are taking. The math class, the science class, the literary class, they each gives you some different pressure. Which class would you most likely fail? I don't know, but I'm sure SVD is involved in this somehow. ☺

We finally end this section with a WARNING: While the singular values and singular vectors are all well defined for A, the decomposition of SVD might NOT be unique. Consider $I = UIU^{\mathrm{T}}$ for any orthogonal matrix U. We see that the U portion is arbitrary.

7.12.3 (Optional) Pseudo Inverse

Given $A = U\Sigma V^{\mathrm{T}}$, where $\Sigma = \begin{bmatrix} D & O \\ O & O \end{bmatrix}$ is an $m \times n$ matrix and D is invertible and diagonal, then we define $\Sigma^+ = \begin{bmatrix} D^{-1} & O \\ O & O \end{bmatrix}$ which is $n \times m$, and we define $A^+ = V\Sigma^+ U^{\mathrm{T}}$ to be the ***pseudo-inverse*** of A.

Proposition 7.12.14. *AA^+ is the orthogonal projection to* $\mathrm{Ran}(A)$.

Proof. Note that $AA^+ = U\Sigma\Sigma^+ U^{\mathrm{T}}$, and $\Sigma\Sigma^+ = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix}$ is an $m \times m$ square diagonal matrix. Therefore AA^+ is obviously symmetric, and $(AA^+)^2 = U(\Sigma\Sigma^+)^2 U^{\mathrm{T}} = U\Sigma\Sigma^+ U^{\mathrm{T}} = AA^+$. So this is an orthogonal projection.

We also see that U represents a change of basis in the product $U\Sigma\Sigma^+ U^{\mathrm{T}}$, and $\Sigma\Sigma^+ = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix}$ means a truncation taking only the first r coordinates. So, AA^+ is a projection to $\mathrm{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r) = \mathrm{Ran}(A)$. □

A very nice aplication is the following. Given a subspace V of $\mathbb{R}^n$, each vector $\boldsymbol{v} \in V$ originally has n coordinates, because it in $\mathbb{R}^n$. However, once we find a basis for V, then $\boldsymbol{v}$ as a vector of V will now have k coordinates under this basis. What is the transition map?

Corollary 7.12.15. *Suppose we have a subspace V in $\mathbb{R}^n$ spanned by linearly independent $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^n$. Then for any $\boldsymbol{w} \in V$, its coordinate under the basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ of V is $A^+\boldsymbol{w}$, where $A = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_k \end{bmatrix}$.*

Proof. If $\boldsymbol{w} \in V$, then $\boldsymbol{w} = AA^+\boldsymbol{w}$, so $\boldsymbol{w}$ is the linear combination of columns of A according to coefficients $A^+\boldsymbol{w}$. □

7.12.4 Low Rank Approximation

Consider the rank one decomposition induced by SVD, $A = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. Recall that we order the singular values so that $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Intuitively, you can think of these $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$ as components of A. Obviously, $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$ is the most important component, and $\sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$ is the second most important component, and so on. Finally, the orthogonality between these components means they are recording some independent aspects of A.

Example 7.12.16. Suppose I have a bunch of data to store in the computer. The data are in the form of an $m \times n$ matrix A, say maybe it is a gray-scale picture. Then I would need to store mn numbers in the computer. But I do not have enough data storage to store these many numbers. What should I do?

As we have noted before, if A have a small rank r, then any rank one decomposition $A = \sum \boldsymbol{x}_i \boldsymbol{y}_i^{\mathrm{T}}$ will store A using only $r(m+n)$ numbers. This is super good when r is tiny.

What if A have big rank? Then to store everything, you have no choice but to store all mn numbers. However, maybe we can find a low rank matrix B, such that B is super close to A, and store B instead? The information in A will be lost somewhat, but B should give a good approximation, and it needs much less space to store.

This is the problem of low rank approximation. Fix a small integer k. Given any matrix A, we want to find the best rank k matrix that is "closest" to A. What is the answer?

Well, here is the answer. If the SVD gives $A = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, then the best rank k approximation is $\sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. (Recall that we order the singular values so that $\sigma_1 \geq \cdots \geq \sigma_r > 0$. So we are simply taking the largest k rank one components of SVD, and we are done.)

Say you are preparing for an exam. There is a lot to memorize, but your memory is finite and probably not enough. What should you do? Ideally, you should perform SVD on all the materials to be memorized, and memorize only the rank one components corresponding to the largest few singular values. This would allow you to most efficiently approximate all the materials given your limited memory capacity. If your teacher tells you some "key points" to memorize in the textbook, those are probably $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$, the most important portion to memorize.

Also notice how orthogonality is the key to efficiency. If you memorize two sentences that are essentially saying the same thing, then you wasted some memory space. The orthogonality of SVD guarantees that each $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$ contains totally independent informations from others, so there is no waste whatsoever. ☺

Now, we want to find the best low rank matrix that is closest to A. But how to measure closeness? One common way is the Frobenius norm.

Definition 7.12.17. *Given two $m \times n$ real matrices A, B, we can perform their inner product such that $\langle A, B\rangle = \sum a_{ij} b_{ij}$, i.e., this is simply entry-wise "dot product". Then the Frobenius norm of the matrix A is simply $\|A\|_F = \sqrt{\langle A, A\rangle}$.*

Lemma 7.12.18. *$\langle A, B\rangle = \mathrm{trace}(A^{\mathrm{T}} B)$, and $\|A\|_F^2 = \mathrm{trace}(A^{\mathrm{T}} A)$.*

Let us have another application of low rank approximation: lines of best fit.

Example 7.12.19. Consider data points $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \dots, \begin{bmatrix} x_n \\ y_n \end{bmatrix}$ on the plane $\mathbb{R}^2$. Find the best line to fit the data.

Now previously, we have a least square method, which is trying to find the best k, b to minimize $\sum (y_i - kx_i - b)^2$. In short, we are minimizing the ***vertical distance*** to the line. The least square is used when we are trying to PREDICT y using x, hence vertical distance to the line is more important.

However, sometimes we are not really interested in prediction, and we only want to look at correlation. So I want to minimize the ***orthogonal distance*** to the line. This is called ***the line of best fit***. What should I do then?

First we shift the data so that the average values are $\mathbb{E}(x) = \mathbb{E}(y) = 0$, i.e., my data is centered around the origin. This way, I just need to find the best line through the origin to fit the data.

Consider any line through the origin L. Then given data points $A = \begin{bmatrix} \boldsymbol{p}_1, \dots, \boldsymbol{p}_n \end{bmatrix}$ which is a $2 \times n$ matrix, we can project each point orthogonally to L, and obtain results $B = \begin{bmatrix} \boldsymbol{q}_1, \dots, \boldsymbol{q}_n \end{bmatrix}$. Since each point $\boldsymbol{p}_i$ is projected to the point $\boldsymbol{q}_i$ on L, we see that the orthogonal distance from the points in A to the line L is $\sum \|\boldsymbol{p}_i - \boldsymbol{q}_i\|^2 = \mathrm{trace}((A-B)^{\mathrm{T}}(A-B)) = \|A - B\|_F$. So this is simply the Frobenius distance between A and B.

Also note that since all $\boldsymbol{q}_i$ lies on the same line L through the origin, B has parallel columns, and it has rank one. So in short, to find the best line L, I just need to find the best rank one matrix B such that $\|A - B\|_F$ is as small as possible. B should be a best rank one approximation to A.

So, if the SVD gives $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, the best rank one approximation is $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$, so this is what B should be. $L = \mathrm{Ran}(B)$ so it is spanned by $\boldsymbol{u}_1$. We have found our line of best fit.

Now suppose the data points are $A = \begin{bmatrix} \boldsymbol{p}_1, \dots, \boldsymbol{p}_n \end{bmatrix}$ but the points are now in $\mathbb{R}^3$. Again suppose $\mathbb{E}(x) = \mathbb{E}(y) = \mathbb{E}(z) = 0$. How to find the line of best fit? Again you just need the best rank one approximation of A. What if you want to find a plane to best fit the data? Well, you just need the best rank two approximation of A.

So this is why SVD is super important in statistics. ☺

7.12.5 Matrix norms and proofs of low rank approximation

We now provide the proof for low rank approximation. First, let us introduce another matrix norm, which is also super useful, especially in physics

Consider a linear map A acting on the unit circle. This means we have an input $\boldsymbol{u}$, some unit vector, and we are interested in $A\boldsymbol{u}$. Since we know the result is an ellipse, how to find the length of the major half-axis of the ellipse? Essentially, we are asking ourselves how to find $\max \|A\boldsymbol{u}\|$.

Definition 7.12.20. *Given a matrix A, its **operator norm** is $\|A\| = \max_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \max_{\|\boldsymbol{u}\|=1} \|A\boldsymbol{u}\|$. In particular, this is the maximal output-input length ratio.*

This is called a norm for a reason. For example, here is the triangle inequality:

Lemma 7.12.21. *$\|A + B\| \leq \|A\| + \|B\|$.*

Proof. For any unit vector $\boldsymbol{u}$, we have $\|(A+B)\boldsymbol{u}\| = \|A\boldsymbol{u}+B\boldsymbol{u}\| \leq \|A\boldsymbol{u}\| + \|B\boldsymbol{u}\| \leq \|A\| + \|B\|$. Since this is true for all $\|(A+B)\boldsymbol{u}\|$, it is also true for their maximum value $\|A+B\|$. □

The main advantage of this norm over the Frobenius norm is that it can be easily generalized to infinite dimensional settings. This is especially useful in physics.

Now, we are interested in the output-input length ratio $\frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$. If you square this, this is $\frac{\boldsymbol{v}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}$. This is called a Rayleigh quotient.

Definition 7.12.22. *Given a symmetric matrix S, a Rayleigh quotient is a quotient $\frac{\boldsymbol{v}^{\mathrm{T}}S\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}$ for some $\boldsymbol{v}$.*

Proposition 7.12.23. *The Rayleigh quotient of S is always between the largest eigenvalue of S and the smallest eigenvalue of S.*

Proof. $a \leq \frac{\boldsymbol{v}^{\mathrm{T}}S\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \leq b$ for all $\boldsymbol{v}$
if and only if $a\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} \leq \boldsymbol{v}^{\mathrm{T}}S\boldsymbol{v} \leq b\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$ for all $\boldsymbol{v}$,
if and only if $\boldsymbol{v}^{\mathrm{T}}(aI-S)\boldsymbol{v} \leq 0 \leq \boldsymbol{v}^{\mathrm{T}}(bI-S)\boldsymbol{v}$ for all $\boldsymbol{v}$,
if and only if $aI-S$ is negative semidefinite and $bI-S$ is positive semidefinite,
if and only if all eigenvalues of S are between a and b. □

Now we connect the operator norm to singular values.

Proposition 7.12.24. *The operator norm of A is its largest singular value σ_1.*

Proof. Since $\|A\boldsymbol{v}_1\| = \sigma_1\|\boldsymbol{v}_1\|$, we see that $\|A\| \geq \sigma_1$.
For any $\boldsymbol{v} \neq \boldsymbol{0}$, $\frac{\|A\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2} = \frac{\boldsymbol{v}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \leq \lambda_1(A^{\mathrm{T}}A) = \sigma_1^2$. So $\|A\| \leq \sigma_1$. □

Alternative proof. First you can prove that $\|UAV\| = \|A\|$, because orthogonal matrices preserve length. Now we do SVD and have $\|A\| = \|U\Sigma V^{\mathrm{T}}\| = \|\Sigma\|$. Now Σ is a very specific matrix and you can simply calculate directly. □

As a notation, we shall now use $\|A\|$ and $\sigma_1(A)$ interchangably, as they refer to the same thing.

Corollary 7.12.25. *Let A acts on an n-dimensional sphere, then the resulting high dimensional ellipsoid has major half-axis length σ_1. (Note that depending on A being injective or not, the resulting ellipsoid might NOT be n dimensional.)*

Remark 7.12.26. *From the ellipsoid result, one can also see the following: If, instead of looking at $\max \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$, we look at $\max_{\boldsymbol{v}\perp\boldsymbol{v}_1} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$ where $\boldsymbol{v}_1$ is the right singular vector for σ_1, then we would get σ_2, i.e. the largest length in the cross section perpendicular to the major half-axis of the ellipsoid. This can be generalized for all singular values.*

Here is a related potential application of Rayleigh quotient.

Example 7.12.27. How to find the maximum and minimum of $\frac{2x^2+4xy+5y^2}{x^2+y^2}$, given that x, y are not both zero?

Well, set $S = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$, then this is a Rayleigh quotient for S. Since S has eigenvalues $1, 6$, this Rayleigh quotient must be between 1 and 6. ☺

Now we go for low rank approximation.

Definition 7.12.28. *Suppose the SVD of A gives $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. The **rank k truncated SVD** of A for $k \leq r$ is $A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$.*

As we shall see A_k will minimize both the operator norm and the Frobenius norm of $A - A_k$. We do the operator norm first.

Proposition 7.12.29. *Among the set of rank k matrices, A_k is one of the closest to A in terms of operator norm. Specifically, $\|A-B\| \geq \sigma_{k+1}(A)$ for all $m \times n$ rank k matrices B, and $\|A - A_k\|$ reaches this minimal value.*

(Here $\sigma_{k+1}(A)$ refers to the $k+1$-th singular value of A, and if A is rank k, $\sigma_{k+1}(A) = 0$.)

Proof. That A_k gives $\|A - A_k\| = \sigma_{k+1}(A)$ is obvious.

If $k = r$, then the statement is trivial as $A_k = A$. Let us assume $k < r$.

(Before the formal proof process, let us do some analysis. We want to show that $\|A-B\| \geq \sigma_{k+1}(A)$. To achieve this, we need to find a unit vector $\boldsymbol{x}$ such that $\|(A-B)\boldsymbol{x}\| \geq \sigma_{k+1}(A)$. So given fixed B, we want to pick unit vector $\boldsymbol{x}$ to make $A\boldsymbol{x}$ and $B\boldsymbol{x}$ to be as far away as possible. As B has low rank, $B\boldsymbol{x}$ is very likely to be zero, in which case we just want $\|A\boldsymbol{x}\|$ to be as large as possible. This would happen if $\boldsymbol{x}$ is a mixture of the singular vectors of A for large singular values.)

To the formal proof now. Suppose $A = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, where the singular values are ordered from large to small as usual. Let $V_{k+1} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_{k+1} \end{bmatrix}$, so these are singular vectors of A for the $(k+1)$ largest singular values. Obviously $\dim \operatorname{Ran}(V_{k+1}) = k+1$, as all these singular vectors are orthonormal.

On the other hand, $\operatorname{Ker}(B) = n - \operatorname{rank}(B) = n - k$. So by inclusion-exclusion principle for subspaces, $\dim(\operatorname{Ker}(B) \cap \operatorname{Ran}(V_{k+1})) \geq 1$. So we pick $\boldsymbol{x}$ to be a unit vector in $\operatorname{Ker}(B) \cap \operatorname{Ran}(V_{k+1})$.

Now $B\boldsymbol{x} = \boldsymbol{0}$, and $\|(A-B)\boldsymbol{x}\| = \|A\boldsymbol{x}\|$. If $\boldsymbol{x} = a_1\boldsymbol{v}_1 + \cdots + a_{k+1}\boldsymbol{v}_{k+1}$, then $A\boldsymbol{x} = \sigma_1 a_1 \boldsymbol{u}_1 + \cdots + \sigma_{k+1} a_{k+1} \boldsymbol{u}_{k+1}$. Since these $\boldsymbol{u}_i$ are orthonormal, we see that $\|A\boldsymbol{x}\| = \sqrt{\sigma_1^2 a_1^2 + \cdots + \sigma_{k+1}^2 a_{k+1}^2} \leq \sqrt{\sigma_{k+1}^2 a_1^2 + \cdots + \sigma_{k+1}^2 a_{k+1}^2} = \sigma_{k+1}\sqrt{\sum a_i^2}$.

However, since $\boldsymbol{x}$ is a unit vector, and $\boldsymbol{x} = a_1\boldsymbol{v}_1 + \cdots + a_{k+1}\boldsymbol{v}_{k+1}$ for orthonormal $\boldsymbol{v}_i$, therefore $\sum a_i = \|\boldsymbol{x}\|^2 = 1$. So we are done. □

Now we move on to Frobenius norm, which is also related to singular values.

Proposition 7.12.30. $\|A\|_F^2 = \sum \sigma_i^2$.

Proof. $\|A\|_F^2 = \operatorname{trace}(A^{\mathrm{T}}A)$, which is the sum of eigenvalues of $A^{\mathrm{T}}A$. □

We have already shown that A_k is the best rank k approximation to A in the operator norm. Let us now show that it is also the best rank k approximation to A in the Frobenius norm.

Lemma 7.12.31 (Generalized triangle inequality for the operator norm)**.** *If $A = B+C$, then $\sigma_i(B)+\sigma_j(C) \geq \sigma_{i+j-1}(A)$. (Note that the regular triangle inequality is the case when $i = j = 1$.)*

Proof. Let B_{i-1}, C_{j-i} be the corresponding rank approximation in the operator norm. Then $\sigma_i(B) = \sigma_1(B - B_{i-1}), \sigma_j(C) = \sigma_1(C - C_{j-1})$. So by the regular triangle inequality of operator norm, $\sigma_i(B)+\sigma_j(C) = \sigma_1(B - B_{i-1}) + \sigma_1(C - C_{j-1}) \geq \sigma_1(B + C - B_{i-1} - C_{j-1}) = \sigma_1(A - B_{i-1} - C_{j-1})$.

However, $B_{i-1} + C_{j-1}$ is a rank $i + j - 2$ matrix. So by the rank approximation for operator norm, $\sigma_1(A - B_{i-1} - C_{j-1}) \geq \sigma_1(A - A_{i+j-2}) = \sigma_{i+j-1}(A)$. So we are done. □

Proposition 7.12.32. *Among the set of rank k matrices, A_k is one of the closest to A in terms of Frobenius norm. Specifically, $\|A - B\|_F \geq \sum_{i=k+1}^n [\sigma_i(A)]^2$ for all $m \times n$ rank k matrices B, and $\|A - A_k\|$ reaches this minimal value.*

(Here $\sigma_i(A)$ refers to the i-th singular value of A, and if A is rank r, $\sigma_i(A) = 0$ for all $i > r$.)

Proof. Consider the decomposition $A = B + (A - B)$. Then $\sigma_i(B) + \sigma_j(A - B) \geq \sigma_{i+j-1}(A)$. Since B has rank k, we see that for $i = k+1$, $\sigma_i(B) = 0$, and hence $\sigma_j(A - B) \geq \sigma_{i+j-1}(A) = \sigma_{j+k}(A)$.

Now $\|A - B\|_F = \sum_{j=1}^n [\sigma_j(A - B)]^2 \geq \sum_{j=1}^n [\sigma_{j+k}(A)]^2 = [\sigma_{k+1}(A)]^2 + \cdots + [\sigma_n(A)]^2$.

Finally, this minimum is reached when $B = A_k$. □

In fact, for all norms defined in terms of singular values, then the operator norm would provide a control over that norm.

7.12.6 Principal Component Analysis

Example 7.12.33. Suppose we want to do facial recognition. Given a person's face, we can store all information about this face into a big list of numbers, i.e., a very high diemnsional vector $\boldsymbol{f}$. So each face is now a vector in the "face space", which is just $\mathbb{R}^m$ for some very big m.

Now to build a facial recognition process, we need to first start by collecting data. Say we collected the face of n people for very large n. Now we have $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n \in \mathbb{R}^m$ for large m, n. These data can fit into a matrix $A = \begin{bmatrix} \boldsymbol{f}_1 & \dots & \boldsymbol{f}_n \end{bmatrix}$. What do we do next?

Next we want to identifa major features of a face. For example, "nose" is a major feature, "eyes" are also a major feature. To see if two faces are the same, we can compare the nose, compare the eyes, and so on. If enough major features are similar, then we say the two face are the same.

Think about the meaning of this. This means we are decomposing each face $\boldsymbol{f}$ into various features. Since we are doing this for each $\boldsymbol{f}_i$, this means we are decomposing A into a sum of matrices $A = \sum A_i$, where each A_i represent some feature. Furthermore, since "nose" would typically involve less pixels than the whole face, each feature matrix A_i would have a low rank. So we are attempting to decompose A into a sum of low rank matrices. By comparing faces using only the most important features, we are hoping that $A_1 + \cdots + A_k$ for some small k would give a good approximation to the whole matrix A.

You can see now how SVD comes in. The best way to do this is via $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. Here we decomposed A into matrices with the smallest rank possible (rank one), and these "features" are all mutually "independent" (orthogonal to each other), and furthermore, a truncation $\sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$ is the BEST rank k approximation of A. These features might not have nice interpretations such as "nose", but they are the most efficient way to capture the essense of A while using minimal rank.

So to do facial recognition, we can perform $A = U\Sigma V = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$. Here columns of U are also in $\mathbb{R}^m$, and they are called "eigenfaces" sometimes, since they are eigenvectors of AA^{T}. You can even draw them out. These $\boldsymbol{u}_i$ form an orthonormal basis for the face space $\mathbb{R}^m$, and for each $\boldsymbol{f} \in \mathbb{R}^m$, then $\boldsymbol{f}$ would be some linear combination of them. Since $\boldsymbol{f} = UU^{\mathrm{T}}\boldsymbol{f}$, you can see that the coordinates under the orthonormal basis is just $\boldsymbol{u}_i$. This is akin to breaking down a face into "features", and note that from $\boldsymbol{u}_1$ to $\boldsymbol{u}_m$, we are going in the order of decreasing importance of features.

So fix some small k. Given two faces $\boldsymbol{f}_1, \boldsymbol{f}_2$, you can simply compare the first k coordinates of $U^{\mathrm{T}}\boldsymbol{f}_1, U^{\mathrm{T}}\boldsymbol{f}_2$. If these coordinates are close enough, then all the most important features are close enough, so we can declare them to be the same face.

Note that V here is less important. It records information about how to combine the input faces $\boldsymbol{f}_i$ into features, and it has little consequence in application.

For more on facial recognition using SVD, check out this paper. ☺

Example 7.12.34. I came across this application in the textbook by Gilbert Strang. The information here is written in this paper published in nature.

In short, for each person, we can write a list of number to represent this person's genetic information, i.e., some vector $\boldsymbol{p}$. So if we collect the genetic information of many people, and use these vectors as columns to form a matrix, we get A_0. Then we center the data (substract each column by the average of all columns, and now each row has average zero). Thus we obtain $A = \begin{bmatrix} \boldsymbol{p}_1 & \dots & \boldsymbol{p}_n \end{bmatrix}$, and perform SVD for $A = U\Sigma V = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$.

Here the two most principal components to explain A (the genetic variation data set) are $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$ and $\sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$. Here $\boldsymbol{u}_1, \boldsymbol{u}_2$ would be the most important genetic "features" to explain genetic variation. What are the meanings of $\boldsymbol{v}_1, \boldsymbol{v}_2$?

Note that $\boldsymbol{u}_1^{\mathrm{T}} A = \sigma_1 \boldsymbol{v}_1^{\mathrm{T}}$ by definition of a left singular vector, yet we also have $\boldsymbol{u}_1^{\mathrm{T}} A = \begin{bmatrix} \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{p}_1 & \dots & \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{p}_n \end{bmatrix}$. So the amount of genetic feature $\boldsymbol{u}_1$ contained by the i-th person is exactly the i-th coordinate of $\sigma_1 \boldsymbol{v}_1^{\mathrm{T}}$.

Suppose we ignore ALL other genetic features than the first two. Then the genetic variation of each person would only need two coordinates, and thus each person becomes a point in $\mathbb{R}^2$, and they are columns of $\begin{bmatrix} \sigma_1 \boldsymbol{v}_1^{\mathrm{T}} \\ \sigma_2 \boldsymbol{v}_2^{\mathrm{T}} \end{bmatrix}$. Graph these points in $\mathbb{R}^2$, and we recover something like a map of Europe. (Note that the paper in question only gathered genetic data in Europe.) Check out the paper for a graph of this comparison.

This means that the most important rank 2 factor influencing genetic variations is approximately the geological location of the person. This makes a lot of sense. If you sample the gene of someone in France, with great probability this person will be of French descent, and this person will most likely have the corresponding genetic variations common to French people.

Think about a study about coronavirus policy. Is lockdown an effective method? If you study this by collecting data from different countries, then your result could be misleading. Say one country locked down, while the other remained open, and we observed a difference in the spread of the virus. Does this imply lockdown is effective? Not by itself. If a person's genetic information effects the spreading of the virus, and geography explains the genetic variation, then what you have detected might be genetic variations caused by geography.

Therefore, a rigorous study must try to "correct" the data against this confounding factor. ☺

7.13 Classification of Quadratic surfaces

(This section should have been right next to definiteness of symmetric matrices. However, since it is the most optional, I put it last in case I have no time and need to skip it.)

Previously, given a square matrix A, we think of it as a linear transformation (domain = codomain). In particular, a change of basis would induce a change in matrix into $X^{-1}AX$ for some invertible X. In this section, we try something different. Let us think of A as a bilinear form.

Definition 7.13.1. *A bilinear form is a function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that $f(-, \boldsymbol{v})$ is linear and $f(\boldsymbol{v}, -)$ is also linear. More specifically, we have $f(a\boldsymbol{u} + b\boldsymbol{v}, \boldsymbol{w}) = af(\boldsymbol{u}, \boldsymbol{w}) + bf(\boldsymbol{v}, \boldsymbol{w})$ for the first input, and same for the second input.*

Proposition 7.13.2. *We have $f(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v}^{\mathrm{T}} G \boldsymbol{w}$ for a matrix G whose (i, j) entry is $f(\boldsymbol{e}_1, \boldsymbol{e}_j)$. (So it is a Gram matrix.)*

Proof. Simply use bilinearity to expand $\boldsymbol{v}, \boldsymbol{w}$ in terms of the standard basis, and we are done. □

Proposition 7.13.3. *Suppose we perform a change of basis, so that $\boldsymbol{v}_{new} = X\boldsymbol{v}_{old}$ for invertible X. Then for the bilinear form with Gram matrix A, we have $A_{old} = X^{\mathrm{T}} A_{new} X$.*

Proof. The bottom line is that $\boldsymbol{v}_{new}^{\mathrm{T}} A_{new} \boldsymbol{w}_{new} = \boldsymbol{v}_{old}^{\mathrm{T}} A_{old} \boldsymbol{w}_{old}$. Now substitute in $\boldsymbol{v}_{new} = X\boldsymbol{v}_{old}$, we see that $\boldsymbol{v}_{old}^{\mathrm{T}} X^{\mathrm{T}} A_{new} X \boldsymbol{w}_{old} = \boldsymbol{v}_{old}^{\mathrm{T}} A_{old} \boldsymbol{w}_{old}$ for all $\boldsymbol{v}_{old}, \boldsymbol{w}_{old} \in \mathbb{R}^n$. So we have $A_{old} = X^{\mathrm{T}} A_{new} X$. □

In particular, the change of basis formula for matrices are now different. If the matrix A represent a bilinear form, then $X^{\mathrm{T}} AX$ is the result after a change of basis.

Definition 7.13.4. *Two matrices A, B are said to be **congruent** if $A = XBX^{\mathrm{T}}$ for some INVERTIBLE X. (Note that we require X to be invertible here, since it should come from some change of basis.)*

Previously, when we study eigenvalues, we would try to find X such that $X^{-1}AX$ is diagonal D, i.e., a decomposition $A = XDX^{-1}$. This is the best basis to study the linear transformation A. Now what is the best basis to study the bilinear form A? We would now try to find X so that $X^{\mathrm{T}} AX$ is as simple as possible.

Thankfully, this is not too bad, when A is symmetric. We have spectral decomposition $A = QDQ^{-1}$ and $Q^{-1} = Q^{\mathrm{T}}$. So by choosing the right basis, we can always assume that A is diagonal. So let us restrict our attention now.

Definition 7.13.5. *A bilinear form is a **symmetric bilinear form** if $f(\boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{w}, \boldsymbol{v})$. (Obviously these are also bilinear forms with symmetric Gram matrix.)*

Remark 7.13.6. *There is also the concept of a **quadratic form**, which are $q : \mathbb{R}^n \to \mathbb{R}$ such that $q(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{x})$ for some symmetric bilinear form f. These corresponds to homogeneous polynomials of degree 2 on n variables.*

The study of quadratic form is pretty much the same as the bilinear form though. If $q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x}$, then by taking transpose on both sides, we also have $q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} A^{\mathrm{T}}\boldsymbol{x}$. So $q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} G\boldsymbol{x}$ where $G = \frac{1}{2}(A + A^{\mathrm{T}})$ is symmetric. So we can always assume that the bilinear form for the quadratic form q is symmetric.

Note that one can also obtain the symmetric bilinear form via polarization identity $f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(q(\boldsymbol{x} + \boldsymbol{y}) - q(\boldsymbol{x}) - q(\boldsymbol{y}))$, which can be verified to be bilinear and symmetric, and $q(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{x})$.

Proposition 7.13.7. *For any symmetric bilinear form, we can choose a basis so that the Gram matrix is diagonal D where the diagonal entries are $0, 1, -1$.*

Proof. Suppose that under the standard basis, the Gram matrix is symmetric A. By spectral decomposition, $A = QBQ^{\mathrm{T}}$ for diagonal B. So our goal now is to find a decomposition so that $B = XDX^{\mathrm{T}}$, and D is as desired.

Say $B = \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_n \end{bmatrix}$. If $b_i \neq 0$, let $x_i = \sqrt{|b_i|}$, and if $b_i = 0$, set $x_i = 1$. Either way, we have $x_i \neq 0$.

Let $X = \begin{bmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{bmatrix}$, and $D = \begin{bmatrix} \operatorname{sign}(b_1) & & \\ & \ddots & \\ & & \operatorname{sign}(b_n) \end{bmatrix}$, where $\operatorname{sign}(x) = 1$ if $x > 0$, -1 if $x < 0$, and 0 if $x = 0$. Then it is easy to verify that $B = XDX^{\mathrm{T}}$ and X is invertible. So we are done. □

Corollary 7.13.8. *Any symmetric matrix A is congruent to $\begin{bmatrix} I_a & & \\ & -I_b & \\ & & O \end{bmatrix}$. (We call this the **congruence canonical form of** A.)*

Now, are the integers a, b uniquely determined by A? They are. In fact, they have some very interesting structural meaning.

Note that there are something interesting going on here. Let's say we picked a nice basis and thus our symmetric bilinear form has Gram matrix $G = \begin{bmatrix} I_a & & \\ & -I_b & \\ & & O \end{bmatrix}$. Then consider the subspace $V = \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_a)$. For any $\boldsymbol{v} \in V$, you can easily verify that $\boldsymbol{v}^{\mathrm{T}} G\boldsymbol{v} = \|\boldsymbol{v}\|^2 \geq 0$ with equality if and only if $\boldsymbol{v} = \boldsymbol{0}$. So even though G itself is NOT positive definite, but on the subspace V, it is positive definite.

Definition 7.13.9. *For a symmetrix matrix A, we define its **positive index of inertia** to be n_+, the size of largest subspace on which A can be positive definite. We define its **negative index of inertia** to be n_-, the size of largest subspace on which A can be negative definite.*

Lemma 7.13.10. *If A is positive definite on a subspace V, then A is positive semi-definite on $V + \operatorname{Ker}(A)$.*

Proof. For any $\boldsymbol{u} \in V + \operatorname{Ker}(A)$, suppose $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$ where $\boldsymbol{v} \in V$ and $\boldsymbol{w} \in \operatorname{Ker}(A)$. (Note that there is a chance of $\boldsymbol{v} = \boldsymbol{0}$ here, when $\boldsymbol{u}$ is entirely in $\operatorname{Ker}(A)$.)

Then $\boldsymbol{u}^{\mathrm{T}} A\boldsymbol{u} = (\boldsymbol{v} + \boldsymbol{w})^{\mathrm{T}} A(\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{v}^{\mathrm{T}} A\boldsymbol{v} \geq 0$. Here all the $\boldsymbol{w}$ disappear because $A\boldsymbol{w} = \boldsymbol{0}$ and (by symmetricity of A) $A^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{0}$. Also note that we merely have $\geq$ rather than $>$, because maybe $\boldsymbol{v} = \boldsymbol{0}$. □

Lemma 7.13.11. *Suppose an $n \times n$ symmetric matrix A has positive index of inertia n_+, negative index of inertial n_-, and kernel of size n_0. Then $n_+ + n_- + n_0 \leq n$.*

Proof. Let V_+ be a subspace of dimension n_+ on which A is positive definite. Then all non-zero $\boldsymbol{v} \in V_+$ will have $\boldsymbol{v}^{\mathrm{T}} A\boldsymbol{v} > 0$, while all non-zero $\boldsymbol{v} \in \operatorname{Ker}(A)$ will have $\boldsymbol{v}^{\mathrm{T}} A\boldsymbol{v} = 0$. In particular, V_+ and $\operatorname{Ker}(A)$ has trivial intersection, and $\dim(V_+ + \operatorname{Ker}(A)) = \dim(V_+) + \dim \operatorname{Ker}(A) = n_+ + n_0$.

Let V_- be a subspace of dimension n_- on which A is negative definite. Then all non-zero $\boldsymbol{v} \in V_+$ will have $\boldsymbol{v}^{\mathrm{T}} A\boldsymbol{v} < 0$. However, in comparison, since A is positive semi-definite on $V_+ + \operatorname{Ker}(A)$, all non-zero $\boldsymbol{v} \in V_+ + \operatorname{Ker}(A)$ will have $\boldsymbol{v}^{\mathrm{T}} A\boldsymbol{v} \geq 0$. In particular, V_- and $V_+ + \operatorname{Ker}(A)$ has trivial intersection, and $\dim(V_- + V_+ + \operatorname{Ker}(A)) = \dim(V_-) + \dim(V_+ + \operatorname{Ker}(A)) = n_- + n_+ + n_0$. But since $V_- + V_+ + \operatorname{Ker}(A) \subseteq \mathbb{R}^n$, we see that $n_+ + n_- + n_0 \leq n$. □

Theorem 7.13.12 (Sylvester's Law of Inertia)**.** *Suppose an $n \times n$ symmetric matrix A has positive index of inertia n_+, negative index of inertial n_-, and kernel of size n_0. We have $n_+ + n_- + n_0 = n$, and n_+ is the number of positive eigenvalues of A, and n_- is the number of negative eigenvalues of A.*

In particular, if A, B are congruent, then they have the same positive index of inertia and same negative index of inertial. If A is congruent to $D_{a,b} = \begin{bmatrix} I_a & & \\ & -I_b & \\ & & O \end{bmatrix}$, then we must have $a = n_+$ and $b = n_-$. So the congruence canonical form of A is unique.

Proof. Since A is normal, its eigenspaces are spanning. Let V_+ be the subspace spanned by all eigenspaces of A for positive eigenvalues of A. Then for any vector $\boldsymbol{v} \in V_+$, it is a sum of eigenvectors of A for positive eigenvalues, say $\boldsymbol{v} = \sum \boldsymbol{v}_i$ where each $\boldsymbol{v}_i$ has eigenvalue $\lambda_i > 0$. Then $\boldsymbol{v}^{\mathrm{T}} A \boldsymbol{v} = \sum \lambda_i \|\boldsymbol{v}_i\|^2 \geq 0$, with equality if and only if $\boldsymbol{v}_i$ are all $\boldsymbol{0}$, if and only if $\boldsymbol{v} = \boldsymbol{0}$. So A is positive definite on V_+.

Since A is normal, algebraic multiplicity equals geometric multiplicity, so $\dim(V_+)$ is exactly the number of positive eigenvalues of A, say m_+. Since A is positive definite on V_+, and n_+ is by definition the largest possible dimension of such a subspace, we have $n_+ \geq m_+$. Similarly, we also see that $n_- \geq m_-$, where m_- is the number of negative eigenvalues of A. Finally, we have $n_0 = m_0$ where m_0 is the algebraic multiplicity of zero as an eigenvalue of A.

Now since A has n eigenvalues, we have $n = m_+ + m_- + m_0 \leq n_+ + n_- + n_0 \leq n$. So we have equality everywhere. □

Remark 7.13.13. *By analyzing the proof above, you can also see that the largest subspace on which A is positive definite is in fact unique. It must be the span of all eigenspaces for all possitive eigenvalues.*

In particular, we see that up to a change of basis, a symmetric bilinear form is completely determined by its positive and negative index of inertia. Now we can proceed to classify quadratic curves and quadratic surfaces.

Example 7.13.14 (Classification of quadratic curves)**.** Consider the solution set to some degree two polynomial equation $ax^2 + bxy + dy^2 + ex + fy + c = 0$ in $\mathbb{R}^2$ for constants a, b, c, d, e, f. What is it?

Let $A = \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & d \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} e \\ f \end{bmatrix}$, then we see that the solution sets are points $\boldsymbol{x}$ such that $\boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} + c = 0$. Suppose for now that A is invertible.

Then we can try to complete the square. Recall that $ax^2 + bx + c$ can be written as $a(x + \frac{b}{2a})^2 + (c - \frac{b^2}{4a})$. So similarly, we change variable with $\boldsymbol{y} = \boldsymbol{x} + \frac{1}{2} A^{-1} \boldsymbol{b}$, and we have $\boldsymbol{y}^{\mathrm{T}} A \boldsymbol{y} = C$ for some constant C. Note that all we did is to translate the solution set around, without changing its shape.

Now we perform a change of basis, and we can assume that A is in congruence canonical form. We have several possibilities.

1. Suppose $n_+ = 2$ and $n_- = 0$, then if $\boldsymbol{y} = \begin{bmatrix} x \\ y \end{bmatrix}$, we are facing $x^2 + y^2 = C$. The solution here is a circle if $C > 0$. Going back to the solution set $\boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} + c = 0$, the change of basis process will make the circle into an ellipse, and then the change of variable process will translate the ellipse around. Anyway, we end up with an ellipse as the solution set. (If $C = 0$, the solution is a single point. If $C < 0$, the solution set is empty.)

2. Suppose $n_+ = 0$ and $n_- = 2$, then if $\boldsymbol{y} = \begin{bmatrix} x \\ y \end{bmatrix}$, we are facing $x^2 + y^2 = -C$. We see that this is the same as above, except that now $C < 0$ gives the ellipse, and $C > 0$ now gives empty solution set.

3. Suppose $n_+ = n_- = 1$. Then if $\boldsymbol{y} = \begin{bmatrix} x \\ y \end{bmatrix}$, we are facing $x^2 - y^2 = C$. We see that this is a hyperbola. Change of basis means the two asymtotes can be in any direction, and the change of variable means we can translate it around.

So all in all, it is either an ellipse, or a hyperbola, or in the degenerate cases, a single point or the empty set.

What if A is NOT invertible? If $A = 0$, then we have a degree one equation $ex + fy + c = 0$, so the solution is a single line, or it is the empty set if $e = f = 0$ while $c = 0$, giving us a contradiction.

We are left with one last case, when A has rank 1. Note that if $\boldsymbol{b} \in \mathrm{Ran}(A)$, say $\boldsymbol{b} = A\boldsymbol{v}$, then we can again change variable $\boldsymbol{y} = \boldsymbol{x} + \frac{1}{2}A^{-1}\boldsymbol{b}$ as before.

1. If $n_+ = 1$ and $n_- = 0$ and $\boldsymbol{b} \in \mathrm{Ran}(A)$. Then by change of variable, we have $\boldsymbol{y}^{\mathrm{T}}A\boldsymbol{y} = C$. By change of basis, we have $x^2 = C$. So we have a pair of parallel lines.

2. If $n_+ = 0$ and $n_- = 1$ and $\boldsymbol{b} \in \mathrm{Ran}(A)$. Then by change of variable, we have $\boldsymbol{y}^{\mathrm{T}}A\boldsymbol{y} = C$. By change of basis, we have $x^2 = -C$. So we have a pair of parallel lines again.

3. If $n_+ = 1$ and $n_- = 0$ and $\boldsymbol{b} \notin \mathrm{Ran}(A)$. Then we change basis directly, and have $\boldsymbol{x}^{\mathrm{T}}D\boldsymbol{x}+(\boldsymbol{b}')^{\mathrm{T}}\boldsymbol{x}+c = 0$, which gives $x^2 + e'x + f'y + c = 0$. Since $\boldsymbol{b} \notin \mathrm{Ran}(A)$, after change of basis, we have $\boldsymbol{b}' \notin \mathrm{Ran}(D)$, so $f' \neq 0$. This gives $y = -\frac{1}{f'}(x^2 + e'x + c)$, so the solution set is a parabola.

4. If $n_+ = 1$ and $n_- = 0$ and $\boldsymbol{b} \notin \mathrm{Ran}(A)$, this is the same as above by changing c into $-c$. So this is a parabola.

Let us now sum up all cases of the solution set to $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + c = 0$. This can be an ellipse, a hyperbola, a parabola, a pair of parallel lines, a single line, a single point or the empty set. Throwing away the degree one cases and the degenerate cases and the contradiction cases, the true degree two cases can only be ellipses, hyperbolas and parabolas. ☺

Example 7.13.15 (Classification of quadric surfaces)**.** We now look at the solution set of degree two polynomials in $\mathbb{R}^3$. Again, we can rearrange so that the equation looks like $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + c = 0$. To save time, let us drop the lower degree cases, degenerate cases and contradictory cases, and only try to classify the rest. This means $A \neq O$ (but A is allowed to be non-invertible), and $\boldsymbol{b}$ and $\mathrm{Ran}(A)$ span the whole $\mathbb{R}^3$. (Otherwise we will be degenerate.)

Suppose A is invertible. Then again we can simplify via translation to $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} = C$. If $C = 0$, we are degenerate. So assume $C \neq 0$. By replacing A via $-A$ if needed, we can assume $C > 0$.

1. If $n_+ = 3$, then after change of basis, we have $x^2 + y^2 + z^2 = C$, so this is a sphere. Going back via change of basis and translation, we have an aribtrary ellipsoid.

2. If $n_+ = 2$ and $n_- = 1$, then after change of basis, we have $x^2 + y^2 - z^2 = C$. This is the famous hyperboloid with one sheet. Take the hyperbola $x^2 - z^2 = C > 0$ and rotate around the z-axis, and you can find this shape. It is a connected surface. (Note that there is a single negative sign in the entire equation, which helps you memorize the number of sheets.)

3. If $n_+ = 1$ and $n_- = 2$, then after change of basis, we have $x^2 - y^2 - z^2 = C$. This is the famous hyperboloid with two sheets. Take the hyperbola $x^2 - z^2 = C > 0$ and rotate around the x-axis, and you can find this shape. It is has two connected surfaces as its components. (Note that there are two negative sign in the entire equation, which helps you memorize the number of sheets.)

4. If $n_- = 3$, then A is negative definite. Since we require $C > 0$, $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} = C > 0$ is impossible. So there is no solution in this case.

5. (Optional) Note that if $C = 0$, then we are degenerate. In the cases $n_+ = 3$ or $n_+ = 0$, the solution is a single point. In the cases $n_+ = 2$ or $n_+ = 1$, the solution set is a pair of opposite cones.

Since we require $\boldsymbol{b}$ and $\mathrm{Ran}(A)$ to span the whole $\mathbb{R}^3$, the only non-invertible case is $\mathrm{rank}(A) = 2$ and $\boldsymbol{b} \notin \mathrm{Ran}(A)$. By a change of basis, we have $\pm x^2 \pm y^2 + ax + by + cz + d = 0$. In fact by a translation in variable x, y, we are left with $\pm x^2 \pm y^2 + cz + d = 0$. We change variable z so that $-cz$ is now replaced by z, then we have $z = \pm x^2 \pm y^2 + d$, and the signs depends on the index of inertia of A.

1. If $n_+ = 2$, then $z = x^2 + y^2 + d$. This is a paraboloid opening upward. (An elliptic paraboloid after a change of basis.)

2. If $n_+ = n_- = 1$, then $z = x^2 - y^2 + d$. This is a hyperbolic paraboloid.

3. If $n_- = 2$, then $z = -x^2 - y^2 + d$. This is a paraboloid opening downward.(An elliptic paraboloid after a change of basis.)

To sum up, the true quardric surfaces are: ellipsoid, hyperboloid with one sheet, hyperboloid with two sheets, elliptic paraboloid, hyperbolic paraboloid.

If you are interested, the quardric degenerate surfaces are: elliptic cones, elliptic cylinder, hyperbolic cylinder, parabolic cylinder. And then we have lower degree cases line planes, or lower dimensional cases like points or the empty set. ☺

7.14 (Optional) Congruence canonical form for skew-symmetric matrices

If A is skew symmetric, then $Q^{\mathrm{T}}AQ$ is block diagonal where each block is 0 or $\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$ for some a. Now when I perform $X^{\mathrm{T}}AX$, by choosing X to be block diagonal as well, I can assume that each block acts independently.

Now since $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}$, up to congruence I can swap the two non-diagonal entries, so I can assume that we have $a > 0$. Finally, by setting $X = \begin{bmatrix} \sqrt{a} & \\ & \sqrt{a} \end{bmatrix}$, we have $\begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} = X \begin{bmatrix} & -1 \\ 1 & \end{bmatrix} X^{\mathrm{T}}$.

So in the end, A is congruent to a block diagonal matrix where each block is $\begin{bmatrix} & -1 \\ 1 & \end{bmatrix}$ or 0. The congruence relation is completely determined by the rank of A.

In particular, for each $\mathbb{R}^{2n}$, all alternating bilinear forms (bilinear forms with $f(\boldsymbol{v}, \boldsymbol{w}) = -f(\boldsymbol{w}, \boldsymbol{v})$) are the SAME up to a change of basis. This is a very important property, and involved in the study of symplectic geometry, classical mechanics, and quantum mechanics. Here we can merely provide a quick taste of the relation.

Example 7.14.1. Suppose a ball is located at $\boldsymbol{p} \in \mathbb{R}^3$ with velocity $\boldsymbol{v} \in \mathbb{R}^3$. Let us say the potential energy is $V(\boldsymbol{p})$ (maybe due to gravity, or due to magnetic field, or whatever), a function depending on the location of the ball, and the kinetic energy is $\frac{1}{2}m\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$. Then the total energy is $E(\boldsymbol{v}, \boldsymbol{p}) = V(\boldsymbol{p}) + \frac{1}{2}m\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$.

Suppose the only force on our object is the potential force. Note that the potential force would always try to push objects to the direction that reduce potential as fast as possible. Let us write $\frac{\partial E}{\partial \boldsymbol{p}}$ as the vector $\begin{bmatrix} \frac{\partial V}{\partial p_1} & \cdots & \frac{\partial V}{\partial p_3} \end{bmatrix}$, then to reduce potential as fast as possible, we skip the physical analysis here, and simply claim that it means $\boldsymbol{F} = -\frac{\partial E}{\partial \boldsymbol{p}}$. But note that $\boldsymbol{F} = m\boldsymbol{a} = m\boldsymbol{v}'$, where $\boldsymbol{v}'$ is the derivative of $\boldsymbol{v}$ with respect of time. So $\frac{\partial E}{\partial \boldsymbol{p}} = -m\boldsymbol{v}'$.

On the other hand, consider $\frac{\partial E}{\partial \boldsymbol{v}} = \frac{\partial}{\partial \boldsymbol{v}}(\frac{1}{2}m\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}) = m\boldsymbol{v} = m\boldsymbol{p}'$, since velocity is the derivative of location.

As a result, we have the following structure: $\begin{bmatrix} \boldsymbol{p}' \\ \boldsymbol{v}' \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{m}I_3 \\ -\frac{1}{m}I_3 & 0 \end{bmatrix}\begin{bmatrix} \frac{\partial E}{\partial \boldsymbol{p}} \\ \frac{\partial E}{\partial \boldsymbol{v}} \end{bmatrix}$. So given the mass and a formula for total energy, this equation here would completely describe the evolution of the location and velocity of our ball. Here we see the involvement of a skew-symmetric matrix.

For more advanced reason (gradients are contravariant and change basis differently), a change of basis would actually induce $X\begin{bmatrix} 0 & \frac{1}{m}I_3 \\ -\frac{1}{m}I_3 & 0 \end{bmatrix}X^{\mathrm{T}}$ on the matrix, so we should think of the matrix here as an alternating bilinear form, rather than a linear transformation. This is how it is related to mechanics.

You can also see that matrices X such that $X \begin{bmatrix} 0 & \frac{1}{m} I_3 \\ -\frac{1}{m} I_3 & 0 \end{bmatrix} X^{\mathrm{T}} = \begin{bmatrix} 0 & \frac{1}{m} I_3 \\ -\frac{1}{m} I_3 & 0 \end{bmatrix}$ are of special interests, since they describe the "symmetries" of our mechanical system.

In general, we say X is symplectic if $X \begin{bmatrix} O & -I \\ I & O \end{bmatrix} X^{\mathrm{T}} = \begin{bmatrix} O & -I \\ I & O \end{bmatrix}$, and the study of geometric properties invariant under such linear transformation X is the study of symplectic geometry. These symplectic properties would be the "essence" of the mechanical system, similar to how eigenstructures are the "essence" of a linear transformation. ☺

Part IV

Review and Introduction

Chapter 8

Complex Matrices

8.1 What is a complex linear combination?

We are entering into the second portion of your linear algebra education, and we are going to see more complex matrices. A complex matrix is, in a very nominal sense, a matrix with possibly complex entries, say $\begin{bmatrix} 1+\mathrm{i} & -\mathrm{i} \\ 2-\mathrm{i} & 3 \end{bmatrix}$. But this should NOT be satisfactory for you, because what does it even mean?

Let us do a little review first.

Recall that a matrix $A = \begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 2 & 0 \end{bmatrix}$ is representing a linear map. In particular, it represents some process that respect linear combinations. As a quick example, say we are doing the famous "hen and rabbit" problem. Each hen has one head, two legs, and two wings. Each rabbit has one head, four legs, and no wing. So given $\begin{bmatrix} x \text{ hens} \\ y \text{ rabbits} \end{bmatrix}$, then you would have a total of $A \begin{bmatrix} x \text{ hens} \\ y \text{ rabbits} \end{bmatrix} = \begin{bmatrix} x+y \text{ heads} \\ 2x+4y \text{ legs} \\ 2x \text{ wings} \end{bmatrix}$. So A is the counting process that tells you how many heads, legs and wings do we have. This process is LINEAR, because the total body parts of "a linear combination of animals" is the corresponding linear combination of the body parts of each type of animal. It RESPECTS the linear combination in the sense that $A(s\boldsymbol{v} + t\boldsymbol{w}) = s(A\boldsymbol{v}) + t(A\boldsymbol{w})$.

For more fun examples, see Chapter 1.

If you forget all about our class last quarter, at least I hope you would remember these. A vector is representing a linear combination, and a matrix is representing a linear map, which is a map that preserves linear combinations. (Personally I think this perspectives on linear combinations and linear maps is WHY we learn linear algebra in college. No other stuff is iimportant.)

Now, under this view, the idea of a complex matrix like $\begin{bmatrix} 1+\mathrm{i} & -\mathrm{i} \\ 2-\mathrm{i} & 3 \end{bmatrix}$ is very disturbing. This seems to be about COMPLEX linear combinations, in contrast the the real linear combinations that we are used to. For example, if I think about all my friends, maybe I have 2 male friends and 3 femail friends, and this corresponds to a vector $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$. But then a complex vector $\begin{bmatrix} 4 \\ \mathrm{i} \end{bmatrix}$ would mean that I have 4 male friends and an imaginary female friend. Huh?

So before we move on, we need a little extra perspective on complex numbers and complex linear combinations.

First of all, why do we even need complex numbers? The answer is obvious: we want a degree n polynomial to have n roots. This is straightforward enough. Over the reals, $x^2 + 1 = 0$ has no solution, which is super annoying. For example, without complex numbers, $\begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}$ has NO eigenvector and no eigenvalues, which is annoying. But over complex numbers, it will have distinct eigenvalues $\pm\mathrm{i}$, and in fact

it will be diagonalizable. Hooray! (If you forget all about eigenvalues, this example is subtly telling you to review.)

So this establishes the necessity of complex numbers. But how can we find roots for $x^2+1=0$ out of thin air? It turns out that this is asking the wrong question. "How to find" is not important. The key is "where to find". In particular, where can we find i?

Example 8.1.1. We are searching for x such that $x^2=-1$. And this is impossible over real numbers. But broaden our minds a little bit. Can we find a real matrix J such that $J^2=-I$?

Yes we can. The 2×2 real matrices are linear transformations on $\mathbb{R}^2$, the plane. On the plane, what is $-I$? That is basically reflecting everything about the origin, i.e., rotation by 180 degree. How can we find an operation J, such that J^2 is rotation by 180 degree? The answer is rotation by 90 degree, easy.

I hope you still remembered how to find this matrix. (Again, if you don't know how, please review Chapter 1.) The answer is (if we rotate counter-clockwise) $J=\begin{bmatrix}0&-1\\1&0\end{bmatrix}$. Of course, $-J$ also satisfies $(-J)^2=-I$, so we in fact have at least two solutions, $\pm J$, just like $x^2=-1$ has two solutions, $\pm i$. (We in fact have infinitely many solutions to the matrix equations $J^2=-I$. Can you find a way to describe them all?)

Now is time to witness magic. Lo and behold the wonders of algebra.

$$(2+3\mathrm{i})(4+\mathrm{i})=5+14\mathrm{i};$$
$$(2I+3J)(4I+J)=\begin{bmatrix}2&-3\\3&2\end{bmatrix}\begin{bmatrix}4&-1\\1&4\end{bmatrix}=\begin{bmatrix}5&-14\\14&5\end{bmatrix}=5I+14J.$$

In fact, you can safely assume that any complex number $a+b\mathrm{i}$ is secretly representing the REAL matrix $\begin{bmatrix}a&-b\\b&a\end{bmatrix}=aI+bJ$. Then addition of complex numbers would corresponds to addition of matrices, and multiplication of complex numbers are simply multiplication of matrices. (Can you prove this yourself?)

Here is something to think about. Suppose some $n\times n$ matrix J satisfies $J^2=-I$, then would we have a similar structure? ☺

Example 8.1.2. Bonus foods for your thought. Compute the following two matrix multiplications. What would you get? How are the two following calculations related?

$$\begin{bmatrix}1&\mathrm{i}\\2\mathrm{i}&1+\mathrm{i}\end{bmatrix}\begin{bmatrix}\mathrm{i}&1-\mathrm{i}\\2&\mathrm{i}\end{bmatrix}=?$$

$$\left[\begin{array}{cc|cc}1&0&0&-1\\0&1&1&0\\\hline0&-2&1&-1\\2&0&1&1\end{array}\right]\left[\begin{array}{cc|cc}0&-1&1&1\\1&0&-1&1\\\hline2&0&0&-1\\0&2&1&0\end{array}\right]=?$$

Suppose some $n\times n$ matrix J satisfies $J^2=-I$, then can you construct similar coincidences? ☺

Example 8.1.3. We have hinted that whenever $J^2=-I$, then you can choose i as representing J, and use complex numbers. What are other possible A? Here is an exotic (but useful) example.

Let V be the space of functions of the form $a\sin(x)+b\cos(x)$. This is a very useful space, because it is the space of all waves with frequency 2π. Let $J:V\to V$ be the linear map "taking derivatives". Then note that $J^2=-I$ in this space. ☺

The above serves to point out that the imaginary unit i has very real meanings, and possibly many many meanings, and you should pick your own meaning depending on the application at hand. Luckily for us, most of the time, when people use complex numbers, they are usually interpreting the imaginary i as some sort of rotation, i.e., $\begin{bmatrix}0&-1\\1&0\end{bmatrix}$.

Under this interpretation, a complex number $a+b\mathrm{i}$ can be interpreted as $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. So a pure real number is like a dilation operation on the plane, while a purely imaginary number is like a rotation operation on the plane. Here is a example copied from the book "From One to Infinity".

Example 8.1.4. A treasure is buried on an island. To find the treasure, we start at a location with a flag (location Z). We then first walk to a building (location A), say with a total distance of x, then we turn right and walk x. Let us call this location A'.

Next we go back to the flag (location Z). We then first walk to a statue (location B), say with a total distance of y, then we turn left and walk y. Let us call this location B'.

The treasure is at the midpoint between A' and B'.

Now some bad guy came and took away the flag (so Z is unknown). Can you still find the treasure? Yes we can.

Note that $A'-A$ is $A-Z$ rotated clockwise, so $A'-A=-\mathrm{i}(A-Z)$. Similarly, $B'-B$ is $B-Z$ rotated counter-clockwise, so $B'-B=\mathrm{i}(B-Z)$. So the treasure locatio $\frac{1}{2}(A'+B')=\frac{1}{2}(A+B)+\frac{1}{2}\mathrm{i}(B-A)$, and no Z is involved in this. So the flag position does not matter at all. I'll leave the interpretation of the final treasure location to yourself.

This is NOT showing you the power of complex numbers. Rather, this is showing you the power of linear algebra. At the center of the entire calculation is the fact that rotation is linear. The complex numbers such as i are merely names that we slap on the operations such as rotations.

So... linear algebra rules, and complex numbers are just names and labels for convenience. ☺

So, when we are dealing with objects that can be "rotated", it would make sense to talk about i times that object. In this sense, we can do complex-linear combinations. No wonder that quantum mechanics, which need to discuss the "spin" of a particle, are using complex numbers.

All in all, for a complex vector such as $\boldsymbol{v}=\begin{bmatrix} 1 \\ \mathrm{i} \\ 1-\mathrm{i} \end{bmatrix}$, it is better to think of each coordinate as representing a point in the plane. So we have three points on the plane. And if we perform a complex scalar multiplication $(2+\mathrm{i})\boldsymbol{v}$, think of this as applying a planar operation $\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ to each coordinate of $\boldsymbol{v}$. So we are stretching and rotating the three points simultaneously.

Here are some other fun applications of complex numbers.

Example 8.1.5 (Complex romantic relation)**.** Romantic relations are indeed complex. Ha, let's dig into this example.

Suppose $f'=kf$, then I'm sure you know that the solution is $f(x)=\mathrm{e}^{kx}f(0)$. That is the prerequisite knowledge of this application.

Suppose two person A, B are in a romantic relation.Their love for each other is a function of time, say $A(t)$ and $B(t)$. Now A is a normal person. For normal people, the more you are loved, the more you love back. In particular, $A'(t)=B(t)$. However, B is an unappreciative person. If you love B, then B take you for granted, and treat you as garbage. If, however, you treat B badly, then B would all of a sudden thinks of you as super charming and attractive. In short, B enjoys things that are hard to get, and think little of the things that are easy to get. In Chinese, we say B is a Jian Ren. Anyway, we see that $B'(t)=-A(t)$.

Now, consider the real vector $\boldsymbol{v}(t)=\begin{bmatrix} A(t) \\ B(t) \end{bmatrix}\in\mathbb{R}^2$. Then for the matrix $J=\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, we see that $\boldsymbol{v}'=-J\boldsymbol{v}$. Now, think of $\mathbb{R}^2$ as simply $\mathbb{C}$, and v would be like some complex number, and J is the rotation counter-clockwise by 90 degree, i.e., multiplication by i. And we have $\boldsymbol{v}'=-\mathrm{i}\boldsymbol{v}$. So the solution is $\boldsymbol{v}(t)=\mathrm{e}^{-\mathrm{i}t}\boldsymbol{v}(0)=(\cos(t)-\mathrm{i}\sin(t))\boldsymbol{v}(0)$.

Then the solution should be $(\cos(t)I-\sin(t)J)\begin{bmatrix} A(0) \\ B(0) \end{bmatrix}=\begin{bmatrix} A(0)\cos(t)+B(0)\sin(t) \\ B(0)\cos(t)-A(0)\sin(t) \end{bmatrix}$. This is indeed the collection of all possible solutions of our system. We have solved the differential equation.

Note that the romantic relation of A and B are necessarily periodic. If you are ever trapped in a relationship which is periodic, (i.e., happy for a week, then fight for a week, and repeat), then maybe you should think about this model a bit more. ☺

8.2 Complex Orthogonality

Procedural-wise, complex linear algebra works in the same way as real linear algebra. The Gaussian elimination works the same way. The matrix multiplication formula, the trace formula and the determinant formula are all the same. Nothing new all in all. However, one thing is crucially different: inner product, and by extension, transpose.

For two real vectors $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}2\\-1\end{bmatrix}$, it is very easy to understand that they are orthogonal to each other. We can draw it, or visualize it in our mind, and so on. But for two complex vectors, what does it mean to be orthogonal to each other?

Example 8.2.1. Consider $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ and $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$. What would happen if we perform the "real dot-product" on these two vectors? We would have $1^2 + (\mathrm{i})^2 = 1 + (-1) = 0$. Huh, this vector is "orthogonal" to itself? How can it be?

It simply cannot be. Quoting Sherlock Holmes, when you have eliminated the impossible, whatever remains, however improbable, must be the truth: we used the wrong "dot product"!

There is a lesson we can learn from this. Blindly apply analogous procedures will usually lead you astray. It is always to guide your scientific exploration with proper intuitions.

What is $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$? Recall that previously, we have talked about the relation between $a + b\mathrm{i}$ and $\begin{bmatrix}a & -b\\b & a\end{bmatrix}$. Using this interpretation, let us think of $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ as $\begin{bmatrix}1 & 0\\0 & 1\\0 & -1\\1 & 0\end{bmatrix}$. So instead of one vector, it is in fact two vectors!

So what is orthogonal to $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$? Well, let us consider $\begin{bmatrix}1\\-\mathrm{i}\end{bmatrix}$. Then the two vectors can be thought of as $\begin{bmatrix}1 & 0\\0 & 1\\0 & -1\\1 & 0\end{bmatrix}$ and $\begin{bmatrix}1 & 0\\0 & 1\\0 & 1\\-1 & 0\end{bmatrix}$. Did you see that? ALL FOUR column vectors are mutually orthogonal to each other. So we conclude that $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ and $\begin{bmatrix}1\\-\mathrm{i}\end{bmatrix}$ are orthogonal to each other.

What does this mean? It means that if n-dimensional complex vectors $\boldsymbol{v}, \boldsymbol{w}$ corresponds to $2n \times 2$ real matrices A, B, then we say $\boldsymbol{v} \perp \boldsymbol{w}$ if and only if $A^{\mathrm{T}}B$ has all four entries zero.

Something funny is going on here. Note that, by interpreting i as $\begin{bmatrix}0 & -1\\1 & 0\end{bmatrix}$, we are interpreting $\boldsymbol{v} = \begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ as $A = \begin{bmatrix}1 & 0\\0 & 1\\0 & -1\\1 & 0\end{bmatrix}$. Then $A^{\mathrm{T}} = \begin{bmatrix}1 & 0 & 0 & 1\\0 & 1 & -1 & 0\end{bmatrix}$, and it does NOT represent $\boldsymbol{v}^{\mathrm{T}}$. Rather, it represents $\overline{\boldsymbol{v}}^{\mathrm{T}}$. Here the line means complex conjugates on each coordinate. So we have

$$\begin{aligned}\boldsymbol{v} \perp \boldsymbol{w} &\iff A^{\mathrm{T}}B = O\\ &\iff \overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{0}.\end{aligned}$$

In particular, the fact that $A^{\mathrm{T}}B$ is the 2×2 zero meatrix corresponds to the fact that $\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}$ is the complex number zero. ☺

Definition 8.2.2. *For two complex vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^n$, then we define their complex dot product to be $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}$.*

Remark 8.2.3. *Here is something fun to think about. If $\boldsymbol{v} \in \mathbb{C}^n$ is represented by a real $2n \times 2$ matrix A, and $\boldsymbol{w} \in \mathbb{C}^n$ is represented by a real $2n \times 2$ matrix B, then $\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}$ is represented by $A^{\mathrm{T}}B$.*

Then what is the meaning of the complex angle $\frac{\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$? Note that columns of A span a 2-dimensional plane of $\mathbb{R}^{2n}$. Similarly, columns of B span a plane of $\mathbb{R}^{2n}$ as well. Is the planar angle between the two planes related to $\frac{\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$?

A generic guideline is that, whenever you take transpose for a real matrix, in the corresponding world of complex matrices, you probably would like to take a transpose conjugate. Think of this as a generalization of the following fact: if $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ represents $a+b\mathrm{i}$, then its transpose actually represents $a-b\mathrm{i}$. For convenience, we shall use the "star" as a shorthand for conjugate transpose, i.e., we define A^* as $\overline{A}^{\mathrm{T}}$.

For example, we have the following result.

Theorem 8.2.4. *For a complex $m \times n$ matrix A, then* $\mathrm{Ran}(A)$ *and* $\mathrm{Ker}(A^*)$ *are orthogonal complements, and* $\mathrm{Ran}(A^*)$ *and* $\mathrm{Ker}(A)$ *are orthogonal complements. Oh, and* $\mathrm{Ran}(A)$ *and* $\mathrm{Ran}(A^*)$ *and* $\mathrm{Ran}(A^{\mathrm{T}})$ *have the same complex dimension, i.e., the rank of A.*

Familar yes? We have a bunch of similar results here. Note that ultimately, everything here involves an orthogonal structure, which is why conjugate transpose is used throughout. Review or read up about their real conterparts if needed.

1. A complex matrix is ***Hermitian*** if $A = A^*$. In this case, it is diagonalizable with real eigenvalues, and the underlying space has an orthogonal basis made of eigenvectors of A.
2. A complex matrix is ***skew-Hermitian*** if $-A = A^*$. In this case, it is diagonalizable with purely-imaginary eigenvalues, and the underlying space has an orthogonal basis made of eigenvectors of A.
3. A complex matrix is ***unitary*** if $A^{-1} = A^*$. In this case, it is diagonalizable with unit complex eigenvalues (complex numbers with absolute value one), and the underlying space has an orthogonal basis made of eigenvectors of A. Note that in particular, such a map would preserve the complex dot product, i.e., $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \langle A\boldsymbol{v}, A\boldsymbol{w} \rangle$.
4. A complex matrix is ***normal*** if $AA^* = A^*A$. In this case, it is diagonalizable, and the underlying space has an orthogonal basis made of eigenvectors of A.

Finally, all these type of matrices have spectral theorems, similar to real symmetric matrices. Check out Section 7.11 if you forget about them.

8.3 Fourier Matrix

Here is a family of matrices that is both super cool, extremely useful in practice, and also illustrates some funny situations mentioned above. It is the famous Fourier matrix.

For any n, let ω be the ***primitive n-th root of unity***, i.e., it is the complex number $\omega = \cos(2\pi/n) + \mathrm{i}\sin(2\pi/n)$. Then as you can check, $1, \omega, ..., \omega^{n-1}$ are all distinct complex numbers, and $\omega^n = 1$. In fact, by thinking of complex numbers as dilations and rotations, it is easy to see that $1, \omega, ..., \omega^{n-1}$ are ALL solutions to the equation $x^n = 1$ over the complex numbers.

We start by looking at the fourier matrix F_n whose (i,j) entry is $\omega^{(i-1)(j-1)}$. For a typical example, we have $F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}$.

As you can see, it appears that $F_n^{\mathrm{T}} = F_n$. However, it is NOT Hermitian. (For example, its diagonal is not real.) In fact, it is the opposite of Hermitian: it is a multiple of a unitary matrix. Feel free to perform $F_4 F_4^*$ to verify the case when $n = 4$. In particular, you can also check that $\frac{1}{n}\overline{F_n} = F_n^{-1}$.

The fourier matrix is closely related to the Fourier series and Fourier Transforms. In Calculus we learned that Fourier series is very important. For a periodic function $f(x)$ with period 2π, you can try to decompose it into different frequencies via fourier series, and write it as a linear combination of sines and cosines. Say we have maybe $f(x) = \sum c_k \mathrm{e}^{\mathrm{k i}x}$. Here note that $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$, so $\mathrm{e}^{\mathrm{i}x}$ is just a lazy way to write sine and cosine simultaneously.

Suppose we have a decomposition $f(x) = c_0 + c_1\mathrm{e}^{\mathrm{i}x} + c_2\mathrm{e}^{2\mathrm{i}x} + c_3\mathrm{e}^{3\mathrm{i}x}$. Given c_0, c_1, c_2, c_3, what do we know about the function $f(x)$? Well, if you apply F_4 to the vector $\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$, then you can verify that you have $\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix}$. As you can see, you get four points on the graph of $f(x)$. By using more fourier coefficients, and larger Fourier matrix, you will get more detailed points on your graph for $f(x)$. This is the forward direction.

But consider the backward direction as well. In pratical cases, we usually have the graph of $f(x)$ by some data gathering. How can we work out the Fourier coefficients? Suppose we have $f(x) = c_0 + c_1\mathrm{e}^{\mathrm{i}x} + c_2\mathrm{e}^{2\mathrm{i}x} + c_3\mathrm{e}^{3\mathrm{i}x}$ where the c_i are unknown. How to find the fourier coefficient of $f(x)$? We could evaluate $f(0), f(\pi/2), f(\pi), f(3\pi/2)$ empirically or experimentially, and then compute $F_4^{-1}\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix} = \frac{1}{n}\overline{F_4}\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix}$. As you can see, by evaluating at merely a few points and apply $\frac{1}{n}\overline{F_n}$, we can conveniently obtain the (approximate) Fourier coefficients. The approximation will get better as we use more data points and larger Fourier matrix.

Suppose you want to compute the first 1000 fourier coefficients (say you know the rest are probably noises or measurement errors). In effect, you want to quickly multiply F_{1000} to a known vector. Wow, that is pretty big! How should you do it? By brute fource, this is a 1000 by 1000 matrix, and calculating with it needs millions of calculations. That would take forever. So a better approach is the Fast Fourier Transfour. We start by looking at F_{1024}, reduce it to F_{512}, then reduce it to F_{256}, and so forth, until we reach $F_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. So in 10 steps, we reduce the problem to a much smaller one. In the end, one million calculations will be reduced to merely 5000 calculations. Imagine the gain in speed in signal processing and etc. This is ranked as the top 10 algorithms of the 20-th centry by the IEEE journal Computing in Science and Engineering.

Example 8.3.1. Consider $F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}$. Observe the relation between its first and third column, and between its second and forth column. You can see that the first and third coordinates of

corresponding columns are the same, and the second and forth coordinates are negated.

Let us now swap the columns to bring the original first and third column togehter, and the original second and forth column together. Then we have $F_4P_{23} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & i & -i \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -i & i \end{bmatrix}$. Hey, note that the upper left corner and lower left corner is exactly $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = F_2$! In fact, let $D_2 = diag(1, i)$, we have $F_4P_{23} = \begin{bmatrix} F_2 & D_2F_2 \\ F_2 & D_2F_2 \end{bmatrix} = \begin{bmatrix} I_2 & D_2 \\ I_2 & -D_2 \end{bmatrix} \begin{bmatrix} F_2 & 0 \\ 0 & F_2 \end{bmatrix}$. So step by step, we have extracted F_2 out of F_4! ☺

Theorem 8.3.2 (Fast Fourier Transform)**.** *We have the following decomposition, where* $D_n = (1, \omega, ..., \omega^{n-1})$ *where* $\omega = \cos(\pi/n) + i\sin(\pi/n)$*, and* P *is a matrix permuting all odd columns to the left and all even columns to the right.*

$$F_{2n} = \begin{bmatrix} I_n & D_n \\ I_n & -D_n \end{bmatrix} \begin{bmatrix} F_n & 0 \\ 0 & F_n \end{bmatrix} P.$$

Proof. Do it yourself. Same idea as Example 8.3.1. □

Example 8.3.3. Here's what happen after a recursion. You will have

$$F_{4n} = \begin{bmatrix} I_{2n} & D_{2n} \\ I_{2n} & -D_{2n} \end{bmatrix} \begin{bmatrix} I_n & D_n & 0 & 0 \\ I_n & -D_n & 0 & 0 \\ 0 & 0 & I_n & D_n \\ 0 & 0 & I_n & -D_n \end{bmatrix} \begin{bmatrix} F_n & 0 & 0 & 0 \\ 0 & F_n & 0 & 0 \\ 0 & 0 & F_n & 0 \\ 0 & 0 & 0 & F_n \end{bmatrix} P.$$

Here P is a permutation matrix that put all (1 mod 4) columns to the left, followed by the (3 mod 4) columns, followed by the (2 mod 4) columns, and followed by the (4 mod 4) columns. ☺

Proof. Do it yourself. □

Example 8.3.4. What would happen to F_{3n}? Can you do something similar? I'll leave this to yourself. ☺

Part V

Basic Matrix Analysis

Chapter 9

Jordan Canonical Form

9.1 Generalized Eigenstuff

We are moving towards Jordan canonical form. For a square matrix A, sometimes it is diagonalizable. And by doing so, we shall find all the eigenvalues and eigenvectors and so on, so that we can completely understand the behavior of this matrix. But what if we cannot diagonalize a matrix?

Maybe we can go for the next best thing, a block-diagonalization for blocks as smaller as possible. But first of all, what are block matrices? What is the geometry behind block matrices?

9.1.1 Subspace decomposition and block matrices

We start by a review of an extremely important old concept: linear independency. The following proposition is simply a recap of old results.

Proposition 9.1.1 (Alternative definitions of linear independence)**.** *For a collection of vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in V$, the following are equivalent:*

1. *For any scalars $a_1, \dots, a_n \in \mathbb{C}$, $\sum a_i \boldsymbol{v}_i = \boldsymbol{0}$ implies that all $a_i = 0$.*
2. *The subspace spanned by $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ has dimension exactly n. In particular, $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is a basis for the subspace they span.*

All these equivalent conditions can be taken as a definition, as they are all equivalent anyway. We skip the proof because they are old news. Say we have a subspace of dimension m spanned by n vectors, then we must always have $m \leq n$. If we have $m < n$, then we have more vectors than needed, so there are redundant vectors. If $m = n$, then no vector is redundant, and this is what linear independency means.

Now we generalize this concept to subspaces.

Definition 9.1.2. *We say a collection of subspaces $V_1, \dots, V_n \subseteq V$ is linearly independent if for any* $\dim(\sum V_i) = \sum(\dim V_i)$.

In short, there is no redundancy in these subspaces. These subspaces would span their sum as efficient as possible. Just by comparing linear independence of vectors with linear independence of subspaces, they are very obviously the same idea. In particular, linear independence of vectors means exactly that the lines spanned by these vectors are linearly independent as subspaces.

Proposition 9.1.3 (Alternative definitions of linear independence)**.** *For a collection of subspaces $V_1, \dots, V_n \subseteq V$, the following are equivalent:*

1. *For any $\boldsymbol{v}_1 \in V_1, \dots, \boldsymbol{v}_n \in V_n$, then $\sum \boldsymbol{v}_i = \boldsymbol{0}$ implies that all $\boldsymbol{v}_i = \boldsymbol{0}$.*

2. *The sum space $V_1 + \cdots + V_n$ has dimension exactly $\dim V_1 + \cdots + \dim V_n$. (More aethetically written as $\dim(\sum V_i) = \sum(\dim V_i)$.) In particular, pick a basis for each V_i, and put them together, we would obtain a basis for the sum space.*

Proof.

Forward direction:

Suppose for any $\boldsymbol{v}_1 \in V_1, \ldots, \boldsymbol{v}_n \in V_n$, then $\sum \boldsymbol{v}_i = \mathbf{0}$ implies that all $\boldsymbol{v}_i = \mathbf{0}$.

Our goal is to show that $\dim(V_1 + \cdots + V_n) = \dim V_1 + \cdots + \dim V_n$. Let us think inductively. We obviously have $\dim V_1 = \dim V_1$. Do we have $\dim(V_1 + V_2) = \dim V_1 + \dim V_2$? Recall that by inclusion exclusion principal, we have $\dim(V_1 + V_2) = \dim V_1 + \dim V_2 - \dim(V_1 \cap V_2)$, so we need to show that this intersection is zero. Intersection is the key for the induction steps.

For each $1 < k \leq n$, consider the subspaces $V_1 + \cdots + V_{k-1}$ and V_k. How do they intersect?

Pick any $\boldsymbol{v}_k \in V_k \cap (V_1 + \cdots + V_{k-1})$. So we have $\boldsymbol{v}_k = \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_{k-1}$ where each $\boldsymbol{v}_i \in V_i$. But our assumption then imply that all vectors involved must be zero. So $\boldsymbol{v}_k = \mathbf{0}$. In particular, $V_k \cap (V_1 + \cdots + V_{k-1}) = \{\mathbf{0}\}$. Consequently,

$$\dim(V_1+\cdots+V_k) = \dim(V_1+\cdots+V_{k-1})+\dim V_k-\dim(V_k\cap(V_1+\cdots+V_{k-1})) = \dim(V_1+\cdots+V_{k-1})+\dim V_k.$$

So we are done by induction. (For a non-inductive proof, see Proposition 7.7.4.)

Backward direction:

Suppose that $\dim(\sum V_i) = \sum(\dim V_i)$. Suppose for contradiction that we have vectors $\boldsymbol{v}_1 \in V_1, \ldots, \boldsymbol{v}_n \in V_n$, and $\sum \boldsymbol{v}_i = \mathbf{0}$, but some of them is non-zero. WLOG say $\boldsymbol{v}_n \neq 0$.

Then $\boldsymbol{v}_n = -\boldsymbol{v}_1 - \cdots - \boldsymbol{v}_{n-1}$, and it is a non-zero vector in the intersection of V_n and $V_1 + \cdots + V_{n-1}$. Hence we have $\dim(V_n \cap (V_1 + \cdots + V_{n-1})) > 0$. In particular,

$$\begin{aligned}\dim(\sum V_n) &= \dim V_n + \dim(V_1 + \cdots + V_{n-1}) - \dim(V_n \cap (V_1 + \cdots + V_{n-1})) \\ &< \dim V_n + \dim(V_1 + \cdots + V_{n-1}) \\ &\leq \dim V_n + (\dim V_1 + \cdots + \dim V_{n-1}).\end{aligned}$$

But this violates the assumption. So we are done. □

Here is a useful lemma for future.

Lemma 9.1.4. *If $V_k \cap (V_1 + \cdots + V_{k-1}) = \{\mathbf{0}\}$ for all $1 < k \leq n$, then $V_1, \ldots, V_n$ are linearly independent.*

Proof. This is a portion of the proof above. □

Example 9.1.5. On the plane $\mathbb{R}^2$, the x-axis is a subspace, and the y-axis is a subspace, and they are linearly independent.

Now, the line $x = y$ is also a subspace. But the three subspaces are NOT independent. Why? First of all, three 1-dimensional subspaces cramped into a 2-dimensional subspace? But $1 + 1 + 1 > 2$, obviously. There is simply not enough space. Some redundancy must occure.

And how would this redundancy manifest? Well, $\begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}-1\\-1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$, violating the alternative definition of linear independence. In particular, this reveals why the subspaces are not as efficient as possible. ☺

Take a long hard look of this example, and we can come to the following interesting observation.

Example 9.1.6. Two subspaces are linearly independent if and only if they have zero intersection. This is simply the inclusion exclusion principal, i.e., $\dim(V_1 + V_2) = \dim V_1 + \dim V_2 - \dim(V_1 \cap V_2)$.

However, for three subspaces or more, pairwise-independence do NOT imply collective independence.

This is not just linear algebra. It is also related to, say, probability theory. If you have many random variables, then pairwise independent random events might not be collectively independent. The following example requires some basic knowledge in probability theory. (So you can feel free to skip it.)

Consider the four vertices of a square. Say we pick a vertex randomly and uniformly, so each vertex is picked witha probability of $\frac{1}{4}$. Let X_1 be the event that the picked vertex is one of the two lower vertices. Let X_2 be the event that the picked vertex is one of the two left vertices. Let X_3 be the event that the picked vertex is either the lower left or the upper right. Then you can check that $\Pr(X_i \text{ and } X_j) = \Pr(X_i)\Pr(X_j)$ for all $i \neq j$, so these events are independent events. However, they are not collectively independent, because $\Pr(X_1 \text{ and } X_2 \text{ and } X_3) \neq \Pr(X_1)\Pr(X_2)\Pr(X_3)$.

Can you see the relation between this probability example and the last linear algebra example? They are secretly the same example. This probability example is simply the case of 2-dimensional plane over the field $\mathbb{F}_2$ instead of $\mathbb{R}$, where each line has only two points. ☺

Now, how is this concept related to diagonalizations? Here is how. Recall that diagonalization is related to eigenvalues and multiplicities. Given a matrix A and an eigenvalue λ, then $\operatorname{Ker}(A - \lambda I)$ is the eigenspace of A for λ, and its dimension is the geometric multiplicity for λ.

Proposition 9.1.7. *All eigenspaces of a matrix A are linearly independent.*

Proof. Suppose $V_1, \ldots, V_k$ are eigenspaces of A for distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Let us show that they are linearly independent. Pick any $\boldsymbol{v}_1 \in V_1, \ldots, \boldsymbol{v}_k \in V_k$ and assume that $\sum \boldsymbol{v}_i = \boldsymbol{0}$.

How to proceed? Well, all we know is that $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$. This is literally the only thing we know, so we use it. Apply A to both sides of $\sum \boldsymbol{v}_i = \boldsymbol{0}$, we have $\sum \lambda_i \boldsymbol{v}_i = \boldsymbol{0}$, another equation! Let us keep going. Apply A again, and we have $\sum \lambda_i^2 \boldsymbol{v}_i = \boldsymbol{0}$, yet another equations! We keep going and eventually we shall obtain k equations about $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Hopefully, with k equations and k unknown vectors, we should be able to solve $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ from these equations.

Now is the time to test whether you truly understand basic linear algebra. We have a system

$$\begin{aligned} \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_k &= \boldsymbol{0}, \\ \lambda_1 \boldsymbol{v}_1 + \cdots + \lambda_k \boldsymbol{v}_k &= \boldsymbol{0}, \\ &\vdots \\ \lambda_1^{k-1} \boldsymbol{v}_1 + \cdots + \lambda_k \boldsymbol{v}_k &= \boldsymbol{0}. \end{aligned}$$

But basic linear algebra means that we despise linear equations. Rather, we should write it in matrix form. Then we have

$$\begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_k \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \ldots & \lambda_1^{k-1} \\ 1 & \lambda_2 & \ldots & \lambda_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_k & \ldots & \lambda_k^{k-1} \end{bmatrix} = 0.$$

Hey, there is the famous Vandermonde matrix! And no surprise there. Just look at the Vandermond matrix. It is literally made for this: iterations of a matrix on many eigenvectors.

And we know that for distinct $\lambda_1, \ldots, \lambda_k$, the Vandermonde matrix is invertible. So $\begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_k \end{bmatrix}$ is the zero matrix. Done. □

Remark 9.1.8. *A side remark for those readers that did not know this before.*

Recall that the Vandermonde matrix $\begin{bmatrix} 1 & \lambda_1 & \ldots & \lambda_1^{k-1} \\ 1 & \lambda_2 & \ldots & \lambda_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_k & \ldots & \lambda_k^{k-1} \end{bmatrix}$ *has determinant* $\prod(\lambda_j - \lambda_i)$. *Why?*

Well, by the big formula, determinant $\det \begin{bmatrix} 1 & \lambda_1 & \ldots & \lambda_1^{k-1} \\ 1 & \lambda_2 & \ldots & \lambda_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_k & \ldots & \lambda_k^{k-1} \end{bmatrix}$ *is some polynomial involving* $\lambda_1, \ldots, \lambda_k$.

What if some $\lambda_i = \lambda_j$? Then two rows of the matrix would be identical, so the determinant is zero. Therefore, the polynomial $\lambda_j - \lambda_i$ must be a factor polynomial of the determinant.

And as it turned out, the determinant is exactly the product of all these polynomials.

Now, note that for vectors, if they are linearly independent and spanning, they form a basis. So here comes the subspace version.

Definition 9.1.9. *We write $V = V_1 \oplus \cdots \oplus V_k$ if the subspaces $V_1, \ldots, V_k$ are linearly independent and they span V. We say V is the direct sum of $V_1, \ldots, V_k$, and we also say V has a subspace decomposition $V_1 \oplus \cdots \oplus V_k$.*

Example 9.1.10. For a matrix A, it is diagonalizable if and only if the domain is the direct sum of eigenspaces of A.

We already know that these eigenspaces are always independent. So we need to show that A is diagonalizable if and only if these eigenspaces are spanning.

Say A is $n \times n$ and all its eigenspaces are $V_1, \ldots, V_k$. Now, A is diagonalizable if and only if all geometric multiplicities adds up to n. This means that $\sum \dim V_1 = n$, which in turn means that $\dim \sum V_i = n$. But this means that $\sum V_i$ is the whole domain. (The logical connection in each step is "if and only if", as you can check easily.) ☺

Example 9.1.11. $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is a basis for V if and only if $V = \bigoplus \operatorname{span}(\boldsymbol{v}_i)$. ☺

Example 9.1.12. Consider $\mathbb{R}^{a+b}$. Sometimes we write $\mathbb{R}^{a+b} = \mathbb{R}^a \oplus \mathbb{R}^b$. This is technically a very ambiguous statement, but people still write this because it looks pretty. What does this mean?

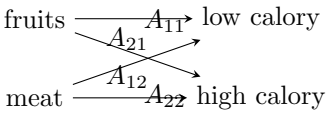
For $\mathbb{R}^{a+b}$, each vector has $a+b$ coordinates. Let $\mathbb{R}^a$ represent the subspace of vectors of the form $\begin{bmatrix} x_1 \\ \vdots \\ x_a \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. You can see why I name it so, because this subspace is pretty much just $\mathbb{R}^a$. Similarly, let $\mathbb{R}^b$ represent the subspace of vectors of the form $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ y_1 \\ \vdots \\ y_b \end{bmatrix}$. Then $\mathbb{R}^{a+b} = \mathbb{R}^a \oplus \mathbb{R}^b$, which is very easy to verify. ☺

Now, consider a linear map $A : V \to W$. When V, W decompose into subspaces, then the linear map A would also decompose into "submaps".

Example 9.1.13. Consider a map sending foods to nutrients. Say we have foods: apples, bananas, meat. And we have nutrients: fibers, proteins, suger. Then this map is a matrix A, such that if we have $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ apples, bananas and meat, then we have the vector $A \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ representing a linear combination of fibers, proteins and suger. Obviously A is a 3 by 3 matrix. Say $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. As you can see, the nutrition table is a linear map. it sends linear combinations of foods into linear combinations of nutrients.

Now consider the block form $A = \left[\begin{array}{cc|c} 1 & 2 & 3 \\ 4 & 5 & 6 \\ \hline 7 & 8 & 9 \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where A_{ij} represent the corresponding blocks. Then each block is a "smaller nutrient table", using only some fruits and some nutrients. For

example, A_{11} only concerns fruits, and as a linear map, it sends fruits to the total amount of sugar they contain. It is a "submap". In total, we have four submaps.



As you can see, the domain decomposes into two subspaces, one spanned by fruits, and one spanned by meat. And the codomain decomposes into two subspaces, one spanned by high calory nutrients, and one spanned by low calory nutrients. As a result, the linear map breaks down into four submaps, from each domain subspace to each codomain subspace.

More computationally, note that $A\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}x+2y+3z\\4x+5y+6z\\7x+8y+9z\end{bmatrix}$. Now A_{11} is about how x, y contributs to the first output coordinat, and we have $A_{11}\begin{bmatrix}x\\y\end{bmatrix} = x+2y$. Similarly, we have $A_{12}(z) = 3z, A_{21}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}4x+5y\\7x+8y\end{bmatrix}$, and $A_{22}(z) = \begin{bmatrix}6z\\9z\end{bmatrix}$.

This corresponds to the decomposition

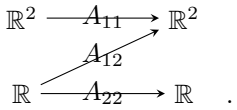
$$\begin{bmatrix}x+2y+3z\\4x+5y+6z\\7x+8y+9z\end{bmatrix} = \begin{bmatrix}x+2y\\0\\0\end{bmatrix} + \begin{bmatrix}3z\\0\\0\end{bmatrix} + \begin{bmatrix}0\\4x+5y\\7x+8y\end{bmatrix} + \begin{bmatrix}0\\6z\\9z\end{bmatrix}.$$

Also check out how the vertical lines splitting the block matrix corresponds to domain decomposition, while the horizontal lines splitting the block matrix corresponds to codomain decomposition. ☺

Intuitively, when we have a block matrix, we are grouping input coordinates and output coordinates. The block A_{ij} records how the j-th group of inputing coordinates effect the i-th group of outputing coordinates.

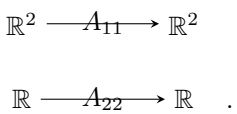
Example 9.1.14. Consider $\left[\begin{array}{cc|c}1&1&1\\1&1&2\\\hline 0&0&1\end{array}\right]$. Note that the lower left block is zero. This means the first two input coodinates does NOT effect the third output coordiante.

Indeed we have $\left[\begin{array}{cc|c}1&1&1\\1&1&2\\\hline 0&0&1\end{array}\right]\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}x+y+z\\x+y+2z\\z\end{bmatrix}$.



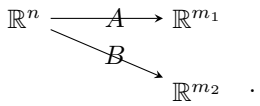
This is a ***block upper triangular matrix***.

In particular, block diagonal means each groups of coordinates only effect themselves. In particular, instead of one system, it is more like many separate independent systems, one for each diagonal block. Here is a picture for $\left[\begin{array}{cc|c}\frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}&0\\\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&0\\\hline 0&0&2\end{array}\right]$, which is a ***block diagonal matrix***.

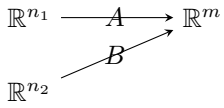


As you can see, a block diagonal matrix happens exactly when the two "linear submaps" are independent of each other. ☺

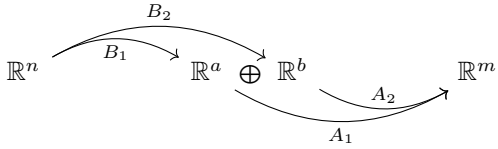
So here is how one can think about block matrices. For example, for the block matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ where A is $m_1 \times n$ and B is $m_2 \times n$, we can think of it as this:



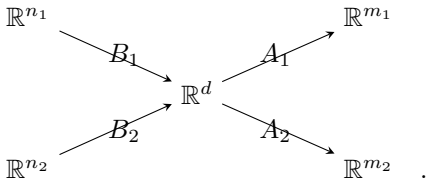
And for the block matrix $\begin{bmatrix} A & B \end{bmatrix}$ where A is $m \times n_1$ and B is $m \times n_2$, we can think of it as this:



Now, why would the block matrices multiply exactly as regular matrices? Let us reprove this via more diagrams. We have $\begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1B_1 + A_2B_2$ because of this:



And we have $\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2 \end{bmatrix}$ because of this:

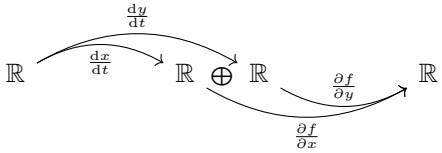


If you like, you can now prove the block multiplication formula by simply drawing these diagrams. Here is another interesting related example.

Example 9.1.15. In multi-variable calculus, we consider functions $f : \mathbb{R}^2 \to \mathbb{R}$, so we consider $f(x, y)$ for two input x, y. We write $\frac{\partial f}{\partial x}$ to mean the derivative with respect to x while holding y constant, and similarly we write $\frac{\partial f}{\partial y}$ accordingly. You will learn about this extensively in multivariable calculus class. They are the "partial derivatives" of f.

However, sometimes x, y are both functions of t, say $x(t), y(t)$. So we have a new function $f(x(t), y(t))$ which is a function of t. What if we are interested in how a change of t would influence f?

The formula is called multivariable chain rule. We have $\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}$. Why is this formula true? Because of the following diagram.



☺

We use mostly $\mathbb{R}^n$ here for visual purpose, but it does not really matter. Replace them all by $\mathbb{C}^n$ if you like.

Now, for block matrices, the domain and codomains must be decomposed by grouping the coordinates, i.e., those that looks like $\mathbb{R}^{a+b} = \mathbb{R}^a \oplus \mathbb{R}^b$. What if they are decomposed in some other way? It turns out that this does not matter much. For any decomposition $V = V_1 \oplus V_2$, we can always change basis. Simply pick a basis for V_1, followed by a basis for V_2. Together this is a basis for V. Then treating this basis as the standard basis, then $V = V_1 \oplus V_2$ would look exactly like $\mathbb{R}^{a+b} = \mathbb{R}^a \oplus \mathbb{R}^b$.

For example, the plane V can decompose as the direct sum of any two linearly independent lines V_1, V_2. Pick any $\boldsymbol{v}_1 \in V_1, \boldsymbol{v}_2 \in V_2$, then $\boldsymbol{v}_1, \boldsymbol{v}_2$ form a basis of V. Using these as the standard basis, $V = V_1 \oplus V_2$ is now exactly $\mathbb{R}^2 = \mathbb{R} \oplus \mathbb{R}$. (Here the first $\mathbb{R}$ is the x-axis and the second $\mathbb{R}$ is the y-axis.)

One can also endeavor to do this abstractly. For a linear map $L : V \to W$, say $V = \bigoplus V_i$ and $W = \bigoplus W_j$, then we define $L_{ij} : V_i \to W_j$ that sends vectors $\boldsymbol{v}$ in V_i to the W_j-component of $L\boldsymbol{v}$. Then L is a "block map" with submaps L_{ij} as the "blocks".

Example 9.1.16 (Optional Example)**.** Recall that we say V is the direct sum of its subspaces V_1, V_2. There are four canonical linear maps involved in this structure.

First of all, we have an inclusion map $\iota_1 : V_1 \to V$ and $\iota_2 : V_2 \to V$. These maps don't change the input at all, but their codomain is larger than the domain. They tell us how the smaller spaces (the domains) is included in the bigger space (the codomain).

Now since $V = V_1 \oplus V_2$, each vector $\boldsymbol{v} \in V$ has a UNIQUE decomposition $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ such that $\boldsymbol{v}_i \in V_i$. (This is because the subspaces are independent. Prove it yourself.) So we also have two projection maps $p_1 : V \to V_1$ and $p_2 : V \to V_2$ such that $p_i(\boldsymbol{v}) = \boldsymbol{v}_i$. These are INDEED projection maps. For example, note that for any $\boldsymbol{v}_1 \in V_1$, then $\boldsymbol{v}_1 = \boldsymbol{v}_1 + \boldsymbol{0}$ must be the unique decomposition according to $V = V_1 \oplus V_2$. Therefore $p_1(\boldsymbol{v}_1) = \boldsymbol{v}_1$. In particular, $p_i^2 = p_i$. (This is the defining algebraic property for projections in any mathematical context.) However, these are NOT necessarily orthogonal projections. They could be oblique projections. See last semester's note for oblique projections. (They are only orthogonal projections when $V_1 \perp V_2$. Otherwise they are oblique projections, where p_i preserves V_i and kills V_j for $j \neq i$.)

Now if we have a linear map $L : V \to W$, and decompositions $V = V_1 \oplus V_2$ and $W = W_1 \oplus W_2$. Then there are four possible linear maps induced from these structures. We can restrict the domain of L to V_i and project the codomain to W_j, and obtain $L_{ij} = p_j \circ L \circ \iota_i : V_i \to W_j$. Then we can write $L = \begin{bmatrix} L_{11} & L_{21} \\ L_{12} & L_{22} \end{bmatrix}$. For each $\boldsymbol{v} \in V$, if the unique decomposition according to $V = V_1 \oplus V_2$ is $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$, then let us write it as $\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix}$, and and we do similar things in W. Then we shall see that $\begin{bmatrix} L_{11} & L_{21} \\ L_{12} & L_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} (L\boldsymbol{v})_1 \\ (L\boldsymbol{v})_2 \end{bmatrix}$. ☺

These whole venture is purely philosophical, and you need to feel no pressure to master these abstract computations. My goal is to address the following question: What is the idea behind a block matrix? It means that as we decompose domain and codomain into subspaces, the linear map is decomposed into submaps. The "blocks" are actually "submaps", or restrictions of the original linear map to corresponding subspaces.

9.1.2 Invariant decompositions and diagonalizations

Which subspace decomposition should we choose to facilitate diagonalization? We need to first understand why we need diagonalization.

Why are diagonal matrices neat? Consider $\begin{bmatrix} d_1 & & \\ & d_2 & \\ & & d_3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} d_1 a_1 \\ d_2 a_2 \\ d_3 a_3 \end{bmatrix}$. As you can see, for a diagonal matrix, treated as a linear map, it acts on each coordinate independently. The i-th coordinate of the output depends only on the i-th coordinate of the input, and vice versa, the i-th coordinate of the input will influence only the i-th coordinate of the output. Coordinates will NOT cross-influence each other, they just each do their own thing during this linear map.

In particular, for block diagonalization of a linear map $A : V \to V$, we are seeking a decomposition of the domain $V = V_1 \oplus \cdots \oplus V_K$ such that A only sends each subspace V_i into V_i itself.

In general, we can define the following:

Definition 9.1.17. *We say a subspace W of a space V is an **invariant subspace** of the linear transformation $L : V \to V$ if $L(W) \subseteq W$. (We do NOT require them to be equal. The point is such that L can be restricted to a linear transformation on W.)*

We say a decomposition $V = V_1 \oplus V_2$ is an invariant decomposition for the linear transformation $L : V \to V$ if both V_1 and V_2 are invariant subspaces.

Indeed, given a diagonalizable matrix, how would we diagonalize it? We need to find eigenvectors. Each eigenvector is like an invariant direction that the matrix must preserve, i.e., each vector spans a line which is an 1-dimensional invariant subspace. Now our matrix acts on each invaraint direction independently, so if we pick a basis made of eigenvectors, then our matrix after a corresponding change of basis will be diagonal.

Proposition 9.1.18. *Given an invariant decomposition $V = V_1 \oplus V_2$ for the linear transformation $L : V \to V$, then the corresponding block structure for L is block diagonal. (I only used two subspaces here, but the case for more subspaces is identical.)*

Proof. Since $L(V_i) \subseteq V_i$, therefore for any $\boldsymbol{v} \in V_i$ and any $j \neq i$, the V_j-component of $L\boldsymbol{v}$ must be zero. So $L_{ij}(\boldsymbol{v}) = \boldsymbol{0}$ for all $\boldsymbol{v} \in V_i$. □

In particular, to block diagonalize a matrix is exactly the same as to find invariant subspace decompositions of the domain. Let us see a concrete example of this, using the same example as before.

Example 9.1.19. Consider a rotation in $\mathbb{R}^3$ around the line $x = y = z$ that sends the positive x-axis to the positive y-axis, and the positive y-axis to the positive z-axis, and the positive z-axis to the positive x-axis. We know its linear map has matrix $R = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. This matrix has non-real eigenvalues, so there is NO REAL diagonalizations. However, maybe we can find a REAL block-diagonalization?

There are two invariant subspaces that R must act on. One is the axis of rotation, the line $x = y = z$. This is a one-dimensional subspace V_1 spanned by $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. R acts on V_1 by simply fixing everyone, i.e., via the 1×1 matrix $R_{11} = \begin{bmatrix} 1 \end{bmatrix}$.

The other is the orthogonal complement of V_1, the subspace V_2 of all vectors $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ such that $x+y+z = 0$. Say we pick basis $\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$. Our linear map acts on V_2 as a rotation of $\frac{2\pi}{3}$, i.e., via some 2×2 matrix R_{22}. To find the matrix $R_{22} : V_2 \to V_2$, note that it depends on the basis we have chosen for V_2!!! So this is NOT going to be the standard rotation matrix, because we forgot to pick an orthonormal basis. Oops. Nevermind, let us just keep going forward.

Using the basis $\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ and $\boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ for V_2, note that $R\boldsymbol{v}_1 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \boldsymbol{v}_2 - \boldsymbol{v}_1$, and $R\boldsymbol{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = -\boldsymbol{v}_1$. So $R_{22} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix}$.

So, under the basis $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$, our matrix will change into $\begin{bmatrix} R_{11} & \\ & R_{22} \end{bmatrix} = \begin{bmatrix} 1 & & \\ & -1 & -1 \\ & 1 & 0 \end{bmatrix}$, which is block diagonal.

So we have $R = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & -1 & -1 \\ & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}^{-1}$.

Of course, as we can see in hind-sight, we can also find an orthonormal basis for V_2, say $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ and $\frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$. Then R_{22} will be the standard rotation matrix $\begin{bmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}$.

So we have$R = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 1 & & \\ & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ & \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix}^{-1}$. We saved a bit of calculations but the numbers are uglier. Also note that the inverse here is also easy to calculate, because that matrix is now an orthogonal matrix, courtesy of picking an orthonormal basis. So the inverse here is just a transpose. In practise, this alone will make this better than the previous calculation, despite the ugly entries. ☺

We want to decompose the domain into as smaller invariant subspaces as possible. Ideally, if we can decompose our domain as the direct sum of many 1-dimensional invariant subspaces, then we have completely diagonalized the matrix. These 1-dimensional subspaces would then corresponds to eigenvectors.

9.1.3 Searching for good invariant decomposition

So this is it. How can we find a good invariant decomposition? Let us first see what kinds of invariant subspaces we have.

Example 9.1.20. Given any matrix A, consider the zero space $\operatorname{Ker}(A)$. obviosuly $A(\operatorname{Ker}(A)) = \{0\} \subseteq \operatorname{Ker}(A)$. So this is indeed an invariant subspace!

Dually, since A sends everything into $\operatorname{Ran}(A)$ by definition, we have $A(\operatorname{Ran}(A)) \subseteq \operatorname{Ran}(A)$ as well. Hooray! Another invariant subspace!

In fact, for $n \times n$ matrices A, we also have $\dim\operatorname{Ker}(A) + \dim\operatorname{Ran}(A) = n$. This is a really good omen. In fact, consider say $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. Then $\operatorname{Ker}(A)$ and $\operatorname{Ran}(A)$ are both invaraint subspaces, and in fact we have $\mathbb{R}^3 = \operatorname{Ker}(A) \oplus \operatorname{Ran}(A)$ in this case, a perfect decomposition into invariant subspaces!

Unfortunately, we do not always have this. Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then $\operatorname{Ker}(A) = \operatorname{Ran}(A)$. So we failed in this case.

In fact, the best complement subspace for $\operatorname{Ker}(A)$ is actually $\operatorname{Ran}(A^{\mathrm{T}})$ (or $\operatorname{Ran}(A^*)$ in the complex case), and we always have $\mathbb{R}^n = \operatorname{Ker}(A) \oplus \operatorname{Ran}(A^{\mathrm{T}})$. However, again consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, you shall see that $\operatorname{Ran}(A^{\mathrm{T}})$ is usually not an invaraint subspace!

We are screwed either way. ☺

What can we do then? Well, recall our original motivation of doing diagonalization. What started us on this path about eigenstuff and diagonalization? The original motivation is to understand iterated applications of the same matrix, i.e., the eventual behavior of the sequence $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^n\boldsymbol{v}, \ldots$. Diagonalization gives us a quick way to calculate A^n for large n.

As a result, maybe we shouldn't focus on the *immediate* kernel and range of A. Rather, we should focus on the *eventual* kernel and range of A.

Example 9.1.21. Consider $A = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$. Then applying A repeatedly, we have:

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \overset{A}{\longmapsto} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \overset{A}{\longmapsto} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \overset{A}{\longmapsto} \mathbf{0}.$$

Then we say $\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ is *eventually* killed by A. Let N_∞ be the subspace of all vectors eventually killed by A.

Also note that $A^2 = \begin{bmatrix} 0 & 0 & 1 & \\ & 0 & 0 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$ and $A^n = \begin{bmatrix} 0 & 0 & 0 & \\ & 0 & 0 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$ for all $n \geq 3$. So eventually, $A^n\boldsymbol{v}$ will be a multiple of $\boldsymbol{e}_4$ for large enough n. So we say the *eventual* range of A is the subspace R_∞ spanned by $\boldsymbol{e}_4$.

Check yourself that in fact $\mathbb{R}^4 = N_\infty \oplus R_\infty$ is an invariant decomposition. ☺

Definition 9.1.22. *Given a linear map or a matrix A, we define $N_\infty(A) = \cup_{k=1}^{\infty} \operatorname{Ker}(A^k)$ and $R_\infty(A) = \cap_{k=1}^{\infty} \operatorname{Ran}(A^k)$.*

In particular, $\boldsymbol{v} \in N_\infty(A)$ if and only if some powers of A will kill $\boldsymbol{v}$. And $\boldsymbol{v} \in R_\infty(A)$ if and only if $\boldsymbol{v}$ is in the range of ALL powers of A.

Example 9.1.23. Let V be the space of smooth real functions. Let D be the map of "taking derivative". Can you show that $N_\infty(D)$ is the space of all polynomials?

☺

It turns out that, for finite dimensional spaces, we don't really have to look at all powers of A. Whatever A kills, then A^2 must kill as well. So as k grows, the subspace $\operatorname{Ker}(A^k)$ will be non-decreasing. However, its dimension is at most n (the dimension of the domain). So it cannot grow forever, and eventually it must stabilize. So we see that $N_\infty(A) = \operatorname{Ker}(A^k)$ for some k. (Note that this is only true for finite dimensional spaces, otherwise as k grows, the subspaces $\operatorname{Ker}(A^k)$ may grow larger and larger forever.)

We in fact have something stronger. It turns out that k does not need to be too large.

Lemma 9.1.24. *For any $n \times n$ matrix A, if $\operatorname{Ker}(A^k) = \operatorname{Ker}(A^{k+1})$ for some natural number k, then $\operatorname{Ker}(A^{k+1}) = \operatorname{Ker}(A^{k+2})$.*

Proof. Suppose $\operatorname{Ker}(A^k) = \operatorname{Ker}(A^{k+1})$. Pick any $\boldsymbol{v} \in \operatorname{Ker}(A^{k+2})$, then $\mathbf{0} = A^{k+2}\boldsymbol{v} = A^{k+1}(A\boldsymbol{v})$. But the assumption of $\operatorname{Ker}(A^k) = \operatorname{Ker}(A^{k+1})$ means that $A^{k+1}(A\boldsymbol{v}) = \mathbf{0}$ if and only if $A^k(A\boldsymbol{v}) = \mathbf{0}$, and hence $A^{k+1}\boldsymbol{v} = \mathbf{0}$. So $\boldsymbol{v} \in \operatorname{Ker}(A^{k+1})$. We are done. □

Proposition 9.1.25. *For any $n \times n$ matrix A, the chain of subspaces $\operatorname{Ker}(A^0) \subseteq \operatorname{Ker}(A^1) \subseteq \operatorname{Ker}(A^2) \subseteq \ldots$ is strictly growing at first, and then completely stabilizes. In particular, we have $N_\infty(A) = \operatorname{Ker}(A^k)$ for some $k \leq n$. (In particular, we always have $N_\infty(A) = \operatorname{Ker}(A^n)$.)*

Proof. Let k be the smallest natural number such that $\operatorname{Ker}(A^k) = \operatorname{Ker}(A^{k+1})$. Then obviously the chain $\operatorname{Ker}(A^0) \subseteq \operatorname{Ker}(A^1) \subseteq \cdots \subseteq \operatorname{Ker}(A^k)$ is strictly growing by definition of k. In particular, $\dim \operatorname{Ker}(A^k) \geq k$. Since all dimensions cannot exceed n, we have $k \leq n$.

Then by the lemma above, $\operatorname{Ker}(A^k) = \operatorname{Ker}(A^{k+1})$ implies that the chain will stabilize forever after. So $N_\infty(A) = \operatorname{Ker}(A^k)$. □

Now, let us turn our attention to R_∞. This is the opposite of N_∞. If you recall, for any $A : \mathbb{C}^n \to \mathbb{C}^n$, we have $n = \dim \mathrm{Ran}(A) + \dim \mathrm{Ker}(A)$. This duality manifest in R_∞ and N_∞ as well.

Proposition 9.1.26. *$N_\infty(A) = \mathrm{Ker}(A^k)$ if and only if $R_\infty(A) = \mathrm{Ran}(A^k)$.*

Proof. Note that as k increases, $\mathrm{Ran}(A^k)$ is a non-increasing chain of subspaces. But since $\dim \mathrm{Ran}(A^k) = n - \dim \mathrm{Ker}(A^k)$, we see that $\dim \mathrm{Ran}(A^k)$ must stabilize as soon as $\dim \mathrm{Ker}(A^k)$ stabilizes, and hence that $\mathrm{Ran}(A^k)$ must stabilize as soon as $\mathrm{Ker}(A^k)$ stabilizes. □

However, unlike $\mathrm{Ran}(A)$ and $\mathrm{Ker}(A)$ which may intersect, R_∞ and N_∞ will never intersect in a non-zero manner.

Theorem 9.1.27 (The Ultimate Invariant Decomposition)**.** *For any $n \times n$ matrix A, we have an invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$.*

Proof. Note that for some $k \leq n$ we have $N_\infty(A) = \mathrm{Ker}(A^k)$ and $R_\infty(A) = \mathrm{Ran}(A^k)$. It is straightforward to verify that $\mathrm{Ker}(A^k), \mathrm{Ran}(A^k)$ are A-invariant subspaces, so we skip that. Also by rank nullity of A^k, we can see that $\dim N_\infty(A) + \dim R_\infty(A) = n$. So we only need to show that they have zero intersection. (Can you see why this is enough?)

Suppose $\boldsymbol{v} \in N_\infty(A) \cap R_\infty(A)$. Since $\boldsymbol{v} \in N_\infty(A)$, we have some $k \leq n$ such that $A^k \boldsymbol{v} = \boldsymbol{0}$. But since $\boldsymbol{v} \in R_\infty(A) \subseteq \mathrm{Ran}(A^n)$, we have $\boldsymbol{v} = A^n \boldsymbol{w}$ for some $\boldsymbol{w}$. Then $A^{k+n}\boldsymbol{w} = \boldsymbol{0}$, so $\boldsymbol{w} \in N_\infty(A)$ as well. But this implies that $\boldsymbol{w} \in \mathrm{Ker}(A^n)$, and hence $\boldsymbol{v} = A^n \boldsymbol{w} = \boldsymbol{0}$. Oops. So we are done.

(Essentially, the key idea is that $N_\infty(A)$ stabilizes after finitely many steps, while $\boldsymbol{v} \in R_\infty(A)$ means we can realize $\boldsymbol{v}$ after arbitrarily many steps, which forces $\boldsymbol{v} \in N_\infty(A)$ to be zero.) □

In particular, we see that R_∞ is truely the "opposite" of N_∞. Here is an alternative way to see this. N_∞ is the maximal subspace collecting all 0-eigenstuff of A, and R_∞ is the maximal subspace collecting all the non-zero eigenstuff of A. In particular, R_∞ is a maximal subspace on which A is invertible.

In this sense, the ultimate decomposition is simply trying to separate the zero-eigenstuff of A and the non-zero-eigenstuff of A.

Proposition 9.1.28. *Consider a linear map $A : V \to V$. Then if $W \subseteq V$ is an A-invariant subspace such that the restriction $A|_W : W \to W$ is invertible, then $W \subseteq R_\infty(A)$. In particular, $R_\infty(A)$ is the largest A-invariant subspace on which A is invertible.*

Proof. If A restricted to W is invertible, then $A(W) = W$. Hence inductively, $A^k(W) = W$ for all k, and thus $W \subseteq \mathrm{Ran}(A^k)$ for all k. Thus $W \subseteq R_\infty(A)$.

Now we show that A restricted to $R_\infty(A)$ is invertible. Note that $R_\infty(A) \cap \mathrm{Ker}(A) \subseteq R_\infty(A) \cap N_\infty(A) = \{\boldsymbol{0}\}$. So A restricted to $R_\infty(A)$ has zero kernel. Hence it is invertible on R_∞. □

Corollary 9.1.29. *If an $n \times n$ matrix A has all eigenvalues zero, then $A^k = 0$ for some natural number $k \leq n$.*

Proof. Block diagonalize A according to the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. Note that the diagonal block corresponding to $R_\infty(A)$ will be invertible, so it will have non-zero eigenvalues, unless that block does not exist, i.e., $R_\infty(A)$ is zero-dimensional. Then $\mathbb{C}^n = N_\infty(A)$, and $A^k = 0$ for some natural number $k \leq n$. □

Proposition 9.1.30. *Consider a linear map $A : V \to V$. Then if $W \subseteq V$ is an A-invariant subspace such that the restriction $A|_W : W \to W$ has all eigenvalues zero, then $W \subseteq N_\infty(A)$. In particular, $N_\infty(A)$ is the largest A-invariant subspace on which A has all eigenvalues zero.*

Proof. Since A restricted to W has all eigenvalues zero, therefore A^k will restrict to the zero map on W for some natural number $k \leq n$. So $W \subseteq \mathrm{Ker}(A^k) \subseteq N_\infty(A)$.

Now we also know that $N_\infty(A) = \mathrm{Ker}(A^k)$ for some natural number $k \leq n$. Let us denote this map as $A|_N : N_\infty(A) \to N_\infty(A)$, then $(A|_N)^k = 0$, and hence any eigenvalue λ of $A|_N$ must satisfy $\lambda^k = 0$. Hence $\lambda = 0$. □

So as we perform the decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$ and change basis accordingly, A will be transformed into a block diagonal matrix $\begin{bmatrix} A_N & \\ & A_R \end{bmatrix}$ where A_N has all eigenvalues zero, and A_R is invertible.

Now, here is one last example of caution. Above analysis only works for finite dimensional spaces.

Example 9.1.31. Let V be the space of all smooth real functions. Let $D : V \to V$ be the map of "taking derivative". Then I claim that $N_\infty(D) \subseteq R_\infty(D)$.

For any k and any polynomial $p(x)$, I claim that $p(x) \in \text{Ran}(D^k)$. Why? Because I can simply integrate $p(x)$ k-times to get some function $P(x)$ such that $P^{(k)}(x) = p(x)$. Hence $p(x) \in \text{Ran}(D^k)$ for all k and thus $p(x) \in R_\infty(D)$. ☺

9.1.4 (Review) Polynomials of Matrices

It has come to my attention that some of our classmates have never seen this. So let us do it here as a review. Note that everything in this section could be over $\mathbb{R}$ or over $\mathbb{C}$, it does not matter much.

Remark 9.1.32. *This remark is not necessary. Feel free to skip this remark entirely.*

Let us define what a polynomial is.

We define a real (or complex) polynomial $p(x)$ toi be a finite sequence of real (or complex) numbers, say $(a_0, \ldots, a_n)$. We also write $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ where the symbol x^k has no specific meaning, and it is simply a place holder.

We add polynomial such that $(a_0, \ldots, a_n) + (b_0, \ldots, b_m) = (a_0 + b_0, \ldots, a_n + b_n, b_{n+1}, \ldots, b_m)$ if $m > n$. We multiply polynomial such that $(a_0, \ldots, a_n)(b_0, \ldots, b_m) = (c_0, \ldots, c_{m+n})$ where $c_k = \sum_{i=0}^{k} a_i b_{k-i}$.

Now, see if you can prove the following:

All polynomials form a vector space V, with a basis $1, x, x^2, \ldots$. For any bilinear map $m : V \times V \to V$ such that $m(x^a, x^b) = x^{a+b}$, then m must be the polynomial multiplication as in our definition.

You do NOT need to remember the formula, or worry about this definition. I want you to see this definition NOT because it is useful. It is not. Writing $p(x) = 4 + 2x + 3x^2$ is strictly better than writing $(4, 2, 3)$.

However, this definition makes clear of the fact that a polynomial does NOT need x to have any meaning. It could be a real number, a complex number, a matrix, a whatever. We can give whatever meaning to x, and as long as x is capable of having a "power structure", then we can define $p(x)$ accordingly as the linear combination of corresponding powers.

Here by power structure, it means that we want x^k to be defined, and we want the property that $x^a x^b = x^{a+b}$.

What is a polynomial, say $p(x) = 4 + 2x + 3x^2$? Well, in the realm of linear algebra, the best answer is that "a polynomial is a linear combination of powers." In our case, $p(x)$ is a linear combination of $1, x, x^2$. (Note that $1 = x^0$, if you like.)

For each square matrix A, we obviously have well-defined powers of A. Therefore, if $p(x)$ is some linear combination of powers of x, we can define $p(A)$ to be the corresponding linear combination of powers of A. Easy peasy.

Proposition 9.1.33. *For any polynomials $p(x), q(x)$, and any square matrix A, then $p(A)+q(A) = (p+q)(A)$ and $p(A)q(A) = (pq)(A)$. (Here $(p + q)(x)$ is the polynomial $p(x) + q(x)$ and $(pq)(x)$ is the polynomial $p(x)q(x)$.)*

Proof. DIY. □

Now, why do we study polynomials of matrices? It is mainly because powers A^k has many good properties related to A, and thus linear combinations of these powers, $p(A)$, would also share such properties. Here let us write some.

Proposition 9.1.34. *For any polynomials $p(x), q(x)$, we have $p(A)q(A) = q(A)p(A)$.*

Proof. First, note that $AA^k = A^{k+1} = A^kA$. Therefore A commutes with powers of A. Therefore A commutes with linear combinations of powers of A, i.e., polynomials of A.

So $p(A)$ commutes with A. Therefore $p(A)$ commutes with powers of A. Therefore $p(A)$ commutes with linear combinations of powers of A, i.e., other polynomials of A, say $q(A)$. So $p(A)q(A) = q(A)p(A)$. □

We also have good results about eigenstuff.

Proposition 9.1.35. *$A\boldsymbol{v} = \lambda\boldsymbol{v}$ implies that $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v}$.*

Proof. If $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then it is easy to see that $A^k\boldsymbol{v} = \lambda^k\boldsymbol{v}$. Now we take linear combinations of various powers, we see that $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v}$. □

Corollary 9.1.36. *If A has eigenvalues $\lambda_1, \ldots, \lambda_n$ counting algebraic multiplicity, then $p(A)$ has eigenvalues $p(\lambda_1), \ldots, p(\lambda_n)$ counting algebraic multiplicity. And each eigenvector of A for some eigenvalue λ is an eigenvector of $p(A)$ for the eigenvalue $p(\lambda)$.*

Now, the eigenvectors of A are all eigenvectors of $p(A)$, but sometimes $p(A)$ has other eigenvectors.

Example 9.1.37. Consider $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Its eigenvectors are vectors on the coordinate-axes. But $A^2 = I$, so ALL vectors are eigenvectors of A^2. As you can see, this is because distinct eigenvalues of A are collapsed into the same eigenvalue of $p(A)$.

Also consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Its eigenvectors are vectors on the x-axis. But $A^2 = O$, so ALL vectors are eigenvectors of A^2. As you can see, this is because A cannot be diagonalized (non-trivial Jordan block....), yet A^2 kills the obstruction to diagonalization (chopped down the bad Jordan block into smaller blocks, i.e., 1×1 blocks), so now A^2 CAN be diagonalized. ☺

So how can we find all eigenvectors of $p(A)$? Under some special cases, the answers are easy.

Proposition 9.1.38. *Suppose A can be diagonalized. Pick any polynomial $p(x)$. For any eigenvalue λ of $p(A)$, let $\lambda_1, \ldots, \lambda_k$ be all eigenvalues of A such that $p(\lambda_i) = \lambda$. Then $p(A)\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $\boldsymbol{v}$ is a linear combination of eigenvectors of A for the eigenvalues $\lambda_1, \ldots, \lambda_k$.*

Proof. Diagonalize $A = BDB^{-1}$. Then $p(A) = Bp(D)B^{-1}$. So up to a change of basis, we can assume that A is diagonal. Then DIY. □

We can have more results if we delve into theory of polynomials. The following are entirely optional. Read on if you like.

Example 9.1.39. Skip this example if you know about the Euclidean algorithm for coprime integers. Otherwise, read on.

Consider 22 and 15. They have no common prime factor. They are coprime.

We divide 22 by 15, and we shall get a remainder. We have $22 = 15 + 7$. Next we divide 15 by 7 and get $15 = 7 \times 2 + 1$. So we eventually reduced to the remainder 1.

Putting these together, we have $1 = 15 - 2 \times 7 = 15 - 2 \times (22 - 15) = 3 \times 15 - 2 \times 22$. So an integer-linear combination of 15 and 22 gives 1. This process is called the Euclidean algorithm, and it shows that two numbers x, y are coprime if and only if we can find integers a, b such that $ax + by = 1$.

Now we do the same thing for polynomials. Note that the polynomial $p(x) = x^3 + 3x^2 + 3x + 1$ and $q(x) = x^2 - 3x + 2$ has no common root, i.e., upon factorization, they shall have no common non-constant factor. They are coprime polynomials.

We divide $x^3 + 3x^2 + 3x + 1$ by $x^2 - 3x + 2$, and we shall get a remainder. We have $x^3 + 3x^2 + 3x + 1 = (x^2 - 3x + 2)(x + 6) + (19x - 11)$. Next we divide $x^2 - 3x + 2$ by $19x - 11$, and we have $x^2 - 3x + 2 = (19x - 11)(\frac{1}{19}x + \frac{46}{19}) + \frac{544}{19}$. So we eventually reduced to a constant remainder $\frac{544}{19}$.

Putting these together, we have $1 = \frac{19}{544}\frac{544}{19} = \frac{19}{544}((x^2 - 3x + 2) - (19x - 11)(\frac{1}{19}x + \frac{46}{19})) = \frac{19}{544}((x^2 - 3x + 2) - (\frac{1}{19}x + \frac{46}{19})((x^3 + 3x^2 + 3x + 1) - (x^2 - 3x + 2)(x + 6)))$. Break down the parenthesis, we see that we can find polynomials $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$. ☺

Theorem 9.1.40. *If two complex polynomial $p(x), q(x)$ has no common root, then we can find polynomials $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$.*

Proof. Outside the scope of this class. Search for Euclidean algorithm online. □

Corollary 9.1.41. *If two complex polynomial $p(x), q(x)$ has no common root, then for any square matrix A, $\mathrm{Ker}(p(A)q(A)) = \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$.*

Proof. Since $p(x), q(x)$ has no common root, we can find polynomials $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$. Then $a(A)p(A) + b(A)q(A) = I$.

Suppose $\boldsymbol{v} \in \mathrm{Ker}(p(A)) \cap \mathrm{Ker}(q(A))$. Then $p(A)\boldsymbol{v} = \boldsymbol{0}$ and $q(A)\boldsymbol{v} = \boldsymbol{0}$. Then $\boldsymbol{v} = I\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} = \boldsymbol{0}$. So we have trivial intersection.

Next, if $\boldsymbol{v} \in \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$, then $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ where $p(A)\boldsymbol{v}_1 = \boldsymbol{0}$ and $q(A)\boldsymbol{v}_2 = \boldsymbol{0}$. Then $p(A)q(A)\boldsymbol{v} = q(A)p(A)\boldsymbol{v}_1 + p(A)q(A)\boldsymbol{v}_2 = \boldsymbol{0}+\boldsymbol{0} = \boldsymbol{0}$. So we see that $\mathrm{Ker}(p(A))\oplus\mathrm{Ker}(q(A)) \subseteq \mathrm{Ker}(p(A)q(A))$.

Conversely, suppose $\boldsymbol{v} \in \mathrm{Ker}(p(A)q(A))$. Then $a(A)p(A)\boldsymbol{v} \subseteq \mathrm{Ker}(q(A))$ and $b(A)q(A)\boldsymbol{v} \subseteq \mathrm{Ker}(p(A))$. Then we have $\boldsymbol{v} = I\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} \in \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$. So we have $\mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A)) \supseteq \mathrm{Ker}(p(A)q(A))$. □

Corollary 9.1.42. *Suppose $p(x) - \lambda$ has distinct roots. Pick any square matrix A. For the eigenvalue λ of $p(A)$, let $\lambda_1, \ldots, \lambda_k$ be all eigenvalues of A such that $p(\lambda_i) = \lambda$. Then $p(A)\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $\boldsymbol{v}$ is a linear combination of eigenvectors of A for the eigenvalues $\lambda_1, \ldots, \lambda_k$.*

Proof. We can WLOG say $\lambda = 0$. Then $p(A)\boldsymbol{v} = \boldsymbol{0}$ implies that $\boldsymbol{v} \in \mathrm{Ker}(p(A))$. On the other hand, by the corollary above, since $p(x) = \prod(x - x_i)$ for distinct roots x_i, $\mathrm{Ker}(p(A)) = \bigoplus \mathrm{Ker}(A - x_iI)$.

So if $\boldsymbol{v} \in \mathrm{Ker}(p(A))$, then $\boldsymbol{v}$ is a linear combination of vectors $\boldsymbol{v}_i \in \mathrm{Ker}(A - x_iI)$. Note that $\mathrm{Ker}(A - x_iI)$ unless x_i is BOTH a root of $p(x)$ AND an eigenvalue, i.e., unless x_i is some λ_i.

So we are done. □

Corollary 9.1.43. *If $p(x)$ has distinct roots $\lambda_1, \ldots, \lambda_n$, then the solutions to the differential equation $p(\frac{d}{dx})f = 0$ are linear combinations of $\mathrm{e}^{\lambda_i x}$.*

Proof. Taking derivative $\frac{d}{dx}$ is a linear operation, and for any complex number λ, $\frac{d}{dx}$ has eigenvalue λ with eigenvectors multiples of $\mathrm{e}^{\lambda x}$. So we are done. □

Example 9.1.44. Consider an object attached to a spring, and it is bouncing around horizontally without friction. Say the elastic coefficient is 1, object mass is 1, and the location of our object at time t is $f(t)$. Then $f''(t) = -f(t)$.

So let $p(x) = x^2 + 1$, we have $p(\frac{d}{dx})f = 0$. Note that $p(x)$ has distinct roots, so the solutions are linear combinations of $\mathrm{e}^{\mathrm{i}t}$ and $\mathrm{e}^{-\mathrm{i}t}$. Taking real solutions only, then we see that the solutions are linear combinations of $\sin t$ and $\cos t$.

So our object moves periodically.

If we have elastic coefficient k, and say we have friction positively correlated to speed with coefficient μ, and object mass m. Then $mf''(t) = -kf(t) - \mu f'(t)$. So let $p(x) = mx^2 + \mu x + k$, and we have $p(\frac{d}{dx})f = 0$ again. Hopefully we have distinct roots (which we almost always have), then we are good to go again. ☺

Finally, there is actually a way to prove Jordan canonical form entirely by juggling polynomials of a matrix A, i.e., dealing with $p(A)$ for various polynomial $p(x)$. We leave that to homework.

9.1.5 Generalized Eigenspace

In the last section, we have separated zero-eigenstuff with nonzero-eigenstuff, by doing invariant decompositions. In this section, we shall separate the eigenstuff of A according to EACH eigenvalue.

Our goal here is the following. For any matrix A, we aim to block diagonalize it, such that each diagonal block is a matrix with all eigenvalues the same. For example, something like this: $\begin{bmatrix} 1 & 2 & 3 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$. Here there are two diagonal blocks, the first one has all eigenvalues 1, and the second one has all eigenvalues 2.

In essense, we are looking for an invariant decomposition $\mathbb{C}^n = V_1 \oplus \cdots \oplus V_k$ such that A retricted to each V_i will be a matrix with all eigenvalues the same λ_i.

Our previous ultimate invariant decomposition is already in this direction. Suppose $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ as the corresponding block-diagonalization of A for the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. Now, A_N is the restriction of A to a linear transformation on $N_\infty(A)$, and it will eventually kill everything in this domain, so A_N can only have zero eigenvalues.

In contrast, Since $\mathrm{Ker}(A) \subseteq N_\infty(A)$ and $N_\infty(A) \cap R_\infty(A) = \{\mathbf{0}\}$, it turns out that A restricted to a linear transformation on $R_\infty(A)$ will have zero kernel, i.e., A_R is an invertible matrix! So it has no zero eigenvalue.

In particular, the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$ has successfully isolated all the zero-eigenvalue behaviors of A in $N_\infty(A)$, and all the non-zero-eigenvalue behaviors of A to $R_\infty(A)$.

Recall that the eigenspace of a matrix A for the eigenvalue λ is simply $\mathrm{Ker}(A - \lambda I)$. We now define the following.

Definition 9.1.45. *The generalized eigenspace of a matrix A for the eigenvalue λ is the subspace $N_\infty(A-\lambda I)$.*

You can easily see the intuition behind this. $N_\infty(A - \lambda I)$ is taking away all zero eigenstuff of $A - \lambda I$, which is exactly all the λ-eigenstuff of A.

Let us show that these subspaces are linearly independent.

Lemma 9.1.46. *If $\lambda \neq \mu$, then $N_\infty(A - \lambda I) \subseteq R_\infty(A - \mu I)$.*

Proof. Note that $A - \lambda I$ restricted to $N_\infty(A - \lambda I)$ has all eigenvalue zero. Therefore A restricted to $N_\infty(A - \lambda I)$ has all eigenvalue λ, and $A - \mu I$ restricted to $N_\infty(A - \lambda I)$ has all eigenvalue $\lambda - \mu \neq 0$. So $A - \mu I$ is invertible on the subspace $N_\infty(A - \lambda I)$. Hence $N_\infty(A - \lambda I) \subseteq R_\infty(A - \mu I)$. □

Note that this immediately implies independence.

Corollary 9.1.47. *Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of A (NOT counting algebraic multiplicity, i.e., they are distinct complex numbers). Let $V_i = N_\infty(A - \lambda_i I)$ be the generalized eigenspace for each i. Then $V_1, \ldots, V_k$ are linearly independent subspaces ,and they are invariant under A.*

Proof. For each k, we need to show that $N_\infty(A-\lambda_k I)$ and $\sum_{i=1}^{k-1} N_\infty(A-\lambda_i I)$ have zero intersection. But we have $N_\infty(A-\lambda_i I) \subseteq R_\infty(A-\lambda_k I)$ for all $i < k$. Therefore, the fact that $N_\infty(A-\lambda_i I) \cap R_\infty(A-\lambda_i I) = \{\mathbf{0}\}$ implies that $N_\infty(A - \lambda_k I)$ and $\sum_{i=1}^{k-1} N_\infty(A - \lambda_i I)$ have zero intersection. □

They are not only independent. They in fact gives us the desired invariant decomposition of the whole domain.

Proposition 9.1.48 (Geometric meaning of algebraic multiplicity)**.** *Let λ be an eigenvalue of a square matrix A with algebraic multiplicity m, and let $V_\lambda = N_\infty(A - \lambda I)$ be the generalized eigenspace. Then* $\dim V_\lambda = m$.

Proof. Replacing A by $A - \lambda I$ if necessary, we can assume that $\lambda = 0$.

Now let $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ be the corresponding block diagonalization of A after a change of basis according to the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. As we have discussed before, A_N will only have eigenvalue zero, while A_R has no zero eigenvalue. But their characteristic polynomials must satisfy $p_A(x) = p_{A_N}(x)p_{A_R}(x)$. So the algebraic multiplicity of 0 in p_A is exactly the same as the degree of p_{A_N}, whcih is $\dim N_\infty(A)$. □

Theorem 9.1.49. *Let $\lambda_1, \dots, \lambda_k$ be the eigenvalues of A (NOT counting algebraic multiplicity, i.e., they are distinct complex numbers). Let $V_i = N_\infty(A - \lambda_i I)$ be the generalized eigenspace for each i. Then we have an invariant decomposition $\mathbb{C}^n = \bigoplus_{i=1}^k V_i$.*

Proof. These subspaces are linearly independent, and their dimensions add up to n (since algebraic multiplicities add up to n). □

Recall that previously, we see that all eigenvalues of A_N must be zero in the block diagonalization $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ corresponding to the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. Similarly, given a block diagonalization of A, say $\begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}$ according to the generalized eigenspaces, then each A_i is the restriction of A to V_i, so all eigenvalues of A_i must be λ_i.

9.2 Nilpotent Matrices

9.2.1 Invariant Filtration and Triangularization

We have now block diagonalized our matrix, where each block is a matrix whose eigenvalues are all the same. What now? Well, we need to understand such matrices whose eigenvalues are all the same! Let us start with a special case. What if all eigenvalues are zero?

Definition 9.2.1. *We say a matrix A is nilpotent if $A^k = O$ for some positive integer k. (I.e., $N_\infty(A)$ is the whole domain.)*

(Tiny remark: "nil" means zero. "potent" means power. "Some power is zero", i.e., nilpotent.)

Remark 9.2.2. *If $A^k = O$ for some positive integer k, then we can in fact require that $k \le n$. This is because of our previous analysis of $N_\infty(A)$. In particular, we always have $A^n = O$.*

Proposition 9.2.3. *A is nilpotent if and only if all eigenvalues of A are zero.*

Proof. Suppose A is nilpotent. Then N_∞ is the whole domain, so all eigenvalues are zero.

If all eigenvalues are zero, then the whole domain is in N_∞. So A is nilpotent. □

Now, these nilpotent matrices are annoying. Many of them has NO good invariant decomposition at all! Instead, they behave like onions: layers of invariant subspaces, each containing the next.

Example 9.2.4. Consider $A = \begin{bmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{bmatrix}$. This is the "shift up" operator that sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} y \\ z \\ 0 \end{bmatrix}$, i.e., it is shifting the coordinates upwards. Therefore, we obviously have $A^3 = O$. It is nilpotent.

Now what are its invariant subspaces? If A is invariant, and $A(V) \subseteq V$ for some subspace V, then A restricted to this linear tranformation on V would be nilpotent as well. Now if $\dim V = k$, any nilpotent linear transformation must die in k steps. So we must have $A^k(V) = \{\mathbf{0}\}$.

(Alternatively, since $A^n = O$, consider the sequence of subspaces $V, A(V), \dots, A^n(V)$, then this sequence must eventually shrink to zero. Now if $A^i(V) = A^{i+1}(V)$, then $A^{i+2}(V) = A(A^{i+1}(V)) = A(A^i(V)) = A^{i+1}(V) = A^i(V)$, and the sequence would stabilize forever. So this sequence must shrink strictly until it hit zero. Each step the dimension must reduce by at least one. So if $\dim V = k$, we must have $A^k(V) = \{\mathbf{0}\}$.)

So $V \subseteq \operatorname{Ker}(A^k)$. However, in our case, note that for any k, $\operatorname{Ker}(A^k)$ is spanned by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_k$. So $\dim \operatorname{Ker}(A^k) = k = \dim V$, wow! So $V = \operatorname{Ker}(A^k)$.

In particular, all invariant subspaces of A are $\operatorname{Ker}(A^k)$ for some k. The invariant subspaces are exactly $\{0\}$, x-axis, xy-plane, and the whole space.

There is no invariant decomposition of the whole domain other than the trivial one. However, you can see that these invariant subspaces come in layers, like an onion, each layer containing the last. Why are Jordan blocks like $\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$? As we shall see later, it is precisely due to this onion structure. ☺

Definition 9.2.5. *Given a vector space V, a filtration for V is a sequence of subspaces $V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n = V$, where $\dim V_k = k$. For any linear tranformation $A : V \to V$, we say this is an (A-)invariant filtration if all V_k are A-invariant subspaces.*

So the idea is this: invariant decomposition leads to block diagonalization. Invariant filtration would lead to triangularization.

Proposition 9.2.6. *If $L : V \to V$ is a linear transformation, and V has an invariant filtration $V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n = V$. Pick any $\boldsymbol{v}_i \in V_i - V_{i-1}$ for each $1 \leq i \leq n$, then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ form a basis of V, under which the matrix for L is upper triangular.*

Proof. Let us first show that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ form a basis. It is enough to show linear independence.

We perform induction. Since $\boldsymbol{v}_1 \in V_1 - V_0$, it is non-zero, so it is linearly independent. For each $i \geq 1$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1} \in V_{i-1}$, yet $\boldsymbol{v}_i \notin V_{i-1}$. By induction hypothesis, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}$ are already linearly independent, so $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i$ are linearly independent as well. We are done.

In fact, it is not hard to see that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i$ form a basis for V_i for each i.

Now $\boldsymbol{v}_i \in V_i$, so by invariance, $L\boldsymbol{v}_i \in V_i$ as well. Say $L\boldsymbol{v}_i = a_{1i}\boldsymbol{v}_1 + \cdots + a_{ii}\boldsymbol{v}_i$ since $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i$ form a basis for V_i.

Now by straight forward calculation, we have:

$$L(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) = (a_{11}\boldsymbol{v}_1, a_{12}\boldsymbol{v}_1 + a_{22}\boldsymbol{v}_2, \ldots, a_{1n}\boldsymbol{v}_1 + \cdots + a_{nn}\boldsymbol{v}_n) = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ & \ddots & \vdots \\ & & a_{nn} \end{bmatrix}.$$

This means that using $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ as basis, the matrix for L is simply the upper triangular matrix above. □

Note that to go from a filtration to a triangularization, we just need to pick a vector from each "gap" between subspaces in the chain, and use them as a basis.

Example 9.2.7. Suppose $A : \mathbb{R}^3 \to \mathbb{R}^3$, and the xy-plane is invariant. What does this mean?

This means that $A\boldsymbol{e}_1, A\boldsymbol{e}_2$ will still have a zero third coordinate. In particular, $A = \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & 0 & * \end{bmatrix}$. As you can see, any k-dimensional invariant subspace corresponds to some upper triangular block structure, decomposing into a k-block and a $(n-k)$-block on the diagonal.

If we have an invariant filtration, then we have a block matrix which is block diagonal by a k-block and a $(n-k)$-block, for all k. This forces the matrix to be (non-block) upper triangular. ☺

The converse is also true. If A is upper triangular, then you can easily check that $\operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k)$ is invariant under A for all k. So we see that a matrix can be triangularized if and only if there is an invariant filtration.

Lemma 9.2.8. *For any linear transformation $L : V \to V$ on a finite dimensional complex vector space V, there is an invariant filtration. (Note that this statement NEEDS V to be a complex vector space.)*

Proof. Due to Schur decomposition (from last semester), triangularization is always possible. hence filtration can be found. In particular, suppose under a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, the map L is upper triangularized, then the desired filtration is simply $\operatorname{span}(\boldsymbol{v}_1) \subseteq \operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2) \subseteq \ldots$. □

Note that, given any invariant filtration for A, simply let $\boldsymbol{v}_i$ be a unit vector orthogonal to V_{i-1} inside of V_i (like finding a normal vector to a plane in the space). Then we shall find a unitary matrix $B = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ such that $A = BTB^{-1}$ where T is upper triangular. This is the Schur decomposition theorem we did last semester. If you look into our proof last semester, you shall see that it is essentially IDENTICAl to what we are doing here.

9.2.2 Nilpotent Canonical Form

Definition 9.2.9. *A matrix J is an $d \times d$ Jordan block for the eigenvalue λ if* $J = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}_{d \times d}$.

(In the case where $\lambda = 0$, we also say it is a nilpotent Jordan block.)

Let us show that all nilpotent matrices can be block diagonalized where the diagonal blocks are nilpotent Jordan blocks.

Theorem 9.2.10. *If A is nilpotent, then we can find B such that $A = BDB^{-1}$ where D is block diagonal, and each diagonal block is a nilpotent Jordan block.*

Note that the nilpotent Jordan blocks are all "shift-up" operators, e.g., $\begin{bmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{bmatrix}$ would sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} y \\ z \\ 0 \end{bmatrix}$, it shifts the coordinates up. If we keep sending all the coordinates upwards, then eventually nothing will survive. In particular, if J is an $n \times n$ Jordan block, then it has a kill chain $\boldsymbol{e}_n \overset{J}{\mapsto} \ldots \overset{J}{\mapsto} \boldsymbol{e}_1 \overset{J}{\mapsto} \boldsymbol{0}$ where the non-zero vectors for a basis.

In particular, our theorem says that any nilpotent matrix can be block diagonalized into such "shift-up" operators. Each Jordan block here has a corresponding "kill chain basis", and our matrix will have several kill chains whose non-zero vectors form a basis.

The next example here will show the algorithm to do the theorem above.

Example 9.2.11. Suppose A is a 7×7 nilpotent matrix. The chain of subspaces $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^2) \subseteq \mathrm{Ker}(A^3) \subseteq \mathrm{Ker}(A^4)$ has a chain of dimensions $3 \leq 5 \leq 6 \leq 7$. Note that this is NOT a filtration by itself, because some adjacent subspaces might differ by more than one dimensions.

Now we fill up the following chart from the bottum upwards:

$$\left(\begin{array}{c|ccc} \mathrm{Ker}(A) - \{\boldsymbol{0}\} & A^3v_1 & Av_2 & v_3 \\ \mathrm{Ker}(A^2) - \mathrm{Ker}(A) & A^2v_1 & v_2 & \\ \mathrm{Ker}(A^3) - \mathrm{Ker}(A^2) & Av_1 & & \\ \mathrm{Ker}(A^4) - \mathrm{Ker}(A^3) & v_1 & & \end{array}\right).$$

How did this work? We start by looking at the gap between $\mathrm{Ker}(A^4)$ and $\mathrm{Ker}(A^3)$. Note that the two subspace differ by exactly one dimension, so one extra vector is enough to extend $\mathrm{Ker}(A^3)$ to $\mathrm{Ker}(A^4)$. So we simply pick any $\boldsymbol{v}_1 \in \mathrm{Ker}(A^4) - \mathrm{Ker}(A^3)$.

Note that if $\boldsymbol{v}_1 \in \mathrm{Ker}(A^4) - \mathrm{Ker}(A^3)$, then we automatically have $A\boldsymbol{v}_1 \in \mathrm{Ker}(A^3) - \mathrm{Ker}(A^2)$, $A^2\boldsymbol{v}_1 \in \mathrm{Ker}(A^2) - \mathrm{Ker}(A)$ and $A^3\boldsymbol{v}_1 \in \mathrm{Ker}(A) - \{\boldsymbol{0}\}$. So we automatically filled a vector into each gap. We have $\mathrm{Ker}(A^4)$ spanned by $\mathrm{Ker}(A^3)$ and $\boldsymbol{v}_1$.

Now consider the gap between $\mathrm{Ker}(A^3)$ and $\mathrm{Ker}(A^2)$. Note that the two subspace differ by exactly one dimension, and we already have $A\boldsymbol{v}_1$ to fill in this gap, so there is nothing to do. We have $\mathrm{Ker}(A^3)$ spanned by $\mathrm{Ker}(A^2)$ and $A\boldsymbol{v}_1$.

Now consider the gap between $\mathrm{Ker}(A^2)$ and $\mathrm{Ker}(A)$. Note that the two subspace differ by two dimensions. We already have $A^2\boldsymbol{v}_1$ in this gap, but we need another vector. Pick any $\boldsymbol{v}_1 \in \mathrm{Ker}(A^2) - (\mathrm{Ker}(A) + \mathrm{span}(A^2\boldsymbol{v}_1))$. Now we have $\mathrm{Ker}(A^2)$ spanned by $\mathrm{Ker}(A)$ and $A^2\boldsymbol{v}_1, \boldsymbol{v}_2$.

Finally consider the gap between $\mathrm{Ker}(A)$ and $\{\boldsymbol{0}\}$. Note that the two subspaces differ by three dimensions. This time, we have $A^3\boldsymbol{v}_1, A\boldsymbol{v}_2$ in this gap already. I claim that they are linearly independent (proven in a later lemma), hence we just need one more. Pick any $\boldsymbol{v}_3 \in \mathrm{Ker}(A) - \mathrm{span}(A^3\boldsymbol{v}_1, A\boldsymbol{v}_2)$. Then we have $\mathrm{Ker}(A)$ spanned by $A^3\boldsymbol{v}_1, A\boldsymbol{v}_2, \boldsymbol{v}_3$.

Now, we see that the following subspaces are spanned by the following vectors:

$$\left(\begin{array}{c|ccccccc} \mathrm{Ker}(A) & A^3v_1 & Av_2 & v_3 & & & & \\ \mathrm{Ker}(A^2) & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 & & \\ \mathrm{Ker}(A^3) & Av_1 & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 & \\ \mathrm{Ker}(A^4) & v_1 & Av_1 & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 \end{array}\right).$$

And furthermore, we have kill chains $\boldsymbol{v}_1 \overset{A}{\mapsto} A\boldsymbol{v}_1 \overset{A}{\mapsto} A^2\boldsymbol{v}_1 \overset{A}{\mapsto} A^3\boldsymbol{v}_1 \overset{A}{\mapsto} \boldsymbol{0}$, and $\boldsymbol{v}_2 \overset{A}{\mapsto} A\boldsymbol{v}_2 \overset{A}{\mapsto} \boldsymbol{0}$, and finally $\boldsymbol{v}_3 \overset{A}{\mapsto} \boldsymbol{0}$. All the vectors in these three kill chains (other than the zero vectors) are linearly independent, and all the important invariant subspaces are spanned by these vectors in very nice manners.

Pick a basis $A^3\boldsymbol{v}_1, A^2\boldsymbol{v}_1, A\boldsymbol{v}_1, \boldsymbol{v}_1, A\boldsymbol{v}_2, \boldsymbol{v}_2, \boldsymbol{v}_3$, then you can check yourself that our matrix A would change into the following:

$$\left[\begin{array}{ccccc|cc|c} 0 & 1 & & & & & & \\ & 0 & 1 & & & & & \\ & & 0 & 1 & & & & \\ & & & 0 & & & & \\ \hline & & & & & 0 & 1 & \\ & & & & & & 0 & \\ \hline & & & & & & & 0 \end{array}\right].$$

☺

Two things could go wrong here. First of all, when we fill in the gap between $\mathrm{Ker}(A)$ and $\{\boldsymbol{0}\}$, we need $A^3\boldsymbol{v}_1$ and $A\boldsymbol{v}_2$ to be linearly independent. Why is that?

Recall that we picked $\boldsymbol{v}_2$ such that $\mathrm{Ker}(A), A^2\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent. It turns out that this is enough.

Lemma 9.2.12. *If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, \mathrm{Ker}(A^t)$ are linearly independent, then $A\boldsymbol{v}_1, \dots, A\boldsymbol{v}_k, \mathrm{Ker}(A^{t-1})$ are linearly independent.*

Proof. Suppose $(\sum a_i A\boldsymbol{v}_i) + \boldsymbol{w} = \boldsymbol{0}$ where $\boldsymbol{w} \in \mathrm{Ker}(A^{t-1})$. Apply A^{t-1} on both sides. Then we have $(\sum a_i A^t\boldsymbol{v}_i) + A^{t-1}\boldsymbol{w} = \boldsymbol{0}$, and here $A^{t-1}\boldsymbol{w}$ would die.

So we have $A^t(\sum a_i\boldsymbol{v}_i) = \boldsymbol{0}$. This implies that $\sum a_i\boldsymbol{v}_i = \boldsymbol{w}'$ for some $\boldsymbol{w}' \in \mathrm{Ker}(A^t)$. But since these $\boldsymbol{v}_i$ and $\mathrm{Ker}(A^t)$ are linearly independent, this means all $a_i = 0$ and $\boldsymbol{w}' = \boldsymbol{0}$.

This in turn means that, from the equation $(\sum a_i A\boldsymbol{v}_i) + \boldsymbol{w} = \boldsymbol{0}$, we must have $\boldsymbol{w} = \boldsymbol{0}$ as well. So we have proven independence. □

This lemma guarantees that our algorithm in the example shall always work, and hence our theorem is correct.

9.3 Jordan Canonical Form

The Jordan canonical form simply combines all previous results. There is one last simple lemma.

Lemma 9.3.1. *If all eigenvalues of A are λ, then $A = BJB^{-1}$ where J is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ.*

Proof. All eigenvalues of $A-\lambda I$ are zero, so this is nilpotent. So $A-\lambda I = BJB^{-1}$ where J is block diagonal, and each diagonal block is a nilpotent Jordan block. Then $A = BJB^{-1} + \lambda I = B(J+\lambda I)B^{-1}$. And we can see that $J + \lambda I$ is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ. □

Theorem 9.3.2 (Jordan canonical form)**.** *For any matrix A, we have $A = BJB^{-1}$ where J is block diagonal, and each diagonal block of J is a Jordan block.*

Proof. Since the domain is the direct sum of generalized eigenspaces, we can assume that $A = XDX^{-1}$ where $D = \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix}$ is block diagonal, and each diagonal block D_i corresponds to a generalized eigenspace for the eigenvalue λ_i.

So all eigenvalues of D_i are λ_i. So $D_i = B_i J_i B_i^{-1}$ where J_i is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ_i.

Then $A = BJB^{-1}$ where $B = X\begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_k \end{bmatrix}$ and $J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{bmatrix}$ is block diagonal, and each diagonal block of J is a Jordan block. □

How to find Jordan canonical form? Let us have some calculation examples.

Lemma 9.3.3. *If λ is an eigenvalue of A with algebraic multiplicity m, then $N_\infty(A-\lambda I) = \operatorname{Ker}(A-\lambda I)^m$.*

Proof. Take $N_\infty(A-\lambda I)$ as the domain, and consider the operator $A-\lambda I$, which is nilpotent. So if $\dim N_\infty(A-\lambda I) = m$, then $(A-\lambda I)^m = 0$ on the space $N_\infty(A-\lambda I)$.

So now using our original domain, we see that $N_\infty(A-\lambda I) \subseteq \operatorname{Ker}(A-\lambda I)^m$. But by definition $\operatorname{Ker}(A-\lambda I)^m \subseteq N_\infty(A-\lambda I)$. So we are done. □

Example 9.3.4. Consider $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$. Then $\det(xI-A) = \det\begin{bmatrix} x-2 & 0 & 0 \\ 1 & x-1 & -2 \\ -3 & 0 & x-1 \end{bmatrix} = (x-2)\det\begin{bmatrix} x-1 & -2 \\ 0 & x-1 \end{bmatrix} = (x-2)(x-1)^2$. So it has eigenvalue 1 with algebraic multiplicity 2 and eigenvalue 2 with algebraic multiplicity 1. So it must has a generalized eigenspace V_1 for the eigenvalue 1 of dimension 2 and a generalized eigenspace V_2 for the eigenvalue 2 of dimension 1.

What is V_1? It is $\operatorname{Ker}(A-I)^2 = \operatorname{Ker}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 2 \\ 3 & 0 & 0 \end{bmatrix}^2 = \operatorname{Ker}\begin{bmatrix} 1 & 0 & 0 \\ 5 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}$, which is spanned by $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. And restricted to this subspace V_1, under this basis, we have $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$, and $A - I = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$ in indeed nilpotent. Now our theorem on nilpotent Jordan normal form tells us that we could pick basis $\boldsymbol{v}_1 = (A-I)\boldsymbol{v}_2$ and $\boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ as the right basis for V_1.

What is V_2? It is $\operatorname{Ker}(A-2I) = \operatorname{Ker}\begin{bmatrix} 0 & 0 & 0 \\ -1 & -1 & 2 \\ 3 & 0 & -1 \end{bmatrix}$ which is spanned by $\boldsymbol{v}_3 = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$. Obviously A restricted to V_2 is just $\begin{bmatrix} 2 \end{bmatrix}$ and there is nothing to do here.

So the best basis for V should be $(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & 0 & 5 \\ 0 & 1 & 3 \end{bmatrix}$. And under this basis, the new matrix for A

should be $\begin{bmatrix}1&1&0\\0&1&0\\0&0&2\end{bmatrix}$. Let us check this. Indeed, we have:

$$\begin{bmatrix}0&0&1\\2&0&5\\0&1&3\end{bmatrix}\begin{bmatrix}1&1&0\\0&1&0\\0&0&2\end{bmatrix}\begin{bmatrix}0&0&1\\2&0&5\\0&1&3\end{bmatrix}^{-1}=\begin{bmatrix}0&0&2\\2&2&10\\0&1&6\end{bmatrix}\begin{bmatrix}-5/2&1/2&0\\-3&0&1\\1&0&0\end{bmatrix}=\begin{bmatrix}2&0&0\\-1&1&2\\3&0&1\end{bmatrix}=A.$$

☺

Example 9.3.5. Let us have a more complicated example. Feel free to have a matrix calculator in hand while reading this example.

Say $A=\begin{bmatrix}-10&9&-7&-1&7\\-17&13&-9&-2&12\\-14&9&-6&-1&10\\-13&9&-7&0&9\\-12&9&-7&-1&9\end{bmatrix}$. You can find its characteristic polynomial and check that its eigenvalues are 1,1,1,1,2.

The eigenvalue 2 is simple. It has algebraic and geometric multiplicity 1, and you can find its corresponding eigenvector is $\boldsymbol{v}_5=\begin{bmatrix}5\\9\\8\\7\\6\end{bmatrix}$.

For the eigenvalue 1, consider $A-I=\begin{bmatrix}-11&9&-7&-1&7\\-17&12&-9&-2&12\\-14&9&-7&-1&10\\-13&9&-7&-1&9\\-12&9&-7&-1&8\end{bmatrix}$. You can check that $\dim\operatorname{Ker}(A-I)=2,\dim\operatorname{Ker}(A-I)^2=3,\dim\operatorname{Ker}(A-I)^3=4$, and we don't need to continue once we reach dimension 4, because 1 only has algebraic multiplicity 4.

Pick any $\boldsymbol{v}_3\in\operatorname{Ker}(A-I)^3-\operatorname{Ker}(A-I)^2$, and set $\boldsymbol{v}_2=(A-I)\boldsymbol{v}_3$ and $\boldsymbol{v}_1=(A-I)\boldsymbol{v}_2$, and find any $\boldsymbol{v}_4$ such that $\boldsymbol{v}_1,\boldsymbol{v}_4$ span $\operatorname{Ker}(A-I)$. One possible choice is $\boldsymbol{v}_3=\begin{bmatrix}1\\2\\2\\1\\1\end{bmatrix}$, then $\boldsymbol{v}_2=\begin{bmatrix}-1\\-1\\-1\\-1\\-1\end{bmatrix}$, and then $\boldsymbol{v}_1=\begin{bmatrix}3\\4\\3\\3\\3\end{bmatrix}$.

Then you can pick say $\boldsymbol{v}_4=\begin{bmatrix}1\\3\\3\\2\\1\end{bmatrix}$. Note that since $\boldsymbol{v}_3$ goes to $\boldsymbol{v}_2$, which goes to $\boldsymbol{v}_1$, and $\boldsymbol{v}_4$ stands alone, therefore the corresponding nilpotent Jordan block is $\left[\begin{array}{ccc|c}0&1&0&0\\0&0&1&0\\0&0&0&0\\\hline0&0&0&0\end{array}\right]$.

So under the basis $B=(\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3,\boldsymbol{v}_4,\boldsymbol{v}_5)=\begin{bmatrix}3&-1&1&1&5\\4&-1&2&3&9\\3&-1&2&3&8\\3&-1&1&2&7\\3&-1&1&1&6\end{bmatrix}$, we have A in Jordan canonical form

$J = \left[\begin{array}{ccc|c|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 2 \end{array}\right]$. In particular, $A = BJB^{-1}$.

(Obviously I designed it so that we could have integer solutions.... Usually we should not be so lucky. A super interesting challenge question here: Can you design another 5×5 integer-entry matrix A such that for $A = BJB^{-1}$, both B and J has integer entries?) ☺

Sometimes we can tell the Jordan normal form right away, just from the geometric and algebraic multiplicity. Here is why.

Proposition 9.3.6. *Suppose λ is an eigenvalue of A with algebraic mulitiplicity m_a and geometric multiplicity m_g. Then m_g is the number of λ-Jordan blocks in the Jordan canonical form of A, while m_a is the sum of the sizes of all these Jordan blocks.*

Proof. WLOG we can assume that A is already in Jordan canonical form. Furthermore, all the blocks not related to λ are irrelavant. It is then clear that each λ-Jordan block contributes to exactly one dimension to $\mathrm{Ker}(A - \lambda I)$, so the statement about geometric multiplicity is done.

The statement about algebraic multiplicity is trivial by just looking at the characteristic polynomial of block diagonal matrices. □

In particular, if a matrix A has all geometric multiplicities equal to algebraic multiplicities, then the number of λ-blocks would equal to the sum of sizes of all these blocks, i.e., each block is 1×1. So the matrix is diagonalizable.

Example 9.3.7. For example, consider $J = \begin{bmatrix} 1 & 1 & & & \\ 0 & 1 & & & \\ & & 1 & & \\ & & & 2 \end{bmatrix}$. Chech that the eigenvalue 1 here has indeed geometric multiplicity 2 and algebraic multiplicity 3.

Conversely, suppose A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 3, m_g(2) = m_a(2) = 1$, then it has two 1-blocks and a single 2-block. Furthermore, since the two 1-blocks have a total size 3, it must be $1 + 2$. So A must have the Jordan canonical form J above.

Of course, if A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 4, m_g(2) = m_a(2) = 1$, then there is no way to tell now. The two 1-blocks could be $2 + 2$ or $1 + 3$, and we may never know. ☺

The last example where we cannot decide is unfortunate. However, there is one more tool we can use: the minimal polynomial. But before we do that, let us do the famous Cayley-Hamilton Theorem as a corollary to our study of generalized eigenspaces.

Corollary 9.3.8 (Cayley-Hamilton Theorem)**.** *For any matrix A, let $p_A(x)$ be its characteristic polynomial. Then $p_A(A)$ is the zero matrix.*

Proof. It is enough to show that $p_A(A)$ kills each generalized eigenspace.

For each eigenvalue λ, let m be its algebraid multiplicity. Then $p_A(x) = q(x)(x - \lambda)^m$. So $p_A(A) = q(A)(A - \lambda I)^m$.

So if $\boldsymbol{v} \in N_\infty(A - \lambda I) = \mathrm{Ker}(A - \lambda I)^m$, then $p_A(A)\boldsymbol{v} = q(A)(A - \lambda I)^m \boldsymbol{v} = \boldsymbol{0}$.

But this is true for all λ. So $p_A(A)$ kills all vectors in all generalized eigenspaces. Oops. □

Definition 9.3.9. *We say a polynomial $p(x)$ is a killing polynomial for A if $p(A) = 0$. We say $p(x)$ is a minimal polynomial for A if any killing polynomial of A must contain $p(x)$ as a factor.*

Proposition 9.3.10. *Any square matrix A has a minimal polynomial.*

Proof. Suppose that A is in Jordan normal form. Then $p(A)$ is simply applying $p(x)$ to each diagonal block, and $p(x)$ is a killing polynomial if and only if it kills all blocks simultaneously. So it is enough to prove this statement for each Jordan block.

Suppose A be a single Jordan block, say $n \times n$ with eigenvalue λ. Then $A - \lambda I$ is the shift up operator, and if $p(x) = a_0 + a_1 x + \cdots + a_k x^k$, then $p(A - \lambda I)$ will have diagonal entries a_0, and entries right above the diagonal a_1, and so on so forth. So $p(A - \lambda I) = 0$ if and only if the coefficients $a_0, \ldots, a_{n-1}$ are zero, i.e., $p(x)$ contains x^n as a factor. So if $q(A) = 0$, then $q(A)$ must contains $(A - \lambda I)^n$ as a factor.

To sum up, to kill a Jordan block, say $n \times n$ with eigenvalue λ, $p(x)$ must contain factor $(x - \lambda)^n$.

So $p(x)$ kills A if and only if it contains $(x - \lambda)^{m_\lambda}$ for all λ, where λ is the size of largest λ-Jordan block for A. □

Example 9.3.11. If A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 4, m_g(2) = m_a(2) = 1$, then there is no way to tell now. The two 1-blocks could be $2 + 2$ or $1 + 3$, and we may never know.

But if we also know that the minimal polynomial is $(x-1)^2(x-2)$, then the 1-blocks must be $2+2$, and we must have two 1-blocks of size 2, and a single 2-block of size 1. If the minimal polynomial is $(x-1)^3(x-2)$, then the 1-blocks must be $1 + 3$, and we must have a 1-blocks of size 3, a 1-blocks of size 1, and a single 2-block of size 1.

Of course, there will be situations where even the minimal polynomial is not enough. Suppose A is 7×7 with $m_a(1) = 7, m_g(1) = 3$, and minimal polynomial $(x - 1)^3$. Then it could be $3 + 3 + 1$ or $3 + 2 + 2$, and we cannot tell anymore. Time to get your hand dirty and actually compute those blasted $\text{Ker}(A - I)^k$. ☺

9.4 (Optional) The geometric interpretation of Jordan canonical form and generalized eigenspaces

Technically we are done. The theorem of Jordan canonical form is saying that, for any linear map, we can decompose it into independent "submaps" that are Jordan blocks. So if we understand all Jordan blocks we would understand every single matrix.

So this raises a new question. How would a Jordan block behave? Let us look at a few to generate some ideas.

Example 9.4.1. What are nilpotent Jordan blocks? Consider the 3×3 nilpotent Jordan block N. It sends the z-axis to the y-axis, and the y-axis to the x-axis. Huh, it seems to be rotating. But then it sends the x-axis to zero. So we are "rotating inwards to zero". (Nei Juan....)

Personally I think of $\mathbb{R}^3$ as the space of all students, and N as some competitive and selective process. Then after N, all students are squeezed into the xy-plane, trying to excel. After another N, now everyone is squeezed into the x-axis, trying to be the best of the best. After yet another N, everyone dies of exhaustion apparently.... ☺

Example 9.4.2. $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is the standard shearing. In general, consider $E = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$. It sends rectangles, with sides parallel to the coordinate-lines, into parallelograms of the same height. Draw a few graphic examples and shapes to see this better. This process would preserve the base and height of the parallelogram, so it preserves the area.

(Also note that EA is a row operation on A. Such row operations corresponds to shearings, so it preserves area, and hence it preserves the determinant. I.e., $\det(EA) = \det(A)$.)

If you repeatedly apply $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ to a vector, say $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, you get $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}$, and so on. Basically the second coordinates are always the same, while the first coordinate keep progressing. The so the orbits of A are lines parallel to the x-axis. ☺

Example 9.4.3. Now consider $J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. It sends $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, then to $\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$, then to $\begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix}$, then to

$\begin{bmatrix}6\\4\\1\end{bmatrix}$, and so on. This is EXACTLY the left three entries of the Pascal's triangule (Yang Hui triangle, or binomial coefficients, etc.)!

So to see $J^k \begin{bmatrix}0\\0\\1\end{bmatrix}$, you can imagine that you are doing $(x+1)^k$, and read out the last three coefficients. You can also see that $J^k \begin{bmatrix}0\\1\\0\end{bmatrix} = \begin{bmatrix}k\\1\\0\end{bmatrix}^T$, which is basically the last three coefficients of $x(x+1)^k$. In general, $J^k \begin{bmatrix}a\\b\\c\end{bmatrix}$ is the last three coefficients of $(ax^2+bx+c)(x+1)^k$. Funny, no?

Is this really true? Well, let P_2 be set of polynomials mod x^3. I.e., we consider two polynomials to be the same as long as they have the same coefficients at degree $2, 1, 0$. For example, we think of x^3+x+1 and x^4+x+1 as the same element in P_2.

Then clearly P_2 is three dimensional, hence we can idenfity it with $\mathbb{R}^3$ via its standard basis $x^2, x, 1$. Then how does J behaves on P_2? It sends 1 to $x+1$, and x to x^2+x, and x^2 to x^2, which is the same as x^3+x^2 since we only care about the coefficients at degree $2, 1, 0$. So J behaves exactly by multiplying polynomials by $(x+1)$. So $J^k(ax^2+bx+c) = (ax^2+bx+c)(x+1)^k \pmod{x}^3$.

This algebraic picture can be generalized to Jordan blocks with eigenvalue 1 of arbitrary size. ☺

Example 9.4.4. What is the geometric behavior of $J = \begin{bmatrix}1&1&0\\0&1&1\\0&0&1\end{bmatrix}$? Say what are its orbits (smooth curves C such that J always maps each point in C back to some point in C)?

Well, in general, $\begin{bmatrix}a\\b\\c\end{bmatrix}$ would goes to $\begin{bmatrix}a+b\\b+c\\c\end{bmatrix}$, and then to $\begin{bmatrix}a+b+b+c\\b+c+c\\c\end{bmatrix}$, and then to $\begin{bmatrix}a+b+(b+c)+(b+c+c)\\b+c+c+c\\c\end{bmatrix}$ and so on. So after k steps, J^k would maps it to $\begin{bmatrix}a+kb+(0+1+\cdots+(k-1))c\\b+kc\\c\end{bmatrix} = \begin{bmatrix}a+kb+\frac{1}{2}(k^2-k)c\\b+kc\\c\end{bmatrix}$.

So generically, to find orbits, I simply replace the integer k by an arbitrary real number t, and we have the orbits $p(t) = \begin{bmatrix}\frac{c}{2}t^2+(b-\frac{c}{2})t+a\\ct+b\\c\end{bmatrix}$. It is easy to verify that any points on this curve shall stay on this curve after J.

As you can see, the third coordinate never change, so the orbit curves stays on a plane (parallel to the xy-plane). On this plane, the first coordintae is in fact a degree two polynomial of the second coordinate. So on this plane, we would actually see a graph of a parabola. So orbits of J are various parabolas parallel to the xy-plane.

Note that for each parabola on a plane $z = c \neq 0$, when $t = -\frac{b}{c}$, then the parabola would go through the xz-plane. So if you want to find all parabolas on the plane $z = c$, then they are $p(t) = \begin{bmatrix}\frac{c}{2}t^2-\frac{c}{2}t+a\\ct\\c\end{bmatrix}$, or the parabola $p(t) = \begin{bmatrix}\frac{c}{2}t^2-\frac{c}{2}t\\ct\\c\end{bmatrix}$ shifted along the x-axis. Furthermore, since we only care about the curve, not how it is parametrized, we can further more substitute t by t/c. Then we have $p(t) = \begin{bmatrix}\frac{1}{2c}t^2-\frac{1}{2}t\\t\\c\end{bmatrix}$ shifted along the x-axis.

So for each constant c, the orbits on $z = c$ are just parabolas obtained by translating this along the x-axis.

I highly recommand you to draw these parabolas on $z = 1, z = 2, z = -1$ to see what would happen. Also fell free to draw the picture on the plane $z = 0$, and see why this is the limiting case for $z > 0$ and $z < 0$.

If you want to see the geometric behavior, you can try to generlize this further. Say you want a size 4 Jordan block with eigenvalue 1. Then for any orbit curve, again the last coordinate is constant for some $d \in \mathbb{C}$. If the third coordinate is t, then the second coordinate would again be $\frac{1}{2d}t^2 - \frac{1}{2}t$ shifted around by some constant. And finally, the first coordinate would be a degree 3 polynomial in t. It would look like some form of spiral. Consider curves like $\begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} \in \mathbb{R}^4$ for a idea of this kind of spirals. ☺

Example 9.4.5. Consider a Jordan block with eigenvalue, say $J = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$. Then it sends $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$, then to $\begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$ and so on. It looks like you are doing $(x+2)^k$.

Indeed, algebraically $J^k \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is the last three coordinates of $(ax^2 + bx + c)(x+2)^k$ for the same reason as before. Now you can generalize this to get the algebraic behavior of all Jordan blocks of all size for all eigenvalues.

What about its geometric behavior? Suppose we start at some vector $\begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix}$, and we construct $J \begin{bmatrix} a_{n-1} \\ b_{n-1} \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix}$. Then we see that $c_n = 2^n c_0$.

We can see that $b_n = 2b_{n-1} + c_{n-1}$. Divide this by 2^n on both sides (because we know all three sequences must be related to 2^n somehow, as 2 is the eigenvalue), we see that $\frac{b_n}{2^n} = \frac{b_{n-1}}{2^{n-1}} + \frac{c_0}{2}$. So the sequence $\frac{b_n}{2^n}$ is arithmetic and $\frac{b_n}{2^n} = \frac{b_0}{2^0} + \frac{c_0}{2}n$. So $b_n = 2^n b_0 + n2^{n-1}c_0$.

Finally, $a_n = 2a_{n-1} + b_{n-1}$. By a similar argument, $\frac{a_n}{2^n} = \frac{a_{n-1}}{2^{n-1}} + \frac{b_0}{2} + (n-1)\frac{c_0}{4}$. So $\frac{a_n}{2^n}$ is a degree two polynomial in n, and specifically you can see that $\frac{a_n}{2^n} = \frac{b_0}{2}n + \frac{c_0}{4}(0+1+2+...+(n-1)) = \frac{c_0}{8}n^2 + (\frac{b_0}{2} - \frac{c_0}{8})n$. So $a_n = n^2 2^{n-3}c_0 + n2^{n-3}(4b_0 - c_0)$.

So a typical curve looks like $p(t) = \begin{bmatrix} t^2 2^{t-3}c_0 + t2^{t-3}(4b_0 - c_0) \\ 2^t b_0 + t2^{t-1}c_0 \\ 2^t c_0 \end{bmatrix}$. By a change in parametrization, we can choose $2^t c_0$ as the new parameter t, then the curve is $p(t) = t \begin{bmatrix} a(t) \\ b(t) \\ c(t) \end{bmatrix}$ where $a(t), b(t), c(t)$ here are polynomials in $\ln t$ of degree 2,1,0.

Also note that, asymptotically for super large n, $\lim \frac{b_n^2}{2a_n c_n} = 1$. Therefore these curves has asymptotic surface $xz = y^2$. What is this surface? It is a cone around the line $\{y = 0\} \cap \{x = z\}$. So all these orbital curves will eventually get closer and closer to this cone. ☺

Example 9.4.6. As shown in the example above, the geometric picture of a Jordan block is not always easy to compute. However, let us try to do another case, $J = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$ for some extremely large λ. Then since

λ is so large, comparatively the ones are ignorable. So $J \approx \lambda I$. This geometric picture is very easy now, it is approximately just stretch everything by λ. So the orbits are approximately just rays shooting from the origin, with some minor perturbations. ☺

The process of finding Jordan canonical form is equivalent to this: First we find generalized eigenspaces of A. Next, for each generalized eigenspace for an eigenvalue λ, we identify linearly independent killing chains of $A - \lambda I$.

With this in mind, what is the generalized eigenspace, i.e., vectors eventually killed by $A - \lambda I$? Here let us formulate an alternative definition for generalized eigenspaces.

The most fundamental motivation for studying eigenstuff is to understand the behavior of sequences like $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$.

Example 9.4.7. Again consider $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. We know that its orbits are parabolas. In particular, the sequence $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$ would tend to produce longer and longer vectors, and they would never converge.

However, even though the vectors do not converge, their DIRECTIONS would in fact converge! The directions of these vectors would get closer and closer to the direction of the opening for the parabola, which is always in the direction of plus or minus x-axis.

In particular, the directions of $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$ converge to $\pm\boldsymbol{e}_1$. ☺

Definition 9.4.8. *Given an inner product space (say, $\mathbb{C}^n$ with the dot product if you prefer), and any linear transformation A, and any vector $\boldsymbol{v}$, we set $\boldsymbol{v}_0 = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$, and set $\boldsymbol{v}_{i+1} = \frac{A\boldsymbol{v}_i}{\|A\boldsymbol{v}_i\|}$. Then if the limit exists and $\lim_{t\to\infty} \boldsymbol{v}_t = \boldsymbol{w}$, then we say $\boldsymbol{v}$ converges in direction to $\boldsymbol{w}$ under iterations of A.*

Proposition 9.4.9. *If $\boldsymbol{v}$ is in a generalized eigenspace for some eigenvalue $\lambda > 0$ of A, then it converges in direction to some eigenvector of λ under iterations of A.*

This can be proven easily with basic topology, which is outside of the scope of this class. (The unit sphere is compact, the rest is easy.) Of course we cannot do that here. So now let us prove this using linear algebra instead.

Proof. Suppose $\boldsymbol{v}$ is in the generalized eigenspace for the eigenvalue λ. Then $A - \lambda I$ would kill it in finitely many steps, say $\boldsymbol{v} \mapsto (A - \lambda I)\boldsymbol{v} \mapsto \cdots \mapsto (A - \lambda I)^{k-1}\boldsymbol{v} \mapsto \boldsymbol{0}$ where $(A - \lambda I)^{k-1}\boldsymbol{v} \neq \boldsymbol{0}$.

Let V be the span of $\boldsymbol{v}, (A-\lambda I)\boldsymbol{v}, \ldots, (A-\lambda I)^{k-1}\boldsymbol{v}$. (Recall that these vectors are linearly independent.) It should be very obvious that V is a k-dimensional $(A-\lambda I)$-invariant subspace. Hence it is also A-invariant. Furthermore, if we restrict the domain and codomain to V, and use basis $(A - \lambda I)^{k-1}\boldsymbol{v}, \ldots, \boldsymbol{v}$, then $A - \lambda I$ would have matrix $\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}$. As a result, A would have a matrix of $\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$. This is a $k \times k$ matrix.

So our problem is reduced to this: we can assume that the space we study is $\mathbb{C}^k$, and the matrix is simple a single Jordan block $\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$. Set $\boldsymbol{v} = \boldsymbol{e}_k$ the last standard basis vector, we aim to show that the sequence $\boldsymbol{v}_i$ as defined would converge to an eigenvector. Also note that the only eigenvectors in this space are multiples of $\boldsymbol{e}_1$.

Note that $\boldsymbol{v}_t$ is in the same direction of $A^t\boldsymbol{v}$, and its coordinates should corresponds to the last n coefficients of the polynomial $(x + \lambda)^t$. By the binomial theorem, it is $\begin{bmatrix} \binom{t}{k-1}\lambda^{t-k+1} \\ \vdots \\ \binom{t}{0}\lambda^t \end{bmatrix}$.

(Notation: Here $\binom{n}{k}$ means the number of ways to choose k objects out of n objects. In Chinese textbooks it is more traditionally written as C_n^k, the combinatorial number to choose k out of n.)

Since we only care about the direction, we can divide all coordinates by the same constant λ^t. Then we are looking at the vector $\begin{bmatrix} \binom{t}{k-1}\lambda^{-k+1} \\ \vdots \\ \binom{t}{0}\lambda^0 \end{bmatrix}$. Note that the coordinates are all polynomials of t, and the i-th coordinate is a polynomial of t of degree $k-i$. In particular, as $t \to \infty$, eventually the first coordinate (polynomial of degree $k-1$) will outgrow everyone else (polynomials of lower degree). So the direction converge towards $\boldsymbol{e}_1$ indeed. □

The proof above should clarify the following idea: generalized eigenspace is where killing chains happen (for the corresponding $A - \lambda I$). And each killing chain corresponds to some indecomposable invariant subspace (cannot be the direct sum of two smaller invariant subspaces), on which the linear map will be a Jordan block. In this sense, Jordan blocks are indeed the "atoms" of a linear map.

What if $\lambda = 0$? Then the sequence $A^t\boldsymbol{v}$ is going to be $\mathbf{0}$ in finitely many steps. So it does not converge to any direction, since it becomes zero.

What if $\lambda < 0$? By basically the same proof the sequence $(A^t\boldsymbol{v})$ for all even t is going to converge to an "eigendirection", while for all odd t the sequence will converge to the negation of the previous direction. It is "alternating", but they all converge to the same "eigenline".

Proposition 9.4.10. *Again suppose we have an inner product space, say $\mathbb{C}^n$ with dot product.*

Suppose V is an A-invariant subspace of $\mathbb{C}^n$ in which all non-zero vectors converge in direction to some eigenvector of $\lambda > 0$, then V is inside the generalized eigenspace of λ for A.

(In short, the generalized eigenspace of $\lambda > 0$ is the UNIQUE LARGEST A-invariant subspace, where all vectors converges in direction to some λ-eigendirection.)

Proof. Pick any $\boldsymbol{v} \in V$. Then since it is A-invariant, linear combinations of $\boldsymbol{v}, A\boldsymbol{v}, \dots$ are all in V. In particular, $(A-\lambda I)^n\boldsymbol{v} \in V$. Suppose $(A-\lambda I)^n\boldsymbol{v}$ is non-zero, then it converges in direction to an eigenvector of λ.

Now note that the whole domain decomposes as a direct sum of $N_\infty(A-\lambda I)$ and $R_\infty(A-\lambda I)$. Then we have a corresponding decomposition $\boldsymbol{v} = \boldsymbol{v}_N + \boldsymbol{v}_R$. Then $(A-\lambda I)^n\boldsymbol{v} = (A-\lambda I)^n\boldsymbol{v}_N + (A-\lambda I)^n\boldsymbol{v}_R = (A-\lambda I)^n\boldsymbol{v}_R$. Since $R_(A-\lambda I)$ is A-invariant, we would still have $(A-\lambda I)^n\boldsymbol{v}_R \in R_\infty(A-\lambda I)$. As a result, we have $(A-\lambda I)^n\boldsymbol{v} \in R_\infty(A-\lambda I)$.

In particular, if $(A-\lambda I)^n\boldsymbol{v}$ converges in direction to some unit vector, that unit vector must still be inside $R_\infty(A-\lambda I)$. But since it is also in V, it must converge in direction to some unit vector in $N_\infty(A-\lambda I)$. Contradiction.

Hence we must conclude that $(A-\lambda I)^n\boldsymbol{v} = \mathbf{0}$, which means $\boldsymbol{v} \in N_\infty(A-\lambda I)$. □

The other cases are similar. We put the result here without proof.

If $\lambda = 0$, then the generalized eigenspace is the UNIQUE LARGEST A-invariant subspace on which A eventually kills everything.

And if $\lambda < 0$, then the generalized eigenspace is the UNIQUE LARGEST A-invariant subspace, where all vectors converges alternatingly to some λ-eigenline.

So we have a geometric description of generalized eigenspaces.

Example 9.4.11. The requirement that A-invariance is important! There are indeed (non-invariant) subspaces OUTSIDE of the generalized eigenspace for λ, where all non-zero vectors converges to λ-eigendirections.

Consider $A = \begin{bmatrix} 1 & 1 & \\ 0 & 1 & \\ & & 2 \end{bmatrix}$. I claim that all vectors NOT in the xy-plane would converge in direction to $\boldsymbol{e}_3$.

To see this, say we started with $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ where $c \neq 0$. Then $A^t\boldsymbol{v} = \begin{bmatrix} a+tb \\ b \\ c2^t \end{bmatrix}$. CLearly the last coorinate would dominate, and the vector's direction would be closer and closer to $\boldsymbol{e}_3$.

In general, if $\boldsymbol{v}$ has non-zero components involving many different generalized eigenspaces, then the λ with largest absolute value would dominate the convergence behavior of $\boldsymbol{v}, A\boldsymbol{v}, \ldots$.

What if some eigenvalues involved with $\boldsymbol{v}$ have the same absolute value? Then something funny might happen. Consider $\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$. Then other than the eigenvectors, nothing else would converge to eigendirections or even eigenlines. They simply bounce. ☺

9.5 Sylvestor's equation

There are many proofs of Jordan canonical form. Our proof here is essentially a geometric proof. We break down into invariant subspaces and yada yada done. There is also a very interesting (but less illuminating) algebraic proof, where we study polynomials and yada yada done. (Maybe I'll type up another optional section about this.)

Finally, here is a computational proof, using Schur decompositions, and row and column operations, we shall achive a block-diagonalization without using generalized eigenstuff.

First, by Schur decomposition, we can always upper triangularize a matrix. Here is a particularly interesting example

Example 9.5.1. Consider $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$. We know that it has eigenvalue 1 with algebraic multiplicity 2 and eigenvalue 2 with algebraic multiplicity 1.

Let us first try to put it in upper triangular form. When we do this, by picking the right filtration, we want to make sure that we are grouping eigenvalues of the same value together. So say we require the resulting upper triangular matrix to have diagonal 1,1,2.

Then first we need a vector $\boldsymbol{v}_1$ for eigenvalue 1, say $\boldsymbol{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. Note that we can do a (non-invariant) decomposition of the domain into $\mathbb{R}^3 = V_y \oplus V_{xz}$ where V_y represents the y-axis, while V_{xy} is the xz-plane. Then since $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$, the corresponding submaps of A would be $A_{y\to y} = \begin{bmatrix} 1 \end{bmatrix}$, $A_{y\to xz} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $A_{xz\to y} = \begin{bmatrix} -1 & 2 \end{bmatrix}$, $A_{xz\to xz} = \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}$.

So to continue our filtration, since we already have V_y chosen, we need to look at V_{xz} and thus the linear map $A_{xz\to xz}$. Let us find an eigenvector $\boldsymbol{v}_2$ of $A_{xz\to xz}$ for eigenvalue 1, say $\boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in V_{xz}$ (the coordinates here are under the basis $\boldsymbol{e}_1, \boldsymbol{e}_3$ for V_{xz}). Then it is in fact the unit vector in the z-axis, i.e., $\boldsymbol{v}_2 = \boldsymbol{e}_3$. You can check that $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is indeed A-invariant.

Now we already have $\boldsymbol{v}_1, \boldsymbol{v}_2$ chosen. To finish the filtration, we just need to pick any $\boldsymbol{v}_3$ that make this into a basis. Since we have $\boldsymbol{v}_1 = \boldsymbol{e}_2, \boldsymbol{v}_2 = \boldsymbol{e}_3$, we might as well just pick $\boldsymbol{v}_3 = \boldsymbol{e}_1$, and we are done.

Under the basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, we have A similar to $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] = \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix}$ with $A_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} 2 \end{bmatrix}$ and $B = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$. ☺

So by choosing the right filtration, our matrix is something like, say, $\begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix}$ where A_1 and A_2 has NO common eigenvalues. Now we would like to kill B to make this block diagonal. How to do this?

We want to perform $C \begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix} C^{-1}$ so that the resulting matrix is block diagonal. Note that since C here is invertible, it must corresponds to some row/column operations that must happen in pairs. Can we find the right row/column operation to do this?

Suppose $C = \begin{bmatrix} I & X \\ & I \end{bmatrix}$, i.e., it is a block operation. Then $C \begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix} C^{-1} = C \begin{bmatrix} A_1 & B + XA_2 - A_1X \\ & A_2 \end{bmatrix} C^{-1}$. So we need to find X such that $A_1X - XA_2 = B$ for given A_1, A_2, B. This is the Sylvester's equation.

Theorem 9.5.2. *Suppose A, B are $m \times m$ matrix and $n \times n$ matrix with no common eigenvalue. Then for any $m \times n$ matrix C, there is a UNIQUE solution X to the matrix equation $AX - XB = C$.*

Proof. First of all, let V be the space of all $m \times n$ matrices. Consider the map $L : V \to V$ such that $L(X) = AX - XB$. Note that, indeed, L would send an $m \times n$ matrix to another $m \times n$ matrix, and it is also linear! This means that it is an linear operator. Our goal is to show that L is a bijection, hence it is enough to check that the kernel of L is trivial.

So we have reduced our problem to this: we need to show that $AX - XB = 0$ must only have the solution $X = 0$. (See how the problem is simplified? THAT is why we do abstract vector spaces. We do not even need V, L from now on, but the abstraction allows us to SEE that we have a simplification.)

Suppose $AX - XB = 0$, then $AX = XB$. In particular, $A^kX = XB^k$ for any positive integer k. Now we take linear combinations of powers, we see that $p(A)X = Xp(B)$ for any polynomial $p(x)$.

Consider $p_A(x)$, the characteristic polynomial of A. Then on one hand, $p_A(A) = 0$. On the other hand, since A, B has no common eigenvalue, for each eigenvalue λ of B, $p_A(\lambda) \neq 0$. So $p_A(B)$ has NO eigenvalue zero. In particular, it is invertible! Hence we have $0 = p_A(A)X = Xp_A(B)$ where $p_A(B)$ is invertible, so $X = 0$ is the only solution. □

Example 9.5.3. We have A similar to $\left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] = \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix}$.

Now, since A_1 and A_2 has NO eigenvalue in common, we know that there is a unique $X \in M_{2\times 1}$ such that $A_1X - XA_2 = B$. Then A is similar to $\begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$.

To be more explicit, if $X = \begin{bmatrix} x \\ y \end{bmatrix}$, then $XA_2 - A_1X = -B$ would translate into $\begin{bmatrix} 2x \\ 2y \end{bmatrix} - \begin{bmatrix} x + 2y \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$, which means that $x = -5$ and $y = -3$. Then A is similar to $\left[\begin{array}{cc|c} 1 & 0 & -5 \\ 0 & 1 & -3 \\ \hline 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$. This corresponds to a spatial decomposition of V into invariant subspaces.

Finally, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is a Jordan block, and $A_2 = (2)$ is already a Jordan block. So A is similar to $\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 2 & 0 \\ \hline 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right] \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$ is block diagonal with Jordan blocks on the diagonal.

If you have been keeping track, we have done the Jordan canonical form of A with exclusively row/column operations that come in inverse paris, i.e., each step is $A \to XAX^{-1}$ for some elementary matrix X. ☺

Chapter 10

Functions of Matrices

10.1 Limit of Matrices

Whenever we have a collection of things, and a concept of "distance" between things, then we can define limits in the sense of shrinking distance. In this way, we can easily define limits of vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$. And treating matrices as vectors in $\mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$, we can define limits of matrices.

So, as an operational definition, one may think of the limit of a sequence of matrices $\{M_k\}_{k\in\mathbb{N}}$ as taking a limit on each entry.

Of course, technically speaking, this definition is bad. The "entries" of a matrix depend on your choice of basis. If you change basis, then all entries are now different. Who can guarantee that the limit will stay the same?

We can ad-hoc verify that this is the case.

Proposition 10.1.1. *If $\lim A_n = A$ and $\lim B_n = B$, then $\lim(A_nB_n)$ exists and it is AB.*

Proof. One line calculation proof. $\lim(\sum_k a_{ik,n}b_{kj,n}) = \sum_k \lim(a_{ik,n})\lim(b_{kj,n})$. □

Corollary 10.1.2. $\lim(BA_nB^{-1}) = B(\lim A_n)B^{-1}$. *So limits are invariant under a change of basis.*

But ad-hoc arguments are like cheating. A GOOD definition should make this clear in the first place. We do not require this good definition, but if you are curious, read the following remark.

Remark 10.1.3. *This is a exposition on how to define limits of linear operators without picking a basis. This portion is optional.*

A sequence of vectors in an abstract vector space has no well-defined limit. This is because there is no way to measure distance (or induce some topology), and therefore there is no way to measure convergence.

But with inner product structures, we are now golden. Given a sequence of vectors $\{\boldsymbol{v}_n\}_{n\in\mathbb{N}}$ in an inner product space V, we say their limit is $\boldsymbol{v}$ if for all $\epsilon > 0$, we can find $N \in \mathbb{N}$ such that $\|\boldsymbol{v} - \boldsymbol{v}_n\| < \epsilon$ whenever $n \geq N$. You know, the obvious way to define this.

Given linear maps $L, L' : V \to W$ between two inner product spaces, how to define distance? It turns out that there are many ways to define this distance. One would be the operator norm, where we define the norm $\|L\|$ to be the largest $\|L\boldsymbol{u}\|$ for all unit vectors $\boldsymbol{u}$. In particular, it is the largest possible length-dilation that can happen, $\max_{\boldsymbol{v}\in V} \frac{\|L\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$. This looks nice, yes? For any input $\boldsymbol{v}$, we shall always have $\|L\boldsymbol{v}\| \leq \|L\|\|\boldsymbol{v}\|$, and the norm $\|L\|$ is exactly the tightest possible constant k for $\|L\boldsymbol{v}\| \leq k\|\boldsymbol{v}\|$ to work for all $\boldsymbol{v}$.

It is even easy to conceptualize: it is exactly the largest singular value of L. (NOT the eigenvalue!) Neat!

Then using $\|L - L'\|$ as a distance between two linear maps, we can then define $L = \lim L_n$ in the obvious way. I.e., for all $\epsilon > 0$, we can find $N \in \mathbb{N}$ such that $\|L - L_n\| < \epsilon$ whenever $n \geq N$.

Note that our operator norm satisfy the condition that $\|LL'\| \leq \|L\|\|L'\|$. (Easy to prove as $\|LL'\boldsymbol{v}\| \leq \|L\|\|L'\boldsymbol{v}\| \leq \|L\|\|L'\|\|\boldsymbol{v}\|$.) As a result, if L_n converge to L and say L'_n converge to L', then $L_nL'_n$ would

converge to LL'. I.e., we have the identity $\lim(L_nL'_n) = (\lim L_n)(\lim L'_n)$ whenever the latter two limits exist.

Now since matrix multiplication respect limits, if we pick orthonormal basis and assume that our domain and codomain are $\mathbb{C}^n, \mathbb{C}^m$, then we see that $\lim(\boldsymbol{e}_i^* L_n \boldsymbol{e}_j) = \boldsymbol{e}_i^*(\lim L_n)\boldsymbol{e}_j$. So if we picked some basis, then linear operator convergence is the same as convergence in all entries.

Let us define this norm in a different way. You may recall (or you can verify) that $\operatorname{trace}(L^*L')$ is an inner product of the space of linear maps from V to W, and we may define $\|L\|^2 = \operatorname{trace}(L^*L)$. To be more clear that this is independent of basis, we actually have $\|L\| = \sqrt{\sum \sigma_i^2}$ where σ_i are all the singular values. If we had picked an orthonormal basis, then we also have $\|L\| = \sqrt{\sum a_{ij}^2}$ where a_{ij} are all the entries. Neat right? This is a very natural way to define a norm, and it is NOT the same as the operator norm.

But worry not. You may verify that we still have $\|LL'\| \leq \|L\|\|L'\|$, and therefore we also have $\lim(L_nL'_n) = (\lim L_n)(\lim L'_n)$ and $\lim(\boldsymbol{e}_i^* L_n \boldsymbol{e}_j) = \boldsymbol{e}_i^*(\lim L_n)\boldsymbol{e}_j$.

So in the end, it does not matter much which norm we pick. The only important property here is $\|LL'\| \leq \|L\|\|L'\|$. As long as this condition is true, then the convergences in different settings mean exactly the same thing. The TOPOLOGY is the same.

Finally, above statements applies strictly to finite dimensional cases. For infinite dimensional spaces, the two norms above would induce different topologies and will have different meaning of convergence.

Now we have a TOPOLOGY (a way to talk about convergence) on matrices. Then we can define dense subsets.

Theorem 10.1.4. *Diagonalizable matrices are dense in $n \times n$ matrices. (I.e., any matrix is a limit of diagonalizable matrices.)*

Proof. Given a matrix A, how to construct a sequence of diagonalizable matrices whose limit is A? First, we change basis and assume that A is in Jordan canonical form (or any upper triangular form).

Say the diagonal entries (eigenvalues) are $a_1, \ldots, a_n$. Note that some of these are the same, while some are not. Let g be the smallest "gap" between distinct diagonal entries, i.e., either $a_i = a_j$, or $|a_i - a_j| \geq g$.

For a tiny real number $t < \frac{g}{2n}$, consider a diagonal matrix $D(t) = \begin{bmatrix} t & & \\ & \ddots & \\ & & nt \end{bmatrix}$, let $A_t = A + D_t$. Then $\lim_{t\to 0} A_t = A$. I only need to show that A_t are diagonalizable.

Note that eigenvalues of A_t are $a_1 + t, \ldots, a_n + nt$. For any $i \neq j$, if $a_i = a_j$, then $a_i + it \neq a_j + jt$. If $|a_i - a_j| \geq g$, then $|(a_i + it) - (a_j + jt)| \geq g - it - jt \geq g - 2nt > 0$ by construction of t, so $a_i + it \neq a_j + jt$. Eitherway, we see that eigenvalues of A_t are all distinct, so it must be diagonalizable. Done. □

Note that we in fact proved something stronger: matrices with distinct eigenvalues are dense. Feel free to prove something even stronger: INVERTIBLE matrices with distinct eigenvalues are dense. (Just throw in distance to zero when you define the "gap" size g.)

This fact is extremely useful. Consider this:

Corollary 10.1.5. *Given a square matrix A, let A_{ij} be its (i,j)-cofactor, and let $\operatorname{Adj}(A)$ be the adjugate matrix of A. (So for invertible matrices, $A^{-1} = \frac{1}{\det(A)}\operatorname{Adj}(A)$. Note that for non-invertible matrices, $\operatorname{Adj}(A)$ is still defined.)*

Then for any square matrices A, B, we havbe $\operatorname{Adj}(AB) = \operatorname{Adj}(B)\operatorname{Adj}(A)$.

Proof. Note that invertible matrices are dense. And for invertible matrices, $\operatorname{Adj}(AB) = \det(AB)(AB)^{-1} = \det(A)B^{-1}\det(A)A^{-1} = \operatorname{Adj}(B)\operatorname{Adj}(A)$. Now take limit and we are done. □

Or for example, let us prove Cayley-Hamilton again. First, if A has distinct eigenvalues, then it is trivial to verify that $p_A(A) = 0$. (A is diagonalizable, so there is a basis made of eigenvectors of A. And $p_A(A)$ will kill all eigenvectors of A.) Then by taking limits of A_n with distinct eigenvalues, we have $p_A(A)$ for any matrix A.

Remark 10.1.6. *The adjugate matrix is NOT useful at all. It is an attempt to relate the matrix A with its inverse. However, the Cayley-Hamilton theorem does a better job at this.*

If A is invertible, then $\det(A) \neq 0$, so $p_A(x)$ has a non-zero constant term. I.e., $p_A(x) = xq(x) + a$ for some $a \neq 0$. Then since $p_A(A) = 0$, we have $Aq(A) + aI = 0$, and thus $A^{-1} = \frac{1}{a}q(A)$. So A^{-1} is ALWAYS a polynomial of A!

A polynomial relation is huge. Whatever you can do using adjugates, you can use Cayley-Hamilton instead. For example, a classical argument for the adjugate matrix goes like this: if entries of A are rational, then entries of A^{-1} are also rational. To see this, note that each cofactor is a sum of products of entries of A, so it is rational. So we are done. We also see that if A has integer entries, then $\det(A)A^{-1}$ has integer entries.

But with Cayley-Hamilton, $A^{-1} = \frac{1}{\det(A)}q(A)$, and the coefficients of $q(x)$ are also sums of products of entries of A. So if A has rational entries, then A^{-1} has rational entries, and if A has integer entries, then $\det(A)A^{-1}$ has integer entries.

10.2 Functions of matrices

What is a function of a matrix? Here is an easy example:

Definition 10.2.1. *We define e^A to be the limit $\lim_{n\to\infty}(I + A + \frac{1}{2!}A^2 + \cdots + \frac{1}{n!}A^n)$. This is the limit of a sequence of matrix.*

This raises an immediate problem. Why would this series converge at all? (Spoiler: it will always converge.) If we were in an analysis class, then we shall then proceed to show convergence. It is not too bad, as entries of A^n grows polynomially while the denominator $n!$ grows faster than exponential.

But as a linear algebra class, let us jump out of this, and think about something bigger. If YOU were to define a function of a matrix, $f(A)$ for some function f, what would you like?

The following principles seem like must-haves:

1. We want it to NOT depend on our choice of basis. So $f(BAB^{-1}) = Bf(A)B^{-1}$. It is really a function of LINEAR TRANSFORMATIONS.
2. We want it to respect independent actions. So $f(\begin{bmatrix} A & \\ & B \end{bmatrix}) = \begin{bmatrix} f(A) & \\ & f(B) \end{bmatrix}$. In particular, for diagonal matrices, $f(D)$ is just applying f on each diagonal entry.
3. If $f : \mathbb{R} \to \mathbb{R}$ or $f : \mathbb{C} \to \mathbb{C}$ is continuous, then the induced function $f : M_{n\times n} \to M_{n\times n}$ should still be continuous. Here $M_{n\times n}$ refers to the space of all $n \times n$ real or complex matrices, depending on context. (We can use real functions when all of our eigenvalues are real.)

Combining these principles, one thing is super clear. If A is diagonalizable $A = BDB^{-1}$ where $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$, then we want $f(A) = Bf(D)B^{-1} = B\begin{bmatrix} f(d_1) & & \\ & \ddots & \\ & & f(d_n) \end{bmatrix}B^{-1}$. So this resulting matrix $f(A)$ is already uniquely defined! It is also not hard to see that, if A changes continuously (i.e., B and each d_i changes continuously), then since f is continuous on $\mathbb{C}$, $f(d_i)$ also changes continuously, and hence $f(A) = Bf(D)B^{-1}$ changes continuously. So we have all the desired result.

BUT what if A is NOT diagonalizable? This is where density comes into play. According to our principles, $f(\lim A_n) = \lim f(A_n)$. So we just use a sequence of diagonalizable matrices to approximate A, and we can get $f(A)$.

Would the limit always exists? Let us see what would happen.

Example 10.2.2. Consider $J = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$. Let $J_t = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda + t \end{bmatrix}$, then clearly J_t is diagonalizable whenever $t \neq 0$, and $\lim_{t\to 0} J_t = J$.

So for a function f, we want $f(J) = f(\lim_{t\to 0} J_t) = \lim_{t\to 0} f(J_t)$. To calculate $f(J_t)$, we need to diagonalize J_t. Note that $J_t = \begin{bmatrix}\lambda & 1\\ 0 & \lambda+t\end{bmatrix}$ can be diagonalized by solving the corresponding sylvesters equation $\lambda x - x(\lambda+t) = 1$, which yeilds $x = -\frac{1}{t}$. So $\begin{bmatrix}1 & \frac{1}{t}\\ 0 & 1\end{bmatrix} J_t \begin{bmatrix}1 & -\frac{1}{t}\\ 0 & 1\end{bmatrix} = \begin{bmatrix}\lambda & \\ & \lambda+t\end{bmatrix}$. In particular, we have $J_t = \begin{bmatrix}1 & \frac{1}{t}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\lambda & \\ & \lambda+t\end{bmatrix}\begin{bmatrix}1 & -\frac{1}{t}\\ 0 & 1\end{bmatrix}$.

So $f(J) = f(\lim_{t\to 0} J_t) = \lim_{t\to 0} f(J_t) = \lim_{t\to 0}\begin{bmatrix}1 & \frac{1}{t}\\ 0 & 1\end{bmatrix}\begin{bmatrix}f(\lambda) & \\ & f(\lambda+t)\end{bmatrix}\begin{bmatrix}1 & -\frac{1}{t}\\ 0 & 1\end{bmatrix} = \lim_{t\to 0}\begin{bmatrix}f(\lambda) & \frac{f(\lambda+t)-f(\lambda)}{t}\\ & f(\lambda+t)\end{bmatrix}$.

Wait, the definition of the derivative is right there!

So we must have $f(J) = \begin{bmatrix}f(\lambda) & f'(\lambda)\\ & f(\lambda)\end{bmatrix}$ when f is differentiable at λ. Otherwise $f(J)$ cannot be defined and $\lim f(J_t)$ does not converge. ☺

This line of logic can easily be generalized to give us a formula for $f(A)$ in general. But for mnemonics sake, let us see an alternative proof.

Proposition 10.2.3. *Assume that f is analytical at λ. (It means f equals to its Taylor expansion at λ.) Consider the $n\times n$ Jordan block $J = \begin{bmatrix}\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda\end{bmatrix}$. Then using previous principles, we must have*

$$f(J) = \begin{bmatrix}f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(n-1)!}f^{(n-1)}(\lambda)\\ & \ddots & \ddots & \vdots\\ & & \ddots & \frac{1}{1!}f'(\lambda)\\ & & & f(\lambda)\end{bmatrix}.$$

Proof. Why do we see coefficients of Taylor expansions? That is not a coincidence. First we have $J = N+\lambda I$ where N is the nilpotent Jordan block.

Now f equals to its Taylor series. So if we expand f at λ, we have $f(x) = a_0 + a_1(x-\lambda) + a_2(x-\lambda)^2 + \dots$ where $a_k = \frac{1}{k!}f^{(k)}(\lambda)$. So $f(J) = a_0 I + a_1 N + a_2 N^2 + \dots$. But as a nilpotent matrix, $N^n = 0$, and N^k is really just the identity matrix shifted up k times. So $f(J) = a_0 I + a_1 N + a_2 N^2 + \dots + a_{n-1}N^{n-1} = \begin{bmatrix}f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(n-1)!}f^{(n-1)}(\lambda)\\ & \ddots & \ddots & \vdots\\ & & \ddots & \frac{1}{1!}f'(\lambda)\\ & & & f(\lambda)\end{bmatrix}$. □

Now we can define functions of matrices.

Definition 10.2.4. *Suppose f is a function defined at all eigenvalues of A, and it is $(m-1)$-times differentiable at the eigenvalue λ when λ-blocks in the Jordan canonical form of A have sizes at most m. Then we define $f(J)$ for each involved Jordan blck as*

$$f(J) = \begin{bmatrix}f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(m-1)!}f^{(m-1)}(\lambda)\\ & \ddots & \ddots & \vdots\\ & & \ddots & \frac{1}{1!}f'(\lambda)\\ & & & f(\lambda)\end{bmatrix}$$

, and we define

$$f(A) = B \begin{bmatrix} f(J_1) & & \\ & \ddots & \\ & & f(J_t) \end{bmatrix} B^{-1}$$

where the Jordan decomposition of A is $B \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_t \end{bmatrix} B^{-1}$.

Here is an obvious result:

Corollary 10.2.5. *If f is infinitely differentiable everywhere, then $f(A)$ is defined for all square matrix A.*

Corollary 10.2.6. *If A has eigenvalues $\lambda_1, \dots, \lambda_n$ counting algebraic multiplicity, then $f(A)$ has eigenvalues $f(\lambda_1), \dots, f(\lambda_n)$ counting algebraic multiplicity.*

Proof. Do it block-wise. □

Corollary 10.2.7. *$f(A)g(A) = h(A)$ if $f(x)g(x) = h(x)$, and $f(A) + g(A) = h(A)$ if $f(x) + g(x) = h(x)$. Finally, if $f(x) = x$, then $f(A) = A$, and if $f = 1$ is a constant function, then $f(A) = I$.*

Proof. Do it block-wise. □

Corollary 10.2.8. *If f is a polynomial, then $f(A)$ is exactly as we have always defined it to be.*

One can then do the boring verification that such a definition satisfy the given principles. We are going to skip those because we might not learn much from that process.

Corollary 10.2.9. *If $f = g$ at all eigenvalues of A and they also equal at enough derivatives that are used in $f(A)$ and $g(A)$, then $f(A) = g(A)$.*

Corollary 10.2.10. *Fix a matrix A, then for any well-defined $f(A)$, there is a polynomial $p(x)$ such that $f(A) = p(A)$. (Be careful, the choice of $p(x)$ here depends on A. For different A and same f, we would need to choose different $p(x)$.)*

Proof. See Hermite interpolations, which is done via Chinese remainder theorem in abstract algebra. (Ring theory.) □

So if we are FIXING A, then there is NO point in studying $f(A)$ at all. They are all just polynomials of A. In particular, we have results like $Af(A) = f(A)A$ always, and etc.

However, be careful here. If we are fixing f, but changing A, then each different A might require a different polynomial. So it is better to study $f(A)$ in terms of f.

Example 10.2.11. For $A = I$, then $\mathrm{e}^A = \mathrm{e}A$, so $f(A) = p(A)$ where $p(x) = \mathrm{e}x$.

But for $A = \begin{bmatrix} 0 & 1 \\ & 0 \end{bmatrix}$, then $\mathrm{e}^A = I + A = q(A)$ where $q(x) = x + 1$. ☺

Finally, let us have some easy and useful propositions.

Proposition 10.2.12. $f(A^{\mathrm{T}}) = f(A)^{\mathrm{T}}$.

Proof. Suppose $A = BJB^{-1}$ where J is the Jordan canonical form, then $f(A) = Bf(J)B^{-1}$. We also see that $A^{\mathrm{T}} = CJ^{\mathrm{T}}C^{-1}$ where $C = (B^{-1})^{\mathrm{T}}$, and thus $f(A^{\mathrm{T}}) = Cf(J^{\mathrm{T}})C^{-1}$, while $f(A)^{\mathrm{T}} = Cf(J)^{\mathrm{T}}C^{-1}$. So it is enough to show that $f(J^{\mathrm{T}}) = f(J)^{\mathrm{T}}$ for any Jordan canonical form. But since J is block diagonal, it is then enough to show this for a single Jordan block.

(This paragraph is NOT part of the proof, merely some explorative exposition.) For a single Jordan block, how are J and J^{T} related? For the sake of clarification, let us assume that J is a nilpotent Jordan

block. Then J is characterized by the killing chain $\boldsymbol{e}_n \mapsto \boldsymbol{e}_{n-1} \mapsto \cdots \mapsto \boldsymbol{e}_1 \mapsto \boldsymbol{0}$. It is easy to see that J^{T} is similarly characterized by the killing chain $\boldsymbol{e}_1 \mapsto \boldsymbol{e}_2 \mapsto \cdots \mapsto \boldsymbol{e}_n \mapsto \boldsymbol{0}$. So to convert between J and J^{T}, we need to flip the entire order of the standard basis!

Let $T = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$, which is the matrix that flips the entire order of the standard basis. Then you may verify that for ANY matrix of the form $X = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_1 \\ & & & a_0 \end{bmatrix}$, then $TXT^{-1} = X^{\mathrm{T}}$. Note that for any Jordan block J, both J and $f(J)$ are matrices of this type. So $f(J^{\mathrm{T}}) = f(TJT^{-1}) = Tf(J)T^{-1} = f(J)^{\mathrm{T}}$. So we are done. □

We do NOT have $f(A^{-1}) = f(A)^{-1}$. Why? Because for functions, we in general do NOT have $f(g(x)) = g(f(x))$ when $g(x) = x^{-1}$. E.g., if $f(x) = x + 1$, then $\frac{1}{x+1} \neq \frac{1}{x} + 1$. In fact, if A is invertible, $f(A)$ may even be NON-invertible, say when f is the constant zero function.

Of course, I am assuming the following fact, which (not surprisingly) is true.

Proposition 10.2.13. *If $f(x) = x^{-1}$, and A is invertible, then $f(A) = A^{-1}$.*

Proof. If A is invertible, then it has no zero eigenvalue, so $f(A)$ is well-defined. Let $g(x) = x$. Then $1 = f(x)g(x)$. So $I = f(A)g(A) = f(A)A$. So $f(A) = A^{-1}$. □

Here is a DANGEROUS thing: do we have $f(A^*) = f(A)^*$ for a complex matrix? Well, we might NOT have this!

Since we already have $f(A^{\mathrm{T}}) = f(A)^{\mathrm{T}}$, all we need now is $f(\overline{A}) = \overline{f(A)}$. However, do we always have $f(\overline{x}) = \overline{f(x)}$ for any complex function f and complex number x? This is NOT always true.

For example, suppose $f(x) = \mathrm{i}x$. Then $f(1 + \mathrm{i}) = \mathrm{i} - 1$, while $f(1 - \mathrm{i}) = \mathrm{i} + 1$. The resulting image is NOT complex conjugates of each other! In particular, $f(\overline{A}) = \overline{f(A)}$ fails for even 1×1 matrices.

Remark 10.2.14. *In fact, if $f(x) = \mathrm{i}x$, then we can define $\overline{f}(x) = -\mathrm{i}x$. The idea is that we take complex conjugate on all coefficients, but NOT on the input. Then we in fact have $\overline{f(x)} = \overline{f}(\overline{x})$, which makes a LOT more sense. The idea is that, when you do $\overline{f(x)}$, the complex conjugate would not only hit x, but also hit f as well.*

Take complex analysis class to properly define $\overline{f}$ and so on.

Our saving grace is the following.

Proposition 10.2.15. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is complex differentiable and $f(\mathbb{R}) \subseteq \mathbb{R}$, then $f(\overline{A}) = \overline{f(A)}$ and $f(A^*) = f(A)^*$ for any complex matrix A.*

Proof. Complex differentiable functions are analytical. So they are infinitely differentiable, and they equal to a power series, i.e., $f(x) = a_0 + a_1 x + \dots$. Furthermore, if $f(\mathbb{R}) \subseteq \mathbb{R}$, then $f(0) \in \mathbb{R}$ which implies that $a_0 \in \mathbb{R}$.

Furthermore, note that $f'(0) = \lim_{t \to 0} \frac{f(t) - f(0)}{t}$. By using only real t to perform the limit $t \to 0$, we see that $f'(0)$ must also be real. So a_1 is real.

Similarly, $2a_2 = f''(0)$ is real, and thus a_2 is real. So on so forth. We see that $(n!)a_n = f^{(n)}(0)$ is real, so a_n is real.

So f is a power series whose coefficients are all real. Now $f(\overline{A}) = a_0 I + a_1(\overline{A}) + \cdots = \overline{a_0 I + a_1 A + \dots} = \overline{(f(A))}$. □

Remark 10.2.16. *Being complex differentiable is a STRONG requirement. For example, $f(x) = \overline{x}$ is NOT complex differentiable, even though it is super nice.*

Consider $\lim_{z\to 0}\frac{f(z)-f(0)}{z}=\lim_{z\to 0}\frac{f(z)}{z}$. *If* z *approaches zero from the real line, then obviously* $f(z)=z$ *for all real* z*, hence the limit is* 1*. However, let* z *approaches zero from the imaginary axis, then* $f(z)=-z$ *for all purely imaginary* z*, so the limit is* -1*. Since the two limits disagree, this complex limit fails to exist. So* f *is NOT differentiable at zero. (In fact, it is differentiable nowhere.)*

In particylar, if $f(z)=\overline{z}$*, then* $f(A)$ *is NOT well-defined for non-diagonalizable* A*!*

Intuitively, a complex function f *being complex differentiable means the function respect angles and orientations locally. If two curves* $a(t), b(t)$ *on the complex plane interset, and their tangent lines at the intersection make an angle of* θ *(positive means counter-clockwise), then the image curves* $f(a(t)), f(b(t))$ *should also intersect and make an angle of* θ*.*

Complex conjugation is NOT differentiable because, while it preserves the absolute value of local angles, it does NOT preserve the orientation. The angle θ *will become* $-\theta$*. Hence it is not complex differentiable.*

In the end, the only complex differentiable functions are power series. Learn more by taking a complex analysis class.

10.3 Applications to functions of Matrices

The obvious application is to solve various differential equations.

Lemma 10.3.1. *If* $AB=BA$*, then* $\mathrm{e}^{A+B}=\mathrm{e}^A\mathrm{e}^B=\mathrm{e}^B\mathrm{e}^A$*.*

Proof. Direct computation using Taylor series of e^x.

$$\mathrm{e}^A\mathrm{e}^B=(\sum_m\frac{1}{m!}A^m)(\sum_n\frac{1}{n!}B^n)=\sum_{m,n}\frac{1}{m!n!}A^mB^n.$$

Now let $k=m+n$. We have

$$\sum_{m,n}\frac{1}{m!n!}A^mB^n=\sum_k\sum_{n=0}^k\frac{1}{n!(k-n)!}A^{k-n}B^n=\sum_k\frac{1}{k!}\sum_{n=0}^k\frac{k!}{n!(k-n)!}A^{k-n}B^n=\sum_k\frac{1}{k!}(A+B)^k.$$

Note that commutativity $AB=BA$ is used in the last step. For example, $A^2+2AB+B^2=(A+B)^2$ is only true when we have commutativity. □

Remark 10.3.2. *When* A *has distinct eigenvalues, then* $AB=BA$ *imples that* A, B *are simultaneously diagonalizable. Hopefully your last linear algebra class has discussed this. But if not, see if you can prove this yourself.*

There are many proofs. If you need a hint, maybe try 2 by 2 matrices. If $\begin{bmatrix}a&\\&b\end{bmatrix}X=X\begin{bmatrix}a&\\&b\end{bmatrix}$ *and* $a\neq b$*, whwy must* X *be diagonal?*

Or one can work abstractly on finding common eigenvectors.

Proposition 10.3.3. $\frac{d}{dt}\mathrm{e}^{At}=A\mathrm{e}^{At}$.

Proof. Compute. Or be cheap and do this for diagonal A, and use density. □

Corollary 10.3.4. *Let* $\boldsymbol{v}(t)$ *be a vector of functions, i.e., each coordinate may change as* t *change. Suppose it satisfy the differential equation* $\boldsymbol{v}'(t)=A\boldsymbol{v}(t)$ *for some linear transformation* A*. Then* $\mathrm{e}^{At}\boldsymbol{c}$ *is a solution for any constant vector* $\boldsymbol{c}$*. (In fact* $\boldsymbol{v}(0)=\boldsymbol{c}$*, so it is the initial condition.)*

I claim that this is in fact the only solution.

Proposition 10.3.5. *The solution space to* $\boldsymbol{v}'(t)=A\boldsymbol{v}(t)$ *is* n *dimensional where* n *is the dimension of the domain. (So columns of* e^{At} *form a basis.)*

Proof. Let us first do a single nilpotent Jordan block. Then we have $\begin{bmatrix} f_1' \\ \vdots \\ \vdots \\ f_n' \end{bmatrix} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ \vdots \\ f_n \end{bmatrix}$. This reads as $f_i' = f_{i+1}$. So the solution space is simply this: f_1 is a polynomial of degree at most $n-1$, and the rest are iterated derivatives of f_1. Obviously the solution space is n dimensional.

Now let us do a single λ-Jordan block J. Let $\boldsymbol{w}(t) = \mathrm{e}^{-\lambda t}\boldsymbol{v}(t)$. Then $\boldsymbol{w}'(t) = \mathrm{e}^{-\lambda t}\boldsymbol{v}'(t) - \lambda \mathrm{e}^{-\lambda t}\boldsymbol{v}(t) = \mathrm{e}^{-\lambda t}(J - \lambda I)\boldsymbol{v}(t) = (J - \lambda I)\boldsymbol{w}(t)$. But $(J - \lambda I)$ is nilpotent, so solutions to $\boldsymbol{w}(t)$ is n dimensional. Hence $\boldsymbol{v}(t) = \mathrm{e}^{\lambda t}\boldsymbol{w}(t)$ has n-dimensions of possibilities as well.

Now suppose A has many Jordan blocks. But being block diagonal means each block behaves independently, so we are reduced to the single block cases and we are done. □

Conclusion: given a differential equation $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$ and initial value $\boldsymbol{v}(0) = \boldsymbol{c}$, then the unique solution is $\mathrm{e}^{At}\boldsymbol{c}$.

You can imagine that things like $\sin(A)$ and such will also help solving other kinds of differential equations. We leave the rest to your future differential equation class.

Let us see another use of functions of matrices.

Definition 10.3.6. *We define the sign function* sign *such that* $\operatorname{sign}(a + b\mathrm{i}) = 1$ *if* $a > 0$, $\operatorname{sign}(a + b\mathrm{i}) = -1$ *if* $a < 0$, *and undefined when* $a = 0$.

It is obvious that this sign function is smooth (infinitely differentiable) whenever the input is NOT purely imaginary. So for any matrix A whose eigenvalues are NOT purely imaginary, then $\operatorname{sign}(A)$ is well-defined. Specifically, for any λ-Jordan block J, then $\operatorname{sign}(J) = \operatorname{sign}(\lambda)I = \pm I$.

Let us consider an application of this sign function.

Example 10.3.7. Consider the following variants of the Sylvester's equation. We want to find X to solve $AX + XB = C$, where A, B have positive eigenvalues.

(This is very possible, because in physics, eigenvalues are usually energy states or some other physical meanings, which we usually want to be positive.)

Solving this equation is the same as finding a diagonalization $\begin{bmatrix} A & -C \\ & -B \end{bmatrix} = \begin{bmatrix} I & X \\ & I \end{bmatrix} \begin{bmatrix} A & \\ & -B \end{bmatrix} \begin{bmatrix} I & -X \\ & I \end{bmatrix}$.

Now apply matrix sign function and watch the magic:

$\operatorname{sign}(\begin{bmatrix} A & -C \\ & -B \end{bmatrix}) = \begin{bmatrix} I & X \\ & I \end{bmatrix} \operatorname{sign}(\begin{bmatrix} A & \\ & -B \end{bmatrix}) \begin{bmatrix} I & -X \\ & I \end{bmatrix} = \begin{bmatrix} I & X \\ & I \end{bmatrix} \begin{bmatrix} I & \\ & -I \end{bmatrix} \begin{bmatrix} I & -X \\ & I \end{bmatrix} = \begin{bmatrix} I & -2X \\ & -I \end{bmatrix}$.

So if we have a magic computer to compute matrix sign function, then to solve $AX + XB = C$, we simply apply the matrix sign function to $\begin{bmatrix} A & -C \\ & -B \end{bmatrix}$ and read the answer from the upper right block. ☺

Remark 10.3.8. *(This part should be moved to earlier sections....)*

The Sylvester's equations are very important. For example, consider the case $AX - XB = C$ where $C = 0$ and B is 1×1. Then we have $AX = Xb$ for some number b, and X is $m \times 1$, a vector! In particular, this is the equation defining eigenvectors and eigenvalues. In general, for the equation $AX = XB$, you may think of the solution X as the B-eigenstuff for A. And $AX - XB = C$ is the inhomogeneous version of this. (Just like how $f' - f = 0$ and $f' - f = x^2$ are related.)

Furthermore, if $AX = XB$, then $\operatorname{Ran}(X)$ *is an invariant subspace of A. Can you see this? (And* $\operatorname{Ran}(X^{\mathrm{T}})$ *is an invariant space of B^{T}.)*

10.4 Matrix exponentials, rotations and curves

Matrix exponentials are super useful. One reason is its ties to rotations.

Example 10.4.1. Consider $A = \begin{bmatrix} 0 & c & -b \\ -c & 0 & a \\ b & -a & 0 \end{bmatrix}$, which is a skew symmetrix matrix. Suppose a, b, c are not all zero. Let us think about the meaning of A and e^A.

First of all, we have $A \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} cy - bz \\ az - cx \\ bx - ay \end{bmatrix}$. This is the formula for a cross product! It is $\begin{bmatrix} x \\ y \\ z \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. According to the geometric meaning of cross product, we can now understand the geometric meaning of A.

Let $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, then for any input $\boldsymbol{x}$, $A\boldsymbol{x}$ is a vector perpendicular to both $\boldsymbol{x}, \boldsymbol{v}$, and under the Euclidean metric, its length is the area of the parallelogram made by $\boldsymbol{x}, \boldsymbol{v}$. But there are two vectors like this! Which one should we pick? Well, we also have the requirement that $\boldsymbol{x}, \boldsymbol{v}, A\boldsymbol{x}$ would make a right-handed system (i.e., $\det(\boldsymbol{x}, \boldsymbol{v}, A\boldsymbol{x}) > 0$).

In partiular, we see that $A\boldsymbol{v} = \boldsymbol{v} \times \boldsymbol{v} = \boldsymbol{0}$. So $\boldsymbol{v}$ is an eigenvector for the eigenvalue zero. And since A us skew-symmetric, we know that all its eigenvalues are purely imaginary, hence the other two eigenvalues are $\pm \mathrm{i}\theta$ for some real number θ.

Now let us try to understand e^A. Since $A\boldsymbol{v} = 0\boldsymbol{v}$, we must have $f(A)\boldsymbol{v} = f(0)\boldsymbol{v}$. Hence $\mathrm{e}^A\boldsymbol{v} = \boldsymbol{v}$. So $\boldsymbol{v}$ is a direction fixed by e^A!

Also note that $A + A^{\mathrm{T}} = O$. Since A and $A^{\mathrm{T}} = -A$ commutes, we have $\mathrm{e}^A(\mathrm{e}^A)^{\mathrm{T}} = \mathrm{e}^A\mathrm{e}^{(A^{\mathrm{T}})} = \mathrm{e}^{A+A^{\mathrm{T}}} = \mathrm{e}^O = I$. Oops! So e^A is an orthogonal matrix! (Also note that since e^x is a power series with real coefficients, it send real matrices to real matrices, so e^A is a real matrix.)

So it is a rotation around $\boldsymbol{v}$. Since A has eigenvalues $0, \mathrm{i}\theta, -\mathrm{i}\theta$, therefore e^A has eigenvalues $1, \mathrm{e}^{\mathrm{i}\theta}, \mathrm{e}^{-\mathrm{i}\theta}$. So e^A is a rotation around $\boldsymbol{v}$ by angle θ.

Finally, let us figure out what θ is. The characteristic polynomial of A is $x^3 + (a^2 + b^2 + c^2)x$. So $\theta = \sqrt{a^2 + b^2 + c^2} = \|\boldsymbol{v}\|$.

In conclusion, if $A = \begin{bmatrix} 0 & c & -b \\ -c & 0 & a \\ b & -a & 0 \end{bmatrix}$, then e^A is a rotation around $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ by an angle $\| \begin{bmatrix} a \\ b \\ c \end{bmatrix} \|$. Neat, yes? ☺

It is not hard to extrapolate the following results from the arguments above.

Lemma 10.4.2. $\det(\mathrm{e}^A) = \mathrm{e}^{\operatorname{trace} A}$.

Proof. If A has eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\det(\mathrm{e}^A) = \prod \mathrm{e}^{\lambda_i} = \mathrm{e}^{\sum \lambda_i} = \mathrm{e}^{\operatorname{trace} A}$. □

Proposition 10.4.3. *If A is real skew-symmetric, then e^A is a real orthogonal matrix with determinant one. (I.e., a rotation matrix.) And if A is skew-Hermitian, then e^A is unitary.*

Proof. DIY. □

Proposition 10.4.4. *If A is a real orthogonal matrix with determinant 1 (i.e., a rotation matrix), then $A = \mathrm{e}^B$ for some real skew-symmetric B.*

Proof. Since A is real orthogonal, $A = BJB^{-1}$ for real B and block diagonal J where each diagonal block of J is either 1, or a 2×2 real rotation matrix. (See spectral theorem for normal matrices. This is in my linear algebra lecture notes last semester.) (Also note that we are pairing up -1 into rotation matrix $\begin{bmatrix} -1 & \\ & -1 \end{bmatrix}$, since -1 must have even algebraic multiplicity.)

So after usual simplification tactics, it is enough to show that the statement is true for a single 2×2 real rotation matrix. Note that $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \mathrm{e}^{\begin{bmatrix} & -\theta \\ \theta & \end{bmatrix}}$. □

The exponential function is not only useful for high dimensional rotations. It also serves as a useful way to "connect" matrices with smooth curves.

Now imagine that an object is rotating. Whatever the object is, at $t = 0$, the object is in its initial state, i.e., we can apply the identity matrix I to the initial state. As t increases, the object will be rotated more and more, in a continuous manner. So at each t, we want to apply some matrix $A(t)$. The matrix $A(t)$ depends on t continuously. In particular, we have a CURVE of matrices.

Given two orthogonal matrices A, B with determinant one, how to find a smooth curve of invertible matrices between them? And how to find the "shortest" or "most efficient" curve between them? (This will be of interests in physics, robotics, making movies, etc..) Note that we want all intermediate matrices to be orthogonal as well.

It turns out that, in the set of orthogonal matrices, e^{tA} when t grows from 0 to 1 is the "straight curve" (geodesic) between the identity matrix and the matrix e^A. (Here A is real skew-symmetric.) And to draw a straight curve between rotations e^A and e^B, it is enough to find a straight curve between I and $\mathrm{e}^{-A}\mathrm{e}^B$ and then apply e^A to the left of everything on this curve, i.e., the desired curve is $\mathrm{e}^A\mathrm{e}^{tC}$ where we find a matrix C such that $\mathrm{e}^C = \mathrm{e}^{-A}\mathrm{e}^B$. (Note that maybe $AB \neq BA$, so usually $C \neq B - A$.)

Intuitively, you can think of e^{tA} as the following. Note that at $t = 0$, we have $\mathrm{e}^{tA}|_{t=0} = I$ and $\frac{\mathrm{d}}{\mathrm{d}t}(e^{tA})|_{t=0} = A$. So this is the curve where we started at the identity matrix, move in the direction of A while remaining inside the set of orthogonal matrices, and go straight in the same direction forever.

Of course, to rigorously prove this, we would need many high dimensional ($\frac{1}{2}(n^2 - n)$ dimensional) geometry. So we leave it as such.

Even though we have no proof that this curve is "straight", we can still do something with it.

Corollary 10.4.5. *The set of real orthogonal matrices has two path-connected components. One component is the set of all orthogonal matrices with determinant* 1*, and the other component is the set of all orthogonal matrices with determinant* -1*.*

Proof. For any real orthogonal matrix e^A with determinant one, it is path-connected to the identity matrix via e^{tA}. So the set of all such matrices is path-connected.

For any real orthogonal matrices A, B with determinant minus one, then $A^{-1}B$ is a real orthogonal matrix with determinant one. Hence there is a path from I to $A^{-1}B$. By applying A to all matrices on this curve, we get a continuous path from A to B. So the set of all such matrices is path-connected.

Finally, how to show that these two components are NOT path-connected? Suppose we have a continuous curve $C : [0, 1] \to M$ where M is the space of all real orthogonal matrices, and $C(0)$ has determinant one while $C(1)$ has determinant minus one. Note that the determinant map is continuous (because it is a sum of products of entries). So $\det \circ C$ is a continuous map. But for each t, either $\det(C(t)) = 1$ or $\det(C(t)) = -1$. So this is a continuous map from $[0, 1]$ to $\{0, 1\}$. So this is a continuous curve on the set $\{0, 1\}$ connecting 0 and 1, which is absurd. So we are done. □

If you like, you can think of these two components as the very definition of "positive orientation" and "negative orientation" in each n-dimensional space.

10.5 Commuting matrices

Matrices are a great source of commutativity. However, most things are not commutative. For example, if $f(x) = 2x + 1$ and $g(x) = 3x + 1$, then in general $f \circ g \neq g \circ f$.

Remark 10.5.1. *Many modern advances in science is essentially the realization that our world is not commutative. By dropping the commutativity assumption, things become unintuitive, ingenius and powerful. For example, general relativity tries to explain various phenomena with the idea of curvature, which is defined in terms of failure of commutativity.*

Suppose we are on a flat world, say $\mathbb{R}^2$*. Then let* A *be moving to the north by 1 unit, and let* B *be moving to the east by 1 unit, then you can see that* $AB = BA$ *since you would end up at the same place. But if we*

live on the sphere (earth?), say we stand on the equator. Then you can verify that $AB \neq BA$. Curvature happens.

Now consider quantum mechanics. In quantum mechanics, the position operator X is defined as a operator that sends a function $f(x)$ to the function $xf(x)$. The momentum operator P is defined as an operator that sends a function to its derivative, say $f(x)$ to $f'(x)$, assuming that the world is one dimensional for simplicity. Then you can verify that $XP \neq PX$, and in fact $(PX - XP)f = (xf(x))' - xf'(x) = f(x)$, so we have $PX - XP = I$ for the identity operator. In physics there will be some extra constant flying around, and this constant is the reduced Planck constant.

The fact that $PX \neq XP$ is at the heart of the uncertainty principle, i.e., you cannot simultaneously measure precisely the position AND the velocity of a particle.

Recently there are also surges of quantum stuff in other fields. Things such as quantum computing are ALL essentially done by dropping commutativity assumption (i.e., use matrices instead of numbers). For example, a recent field in cognitive science is quantum inference. Suppose we are the judge, and we are going to decide if a suspect is guilty or not. If we first see evidence A, then see evidence B, then we may have some idea. But if we first see evidence B, then see evidence A, then we may have a different idea. This does NOT go well with traditional probability, since Pr(Guilty—A and B) is the same as Pr(Guilty—B and A). We need some non-commutative model to handle this.

Let us say we are going to genuinly invent quantum speed reading. What would we do? It must be some non-commutative (and thus non-linear) form of reading. Maybe we look at one word from each line, and then look at a different word from each line, and repeat this several times for a single page, and then try to infer the meaning of the whole page? Maybe if we are proficient enough, hopefully this might yeild a faster way of reading things (but with non-zero chance of misunderstanding the content...).

So in this section, we aim to explore some commutative and non-commutative behaviors.

10.5.1 Totally dependent commutativity

What commutes with A? Well, A commutes with A. In fact, all powers of A commutes with A. Furthermore, all polynomials of A commutes with A. Finally, all functions of A (which are essentially polynomials of A if we fix A) must commutes with A. So here is a super easy result:

Proposition 10.5.2. *For any functions f, g, suppose $f(A), g(A)$ are both defined, then $f(A)g(A) = g(A)f(A)$.*

Here comes a question: Are these all? In general, then answer is no.

Example 10.5.3. The identity matrix commute with ALL matrices. But are all matrices functions of the identity matrix? Obviously no. For any function f, we have $f(I) = f(1)I$, always a multiple of identity. ☺

Luckily, there are cases where all matrices that commutes with A are functions of A. For example, if A has distinct eigenvalues, then $AB = BA$ implies that A, B are simultaneously diagonalizable. WLOG suppose they are both diagonal. Say A has eigenvalues $a_1, \ldots, a_n$ while B has eigenvalues $b_1, \ldots, b_n$. Since $a_1, \ldots, a_n$ are all distinct, we can simply find any function f such that $f(a_i) = b_i$ for all i, then $f(A) = B$.

More generally, we have the following.

Proposition 10.5.4. *Suppose A has a single Jordan block for each eigenvalue. (I.e., all geometric multiplicity are one.) The $AB = BA$ implies that $B = p(A)$ for some polynomial p.*

Proof. Suppose A is a single nilpotent Jordan block. Then $AB = BA$ means entries of B shifted up and entries of B shifted right shall have the same results. Use this and you can show that we must have $B = \begin{bmatrix} a_0 & a_1 \ldots & a_{n-1} & \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_1 \\ & & & a_0 \end{bmatrix} = a_0 I + a_1 A + \cdots + a_{n-1} A^{n-1}$. We are good.

Now suppose A is a single λ Jordan block. Then $A = \lambda I + N$ for a nilpotent Jordan block N. So $AB = BA$ implies that $(\lambda I + N)B = B(\lambda I + N)$, and simplification gives $NB = BN$. So $B = p(N)$ for some polynomial p. Let $q(x) = p(x - \lambda)$, then $B = p(A - \lambda I) = q(A)$.

Now consider the generic case. By changing basis, I assume that A is in Jordan canonical form, say $A = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}$. Now we write B into a block matrix in the same manner, and let the (i,j)-block be B_{ij}. Then $AB = BA$ implies that $A_i B_{ij} = B_{ij} A_j$. But by our assumption, A_i, A_j has no common eigenvalues! Hence the only solution to the Sylvester's equation $A_i X - X A_j = 0$ is zero. So B is block diagonal as well.

So $B = \begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_k \end{bmatrix}$, and we have $A_i B_i = B_i A_i$. Since each A_i is a single Jordan block, we see that $B_i = p_i(A_i)$ for some polynomial p_i. So our goal is now the following: we want to find a polynomial $p(x)$ such that $p(A_i) = p_i(A_i)$ for all i.

So we want to find $p(x)$ such that $p \equiv p_i$ modulus the killing polynomial of A_i. We are done by the lemma below. □

Lemma 10.5.5. *Given polynomials $q_1, \dots, q_k$ and coprime polynomials $p_1, \dots, p_k$, there is a polynomial p such that $p \equiv p_i \pmod{q_i}$ for all i. (We can in fact require this polynomial to have degree less than $\sum \deg(p_i)$, and in this case such p is unique.)*

Proof. Chinese Remainder Theorem (Sun Zi Ding Li).

Alternatively, say q_i has degree d_i, and let $d = \sum d_i$. Consider the space V of all polynomials of degree less than d, and let V_i be the space of all polynomials of degree less than d_i.

Now for each i, we have a map $Q_i : V \to V_i$, such that $Q_i(p)$ is the remainder of p divided by p_i. You can check that Q_i is linear. So we have a linear map $Q : V \to \prod V_i$. (Here $\prod V_i$ is the space of $(p_1, \dots, p_k)$ where each $p_i \in V_i$.) It is enough to show that Q is surjective. Note that $\dim(V) = d = \sum d_i = \sum \dim V_i = \dim \prod V_i$, so it is enough to show that the map is injective.

Finally, if $Q(p) = 0$, then p_i divides p for all i. So $\prod p_i(x)$ divides $p(x)$. But since $p \in V$, it has degree less than d, while $\prod p_i(x)$ has degree exactly $\sum d_i = d$. Hence we can only have $p = 0$. So Q has trivial kernel and is injective (hence bijective). □

Example 10.5.6. The condition here that A had all geometric multiplicity one is the best possible.

Suppose A has geometric multiplicity larger than one for some eigenvalue λ. By usual simplification method, we only need to consider the case where A is made of two nilpotent blocks. Say $A = \begin{bmatrix} N_1 & \\ & N_2 \end{bmatrix}$ and say N_1, N_2 are $m \times m$ and $n \times n$ and $m \leq n$. Let X be an $m \times n$ matrix such that $X = \begin{bmatrix} 0 & N \end{bmatrix}$ where 0 is $m \times (n - m)$ and N is the $m \times m$ nilpotent Jordan block. Then $N_1 X = X N_2$, so we have a non-trivial solution to the Sylvester's equation. So $\begin{bmatrix} I & X \\ & I \end{bmatrix} A = A \begin{bmatrix} I & X \\ & I \end{bmatrix}$, yet any function of A must remain block diagonal. ☺

10.5.2 Totally INdependent commutativity

There is an alterntive case where things always commute. If A, B acts on completely different things, and do not interfere with each other, then we should have $AB = BA$.

Example 10.5.7. $\begin{bmatrix} A & \\ & I \end{bmatrix}, \begin{bmatrix} I & \\ & B \end{bmatrix}$ always commute, because they act on independent subspaces and they do not interfere with each other.

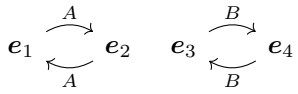
For any distinct i, j, k, l, consider the elementary matrix E_{ij}, E_{kl}, then they commute. Because as row operations, they do their things independently and do not touch each other.

Note that in both cases, neither matrix is a function of the other. Consider $A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{bmatrix}$ and $B = \begin{bmatrix} & 1 & & \\ 1 & & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$. Then any $f(B)$ will have the lower right block diagonal, so it is never A, while any $f(A)$ will have the upper left block diagonal, so it is never B.

However, it is possible to reconcile the totally independent case with the totally dependent case. We can in fact find C such that $A = f(C), B = g(C)$ for some polynomials f, g. Can you find them?

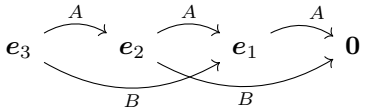
Also, for another challenge, can you find a matrix C such that $f(C)$ is the elementary matrix E_{12} and $g(C)$ is the elementary matrix E_{34}? ☺

Graphically, one can see that if $A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{bmatrix}$ and $B = \begin{bmatrix} & 1 & & \\ 1 & & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$, then their action lookes like this:



You can sort of see why they commute.

For a totally dependent case, say if $A = \begin{bmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{bmatrix}$ and $B = A^2$, then their action is like the following:



You can see that they are the same flow on the same killing chain, except that B is a faster version of A.

Both cases are ultimately described by the fact that we can find C, and $A = f(C), B = g(C)$ for some function f, g. You may think of this phenomena as such: at some places they do not meet at all, and at those places where they meet each other, they shall essentially be different versions of the same thing.

Unfortunately, NOT all commutativities are like this. Read on.

10.5.3 Entangled commutativity and non-commutativity

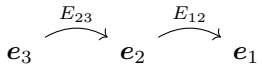
Totally dependent and independent things commute. But things will be bad if they are "entangled", a status between dependent and independent.

Example 10.5.8 (Entangled things might not commute)**.** Consider the 3×3 elementary matrices E_{12}, E_{23}. The following to operations are different:

1. Add second row to first row, then add third row to second row.
2. Add third row to second row, then add second row to first row.

The difference here is that, in the first one $E_{23}E_{12}$, the original third row did NOT contribute to the first row. While in the second one $E_{12}E_{23}$, the original third row DID end up contributing to the first row. This is also evident in the calculation $E_{12}E_{23} - E_{23}E_{12} = \begin{bmatrix} 1 & 1 & 1 \\ & 1 & 1 \\ & & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 \\ & 1 & 1 \\ & & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ & 0 & 0 \\ & & 0 \end{bmatrix}$. So you see that the difference is exactly this: whether the third row made it to the first or not.

Graphically, it looks like the following. Note how they entangle at $\boldsymbol{e}_2$. The order of multiplication determines whether the two arrow "connect" or "disconnect" at $\boldsymbol{e}_2$. (These graphs are NOT rigorous. They are just what I do to make things clear to myself.)



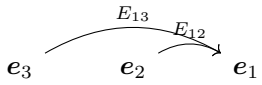
The "passing of the baton" thing is usually a bad sign that things are not going to commute. ☺

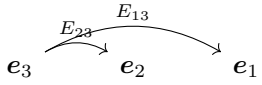
So, are all entangled things bad? Not always. Here is an entanglement that is actually quite nice.

Example 10.5.9 (Parallel things commute)**.** Consider the 3×3 elementary matrices E_{12}, E_{13}. They commute, because both $E_{12}E_{13}$ and $E_{13}E_{12}$ says the same thing: add the bottom two rows to the first row.

Similarly, $E_{13}E_{23} = E_{23}E_{13}$, since they are both saying the same thing: add the third row to the top two rows.

Graphically they look like this:





Note that in these cases, you would NOT be able to find C such that $E_{12} = f(C)$ and $E_{13} = g(C)$. ☺

Proposition 10.5.10. *There is no matrix C such that $E_{12} = f(C)$ and $E_{13} = g(C)$.*

Proof. Suppose for contradiction that there is such a matrix C. Then if $\boldsymbol{v}$ is an eigenvector for C, it must also be an eigenvector for $f(C)$ and for $g(C)$.

However, the only common eigenvectors of $E_{12}E_{13}$ are multiples of $\boldsymbol{e}_1$. So any eigenvector of C must be a multiple of $\boldsymbol{e}_1$. In particular, C in its Jordan form has a single Jordan block.

Suppose $C = XJX^{-1} = X \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} X^{-1}$. Now note that $E_{12} - I$ has rank one. So $X(f(J) - I)X^{-1}$ has rank one, and thus $f(J) - I$ has rank one. But $f(J) - I$ must looke like $\begin{bmatrix} a & b & c \\ & a & b \\ & & a \end{bmatrix}$, so the only rank one possibility is $\begin{bmatrix} 0 & 0 & c \\ & 0 & 0 \\ & & 0 \end{bmatrix}$.

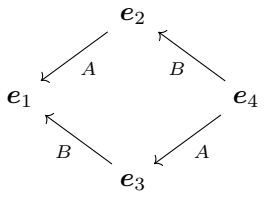
Similarly, $E_{13} - I$ also has rank one. So by identical logic, $g(J) - I$ is also $\begin{bmatrix} 0 & 0 & d \\ & 0 & 0 \\ & & 0 \end{bmatrix}$. But this means $\mathrm{Ker}(f(J) - I) = \mathrm{Ker}(g(J) - I)$, which, after change of basis, implies that $\mathrm{Ker}(E_{12} - I) = \mathrm{Ker}(E_{13} - I)$, which is false. □

Now, parallel things are not the only non-functional commuting behavior. Here is another, where the entanglement "balanced out"

Example 10.5.11 (Balancing Entanglement)**.** Consider $A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. You can check that they commute. However, they are non-parallel and there is no polynomial relation.

(Same proof essentially. Suppose there is a C. Then for the same reason, all eignevalues of C must be multiples of v_1, so under some basis C is just a single Jordan block. Then $A-I, B-I$ are rank 2 and nilpotent, so under the new basis they look like $\begin{bmatrix} 0 & 0 & a & b \\ 0 & 0 & 0 & a \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. Then they have the same range, contradiction.)

Graphically, they look like this:



As you can see, there are two "passing the baton" phenomena, but they balanced out. No matter AB or BA, $\boldsymbol{e}_4$ is carried over to $\boldsymbol{e}_1$ exactly once.

This is also the case of $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ on the space of functions spanned by $1, x, y, xy$. ☺

Example 10.5.12. Let us also try to understand the above phenomena from yet another perspective, by looking at a related phenomena. Consider the nilpotent version $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, and then we symmetrize B so that $B_s = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$.

Then note that $B_s A B_s^{-1} = A$, and B_s essentially permute the two Jordan blocks of A. Yet $B_s A B_s^{-1} = A$ is equivalent to $B_s A = A B_s$. So we see that they commute. ☺

Now, let us give one last explanation of the commutativity here. Read on.

10.5.4 Kronecker tensor product

Definition 10.5.13. *Given two (not necessarily square) matrix A, B, let a_{ij} be the (i, j) entry of A. Then we define their Kronecker tensor product $A \otimes B$ to be the matrix whose (i, j) block is $a_{ij}B$.*

Proposition 10.5.14. *The Kronecker tensor product is bilinear, i.e., $(kA) \otimes B = k(A \otimes B) = A \otimes (kB)$, and we also have $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B$, and $A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$.*

Proof. Straightforward verification by definition. □

Example 10.5.15. Note that in our last example, $A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I \otimes X$ where I is the 2 by 2 identity matrix and $X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. We also see that $B = X \otimes I$. ☺

So from another perspective, the commutativity here can also be explained as the following:

Proposition 10.5.16. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. *(So, if A, C commute and B, D commute, then $A \otimes B, C \otimes D$ commute.)*

To prove this, let us first try to figure out what exactly is $A \otimes B$ tries to do. Note that there is no size requirement for A and B. If A is $m_A \times n_A$ and B is $m_B \times n_B$, then $A \otimes B$ is $(m_A m_B) \times (n_A n_B)$.

In particular, for two vectors $\boldsymbol{v} \in \mathbb{C}^m$ and $\boldsymbol{w} \in \mathbb{C}^n$, then $\boldsymbol{v} \otimes \boldsymbol{w}$ is $1 \times (mn)$, hence it is a vector in $\mathbb{C}^{mn}$.

Example 10.5.17. Suppose $m = n = 3$. Then $\boldsymbol{e}_1 \otimes \boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, and $\boldsymbol{e}_1 \otimes \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, while $\boldsymbol{e}_2 \otimes \boldsymbol{e}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

In general, $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$ is the $(3i - 3 + j)$-th standard basis vector of $\mathbb{C}^9$.

Note that NOT all vectors in $\mathbb{C}^{mn}$ are of the form $\boldsymbol{v} \otimes \boldsymbol{w}$. For example, $\boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2$ will have no such representation. Note that $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$ form a basis. You may think of $\mathbb{C}^{mn}$ with this basis as "the space of matrices", and for each element $\sum a_{ij} \boldsymbol{e}_i \otimes \boldsymbol{e}_j$, you can make a matrix A whose (i, j) entry is a_{ij}. Then as far as addition and scalar multiplication go, it works as expected. Now $\boldsymbol{v} \otimes \boldsymbol{w}$ corresponds to the matrix $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$ for any $\boldsymbol{v}, \boldsymbol{w}$, so its matrix has rank one, while $\boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2$ corresponds to a matrix of rank two, so it is never $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$ for any $\boldsymbol{v}, \boldsymbol{w}$. ☺

Lemma 10.5.18. $(\boldsymbol{v} \otimes B)\boldsymbol{w} = \boldsymbol{v} \otimes (B\boldsymbol{w})$.

Proof. $(\boldsymbol{v} \otimes B)\boldsymbol{w} = \begin{bmatrix} v_1 B \\ \vdots \\ v_m B \end{bmatrix} \boldsymbol{w} = \begin{bmatrix} v_1 B\boldsymbol{w} \\ \vdots \\ v_m B\boldsymbol{w} \end{bmatrix} = \boldsymbol{v} \otimes (B\boldsymbol{w})$. □

Proposition 10.5.19. $(A \otimes B)(\boldsymbol{v} \otimes \boldsymbol{w}) = (A\boldsymbol{v}) \otimes (B\boldsymbol{w})$.

Proof. Mostly by direct computation, which you can DIY. Here is a slightly (and hopefully) less boring presentation of the calculations.

Note that we have the identity $\begin{bmatrix} A_1 & A_2 \end{bmatrix} \otimes B = \begin{bmatrix} A_1 \otimes B & A_2 \otimes B \end{bmatrix}$ just by definition of this block matrix. So if $A\boldsymbol{e}_i = \boldsymbol{a}_i$, i.e., $A = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_m \end{bmatrix}$, then $A \otimes B = \begin{bmatrix} \boldsymbol{a}_1 \otimes B & \dots & \boldsymbol{a}_m \otimes B \end{bmatrix}$.

Now $(A \otimes B)(\boldsymbol{v} \otimes \boldsymbol{w}) = \begin{bmatrix} \boldsymbol{a}_1 \otimes B & \dots & \boldsymbol{a}_m \otimes B \end{bmatrix} \begin{bmatrix} v_1 \boldsymbol{w} \\ \vdots \\ v_m \boldsymbol{w} \end{bmatrix} = \sum (\boldsymbol{a}_i \otimes B)(v_i \boldsymbol{w}) = \sum (\boldsymbol{a}_i) \otimes (v_i B\boldsymbol{w}) = (\sum v_i \boldsymbol{a}_i) \otimes (B\boldsymbol{w}) = (A\boldsymbol{v}) \otimes (B\boldsymbol{w})$. □

Note that, if we were to think of $\boldsymbol{v} \otimes \boldsymbol{w}$ as the matrix $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$, then $(A \otimes B)$ acts by mutliplying A on the left, and multiply B^{T} on the right. In particylar, the matrix for $(A \otimes B)(\boldsymbol{v} \otimes \boldsymbol{w})$ would be $A\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}B^{\mathrm{T}}$.

Corollary 10.5.20. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Proof. For any basis vector $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$, then $((AC) \otimes (BD))(\boldsymbol{e}_i \otimes \boldsymbol{e}_j) = (AC\boldsymbol{e}_i) \otimes (BD\boldsymbol{e}_j)$, while $(A \otimes B)(C \otimes D)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j) = (A \otimes B)((C\boldsymbol{e}_i) \otimes (D \otimes \boldsymbol{e}_j)) = (AC\boldsymbol{e}_i) \otimes (BD\boldsymbol{e}_j)$. So the two agree on a basis. They must be the same map.

Alternatively, we can also prove this using the matrix interpretation of the input $\sum x_{ij} \boldsymbol{e}_i \otimes \boldsymbol{e}_j$. Let this corresponds to the matrix X. Then $(A \otimes B)(C \otimes D)$ sends this to the matrix $A(CXD^{\mathrm{T}})B^{\mathrm{T}}$, while $(AC) \otimes (BD)$ sends this to the matrix $(AC)X(BD)^{\mathrm{T}}$. You can see that the two resulting image are the same. □

So, by picking commuting A, C and commuting B, D, we can create commuting $A \otimes B, C \otimes D$, which might look surprising before you realize the tensor structure.

But are all commuting matrices like this? The answer is still no. Here is an example of commuting matrices that cannot be explained by anything we've done.

Example 10.5.21. Consider $A = \left[\begin{array}{ccc|cc} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & & \\ \hline & & & 0 & 1 \\ & & & & 0 \end{array}\right]$. Any matrix that commute with it must have the following form $\left[\begin{array}{ccc|cc} a & b & c & d & e \\ & a & b & & d \\ & & a & & \\ \hline & f & g & h & i \\ & & f & & h \end{array}\right]$. (Can you prove this?)

Now let $B = \left[\begin{array}{ccc|cc} 0 & 1 & 1 & 1 & 1 \\ & 0 & 1 & & 1 \\ & & 0 & & \\ \hline & & & 0 & 1 \\ & & & & 0 \end{array}\right]$. Then we have $AB = BA$. Furthermore, they are not functions of some common matrix.

(Usual proof. A, B has only multiples of $\boldsymbol{e}_1$ as common eigenvectors, so any such C must have only one Jordan block. Then since both A, B has rank three, $f(C), g(C)$ must also have rank 3. By change of basis to make C into canonical form, we see that $f(C)^2, g(C)^2$ have the same kernel. Which is not the case, as $\mathrm{Ker}(A^2) \neq \mathrm{Ker}(B^2)$.)

Furthermore, there can be no tensor decomposition, since both matrices are 5×5 and 5 is a prime number. And finally, some arrows in the graph would disagree with others.

So while they commute, the situation does not fall into any categories we have discussed about. ☺

10.5.5 Simultaneously nice

If $AB = BA$ and one of them has distinct eigenvalues, then they can be simultaneously diagonalized. This is true. And in general, if $AB = BA$, then they can be simultaneously triangularized. This is HW.

However, they might not be simultaneously Jordanized. In fact, it might not even be possible to put one in Jordan canonical form and put another in upper triangular form.

Here let us see some examples, with varying degree of niceness.

Example 10.5.22. Say $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Then $AB = BA = 0$ and in fact they have parallel behaviors. (Just consider the corresponding row operations of $I + A$ and $I + B$.)

Also note that both matrices are nilpotent. They also both have the same Jordan normal form $J_1 = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$, or alternatively $J_2 = \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right]$. They also have the same kernel. Their range are both 1-dimensional.

Now, under whatever basis, they must not be equal. So if they are simultaneously Jordanized, then one must be J_1 while the other must be J_2. But then $J_1 J_2 \neq J_2 J_1$, which contradict the fact that $AB = BA$. So they CANNOT be simultaneously Jordanized.

Pick $\boldsymbol{v}_1$ that span $\mathrm{Ran}(A)$ and pick $\boldsymbol{v}_3$ such that $A\boldsymbol{v}_3 = \boldsymbol{v}_1$. Since $A\boldsymbol{v}_3 \neq 0$,and $\mathrm{Ker}(A) = \mathrm{Ker}(B)$, it follows that $\boldsymbol{v}_2 := B\boldsymbol{v}_3$ is non-zero. And under the basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, we would turn A into uppoer triangular $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and B into its Jordan canonical form J_2. ☺

Example 10.5.23. Consider $C = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, and $A = I \otimes C$ and $B = S \otimes C$, where $S = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}$. Obviously $AB = BA$, and A is already in Jordan normal form.

Now $B^2 = 0$, so all eigenvalues of B are zero. B has a two dimensional kernel, so it has two Jordan blocks. Finally, $B^2 = 0$ means each block is 2 by 2. Hence B in fact has the same Jordan normal form as A.

Suppose we want to change basis simultaneously by some matrix T, such that A is still in JNF but B is upper triangular. If afterwards A is still in Jordan normal form, then $A = TAT^{-1}$. In particular, $TA = AT$.

To make the latter process rigorous, let P be the matrix that swaps the second and third row as a row operation. Then $P(X \otimes Y)P^{-1} = Y \otimes X$ always, as you can verify. Then $PAP^{-1} = C \otimes I = \begin{bmatrix} 0 & I \\ & 0 \end{bmatrix}$, and PTP^{-1} must commute with PAP^{-1}. Hence $PTP^{-1} = \begin{bmatrix} X & Y \\ & X \end{bmatrix}$ for some X, Y. Its inverse is $\begin{bmatrix} X^{-1} & -X^{-1}YX^{-1} \\ & X^{-1} \end{bmatrix}$.

Now $PBP^{-1} = C \otimes S = \begin{bmatrix} 0 & S \\ & 0 \end{bmatrix}$. So $P(TBT^{-1})P^{-1} = (PTP^{-1})(PBP^{-1})(PTP^{-1})^{-1} = \begin{bmatrix} X & Y \\ & X \end{bmatrix} \begin{bmatrix} 0 & S \\ & 0 \end{bmatrix} \begin{bmatrix} X^{-1} & -X^{-1}Y \\ & X^{-1} \end{bmatrix} = \begin{bmatrix} 0 & XSX^{-1} \\ & 0 \end{bmatrix}$. Say $XSX^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then we see that $TBT^{-1} = P^{-1}(C \otimes (XSX^{-1}))P = XSX^{-1} \otimes C = \begin{bmatrix} 0 & a & 0 & b \\ & 0 & & 0 \\ 0 & c & 0 & d \\ & 0 & & 0 \end{bmatrix}$. This is triangular if and only if $c = 0$.

So pick any X that upper triangularize S, say $X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Then $P(TBT^{-1})P^{-1} = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$.

As a result, we have $TBT^{-1} = P^{-1} \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] P = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right]$.

So as you can see, you can pick any $T = P \begin{bmatrix} X & Y \\ 0 & X \end{bmatrix} P^{-1}$ with any X upper-triangularizing S and take any Y. These are all possible choices of basis T. ☺

Now consider the next example, where we can do this, but there is no control over WHICH matrix get put into Jordan normal form.

Example 10.5.24. Consider $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ and $B = A^2$. We are already good since A is in Jordan normal form and B is upper triangular.

However, there is no way to put B in Jordan normal form while keeping A upper triangular. Suppose TAT^{-1} is upper triangular. Then it is still nilpotent, and its rank is still 2. So $TAT^{-1} = \begin{bmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{bmatrix}$ for some unknown $a, b, c \in \mathbb{C}$. Then $TBT^{-1} = (TAT^{-1})^2 = \begin{bmatrix} 0 & 0 & ac \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. So if A is upper triangular, B may never be in its Jordan normal form. ☺

Finally, consider the following example where this is not possible at all. Whenever one is in JNF, the other cannot be upper triangular.

Example 10.5.25. Let $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ and $D = C^2$ be as in the last example. Let $A = diag(C, D)$ and $B = diag(D, C)$. The intuition is that, in the first block, you cannot Jordanize the D portion withouth

ruining the upper triangular structure of C, but in the second block, the situation is reversed. So it turned out that neither can be Jordanized without ruining the other.

Of course, who knows if there is some super weird change of basis that end up achieving the desired result? To rigorously prove the impossibility, first note that $AB = BA = 0$. Suppose, for contradiction, that we find a basis A^2u, Au, u, Av, v, w, such that A becomes its Jordan normal form $\left[\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$ and B is upper triangular. Note that since $BA = 0$, we see that $B(A^2u) = B(Au) = B(Av) = 0$. So the first, second and forth columns of B must be all 0. Since $AB = 0$, by looking at A, the second, third and fifth rows of B are also all 0. Since B is nilpotent and upper triangular, its entries on the diagonal and below the diagonal are all 0. So B has the form $\begin{bmatrix} 0 & 0 & * & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$. Then you can check that $B^2 = 0$. However, this is not true. In the original basis, $B^2 = diag(0, C^2) = diag(0, D) \neq 0$. Contradiction ☺

Part VI

Multilinear Algebra

Chapter 11

Dual Space

11.1 The Dual Phenomena

Before any formal exploration of tensors, it is important to realize that there are two kinds of vectors. In some informal sense, I guess we can call them "column vectors" and "row vectors". But more formally, we call them "vectors" and "dual vectors".

Example 11.1.1. Consider the following example. We go to a McDonald store and buy food. I can buy burgers, wings and cokes. Then my orders are linear combinations of these things, i.e., vectors. Say if I buy a burgers, b wings and c cokes, I can simply say that I am buying $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. So I am working on a vector space $V = \mathbb{R}^3$, where each vector represent a potential order I could make.

Now, after I made my order, I need to pay. The process of "paying" is like a map $\alpha : V \to \mathbb{R}$. Further more, this map is linear. The total cost of ordering $\boldsymbol{v} + \boldsymbol{w}$ is exactly $\alpha(\boldsymbol{v} + \boldsymbol{w}) = \alpha(\boldsymbol{v}) + \alpha(\boldsymbol{w})$ and so on. Yeah, this is a real life linear algebra phenominon. Don't let any tell you that linear algebra is not related to everyday life.

So what is this α? Well, it is a linear map from $\mathbb{R}^3$ to $\mathbb{R}$, so it is a 1×3 matrix, i.e., a row vector $\alpha = \begin{bmatrix} p & q & r \end{bmatrix}$. If you think about it, the three coordnates have well-defined meanings: p, q, r are the prices of a single burger, a single wing and a singel coke. So if we order $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ burgers, wings and cokes, the total cost is $\begin{bmatrix} p & q & r \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = ap + bq + cr$.

What if we go to a different fast food store? Then they might have a different prices for burgers, wings and cokes, so it will have a different row vector. Now, let V^* be the space of all 1×3 row vectors, i.e., the space of potential prices. Given any order $\boldsymbol{v} \in V$ and any pricing $\alpha \in V^*$, the total cost would be $\alpha(\boldsymbol{v})$, which is the multiplication of a row vector to a column vector.

Now, suppose McDonald gives us options to buy combos! Say Combo A contains 2 burgers plus one coke, and Combo B contains 1 burger, 2 wings and 2 cokes. Then if I purchase $\begin{bmatrix} x \\ y \end{bmatrix}$ Combo A's and Combo B's, then it contains a total of $\begin{bmatrix} 2 & 1 \\ 0 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ burgers, wings and cokes. As you can see, we have a linear map L, called "counting the ingredients", that goes from the combo space $W = \mathbb{R}^2$ to the food space $V = \mathbb{R}^3$.

Now let us look at the pricing of combos. If the food price are $\begin{bmatrix} p \\ q \\ r \end{bmatrix}$ (we write them vertically now for convenience), then the combo prices for combo A and combo B would make a vector $\begin{bmatrix} 2 & 0 & 1 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix}$. (I am assuming that there is no discount.) As you can see, we have a linear map L^* that goes from the food price space V^* to the combo price space W^*.

I would like to draw your attention to this phenomena, which is the key to the understanding of dual spaces:

1. The food space V and the food price space V^* evaluate each other. Given a food vector, you can evaluate the total cost using a pricing (row) vector. But it also goes the other way: given a pricing (row) vector, you can use a food vector to get a total cost.

2. Similarly, the combo space W and the combo price space W^* have the same relation.

3. If we are putting foods into combos, it actually gives a map L from the combo space to the food space. It goes in the counter-intuitive direction. However, in the price spaces, things would go in the intuitive direction. Putting foods into combos induce a linear map L^* from the food price space to the combo price space.

4. Finally, not only the two maps L, L^* goes in the opposite direction, in fact they are transposes of each other!

☺

So the key question is this: what is the meaning of transpose? The above example hopefully gives you some idea about this. We now start the formal process of building these things.

Definition 11.1.2. *Given a vector space V, its dual space V^* is the space of all linear maps from V to $\mathbb{R}$ (or to $\mathbb{C}$ if we were doing complex vector spaces).*

People call elements of V^* many things. Some popular choices are "dual vectors" and "linear functionals".

Intuitively, a dual vector is something used to evaluate vectors. And usually, despite its abstract construction, dual vectors shall turn out to be more intuitive than vectors. In fact, we all actually udnerstand dual vectors way before we understand vectors. Let us see some examples.

Example 11.1.3. Consider $\mathbb{R}^3$. Given a vector $\boldsymbol{v} \in \mathbb{R}^3$, we sometimes say we want to take its x-coordinates. But what is "taking the x-coordinate"? It is in fact a linear map $x : \mathbb{R}^3 \to \mathbb{R}$. As you can see, taking coordinates are dual vectors.

When we first encouter vectors in high school, we usually start with coordinates. Why? Because dual vectors are the only way for us to understant these vectors. If I just say "we have a generic vector $\boldsymbol{v}$", then you might feel that it is a bit abstract. But if I say "look at the vector $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$", now you feel a little better. By using three dual vectors to "evaluate" and "locate" a vector, we now feel a little more comfortable. ☺

Example 11.1.4. This is an informal example. Let X be the space of all students. Then what should a "dual student" be? It should be an evaluation of students. How can we evaluate students? Well, through exams of course. So the dual student space X^* should be the space of all exams. ☺

Here are several rather important examples.

Example 11.1.5. Let V be the space of all real functions. Then for any real number $a \in \mathbb{R}$, we can use this to evaluate functions at a, i.e., we send function f to $f(a)$.

Let us write this map as $\mathrm{ev}_a : V \to \mathbb{R}$. Then $\mathrm{ev}_a(f+g) = (f+g)(a) = f(a) + g(a) = \mathrm{ev}_a(f) + \mathrm{ev}_a(g)$, and $\mathrm{ev}_a(kf) = (kf)(a) = k\mathrm{ev}_a(f)$. Well, this is linear! (Be careful here. $a \mapsto f(a)$ is not linear unless f is linear itself. But $f \mapsto f(a)$ is always linear, even if f is non-linear.)

So for each $a \in \mathbb{R}$, the corresponding evaluation map ev_a is a dual vector (or dual functions, or "functionals") in V^*. These are "local" evaluations of functions, since we only care about the value of the function at a single point. Similarly, we have other local evaluations. For example, the map $D|_a : f \mapsto f'(a)$ is also a dual vector, and it is also a local evaluation. ☺

Example 11.1.6. Let V be the space of all real integrable functions. What is integration?

Given any $a, b \in \mathbb{R}$, consider the map $\int_a^b : V \to \mathbb{R}$. Well, it is easy to verify that this is linear! So the definite integral $\int_a^b$ is an element of V^*. This is a more "global" evaluation, in the sense that it ignores local information. Indeed, if we change the value of f at a single point, then it shall have NO effect in these integral evaluations. ☺

Remark 11.1.7. *(This remark is optional.) What is an indefinite integral?*

If we were to study derivatives from the prespectives of linear algebra, we usually just think of it as a linear map, sending functions to functions. However, it is a bad idea to do so for integrations.

For example, the indefinite integral $\int x\,\mathrm{d}x$ is NOT a function. It is $\frac{1}{2}x^2 + C$ for some undetermined constant C, and this undetermined constant means the result is NOT a well-defined element of the function space!

The source of trouble is this: given a function f, we need only one input a to evaluate its derivative into a real number $f'(a)$. But we need two inputs a, b to evaluate its anti-derivative into a real number $\int_a^b f(x)\,\mathrm{d}x$. Furthermore, it might be better to NOT think of a, b as two numbers, but rather think of it as the closed interval $[a, b]$, a subset of $\mathbb{R}$. For multivariable calculus, integration would be done as $\int_S f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ where S is some subset of $\mathbb{R}^3$.

In particular, an indefinite integral $\int$ is a pairing. Given a domain $S \subseteq \mathbb{R}^n$ and a function f, it shall send them to a number $\int_S f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. With more advanced tools from algebraic topology, we can in fact make the "space of domains" into a vector space, then $\int$ is in fact a bilinear map (the evaluation process), and "domains" and "functions" are duals to each other.

Example 11.1.8. So far, we have seen two kinds of evaluations of functions, a local one and a global one. They have vastly different behaviors and they measure very different aspects of a function. How can they be so different? Is there a way to think of both in a single perspective?

There indeed is one. Let X be a random real number for some probability distribution. Then for each function $f \in V$, $f(X)$ is also a random real number for some probability distribution. Then one can look at the expected value (i.e., "average value") $\mathbb{E}(f(X))$ as a linear evaluation of f. Let us call this ev_X.

If X is a random number with 100% chance to have value $a \in \mathbb{R}$, then $\mathbb{E}(f(X)) = f(a)$. So ev_X is exactly the local evaluation ev_a. On the other hand, let X be a random number uniformly distributed in the closed interval $[a, b]$. (Uniformly distributed means each number happen with the same probability.) Then $\mathbb{E}(f(X)) = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x$. So ev_X is a global evaluation.

In this sense, we can have the following funny interpretation of a random real variable: it is simply an element of V^*. Some subjects these days require "non-classical" probability theory, like quantum computations and such. And thinking of random variables as elements in a dual space is a very important idea to have. ☺

Here is one last example, and it is pretty imporant. Say the dual of V is V^*. What is the dual of V^*?

Example 11.1.9. Given $\alpha \in V^*$, then it is a linear map from V to $\mathbb{R}$. In particular, given any $\boldsymbol{v} \in V$, we can form an evaluation map $\mathrm{ev}_{\boldsymbol{v}}$ that sends each α to $\alpha(\boldsymbol{v})$. In this sense, each $\boldsymbol{v}$ corresponds to an element of $(V^*)^*$. Wow!

To be more elaborate, we now think of the evaluation process ev as a two-input function, $\mathrm{ev}_{(-)}(-)$. By put in some $\boldsymbol{v} \in V$ and some $\alpha \in V^*$, we have a real number $\mathrm{ev}_{\boldsymbol{v}}(\alpha) = \alpha(\boldsymbol{v})$. If we only put in a dual vector α but leave the vector slot open, then we have $\mathrm{ev}_{(-)}(\alpha)$. It is waiting to eat a vector and then spit

out a number, i.e., it is a map from V to $\mathbb{R}$, and in fact it is easy to see that it is exactly α itself (because $\mathrm{ev}_{\boldsymbol{v}}(\alpha) = \alpha(\boldsymbol{v})$).

But if we only put in a vector $\boldsymbol{v}$ but leave the dual vector slot open, then we are left with $\mathrm{ev}_{\boldsymbol{v}}(-)$, and it is waiting to eat a dual vector and spit out a number. I.e., it is a map from V^* to $\mathbb{R}$. So this is an element of $(V^*)^*$. We simply write $\mathrm{ev}_{\boldsymbol{v}}$ for $\mathrm{ev}_{\boldsymbol{v}}(-)$.

So, $\boldsymbol{v}$ is an element of V while $\mathrm{ev}_{\boldsymbol{v}}$ is an element of $(V^*)^*$. So we in fact have a map $\mathrm{ev} : V \to (V^*)^*$ that sends $\boldsymbol{v}$ to $\mathrm{ev}_{\boldsymbol{v}}$.

So, many elements of $(V^*)^*$ are in correspondence of elements of V as some local evaluation. Are these all? ☺

Lemma 11.1.10. *If* $\dim V = n$*, then* $\dim V^* = n$*.*

Proof. The cheap way is to pick a basis, and pretent V is $\mathbb{R}^n$. Then it is the space of $n \times 1$ column vectors. Then the space V^* is the space of linear maps from $\mathbb{R}^n$ to $\mathbb{R}$, so it is the space of $1 \times n$ row vectors, and immediately $\dim V^* = n$. □

Proposition 11.1.11. *If V is finite dimensional, then* $\mathrm{ev} : V \to (V^*)^*$ *is an isomorphism of vector spaces (i.e., it is a linear bijection).*

Proof. If V is n dimensional, then its dual V^* is n dimensional. But applying this logic again, if V^* is n dimensional, then its dual is also n dimensional. So the domain and codomain of ev have the same dimension. So to show that it is a linear bijection, it is enough to show that it is a linear injection.

The verification that it is linear is routine so we skip it here (but do this yourself). Now suppose $\boldsymbol{v} \in \mathrm{Ker}(\mathrm{ev})$. Then $\mathrm{ev}_{\boldsymbol{v}}$ is the zero map from V^* to $\mathbb{R}$. This means $\mathrm{ev}_{\boldsymbol{v}}(\alpha) = 0$ for all α, i.e., $\alpha(\boldsymbol{v}) = 0$ for all α.

By picking basis for V, we may assume that $V = \mathbb{R}^n$. Recall that each coordinate is a dual vector in V^*. So $\alpha(\boldsymbol{v}) = 0$ for all α implies that all coordinates of $\boldsymbol{v}$ are zero, so $\boldsymbol{v} = \boldsymbol{0}$. □

To paint a complete picture, if $V = \mathbb{R}^n$ is the space of column vectors, then V^* is the space of row vectors, and then $(V^*)^*$ is the space of column vectors again. In some sense, this is similar to the fact that taking transpose twice would go back to the original matrix.

Example 11.1.12. As you can see, the above proofs are essentially built upon the fact that $\dim V = \dim V^*$. Unfortunately, this is only true for finite dimensional spaces. For infinite dimensional spaces, $\dim V^*$ is always larger than $\dim V$. (And then $(V^*)^*$ would be even bigger, so we would have $V \neq (V^*)^*$.)

Let us see an example. Let V be the space of finite sequences, i.e., $\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$ such that after finitely many terms, all later terms are zero. Then for each INFINITE sequence $\boldsymbol{b}^* = \begin{bmatrix} b_1 & b_2 & \dots \end{bmatrix}$, we can think of it as a linear map $\boldsymbol{a} \mapsto \sum a_i b_i$ from V to $\mathbb{R}$. Note that the sum is always defined, because it is in fact a finite sum, as only finitely many a_i are non-zero.

We see that all infinite sequences are in V^*! In fact that is everything. Informally, we can say that the dual to the space of finite sequences is the space of all infinite sequences.

(If you have the extra knowledge, you can further verify that V is coutable-dimensional while V^*, the space of all infinite sequences, is uncountable-dimensional.) ☺

Now we usually do computations by picking a basis. If we have a basis for V, we would like to pick a "corresponding" basis for V^* which would hopefully make my computations easier.

Definition 11.1.13. *Given a basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ in V, then we say a basis $\alpha_1, \dots, \alpha_n$ for V^* is its* ***dual basis*** *if $\alpha_i(\boldsymbol{v}_j) = \delta_{ij}$. (Here as usual, δ_{ij} is 1 if $i = j$, and 0 if $i \neq j$.)*

Example 11.1.14. Let V be the space of polynomials of degree at most 2. If we pick basis $1, x, x^2$ for V, then what is the dual basis?

The dual basis are $\alpha : p \mapsto p(0)$, $\beta : p \mapsto p'(0)$ and $\gamma : p \mapsto \frac{1}{2!}p''(0)$. I shall leave the verification for you.

By the way, hopefully you can see the pattern here: they are Taylor expansion coefficients at zero. A polynomial is simply a function whose Taylor expansion terminates after finitely many steps. ☺

"Dual basis" looks like something new, but it is actually quite familiar to us. $\alpha_i(\boldsymbol{v})$ simply means "the $\boldsymbol{v}_i$-coordinate of $\boldsymbol{v}$". For example, if $\boldsymbol{v} = a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n$, you can easily see that $\alpha_i(\boldsymbol{v}) = a_i$. So given a basis, its dual basis are simply the "coordinates" under this basis.

Example 11.1.15. If we use the standard basis for $\mathbb{R}^3$, then its dual basis vectors are "x-coordinate", "y-coordinate" and "z-coordinate" maps.

In general, fix the basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ for $\mathbb{R}^n$. Then its dual basis is obviously $\boldsymbol{e}_1^{\mathrm{T}}, \dots, \boldsymbol{e}_n^{\mathrm{T}}$ in the space of row vectors $(\mathbb{R}^n)^*$. These corresponds to taking the corresponding coordinates, as always.

But what if the basis are ugly? Now consider the basis $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}$. You may verify that its dual basis is in fact $\begin{bmatrix}1 & -1 & 0\end{bmatrix}, \begin{bmatrix}0 & 1 & -1\end{bmatrix}, \begin{bmatrix}0 & 0 & 1\end{bmatrix}$. Look, if we put the basis and dual basis into matrices, we see that $\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 1\\0 & 0 & 1\end{bmatrix}$ while the dual basis is $\begin{bmatrix}\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix} = \begin{bmatrix}1 & -1 & 0\\0 & 1 & -1\\0 & 0 & 1\end{bmatrix}$! Wow, coincidence?

Furthermore, note that in the last two examples, $\boldsymbol{v}_1$ are the same, while α_1 are different! This is a very important thing to remember: α_i does NOT depend on $\boldsymbol{v}_i$. In fact, the opposite is true (which we shall prove below): α_i depends completely on $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_n$. (I.e., anything but $\boldsymbol{v}_i$.) ☺

Proposition 11.1.16. *If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ form a basis in $\mathbb{C}^n$, let $\alpha_1, \dots, \alpha_n$ be the corresponding dual basis in the row vector space $(\mathbb{C}^n)^*$. Then the complex matrices $A = \begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$ and $B = \begin{bmatrix}\alpha_1\\\vdots\\\alpha_n\end{bmatrix}$ are inverse of each other.*

Proof. The block matrix multiplication shows $BA = \begin{bmatrix}\alpha_1\boldsymbol{v}_1 & \dots & \alpha_1\boldsymbol{v}_n\\\vdots & \ddots & \vdots\\\alpha_n\boldsymbol{v}_1 & \dots & \alpha_n\boldsymbol{v}_n\end{bmatrix}$. Since $\alpha_i(\boldsymbol{v}_j) = \delta_{ij}$, we see that $BA = I$. □

So finding the dual basis is exactly the same as finding inverse. In particular, given a basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, say the dual basis is $\alpha_1, \dots, \alpha_n$. What determines α_1? Well, let $A = \begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$, then we are wondering about what determines the first ROW of A^{-1}. Look back at your linear algebra one notes, you shall realize that the first row of A^{-1} is exactly determined by $\boldsymbol{v}_2, \dots, \boldsymbol{v}_n$, i.e., all columns but $\boldsymbol{v}_1$.

(In fact, if you were in my linear algebra class, you can even see the geometric relation. Say we are over $\mathbb{R}^3$, and we use the Euclidean length. Then $\boldsymbol{v}_2, \boldsymbol{v}_3$ form a parallelogram in the space. And α_1 (flipped into a column vector) is in the perpendicular direction to this parallelogram, while its length is the same as the area of the parallelogram. In general, α_1 is perpendicular to the hyperplane spanned by $\boldsymbol{v}_2, \dots, \boldsymbol{v}_n$, and its length is the $(n-1)$-dim volumn of the $(n-1)$-dim parallelotope made of $\boldsymbol{v}_2, \dots, \boldsymbol{v}_n$.)

Example 11.1.17 (Polynomial interpolation)**.** Suppose I want to find all polynomials p such that $p(1) = a$, $p(2) = b$ and $p'(1) = c$ for some given constants a, b, c. What should I do?

The central ideal of linear algebra is the reduction to zero. Let us first find all solutions to the requirements $p(1) = p'(1) = p(2) = 0$. Well, this is not so bad. $p(1) = p(2) = 0$ means we have roots at 1 and 2. So p must have factors $(x-1)$ and $(x-2)$. Furthermore, $p'(1) = 0$ means 1 is in fact a "double roots" for p, so $(x-1)^2$ must be a factor of p. It is easy to verify that these are all necessary and sufficient conditions. So p satisfies $p(1) = 0$, $p'(1) = 0$ and $p(2) = 0$ if and only if p is a multiple of $(x-1)^2(x-2)$. So in this case, all the solutions are $q(x)(x-1)^2(x-2)$ for an arbitrary polynomial q.

Now back to our problem. We do not know how to find a polynomial p such that $p(1) = a$, $p(2) = b$ and $p'(1) = c$. But any two solutions p_1, p_2 must have $(p_1 - p_2)(1) = (p_1 - p_2)'(1) = (p_1 - p_2)(2) = 0$. In particular,

if we have found one solution p_0, then we know that all solutions must be $p_0(x) + q(x)(x-1)^2(x-2)$ for an arbitrary polynomial q.

So how to locate our polynomial p_0? Well, since we are only interested "modulus of $(x-1)^2(x-2)$", it is enough to search for solutions among polynomials of degree at most 2. Let V be the space of polynomials of degree at most 2. Note that $\dim V = 3$.

Our requirements are essentially dual vectors. Let α_1 be the dual vector in V^* that sends p to $p(1)$, and let α_2 be the dual vector in V^* that sends p to $p(2)$, and finally let α_3 be the dual vector in V^* that sends p to $p'(1)$. We want to find $p \in V$ such that $\alpha_1(p), \alpha_2(p), \alpha_3(p)$ gives us the desired values. Or in vector language, we want to find p such that $\begin{bmatrix}\alpha_1(p)\\ \alpha_2(p)\\ \alpha_3(p)\end{bmatrix} = \begin{bmatrix}a\\ b\\ c\end{bmatrix}$. How to do that?

I think we have had this investigation before. What we did was to notice that $L : V \to \mathbb{R}^3$ via $p \mapsto \begin{bmatrix}\alpha_1(p)\\ \alpha_2(p)\\ \alpha_3(p)\end{bmatrix}$ is a linear map from a 3-dimensional space to a 3-dimensional space, and it is easy to verify that it is injective (since $p(1) = p'(1) = p(2) = 0$ implies p must be a multiple of a degree 3 polynomial). So it is bijective, and thus a solution exists. But WAIT! This only shows the existence of a solution. It does not FIND the actual solution!

What do we do to FIND the actual solution? Well, let us start some day-dreaming. Suppose we have magically found a polynomial p_1, p_2, p_3 such that L sends them to the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. Then we immediately see that $ap_1 + bp_2 + cp_3$ would be sent to $a\boldsymbol{e}_1 + b\boldsymbol{e}_2 + c\boldsymbol{e}_3 = \begin{bmatrix}a\\ b\\ c\end{bmatrix}$. YES!

What are these p_1, p_2, p_3? To be sent to $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, they need to satisfy the condition of $\begin{bmatrix}\alpha_1(p_1) & \alpha_1(p_2) & \alpha_1(p_3)\\ \alpha_2(p_1) & \alpha_2(p_2) & \alpha_2(p_3)\\ \alpha_3(p_1) & \alpha_3(p_2) & \alpha_3(p_3)\end{bmatrix} = I$, the identity matrix. In particular, you can see that this requires $\alpha_i(p_j) = \delta_{ij}$. So $\alpha_1, \alpha_2, \alpha_3 \in V^*$ should be a dual basis to $p_1, p_2, p_3 \in V$!

(Optional paragraph.) Indeed, if $\alpha_1, \alpha_2, \alpha_3 \in V^*$ is a dual basis to $p_1, p_2, p_3 \in V$, then $\begin{bmatrix}\alpha_1(p_1) & \alpha_1(p_2) & \alpha_1(p_3)\\ \alpha_2(p_1) & \alpha_2(p_2) & \alpha_2(p_3)\\ \alpha_3(p_1) & \alpha_3(p_2) & \alpha_3(p_3)\end{bmatrix} = I$, the identity matrix. Then $ap_1 + bp_2 + cp_3$ would evaluate into

$$\begin{bmatrix}\alpha_1(ap_1+bp_2+cp_3)\\ \alpha_2(ap_1+bp_2+cp_3)\\ \alpha_3(ap_1+bp_2+cp_3)\end{bmatrix} = \begin{bmatrix}a\alpha_1(p_1)+b\alpha_1(p_2)+c\alpha_1(p_3)\\ a\alpha_2(p_1)+b\alpha_2(p_2)+c\alpha_2(p_3)\\ a\alpha_3(p_1)+b\alpha_3(p_2)+c\alpha_3(p_3)\end{bmatrix} = \begin{bmatrix}\alpha_1(p_1) & \alpha_1(p_2) & \alpha_1(p_3)\\ \alpha_2(p_1) & \alpha_2(p_2) & \alpha_2(p_3)\\ \alpha_3(p_1) & \alpha_3(p_2) & \alpha_3(p_3)\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}a\\ b\\ c\end{bmatrix}.$$

So let me summarize. How to find p such that $p(1) = a$, $p(2) = b$ and $p'(1) = c$? All we need to do is to find a basis $p_1, p_2, p_3 \in V$ whose dual basis is $\alpha_1 : p \mapsto p(1), \alpha_2 : p \mapsto p(2), \alpha_3 : p \mapsto p'(1)$.

How to find the basis p_1, p_2, p_3? We start from any easy basis, say $1, x, x^2$ for V, then we can think of V as the space of column vectors $\mathbb{R}^3$ and V^* as the corresponding space of row vectors. Then the row-vector coordinates for α_1 is $\begin{bmatrix}\alpha_1(1) & \alpha_1(x) & \alpha_1(x^2)\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}$. Similarly, we can calculate the row-vector coordinates for α_2 and α_3, and we get $\begin{bmatrix}1 & 2 & 4\end{bmatrix}, \begin{bmatrix}0 & 1 & 2\end{bmatrix}$. So $\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 4\\ 0 & 1 & 2\end{bmatrix}$. But we should have

$$\begin{bmatrix}p_1 & p_2 & p_3\end{bmatrix} = \begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix}^{-1} = \begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 4\\ 0 & 1 & 2\end{bmatrix}^{-1} = \begin{bmatrix}0 & 1 & -2\\ 2 & -2 & 3\\ -1 & 1 & -1\end{bmatrix}.$$

So $p_1(x) = 2x - x^2, p_2(x) = 1 - 2x + x^2, p_3(x) = -2 + 3x - x^2$. You may verify that indeed $\alpha_i(p_j) = \delta_{ij}$.

So the all solutions are $[(-a+b-c)x^2 + (2a-2b+3c)x + b - 2c] + q(x)(x-1)^2(x-2)$ for an arbitrary polynomial $q(x)$.

Our example is long, but it is mostly explanatory. The actual solution process is very short: first identify $\alpha_1, \alpha_2, \alpha_3$, then write them in coordinates and put them into a matrix. Next find inverse, and read the columns, and we get p_1, p_2, p_3, and we are done. ☺

Remark 11.1.18. *It is easy to see from these discussion that, for any basis $\alpha_1, \ldots, \alpha_n \in V^*$, then we can find a unique "dual basis" $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$. To do this, simply put the rows $\alpha_1, \ldots, \alpha_n$ into a matrix, take inverse, and look at the columns.*

This gives rise to a very nice intuition about dual vectors: what is the meaning of a dual vector $\alpha \in V^*$? α is always "taking a coordinate" under some basis of V. All dual vectors are coordinate maps.

11.2 Dual Maps

Recall the starting example from last time. We are combining foods into combos. And this induces two maps, one is the "counting map" L from the combo space V to the food space W, and one is the "combining map" L^* from the food price space V^* to the combo price space W^*. Surprisingly, the corresponding forms look like transposes of each other!

And say we are purchasing a meal from a store where the food price is $\alpha \in V^*$. We can buy a combo $\boldsymbol{w} \in W^*$ (again without discount). Then we can check out via the combo price $L^*(\alpha)(\boldsymbol{w})$, and we can also check out via the food price $\alpha(L\boldsymbol{w})$.

Proposition 11.2.1. *For any linear map $L : V \to W$, then $\alpha \mapsto \alpha \circ L$ is a linear map from W^* to V^*.*

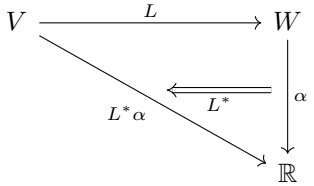
Proof. For any $\alpha \in W^*$ and $\boldsymbol{v} \in V$, since $L : V \to W$ and $\alpha : W \to \mathbb{R}$ (or $\mathbb{C}$ if we were over complex spaces) are well-defined linear maps, therefore their composition $\alpha \circ L(\boldsymbol{v})$ is a well-defined linear map from V to $\mathbb{R}$, i.e., $\alpha \circ L$ is an element of V^*. □

Definition 11.2.2. *For any linear map $L : V \to W$, we define its dual map to be the linear map $L^* : W^* \to V^*$ such that $L^*(\alpha) = \alpha \circ L$.*

It is unfortunate that it shares the same notation as adjoints (conjugate transposes). But I assure you that this is the standard notation. The two are related, but not to be confused. Sometimes standard notations are confusing and not too smart, and we have to live with it because everyone else is using it. From now on, we shall use A^* to denote the dual of A, and only use A^{ad} to denote the adjoint of A.

(Usually you can tell the difference between a dual and an adjoint by looking at the domain and codomain. Given a map between inner product spaces $L : V \to W$, the its adjoint is $L^{ad} : W \to V$ while its dual is $L^* : W^* \to V^*$.)

Graphically, this process looks like this:



As you can see, L sends vectors to vectors, while L^* sends arrows to arrows. And while L would "push forward" vectors from V to W, the map L^* would "pull back" dual vectors from W^* to V^*.

Example 11.2.3. The "pushforward" and "pullback" phenomena is very common. Here are some real life examples.

Say X is the set of all people, and Y is the set of all jobs, and we have an assignment map $f : X \to Y$ that sends people to jobs.

Now what should Y^* be? What is an evaluation of jobs? The salary. Consider $\alpha \in Y^*$ which is a map from Y to $\mathbb{R}$ that send jobs to salaries. Now if all jobs now have well-defined salaries, then immediately for

each person, we can first find a job via f, and the look at the salary. So this map $\alpha \circ f$ is evaluation people to their personal income.

This is exactly what the dual map $f^* : Y^* \to X^*$ should do: pull back evaluations of jobs to evaluations of people. In particular, f^* would send the salary of a job α to the income of a person $\alpha \circ f$.

The slogan is the following: If you push stuff forward, then you are pulling back evaluations of the stuff. For another non-rigorous example of some people's parenting style: if parents push their world view onto their children, then they can pull back achievements off of their children. ("If the child did what I said, and then achieved something, then it is my achievement!") (Achievements can be considered as an evaluation of one's world view.) ☺

The important thing to keep in mind is that L and L^* are essentially the same process. In some sense, you can think of L as a matrix that is going to multiply some column vector, i.e., the process $\boldsymbol{v} \mapsto L\boldsymbol{v}$. And you can think of L^* as the very same matrix but it is now going to multiply a row vector, i.e., the process $\alpha \mapsto \alpha L$ where $\alpha \in V^*$ is the row vector.

However, even though it might be tempted to use the same matrix to represent both, they have drastically different behaviors. L wants to multiply (column) vectors to its right, while L^* wants to multply (row) vectors to its left! The direction of multiplication is different!

In particular, $(L_1L_2)\boldsymbol{v}$ would have L_2 happening first, and then L_1. But $\alpha(L_1L_2)$ will have L_1 multiplied to α first, and then L_2, so the order of multiplication is different! (In particular, you immediately see that $(L_1L_2)^* = L_2^*L_1^*$.)

Proposition 11.2.4. $(AB)^* = B^*A^*$.

Proof. $(AB)^*(\alpha) = \alpha \circ (AB) = (\alpha \circ A) \circ B = (A^*(\alpha)) \circ B = B^*(A^*(\alpha)) = (B^*A^*)\alpha$. Be ware of the parenthesis. The only tools we used here are the law of associativity for function composition, and the definition of a dual map. □

What if we want to write the linear map $\alpha \mapsto \alpha L$ the usual way, as $L^*(\alpha)$ where we treat α as a (column) vector? To make a row vector α into a column vector, we need to take transpose. Then the process $\alpha \mapsto \alpha L$ now looks like $\alpha^{\mathrm{T}} \mapsto (\alpha L)^{\mathrm{T}} = L^{\mathrm{T}}\alpha^{\mathrm{T}}$.

So as it turns out, if we FORCE the notation as $L^*(\alpha)$ (the "correct" order) to denote the process αL, then the resulting matrix for L^* would be the transpose of L. Let us now establish these claims rigorously.

Proposition 11.2.5. *If the matrix for $L : V \to W$ under some basis is A, then the matrix for $L^* : W^* \to V^*$ under the dual basis is A^{T}. (Even over complex vector spaces!)*

Proof. Pick basis for V, W and dual basis for V^*, W^*, then we can pretend that V, W are $\mathbb{C}^n, \mathbb{C}^m$ with the standard basis and V^*, W^* are the space of row vectors, with basis $\boldsymbol{e}_1^{\mathrm{T}}, \ldots, \boldsymbol{e}_n^{\mathrm{T}}$ and basis $\boldsymbol{e}_1^{\mathrm{T}}, \ldots, \boldsymbol{e}_m^{\mathrm{T}}$.

Now the (i,j)-entry of L is the i-th coordinate of $L(\boldsymbol{e}_j) = A\boldsymbol{e}_j$, so it is $\boldsymbol{e}_i^{\mathrm{T}}L\boldsymbol{e}_j$. While the (i,j)-entry of L is the i-th coordinate of $L^*(\boldsymbol{e}_j^{\mathrm{T}}) = \boldsymbol{e}_j^{\mathrm{T}}A$, so it is $\boldsymbol{e}_j^{\mathrm{T}}A\boldsymbol{e}_i$. Now it is clear that the (i,j)-entry of L is exactly the (j,i)-entry of L^*. So the matrix for L^* is A^{T}. □

A super important distinction here. The "dual" operation (or "transpose") is linear, i.e., $(kL)^* = kL^*$ for all complex k, while taking adjoint is NOT complex linear, i.e., $(kL)^{ad} = \overline{k}L^{ad}$.

Since we have now established the relation between "dual" and "transpose", we technically do not need anything below. However, I present them here nonetheless as an approach that is independent of basis. After all, the "proper" way to do things is NOT to think of dual as transpose. Rather, we should do the opposite. We should think of transpose (basis dependent concept) as a representation of the dual process (basis independent concept). All properties of transpose should be derived from properties of dual, not the other way around.

(Also, hopefully this shall explain some of the mysterious phenomena surrounding transpose. Why would A, A^{T} have the same rank? Why must they have the same Jordan form? Why must $f(A^{\mathrm{T}}) = f(A)^{\mathrm{T}}$? The following perspective would hopefully make these things less "mysterious" and more "obvious".)

Proposition 11.2.6. *(Note that for fixed vector spaces V, W, all linear maps from V to W would form a vector space. We denote this as $\mathcal{L}(V, W)$.) For finite dimensional spaces V, W, the "dual" operator $(-)^* : \mathcal{L}(V, W) \to \mathcal{L}(W^*, V^*)$ is a linear isomorphism.*

Proof. The domain and codomain of $(-)^*$ both have dimension $(\dim V)(\dim W)$, so we only need to establish injectivity.

Injectivity requires only the lemma below. □

Lemma 11.2.7. *For any linear map L between finite dimensional spaces, $L = 0$ if and only if $L^* = 0$.*

Proof. $L = 0$ if and only if $L\boldsymbol{v} = \mathbf{0}$ for all $\boldsymbol{v} \in V$,
if and only if $\alpha L\boldsymbol{v} = 0$ for all $\boldsymbol{v} \in V, \alpha \in V^*$,
if and only if $(L^*(\alpha))(\boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in V, \alpha \in V^*$,
if and only if $L^*(\alpha) = 0$ for all $\alpha \in V^*$,
if and only if $L^* = 0$. □

Before we move on, here is a cute characterization of injectivity and surjectivity for linear maps. The content is not that useful, but the perspective is super interesting.

Lemma 11.2.8 (Dual vector knows the difference)**.** *In a vector space V, let W be any proper subspace (i.e., $W \neq V$) and pick any $\boldsymbol{v} \in V - W$. Then we can find $\alpha \in V^*$ such that $\alpha(W) = 0$ and $\alpha(\boldsymbol{v}) = 1$.*

Proof. (If you are a different person form the rest of the class, then somewhere there MUST BE an exam that you will score 100/100 while all your classmates get zero....)

It is harmless to make W larger. So WLOG we can assume that W has dimension $\dim V - 1$, and thus W and $\boldsymbol{v}$ would span V.

For each $\boldsymbol{u} \in V$, note that $W, \boldsymbol{v}$ spans V, so $\boldsymbol{u}$ has a unique decomposition $\boldsymbol{u} = \boldsymbol{w} + k\boldsymbol{v}$ for some $\boldsymbol{w} \in W$ and scalar k. We define $\alpha(\boldsymbol{u}) = k$. I claim that this is linear.

Indeed, if $\boldsymbol{u}_1 = \boldsymbol{w}_1 + k_1\boldsymbol{v}$ and $\boldsymbol{u}_2 = \boldsymbol{w}_2 + k_2\boldsymbol{v}$, then $\boldsymbol{u}_1 + \boldsymbol{u}_2 = (\boldsymbol{w}_1 + \boldsymbol{w}_2) + (k_1 + k_2)\boldsymbol{v}$, so $\alpha(\boldsymbol{u}_1 + \boldsymbol{u}_2) = \alpha(\boldsymbol{u}_1) + \alpha(\boldsymbol{u}_2)$. Similarly, if $\boldsymbol{u} = \boldsymbol{w} + k\boldsymbol{v}$, then $a\boldsymbol{u} = a\boldsymbol{w} + ak\boldsymbol{v}$, so $\alpha(a\boldsymbol{u}) = a\alpha(\boldsymbol{u})$. So α is a linear map from V to $\mathbb{R}$ (or $\mathbb{C}$). So $\alpha \in V$.

Now we can easily verify that $\alpha(W) = 0$ while $\alpha(\boldsymbol{v}) = 1$. □

Now we establish the following perspective: injectivity and surjectivity is really about the law of cancellations. Recall that in arithmetic calculations, for any non-zero $a \in \mathbb{R}$ and any $b, c \in \mathbb{R}$, we know that $ab = ac$ implies $b = c$, and $ba = ca$ implies $b = c$. This is called the law of cancellation, and I'm sure you all love this. Sadly, this is false for matrices. In general, $AB = AC$ might NOT imply $B = C$, and $BA = CA$ might NOT imply $B = C$, unless A is invertible.

Here we show that injectivity is the same as left-cancellation, and surjectivity is the same as right-cancellation.

Lemma 11.2.9 (Categorical characterization of injectivity and surjectivity)**.** *A linear map L is injective if and only if $LT_1 = LT_2$ implies $T_1 = T_2$ for any linear maps T_1, T_2. (Assuming that domains and codomains match so that everything is well-defined.)*

Similarly, A linear map L is surjective if and only if $T_1L = T_2L$ implies $T_1 = T_2$ for any linear maps T_1, T_2. (Assuming that domains and codomains match so that everything is well-defined.)

Intuitively, injectivity is defined as $L\boldsymbol{v} = L\boldsymbol{w}$ implies $\boldsymbol{v} = \boldsymbol{w}$, which is already a special case of left-cancellation. For the surjectivity portion, note that $T_1L = T_2L$ means T_1, T_2 agrees on $\mathrm{Ran}(L)$. But they could potentially disagree outside of $\mathrm{Ran}(L)$. So this implies $T_1 = T_2$ if and only if there is NO "outside", i.e., $\mathrm{Ran}(L)$ is the whole space.

Proof. Suppose L is injective, and $LT_1 = LT_2$. Then T_1, T_2 have common domain. And for any $\boldsymbol{v}$ in this common domain, $L(T_1\boldsymbol{v}) = L(T_2\boldsymbol{v})$. By injectivity of L, this means $T_1\boldsymbol{v} = T_2\boldsymbol{v}$. But this $\boldsymbol{v}$ is arbitrary, so T_1, T_2 are the same linear map.

Conversely, suppose $LT_1 = LT_2$ implies $T_1 = T_2$ for any linear maps T_1, T_2. We want to show that L is injective. (Pick any $\boldsymbol{v}, \boldsymbol{w}$, assuming $L\boldsymbol{v} = L\boldsymbol{w}$, we want to show that $\boldsymbol{v} = \boldsymbol{w}$.)

Then for any $\boldsymbol{v}, \boldsymbol{w}$ in the domain V of L, suppose $L(\boldsymbol{v}) = L(\boldsymbol{w})$. Set $T_1 : \mathbb{C} \to V$ such that $T_1(k) = k\boldsymbol{v}$, and set $T_2 : \mathbb{C} \to V$ such that $T_2(k) = k\boldsymbol{w}$. We then have $LT_1(k) = L(k\boldsymbol{v}) = kL(\boldsymbol{v}) = kL(\boldsymbol{w}) = L(k\boldsymbol{w}) = LT_2(k)$, and this is true for all k. So $LT_1 = LT_2$ as linear maps. So $T_1 = T_2$ as linear maps. So $\boldsymbol{v} = \boldsymbol{w}$.

The surjectivity portion is similar. Suppse L is surjective and $T_1 L = T_2 L$. Then for any $\boldsymbol{v}$ in the codomain of L (which is also the common domain of T_1 and T_2), then $\boldsymbol{v} = L\boldsymbol{w}$ for some $\boldsymbol{w}$. So $T_1 L\boldsymbol{w} = T_2 L\boldsymbol{w}$, which implies that $T_1\boldsymbol{v} = T_2\boldsymbol{v}$. But since $\boldsymbol{v}$ is arbitrary, the two linear maps T_1, T_2 are the same.

Conversely, suppose L is NOT surjective. Then we shall show that $T_1 L = T_2 L$ might not imply $T_1 = T_2$. Pick any $\boldsymbol{v} \notin \operatorname{Ran}(L)$, then we can find α such that $\alpha(\boldsymbol{v}) = 1$ and $\alpha(\operatorname{Ran}(L)) = 0$.

Now clearly $\alpha \neq 2\alpha$. However, $\alpha(L(\boldsymbol{v})) \in \alpha(\operatorname{Ran}(L)) = 0$, and $2\alpha(L(\boldsymbol{v})) \in \alpha(\operatorname{Ran}(L)) = 0$, so $\alpha \circ L = (2\alpha) \circ L$. So $T_1 L = T_2 L$ cannot imply $T_1 = T_2$ for any linear maps T_1, T_2. □

As you can see here, the difference between injectivity and surjectivity lies in the "order of multiplication". I.e., whether it is LEFT cancellation or RIGHT cancellation. So if something, say the "dual" process, would switch up the order of multiplication, then it would swap the two concepts.

Proposition 11.2.10. *L is injective if and only if L^* is surjective, and L is surjective if and only if L^* is injective. (The transpose version is obvious.)*

Proof. L is injective if and only if $LT_1 = LT_2$ implies $T_1 = T_2$ for any linear maps T_1, T_2,
if and only if $(LT_1)^* = (LT_2)^*$ implies $T_1^* = T_2^*$ for any linear maps T_1^*, T_2^*,
if and only if $T_1^* L^* = T_2^* L^*$ implies $T_1^* = T_2^*$ for any linear maps T_1^*, T_2^*,
if and only if L^* is surjective.

The other one is identical. □

Corollary 11.2.11. *For any linear map L, $\dim \operatorname{Ran}(L) = \dim \operatorname{Ran}(L^*)$. (The basis-dependent expression is that A and A^{T} have the same rank.)*

Proof. We decompose $L : V \to W$ into two parts, the "essense of L" which is $L_e : V \to \operatorname{Ran}(L)$. This is essentially just L, except that we throw away the untouched portion of the codomain. Let $\iota : \operatorname{Ran}(L) \to W$ be the inclusion map. Then $L = \iota \circ L_e$ where ι is injective and L_e is surjective.

Now we take dual. Then $L^* = L_e^* \circ \iota^*$, where L_e^* is injective and ι^* is surjective. Then $\dim \operatorname{Ran}(L^*) = \dim \operatorname{Ran}(L_e^* \circ \iota^*) = \dim \operatorname{Ran}(\iota^*)$, since injective linear map do not change dimensions. Finally, note that ι^* is a surjective map from W^* to $\operatorname{Ran}(L)^*$, so $\dim \operatorname{Ran}(\iota^*) = \dim \operatorname{Ran}(L)^* = \dim \operatorname{Ran}(L)$. So we are done. □

The slogan is this: RANK is the dimension of the middle space.

What is the rank of a linear map L? For any linear map $L : V \to W$, we can decompose it as $L = AB$ where $A : U \to W$ is injective and $B : V \to U$ is surjective, and rank of L is the dimension of the middle space $\dim U$. From this perspective, it is trivially obvious that L and L^* have the same rank. The dual process simply flips everything, and the middle space U and U^* have the same dimension. (Remember how bothersome it is to show that A and A^{T} have the same rank? Now it is just trivial word game.)

Corollary 11.2.12. *If $L : V \to V$ is a linear transformation, then $\dim \operatorname{Ker}(L - \lambda I)^k = \dim \operatorname{Ker}(L^* - \lambda I)^k$.*

Proof. $\dim \operatorname{Ran}(L - \lambda I)^k = \dim \operatorname{Ran}((L - \lambda I)^k)^* = \dim \operatorname{Ran}(L^* - \lambda I)^k$. Now $\dim \operatorname{Ker} = \dim V - \dim \operatorname{Ran}$, so we are done. □

Corollary 11.2.13. *L and L^* have the same Jordan canonical form.*

Proof. The Jordan canonical form is defined entirely by the generalized eigenstructures, i.e., the $\dim \operatorname{Ker}(L - \lambda I)^k$ stuff. But L, L^* have the same generalized eigenstructures, according to the last corollary. □

Corollary 11.2.14. $f(L)^* = f(L^*)$.

Proof. Oops, this is negligence on my part. In general, if we fix A, then $f(A) = p(A)$ for some polynomial p. But actually p does not depends on A, it only depends on the Jordan canonical form of A. For example, if $f(A) = p(A)$, then obviously $f(BAB^{-1}) = Bf(A)B^{-1} = Bp(A)B^{-1} = p(BAB^{-1})$.

Anyway, since L and L^* have the same Jordan canonical form, there is a polynomial p such that $f(L) = p(L)$ and $f(L^*) = p(L^*)$. So we only need to prove the statement when f is a polynomial.

Now the statement is true for powers. (E.g., $(L^k)^* = (L \dots L)^* = L^* \dots L^* = (L^*)^k$.) Thus it is true for polynomials (i.e., linear combinations of powers). □

Of course, everyting we've compiled here can also be proven if we simply think of dual as transpose.

11.3 Double Dual and Canonical Isomorphisms

In the world of linear algebra, "isomorphism" just means "bijective linear map", and we say two space are isomorphic if there is a bijective linear map, i.e., if they have the same dimension.

But now let us look at a stronger term, "canonical isomorphism" (or also "natural isomorphism" in some textbooks). For any vector space V, we can constructits dual space V^* and its double dual space $(V^*)^*$. In the finite dimensional world, all three spaces $V, V^*, (V^*)^*$ have the same dimension, so they are all isomorphic. However, we say V and $(V^*)^*$ are canonically isomorphic, while V and V^* has NO canonical isomorphism. What do we mean by this?

Vaguely, we have have the following feeling: even though V and V^* have the same dimension, but there is an "unseen order flip", i.e., the order of multiplication of linear transformations are reversed. If we have $A, B : V \to V$, then we have corresponding $A^*, B^* : V^* \to V^*$. But we do NOT have $(AB)^* = A^*B^*$. Rather, we have $(AB)^* = B^*A^*$. So the "unseen hidden-structure" of the two spaces are different.

But if we take dual twice, looking at V and $(V^*)^*$, then not only their linear structure match (same dimension), their "unseen hidden-structure" also match (because the order of multiplication flipped twice, which is the same as not flipped at all). We indeed have $(AB)^{**} = A^{**}B^{**}$ always.

So what does it mean to be canonically isomorphism? It is NOT just the relation between two spaces. It means that not only we can identify the two space, we can also identify all related linear maps.

Proposition 11.3.1. *Take the "evaluation map"* $\mathrm{ev} : V \to (V^*)^*$ *that sends* $\boldsymbol{v}$ *to* $\mathrm{ev}_{\boldsymbol{v}}$*. By abuse of notation, we also use the same symbol for the evaluation map* $\mathrm{ev} : W \to (W^*)^*$*. Then for any linear map* $A : V \to W$*, the linear maps* $\mathrm{ev} \circ A = (A^*)^* \circ \mathrm{ev}$*.*

In particular, we have the following diagram, where going down right is the same as going right down. You can see that not only ev identifies spaces, it also identifies maps. This is what we mean by the term "canonical isomorphism". It is not just about the spaces being in correspondence, but the fact that maps are also in correspondence.

$$\begin{array}{ccc} V & \xrightarrow{A} & W \\ \downarrow{\scriptstyle \mathrm{ev}} & & \downarrow{\scriptstyle \mathrm{ev}} \\ V^{**} & \xrightarrow{A^{**}} & W^{**} \end{array}$$

Before we proceed with the proof, note that the parenthesis here is a nightmare. Here are some calculational remark to keep in mind. These are NOT just for linear algebra, they are also true in other settings of mathematics. Here we use x for an element or a vector, f, g for functions or dual vectors or linear maps, A for linear maps or operators.

1. By definition of function composition, $f \circ g(x) = f(g(x))$.
2. By definition of dual, $A^*(f) = f \circ A$. (Applying this to the dual of A, we have $A^{**}(f) = f \circ A^*$.)
3. A more fancy way of writing above is $[A^*(f)](x) = f(Ax)$.
4. By definition of the evaluation map, $\mathrm{ev}(x) = \mathrm{ev}_x$, and $\mathrm{ev}_x(f) = f(x)$.

5. By definition of functions, if $f(x) = g(x)$ for all x, then $f = g$.

Proof of the last proposition. Pick any $\boldsymbol{v} \in V$. Then $\text{ev} \circ A(\boldsymbol{v})$ and $(A^*)^* \circ \text{ev}(\boldsymbol{v})$ are elemente of W^{**}, which evaluate elements of W^*. To show that they are the same, we need to show that they give the same evaluation for any $\alpha \in W^*$.

Take any $\alpha \in W^*$. Then we have

$$[\text{ev} \circ A(\boldsymbol{v})](\alpha) = [\text{ev}(A\boldsymbol{v})](\alpha) = \text{ev}_{A\boldsymbol{v}}(\alpha) = \alpha(A\boldsymbol{v}).$$

Now we tackle $(A^*)^* \circ \text{ev}(\boldsymbol{v})$ applied to α. Note that the dual map is defined as $L^*(\alpha) = \alpha \circ L$. So the dual of A^* will give us $(A^*)^* \circ \text{ev}(\boldsymbol{v}) = (A^*)^*(\text{ev}(\boldsymbol{v})) = \text{ev}(\boldsymbol{v}) \circ A^* = \text{ev}_{\boldsymbol{v}} \circ A^*$ by definition.

So we have

$$[(A^*)^* \circ \text{ev}(\boldsymbol{v})](\alpha) = (\text{ev}_{\boldsymbol{v}} \circ A^*)(\alpha) = \text{ev}_{\boldsymbol{v}}(A^*\alpha) = \text{ev}_{\boldsymbol{v}}(\alpha \circ A) = \alpha \circ A(\boldsymbol{v}) = \alpha(A\boldsymbol{v}).$$

Hey, so we see that $\text{ev} \circ A(\boldsymbol{v})$ and $(A^*)^* \circ \text{ev}(\boldsymbol{v})$ evaluate arbitrary $\alpha \in W^*$ to the same value $\alpha(A\boldsymbol{v})$. So $\text{ev} \circ A(\boldsymbol{v}) = (A^*)^* \circ \text{ev}(\boldsymbol{v})$. But since this is true for all $\boldsymbol{v}$, we have $\text{ev} \circ A = (A^*)^* \circ \text{ev}$. □

In summary, the "evaluation process" gives us the canonical isomorphism between V and V^{**}. Not only we can identify V and V^{**} as the same space for all V, we can also simultaneously identify A and A^{**} for all linear map A!

In comparison, here is the situation for a single dual, as opposed to double duals.

Example 11.3.2. V and V^* are NOT canonically isomorphic.

Suppose there is an canonical isomorphism between spaces and their duals. Say we have canonical isomorphisms $L_V : V \to V$ and $L_W : W \to W$. Then for any $A : V \to W$, we should have the diagram

$$\begin{array}{ccc} V & \xrightarrow{A} & W \\ \downarrow{\scriptstyle L} & & \downarrow{\scriptstyle L} \\ V^* & \xleftarrow[A^*]{} & W^* \end{array}$$

The above diagram is supposed to be true for all A. However, pick $A = 0$, and then $L_V = A^* L_w A = 0$ is NOT an isomorphism. Contradiction. ☺

11.4 Inner products and Dual space

Here we connect the dual space and the inner product structure. Depending on how your last semester was taught, this might be a review or new knowledge. For the moment, let us first restrict our attention to real numbers.

Think about dot products. The motivation of defining dot product is to define length of vectors and angles between vectors. For example, in an arbitrary abstract vector space,say P_2 the space of polynomials of degree at most 2, would you set x^2 to have length one? Would you set $\frac{1}{2}x^2$ to have length one? Would you set $1, x, x^2$ to be an orthonormal basis? Or would you rather set $1, x-1, (x-1)^2$ to be an orthonormal basis? There is no "unique best way" to do this. There is no innate "dot product" structure.

Definition 11.4.1. *Given a real vector space V, an inner product structure is a map $\langle -, - \rangle : V \times V \to \mathbb{R}$ that sends pairs of vectors to a complex number, such that the following is true:*

1. *(Bilinear) $\langle k\boldsymbol{v}, \boldsymbol{w} \rangle = k\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ and $\langle \boldsymbol{v}, k\boldsymbol{w} \rangle = k\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ and $\langle \boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{w} \rangle + \langle \boldsymbol{v}, \boldsymbol{w} \rangle$ and $\langle \boldsymbol{u}, \boldsymbol{v}+\boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle + \langle \boldsymbol{u}, \boldsymbol{w} \rangle$.*
2. *(Symmetric) $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{w}, \boldsymbol{v} \rangle$.*
3. *(Positive-Definite) $\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$ for all $\boldsymbol{v}$, and it is zero if and only if $\boldsymbol{v} = \boldsymbol{0}$.*

Then we define $\|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}$*, and the angle between any two non-zero vectors* $\boldsymbol{v}, \boldsymbol{w}$ *to be* $\arccos \frac{\langle \boldsymbol{v}, \boldsymbol{w}\rangle}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$.

Proposition 11.4.2. *For any finite dimensional real inner product space* V*, the inner product is some dot product under the coordinates of some basis.*

Proof. We would not do the whole proof here, since it is more relevant to last semester. However, we can make some short remark on the idea behind the proof.

First, we want to find an orthonormal basis, i.e., a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ such that $\langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle = \delta_{ij}$, i.e., the Gram matrix is the identity matrix. (The existance of an orthonormal basis is guaranteed by the Gram-Schmidt orthogonalization.) Now under this basis, we can then verify that the inner product is the dot product. □

This is nice in the following sense: If you are not familiar with inner product spaces, do not worry. It is simply "dot product" under some basis. In fact, check the following: a basis is an orthonormal basis if and only if the inner product is the dot product under coordinates of this basis.

Example 11.4.3. We here show an exotic example of inner product on an infinite dimensional space. Let $\mathcal{F}(\mathbb{R})$ be the space of all real integrable functions from $\mathbb{R}$ to $\mathbb{R}$.

For each function f, if we think of f as a "vector", then we may think of numbers such as $f(1), f(2), f(1.23)$ and such as "coordinates" of f. Then a "dot product" between two functions f and g will try to multiply these "coordinates", and try to "add" these products $f(x)g(x)$, i.e., $\int_{-\infty}^{\infty} f(x)g(x)\,\mathrm{d}x$. We define this to be $\langle f, g\rangle$ the "dot product" on $\mathcal{F}(\mathbb{R})$.

Note that if $\mathcal{F}([a, b])$ is the space of all real integrable functions from an interval $[a, b]$ to $\mathbb{R}$, then we can also define $\langle f, g\rangle = \int_a^b f(x)g(x)\,\mathrm{d}x$.

Now, is this "dot product" an inner product structure? The answer is no. We have bilinearity and symmetry obviously. But we only have positive SEMI-definiteness. Even thouhg $\int_{-\infty}^{\infty} f(x)f(x)\,\mathrm{d}x \geq 0$ always, but we might have non-zero integrable functions whose square integrate to zero, say if $f(x) = 0$ everywhere except that $f(x) = 1$ when $x = 0$.

One solution is to restrict our attention to continuous functions. Then $\int_{-\infty}^{\infty} f(x)^2\,\mathrm{d}x = 0$ would necessarily implies that $f = 0$ everywhere, and hence this becomes a genuin inner product.

Another get-away is to say that two functions are "almost the same" if their difference has "zero-length". I.e., we employ an equivalence relation where $f \cong g$ to NOT mean that $f(x) = g(x)$ for all x, but rather to mean that $\int_{\text{domain}}(f(x)-g(x))^2\,\mathrm{d}x = 0$. This yeilds a new vector space $\mathcal{L}_2(\mathbb{R})$ whose elements are "equivalent classes". On this space, $\langle f, g\rangle$ would be a genuin inner product.

Of course, another solution is to simply define "semi-inner product", and just be careful. Either way, the notion $\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)g(x)\,\mathrm{d}x$ is a very useful one. ☺

Now note that each inner product needs a pair of inputs, $\langle \boldsymbol{v}, \boldsymbol{w}\rangle$. If we FIX $\boldsymbol{v}$ and let $\boldsymbol{w}$ be an unknown input, then we have a linear map $\langle \boldsymbol{v}, -\rangle : V \to \mathbb{R}$. Hey, this is a dual vector!

Due to conventions in physics, who call the symbol $\langle -, -\rangle$ a "braket", people (especially in physics) sometimes use the following notations:

1. We think of $\langle \boldsymbol{v}, \boldsymbol{w}\rangle$ as "the bra" $\langle \boldsymbol{v}|$ and "the ket" $|\boldsymbol{w}\rangle$.
2. The bra of $\boldsymbol{v}$ refers to the linear map $\langle \boldsymbol{v}, -\rangle : V \to \mathbb{R}$. The ket of $\boldsymbol{w}$ means simply $\boldsymbol{w}$ itself, and we are only giving it this name to make the duality clearer.
3. So we may also think of $\langle \boldsymbol{v}, \boldsymbol{w}\rangle$ as the dual vector $\langle \boldsymbol{v}|$ applied to the vector $|\boldsymbol{w}\rangle$.

So what does an inner product do? Think about the bra process $\boldsymbol{v} \mapsto \langle \boldsymbol{v}|$. It gives us a canonical way to change vectors into dual vectors!

Theorem 11.4.4. *Given a real inner product space* V*, the bra map* $\langle -| : V \to V^*$ *is the unique linear bijection such that* $\langle \boldsymbol{v}|(\boldsymbol{w}) = \langle \boldsymbol{v}, \boldsymbol{w}\rangle$.

Proof. Uniqueness is obvious because we literally defined $\langle \boldsymbol{v}|(-)$ as $\langle \boldsymbol{v}, - \rangle$. Linearity is also obvious because the inner product is linear in its left input. Finally, to show bijectiveness, note that $\dim V = \dim V^*$, so we only need to show that $\langle -|$ is injective. Suppose $\langle \boldsymbol{v}|$ is the zero map. Then $\langle \boldsymbol{v}|(\boldsymbol{v}) = 0$, and thus $\|\boldsymbol{v}\| = 0$, and thus $\boldsymbol{v} = \boldsymbol{0}$. □

Previously, without any inner product structure, we discussed that there is NO canonical bijection between V and V^*. There are bijections between V and V^*, but they all suck. None of them have good properties. However, given an inner product structure, now we have a UNIQUE BEST bijection between V and V^* that could play well with the given an inner product structure!

In this sense, finding an inner product structure is like this: since there is no canonical isomorphism between V and V^*, we just pick a bijection artificially, and claim it to be canonical.

Btw, to make the resulting "canonical isomorphism" nice, we need the artificial bijection $L : V \to V^*$ to be "symmetric" in the sense of $L = L^*$. (Funny thing: L and L^* do have the same domain and codomain in this case.)

We also want L to be "positive definite", i.e., the linear evaluation $L(\boldsymbol{v})$ should not screw up $\boldsymbol{v}$ itself. In particular, the linear map $L(\boldsymbol{v})$ should always send $\boldsymbol{v}$ to a positive number (unless $\boldsymbol{v} = \boldsymbol{0}$, in this case $L(\boldsymbol{0})$ sends everyone to zero).

For any "symmetric" and "positive-definite" linear bijection $L : V \to V^*$, you may verify that $[L(\boldsymbol{v})](\boldsymbol{w})$ is indeed a inner product structure.

But now comes the moment of realization: if we have a linear bijection $L : V \to V^*$, then its inverse is also linear bijection from V^* to $V^{**} = V$! This means the following:

Corollary 11.4.5. *Given any inner product on V, there is a unique induced inner product on V^* such that the dual basis to an orthonormal basis is orthonormal. (I.e., dot product of column vectors would induce the dot product on row vectors.)*

Proof. Let L be the inverse of $\langle -| : V \to V^*$. Then we can verify that $[L(-)](-)$ is an inner product structure on V^*. (Here we identify V and V^{**} via the evaluation process as usual.) Let us do the verification now.

$[L(-)](-)$ is obvious bilinear by construction. Let us now verify symmetry. For any $\alpha, \beta \in V^*$, let $\boldsymbol{v} = L(\alpha)$ and $\boldsymbol{w} = L(\beta)$. Then $[L(\alpha)](\beta) = \mathrm{ev}_{\boldsymbol{v}}(\beta) = \beta(\boldsymbol{v})$. But since $\boldsymbol{w} = L(\beta)$, by definition we have $\beta = \langle \boldsymbol{w}|$. So $\beta(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \alpha(\boldsymbol{w})$. Finally, $[L(\beta)](\alpha) = \alpha(\boldsymbol{w})$ by the same process as before. So $[L(\alpha)](\beta) = [L(\beta)](\alpha)$.

Finally, we want to establish positive definiteness. For any $\alpha \in V^*$, let $\boldsymbol{v} = L(\alpha)$. Then we have $[L(\alpha)](\alpha) = \alpha(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{v} \rangle$. So this is positive unless $\boldsymbol{v} = \boldsymbol{0}$, which could happen if and only if $\alpha = 0$.

Finally, let us verify that dual basis to an orthonormal basis is orthonormal. Note that, along our previous arguments, we have proven something very funny: $[L(\alpha)](\beta) = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$. So if $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is an orthogonal basis, then immediately we see that $\langle \boldsymbol{v}_1|, \dots, \langle \boldsymbol{v}_n|$ is also an orthonormal basis. I claim that $\langle \boldsymbol{v}_1|, \dots, \langle \boldsymbol{v}_n|$ is also the dual basis.

To see this, note that $\langle \boldsymbol{v}_i|(\boldsymbol{v}_j) = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}$, so we are done. □

We explore one last idea stemming from the identification of V and V^*. It states that for each dual vector α, there is a unique vector $\boldsymbol{v}$ such that $\alpha = \langle \boldsymbol{v}|$. This is the inverse of the "bra" map, which we also call the Riesz map.

Theorem 11.4.6 (Riesz representation theorem)**.** *For any $\alpha \in V^*$ on an inner product space V, there is a unique $\boldsymbol{v}$ such that $\alpha = \langle \boldsymbol{v}|$.*

We call this map $V^* \to V$ (inverse of the bra map) as the Riesz map. This has some very interesting applications.

Example 11.4.7. Let X be a random real number. The standard way to study such a random number is to define its "probability distribution function". We say X has a probability function p_X if $\int_a^b p_X(x)\,\mathrm{d}x = \Pr(a \le X \le b)$. One might loosely think of $p_X(x)$ as the "probability" of $X = x$. We add this up for all $a \le x \le b$, and the result would be $\int_a^b p_X(x)\,\mathrm{d}x = \Pr(a \le X \le b)$.

Now note that X can also be thought of as a dual vector $f \mapsto \mathbb{E}(f(X))$. How is this calculated? Well, since X has "probability" $p_X(x)$ to be x, therefore $f(X)$ has "probability" $p_X(x)$ to be $f(x)$. To find the "average value" of the random value $f(X)$, we want to informally do something like $\sum_x f(x)\Pr(X = x)$, whose integration version is $\int_{-\infty}^{\infty} f(x)p_X(x)\,\mathrm{d}x$. So we usually have $\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(x)p_X(x)\,\mathrm{d}x = \langle p_X, f\rangle$.

In particular, if we think of X as a dual vector, then note that $f \xmapsto{X} \mathbb{E}(f(X))$ is the same as $\langle p_X|$. So we have $Riesz(X) = p_X$ the probability distribution function.

So going from a random variable to its probability distribution function is just trying to do the Riesz representation theorem. ☺

Example 11.4.8. Let us see an application of an infinite dimensional version of the Riesz representation theorem. We do not actually prove this theorem, since it needs a lot more set up. However, it illustrate perfectly how the idea of "dual vectors are represented as bra of vectors" could be useful.

Suppose we have a differential equation $-f''(x) + b(x)f(x) = q(x)$, where $b(x)$ is a known function and we always have $b(x) \geq 0$, and $q(x)$ is another known function. We are trying to solve for possible f. For simplicity, say f is defined on the interval $[0, 1]$, satisfying the initial condition $f'(0) = f'(1) = 0$. Let us show that a solution exist.

First, via integration by parts, for any function $\phi(x)$ we have $\int_0^1 f''(x)\phi(x)\,\mathrm{d}x = -\int_0^1 f'(x)\phi'(x)\,\mathrm{d}x$. We consider the dot product of both sides of our differential equation with an arbitrary function ϕ, and we have the computation:

$$\int_0^1 [-f''(x)\phi(x) + b(x)f(x)\phi(x)]\,\mathrm{d}x = \int_0^1 q(x)\phi(x)\,\mathrm{d}x$$
$$\int_0^1 [f'(x)\phi'(x) + b(x)f(x)\phi(x)]\,\mathrm{d}x = \int_0^1 q(x)\phi(x)\,\mathrm{d}x.$$

Let us define that $\langle f, g\rangle$ as $\int_0^1 [f'(x)g'(x) + b(x)f(x)g(x)]\,\mathrm{d}x$ on the space V of differntiable functions, then you can easily see that this is symmetric and positive definite. In fact, if $\langle f, f\rangle = 0$, then we must have $f = 0$.

Let us also define a dual vector $\alpha : V \to \mathbb{R}$ such that $f \mapsto \int_0^1 q(x)f(x)\,\mathrm{d}x$.

Then our differential equation is now this: $\langle f, \phi\rangle = \alpha(\phi)$.

Recall our goal: we want to find a solution f to the differential equation $-f''(x)+b(x)f(x) = q(x)$. Now by computations above, we have transformed our goal into the following: we want to find a solution f such that $\langle f, \phi\rangle = \alpha(\phi)$ for all ϕ? Or in short, given a dual vector α, can we find f such that $\alpha = \langle f|$? Well, by some corresponding Riesz representation theorem, we can. So there you go, a solution exists. ☺

11.5 (Optional) Complex Riesz map

Now let us consider the case of a complex space. For complex spaces, we have the following distinctions.

Definition 11.5.1. *Given a real vector space V, an inner product structure is a map $\langle -, -\rangle : V \times V \to \mathbb{C}$ that sends pairs of vectors to a complex number, such that the following is true:*

1. *(Sesquilinear) $\langle k\boldsymbol{v}, \boldsymbol{w}\rangle = \overline{k}\langle \boldsymbol{v}, \boldsymbol{w}\rangle$ and $\langle \boldsymbol{v}, k\boldsymbol{w}\rangle = k\langle \boldsymbol{v}, \boldsymbol{w}\rangle$ and $\langle \boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{u}, \boldsymbol{w}\rangle + \langle \boldsymbol{v}, \boldsymbol{w}\rangle$ and $\langle \boldsymbol{u}, \boldsymbol{v} + \boldsymbol{w}\rangle = \langle \boldsymbol{u}, \boldsymbol{v}\rangle + \langle \boldsymbol{u}, \boldsymbol{w}\rangle$.*
2. *(Conjugate Symmetric) $\overline{\langle \boldsymbol{v}, \boldsymbol{w}\rangle} = \langle \boldsymbol{w}, \boldsymbol{v}\rangle$. (Note that this implies that $\langle \boldsymbol{v}, \boldsymbol{v}\rangle$ is always real.)*
3. *(Positive-Definite) $\langle \boldsymbol{v}, \boldsymbol{v}\rangle \geq 0$ for all $\boldsymbol{v}$, and it is zero if and only if $\boldsymbol{v} = \boldsymbol{0}$.*

Then we define $\|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}$.

In many sense, this is just as before. For example, we have this result

Theorem 11.5.2. *Given a complex inner product space V, the bra map $\langle -| : V \to V^*$ is the unique linear bijection such that $\langle \boldsymbol{v}|(\boldsymbol{w}) = \langle \boldsymbol{v}, \boldsymbol{w}\rangle$.*

However, the main difference is the following: the complex inner product is NOT bilinear. It is only linear in the right input, and it is merely conjugate-linear in the left input. (Think about the dot product $\boldsymbol{v}^{ad}\boldsymbol{w}$.)

In particular, the identification between V and V^* via $\boldsymbol{v} \mapsto \langle \boldsymbol{v}|$ is NOT complex linear, only real linear. We have $\langle k\boldsymbol{v}| = \overline{k}\langle \boldsymbol{v}|$.

Nevertheless, we have maps $\langle -| : V \to V^*$ and its inverse $Riesz : V^* \to V$. They are not complex linear, but they still provide bijective identification of V and V^*. So for any linear transformation $L : V \to V$, for its dual map $L^* : V^* \to V^*$, we can identify the domain and codomain of L^* as V via the Riesz map, and thus obtain a linear map $L^{ad} : V \to V$.

Definition 11.5.3. *On an inner product space (real or complex), we define the adjoint of a linear transformation $L : V \to V$ to be $L^{ad} : V \to V$ such that $L^{ad}\boldsymbol{v} = Riesz(L^*\langle \boldsymbol{v}|)$.*

Note that, in a sense, $L^{ad} = Riesz \circ L^* \circ Riesz^{-1}$, so L^{ad} and L^* are the same map. Just like "similar matrices", they differ only via the "change of basis" which is the Riesz map. However, in the complex case, the "change of basis" here is NOT complex linear! This causes some computation trouble.

Lemma 11.5.4. *For any complex inner product space V, we pick an orthonormal basis for V, so we may treat V as the column vector space $\mathbb{C}^n$ and V^* as the row vector space. Then the bra map would send $\boldsymbol{v} \in \mathbb{C}^n$ to $\boldsymbol{v}^{ad}$. As the inverse of the bra map, the Riesz map send a row vector α to the column vector α^{ad}.*

Proof. This is obvious, since $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^{ad}\boldsymbol{w}$ under an orthonormal basis. □

Proposition 11.5.5. *For any complex inner product space V, we pick a basis for V and pick the dual basis for V^*. Consider any linear transformation $L : V \to V$, and suppose its matrix under the chosen basis is A. Then the matrix for L^* under the dual basis is A^{T}, but the matrix for L^{ad} is A^{ad}.*

Proof. $L^{ad}(\boldsymbol{v}) = Riesz(L^*\langle \boldsymbol{v}|) = Riesz(L^*(\boldsymbol{v}^{ad})) = Riesz(\boldsymbol{v}^{ad}A) = (\boldsymbol{v}^{ad}A)^{ad} = A^{ad}\boldsymbol{v}$. □

Corollary 11.5.6 (Alternative definition)**.** *The adjoint $L^{ad} : V \to V$ is the unique linear transformation on the inner product space V such that $\langle L\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{v}, L^{ad}\boldsymbol{w}\rangle$.*

Definition 11.5.7. *A linear transformation $L : V \to V$ is self-adjoint (Hermitian, or symmetric in the real case) if $L = L^{ad}$, skew-adjoint if $-L = L^{ad}$, unitary if $L^{-1} = L^{ad}$.*

Then one may proceed to do spectral theorems and so on. These should already be done in the last semester, so we stop here.

Chapter 12

Tangent Space and cotangent space

12.1 Tangent vectors and push forwards

The goal here is to use what we have learned to about dual spaces to study geometry. Technically, we want to talk about manifolds, but their formulation is a bit abstract and unwieldy. Nevertheless, let us take a look at an informal characterization of it.

Informally, a embedded n-manifold is a subset M of $\mathbb{R}^m$ for some m, such that it locally looks like $\mathbb{R}^n$ for some $n \leq m$. For example, we say a curve in $\mathbb{R}^m$ is a 1-manifold, because it is locally "line-like", i.e., $\mathbb{R}^1$. A surface is a 2-manifold, because it is locally "plane-like" and so on. We say it is a differentiable manifold if it has "tangent stuff" everywhere. For example, a differentiable curve in $\mathbb{R}^m$ is a curve with a well-defined tangent line everywhere. And a differentiable surface is a surface with a well-defined tangent plane everywhere.

Remark 12.1.1 (Formal Definition of a Differentiable Manifold)**.** *Skip this entirely, unless you are super curious. The main trouble of a formal definition is that of topology. One needs to learn topology before talking about manifolds.*

We define an open ball in $\mathbb{R}^m$ to be $B_r(\boldsymbol{p}) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{p}\| < r\}$. We define open subsets of $\mathbb{R}^m$ to be arbitrary unions of open balls, and define closed subsets of $\mathbb{R}^m$ to be complements of open subsets. Here are some observations, which hopefully gives you some intuition about the concept of open stuff and closed stuff.

1. *Arbitrary unions and finite intersections of open subsets are open.*
2. *Arbitrary intersections and finite unions of closed subsets are closed.*
3. *Given any subset S of $\mathbb{R}^m$, its "interior" is the largest open subset inside of it. (This is the definition.)*
4. *Given any subset S of $\mathbb{R}^m$, its "closure" is the smallest closed subset containing it. (This is the definition.)*
5. *A subset S of $\mathbb{R}^m$ is closed if and only if for any converging sequence inside of S, the limit is in S. (See if you can prove this.)*
6. *A subset S of $\mathbb{R}^m$ is open if and only if for any $\boldsymbol{p} \in S$, a "neighborhood" of p is inside of S, i.e., we can find $r > 0$ such that the open ball $B_r(\boldsymbol{p}) \subseteq S$.*
7. *A subset S is defined to be disconnected if we can find two open subsets U, V, $U \cap V = \varnothing$ and $U \cup V$ contains S.*
8. *The concept of "open intervals" and "closed intervals" are really open connected subsets and closed connected subsets of $\mathbb{R}$.*

For any open subset U, a diffeomorphic (i.e., "differentiably isomorphic") image of U is the image $f(U)$ of a continuously differentiable bijective function $f : U \to \mathbb{R}^m$ such that the total derivative of f (Jacobi matrix) is everywhere invertible. (This is to guarantee that its inverse is also differentiable.)

For any subset M of $\mathbb{R}^m$, we say it is an embedded differentiable n-manifold if for any $\boldsymbol{p} \in M$, we can find $r > 0$ such that $M \cap B_r(\boldsymbol{p})$ is the diffeomorphic image of an open ball in $\mathbb{R}^n$. (Thus the idea of "locally the same as $\mathbb{R}^n$".)

If you are even more hardcore, one can define manifold abstractly as "second countable Hausdorff space" with an "atlas" of open subsets each "homeomorphic" to an open ball in $\mathbb{R}^n$. (Don't worry if these words sound like gibberish....) And the manifold is a differentiable manifold if the transition maps for the atlas are all differentiable. This would require more definitions though, e.g., what is an atlas and what are the transition maps etc..

All these talks about Hausdorff space and atlas and such sound scary. However, by an advanced geometric theorem called the Whitney's embedding theorem, any abstract n-manifold can be "put" into $\mathbb{R}^{2n+1}$. So the definition is essentially the same to our previous "concrete" versions of manifolds.

Here I would propose an alternative way to study "manifolds", rather than cramming an entire course of topology here. Our goal is to define what is a "differentiable" thing, i.e., smooth curve, smooth surface and etc.. And the ultimate definition for that is to have a corresponding "tangent stuff". So we need to define tangent vectors. Intuitively, a tangent vector on a geometric object is a direction that, if I move a tiny bit along that direction, then I "approximately" would stay in the geometric object.

How to define this "tiny movement"? We start with curves.

Definition 12.1.2. *Given a subset X of $\mathbb{R}^m$, a **curve (segment)** is a continuous map $\gamma : [0, 1] \to X$. A curve is differentiable if, well, the map is differentiable.*

Here continuity or differentiability means that if we treat γ as a map from $[0, 1]$ to $\mathbb{R}^m$, then it is continuous or differentiable.

Remark 12.1.3. *Our requirement that the domain of γ is the closed interval $[0, 1]$ is unimportant. Change it into any closed interval $[a, b]$, you will be just fine.*

Just like linear maps $\boldsymbol{v} : \mathbb{R} \to V$ corresponds to "elements" of the space V, we study curves on X because they are the "continious elements" of X. If our goal is to study continuous structure of X, then mere discrete points are NOT enough. Curves (i.e., how two arbitrary points $\gamma(0)$ and $\gamma(1)$ "connect") gives us a primitive way to study such continuous structures.

It is then natural to do the following definitions.

Definition 12.1.4. *Given a subset X of $\mathbb{R}^m$ and a point $\boldsymbol{p} \in X$, we say $\boldsymbol{v} \in \mathbb{R}^m$ is a **tangent vector** to X at $\boldsymbol{p}$ if there is a differentiable curve $\gamma : [0, 1] \to X$ such that $\gamma(0) = \boldsymbol{p}$ and $\gamma'(0) = \lim_{t\to 0^+} \frac{\gamma(t)-\gamma(0)}{t} = \boldsymbol{v}$. (Since 0 is on the boundary, we only require the one-sided limit to exist.)*

Definition 12.1.5. *(I made this definition myself.) We say $X \in \mathbb{R}^m$ is a diffferential k-set if for each $\boldsymbol{p} \in X$, all possible tangent vectors to X at $\boldsymbol{p}$ form a k-dimensional subspace, i.e., the **tangent space** to X at $\boldsymbol{p}$, written as $T_{\boldsymbol{p}}(X)$.*

In short, tangent directions at $\boldsymbol{p}$ in X are "possible velocities" if we move inside of X along some curve, starting at $\boldsymbol{p}$. Seems natural enough, right? Here let us see some fun and weird examples. The "weirdness" is mostly for fun, and will not be tested, at least not in our class. The point is to do some mental exercises with our newly defined concepts.

Example 12.1.6. Consider the famous curve $\gamma : \mathbb{R} \to \mathbb{R}^2$ such that $f(t) = \begin{bmatrix} t \\ t^2 \sin(\frac{1}{t}) \end{bmatrix}$. The geometric object we study is simply $X = \gamma(\mathbb{R})$, the image of the curve. (I am using $\mathbb{R}$ as the domain of the curve, but it matters little. If you are feeling pedantic, then restrict to some closed intervals $[a, b]$.)

Note that this is also the graph of the function $t \mapsto t^2 \sin(\frac{1}{t})$, which is famously differentiable everywhere but NOT continuously differentiable. We can take derivative and see that $\gamma'(t) = \begin{bmatrix} 1 \\ 2t \sin(\frac{1}{t}) - \cos(\frac{1}{t}) \end{bmatrix}$ when

$t \neq 0$, and $\gamma'(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. And by using the curve $t \mapsto \gamma(kt)$, we see that $k\gamma'(t)$ is a tangent vector at $\gamma(t)$ for all t. So at each point in X, the tangent directions form a 1-dimensional subspace. $T_{\boldsymbol{p}}(X)$ is always a one dimensional subspace. X is a 1-dim differentiable set.

However, just as the function $t \mapsto t^2 \sin(\frac{1}{t})$ is NOT continuously differentiable, X has some trouble. The tangent line to X at the origin $T_{\mathbf{0}}(X)$ is horizontal, but points of X near the origin will have wildly osciliating tangent lines!

These are the "tangent directions" on the subset $X = \gamma(\mathbb{R})$, and they do NOT change continuously!

In particular, if we merely study differentiable subsets X of $\mathbb{R}$, then adjacent points might have "non-continuous" tangent directions.

Lucky for us, this example shall give us no trouble, because at least in our class, we would NEVER compare tangent vectors at different points. We only compare different tangent vectors at the same points. ☺

Example 12.1.7. Let us do a favorite geometric object of mine, the Hawaiian earring. Take the circle of radius one touching the origin, say e.g. $(x-1)^2 + y^2 = 1$. Then "shrink" this circle towards the origin to get a smaller circle, say $(x - \frac{1}{2})^2 + y^2 = \frac{1}{4}$ which now has radius $\frac{1}{2}$. Then shrink again to have an even smalle circle, say $(x - \frac{1}{4})^2 + y^2 = \frac{1}{16}$ which now has radius $\frac{1}{4}$. So on so forth. Let X be the union of ALL these infinitely many circles, then X is called the Hawaiian earring. It is a 1-dimensional differentiable set.

The tangent structure of X are very obvious. At any point $\boldsymbol{p} \in X$, if $\boldsymbol{p}$ is not the origin, then the tangent lines $T_{\boldsymbol{p}}(X)$ are the obvious tangent lines to the corresponding circle containing $\boldsymbol{p}$. If $\boldsymbol{p}$ is the origin, then the tangent line $T_{\mathbf{0}}(X)$ is the vertical line. So as far as we are concerned, this is a super nice thing to study.

The Hawaiian earring is famous because it gives many trouble to topologists, who want to understand all curves on this thing. Let me show you a weird curve that goes through ALL the circles. Note that the largest circle has circumference 2π. Then the next one has circumference π. Then the next one has circumference $\frac{\pi}{2}$. So the TOTAL circumference of all circles is $2\pi + \pi + \frac{\pi}{2} + \cdots = 4\pi$ via our knowledge on geometric series.

So imagin that I am holding a thread of length 4π. Then I can simple wind the thread around each circle, one by one, and I would have enough length to "eventually" cover all infinitely many circles. This gives a map $\gamma : [0, 4\pi] \to X$ which is surjective (all circles are covered), and continous (the thread is not "broken in two") and differentiable (no "sharp turns").

(Note that continuity at $t = 4\pi$ for γ is actually quite tricky, but can be proven. But we leave that to an actual geometry class. If you attempt to wind the same circle infinitely times using similar techniques, you shall fail, and the curve would NOT be continuous in the end.)

Another annoying thing about the Hawaiian earring is that, it is NOT a manifold. Take the orgin. NO neighborhood around the origin is "line-like", because any ball around the origin, no matter how small, must contain some even smaller "loop". The construction of the Hawaiian earring deviously sneaks in a loop into EVERY neighborhood of the origin. ☺

Example 12.1.8. Since we are at the topic of curves, have you heard of a space-filling curve? There is a surjective continuous map $\gamma : [0, 1] \to [0, 1]^2$, i.e., the curve segment actually fills up a square. Search for it yourself.

Luckily such a curve must be non-differentiable. We only do differentiable things, so we are fine. ☺

Example 12.1.9. All previous examples are 1-dimensional. Let us see a 2-dimensional example, which turns out NOT to be a differentiable set. Let X be the "cone" in $\mathbb{R}^3$, i.e., $z = \sqrt{x^2 + y^2}$, or the "upper half" of $x^2 + y^2 = z^2$. This is an upward-opening cone with the "tip" at the origin.

At any $\boldsymbol{p} \in X$ that is NOT the origin, the tangent plane $T_{\boldsymbol{p}}(X)$ is very obvious. However, at the origin, all possible "velocities" are all the directions along the cone itself, so $T_{\boldsymbol{p}}(X)$ is in fact X itself! In particular, they do NOT form a subspace. So X does NOT have a tangent plane at the origin. It is NOT a 2-dim differentiable set.

Sad.... However, do not despire. Let $X' = X - \{\mathbf{0}\}$, then X' is a 2-dim differentiable set, and we simply deal with X' whenever we want to deal with X. ☺

Example 12.1.10. Here is a weird example. Let $X = \mathbb{R} - \mathbb{Q} \subseteq \mathbb{R}$, the set of all irrational real numbers. At each $p \in X$, where could you go? NOWHERE! By removing $\mathbb{Q}$, we have made sure that it is disconnected everywhere. The only possible conitnuous curve $\gamma : [0,1] \to X$ is a constant curve, i.e., $\gamma(t)$ is the same point for all t, and $\gamma'(t) = 0$ always.

This is obviously NOT a manifold, and NOT a nice geometric object at all. Nevertheless, it is a 0-dim differentiable set, since all "tangent stuff" $T_{\boldsymbol{p}}(X)$ are zero-dimensional subspaces. ☺

Example 12.1.11. Let us do a "vanilla" example. Say $U \subseteq \mathbb{R}^m$ is "open", i.e., it is a union of open balls in $\mathbb{R}^m$. Here open balls means sets like $B_r(\boldsymbol{p}) := \{\boldsymbol{q} \in \mathbb{R}^n : \|\boldsymbol{p} - \boldsymbol{q}\| < r\}$, i.e., a ball of radius r around some point $\boldsymbol{p}$, without boundary.

For each point $\boldsymbol{p} \in U$, then $\boldsymbol{p}$ is in one of the open balls that make up U. In particular, we can find $r > 0$ such that $B_r(\boldsymbol{p}) \subseteq U$. In particular, we see that starting from $\boldsymbol{p}$, all directions are possible to make "tiny movements". So all vectors are tangent vectors, and $T_{\boldsymbol{p}}(X) = \mathbb{R}^m$ for all $\boldsymbol{p}$.

(For any $\boldsymbol{p} \in U$ and $\boldsymbol{v} \in \mathbb{R}^m$, try yourself and see if you can construct a curve γ such that $\gamma(0) = \boldsymbol{p}$ and $\gamma'(0) = \boldsymbol{v}$.)

In particular, all open subsets of $\mathbb{R}^m$ are m-dim differentiable sets, since all of its points have the entire $\mathbb{R}^m$ as the tangent space. ☺

Anyway, despite some weird looking examples, for these differentiable sets, at least the concept of tangent vectors and tangent spaces are well-defined. Now, let me show you what are derivatives through some examples.

Imagine that we have a map $f : X \to Y$ between differentiable sets. Then it will send points $\boldsymbol{p} \in X$ to a point $f(\boldsymbol{p}) \in Y$. Now, if I perform some tiny movement starting at $\boldsymbol{p}$ inside X, i.e., some tangent vector to X, then the image $f(\boldsymbol{p})$ would also change into something new, inducing a tangent vector to Y. This is the idea of a directional derivative.

Example 12.1.12. Let X be the unit circle in $\mathbb{R}^2$, and Y be the unit spher in $\mathbb{R}^3$. For each point $\begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix}$ on X, we can map it to $\begin{bmatrix}\frac{1}{\sqrt{2}}\cos\theta \\ \frac{1}{\sqrt{2}}\sin\theta \\ \frac{1}{\sqrt{2}}\end{bmatrix}$. So this is a map $f : X \to Y$.

(Tip: the formula is not that useful. We keep it here for to make things precise, but you'd better get used to the geometry. Understanding what goes where is more important than tracking the formula. Try to visualize this map.)

For any curve $\gamma : [0,1] \to X$, then f would immediately push it into a curve $f_*(\gamma) = f \circ \gamma : [0,1] \to Y$. For the sake of example, let us say we have $\gamma(t) = \begin{bmatrix}\cos(2t\pi) \\ \sin(2t\pi)\end{bmatrix}$. Then at "time" $t = \frac{1}{4}$, γ would induce a tangent vector $\gamma'(\frac{1}{4}) = \begin{bmatrix}2\pi\sin(\frac{\pi}{2}) \\ -2\pi\cos(\frac{\pi}{2})\end{bmatrix} = \begin{bmatrix}2\pi \\ 0\end{bmatrix}$ to the unit circle X at the point $\gamma(\frac{1}{4}) = \begin{bmatrix}0 \\ 1\end{bmatrix}$.

Now if any tangent vector to X is induced by γ (like in the example above), then in fact $f_*(\gamma)$ would also induce a tangent vector. Note that $f \circ \gamma$ is the curve $t \mapsto \begin{bmatrix}\frac{1}{\sqrt{2}}\cos(2t\pi) \\ \frac{1}{\sqrt{2}}\sin(2t\pi) \\ \frac{1}{\sqrt{2}}\end{bmatrix}$. Now at the same "time" $t = \frac{1}{4}$, $f_*(\gamma)$ would induce a tangent vector $(f \circ \gamma)'(\frac{1}{4}) = \begin{bmatrix}\sqrt{2}\pi \\ 0 \\ 0\end{bmatrix}$ at the point $f(\gamma(\frac{1}{4}))$.

We say that the tangent vector $\boldsymbol{v} = \begin{bmatrix}2\pi \\ 0\end{bmatrix}$ to X at $\boldsymbol{p}$ is pushed forward to the tangent vector $\begin{bmatrix}\sqrt{2}\pi \\ 0 \\ 0\end{bmatrix}$ to Y at $f(\boldsymbol{p})$.

Computations aside, hopefully this is graphically trivial. The function f simply "shrink" the circle by a factor of $\sqrt{2}$, and then put the shrunken loop on top of the unit sphere. So all tangent vectors "shrunk" by the same factor. ☺

Given a map $f : X \to Y$ between subsets $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$, if f is nice enough, then we usually should expect to have a well-defined map $f_*|_{\boldsymbol{p}} : T_{\boldsymbol{p}}X \to T_{f(\boldsymbol{p})}Y$, where tangent vectors are "pushed forward" according to how differential curves are pushed forward.

If you choose to understant tangent space $T_{\boldsymbol{p}}X$ as representing an "infinitesimally small neighborhood" around $\boldsymbol{p}$, then $f_*|_{\boldsymbol{p}}$ is basically the restriciton of f to this tiny neighborhood.

Furthermore, if $T_{\boldsymbol{p}}X$ and $T_{f(\boldsymbol{p})}Y$ are subspaces, then we hope that $f_*|_{\boldsymbol{p}}$ is linear. I.e., we hope that f behaves "linearly" around each infinitesimally small neighborhood. If a function f is "locally linear" like this, then we say f is differentiable.

Definition 12.1.13. *A map $f : X \to Y$ between differentiable sets is differentiable at $\boldsymbol{p} \in X$ if there is a linear map $L : T_{\boldsymbol{p}}X \to T_{f(\boldsymbol{p})}Y$, such that for any curve γ on X with $\gamma(0) = \boldsymbol{p}$ and $\gamma'(0) = \boldsymbol{v}$, then the derivative of $f_*(\gamma)(t)$ at $t = 0$ is $L\boldsymbol{v}$. We write $f_*|_{\boldsymbol{p}}$ for this linear map L.*

Example 12.1.14. Consider a function $\mathbb{R} \to \mathbb{R}$, say $f(x) = |x|$. Note that for any $x \in \mathbb{R}$, the tangent space $T_x\mathbb{R}$ is simlpy $\mathbb{R}$ itself. If we have a tiny change $\mathrm{d}x$ from x in the domain, and $x < 0$, then it would be mapped to a tiny change $-\mathrm{d}x$ at $-x$ in the codomain. This map $f_*|_x$ is simply "multiplication by -1", and it is linear. Similarly, if $x > 0$, then the map $f_*|_x$ is simply "multiplication by 1", and it is also linear.

However, at the point 0 in the domain, a tiny change forward and a tiny change backward would BOTH be mapped to a tiny change forward in the codomain. So $f_*|_0 : T_0\mathbb{R} \to T_0\mathbb{R}$ actually maps both 1 and -1 to 1. So it cannot be linear. ☺

Example 12.1.15. Consider a function $\mathbb{R}^2 \to \mathbb{R}^3$, such that $\begin{bmatrix} x \\ y \end{bmatrix}$ is mapped to $\begin{bmatrix} x \\ y \\ \sqrt{x^2+y^2} \end{bmatrix}$. Geometrically, we are folding the plane into the cone. Then $f_*|_{\boldsymbol{0}}$ would not be linear. Can you see this visually?

Intuitively, the "sharp corner" ruins differentiability. ☺

12.2 Cotangent vectors and pullbacks

Let us first see a curious special case of a differentiable map.

Example 12.2.1. Suppose $f : X \to \mathbb{R}$ is differentiable for some differentiable set X. Then we have an induced map $f_*|_{\boldsymbol{p}} : T_{\boldsymbol{p}}X \to T_{f(\boldsymbol{p})}\mathbb{R}$. However, note that at any point on $\mathbb{R}$, the corresponding tangent line is simply $\mathbb{R}$ itself. So we in fact have a map $f_*|_{\boldsymbol{p}} : T_{\boldsymbol{p}}X \to \mathbb{R}$. Hey, so $f_*|_{\boldsymbol{p}}$ is a "dual tangent vector" (or more traditionally, a cotangent vector, or sometimes a covector for short)!

Things are even more uncanny than you think. If we have a "tiny" change inside of X, i.e., a tangent vector $\boldsymbol{v}$ to X, then this dual vector would give us a corresponding "tiny" change in the codomain $\mathbb{R}$, i.e., the change in the f-value. This is exactly how the "directional derivative" is defined.

We sometimes also write $f_*|_{\boldsymbol{p}}$ as $\mathrm{d}f|_{\boldsymbol{p}}$ which literally reads as "change in f-value at $\boldsymbol{p}$". For the directional derivative, we have many notations but they are all the same, like $\nabla_{\boldsymbol{v}} f(\boldsymbol{p})$ or $\mathrm{d}f|_{\boldsymbol{p}}(\boldsymbol{v})$. They mean the same thing (but as you can see, one focus on changing $\boldsymbol{p}$ while the other is on changing $\boldsymbol{v}$).

Finally we have formulas such as $\lim_{t\to 0} \frac{f(\boldsymbol{x}+t\boldsymbol{v})-f(\boldsymbol{x})}{t}$. What is $\boldsymbol{x} + t\boldsymbol{v}$ here? It is simply a "curve" at $\boldsymbol{x}$ with tangent vector $\boldsymbol{v}$. We simply picked this curve, and see how a tiny change along this curve would induce a tiny change in f-value.

In general, if we have any curve γ such that $\gamma(0) = \boldsymbol{x}$ and $\gamma'(0) = \boldsymbol{v}$, then the directional derivative should be defined as $\nabla_{\boldsymbol{v}} f := \lim_{t\to 0} \frac{f(\gamma(t))-f(\gamma(0))}{t}$. The traditional formula is simply a special case of this when we pick γ to be a straight line. ☺

Last section we were mainly studying the "continuous elements" of X, i.e., curves $\mathbb{R} \to X$. This section, we study the "dual" of these, i.e., real functions on X, $f : X \to \mathbb{R}$. Just as curves induce tangent vectors, real functions would induce cotangent vectors. And just as maps pushforward tangent vectors, they would pull back cotangent vectors.

But we are getting ahead of ourselves. Let us see some more examples.

Example 12.2.2. Hmmm, an intersting notion in calculus is $\mathrm{d}x$. What is $\mathrm{d}x$?

Well, on the differential set $X = \mathbb{R}^2$, we have a map $x : \mathbb{R}^2 \to \mathbb{R}$ sending each point to their x-component. Then $\mathrm{d}x$ at each point is simply a covector, sending each "tiny change" (tangent vector) in the domain $\mathbb{R}^2$ to the corresponding change in x-value. In particular, you can see that $\mathrm{d}x|_{\boldsymbol{p}}(\boldsymbol{v}) = v_x$, the x-coordinate of $\boldsymbol{v}$.

Now given a differentiable curve γ on X, say $\gamma : [0,1] \to X$ with $\gamma(0) = \boldsymbol{a}$ and $\gamma(1) = \boldsymbol{b}$, suppose we want to do the integral $\int_\gamma \mathrm{d}x$. What does this mean? Well, on each point of the curve, the covector dx at $\gamma(t)$ would evaluate the tangent vector $\gamma'(t)$. I "sum over" all such evaluations (Riemann-type sum), and I would get the "total change in x-values along γ" as a result.

So calculations goes like $\int_\gamma \mathrm{d}x := \int_0^1 \mathrm{d}x|_{\gamma(t)}(\gamma'(t))dt$. Note that $\mathrm{d}x|_{\gamma(t)}(\gamma'(t))$ is the directional derivative of x along $\gamma'(t)$, so it is simply the x-coordinate of $\gamma'(t)$.

So we have $\int_\gamma \mathrm{d}x := \int_0^1 \mathrm{d}x|_{\gamma(t)}(\gamma'(t))dt = \int_0^1 \gamma_x'(t)dt = \gamma_x(t)|_0^1 = \gamma_x(1) - \gamma_x(0) = b_x - a_x$, the difference in x-coordinates. ☺

In general, we would like to think of $\mathrm{d}f$ as a "covector field", i.e., each point $\boldsymbol{p}$ has a corresponding covector $\mathrm{d}f|_{\boldsymbol{p}}$. Here is a very funny observation, which makes a "intuitive equality" into an actual concrete equality. (Btw, people also call these covector fields "differential 1-form".)

Proposition 12.2.3. *For any continuously differentiable $f : \mathbb{R}^2 \to \mathbb{R}$, then $\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y$.*

Here I have three functions, $f : \mathbb{R}^2 \to \mathbb{R}, x : \mathbb{R}^2 \to \mathbb{R}, y : \mathbb{R}^2 \to \mathbb{R}$. They induce a covector field $\mathrm{d}f, \mathrm{d}x, \mathrm{d}y$. I am claiming that, at each point $\boldsymbol{p}$, $\mathrm{d}f$ is a linear combination of $\mathrm{d}x, \mathrm{d}y$. The precise combination may change depending on $\boldsymbol{p}$. At different points, $\mathrm{d}f$ might be a different linear combination of $\mathrm{d}x, \mathrm{d}y$, hence the coefficients of linear combination $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are functions whose value depends on $\boldsymbol{p}$.

Proof. Fix a point $\boldsymbol{p} \in \mathbb{R}^2$. Note that $T_{\boldsymbol{p}}\mathbb{R}^2 = \mathbb{R}^2$, so covectors at $\boldsymbol{p}$ are all row vectors. Then it is easy to see that $\mathrm{d}x|_{\boldsymbol{p}}$ is the row vector $\begin{bmatrix}1 & 0\end{bmatrix}$, $\mathrm{d}y|_{\boldsymbol{p}}$ is the row vector $\begin{bmatrix}0 & 1\end{bmatrix}$.

What about $\mathrm{d}f$? Note that given a tangent vector $\boldsymbol{v} = \begin{bmatrix}v_x \\ v_y\end{bmatrix}$, the directional derivative is $\frac{\partial f}{\partial x}(\boldsymbol{p})v_x + \frac{\partial f}{\partial y}(\boldsymbol{p})v_y$, so $\mathrm{d}f|_{\boldsymbol{p}} = \begin{bmatrix}\frac{\partial f}{\partial x}(\boldsymbol{p}) & \frac{\partial f}{\partial y}(\boldsymbol{p})\end{bmatrix}$.

So we see that $\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y$. □

Basically the heavy lifting was already done by your calculus class. I simply cheat by using the directional derivative formula. But the realization that $\mathrm{d}x, \mathrm{d}y, \mathrm{d}f$ are covector fields are very handy sometimes.

Example 12.2.4. For example, given a differential curve γ from $\begin{bmatrix}1\\2\end{bmatrix}$ to $\begin{bmatrix}3\\4\end{bmatrix}$, how to calculate $\int_\gamma (y\,\mathrm{d}x + (x+1)\,\mathrm{d}y)$? Well, we have $y\,\mathrm{d}x + (x+1)\,\mathrm{d}y = \mathrm{d}(xy) + \mathrm{d}y$. So since the whole thing is a (Riemann) sum of covectors evaluating tangent vectors, everything is linear, so $\int_\gamma (y\,\mathrm{d}x - x\,\mathrm{d}y) = \int_\gamma \mathrm{d}(xy) + \int_\gamma \mathrm{d}y = (xy)|_{\gamma(0)}^{\gamma(1)} + y|_{\gamma(0)}^{\gamma(1)} = 10 + 2 = 12$. ☺

Here is another handy result.

Proposition 12.2.5. *All covector fields on $\mathbb{R}^2$ are $f\,\mathrm{d}x + g\,\mathrm{d}y$ for some functions $f, g : \mathbb{R}^2 \to \mathbb{R}$.*

Proof. Pick any covector field α. At each point $\boldsymbol{p}$, obviously the covectors $\mathrm{d}x|_{\boldsymbol{p}} = \begin{bmatrix}1 & 0\end{bmatrix}$ and $\mathrm{d}y|_{\boldsymbol{p}} = \begin{bmatrix}0 & 1\end{bmatrix}$ span the entire dual space. So $\alpha|_{\boldsymbol{p}}$ is a linear combination of $\mathrm{d}x|_{\boldsymbol{p}}$ and $\mathrm{d}y|_{\boldsymbol{p}}$ for each $\boldsymbol{p}$. So we are done. □

12.3 Integration on covector fields

Your calculus class should teach you how to do this. But I shall attempt to teach you how to see this.

How to visualize a vector field? This is easy. We just think of it as "tiny arrows" at each point of the domain. But how should one visualize covector field? First let us try to visualize a single dual vector.

Example 12.3.1. In the space $\mathbb{R}^n$, say $n = 3$, we may think of a vector as an "arrow", i.e., a one-dimensional object with a real number associated to it. For the vector $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$, we can draw an arrow from the origin to the end point $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$. It can be thought of as a one-dimensional subspace (the direction) with a number (the length).

Then in an analogous manner, we may think of a dual vector as an $(n-1)$-dimensional object with a number associated to it. The idea is that, for the row vector $\begin{bmatrix} a & b & c \end{bmatrix}$, we think of the equation $ax+by+cz = 0$, which gives an $(n-1)$-dimensional hyperplane (in fact a subspace).

If you think about this, it makes natural sense. A dual vector is, first and foremost, a linear map $\alpha : V \to \mathbb{R}$. The row vector $\begin{bmatrix} a & b & c \end{bmatrix}$ literally refers to the linear map that sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $ax + by + cz$. So far so good.

However, note that the equations $x + 2y + 3z = 0$ and $2x + 4y + 6z = 0$ refers to the SAME hyperplane! So a dual vector is not just a hyperplane. It is a hyperplane with a number (density)!

Consider the dual vector $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, for example. Then think of this as such: we are layering the space $\mathbb{R}^3$ with hyperplanes parallel to the plane $z = 0$. It has "density one", which means that we have one such plane in each vertical unit distance. Then why would $\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$ be sent to 5? Because it puctures through 5 layers. (Similarly, the vector $\begin{bmatrix} 1 \\ 1 \\ 0.5 \end{bmatrix}$ be sent to 0.5, because it puctures through half a layer.)

Now consider the dual vector $\begin{bmatrix} 0 & 0 & 4 \end{bmatrix}$. Then think of this as such: we are layering the space $\mathbb{R}^3$ with hyperplanes parallel to the plane $z = 0$. It has "density four", which means that we have FOUR such plane in each vertical unit distance. Then why would $\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$ be sent to 20? Because it now puctures through 20 layers.

This phenomenon is sometimes called a foliation. ("Falling of leaves", which makes layers and layers of leaves....) A dual vector $\begin{bmatrix} a & b & c \end{bmatrix}$ can be though of as a foliation of $\mathbb{R}^3$ by planes parallel to $ax+by+cz = 0$, with a density $\sqrt{a^2 + b^2 + c^2}$. ☺

In general, for any dual vector $\alpha : V \to \mathbb{R}$, we may think of α as the hyperplane $\mathrm{Ker}(\alpha)$ with a number (density). It corresponds to a picutre of foliation of the whole space V, and it measures vectors by counting how many layers in the foliation were punctured.

(Note that if $\alpha = 0$, then $\mathrm{Ker}(\alpha)$ is no longer a hyperplane. It is the whole space (an n-dim object). You would never puncture through the whole space, so all vectors are measured to be zero. This corresponds to the fact that $\mathbf{0}$ no longer corresponds to any 1-dim object, but instead becomes a point (a 0-dim object). So it cannot puncture through anything. It is always measured to be zero.)

Now let us look at covector fields. Say we are looking at covector fields on $\mathbb{R}^n$. Then we have a covector at each point, i.e., a "tiny $(n-1)$-dim hyperplane' with a number (density) at each point. By connecting these "tiny $(n-1)$-dim hyperplanes", we can obtain a foliation of $\mathbb{R}^n$ by hypersurfaces with densities everywhere.

Example 12.3.2. Suppose $n = 2$. Then covector fields are supposed to look like a foliation of $\mathbb{R}^2$ by hypersurfaces (i.e., curves).

Say we have a function $f : \mathbb{R}^2 \to \mathbb{R}$, and it induces a covector field $\mathrm{d}f$. What is the corresponding foliation of curves? It is exactly the level curves!

Consider $f(x, y) = x^2 + y^2$. Then the covector field is $\mathrm{d}f = 2x\,\mathrm{d}x + 2y\,\mathrm{d}y$. So at each point $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$,

the corresponding covector is simply $\begin{bmatrix}2x & 2y\end{bmatrix}$. It kills vectors perpendicular to the "radial direction", and its kernel is always tangent to some circle around the origin.

If we connect these "tiny line segments tangent to some circle around the origin", then we would actually get these circles around the origin. We have a foliation of $\mathbb{R}^2$ by concentric curcles around the origin, i.e., the level curves of f. Note that $\begin{bmatrix}2x & 2y\end{bmatrix}$ grows larger and larger as we move away from the origin, so the "densities" of these level curves would increas as we move away from the origin. If you draw these level curves at constant "f-speed", e.g., say we draw $f = 1, f = 2, f = 3, \dots$ and so on, then you can see how the circles get denser and denser away from the origin.

Consider $f(x, y) = \sqrt{x^2 + y^2}$, which induces a covector field $\mathrm{d}f = \frac{x\,\mathrm{d}x}{\sqrt{x^2+y^2}} + \frac{y\,\mathrm{d}y}{\sqrt{x^2+y^2}}$ on the geometric object $\mathbb{R}^2 - \{\mathbf{0}\}$. This is also a foliation of the plane by concentric circles, but the densities are now different. Note that the covector $\begin{bmatrix}\frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}}\end{bmatrix}$ is always a unit covector, and hence the "density" is the same everywhere. If you draw these level curves at constant "f-speed", e.g., say we draw $f = 1, f = 2, f = 3, \dots$ and so on, then you can see how the "distances" between layers stay constant.

For whatever function f on X, its level curves is the foliation of X, which is a graphic representation of the covector fied $\mathrm{d}f$. And if we integrate a curve $\gamma : [0, 1] \to X$ on this covector field, then we are asking ourselves: how many layers (level curves) are punctured by the curve γ? And the answer is $\int_\gamma \mathrm{d}f = f(\gamma(1)) = f(\gamma(0))$.

In particular, if γ is a closed curve, then $\int_\gamma \mathrm{d}f$ is always zero, because whatever γ would puncture, it would eventually "unpuncture" so that it can be a closed curve. It also makes a lot of sense: if you end where you started, then f has no change, so $\mathrm{d}f$ integrates to zero. ☺

What if our covector field is NOT induced by a function? This is perfectly possible. Then something weird might happen.

Example 12.3.3 (A non-conservative vector field)**.** Consider the following foliation of $\mathbb{R}^2$. The foliation curves are actually rays shooting away from the origin. Which covector field does it represent?

At each point $\boldsymbol{v} = \begin{bmatrix}x\\y\end{bmatrix} \in \mathbb{R}^2$, we wish the kernel to be the radial direction, i.e., the corresponding covector should be parallel to $\begin{bmatrix}-y & x\end{bmatrix}$. What about density? Well, suppose going around the unit circle (curve of length 2π) would puncture 2π layers. Then you see that, around any circle around the origin, we would always puncture the SAME amount of layers, i.e., 2π layers. So the density must drop as we move away from the origin. By the circumference formula of circles, we see that the density of the foliation at $\boldsymbol{v}$ should be $\frac{1}{\|\boldsymbol{v}\|}$.

So the desired covector field should be $\omega = \frac{-y\,\mathrm{d}x}{x^2+y^2} + \frac{x\,\mathrm{d}y}{x^2+y^2}$. ☺

Example 12.3.4 (Winding Number)**.** Consider the covector field $\omega = \frac{-y\,\mathrm{d}x}{x^2+y^2} + \frac{x\,\mathrm{d}y}{x^2+y^2}$ defined on $\mathbb{R}^2 - \{\mathbf{0}\}$.

Consider the integration of this covector field around the circle around the origin with radius r, $\gamma : [0, 2\pi] \to \mathbb{R}^2 - \{\mathbf{0}\}$ such that $\gamma(\theta) = \begin{bmatrix}r\cos\theta & r\sin\theta\end{bmatrix}$. To do $\int_\gamma \omega = \int_\gamma (\frac{-y\,\mathrm{d}x}{x^2+y^2} + \frac{x\,\mathrm{d}y}{x^2+y^2})$, note that on the unit circle, we always have $x^2 + y^2 = 1$. So we can simplify this to $\frac{1}{r^2}\int_\gamma (-y\,\mathrm{d}x + x\,\mathrm{d}y)$.

How to do such an integration? We integrate θ from 0 to 2π, and at each point we use the corresponding covector to evaluate the corresponding velocity vector for γ, i.e., $\int_0^{2\pi} (-y\,\mathrm{d}x + x\,\mathrm{d}y)|_{\gamma(\theta)}(\gamma'(\theta))\,\mathrm{d}\theta$.

Now $\gamma'(\theta) = \begin{bmatrix}-r\sin\theta & r\cos\theta\end{bmatrix}$, and $-y\,\mathrm{d}x + x\,\mathrm{d}y$ at $\gamma(\theta)$ is $\begin{bmatrix}-r\sin\theta & r\cos\theta\end{bmatrix}$. So $(-y\,\mathrm{d}x + x\,\mathrm{d}y)|_{\gamma(\theta)}(\gamma'(\theta)) = r^2\sin^2 + r^2\cos^2 = r^2$. So $\int_\gamma \omega = \frac{1}{r^2}\int_0^{2\pi} r^2\,\mathrm{d}\theta = 2\pi$. Hooray!

In fact, let us forget about the calculations. Take any closed curve. If it winds around the origin once (counter-clockwise), then it shall ALWAYS puncture through the same amount of layers as did the unit circle. Hence $\int_\gamma \omega = 2\pi$.

Given any closed curve γ, we define $\frac{1}{2\pi}\int_\gamma \omega$ to be the ***winding number*** of γ. It is quite amazing that $\frac{1}{2\pi}\int_\gamma (\frac{-y\,\mathrm{d}x}{x^2+y^2} + \frac{x\,\mathrm{d}y}{x^2+y^2})$ is always an integer! And it counts the number of times the curve winds around the origin (where positive = counter-clockwise and negative = clockwise).

The covector field ω is also sometimes written as $\mathrm{d}\theta$, because it measures the change in angles. You can loosely think of θ as a "function" that sends vectors to their "angles". However, θ on $\mathbb{R}^2$ is NOT a well-defined function. A point at angle π also has angle 3π. There is NO function θ. ☺

Note that, in the case of $\mathrm{d}f$ for some f, the level curves are always closed curves or bi-infinite curves (extends on both sides of the curves infinitely). However, in the case of $\mathrm{d}\theta$ as above, the foliation curves are NOT bi-infinite. They are rays, i.e., only infinite on one side of the curve.

Here is another covector field whose foliation is made of rays.

Example 12.3.5 (The "middle-aged man" covector field)**.** The covector field $\mathrm{d}\theta$ has an annoying denominator. Let us throw it away. Say $\omega = -y\,\mathrm{d}x + x\,\mathrm{d}y$. Hooray! It is now defined on all of $\mathbb{R}^2$. Now how can we visualize this new covector field?

Well, again the kernel of $\begin{bmatrix}-y & x\end{bmatrix}$ is always the radial direction. So it looks a bit like $\mathrm{d}\theta$. However, the densities are different. ω has zero density at the origin, and it become denser and denser as we move away from the origin!

So you have more and more rays shooting away from the origin. Unlike the case of $\mathrm{d}\theta$, where all rays started at the origin, now for ω, all the rays must start somewhere away from the origin. And we have more and more new "starting rays" as we move away from the origin, creating a denser and denser foliation.

It looks like the hair of a "middle-aged man", with less and less hair towards the center.... ☺

Example 12.3.6 (Rays to the right)**.** Consider the covector field $x\,\mathrm{d}y$. The kernel of $\begin{bmatrix}0 & x\end{bmatrix}$ are in the horizontal direction, and the density increases as we move to the right. So the covector fields actually looks like rays shooting to the right. If we move to the right at a constant speed, then the density ("number of rays") increases (new rays are emerging/starting) at a constant speed.

This means that we have the same number of new curves at every region of the same area size. In particular, the starting points of these rays are dots in $\mathbb{R}^2$, and they are uniformly distributed all over $\mathbb{R}^2$ with "unit density".

Now let γ be a curve that goes around the unit square in $\mathbb{R}^2$ counterclockwise, and consider $\int_\gamma x\,\mathrm{d}y$. How many rays does it puncture?

Well, for the rays above and below the square, they do not touch the curve, so they do not contribute. For the rays started to the right of the square, they also do not touch the curve. For the rays started to the left of the square, they are punctured but then "unpunctured" by the curve, so they do not contribute.

And we have come to the following conclusion: the only thing that contribute to $\int_\gamma x\,\mathrm{d}y$ are rays STARTING INSIDE the square! In fact, this is exactly what $\int_\gamma x\,\mathrm{d}y$ is counting.

How many rays started in the unit square? Well, we can just count the number of starting points of these rays inside the square. So this is the area of the square.

Now look at the famous special case of Green's theorem: if ∂M is the boundary of a region M, then $\int_{\partial M} x\,\mathrm{d}y = \mathrm{Area}(M)$.

In general, the Green's theorem is pointing out that $\int_{\partial M}\omega$ is asking the number of "starting points" of rays are inside of M, which is the integration over the inside of M, i.e., $\int_M \mathrm{d}\omega$. Here you can interpret d as "taking boundarys of each foliation curve" to get a "dot field with density", on which you can integrate. ☺

Remark 12.3.7. *(Optional proof of Green's theorem.)*

As Green's theorem has told us, the formula for $\mathrm{d}\omega$ *is that* $\mathrm{d}(f\,\mathrm{d}x + g\,\mathrm{d}y) = (g_x - f_y)\,\mathrm{d}x\,\mathrm{d}y$*. You can actually also establish this formula using the idea we have presented here. Think of this as* $\mathrm{d}(f\,\mathrm{d}x + g\,\mathrm{d}y) = \mathrm{d}(f\,\mathrm{d}x) + \mathrm{d}(g\,\mathrm{d}y)$ *and do each independently. Then* $g\,\mathrm{d}y$ *also corresponds to a foliation of rays going to the right, and the new rays emerge at a speed of* g_x*. So* $\mathrm{d}(g\,\mathrm{d}y) = g_x\,\mathrm{d}x\,\mathrm{d}y$*.*

Similar argument can show that $\mathrm{d}(f\,\mathrm{d}x) = -f_y\,\mathrm{d}x\,\mathrm{d}y$*. Note that the negative sign is due ot orientation. Imagin that you are cutting the foliation curves for* $f\,\mathrm{d}x$ *counter-clockwise, then higher edge actually cuts in the negative* x*-direction, while lower edge cuts in the positive* x*-direction. So if we have more curves as we move upward, then it contribute negatively towards the integration.*

So far, all examples are in $\mathbb{R}^2$. If we were in $\mathbb{R}^3$, then the foliations for a covector field would by via surfaces. And in $\mathbb{R}^n$, it would be hyper-surfaces.

In whatever case, $\mathrm{d}f$ always have a foliation made by the level-hypersurfaces. These are always closed-up or extending infinitely, i.e., no boundary. And for covector fields which are NOT derived from a function, then some foliation hypersurfaces might have boundaries. (Just like the "rays" which now has starting points or ending points.)

(And we would always have the formula $\int_{\partial M} \omega = \int_M \mathrm{d}\omega$, which is simultaneously Green's theorem and Stoke's theorem and divergence theorem. here ∂ means taking the boundary of M, while $\mathrm{d}\omega$ means we take the boundary of each "foliating thing" for ω. For example, suppose ω is a foliation of surfaces in $\mathbb{R}^3$, and a disc M is cutting the foliation of closed curves $\mathrm{d}\omega$. How many closed curves were cut? Well, this equal to the number of foliation surfaces ω cut by the curve ∂M. Think on this if you like.)

Calculations will become more annoying though, so we stop here.

Chapter 13

Tensor

13.1 Motivating Examples

Maybe some of you have thought about something like this. (If not, then maybe you don't play enough)

Example 13.1.1. A matrix is, somewhat crudely speaking, a rectangular array of numbers. It is a two dimensional array of numbers. What if we have a "cubic array" of numbers (a three dimensional array of numbers)? It would feel like a "box matrix". Say, maybe we can have a $2 \times 4 \times 3$ array of numbers.

How would we write such a "box matrix"? Well, we can do it one layer at a time. We have three layers of 2×4 matrices A_1, A_2, A_3. (Draw a stack of three papers and write A_1, A_2, A_3 on them respectively, if you like.)

So far so good. But we need to give it a meaning. How would such a thing acts on a vector?

There are three possible ways. First of all, given a vector $\boldsymbol{v} \in \mathbb{R}^4$, we can multiply each layer matrix to $\boldsymbol{v}$. Then we get a resulting 2×3 matrix $\begin{bmatrix} A_1\boldsymbol{v} & A_2\boldsymbol{v} & A_3\boldsymbol{v} \end{bmatrix}$. As you can see, a $2 \times 4 \times 3$ box matrix sends vectors in $\mathbb{R}^4$ to 2×3 matrices.

But if $\boldsymbol{v} \in \mathbb{R}^2$, then we can apply the row vector $\boldsymbol{v}^{\mathrm{T}}$ to each layer matrix, and get a 3×4 matrix $\begin{bmatrix} \boldsymbol{v}^{\mathrm{T}}A_1 \\ \boldsymbol{v}^{\mathrm{T}}A_2 \\ \boldsymbol{v}^{\mathrm{T}}A_3 \end{bmatrix}$. This is a different way of sending a vector (in $\mathbb{R}^2$) to a matrix (of dimension 3×4).

Finally, we can simply do a linear combination of the three layers according to the coordinates of a vector $\boldsymbol{v} \in \mathbb{R}^3$. Say if $\boldsymbol{v} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$, then we do the linear combination of layer matrices, and get a 2×4 matrix $2A_1 + 3A_2 + 4A_3$. This is yet a different way of sending a vector (in $\mathbb{R}^3$) to a matrix (of dimension 2×4).

If you draw the array of numbers, you shall see that the three ways above correspond to the three ways of collapsing a 3D object into 2D. From a 3D array, we can collapse the columns of each matrix (the first way) to get a 2D array, or we can collapse the rows of each matrix (the second way) to get a 2D array, or we can collapse vertically all the layers (the third way) to get a 2D array.

So, we have three ways to collapse a 3D array. Which way should we choose? Well, we choose ALL OF ABOVE!

Consider a 2D array, i.e., a matrix A. Rather than think of it as a linear map $\boldsymbol{v} \mapsto A\boldsymbol{v}$, we can think of it as a bilinear map $(\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v}^{\mathrm{T}}A\boldsymbol{w}$.

Similarly, for a cube array, say the $2 \times 4 \times 3$ box array B with layers A_1, A_2, A_3, then we think of it as a TRI-LINEAR map, sending $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ to $\begin{bmatrix} \boldsymbol{u}^{\mathrm{T}}A_1\boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}}A_2\boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}}A_3\boldsymbol{v} \end{bmatrix} \boldsymbol{w}$. Here $\boldsymbol{u} \in \mathbb{R}^2, \boldsymbol{v} \in \mathbb{R}^4, \boldsymbol{w} \in \mathbb{R}^3$, and the thing to the left of $\boldsymbol{w}$ is a row vector, with entries $\boldsymbol{u}^{\mathrm{T}}A_i\boldsymbol{v}$ for $i = 1, 2, 3$. So we have a map $B : \mathbb{R}^2 \times \mathbb{R}^4 \times \mathbb{R}^3 \to \mathbb{R}$ such that it is tri-linear, i.e., linear in each input if we fix the other two inputs. ☺

As you can see there, the generalization of a matrix (which corresponds to bilinear maps) is a MULTI-LINEAR map. In the example above I choose $\mathbb{R}$ as the codomain. But in general, we can pick any vector

space as the codomain.

Definition 13.1.2. *Given vector spaces $V_1, \dots, V_n$ and W, then the "n-input" map $M : V_1 \times \cdots \times V_n \to W$ is multilinear if it is linear in each input.*

(Technically, mathematicians call an "n-input" map as an n-ary map. Unary means one input, i.e., a regular function as we are used to. Binary means two input. Trinary means three input, and n-ary means n inputs. There are also related weird names. For example, the "arity" of a map is the number of inputs.... A constant map is sometimes also called a 0-ary map, since it does not really need any input.) (We do not need these Jargons though.)

Here are some examples of multilinear maps (tensors), as motivations to study tensors.

Example 13.1.3 (Real number multiplication)**.** The multiplication map $M : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by sending (a, b) to ab is a multilinear map, in fact a bilinear map. This seems pretty straight forward. ☺

Example 13.1.4 (Cross Product)**.** By now you must know what a cross product is. Given two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$, their cross product $\boldsymbol{v} \times \boldsymbol{w}$ is a unique vector such that $\det \begin{bmatrix} \boldsymbol{v} & \boldsymbol{w} & \boldsymbol{v} \times \boldsymbol{w} \end{bmatrix}$ is positive (they make a "right-handed" system), $\boldsymbol{v} \times \boldsymbol{w}$ is perpendicular to $\boldsymbol{v}$ and $\boldsymbol{w}$, and $\|\boldsymbol{v} \times \boldsymbol{w}\|$ is the area of the parallelogram made by $\boldsymbol{v}$ and $\boldsymbol{w}$. It also has a formula that I'm sure you have memorized.

And this is a bilinear map $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$. ☺

Example 13.1.5 (Determinant)**.** The determinant map send a matrix to a number. It is NOT linear, since in general $\det(kA)$ is NOT the same as $k \det(A)$. Rather, it is $k^n \det(A)$ if A is $n \times n$.

However, it is multilinear! Do NOT think of A as a matrix. Rather, think of it as n column vectors. Then the determinant map sends n vectors $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ to a number, and it is linear in each input (since determinant is linear in each column). So it is a multilinear map. (n-linear map.)

To exploit the geometric meaning, higher-dimensional volumn is multilinear in the edge vectors. Can you see this geometrically?

Here's a challenge problem for you to do: For any n-linear map from $\mathbb{R}^n \times \cdots \times \mathbb{R}^n$ to $\mathbb{R}$, if it sends the identity matrix to 1 and it is "alternating", i.e., swapping two inputs would negate the resulting value, then it MUST be the determinant map.

So the determinant map is in fact the UNIQUE alternating multilinear map on $\mathbb{R}^n$ that sends the identity matrix to one. Ha! ☺

A tensor is basically a multilinear map where all vector spaces involved are the same space. Here are some examples.

Example 13.1.6 (Tri-focal tensor)**.** Say I take a photo of an object, say a ball. And by looking at the picture, we can see a fly at the center of the ball. However, since the picture is 2D, I do not know how far the fly is from me. It's simply a dot in the picture, and it could be touching the ball, or maybe it is staying mid-air. Just by looking at the picture, I do NOT know the 3D location of the fly, and I only know that it rests on the line connecting my camera to the center of the ball. By taking a picutre, I am squashing a 3D scene into a 2D picture, and I have "one-dimension" of ambiguity.

To figure out the location of the fly, I take another photo at a different angle. Now from the new picutre, I also have "one-dimension" of ambiguity. By looking at the new picture, again I only know that the fly lies on some new line. However, by looking at both pictures, I can intersect the two "lines of ambituity", then I would obtain the actual 3D location of the fly.

In particular, if I want to take a third photo at yet another different angle, then I don't even need to take the photo. I can already predict where the fly shall appear on the third photo.

The tri-focal tensor thus refers to the following process. We fix a scene to take photos, and we fix three different locations for cameras. Say the three photos are X, Y, Z. Then given X, Y, I can already infer the location of everything, so even without looking at Z, I can infer what Z shall look like.

The map $(X, Y) \mapsto Z$ is multilinear. (If the fly moves in a straight line, then it shall move in a straight line in all pictures X, Y, Z simultaneously.) This is the tri-focal tensor, and it is used for photography, rental websits (to give "virtual tours" online), movie making (make two photos of a 3D animated object, then we can compute the photo of the animated object from all angles, etc.). ☺

Example 13.1.7 (Cauchy Stress tensor)**.** In engineering, Cauchy stress tensor is used to describe some internal tension within an object. How to study tension forces inside an object? Well, the answer is to use tensors. Not surprising, huh? (This is the reason for the name.)

Given a cylinder, let us pull both ends. Then it shall bread along some cross section. As you can imagine, before the break, cross sections within the cylinder must be suffering from forces to pull them apart!

Now imagine that I have a chocolate bar. I can snap it into two. It will also break along some cross section. As you can imagine, before the break, cross sections within the cylinder must be suffering from forces to break them. However, the direction of the forces are different! Now the forces are not apart, but rather trying to shear apart. (I.e., a sliding-like tension.)

We now attempt to describe stress at a point $\boldsymbol{p} \in X$. First we need to describe a cross section. For any pair of tangent vectors $\boldsymbol{u}, \boldsymbol{v}$ at $\boldsymbol{p}$, it would span some "cross section". Then this cross section would face a stress force, which is another tangent vector $\boldsymbol{w}$ at $\boldsymbol{p}$. So we have a multilinear map $(\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{w}$, which goes from $T_{\boldsymbol{p}}(X) \times T_{\boldsymbol{p}}(X)$ to $T_{\boldsymbol{p}}(X)$. This is the stress tensor at $\boldsymbol{p}$. What we would end up having is a tensor field on X (so that you can integrate or manipulate etc.).

Just like the shape of X might influence possibilities of continuous vector fields, the shape of X would also influence possiblilities of continuous tensor fields. So by designing the shape of X carefully, stress at some point against some cross-sections would be maximal, and the object is more likely to break that way. (E.g., a chocolate bar would snap along the "dents".)

Note that since most engineering applications are in $\mathbb{R}^3$, we do not actually need to use a pair of vectors to describe a cross section. We can just pick a normal vector. Then the Cauchy stress tensor would be sending $\boldsymbol{n}$ to $\boldsymbol{w}$, i.e., it is some linear map from $T_{\boldsymbol{p}}(X)$ to $T_{\boldsymbol{p}}(X)$. Then actually it can be described by a matrix. (Which is apprarently what all the engineering textbooks are doing.... A matrix is less scary than a tensor I guess.) ☺

Example 13.1.8 (Curvature)**.** General relativity states that gravity can be modeled as curvaturs of a 4-dimensional geometric object. How can we define curvature on a high-dimensional object?

Suppose we are in $\mathbb{R}^2$. I hold my arm towards the north. Now I walk along the unit square, all the way back to my starting position, while trying to KEEP my arm pointing to the North. This is called a "parallel translation". And when I finish my walk, my arm is still pointing the North. So far so good.

Suppose now that we are on the unit sphere. Say I start at somewhere on the equator, and I hold my arm towards the direction of the north pole. I move towards the north pole while holding my arm in the same direction, and then turn my body by 90 degree without changing my arm direction. Then I walk straight until I hit the equator again while holding my arm in the same direction, and turn my body to face my starting point without changing arm direction. Then I walk back to my starting point while holding my arm in the same direction. Wait! Now my arm is no longer pointing at the north pole! What happened?

The idea of curvature is based on this. Suppose I am standing inside an object X, at some point $\boldsymbol{p} \in X$. If I hold my arm in some direction (pick a tangent vector $\boldsymbol{v}$ at $\boldsymbol{p}$), and perform a "parallel translation" while walking around a loop, my arm might end up at a different direction (a resulting tangent vector $\boldsymbol{w}$ at $\boldsymbol{p}$). If X is flat without curvature, then I expect $\boldsymbol{v} = \boldsymbol{w}$. But if X is not flat but curved, then around certain loops, we might have $\boldsymbol{v} \neq \boldsymbol{w}$. The difference between $\boldsymbol{v}$ and $\boldsymbol{w}$ is the curvature.

To define this more rigorously, first we pick two tangent vectors $\boldsymbol{u}, \boldsymbol{v}$ at $\boldsymbol{p}$. Then we have a tiny parallelogram made by $t\boldsymbol{u}, t\boldsymbol{v}$ for any number t. If t is tiny enough, then we can approximately imagine that this is a tiny parallelogram inside of X. We pick a third tangent vector $\boldsymbol{w}$ at $\boldsymbol{p}$, which is my "arm direction". We perform this parallel transport of $\boldsymbol{w}$ while walking around the tiny parallelogram loop made by $t\boldsymbol{u}, t\boldsymbol{v}$, and my arm shall end up at some direction $\boldsymbol{w}_t$ depending on t.

If t is tiny, then the differences between $\boldsymbol{w}$ and $\boldsymbol{w}_t$ shall also be tiny. So we can take $\lim_{t\to 0} \frac{\boldsymbol{w}_t - \boldsymbol{w}}{t}$ as the curvature.

So what is the curvature? It takes three vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ and output a vector $\lim_{t\to 0} \frac{\boldsymbol{w}_t - \boldsymbol{w}}{t}$. It is a multilinear map. This is the Riemann curvature tensor, which goes from $T_{\boldsymbol{p}}(X) \times T_{\boldsymbol{p}}(X) \times T_{\boldsymbol{p}}(X)$ to $T_{\boldsymbol{p}}(X)$. ☺

13.2 Kronecker tensor product of $\mathbb{R}^n$

Multilinear maps are usually NOT linear. (E.g., determinant.) In particular, if a map $L : V \times V \to W$ is linear, then $L(k\boldsymbol{u}, k\boldsymbol{v}) = kL(\boldsymbol{u}, \boldsymbol{v})$. But if it is bilinear , then $L(k\boldsymbol{u}, k\boldsymbol{v}) = k^2 L(\boldsymbol{u}, \boldsymbol{v})$. As such, it is not obvious how we can study them using linear algebra.

But the following is a starting point. How did we start studying linear maps? Well, it all started with the following observation: a linear map is COMPLETELY determined by where would a basis goes. If I know the image of a linear map on the basis vectors, then the whole linear map is uniquely determined.

Observe something similar:

Example 13.2.1. Consider a bilinear map $B : \mathbb{R} \times \mathbb{R} \to V$ for whatever vector space V. Say we find out that $B(1, 1) = \boldsymbol{v}$ for some vector $\boldsymbol{v} \in V$.

Now, for any $a, b \in \mathbb{R}$, according to bilinearity, $B(a, b) = abB(1, 1) = ab\boldsymbol{v}$. Hey, everything is decided!

Just like a linear map is completely decided if we know where a basis would go, the object $(1, 1)$ is like a "tensor basis". Any bilinear map from $\mathbb{R} \times \mathbb{R}$ to V must be completely determined by where $(1, 1)$ goes.

In particular, any bilinear map $B : \mathbb{R} \times \mathbb{R} \to V$ corresponds to a linear map $L : \mathbb{R} \to V$, where we define $L(1) = B(1, 1)$. This is a one-to-one-correspondence.

Furthermore, the relation between the domain of B and L is funny: it is simply the multiplication map. Let $M : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the multiplication map, the it is easy to see that $B = L \circ M$.

In short, the study of bilinear maps from $\mathbb{R} \times \mathbb{R}$ is identical to the study of linear maps from $\mathbb{R}$, via the multiplication process. ☺

Proposition 13.2.2. *Any bilinear map $B : \mathbb{R}^m \times \mathbb{R}^n \to V$ is uniquely determined by its images $B(\boldsymbol{e}_i, \boldsymbol{e}_j)$ in U for all i, j.*

Proof. For any $\boldsymbol{v} \in \mathbb{R}^m$ and $\boldsymbol{w} \in \mathbb{R}^n$, then $\boldsymbol{v} = \sum a_i \boldsymbol{e}_i$ and $\boldsymbol{w} = \sum b_j \boldsymbol{e}_j$. Then using bilinearity, we see that $B(\boldsymbol{v}, \boldsymbol{w}) = B(\sum a_i \boldsymbol{e}_i, \sum b_j \boldsymbol{e}_j) = \sum_{i,j} a_i b_j B(\boldsymbol{e}_i, \boldsymbol{e}_j)$. So we are done. □

We only proved the binary version here. But the n-ary version is identical. For example, a trinary map B is completely determined by values $B(\boldsymbol{e}_i, \boldsymbol{e}_j, \boldsymbol{e}_k)$ for all i, j, k.

We now go one step further. We have seen that any bilinear maps from $\mathbb{R} \times \mathbb{R}$ can ba factored through the multiplication map. What is the higher dimensional analogue? Well, recall the Kronecker tensor product for vectors $\mathcal{K} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{mn}$ that sends $\boldsymbol{v} \in \mathbb{R}^m, \boldsymbol{w} \in \mathbb{R}^n$ to their Kronecker product $\boldsymbol{v} \otimes \boldsymbol{w} \in \mathbb{R}^{mn}$.

Proposition 13.2.3. *Any bilinear map $B : \mathbb{R}^m \times \mathbb{R}^n \to V$ can be decomposed into the Kronecker product $\mathcal{K} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{mn}$ and then a linear map $L : \mathbb{R}^{mn} \to V$.*

Proof. Note that vectors like $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$ form a basis for $\mathbb{R}^{mn}$. By sending each basis vector $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$ to $B(\boldsymbol{e}_i, \boldsymbol{e}_j)$, we have a linear map $L : \mathbb{R}^{mn} \to V$.

Now it is easy to check that $L \circ \mathcal{K}(\boldsymbol{e}_i, \boldsymbol{e}_j) = B(\boldsymbol{e}_i, \boldsymbol{e}_j)$, hence the two bilinear maps are the same. □

We sometimes also write $\mathbb{R}^m \otimes \mathbb{R}^n$ for $\mathbb{R}^{mn}$. Then multilinear maps from $\mathbb{R}^m \times \mathbb{R}^n$ to V would be in one-to-one correspondence with linear maps from $\mathbb{R}^m \otimes \mathbb{R}^n$ to V.

As you can imagine, if I want to study multilinear maps from $\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_k}$ to V, it is enough to study linear maps from $\mathbb{R}^{\prod n_i} = \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_k}$ to V. This way, we have successfully converted multilinear algebra into linear algebra. The dimensions are big, but all the theories shall still be there for us to use.

Here is a fast application.

Example 13.2.4 (Determinant is well-defined)**.** Remember how we use to establish determinants? It is a major headache no matter what! Here are some popular choices.

1. Use the big formula to define determinants. Oh boy is it ugly and scary. And when is the last time you ever used this? In fact, I won't be surprised if you never ever use the big formula again in your life.

2. Define the case for dimension one and two (and maybe three), and then use Laplace expansion to define determinant inductively. This one is better because sometimes we do use this method to compute determinant. (And how often do you use the big formula for 4×4 matrices?) It also has a geometric meaning hidden inside, which is a nice bonus. However, the proof is lengthy, and it is annoying to establish all the column/row operation properties.

3. I like to define it as "higher-dimensional oriented volumn". However, who can say such a thing even exist in the first place? This is not a rigorous way.

Now let us re-establish the existence of determinants in a super easy way. The key is to NOT treat $\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$ as a matrix. Rather, treat it as n inputs from $\mathbb{R}^n$, and consider multilinear maps into $\mathbb{R}$.

Consider the space $\mathbb{R}^{(n^n)} = \mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n$. Then a basis is $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_n}$ for all $i_1, \dots, i_n$. Define $L : \mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n \to \mathbb{R}$ such that

1. $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_n}$ goes to zero if some indices among $i_1, \dots, i_n$ are repeating.

2. $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_n}$ goes to 1 if the tuple $(i_1, \dots, i_n)$ can be obtained from $(1, \dots, n)$ via an even number of swaps. (I.e., the positively-oriented unit cube $\begin{bmatrix}\boldsymbol{e}_1 & \dots & \boldsymbol{e}_n\end{bmatrix}$ reflected even times could yield the cube $\begin{bmatrix}\boldsymbol{e}_{i_1} & \dots & \boldsymbol{e}_{i_n}\end{bmatrix}$.)

3. $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_n}$ goes to -1 if the tuple $(i_1, \dots, i_n)$ can be obtained from $(1, \dots, n)$ via an odd number of swaps. (I.e., the positively-oriented unit cube $\begin{bmatrix}\boldsymbol{e}_1 & \dots & \boldsymbol{e}_n\end{bmatrix}$ reflected odd times could yield the cube $\begin{bmatrix}\boldsymbol{e}_{i_1} & \dots & \boldsymbol{e}_{i_n}\end{bmatrix}$.)

Then this induces a multilinear map from $\mathbb{R}^n \times \cdots \times \mathbb{R}^n$ to $\mathbb{R}$ sending n (column) vectors to a real number, which we call the determinant. It is immediate to see that this map from $\begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_n\end{bmatrix}$ to a real number is linear in each column, and alternating (because it is alternating on abasis).

The row operation properties are also easy to do: just check on the basis $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_n}$. ☺

13.3 The abstract tensor product

Before we move on, let us note that linear maps from V to W form a vector space. We write this as $\mathcal{L}(V; W)$. Now, multilinear maps from $V_1, \dots, V_k$ to W also form a vector space, which we shall call $\mathcal{M}(V_1, \dots, V_k; W)$.

Example 13.3.1. Say we have $M_1, M_2 \in \mathcal{M}(V_1, \dots, V_k; W)$. Then a linear combination $aM_1 + bM_2$ is the obvious multilinear map that sends $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ to $aM_1(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) + bM_2(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$. Verify yourself that this is still multilinear.

Consider the space $\mathcal{M}(\mathbb{R}^n, \dots, \mathbb{R}^n; \mathbb{R})$, where we have n-copies of $\mathbb{R}^n$. In this space, all alternating multilinear maps must form a subspace $\mathcal{A}$ ("alternating" means swapping a pair of input would negate the output). You can verify that this subspace is actually one-dimensional, and thus all alternating multilinear maps here are multiples of the determinant map.

It is easy to see from our discussion in $\mathbb{R}^n$ that $\dim \mathcal{M}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m) = \dim \mathcal{L}(\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_k}; \mathbb{R}^m) = n_1 \dots n_k m$. ☺

We now attempt to do the tensor product WITHOUT using a basis. First let us observe: given a matrix A, there are three interpretations of it. We can multiply a (column) vector to its right, and think of it as a linear map $\boldsymbol{v} \mapsto A\boldsymbol{v}$. We can also multiply a (row) vector to its left, and think of it as a linear map $\boldsymbol{w}^{\mathrm{T}} \mapsto \boldsymbol{w}^{\mathrm{T}} A$. Finally, we can multiply a row vector and a column vector to both sides and think of it as a bilinear map $(\boldsymbol{v}, \boldsymbol{w}^{\mathrm{T}}) \mapsto \boldsymbol{w}^{\mathrm{T}} A \boldsymbol{v}$.

As you can see, all three things corresponds to the same matrix A. It seems that the following three things are all actually the same thing! (As they are all the same matrices.)

1. Linear maps $\mathbb{R}^m \to \mathbb{R}^n$.

2. Linear maps $(\mathbb{R}^n)^* \to (\mathbb{R}^m)^*$.

3. Bilinear maps $\mathbb{R}^m \times (\mathbb{R}^n)^* \to \mathbb{R}$.

4. $n \times m$ matrices.

Proposition 13.3.2 (What are matrices)**.** *Let V, W be finite dimensional spaces. Then the following three things are the same, in the sense that they all have (canonical and linear) one-to-one correspondings with each other.*

1. *Linear maps $V \to W$. I.e., $\mathcal{L}(V; W)$.*

2. *Linear maps $W^* \to V^*$. I.e., $\mathcal{L}(W^*; V^*)$.*

3. *Bilinear maps $V \times W^* \to \mathbb{R}$. I.e., $\mathcal{M}(V, W^*; \mathbb{R})$.*

(Note that these three are three vector spaces, and their identifications with each other are (canonical) linear bijections.)

Proof. The first two things are in correspondence by taking dual map of each other.

The third one is in correspondence with previous two by taking an analogue of the "bra map". Given a bilinear map $B : V \times W^* \to \mathbb{R}$, we define $L_B : V \to W^{**}$ to be $B(\boldsymbol{v}, -)$, and $L_B^* : W^* \to V^*$ to be $B(-, \alpha)$. I shall leave the verification that these are indeed linear for yourself. □

Remark 13.3.3 (Why canonical)**.** *(Optional)*

*This portion deal with the use fo the term **canonical** here. Suppose we have spaces V_1, V_2, W_1, W_2. For any $L_V : V_2 \to V_1$ and $L_W : W_1 \to W_2$, then we have three induced linear maps:*

1. *Elements in the space $\mathcal{L}(V_1; W_1)$ of linear maps from V_1 to W_1 could be send (via composition with L_V, L_W) to elements in the space $\mathcal{L}(V_2; W_2)$.*

2. *Elements in the space $\mathcal{L}(W_1^*; V_1^*)$ could be send (via composition with L_V^*, L_W^*) to elements in the space $\mathcal{L}(V_2; W_2)$.*

3. *Elements in the space $\mathcal{B}(V_1, W_1^*; \mathbb{R})$ of bilinear maps from $V_1 \times W_1^*$ to $\mathbb{R}$ can be sent (via composition with L_V, L_W^*) to elements of $\mathcal{B}(V_2, W_2^*; \mathbb{R})$.*

(Draw the diagram to make this easier....) If we treat the spaces $\mathcal{L}(V_1; W_1), \mathcal{L}(W_1^*; V_1^*), \mathcal{B}(V_1, W_1^*; \mathbb{R})$ as the same space, and the spaces $\mathcal{L}(V_2; W_2), \mathcal{L}(W_2^*; V_2^*), \mathcal{B}(V_2, W_2^*; \mathbb{R})$ as the same space, then by straight forward calculation, the three maps induced by L_V, L_W are the same linear map.

This way, we see that not only spaces are isomorphic, maps between pairs of spaces are also isomorphic. Hence we can say that this is canonical.

Definition 13.3.4. *For any two finite dimensional vector spaces V, W, we define their tensor space $V \otimes W$ as the space $\mathcal{L}(V^*; W)$ of linear maps from V^* to W, or the space $\mathcal{L}(W^*; V)$ of linear maps from W^* to V, or the space $\mathcal{B}(V^*, W^*; \mathbb{R})$ of bilinear maps from $V^* \times W^*$ to $\mathbb{R}$.*

We have a standard bilinear map $\otimes : V \times W \to V \otimes W$, such that $\otimes(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v} \otimes \boldsymbol{w}$ should represent the linear map that sends each $\beta \in W^*$ to $\beta(\boldsymbol{w})\boldsymbol{v}$, or the linear map that sends each $\alpha \in V^*$ to $\alpha(\boldsymbol{v})\boldsymbol{w}$, or the bilinear map that sends $(\alpha, \beta) \in V^* \times W^*$ to the number $\alpha(\boldsymbol{v})\beta(\boldsymbol{w})$.

In short, $\boldsymbol{v} \otimes \boldsymbol{w}$ is just two vectors waiting to be eaten by the corresponding dual vectors! For example, you can eat $\boldsymbol{v}$ via something in V^*, and what you are left with is a multiple of $\boldsymbol{w}$. You can also eat $\boldsymbol{w}$ via something in W^*, and what you are left with is a multiple of $\boldsymbol{v}$.

A vector and a dual vector will eat each other. So what is $\boldsymbol{v} \otimes \boldsymbol{w}$? We have two vectors, and they are waiting to eat corresponding dual vectors (and then the result is multiplied together). So they induce a bilinear map from $V^* \times W^*$ to $\mathbb{R}$.

Example 13.3.5. Consider $\mathbb{R}^n \otimes (\mathbb{R}^m)^*$. This is also the space of linear maps from $\mathbb{R}^m \to \mathbb{R}^n$, i.e., the space of $m \times n$ matrices!

So elements of $\mathbb{R}^n \otimes (\mathbb{R}^m)^*$ are just matrices. the element $\boldsymbol{v} \otimes \boldsymbol{w}^{\mathrm{T}}$ in it is simply just the matrix $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$ in the most literal sense! This is a very nice idea to keep in mind. The symbol $\otimes$ might looks strange, but it is usually just the obvious multiplication under whatever setting. E.g., in the space $\mathbb{R} \otimes \mathbb{R}$, then $a \otimes b$ is just ab.

Note that not all matrices are rank one. There are many elements in $\mathbb{R}^n \otimes (\mathbb{R}^m)^*$ that CANNOT be written as $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$. Similarly, not all elements of $V \otimes W$ are of the form $v \otimes w$. Those are simply the "rank one" tensors. In general, we say a tensor $\omega \in V \otimes W$ to have rank k if we need at least k "rank one" tensors to linearly combine into ω. ☺

Let us go back to the point of establishing tensor spaces.

Proposition 13.3.6. *There is a (canonical) linear one-to-one corresponding between bilinear maps from $V \times W$ to U and linear maps from $V \otimes W$ to U.*

Proof. The cheap way: pick basis, and turn everything into $\mathbb{R}^m, \mathbb{R}^n$. Then we have already done this.

The honest way: Let us build a linear isomorphism from $\mathcal{L}(V \otimes W; U)$ to $\mathcal{B}(V, W; U)$. For any linear $L : V \otimes W \to U$, note that we have a natural bilinear map $\otimes : V \times W \to V \otimes W$, then $L \circ \otimes$ is a bilinear map. It is straight forward to verify that $L \mapsto L \circ \otimes$ is linear.

Suppose $L \circ \otimes$ is zero. Then $L(\boldsymbol{v} \otimes \boldsymbol{w}) = L \circ \otimes(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{0}$. So the process $L \mapsto L \circ \otimes$ is injective. We count dimension and see that it must be a bijection. ($\dim \mathcal{B}(V, W; U) = \dim V \dim W \dim U = \dim \mathcal{L}(V \otimes W; U)$.) □

Example 13.3.7. Let us study the cross product, which is bilinear from $\mathbb{R}^3 \times \mathbb{R}^3$ to $\mathbb{R}^3$. This corresponds to a linear map from $\mathbb{R}^3 \otimes \mathbb{R}^3$ to $\mathbb{R}^3$. How is it built?

Note that the cross product is defined so that $\boldsymbol{e}_i \times \boldsymbol{e}_{i+1} = \boldsymbol{e}_{i+2}$ (where the indices are taken mod three), and it is alternating (skew-symmetric). Hence let us build a linear map $C : \mathbb{R}^3 \otimes \mathbb{R}^3 \to \mathbb{R}^3$ by declaring its value at a basis. The rule is this: for each pair of basis vectors, the output is the third basis vector. (And we choose directions carefully to achieve skew-symmetry.)

1. $C(\boldsymbol{e}_1 \otimes \boldsymbol{e}_1) = \boldsymbol{0}$.
2. $C(\boldsymbol{e}_1 \otimes \boldsymbol{e}_2) = \boldsymbol{e}_3$.
3. $C(\boldsymbol{e}_1 \otimes \boldsymbol{e}_3) = -\boldsymbol{e}_2$.
4. $C(\boldsymbol{e}_2 \otimes \boldsymbol{e}_1) = -\boldsymbol{e}_3$.
5. $C(\boldsymbol{e}_2 \otimes \boldsymbol{e}_2) = \boldsymbol{0}$.
6. $C(\boldsymbol{e}_2 \otimes \boldsymbol{e}_3) = \boldsymbol{e}_1$.
7. $C(\boldsymbol{e}_3 \otimes \boldsymbol{e}_1) = \boldsymbol{e}_2$.
8. $C(\boldsymbol{e}_3 \otimes \boldsymbol{e}_2) = -\boldsymbol{e}_1$.
9. $C(\boldsymbol{e}_3 \otimes \boldsymbol{e}_3) = \boldsymbol{0}$.

This gives the values of C at a basis, so they extends to a linear map, whose corresponding bilinear map is the cross product. Cross product is the unique thing that linearly extends the declarations above. ☺

So indeed $V \otimes W$ serves the purpose of an abstract version of Kronecker product, and the bilinear map $\otimes$ is the "first process" that any bilinear map must go through.

Note that above process is true for all multilinear maps. We just do the bilinear version because it is easier to present. In general, we have $\mathcal{M}(V_1, \ldots, V_k; W)$ canonically isomorphic to $\mathcal{L}(V_1 \otimes \cdots \otimes V_k; W)$.

Remark 13.3.8. *There are some hidden things. For example, one can also prove that the spaces $U \otimes V \otimes W$ and $(U \otimes V) \otimes W$ and $U \otimes (V \otimes W)$ are canoncially the same.*

$V \otimes W$ and $W \otimes V$ are also canonically the same, but note that this canonical identication is NOT identity map. For example, $\mathbb{R}^2 \otimes \mathbb{R}^2$ under this "swapping" canonical isomorphism to itself is like taking a "transpose".

At the level of spaces, you may just think of $V \otimes W$ as the same space as $W \otimes V$. But when $V = W$, then at an level of elements, we almost always expect $\boldsymbol{v} \otimes \boldsymbol{w} \neq \boldsymbol{w} \otimes \boldsymbol{v}$.

It is now time to loosen up. It turns out that domain and codomain should be treated "fluidly" in the following sense:

Proposition 13.3.9. *All these spaces are canonically isomorphic to $\mathcal{B}(V, W; U)$.*

1. $V^* \otimes W^* \otimes U$.
2. $\mathcal{L}(\mathbb{R}; V^* \otimes W^* \otimes U)$
3. $\mathcal{L}(V; W^* \otimes U)$.
4. $\mathcal{L}(W; V^* \otimes U)$.
5. $\mathcal{L}(U^*; V^* \otimes W^*)$.
6. $\mathcal{L}(V \otimes W; U)$.
7. $\mathcal{L}(V \otimes U^*; W^*)$.
8. $\mathcal{L}(W \otimes U^*; V^*)$.
9. $\mathcal{L}(V \otimes W \otimes U^*; \mathbb{R})$.

Proof. The actual proof is not hard, just boring and long. Here let us see this intuitively.

An element $\omega \in V^* \otimes W^* \otimes U$ means we have a dual vector in V^*, a dual vector in W^*, and a vector in U, combined in a multilinear fashion.

If the V^* portion of ω eat a vector in V, then we are left with a dual vector in W^* and a vector in U, combined in a multilinear fashion. So ω can be seen as an element of $\mathcal{L}(V; W^* \otimes U)$.

Similarly, let ω eat things selectively, and we have all these interpretations of ω as listed. □

So if some tensor-factor is in the domain, you can throw it to the codomain by taking dual. If something tensor-factor is in the codomain, you can throw it back to the domain by taking dual. If the domain or codomain has nothing left, then it becomes $\mathbb{R}$. These are the rules of tensor algebra.

So, how to study a multilinear map $\mathcal{M}(V_1, \dots, V_k; W)$? Well, we simply study the vector space $V_1^* \otimes \cdots \otimes V_k^* \otimes W$.

So from now on, we never need to study multilinear maps anymore. We simply study vector spaces.

Example 13.3.10. Recall our examples of 3D array of numbers in the begining, which is a trilinear map in $\mathcal{M}(\mathbb{R}^2, \mathbb{R}^4, \mathbb{R}^3; \mathbb{R})$. For such a 3D box array B, say $2 \times 4 \times 3$, then these arrays form a vector space $(\mathbb{R}^2)^* \otimes (\mathbb{R}^4)^* \otimes (\mathbb{R}^3)^*$.

Recall that we have three ways to collapse the 3D array into 2D arrays. This corresponds to the interpretation of B as an element of $\mathcal{L}(\mathbb{R}^2; (\mathbb{R}^4)^* \otimes (\mathbb{R}^3)^*), \mathcal{L}(\mathbb{R}^4; (\mathbb{R}^2)^* \otimes (\mathbb{R}^2)^*), \mathcal{L}(\mathbb{R}^3; (\mathbb{R}^2)^* \otimes (\mathbb{R}^4)^*)$.

A basis for $(\mathbb{R}^2)^* \otimes (\mathbb{R}^4)^* \otimes (\mathbb{R}^3)^*$ is made in the form of $\boldsymbol{e}_i^{\mathrm{T}} \otimes \boldsymbol{e}_j^{\mathrm{T}} \otimes \boldsymbol{e}_k^{\mathrm{T}}$. And the "entries" in the box array is the coordinates under this basis. For example, $2\boldsymbol{e}_1^{\mathrm{T}} \otimes \boldsymbol{e}_1^{\mathrm{T}} \otimes \boldsymbol{e}_1^{\mathrm{T}} + 4\boldsymbol{e}_2^{\mathrm{T}} \otimes \boldsymbol{e}_3^{\mathrm{T}} \otimes \boldsymbol{e}_3^{\mathrm{T}}$ corresponds to the box array whose three layers are $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 \end{bmatrix}$. You can think of the i, j, k index-values as saying "this coefficient is the entry in the (i, j)-location of the k-th layer".

You can now verify that if the three layers are A_1, A_2, A_3, then it indeed send $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ to $\begin{bmatrix} \boldsymbol{u}^{\mathrm{T}} A_1 \boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}} A_2 \boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}} A_3 \boldsymbol{v} \end{bmatrix} \boldsymbol{w}$.

For example, $\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\2\\3\\4\end{bmatrix}, \begin{bmatrix}1\\2\\3\end{bmatrix}$ is sent by the example above to $\begin{bmatrix}\boldsymbol{u}^{\mathrm{T}}A_1\boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}}A_2\boldsymbol{v} & \boldsymbol{u}^{\mathrm{T}}A_3\boldsymbol{v}\end{bmatrix}\boldsymbol{w} = \begin{bmatrix}2 & 0 & 24\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} =$ 74.

However, there is another way to calculate this. From $2\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}+4\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}$, note that $\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}$ and $\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}$ would simply multiply the corresponding coordinates of input vectors, according to their indices. So $\omega(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) = 2u_1v_1w_1 + 4u_2v_3w_3 = 74$. ☺

Example 13.3.11. Consider the cross product again. It is bilienar from $\mathbb{R}^3\times\mathbb{R}^3$ to $\mathbb{R}^3$. Hence the cross product is an element C of the vector space $(\mathbb{R}^3)^*\otimes(\mathbb{R}^3)^*\otimes\mathbb{R}^3$.

In fact, since C is defined as the following:

1. $C(\boldsymbol{e}_1\otimes\boldsymbol{e}_1) = \boldsymbol{0}$.
2. $C(\boldsymbol{e}_1\otimes\boldsymbol{e}_2) = \boldsymbol{e}_3$.
3. $C(\boldsymbol{e}_1\otimes\boldsymbol{e}_3) = -\boldsymbol{e}_2$.
4. $C(\boldsymbol{e}_2\otimes\boldsymbol{e}_1) = -\boldsymbol{e}_3$.
5. $C(\boldsymbol{e}_2\otimes\boldsymbol{e}_2) = \boldsymbol{0}$.
6. $C(\boldsymbol{e}_2\otimes\boldsymbol{e}_3) = \boldsymbol{e}_1$.
7. $C(\boldsymbol{e}_3\otimes\boldsymbol{e}_1) = \boldsymbol{e}_2$.
8. $C(\boldsymbol{e}_3\otimes\boldsymbol{e}_2) = -\boldsymbol{e}_1$.
9. $C(\boldsymbol{e}_3\otimes\boldsymbol{e}_3) = \boldsymbol{0}$.

This means we have $C = \boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3-\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_3+\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_1-\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_1+\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_2-\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_2$. And the three layers are $A_1 = \begin{bmatrix}0&0&0\\0&0&1\\0&-1&0\end{bmatrix}, A_2 = \begin{bmatrix}0&0&-1\\0&0&0\\1&0&0\end{bmatrix}, A_3 = \begin{bmatrix}0&1&0\\-1&0&0\\0&0&0\end{bmatrix}$. And as a bilinear map from $\mathbb{R}^3\times\mathbb{R}^3$ to $\mathbb{R}^3$, it would sends $\boldsymbol{v},\boldsymbol{w}$ to $\begin{bmatrix}\boldsymbol{v}^{\mathrm{T}}A_1\boldsymbol{w}\\\boldsymbol{v}^{\mathrm{T}}A_2\boldsymbol{w}\\\boldsymbol{v}^{\mathrm{T}}A_3\boldsymbol{w}\end{bmatrix}$. This map action is a bit different from before, because the last space of $(\mathbb{R}^3)^*\otimes(\mathbb{R}^3)^*\otimes\mathbb{R}^3$ is not dualed.

Also note that you can litereally read the cross product formula in your calculus class from the expression $C = \boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3 - \boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_3 + \boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_1 - \boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_2^{\mathrm{T}}\otimes\boldsymbol{e}_1 + \boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_2 - \boldsymbol{e}_1^{\mathrm{T}}\otimes\boldsymbol{e}_3^{\mathrm{T}}\otimes\boldsymbol{e}_2$. Can you see the correspondence? ☺

Example 13.3.12. Consider the determinant map $\det\in(\mathbb{R}^3)^*\otimes(\mathbb{R}^3)^*\otimes(\mathbb{R}^3)^*$. What are the coordinates of det in terms of standard basis vectors? What are the layer matrices? Which element of $(\mathbb{R}^3)^*$ (the third tensor factor space) would it send $\boldsymbol{v},\boldsymbol{w}$ to? Compare the result with the last example, and enjoy the surprise. ☺

Example 13.3.13 (Complexification)**.** If V is a real vector space, say of dimension n. Treat $\mathbb{C}$ also as a real vector space. Then $V\otimes\mathbb{C}$ has real dimension $2n$.

For any $\boldsymbol{v}\otimes z$ and any $k\in\mathbb{C}$, let us define scalar multiplication as $k(\boldsymbol{v}\otimes z) = \boldsymbol{v}\otimes kz$. Then you may verify that this makes $V\otimes\mathbb{C}$ into a complex vector space of complex dimension n. If $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ is a basis for the real vector space V, then $\boldsymbol{v}_1\otimes 1,\ldots,\boldsymbol{v}_n\otimes 1$ is a basis for the complex vector space V.

You may also check that $\mathbb{R}^n\otimes\mathbb{C}$ is canonically the same as $\mathbb{C}^n$. In general, if V has property blah, then $V\otimes\mathbb{C}$ is expected to have the corresponding complex property blah.

Suppose you have been working with a real vector space V, but for whatever reason, all of a sudden you need to expand the coefficients to include complex numbers. Then you can just do the complex vector space $V\otimes\mathbb{C}$, and translate whatever result you have from V over to $V\otimes\mathbb{C}$ in a very simple way. ☺

Suppose we have $A : V_1 \to V_2$ and $B : W_1 \to W_2$. Then for each $\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \in V_2 \otimes V_2$, we can send it to $(A\boldsymbol{v}_1) \otimes (A\boldsymbol{v}_2)$. Since the tensor space is spanned by rank one elements, this induces a well-defined map $A \otimes B : V_1 \otimes V_2 \to W_1 \otimes W_2$. (We do not establish it here, but as you can imagine, in coordinates, the matrix of $A \otimes B$ would be the Kronecker tensor product of the matrices for A and B.)

Example 13.3.14 (Tensor multiplication and real number multiplication)**.** Why would $\boldsymbol{v} \otimes \boldsymbol{w}$ in $V \otimes W$ corresponds to the bilinear map $\boldsymbol{v} \otimes \boldsymbol{w}(\alpha \otimes \beta) = \alpha(\boldsymbol{v})\beta(\boldsymbol{w})$? In particular, why does the tensor multiplication between $\boldsymbol{v}$ and $\boldsymbol{w}$ translate into the multiplication of real numbers?

Note that dual vectors are linear maps by nature, i.e., $\alpha : V \to \mathbb{R}$ and $\beta : W \to \mathbb{R}$. Therefore $\alpha \otimes \beta$ is a linear map from $V \otimes W$ to $\mathbb{R} \otimes \mathbb{R} = \mathbb{R}$. The last equality here is exactly real number multiplication. ☺

13.4 Tensor powers and calculations

Definition 13.4.1. *An (a, b)-tensor over a vector space V is an element of $V^{\otimes a} \otimes (V^*)^{\otimes b}$. (Here the tensor power notation means the tensor of this space with itself many times.)*

Note that all (a, b) tensors over a vector space V form a vector space, which we write as $\mathcal{T}_b^a(V)$.

For example, the cross product of 3D vectors is a $(1, 2)$ tensor over $\mathbb{R}^3$, while a determinant map is an $(0, n)$ tensor over $\mathbb{R}^n$. The inner product is a $(0, 2)$-tensor. The Riemann curvature tensor, as you might have recall, would eat three vectors of $T_{\boldsymbol{p}}X$ and spit out a single vector of $T_{\boldsymbol{p}}X$. Therefore it is a $(1, 3)$ tensor over $T_{\boldsymbol{p}}X$.

(Also, sometimes people get careless and simply say $a + b$ tensor. Then a 2-tensor can always be represented by a matrix, a 3-tensor is always a 3D array, and so on.)

Example 13.4.2. Say M is a geometric object (manifold), then at each point $\boldsymbol{p} \in M$, the Riemann curvature tensor at this point is a $(1, 3)$ tensor. Since we have such a tensor at each point, we see that curvature on M is in fact a $(1, 3)$-tensor field!

Similarly, say we want to endow M with distance structure (called a Riemannian manifold). To measure angle between intersecting curves, we need an inner product between tangent vectors at the intersection point! Furthermore, to get the length of a curve, we simply integrate the length of the velocity along the curve. So again, we need an inner product structure on each tangent space.

So the distance structure of M is equivalent to picking an inner product for each tangent space. I.e., this is a $(0, 2)$-tensor field. ☺

Also note that by convention, we define $V^{\otimes 0}$ to be $\mathbb{R}$. (Because $V \otimes \mathbb{R}$ is just V itself. Can you see why?)

This section focus on calculation of things. Say you have a $(0, b)$-tensor. Then this corresponds to a multilinear map to $\mathbb{R}$ who needs to eat b vectors. So I can feed k vectors (or a $(k, 0)$-tensor) to it, and get a $(0, b - k)$-tensor.

Similarly, say you have a $(a, 0)$-tensor. Then this corresponds to a multilinear map to $\mathbb{R}$ who needs to eat a dual vectors. So I can feed k dual vectors (or a $(0, k)$-tensor) to it, and get a $(a - k, 0)$-tensor.

In general, for an (a, b)-tensor, you can feed it a (b, a)-tensor to get a number. I.e., you can think of $\mathcal{T}_b^a(V)$ and $\mathcal{T}_a^b(V)$ as dual spaces of each other. So the calculation works like this:

$$\boldsymbol{v}_1 \otimes \cdots \otimes \boldsymbol{v}_a \otimes \alpha_1 \otimes \cdots \otimes \boldsymbol{\alpha}_b(\boldsymbol{w}_1 \otimes \cdots \otimes \boldsymbol{w}_b \otimes \beta_1 \otimes \cdots \otimes \beta_a) = \beta_1(\boldsymbol{v}_1) \dots \beta_a(\boldsymbol{v}_a)\alpha_1(\boldsymbol{w}_1) \dots \alpha_b(\boldsymbol{w}_b).$$

Note that this formula is for rank 1 tensors. A generic (a, b)-tensor or (b, a)-tensor might not have rank one, but they are always a linear combination of rank 1 tensors.

Any, we have the following.

Proposition 13.4.3. $(\mathcal{T}_b^a(V))^* = \mathcal{T}_a^b(V)$.

Here are some easy calculations involving tensors.

Example 13.4.4 (What is dot product)**.** Consider $V = \mathbb{R}^2$ and let g be the $(0,2)$-tensor which is the dot product.

Then $g(\boldsymbol{v} \otimes \boldsymbol{w}) = v_1 w_1 + v_2 w_2$. Well, it is easy to see that the tensor $\boldsymbol{e}_1^{\mathrm{T}} \otimes \boldsymbol{e}_1^{\mathrm{T}} + \boldsymbol{e}_2^{\mathrm{T}} \otimes \boldsymbol{e}_2^{\mathrm{T}}$ does exactly this! So $g = \boldsymbol{e}_1^{\mathrm{T}} \otimes \boldsymbol{e}_1^{\mathrm{T}} + \boldsymbol{e}_2^{\mathrm{T}} \otimes \boldsymbol{e}_2^{\mathrm{T}}$. ☺

Example 13.4.5 (What is a Euclidean space)**.** We treat the Euclidean space $\mathbb{R}^2$ as a Riemannian manifold. So it has a "distance structure", i.e., a $(0,2)$-tensor field, i.e., a $(0,2)$-tensor at each point. What is its $(0,2)$-tensor field?

Note that a $(0,2)$-tensor on $T_{\boldsymbol{p}}\mathbb{R}^2$ is a linear combination of $\alpha \otimes \beta$ for dual vectors of $T_{\boldsymbol{p}}\mathbb{R}^2$. For each tangent vector $\boldsymbol{v}, \boldsymbol{w} \in T_{\boldsymbol{p}}\mathbb{R}^2$, we want $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = v_x w_x + v_y w_y$ obviously. This corresponds to the dual vector tensor $\begin{bmatrix} 1 & 0 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \end{bmatrix}$.

Now note that $\mathrm{d}x, \mathrm{d}y$ are covector fields. So $\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y$ gives a $(0,2)$-tensor field, and at each point it will be exactly what we want. ☺

Example 13.4.6 (The hyperbolic plane)**.** Let H be the open half plane, i.e., the collection of points with positive y-coordinate. Consider the $(0,2)$-tensor field $\frac{\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y}{y^2}$. This is called the hyperbolic plane.

Consider the curve $\gamma : (-\infty, \infty) \to H$ such that $\gamma(t) = \begin{bmatrix} 0 \\ \mathrm{e}^{-t} \end{bmatrix}$. Then its tangent vector at t is $\gamma'(t) = \begin{bmatrix} 0 \\ -\mathrm{e}^{-t} \end{bmatrix}$. What is the "speed" according to my distance structure? Well, we have to use the $(0,2)$-tensor $\frac{\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y}{y^2}$ to evaluate this tangent vector with itself (like dot product) and then take square root.

The tensor field $\frac{\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y}{y^2}$ would tell me that this tangent vector has "length" $\sqrt{\frac{0 \times 0 + (-\mathrm{e}^{-t})(-\mathrm{e}^{-t})}{\mathrm{e}^{-t}\mathrm{e}^{-t}}} = 1$. So what is γ? It is a walk towards the origin along the y-axis at "constant speed"!

As you can see, as we move towarts the x-axis, things will become "sluggish". You feel like you are moving at constant speed, yet "Euclideanly" you seem to be slower and slower, and you will in fact never be able to reach the x-axis. You may also imagine that this is a magical place where, as you move towards the x-axis, your leg become shorter and shorter.

It can be shown that "Straight lines" in H are of the following two kinds: either a half-circle around some point on the x-axis, or a vertical line perpendicular to the x-axis. Given any two points $\boldsymbol{p}, \boldsymbol{q}$, what is the "shortest path" connecting them? Well, you draw their "perpendicular bisector" which intersects with the x-axis, and then use the intersection as a center of a half circle going through $\boldsymbol{p}, \boldsymbol{q}$. Then the shortest path from $\boldsymbol{p}$ to $\boldsymbol{q}$ is the corresponding arc on the half-circle.

Intuitively, to go from $\boldsymbol{p}$ to $\boldsymbol{q}$ as fast as possible, you need to first move somewhat "upward" so that your leg can grow longer.

The hyperbolic plane is very famous, because it is the opposite of a sphere. A sphere has constant positive curvature everywhere, while a hyperbolic plane has constant negative curvature everywhere. Triangles in H will have internal angles less than π, and "circles" in H or radius r (i.e., points with distance r to a center point in the H-distance.) will have circumference $2\pi \frac{\mathrm{e}^r - \mathrm{e}^{-r}}{2} \geq 2\pi r$. As you can imagine, it grows approximately exponentially as r grows. If you stand in H and look away, then "things with distance r to you" will be exponentially more as r grows.

If you shoot at a target with distance r away from you, then it is easier to do in a Euclidean space ($2\pi r$ things are distance r away from you) than in the hyperbolic space (exponentially more things are distance r away from you). If we do chemistry in the hyperbolic plane (or its generalization, the hyperbolic space), then everything is super stable, because it is so much harder for molecules to accurately collide with each other.

For someone standing inside the hyperbolic plane, in their eyes, the hyperbolic plane would actually look like a disk (search for "Poincaré circle model".) This is becasue their vision border at distance r will grow exponentially as r grows. This growth is so fast, that infinity is right in front of your eyes! ☺

The examples here are not too bad. However, that is because $\mathbb{R}^2$ is rather easy. For more complicated geometric object, say a 4-dim black hole model, then the computations of distances ($(0,2)$-tensor field) and

curvatures ($(1,3)$-tensor field) will become increasingly annoying. Therefore, the famous Albert Einstein invented a whole new notation to calculate tensors over $\mathbb{R}^n$!

So you see, sometimes the key to discovering a great theory, like general relativity, is to invent super nice notation, so that calculations are easier.

The key idea behind it are the followings:

1. We write coordinates of vectors with upper indices, and coordinate of dual vectors using lower indices. So we write $\boldsymbol{v} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix}$ and $\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}$ and so on.

2. Basis-dependent things should NOT matter. So given $\boldsymbol{v} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix}$ and $\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}$, will we be interested in the quantity $\alpha_1 v^1$? NEVER! That would depends on a choice of basis. However, could we be interested in the quantity $\sum \alpha_i v^i$? Yes. This is simply $\alpha(\boldsymbol{v})$, and it does not depend on the choice of basis.

3. As a result, if we see $\alpha_1 v^1$ in a calculation process, it must never be alone. It must come as a cluster $\alpha_1 v^1 + \alpha_2 v^2 + \alpha_3 v^3$ so that it is independent of basis!

4. Writing clusters or summation symbols are tiring. So from now on, if we write $\alpha_i v^i$, it is implied that we are adding over i.

Continuing this idea, we have the following notations when it comes to vectors and dual vectors.

1. For a vector $\boldsymbol{v}$, we sometimes just write (v^i). Similarly, for a dual vector α, we sometimes just write (α_i).

2. If we evaluate a dual vector on a vector, we have $(\alpha_i)(v^i) = a_i v^i$, where in the last expression, it is understood that we take the sum over all possible i.

3. Note that if you see $v^i \alpha_i$, then it is the same thing as $\alpha_i v^i$. We are just adding these products of these coordinates over all possible i.

As a rule of thumb, if an index i appear somewhere as an upper index (some vector-coordinates have this index), and then somewhere else as a lower index (some dual vector-coordinates have this index), then it is implied that we are summing up over i (the vector index and the dual vector index eat each other). We do not write the summation symbol. All other notations are invented to keep this rule alive.

For example, consider this:

Example 13.4.7. If we are writing a sequence of vectors, we write $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$. If we are writing a sequence of dual vectors, we write $\alpha^1, \dots, \alpha^j$.

For example, we write the standard basis vectors as $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$. Then any $\boldsymbol{v}$ is a linear combination of these. We simply write $\boldsymbol{v} = v^i \boldsymbol{e}_i$ where we don't need to write the summation symbol. Whenever you see i as both an upper index and a lower index, then you just sum over it without hesitation.

Similarly, the dual stardard basis would be $\boldsymbol{e}^1, \dots, \boldsymbol{e}^n$, and $\alpha = \alpha_i \boldsymbol{e}^i$. ☺

Now what about tensors? We generalize the following convension: vector indices goes to the top, while dual vector indices goes to the bottom. So an (a,b)-tensor T will have coordinates $T^{i_1,\dots,i_a}_{j_1,\dots,j_b}$. This means we can write T as a linear combination $T = \sum T^{i_1,\dots,i_a}_{j_1,\dots,j_b} \boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_a} \otimes \boldsymbol{e}^{\mathrm{T}}_{j_1} \otimes \cdots \otimes \boldsymbol{e}^{\mathrm{T}}_{j_b}$. Sometimes the standard basis vector $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_a} \otimes \boldsymbol{e}^{\mathrm{T}}_{j_1} \otimes \cdots \otimes \boldsymbol{e}^{\mathrm{T}}_{j_b}$ for the tensor space $\mathcal{T}^a_b(V)$ is simply written as $\boldsymbol{e}^{j_1,\dots,j_b}_{i_1,\dots,i_a}$. Then we simply write $T = T^{i_1,\dots,i_a}_{j_1,\dots,j_b} \boldsymbol{e}^{j_1,\dots,j_b}_{i_1,\dots,i_a}$, where we sum up corresponding things in the obvious manner.

Consider the following

Example 13.4.8. 1. A matrix $A : V \to V$ is a tensor in $V \otimes V^*$. So we write its entries as $A = (A^i_j)$. Note that i is indexing the coordinates of the output, while j is indexing the coordinate of the input. Might as well keep this in mind: upper-index=output, lower-index=input.

2. For any input vector (v^j), the rule of matrix-vector multiplication is simply $(A^i_j)(v^j) = (A^i_j v^j)$. Note that the index j is summed over, so $A^i_j v^j$ is simply a number for each i. Hence $(A^i_j v^j)$ is a vector whose coordinates are indexed by i. Look at how pretty this multiplication formula is: $(A^i_j)(v^j) = (A^i_j v^j)$!

3. What is A^i_i? Well, it is obviously the trace of the matrix A. This is probably the fastest proof that trace is independent of coordinates.

4. Consider matrix multiplications. Say $A = (A^i_j)$ and $B = (B^i_j)$. To do matrix multiplications AB, we need to feed the output of B to the input of A. So we just "connect the lower index of A (the input of A) with the upper index of B (the output of B)". So $(AB)^i_j = (A^i_k)(B^k_j) = (A^i_k B^k_j)$. Wow, matrix multiplication formula is so easy!

5. What if we see an expression $B^k_j A^i_k$? Well, it is STILL the (i, j)-coordinate for AB! Because by looking at the index, we are still identifying the input of A with the output of B. So as you can see, the Einstein notation has NO confusion about the order of multiplication. You can feel free to write $B^k_j A^i_k$ or $A^i_k B^k_j$, and they both corresponds to the same entry of AB. To get an entry of BA, the you would need to connect the output of A with the input of B, so you are looking at $A^k_j B^i_k$.

6. Suppose A, B are inverse matrix of each other. Then we expect $A^i_j B^j_k = A^j_k B^i_j = \delta^i_k$, where δ^i_k is the (i, k)-entry of the identity matrix.

7. Given a vector $\boldsymbol{v} = (v^i)$ and a row vector $\boldsymbol{w}^{\mathrm{T}} = (w_i)$, then $v^i w_i$ means $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}$, while $v^i w_j$ is a $(1, 1)$-tensor, and it is the linear map $\boldsymbol{v}\boldsymbol{w}^{\mathrm{T}}$.

8. Given $\boldsymbol{v} = (v^i), \boldsymbol{w} = (w^i)$, then what is $v^i w^j$? It is $\boldsymbol{v} \otimes \boldsymbol{w}$, a $(2, 0)$-tensor. What is $v^i w^i$? It is NOT the dot product. It is in fact MEANINGLESS. NEVER put the same index as upper index twice, or as lower index twice. But how would we do dot product then? Well, look below.

9. What is a bilinear map? Well, it eats two vectors, so it is a $(0, 2)$-tensor, i.e., its entry shall have two lower indices. Say A is a bilinear map, then $A = (A_{i,j})$. Then $A(\boldsymbol{v}, \boldsymbol{w}) = A_{ij} v^i w^j$, a number. The dot product is (δ_{ij}) where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. So how to take dot product between two vectors? DO NOT WRITE $\sum_i v^i w^i$, which looks stupid. We NEVER want the same index appear twice as an upper index, or twice as a lower index. Instead, we write $v^i w^j \delta_{ij}$ for the dot product. The symmetricity is super clear in this formula, much better than, say, $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$, which makes symmetricity unclear.

☺

So, given entries of matrices A, B, C, what is the multiplication formula for ABC? Easy. Hook up the output of C to the input of B, and the output of B to the input of A, and we are done. So we can just do $C^k_j B^l_k A^i_l$ (the order of multiplication of C^k_j, B^l_k, A^i_l does not matter), and this is the (i, j)-entry of ABC.

13.5 Inner products of tensors

In this section, we FIX a vector space V with an inner product structure. Say the inner product is induced by the $(0, 2)$-tensor $g = (g_{ij})$, i.e., $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = g(\boldsymbol{v}, \boldsymbol{w}) = v^i w^j g_{ij}$. Note that we require g to be symmetric and positive definite, i.e., $g_{ij} = g_{ji}$, and $g_{ij} v^i v^j \geq 0$ with equality if and only if all v^i are zero.

First we need a calculational version of Riesz representation theorem. The bra map is easy.

Proposition 13.5.1 (Lowering the index)**.** *The map $\boldsymbol{v} \mapsto \langle \boldsymbol{v}|$ would send the vector (v^i) to the dual vector $(v^i g_{ij})$.*

Proof. $\langle \boldsymbol{v}|((w^j)) = v^i g_{ij} w^j = (v^i g_{ij})(w^j)$. So we are done. □

So under the inner product on V, given a vector, how to find the corresponding dual vector? We simply multiply by g_{ij}.

Then conversely, given a dual vector, how would we go back to the vector? Well, we would like to multiply the inverse matrix of g_{ij}.

Definition 13.5.2. *Let (g_{ij}) be a symmetric positive definite $(0,2)$-tensor on $\mathbb{R}^n$. Then we write (g^{ij}) to be the $(2,0)$-tensor such that the matrix with entries g_{ij} and the matrix with entries (g^{ij}) are inverse of each other. (Note that it does NOT matter whether we choose i or j to be the row index or column index, because everything is symmetric.)*

In particular, we have the following identity:

Proposition 13.5.3 (The duality between inner products)**.** *Let (δ_i^j) be the identity map.*

1. *$g^{ij} = g^{ji}$.*
2. *$g^{ij} g_{jk} = \delta_k^i$. (Note that by exploiting the symmetricity of g_{ij} and g^{ij}, we would get many similar identities.)*
3. *$g^{ij} g_{ij} = n$, the dimension of V.*

Proof. Note that (g_{ij}) form a symmetric matrix, and the inverse of a symmetric matrix is symmetric, so (g^{ij}) is symmetric.

By construction, (g^{ij}) and (g_{ij}) are inverse matrix of each other. So $(g^{ij} g_{jk}) = (g^{ij})(g_{jk}) = I = (\delta_k^i)$.

Finally, $g^{ij} g_{ij} = \delta_i^i = \operatorname{trace}(I) = \dim V$. □

Proposition 13.5.4 (Raising the index)**.** *The Riesz map would send the dual vector (a_i) to the vector $(a_i g^{ij})$.*

Proof. We only need to check that this is the inverse of the bra map. We have $a_i g^{ij} g_{jk} = a_i \delta_k^i = a_i$. Yay! □

Remark 13.5.5 (Transpose)**.** *In general, transpose means "raise all lower indices and lower all upper indices." So A^i_j has a transpose of A^j_i, and the standard basis vector $\boldsymbol{e}_i$ has a transpose $\boldsymbol{e}^i$ (i.e., $\boldsymbol{e}_i^{\mathrm{T}}$).*

Now let us consider an inner product structure on V^*. Such an inner product shall eat two DUAL vectors and output a number. Hence it is a $(2,0)$-tensor over V!

Definition 13.5.6. *Given an inner product space V with inner product (g_{ij}), we choose (g^{ij}) as the corresponding inner product on V^*.*

Proposition 13.5.7. *Under the inner product on V and the corresponding inner product on V^*, the bra map and the Riesz map are isometric bijections. (I.e., they preserve inner product in the domain and codomain.)*

Proof. $\langle \langle \boldsymbol{v}|, \langle \boldsymbol{w}| \rangle = (v^i g_{ij})(w^k g_{kl}) g^{jl} = v^i g_{ij} w^k (g_{kl} g^{jl}) = v^i w^k g_{ij} (\delta_k^j) = v^i w^k (g_{ij} \delta_k^j) = v^i w^k g_{ik} = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$. □

Now we are ready to define inner products on tensors. (And then length of and angles between tensors are defined.)

Definition 13.5.8. *We define inner products on (k,t)-tensor space over V as $\langle (T^{i_1,\dots,i_k}_{j_1,\dots,j_t}), (S^{m_1,\dots,m_k}_{n_1,\dots,n_t}) \rangle = T^{i_1,\dots,i_k}_{j_1,\dots,j_k} S^{m_1,\dots,m_k}_{n_1,\dots,n_k} g_{i_1 m_1} \dots g_{i_k m_k} g^{j_1 n_1} \dots g^{j_t n_t}$. Here g_{ij} is the inner product on V.*

In short, we literally just do the inner product for each corresponding components. For example, $\langle \alpha \otimes \boldsymbol{v}, \beta \otimes \boldsymbol{w} \rangle = \langle \alpha, \beta \rangle \langle \boldsymbol{v}, \boldsymbol{w} \rangle$.

13.6 Alternating tensor and alternization

We now focus our study on $(0,k)$-tensors and $(k,0)$-tensors. A $(0,k)$-tensor T is both an element of $(V^*)^{\otimes k}$, and a multilinear map $T : V \times \cdots \times V \to \mathbb{R}$. Dually, a $(k,0)$-tensor T is both an element of $V^{\otimes k}$, and a multilinear map $T : V^* \times \cdots \times V^* \to \mathbb{R}$. In the following discussions, we first and foremost treat them as multilinear maps into $\mathbb{R}$.

Definition 13.6.1. *A $(0,k)$-tensor or a $(k,0)$-tensor is alternating if swapping a pair of inputs would negate the output.*

Determinant is the most prominant example of this. But let us review a bit about what's going on.

Example 13.6.2. In $\mathbb{R}$, given a vector $\boldsymbol{v} \in \mathbb{R}$ (which is secretly just a single number...), we can talk about its length. Note that $\mathbb{R}$ is naturally endowed with a positive axis direction. Therefore, some vectors have positive length, and some have negative length. The oriented-length measurement is simply $\boldsymbol{e}^1$ (i.e., $\boldsymbol{e}_1^{\mathrm{T}}$), the unit dual vector.

Note that for any multilinear map, if it is linear (only one input), then there is no other input to swap with, so we treat it as alternating by convention.

Now in $\mathbb{R}^2$, given a pair of vectors $(\boldsymbol{v}, \boldsymbol{w})$, they form a parallelogram. We say the parallelogram is positively oriented if doing $\boldsymbol{v}$ first and then $\boldsymbol{w}$ is going counter-clockwise around this parallelogram. Then the parallelogram will have an oriented area! Note that $(\boldsymbol{v}, \boldsymbol{w})$ and $(\boldsymbol{w}, \boldsymbol{v})$ would represent the SAME parallologram with OPPOSITE orientation. Therefore the measurement of oriented area is alternating! (And it is also easily seen as bilinear.)

Of course, this area is just the determinant $\det(\boldsymbol{v}, \boldsymbol{w})$. You can review last semester's notes on determinants. By the determinant formula, oriented area is $\boldsymbol{e}^{12} - \boldsymbol{e}^{21}$, where $\boldsymbol{e}^{ij}$ refers to $\boldsymbol{e}_i^{\mathrm{T}} \otimes \boldsymbol{e}_j^{\mathrm{T}}$ as usual.

In $\mathbb{R}^3$, pick three vectors and we have oriented volume, also given by the determinant. And the oriented volume tensor is $\boldsymbol{e}^{123} + \boldsymbol{e}^{231} + \boldsymbol{e}^{312} - \boldsymbol{e}^{132} - \boldsymbol{e}^{213} - \boldsymbol{e}^{321}$. So on so forth to higher dimensions and oriented higher dimensional volumes. ☺

Let us have an exotic example here, which is the "dual" to the ideas above. In the example above, say in $\mathbb{R}^2$, then two one-dimensional things (vectors) span an area (the parallelogram). Dually speaking, in $\mathbb{R}^2$ we can also intersect two one-dimensional things to get a point (zero dimension).

Example 13.6.3. (This example still has some troubles.... Consider $y = x^3$ intersecting with $y = -x^3$.)

Instead of using an a-dim thing and a b-dim thing to span an $(a+b)$-dim thing, we can also intersect an $(n-a)$-dim thing and a $(n-b)$-dim thing to obtain an $(n-a-b)$-dim thing. Here we show a simple example of this where $n = 2$ and $a = b = 1$.

In $\mathbb{R}^2$, curves wouold intersect at points. However, sometimes it is useful to study the oriented-intersections. For example, let's say that γ_1 is in fact a Jordan-curve, i.e., a closed curve that will unambiguously bound a region. (But it might still look windy and complicated.) In short, there are well-defined concepts of "region inside of the curve" and "region outside of the curve". Then given a point $\boldsymbol{p} \in \mathbb{R}^2$, how can we tell if it is inside the region or outside of the region? Well, we just draw a straight ray from $\boldsymbol{p}$ to infinity γ_2. Then when γ_2 intersects with γ_1, it will "cut", then "uncut", then "cut", and then "uncut", and so on. Each time γ_2 will change its status of being "inside" or "outside" of γ_1. If in the end we have even number of cuts, then $\boldsymbol{p}$ is outside of γ_1, and we say the total oriented-intersection of γ_1 and γ_2 is zero.

To formally define this idea, given two curves γ_1 and γ_2 on $\mathbb{R}^2$, if they intersect somewhere, i.e., $\gamma_1(s) = \gamma_2(t)$ for some s, t, then we say they intersect positively here if $\det(\gamma_1'(s), \gamma_2'(t)) > 0$, and intersect negatively here if $\det(\gamma_1'(s), \gamma_2'(t)) < 0$. What if they intersect with $\det(\gamma_1'(s), \gamma_2'(t)) = 0$, i.e., tangentially? Well, note that a tangent intersection is actually always the limit of a pair of oppositely-oriented intersections. So if $\det(\gamma_1'(s), \gamma_2'(t)) = 0$, we do not count this as an intersection.

In particular the total number of intersections of a curve γ with itself is zero. Even though γ intersect with itself everywhere, but all these intersections are tangential. Hence they all do not count. Similarly, from definition you can easily see that the total number of oriented-intersections of γ_1 with γ_2 and the total number of oriented-intersections of γ_2 with γ_1 must be negations of each other.

So this "intersection number" between curves is an "alternating" thing, say if $i(\gamma_1, \gamma_2)$ denotes the total number of oriented-intersections between the two curves, then $i(\gamma_1, \gamma_2) = -i(\gamma_2, \gamma_1)$.

We now make this bilinear. To avoid infinite intersections, suppose we only consider straight lines and quadratic curves. (You can safely generalize this to all algebraic curves.) Then let V be the "formal linear combination of these curves". So elements of V are a finite sum $a_1\gamma_1 + \cdots + a_k\gamma_k$, where you may interpret $k\gamma$ as a "curve where each point has mass k" if you like. Then given a linear combination of curves $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$, then we can ask how many oriented-intersections do they have.

Then we can extend the idea of oriented-intersections linearly to V. Then we have a bilinear alternating map $i : V \to V$ such that $i(\gamma_1, \gamma_2)$ is the total number of oriented-intersections, and $i(a\gamma_1, b\gamma_2) = abi(\gamma_1, \gamma_2)$ is the total oriented-weight of the intersections. ☺

The above examples are all $(0, k)$-tensors. What about $(k, 0)$-tensors?

Example 13.6.4. Suppose we are in $\mathbb{R}^2$. Suppose we have a $(2, 0)$ tensor P. Then the $(0, 2)$-tensor det would be able to evaluate it into a number, i.e., P is an object with "oriented area"!

So given an oriented-parallelogram, P made by two edges $(\boldsymbol{v}, \boldsymbol{w})$ (so the orientation is done by going along $\boldsymbol{v}$ and then $\boldsymbol{w}$), which tensor represent it? Well, note that if we switch the order of the two edges, the orientation of P is flipped. So we want an alternating $(2, 0)$-tensor.

It turns out that there is a unique alternating tensor using $\boldsymbol{v}$ and $\boldsymbol{w}$ such that det would evaluate P to the desired area. For example, consider the positively-oriented unit square. Well, note that $\det(\boldsymbol{v} \otimes \boldsymbol{w}) = \det(\boldsymbol{v}, \boldsymbol{w})$, so $\boldsymbol{e}_{1,2} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_2$ and $-\boldsymbol{e}_{21} = -\boldsymbol{e}_2 \otimes \boldsymbol{e}_1$ are BOTH objects constructed using the edge $\boldsymbol{e}_1, \boldsymbol{e}_2$, so they are both candidates to represent the unit square. In fact, any linear combination of then $a\boldsymbol{e}_{12} - b\boldsymbol{e}_{21}$ with $a + b = 1$ would also represent the unit square. However, most of these are NOT alternating. The only alternating one is $\frac{1}{2}(\boldsymbol{e}_{12} - \boldsymbol{e}_{21})$. So this is what we choose.

In general, you may think of $\frac{1}{2}(\boldsymbol{v} \otimes \boldsymbol{w} - \boldsymbol{w} \otimes \boldsymbol{v})$ as representing the parallelogram made by $(\boldsymbol{v}, \boldsymbol{w})$. ☺

Definition 13.6.5. *Given a $(k, 0)$-tensor T, its **alternization** is $\mathrm{Alt}(T) = \frac{1}{k!}\sum_{\sigma \in S_k} \mathrm{sign}(\sigma)\sigma(T)$, where $\sigma(T)$ means we are permuting the inputs of the multilinear map T by the permutation σ.*

Here S_k is the set of all permutations of k things.

Think of this as an "anti-symmetrization" of T, so that we get an alternized version of T.

Proposition 13.6.6. *In $\mathbb{R}^n$, the map* Alt *from $(k, 0)$-tensors to alternating $(k, 0)$-tensors is a linear projection map. Here projection means $\mathrm{Alt}(\mathrm{Alt}(T)) = \mathrm{Alt}(T)$.*

Proof. It is easy to see that it is linear from the definition. For the second statement, we just need to show that if T is alternating, then $\mathrm{Alt}(T) = T$.

If T is alternating, then $\sigma(T) = \mathrm{sign}(\sigma)T$. So by direct computation we have $\mathrm{Alt}(T) = T$. □

Corollary 13.6.7. *Let $*^k(V)$ be the space of all alternating $(k, 0)$ tensors on V. Then if $V = \mathbb{R}^n$, a basis is $\mathrm{Alt}(\boldsymbol{e}_{i_1,\ldots,i_k})$ where $1 \le i_1 < \cdots < i_k \le n$. In particular, $\dim *^k(V) = \binom{n}{k}$ where $n = \dim V$. Here $\binom{n}{k}$ means the number of ways to pick k elements out of a set of n elements, i.e., $\frac{n!}{k!(n-k)!}$.*

Proposition 13.6.8. *In $\mathbb{R}^n$, for any $(k, 0)$-tensor P and $(0, k)$ ALTERNATING tensor T, then $T(P) = T(\mathrm{Alt}(P))$.*

Proof. Note that T is alternating. So for each $\sigma \in S_k$, $T(\sigma(P)) = \mathrm{sign}(\sigma)T(P)$.

$T(\mathrm{Alt}(P)) = \frac{1}{k!}\sum_{\sigma \in S_k} \mathrm{sign}(\sigma)T(\sigma(P)) = \frac{1}{k!}\sum_{\sigma \in S_k} \mathrm{sign}(\sigma)^2 T(P) = \frac{1}{k!}(k!)T(P)$.

So we are done. □

Corollary 13.6.9. *In $\mathbb{R}^n$, let* det *be the n-dim volume tensor, then $\det(\mathrm{Alt}(\boldsymbol{v}_1 \otimes \cdots \otimes \boldsymbol{v}_n)) = \det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$.*

In general, in $\mathbb{R}^n$, you may think of $\mathrm{Alt}(\boldsymbol{v}_1 \otimes \cdots \otimes \boldsymbol{v}_k)$ as representing the oriented k-dim parallelotope. Here is the final piece that might convince you of this.

Proposition 13.6.10. *In the Euclidean space $\mathbb{R}^n$ (so all vectors and all tensors have "length"), let P be the tensor representing a k-dim parallelotope. Then the k-dim volume of this parallelotope is simply $(\sqrt{k!})\|P\|$.*

Proof. We only prove $k = 2$ here for simplicity. The generic case is pretty much the same.

If P is made of $(\boldsymbol{v}, \boldsymbol{w})$, a pair of orthogonal vectors, then $\|2P\|^2 = \langle \boldsymbol{v} \otimes \boldsymbol{w} - \boldsymbol{w} \otimes \boldsymbol{v}, \boldsymbol{v} \otimes \boldsymbol{w} - \boldsymbol{w} \otimes \boldsymbol{v} \rangle$. Here I use $2P$ to get rid of the annoying alternization coefficient.

Note that $\langle \boldsymbol{v} \otimes \boldsymbol{w}, \boldsymbol{v} \otimes \boldsymbol{w} \rangle = \langle \boldsymbol{w} \otimes \boldsymbol{v}, \boldsymbol{w} \otimes \boldsymbol{v} \rangle = \|\boldsymbol{v}\|^2 \|\boldsymbol{w}\|^2 = \text{Area}(P)^2$, while $\langle \boldsymbol{v} \otimes \boldsymbol{w}, \boldsymbol{w} \otimes \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{w} \rangle^2 = 0$. Hence we are done.

Now let me show that both area and $\|P\|$ are invaraint under shearing. Suppose P is made of $(\boldsymbol{v}, \boldsymbol{w})$, any pair of vectors. Then let P' be the parallelogram made by $(\boldsymbol{v}, \boldsymbol{w} - k\boldsymbol{v})$ for some constant k. Then treating $\boldsymbol{v}$ as the "base", you can see that P, P' have the same "base" and "height", therefore they have the same area.

On the other hand, note that $\text{Alt}(\boldsymbol{v} \otimes (\boldsymbol{w} - k\boldsymbol{v})) = \text{Alt}(\boldsymbol{v} \otimes \boldsymbol{w}) - k\text{Alt}(\boldsymbol{v} \otimes \boldsymbol{v}) = \text{Alt}(\boldsymbol{v} \otimes \boldsymbol{w})$. So P, P' are the same tensor, and $\|P\| = \|P'\|$.

Since any parallelogram can be obtained by shearing a rectangle, we are done. □

Example 13.6.11 (3D Pythagorean theorem)**.** Let us prove the following fact. Given a right tetrahedron, say the three right triangles are on yz, zx, xy-planes respectively, with area S_x, S_y, S_z. And let S be the area of the slant face. Then $S^2 = S_x^2 + S_y^2 + S_z^2$.

How to show this? Let the three edges on the coordinate axis be $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$. Then the three right triangles are represented by tensors $T_x = \frac{1}{2}\text{Alt}(\boldsymbol{u} \otimes \boldsymbol{v}), T_y = \frac{1}{2}\text{Alt}(\boldsymbol{v} \otimes \boldsymbol{w}), T_z = \frac{1}{2}\text{Alt}(\boldsymbol{w} \otimes \boldsymbol{u})$.

The slant triangle is made by $T = \frac{1}{2}\text{Alt}((\boldsymbol{v} - \boldsymbol{w}) \otimes (\boldsymbol{v} - \boldsymbol{u})) = \frac{1}{2}[\text{Alt}(\boldsymbol{v} \otimes \boldsymbol{v}) - \text{Alt}(\boldsymbol{w} \otimes \boldsymbol{v}) - \text{Alt}(\boldsymbol{v} \otimes \boldsymbol{u}) + \text{Alt}(\boldsymbol{w} \otimes \boldsymbol{u})] = T_x + T_y + T_z$. As you can see here, the fact that the four triangles "closed up" is translated into the fact that $T = T_x + T_y + T_z$.

Now we think of $\mathbb{R}^n$ as the Euclidean space. Then $S_x = \sqrt{2}\|T_x\|$ and so on. So we aim to show that $\|T\|^2 = \|T_x\|^2 + \|T_y\|^2 + \|T_z\|^2$.

However, a direct calculation (or a geometric observation) shows that T_x, T_y, T_z are mutually orthogonal tensors. Hence we are done by regular Pythagorean theorem. ☺

Note that the correspondence here between alternating $(k, 0)$-tensors and k-dim parallelotopes is NOT one-to-one, but one-to-many. For example, $\boldsymbol{v} \otimes (2\boldsymbol{w})$ and $(2\boldsymbol{v}) \otimes \boldsymbol{w}$ are the SAME tensor. In the end, a $(k, 0)$ alternating tensor is simply something with a k-dim "direction" (the k-dim subspace spanned by the parallelotope) and a "magnitude" (the k-dim volume), like a generalized concept of vector. It does NOT really care about the specific shape.

For example, $\pi\text{Alt}(\boldsymbol{e}_1 \otimes \boldsymbol{e}_2)$ could either refer to a $1 \times \pi$ rectangle on the xy-plane, or a unit circle on the xy-plane. It only records the "direction", i.e., it lies on the xy-plane, and the "magnitude", which is the "oriented-area".

13.7 Wedge product

We now try to figure out intersections.

Example 13.7.1. Recall that in a vector space V, a dual vector α induces a foliation of V by hyperplanes parallel to $\text{Ker}(\alpha)$.

Now consider $\mathbb{R}^2$ with two dual vectors $\alpha = \boldsymbol{e}^1, \beta = \boldsymbol{e}^2$. If you like, we can also write $\mathrm{d}x, \mathrm{d}y$ for them.

Either way, $\mathrm{d}x$ induces a unit density foliation of $\mathbb{R}^2$ by lines parallel to $x = 0$, i.e., the y-axis. Similarly, $\mathrm{d}y$ induces a unit density foliation of $\mathbb{R}^2$ by lines parallel to $y = 0$, i.e., the x-axis. If we intersect the lines in the $\mathrm{d}x$ foliation with lines in the $\mathrm{d}y$ foliation, what would we get? We would get a foliation of $\mathbb{R}^2$ by points, and we have unit density of points everywhere. Write this as $\mathrm{d}x \wedge \mathrm{d}y$. (In calculus, people simply write $\mathrm{d}x\,\mathrm{d}y$.)

What is this object $\mathrm{d}x \wedge \mathrm{d}y$? Well, for each parallelogram, say represented by the $(2, 0)$ alternating tensor P, then we can ask ourself: how many points of the foliation $\mathrm{d}x \wedge \mathrm{d}y$ are inside of P? Since this is a unit density foliation, we would simply get the oriented-area. In particular, $\mathrm{d}x \wedge \mathrm{d}y$ send $(2, 0)$ tensors to numbers, so it is a $(0, 2)$-tensor itself. In fact, since $\mathrm{d}x \wedge \mathrm{d}y$ means the oriented-area, we have $\mathrm{d}x \wedge \mathrm{d}y = \det$ as a $(0, 2)$-alternating tensor over $\mathbb{R}^2$.

In terms of the standard basis, $\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \boldsymbol{e}^{12} - \boldsymbol{e}^{21}$, since $\det(\boldsymbol{v} \otimes \boldsymbol{w}) = \det(\boldsymbol{v}, \boldsymbol{w}) = v^1 w^2 - v^2 w^1$. ☺

Example 13.7.2. What about $\mathbb{R}^3$? We have standard dual basis $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z$. Then $\mathrm{d}x \wedge \mathrm{d}y$ is the intersection of planes perpendicular to x-axis with planes perpendicular to y-axis. Hence it is a foliation of $\mathbb{R}^3$ by lines parallel to the z-axis (with unit density throughout).

Given any parallelogram, say represented by the $(2,0)$ alternating tensor P over $\mathbb{R}^3$, then we can ask ourself: how many lines in the foliation of $\mathrm{d}x \wedge \mathrm{d}y$ would P cut? As you can imagine, since $\mathrm{d}x \wedge \mathrm{d}y$ have unit density, $\mathrm{d}x \wedge \mathrm{d}y(P)$ is the oriented-area if we project P to the xy-plane. In particular, $\mathrm{d}x \wedge \mathrm{d}y(\mathrm{Alt}(\boldsymbol{v} \otimes \boldsymbol{w})) = \mathrm{d}x \wedge \mathrm{d}y(\boldsymbol{v}, \boldsymbol{w}) = v^1 w^2 - v^2 w^1$ again.

If you want the foliation of $\mathbb{R}^3$ by unit density dots, then you would have to do $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$. Then it measures volumes of parallelopiped. It is the $(0,3)$ alternating tensor det, and we have $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = \boldsymbol{e}^{123} + \boldsymbol{e}^{231} + \boldsymbol{e}^{312} - \boldsymbol{e}^{132} - \boldsymbol{e}^{213} - \boldsymbol{e}^{321}$. ☺

So the wedge product is induced by the geometric idea of taking intersections. Now let us figure out the formula for wedge products.

The basic idea is that we want $\boldsymbol{e}^1 \wedge \cdots \wedge \boldsymbol{e}^n$ to corresponds to the tensor det over $\mathbb{R}^n$. By the big formula, this should be $\boldsymbol{e}^1 \wedge \cdots \wedge \boldsymbol{e}^n = \sum_{\sigma \in S_n} \mathrm{sign}(\sigma) \sigma(\boldsymbol{e}^1 \otimes \cdots \otimes \boldsymbol{e}^n)$. But wait! This almost look like an alternization, except that we do NOT divide by the factorial $n!$.

Therefore, for $(0,1)$-tensors $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^k$, we define $\boldsymbol{a}^1 \wedge \cdots \wedge \boldsymbol{a}^k = \sum_{\sigma \in S_k} \mathrm{sign}(\sigma) \sigma(\boldsymbol{a}^1 \otimes \cdots \otimes \boldsymbol{a}^k) = (k!)\mathrm{Alt}(\boldsymbol{a}^1 \otimes \cdots \otimes \boldsymbol{a}^k)$.

What if we have a $(0,a)$-tensor T and a $(0,b)$-tensor K? Well, consider $(\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 \wedge \boldsymbol{e}^3) \wedge (\boldsymbol{e}^4 \wedge \boldsymbol{e}^5)$, which we would like to become $\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 \wedge \boldsymbol{e}^3 \wedge \boldsymbol{e}^4 \wedge \boldsymbol{e}^5$. Then we would want the following identity

$$[(3!)\mathrm{Alt}(\boldsymbol{e}^1 \otimes \boldsymbol{e}^2 \otimes \boldsymbol{e}^3)] \wedge [(2!)\mathrm{Alt}(\boldsymbol{e}^4 \otimes \boldsymbol{e}^5)] = (5!)\mathrm{Alt}(\boldsymbol{e}^1 \otimes \boldsymbol{e}^2 \otimes \boldsymbol{e}^3 \otimes \boldsymbol{e}^4 \otimes \boldsymbol{e}^5).$$

So we make this definition here:

Definition 13.7.3. *For a $(0,a)$-tensor T and a $(0,b)$-tensor K, we define $T \wedge K$ as $\frac{(a+b)!}{a!b!}\mathrm{Alt}(T \otimes K)$.*

Proposition 13.7.4. *For any non-negative integer a, b, the wedge map $\wedge : *_a(V) \times *_b(V) \to *_{a+b}(V)$ is a bilinear alternating map. It is also associative as a binary operation.*

We skip the proof because it is just boring calculations.

Some other things to keep in mind is that the tensors $\boldsymbol{e}^{i_1} \wedge \cdots \wedge \boldsymbol{e}^{i_k}$ for $1 \leq i_1 < \cdots < i_k \leq n$ form the standard basis of $*_k(\mathbb{R}^n)$. Note that we usually write $*^k(V)$ for alternating $(k,0)$ tensors, and $*_k(V)$ for alternating $(0,k)$ tensors. It is also easy to verify that $*^k(V)$ and $*_k(V)$ are canonically dual to each other.

13.8 Differential form and exterior derivative

Definition 13.8.1. *Given a differential set M, we say ω is a differential k-form if it is a smooth alternating $(0,k)$ tensor field on M. (The word "smooth" here is tricky. But we shall clarify later.) In particular, for each $\boldsymbol{p} \in M$, ω_p is an element of $*_k(T_{\boldsymbol{p}}M) = *^k(T_{\boldsymbol{p}}M)^*$.*

Differential k-forms on M form a vector space, which we shall denote as $\Omega^k(M)$. The use of upper index here might be a bit annoying, but it is so for topological reasons.

Example 13.8.2. A differential 0-form is assigning each point a $(0,0)$-tensor, i.e., a number. So it is simply a function. $\Omega^0(M)$ is the space of all smooth functions $X \to \mathbb{R}$. Here smooth means infinitely differentiable.

A differenital 1-form is assigning each point a $(0,1)$-tensor, i.e., a covector. So it is simply a covector field. $\Omega^1(M)$ is the space of all smooth covector fields ω on X. Note that if $X = \mathbb{R}^n$, then $\omega = \omega_i \, \mathrm{d}x^i$ as per the Einstein's notation. Here $\mathrm{d}x^i$ is the covector field that pick the covector $\boldsymbol{e}^i$ everywhere, and $\omega_i : X \to \mathbb{R}$ are functoions. The smoothness of ω refers to the fact that all ω_i are infinitely differentiable. ☺

Example 13.8.3. If X is an open subset of $\mathbb{R}^n$, then we may simply choose the alternating $(0,k)$-tensor $\boldsymbol{e}^{i_1} \wedge \cdots \wedge \boldsymbol{e}^{i_k}$ everywhere. We write this differential k-form as $\mathrm{d}x^{i_1, \ldots, i_k}$.

Then at each point, these $\mathrm{d}x^{i_1,\dots,i_k}$ form a basis for the corresponding alternating tensor space. So for any differential k-form ω on X, then $\omega_{\boldsymbol{p}}$ is a linear combination of $(\mathrm{d}x^{i_1,\dots,i_k})_{\boldsymbol{p}}$. Note that at different $\boldsymbol{p}$, the coefficient for the linear combination might be different though.

So we have $\omega = \omega_{i_1,\dots,i_k}\,\mathrm{d}x^{i_1,\dots,i_k}$ as per the Einstein's notation, and $\omega_{i_1,\dots,i_k} : X \to \mathbb{R}$ is some function on X. The smoothness of ω means all $\omega_{i_1,\dots,i_k}$ are infinitely differentiable. ☺

Now intuitively, you may think of a differentiable k-form on an n-dim manifold M as some $(n-k)$-dim foliations of M with density. And given two differential form ω, η, we may take their wedge product $\omega \wedge \eta$ such that $(\omega \wedge \eta)_{\boldsymbol{p}} = \omega_{\boldsymbol{p}} \wedge \eta_{\boldsymbol{p}}$. Then wedge products of differential forms corresponds to intersections of foliations, just like before.

Remark 13.8.4. *The paragraph above is just an intuitive guide. It is entirely accurate for $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$, but starting from $\mathbb{R}^4$ there might be some weird things. For example, $\dim \Lambda^2(\mathbb{R}^4) = 6$. However, the "2-dim planes in $\mathbb{R}^4$" form a geometric object with dimension 4 (the Grassmanian manifold). Therefore, some alternating $(2,0)$-tensors cannot be seen as "parallelograms".*

Nevertheless, you may think of elements of $\Lambda^2(\mathbb{R}^4)$ as linear combinations of parallelograms. Then everything is fine again. It is the same thing for differential forms. Some ω cannot be realized as actual foliations. But it is always a linear combination of things that could be realized as actual foliations. Then $\omega \wedge \eta$ is still the corresponding intersection (of linear combinations of foliations).

Example 13.8.5. In $\mathbb{R}^3$, a differential 0-form is some function f.

A differential 1-form is something like $f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z$. People also write this as $\begin{bmatrix} f \\ g \\ h \end{bmatrix} \cdot \mathrm{d}\boldsymbol{r}$. It is a foliation of $\mathbb{R}^3$ by surfaces, and the surface has normal vector $\begin{bmatrix} f \\ g \\ h \end{bmatrix}$ everywhere, and density $\| \begin{bmatrix} f \\ g \\ h \end{bmatrix} \|$ everywhere.

A differential 2-form is something like $f\,\mathrm{d}y \wedge \mathrm{d}z + g\,\mathrm{d}z \wedge \mathrm{d}x + h\,\mathrm{d}x \wedge \mathrm{d}y$. Note that in calculus, people usualy obmit the wedge sign, and simply write stuff like $\mathrm{d}x\,\mathrm{d}y$ instead. People also write the 2-form as $\begin{bmatrix} f \\ g \\ h \end{bmatrix} \cdot \mathrm{d}S$. It is a foliation of $\mathbb{R}^3$ by curvess, and the curves has tangent vector $\begin{bmatrix} f \\ g \\ h \end{bmatrix}$ everywhere, and density $\| \begin{bmatrix} f \\ g \\ h \end{bmatrix} \|$ everywhere.

A differential 3-form is something like $f\,\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$. It is a foliation of $\mathbb{R}^3$ by points, with density f everywhere.

On a k-dim "integrable oriented object" M, whatever that means, we can then integrate a differential k-form $\int_M \omega$, which counts how many layers of the $(n-k)$-dim foliation by ω is cut by the k-dim object M (counting orientation, of course). ☺

Example 13.8.6. An interesting yet somewhat disturbing fact is $\mathrm{d}x \wedge \mathrm{d}y = -\,\mathrm{d}y \wedge \mathrm{d}x$. However, I'm sure calculus taught you the formula $\int_a^b \int_c^d f\,\mathrm{d}x\,\mathrm{d}y = \int_c^d \int_a^b f\,\mathrm{d}y\,\mathrm{d}x$. Well, why don't they agree with each other?

The hidden reason is that when we learn $\int_a^b \int_c^d f\,\mathrm{d}x\,\mathrm{d}y = \int_c^d \int_a^b f\,\mathrm{d}y\,\mathrm{d}x$, we were not careful about the orientation. $\int_a^b \int_c^d$ refers to a parallelogram with side vectors $((b-a)\boldsymbol{e}_1, (d-c)\boldsymbol{e}_2)$, while $\int_c^d \int_a^b$ refers to a parallelogram with side vectors $((d-c)\boldsymbol{e}_2, (b-a)\boldsymbol{e}_1)$. Hey! They are the same parallelogram with the OPPOSITE orientation!

So the proper way to do this is like this: $\int_a^b \int_c^d f\,\mathrm{d}x\,\mathrm{d}y = \int_P f\,\mathrm{d}x\,\mathrm{d}y = -\int_P f\,\mathrm{d}y\,\mathrm{d}x = \int_{-P} f\,\mathrm{d}y\,\mathrm{d}x = \int_c^d \int_a^b f\,\mathrm{d}y\,\mathrm{d}x$. Here P refers to the oriented parallelogram for $\int_a^b \int_c^d$, i.e., the parallelogram $[a,b] \times [c,d] \subseteq \mathbb{R}^2$ with positive orientation. ☺

Example 13.8.7. Consider a smooth function $f : \mathbb{R}^2 \to \mathbb{R}$. What is $\mathrm{d}f \wedge \mathrm{d}y$? Note that $\mathrm{d}f = f_x\,\mathrm{d}x + f_y\,\mathrm{d}y$ where f_x, f_y are the partial derivatives. So $\mathrm{d}f \wedge \mathrm{d}y = (f_x\,\mathrm{d}x + f_y\,\mathrm{d}y) \wedge \mathrm{d}y = f_x\,\mathrm{d}x \wedge \mathrm{d}y + f_y\,\mathrm{d}y \wedge \mathrm{d}y = f_x\,\mathrm{d}x \wedge \mathrm{d}y$.

Here $\mathrm{d}y \wedge \mathrm{d}y = 0$ because the wedge product is alternating (and because the foliation curves of $\mathrm{d}y$ is tangent to itself everywhere).

$\mathrm{d}f$ is the foliation by level curves of f, while $\mathrm{d}y$ is the foliation by lines parallel to the x-axis. You can indeed see that the density of $\mathrm{d}f$ foliation curves in the x-direction f_x is the density of the intersections. ☺

Now let us define exterior derivatives. Given a differential k-form, interpreted as some $(n-k)$-dim foliations, then we can take the boundary of each foliation layer, and get a $(n-k-1)$-dim foliation, i.e., a differential $(k+1)$-form. This operation is called the exterior derivative $\mathrm{d}\omega$ of a differential form ω. Let us see how one might define this.

First of all, $\mathrm{d}\colon \Omega^k(M) \to \Omega^{k+1}(M)$ must be linear.

Furthermore, we know already that, when $k = 0$, we want $\mathrm{d}(f)$ to be $f_x\,\mathrm{d}x + f_y\,\mathrm{d}y + f_z\,\mathrm{d}z$. This could be the start of some inductive definition.

Also note that, since d is taking boundary of the layers of the foliation, $\mathrm{d}\circ\mathrm{d} = 0$. This is tied to the geometric fact that the boundary of boundary is always empty, i.e., $\partial(\partial M) = \varnothing$. (E.g., if a curve is not even closed up, then how could it bound a region? If a surface is not even closed up, how could it bound a volume?)

The last ingredient, as with all derivatives, is the Leibniz rule.

Definition 13.8.8. *We define the exterior derivative* d *to be the unique linear map such that:*

1. *(Start of the induction.)* $\mathrm{d}(f)$ *is the covector field* $\mathrm{d}f$ *for all* $f \in \Omega^0(M)$.
2. *(Boundary of the boundary is empty.)* $\mathrm{d}\circ\mathrm{d} = 0$ *always.*
3. *(Leibniz rule.)* $\mathrm{d}(\alpha\wedge\beta) = (\mathrm{d}\alpha)\wedge\beta + (-1)^p\alpha\wedge(\mathrm{d}\beta)$ *where* α *is an* p*-form.*

Example 13.8.9 (An explicit formula for the exterior derivative)**.** Consider $f\,\mathrm{d}x\,\mathrm{d}y$ on $\mathbb{R}^3$. This is a foliation of $\mathbb{R}^3$ by lines or rays or line segments parallel to the z-axis. At a point $\boldsymbol{p}$, if f increases in the z-direction, i.e., $f_z > 0$, then it means more rays along the positive z-axis direction are emerging (positive boundary point). If f decreases in the z-direction, i.e., $f_z < 0$, then it means rays along the positive z-axis direction are ending (negative boundary point).

In particular, it only makes sense that $\mathrm{d}(f\,\mathrm{d}x\,\mathrm{d}y) = f_z\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y = -f_z\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}y = f_z\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

In general, if we have $f\,\mathrm{d}x^I$ for some $I = (i_1,\dots,i_k)$, then $\mathrm{d}(f\,\mathrm{d}x^I) = (\mathrm{d}f)\wedge(\mathrm{d}x^I)$, since $\mathrm{d}(\mathrm{d}x^I)$ is a double d, hence it is zero. Since all k-forms are linear combinations of things like $f\,\mathrm{d}x^I$, this actually gives a more explicit formula to calculate exterior derivatives on $\mathbb{R}^n$. ☺

Example 13.8.10 (Gradient, Curl, Divergence)**.** The following three computations reveals that tha nature of gradient, curl and divergence are ALL the same: they are all exterior derivatives in disguise!

$\mathrm{d}f = f_x\,\mathrm{d}x + f_y\,\mathrm{d}y + f_z\,\mathrm{d}z$, where f_x, f_y, f_z refers to the corresponding partial derivatives. This is the "gradient" process. In the language of calculus, we have $\mathrm{d}f = (\nabla f)\cdot\mathrm{d}\boldsymbol{r}$.

$\mathrm{d}(f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z) = (h_y - g_z)\,\mathrm{d}y\,\mathrm{d}z + (f_z - h_x)\,\mathrm{d}z\,\mathrm{d}x + (g_x - f_y)\,\mathrm{d}x\,\mathrm{d}y$. This is the "curl" process. In the language of calculus, we have $\mathrm{d}(\boldsymbol{F}\cdot\mathrm{d}\boldsymbol{r}) = (\nabla\times\boldsymbol{F})\cdot\mathrm{d}\boldsymbol{S}$.

$\mathrm{d}(f\,\mathrm{d}y\,\mathrm{d}z + g\,\mathrm{d}z\,\mathrm{d}x + h\,\mathrm{d}x\,\mathrm{d}y) = (f_x + g_y + h_z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$. This is the "divergence" process. In the language of calculus, we have $\mathrm{d}(\boldsymbol{F}\cdot\mathrm{d}\boldsymbol{S}) = (\nabla\cdot\boldsymbol{F})\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

As you can see, calculus class basically would try to do these in a way to avoid mention tensors. They end up needing many different names for these different formulas. But the concept of tensor unify all these calculations.

Furthermore, since $\mathrm{d}\circ\mathrm{d} = 0$, we immediately see that $\nabla\times\nabla f = 0$ and $\nabla\cdot(\nabla\times\boldsymbol{F}) = 0$. Also note that $\nabla\times\nabla f = 0$ is the SAME statement as the fact that all mixed derivatives are the same. All of these are consequences of $\mathrm{d}\circ\mathrm{d} = 0$.

Finally, the Green's theorem, Stoke's theorem and Gauss theorem are all unified under the same equation: $\int_{\partial M}\omega = \int_M \mathrm{d}\omega$. In some sense, this is pointing out the fact that the geometric "boundary" map ∂ and the algebraic "boundary" map d are dual to each other. ☺

Remark 13.8.11 (Optional Proof of Stoke's theorem)**.** *See Class Video 142-3.*

13.9 Poincaré duality and de Rham cohomology

Note that we have a combinatorial identity $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{n-k}$. In particular, the spaces $\Lambda_k(\mathbb{R}^n)$ and $\Lambda_{n-k}(\mathbb{R}^n)$ have the SAME DIMENSION! Do they have some sorts of correspondence?

Example 13.9.1. If we switch up k-dim things with $(n-k)$-dim things, there would be some fun phenomena. An interesting examlpe are the platonic solids, i.e., the most "regular" possible 3D shapes.

A regular polygon is a polygon where all side lengths and all inner angles are the same, i.e., equilateral triangles, squares, regular pentagons etc.. A platonic solid is a 3D polyhedron where all faces are the same regular polygon of the same sizes, and all vertices looks the same. So, in some sense, this is a 3D shape with maximal symmetry.

(Note that even though polyhedrons are 3D shapes, i.e., shapes in $\mathbb{R}^3$, we actually treat them as 2-dim "piecewise-differentiable" surfaces.)

The most well-known examples of pPlatonic solids are tetrahedrons (made of four triangles) and cubes (made of six squares). Lesser known Platonic solids are octahedrons (made of eight triangles), dodecahedron (made of 12 pentagons. Here "do-dec-ahedron" breaks down into "do", which means two [e.g. double], and "dec", which means 10 [e.g. "decade"], so do-dec-ahedron means "12-polyhedron"), and the icosahedron (made of 20 triangles).

Well, it turns out that this is it! There are only five Platonic solids, no more. You might take some topology or graph theory class to see this.

(Also for those interested in dodecahedrons, I highly recomment the exposition website `https://math.ucr.edu/home/baez/dodecahedron/1.html`.)

Let us switch up k-dim things with $(n-k)$-dim things. Since we treat these things as 2-dim "piecewise-differentiable" surfaces, $n = 2$.

Given a platonic solid P, for each face, place a vertex in the center, and connect these vertices. You will obtain the dual Platonic solid P^*. If you do this again, you will realize that P^{**} is just the shap P again. (Try this on the cube to get an octahedron.)

By comparing P and P^*, you can see that each k-dim thing of P corresponds to some $(n-k)$-dim thing of P^*! Furthermore, if an edge e is on the boundary of a face f of P, then the point f^* is on the boundary of the edge e^* of P^*. So "boundary" maps goes in the opposite direction! This duality is not just some one-to-one correspondence of objects, but in fact a total duality of the internal geometric structure.

Under this idea, the cube and the octahedron are dual, the dodecahedron and the icosahedron are dual, and finally, the tetrahedron is self-dual. ☺

As you can feel from the example of the Platonic solids (which are 2-dim surfaces), k-dim things with $(n-k)$-dim things should in some sense be the "dual" of each other. This is not just a geometric duality. Here is a combinatorial duality to think about.

Example 13.9.2. How many k-element subset does an n-element set S has? Well, the answer is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. However, how many $(n-k)$-element subset doew an n-element set has? Well, the answer is $\binom{n}{n-k} = \frac{n!}{k!(n-k)!}$.

Furthermore, there is indeed a "mutual evaluation" going on. Let S_k be the set of k-element subsets of S, and let S_{n-k} be the set of $(n-k)$-element subsets of S. Then eash element $x \in S_{n-k}$ would send elements $y \in S_k$ to the number $|x \cap y|$, the number of elements in the intersection. This way, we see that S_k, S_{n-k} is indeed "dual" to each other, in the sense that they send each other to numbers. ☺

Now consider the fact that $\dim \Lambda_k(\mathbb{R}^n) = \binom{n}{k} = \binom{n}{n-k} = \dim \Lambda_{n-k}(\mathbb{R}^n)$, it is very natural to conjecture that the two are actually dual spaces of each other! I.e., alternating k-tensors should be naturally dual to alternating $(n-k)$-tensors.

But how is this duality achieved? Well, as suggested by the combinatorial case, it should be done via "intersection", which in the case of alternating tensors means the wedge product.

Consider the wedge map $\wedge : \Lambda_k(\mathbb{R}^n) \times \Lambda_{n-k}(\mathbb{R}^n) \to \Lambda_n(\mathbb{R}^n)$. What is this codomain? Well, $\dim \Lambda_n(\mathbb{R}^n) = 1$, so it is $\mathbb{R}$. Therefore elements of $\Lambda_k(\mathbb{R}^n)$ and $\Lambda_{n-k}(\mathbb{R}^n)$ send each other to $\mathbb{R}$, i.e., they are dual to each other!

Before we go into the proof, let us standardize a notation. For any multi-index $I = (i_1, \dots, i_k)$, we use ξ^I to denote $\boldsymbol{e}^{i_1} \wedge \cdots \wedge \boldsymbol{e}^{i_k}$. Then a basis for $\Lambda_k(\mathbb{R}^n)$ is ξ^I for all strictly increasing k-multi-indices I.

Proposition 13.9.3. *We have a linear isomorphism $P : \Lambda_{n-k}(\mathbb{R}^n) \to \Lambda^k(\mathbb{R}^n)$ such that, for each $\omega \in \Lambda_{n-k}(\mathbb{R}^n)$ and $\eta \in \Lambda_k(\mathbb{R}^n)$, we have $\omega \wedge \eta = P(\omega)(\eta) \det$. Here $\det$ is the determinant tensor in $\Lambda_n(\mathbb{R}^n)$.*

Proof. Since $\Lambda_n(\mathbb{R}^n)$ is one-dimensional and $\det \in \Lambda_n(\mathbb{R}^n)$ is non-zero, therefore any element of $\Lambda_n(\mathbb{R}^n)$ is a unique multiple of $\det$. This gives a linear bijection $L : \Lambda_n(\mathbb{R}^n) \to \mathbb{R}$.

We now define $P = L \circ \wedge$, so $P(\omega, \eta) = L(\omega \wedge \eta)$ is a bilinear map. Then $P : \Lambda_{n-k}(\mathbb{R}^n) \times \Lambda_k(\mathbb{R}^n) \to \mathbb{R}$ is bilinear. So $P \in \mathcal{M}(\Lambda_{n-k}(\mathbb{R}^n), \Lambda_k(\mathbb{R}^n); \mathbb{R}) = \mathcal{L}(\Lambda_{n-k}(\mathbb{R}^n); \Lambda_k(\mathbb{R}^n))$. So we may treat P as a linear map from $\Lambda_{n-k}(\mathbb{R}^n)$ to $\Lambda^k(\mathbb{R}^n)$. By construction, we have $\omega \wedge \eta = P(\omega)(\eta) \det$.

It remains to show that $P : \Lambda_{n-k}(\mathbb{R}^n) \to \Lambda^k(\mathbb{R}^n)$ is a bijection. We already know that the domain and codomain have the same dimension. So we only need injectivity.

Suppose $P(\omega) = 0$, then $\omega \wedge \eta = 0$ for all $\eta \in \Lambda_k(\mathbb{R}^n)$. But this means $\omega \wedge \xi^I = 0$ for all k-multi-indices I. In particular, by the lemma below, this means under the basis ξ^I for $\Lambda_{n-k}(\mathbb{R}^n)$, all coordinates of ω are zero. So we are done. □

Lemma 13.9.4. *Suppose $\omega = a_I \xi^I$ where I ranges over all possible strictly increasing $(n-k)$-multi-indices. For any strictly increasing k-multi-index J, let J' be the strictly increasing $(n-k)$-multi-index such that J, J' has no common index. (Note that such J' must be unique.) In particular, (J', J) is a permutation of $(1, \dots, n)$.*

Then $\omega \wedge \xi^J = (\operatorname{sign}(J, J') a_{J'}) \det = \pm a_{J'} \det$. In particular, if $\omega \wedge \xi^J = 0$ for all J, then all coordinates a_I of ω are zero, and thus $\omega = 0$.

Proof. Note that for any strictly increasing $(n-k)$-multi-indices I, either it has common index with J and thus $\xi^I \wedge \xi^J = 0$. Or it does not has common index with J, and thus $I = J'$.

So $\omega \wedge \xi^J = a_I \xi^I \wedge \xi^J = a_I \xi^{(I,J)} = a_{J'} \xi^{J',J}$. Since $\xi^{J',J}$ is the result of the permutation (J', J) acting on $\det$, we see that $a_{J'} \xi^{J',J} = (\operatorname{sign}(J, J') a_{J'}) \det = \pm a_{J'} \det$. □

The isomorphism P here is called the Poincaré duality, i.e., the k and $(n-k)$ alternating tensor space are dual to each other.

In general, for any abstract vector space V, then $\Lambda^n V^*$ is always one dimensional, and thus isomorphic to $\mathbb{R}$. However, there is no canonical isomorphism if we do not pick a basis. For example, for any linearly independent vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, then if we pick these as basis to do determinant, then $\det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = 1$. But if we pick $\boldsymbol{v}_2, \boldsymbol{v}_1, \boldsymbol{v}_3, \boldsymbol{v}_4, \dots, \boldsymbol{v}_n$ as the standard basis, then $\det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = -1$.

Remark 13.9.5. *If we treat determinant as a function on matrices or linear maps, then they are independent of "basis change for matrices". However, if we treat it as a multilinear map on columns, then it depends on the basis of each column. A change of basis for these columns is like $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \mapsto B(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ for the change of coordinate matrix B, and thus $\det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ will change into $\det(B) \det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$, which is different.*

Definition 13.9.6. *An orientation on V is a choice of non-zero vector ω on $\Lambda_n V$. (Note that $\Lambda_n V$ is a one-dimensional space, so it is spanned by any non-zero vector.)*

You might intuitively think of this as "declaring unit volume". Elements of $\Lambda_n V$ are sending $\Lambda^n V$ to numbers, i.e., they send parallelotope to numbers in an alternating multilinear way (i.e., multiples of the volume measurement). We know $\Lambda_n V$ is isomorphic to $\mathbb{R}$, but which one do you pick to be 1? The ω serves the purpose of 1, i.e., we choose it as the standard volume measurement of the parallelotopes.

For any linearly independent vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, let its dual basis be $\alpha^1, \dots, \alpha^n$. Then choosing $\alpha_1 \wedge \cdots \wedge \alpha_n$ as orientation is the same as declaring that we treat the volume of the parallelotope $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ as the unit volume. (Because using $\alpha_1 \wedge \cdots \wedge \alpha_n$ as the multilinear map results in $(\alpha_1 \wedge \cdots \wedge \alpha_n)(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = 1$.)

Proposition 13.9.7. *Let V be an oriented vector space with orientation $\omega \in \Lambda_n V$. Then we have a linear isomorphism $P : \Lambda_{n-k}(\mathbb{R}^n) \to \Lambda^k(\mathbb{R}^n)$ such that, for each $\eta_1 \in \Lambda_{n-k}(\mathbb{R}^n)$ and $\eta_2 \in \Lambda_k(\mathbb{R}^n)$, we have $\eta_1 \wedge \eta_2 = P(\omega)(\eta)\omega$.*

Example 13.9.8. An n-dim manifold or differential set M is said to be oriented if we pick a non-vanishing smooth n-form on it, called the orientation form. (Note that an n-form is a $(0,n)$-alternating tensor field. So this means we "smoothly" picked an orientation for all $T_{\boldsymbol{p}}M$, where "smooth" means adjacent points $\boldsymbol{p}$ has adjacent orientation for $T_{\boldsymbol{p}}M$, hopefully in a differentiable manner.)

Say the orientation is ω. Then for any k-form η_1 and $(n-k)$-form η_2, at each point, $\eta_1 \wedge \eta_2$ must be a multiple of ω. So as a whole, $\eta_1 \wedge \eta_2 = f\omega$ for some function f. This is the Poincaré duality for manifolds.

For example, consider when $M = \mathbb{R}^3$ and the orientation is the standard $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$. Then note that 0-forms and 3-forms can BOTH be described by a single function, while 1-forms and 2-forms can BOTH be described by three functions. Poincaré duality is the main reason for this.

(This is used rather subtly. For example, we say we want to integrate a function f on M. But if M is a k-dimensional domain, then you can only integrate k-forms, not 0-forms. So what do we mean? Well, if M is oriented by a non-vanishing k-form ω, then what we mean is $\int_M f\omega$. Inside $\mathbb{R}^3$, we almost always use $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ as the orientation form, so integrationg a function f just means $\int f \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.)

Now, sometimes it is IMPOSSIBLE to pick an orientation form. For example, on the Mobius strip, any n-form must vanish at some point, hence it is NOT orientable.

Integrations on things such as the Mobius strip will have SERIOUS troubles. For example, the Stoke's theorem is no longer true. See for example `https://sites.icmc.usp.br/szani/sma332/material/moebius.pdf`. ☺

13.10 Hodge dual and Maxwell's equation

The Poincaré duality states that $\Lambda_k(\mathbb{R}^n)$ and $\Lambda_{n-k}(\mathbb{R}^n)$ are dual spaces of each other. However, in cases such as $\mathbb{R}^n$, we actually have an inner product structure on all tensor spaces. Hence each space is isomorphic to its dual via Riesz representation theorem!

Of course, first we shall standardize the inner product structure on $\Lambda_k(\mathbb{R}^n)$. This is slightly different from the tensor inner product. For example, the tensor $\xi^{12} = \boldsymbol{e}^{12} - \boldsymbol{e}^{21}$ has length $\sqrt{2}$, yet we prefer it to form an orthonormal basis. So from now on, to avoid unnecessary ugly constants, we use the following inner product.

Definition 13.10.1. *For $\eta_1, \eta_2 \in \Lambda_k(\mathbb{R}^n)$, we define $\langle \eta_1, \eta_2\rangle$ to be $\frac{1}{k!}\langle \eta_1, \eta_2\rangle_\otimes$, where $\langle -,-\rangle_\otimes$ is the regular tensor inner product.*

You may also think of this as this: for skew-symmetric matrices, we do not really need all $n^2 - n$ non-diagonal entries. We only need half of that, say the upper half. In this sense, the fact that $\langle \xi^{12}, \xi^{12}\rangle_\otimes = 2$ instead of 1 is exactly caused by using those redundant coordinates. For alternating k-tensors, the "redundant repetition" is $k!$ fold. If we simply ignore the "redundant coordinates", and only use "essential coordinates" to calculate inner product, then the inner product will be shrunk by the factor $\frac{1}{k!}$.

Proposition 13.10.2. *For $\alpha^1, \dots, \alpha^k, \beta^1, \dots, \beta^k \in \Lambda_1(\mathbb{R}^n)$, we have $\langle \alpha^1 \wedge \cdots \wedge \alpha^k, \beta^1 \wedge \cdots \wedge \beta^k\rangle = \det((\langle \alpha^i, \beta^j\rangle))$.*

Proof. This is a direct computation. Time to test your mastery of the determinant big formula!

Alternatively, you may verify that both $\frac{1}{k!}\langle -,-\rangle_\otimes$ and the determinant definition above are inner products, and ξ^I form an orthonormal basis for all k-multi-indices I. □

Hence, for inner product spaces, we should have a "compatible" isomorphism between $\Lambda_k(\mathbb{R}^n)$ and $\Lambda_{n-k}(\mathbb{R}^n)$. This is Hodge duality. We write this isomorphism as the ***Hodge star operator*** $* : \Lambda_k(\mathbb{R}^n) \to \Lambda_{n-k}(\mathbb{R}^n)$.

Definition 13.10.3. *For any n-dim inner product space V with orientation $\omega \in \Lambda_n V$, we define the Hodge dual of a $(0,k)$-alternating tensor η to be a $(0, n-k)$-alternating tensor $*(\eta)$, such that $\eta' \wedge *(\eta) = \langle \eta', \eta\rangle\omega$ for all $(0,k)$-alternating tensor η'.*

To figure out $*(\eta)$, usually we just take η' to be all possible standard basis tensors, and we shall have $*(\eta)$. To keep things simple, let us focus on $\mathbb{R}^3$, where the Hodge duality is defined in a very simple manner.

Example 13.10.4. The Hodge duality identifies $\Lambda_k(\mathbb{R}^n)$ and $\Lambda_{n-k}(\mathbb{R}^n)$, and thus they identifies k-forms with $(n-k)$-forms. In the context of $\mathbb{R}^3$, it works like this:

For any 0-form (i.e., functions) f, we have $*f = f\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

For any 1-form $f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z$, we have $*(f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z) = f\,\mathrm{d}y\,\mathrm{d}z + g\,\mathrm{d}z\,\mathrm{d}x + h\,\mathrm{d}x\,\mathrm{d}y$. (We simply send $\boldsymbol{e}^i$ to $\boldsymbol{e}^{i+1} \wedge \boldsymbol{e}^{i+2}$.)

For any 2-form $f\,\mathrm{d}y\,\mathrm{d}z + g\,\mathrm{d}z\,\mathrm{d}x + h\,\mathrm{d}x\,\mathrm{d}y$, we have $*(f\,\mathrm{d}y\,\mathrm{d}z + g\,\mathrm{d}z\,\mathrm{d}x + h\,\mathrm{d}x\,\mathrm{d}y) = f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z$.

Finally, for any 3-form $f\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$, we have $*(f\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z) = f$.

As you can see clearly, the "Hodge" process goes like this. Given a 1-form (i.e., a covector field), it is a foliation of $\mathbb{R}^3$ by surfaces. Note that these surfaces have NORMAL VECTORS $\begin{bmatrix} f \\ g \\ h \end{bmatrix}$ at each corresponding point.

By taking Hodge star, the 1-form is sent to the two form $f\,\mathrm{d}y\,\mathrm{d}z + g\,\mathrm{d}z\,\mathrm{d}x + h\,\mathrm{d}x\,\mathrm{d}y$. This is a foliation of $\mathbb{R}^3$ by curves, and the curves have tangent direction $\begin{bmatrix} f \\ g \\ h \end{bmatrix}$ at each corresponding point. In particular, the foliation of $\eta = f\,\mathrm{d}x + g\,\mathrm{d}y + h\,\mathrm{d}z$ and the foliation of $*(\eta)$ are EVERYWHERE PERPENDICULAR to each other.

Finally, since both have coordinates f, g, h, the foliation of η and the foliation of $*(\eta)$ must have the SAME density everywhere.

So the Hodge star is very geometric. Given a foliation, we simply take the "orthogonal complement" everywhere, while trying to preserve density.

Think about foliation of $\mathbb{R}^3 - \{\mathbf{0}\}$ by concentric spheres around the origin, and foliation of $\mathbb{R}^3 - \{\mathbf{0}\}$ by rays shooting away from the origin. Can you see that they are Hodge dual of each other? ☺

Example 13.10.5. In the context of $\mathbb{R}^2$, it works like this:

For any 0-form (i.e., functions) f, we have $*f = f\,\mathrm{d}x\,\mathrm{d}y$. And for 2-forms, things just go back.

For any 1-form $f\,\mathrm{d}x + g\,\mathrm{d}y$, we have $*(f\,\mathrm{d}x + g\,\mathrm{d}y) = f\,\mathrm{d}y - g\,\mathrm{d}x$, as you can verify computationally.

But hey! Note that $*(*(f\,\mathrm{d}x + g\,\mathrm{d}y)) = -(f\,\mathrm{d}x + g\,\mathrm{d}y)$. It fails to go back!

This should be somewhat disturbing. Why won't it go back? The reason is orientation. Hodge dual is not just the Riesz identification of dual spaces. It is a combination of Riesz and the Poincaré duality, i.e., it involves the orientation form. In particular, we have the following formula.

In particular, $\mathrm{d}x$ is a foliation of $\mathbb{R}$ by rays going to the right. Then when we take "orthogonal complement" everywhere, we try to rotation counter-clockwise by 90 degree, and obtian the foliation of $\mathrm{d}y$. But if we do the Hodge star agian, then we rotate by 90 degree again, and we actually obtained $-\mathrm{d}x$ instead.

So if η are concentric circles around the origin counter clockwise, then $*(\eta)$ would be rays INTO the origin (since motion direction is rotated counter clockwise by 90 degree). ☺

So, even though the Hodge star is an isomorphism, it is technically "not a duality", in the sense that it is not its own inverse. It is only a duality up to a sign sometimes.

Proposition 13.10.6. *$*(*(\eta)) = (-1)^{k(n-k)}\eta$ for all $(0,k)$ alternating tensor η on an n-dim space. (So if n is odd, Hodge star is its own inverse always.)*

Proof. Direct computation. □

Example 13.10.7 (Maxwell's equation)**.** The full Maxwell's equation in vacuum is the following: we can describe the electromagnetic structure using a 2-form, the so-called electromagnetic tensor. Note that the space-time is 4-dim, and thus a 2-tensor can be represented as a 4 by 4 matrix, give as $F = \begin{bmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{bmatrix}$, where E_i and B_i are describing the electric field and the magnetic fields.

(Note that the entries related to time is electrical, whereas the purely spatial entries are all about the magnetics.) Note that as a matrix it is skew-symmetric, so as a tensor it is alternating, and thus a 2-form.

The Maxwell's equation in vacuum says that any such F must have $\mathrm{d}F = \mathrm{d}(*F) = 0$. No proof is required, since all you need in physics are experiments. However, note that $\mathrm{d}F = 0$ implies that F is closed. Note that for a closed foliation, say it is made of closed curves, then we can "fill up the inside" and get a higher dimensional foliation A^{em}. This can always be done in a domain without holes, e.g., $\mathbb{R}^n$. So by this "fill up the inside" process, we see that $F = \mathrm{d}A^{em}$.

Here A^{em} is the electromagnetic 4-potential, defined as (ϕ, A) where ϕ is the electric potential and A is the magnetic potential (a vector potential), such that $curl(A) = B$ the magnetic field, and $-\nabla\phi - \frac{\partial}{\partial t}A = E$, the electic field.

Let us see how this acts out. Consider $\mathrm{d}F = 0$. Recall that as a (0,2)-tensor, we write a matrix for F as $\begin{bmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{bmatrix}$. (Btw, I am taking the speed of light as the unit here.) This means that $F = \mathrm{d}t \wedge (E_x\,\mathrm{d}x + E_y\,\mathrm{d}y + E_z\,\mathrm{d}z) - B_z\,\mathrm{d}x\,\mathrm{d}y - B_y\,\mathrm{d}z\,\mathrm{d}x - B_x\,\mathrm{d}y\,\mathrm{d}z$. Taking exterior derivative, the electric portion leaves $-\mathrm{d}t \wedge \mathrm{d}(E_x\,\mathrm{d}x + E_y\,\mathrm{d}y + E_z\,\mathrm{d}z) = -\mathrm{d}t \wedge (curl(E) \cdot \begin{bmatrix} \mathrm{d}y\,\mathrm{d}z \\ \mathrm{d}z\,\mathrm{d}x \\ \mathrm{d}x\,\mathrm{d}y \end{bmatrix})$. The magnetic portion leaves $-\mathrm{d}t \wedge (\frac{\partial B}{\partial t} \cdot \begin{bmatrix} \mathrm{d}y\,\mathrm{d}z \\ \mathrm{d}z\,\mathrm{d}x \\ \mathrm{d}x\,\mathrm{d}y \end{bmatrix}) - div(B)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$. Now, the coefficients for the basis $\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y, \mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z, \mathrm{d}t\,\mathrm{d}z\,\mathrm{d}x, \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ must be zero. So in particular, we obtained the Gauss's law for magnetism $div(B) = 0$ and the Maxwell-Faraday equation $\frac{\partial B}{\partial t} + curl(E) = 0$.

Here the Gauss's law says that, if you treat $B_z\,\mathrm{d}x\,\mathrm{d}y + B_y\,\mathrm{d}z\,\mathrm{d}x + B_x\,\mathrm{d}y\,\mathrm{d}z$ as a 2-form in $\mathbb{R}^3$, then its exterior derivative is zero. In particular, the magnetic lines are all closed loops, since they cannot have boundaries! And the Maxwell-Faraday equation says that a changing magnetic field is related to the exterior deriavtive of the 1-form $E_x\,\mathrm{d}x + E_y\,\mathrm{d}y + E_z\,\mathrm{d}z$ in $\mathbb{R}^3$. Note that boundaries are where things start. So changing magnetic field would GENERATE electric field.

What about the Hodge dual? The hodge dual sends, say, $\mathrm{d}t\,\mathrm{d}x$ to $\mathrm{d}y\,\mathrm{d}z$. So the $(1,2)$ entry of the matrix for F is sent to the $(3,4)$-entry. In general, there might be negative signs involved. You can verify that $*F$ can be represented by the matrix $\begin{bmatrix} 0 & -B_x & -B_y & -B_z \\ B_x & 0 & -E_z & E_y \\ B_y & E_z & 0 & -E_x \\ B_z & -E_y & E_x & 0 \end{bmatrix}$. And as a 2-form we have $*F = -\mathrm{d}t \wedge (B_x\,\mathrm{d}x + B_y\,\mathrm{d}y + B_z\,\mathrm{d}z) + E_z\,\mathrm{d}x\,\mathrm{d}y + E_y\,\mathrm{d}z\,\mathrm{d}x + E_x\,\mathrm{d}y\,\mathrm{d}z$. As you can see, the hodge dual does have a feeling of duality, yes? Now in vacuum, we have $\mathrm{d}*F = 0$, which dually implies that $div(E) = 0$ and $\frac{\partial E}{\partial t} - curl(B) = 0$. The first one is Gauss's law for electricity, which states that electrical circuits are closed, as they have no boundary. The second one is the Ampère's circuit law, which states that a change in electric field would GENERATE magnetic field.

Note that the term "boundary" here has multiple meanings. For a vector field, say B, you may think of it as $B_x\,\mathrm{d}x + B_y\,\mathrm{d}y + B_z\,\mathrm{d}z$ and take curl. Then you are looking at a foliation of $\mathbb{R}^3$ by surfaces, orthogonal to the magnetic lines everywhere. If you think of it as $B_z\,\mathrm{d}x\,\mathrm{d}y + B_y\,\mathrm{d}z\,\mathrm{d}x + B_x\,\mathrm{d}y\,\mathrm{d}z$, then you are looking at a foliation of $\mathbb{R}^3$ by curves, i.e., the familiar magnetic lines. The latter view usually feels more natural for us, whereas for the electric field, the first view usually feels more natural.

The above situations are only true in vacuum. In general, $\mathrm{d}*F = 0$ may fail due to the so-called charge density and current density, together they form a 3-form $J = \phi\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + j_1\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z + j_2\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}x + j_3\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y$. As you can see, the "time" coordinate, the first one, is actually completely spatial, where as the other three coordinates must form a spatial 2-form that is also related to time, i.e., it has field lines and they are flowing. Then the Maxwell's equation becomes $\mathrm{d}*F = J$. ☺

Example 13.10.8. What does it mean to have $\mathrm{d}F = \mathrm{d}*F = 0$? Consider the case of a 1-form in $\mathbb{R}^2$ with

$\mathrm{d}F = \mathrm{d}{*}F = 0$. Then F and ${*}F$ are both foliations of curves in $\mathbb{R}^2$, and by definition of the Hodge star, the curves for F are orthogonal to the curves of ${*}F$, and they should have the same density everywhere. Furthermore, none of the curves involved could have boundary, because $\mathrm{d}F = \mathrm{d}{*}F = 0$.

One example of solution is $F = \mathrm{d}x$, where all curves are verticle, and then ${*}F = \mathrm{d}y$, where all curves are horizontal. Another example with a singularity is defined on $\mathbb{R}^2$, with $F = \mathrm{d}r/r$. Here r is the function sending each $\boldsymbol{p}$ to the distance to the origin. Then F corresponds to circles with greater density as you get closer to the origin, and ${*}F$ are rays from the origin to infinity. ☺

13.11 Pulling tangents and pushing forms

See Video 151-1, 151-2 and 151-3.

13.12 de Rham cohomology

See Video 152-1 and 152-2.