Linear Algebra

NOTES BY DYLAN PENTLAND

Introduction

These are a set of notes for the Summer 2019 HSSP linear algebra class. These are work in progress, and the current form is just a sketch of how the course might go (content could be removed or added depending on how it goes).

Lecture 1: So what is a vector, anyway?

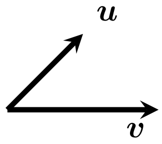
Vectors in $\mathbf{R}^n$

If you've taken a physics class, you might have seen the following definition of a vector:

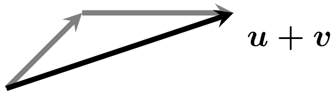
Definition. A vector is a quantity having direction as well as magnitude.

This is good for giving some intuition about how you might interpret a vector, but it's a bit tricky to do anything with this without any extra structure. For example, how will we add vectors? How would we describe functions that take in vectors?

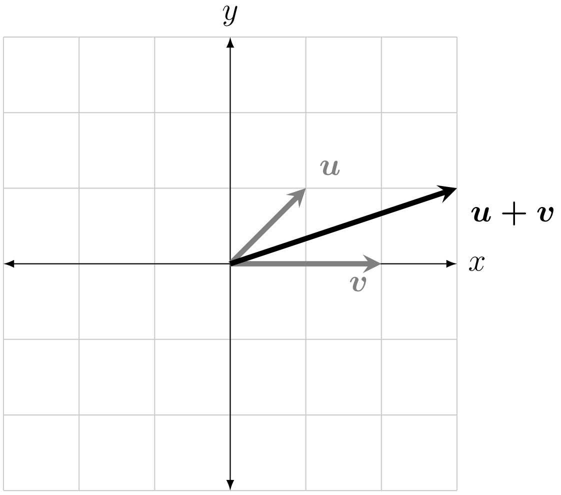
Let's think about how we should add vectors. Let's call our vectors $\boldsymbol{u}, \boldsymbol{v}$.



Their sum should look something like this, where we start $\boldsymbol{v}$ at the tail of $\boldsymbol{u}$:



If we add in coordinate axes, we will get something like this:



What we get from this is that we can represent a vector as a pair (x, y). In the picture above, this is $\boldsymbol{u}=(1,1), \boldsymbol{v}=(2,0)$ and $\boldsymbol{u}+\boldsymbol{v}=(3,1)$. We can formulate a general rule for adding vectors from this:

$$(x, y)+(x', y')=(x+x', y+y').$$

But this is only two dimensions. Can we do better? Yes we can.

DEFINITION. We will use $\mathbf{R}$ to denote the set of real numbers.

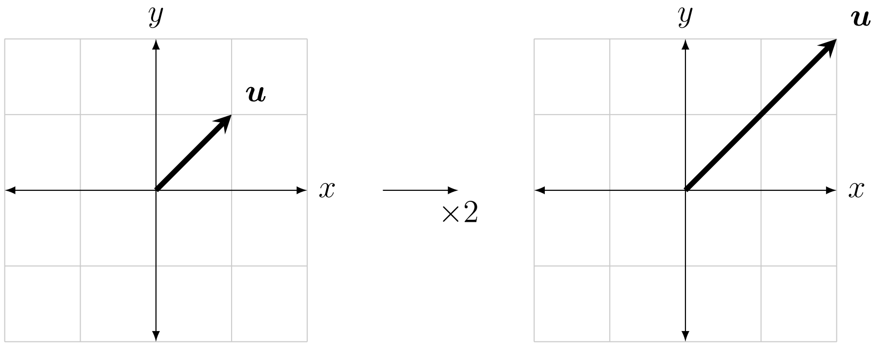
DEFINITION. ($\mathbf{R}^n$) As a set, $\mathbf{R}^n$ consists of sets of n real numbers:

$$(v_1, \dots, v_n) \in \mathbf{R}^n.$$

We have two main operations we can perform on these.

- Addition. If $v = (v_1, \dots, v_n)$ and $w = (w_1, \dots, w_n)$ then
$$v + w = (v_1 + w_1, \dots, v_n + w_n).$$
- Scalar multiplication. If $v = (v_1, \dots, v_n)$, and $c \in \mathbf{R}$, we define
$$cv = (cv_1, \dots, cv_n).$$

The example we did above was $\mathbf{R}^2$. The addition operation is what we did by putting the other vector at the end of the first, and scalar multiplication just makes a vector longer:



This is usually what people mean when they say "n-dimensional space". Don't get intimidated by the fact that you can't draw pictures anymore - since we have the definition above, we still know how to work with vectors even when it's hard to visualize them. We can get our intuition from the $\mathbf{R}^2$ or $\mathbf{R}^3$ case.

Now that we know what vectors are, let's talk about some of the things we can do with them before we get to linear maps.

ORTHOGONAL VECTORS

First, we'll give an intuitive definition of what it means for two vectors to be orthogonal. I will assume you know what this means in $\mathbf{R}^2$.

DEFINITION. Given two vectors in $\mathbf{R}^n$, there is a unique plane Π containing both of them [1]. In this plane, we say v and w are orthogonal they are orthogonal in $\Pi \simeq \mathbf{R}^2$.

[1] Why? Describe the plane.

What does this mean in $\mathbf{R}^2$ in terms of coordinates? If we let $v = (v_1, v_2), w = (w_1, w_2)$, we have following:

LEMMA. *The vectors $v, w \in \mathbf{R}^2$ are orthogonal if and only if $v_1w_1 + v_2w_2 = 0$.*

If we work it out in $\mathbf{R}^3$, we can see that the condition becomes $v_1w_1 + v_2w_2 + v_3w_3 = 0$. So, we might guess the following:

$$v \perp w \iff \sum_i v_iw_i = 0.$$

We denote the sum $\sum_i v_iw_i$ as $v \cdot w$. That would be really nice if it were true, because then we wouldn't have to bend our minds in $\mathbf{R}^n$ trying to figure out whether or not two vectors are orthogonal.

REMARK. Note that $u \cdot (v + w) = u \cdot v + u \cdot w$, and $cv \cdot w = c(v \cdot w)$. Also, $w \cdot v = v \cdot w$.

LEMMA. *We have*

$$v \cdot w = |v| \cdot |w| \cos(\theta),$$

where θ is the angle between the two vectors in the plane Π.

Proof. Given vectors v, w, if we place the start of the vector $v - w$ at w then it connects to v since $w + (v - w) = v$. Thus, we have formed a triangle. The law of cosines tells us that

$$|v - w|^2 = |v|^2 + |w|^2 - 2|v| \cdot |w| \cos\theta.$$

From the Pythagorean theorem, note that

$$|v - w|^2 = (v - w) \cdot (v - w).$$

Using our remark, we have

$$\begin{aligned}(v - w) \cdot (v - w) &= (v \cdot v) - 2(v \cdot w) + (w \cdot w)\\ &= |v|^2 + |w|^2 - 2(v \cdot w).\end{aligned}$$

Comparing this with the law of cosines from before, we see that $v \cdot w = |v| \cdot |w| \cos\theta$. □

THEOREM. *Our conjecture was right! For nonzero vectors v, w, we have $v \perp w$ if and only if $v \cdot w = 0$.*

Proof. It's easy with the lemma: we see that $v \cdot w = 0$ if and only if $\cos(\theta) = 0$. This only happens for $\theta = \pm\pi/2$. □

EXAMPLE. Let $\hat{n}$ be a unit vector, that is $\hat{n} \cdot \hat{n} = 1$. Set $\Pi_{\hat{n}} = \{v \in \mathbf{R}^n : v \cdot \hat{n} = 0\}$. This will be a hyperplane of dimension $n - 1$, and in fact all hyperplanes arise this way.

This theorem is actually quite useful. It translates a geometric idea into an algebraic one, which makes dealing with higher dimensions much easier.

For this reason, many linear algebra books will actually have our theorem as a *definition*. It is good to see the geometric meaning, but this probably a good way to view it in general. The reason for this will become clearer later, but the idea is that we want to eventually discard ideas from classical geometry in favor of more algebraic ideas that allow us to more easily deal with abstract vector spaces.

ABSTRACTING GEOMETRY

One thing the last part of the lecture demonstrated was that we can discard geometric ideas in favor of algebraic ones, and this can actually be very useful because it allows us to more easily make computations and check if things are true. If you didn't know about the dot product, would you have a reasonable way to check if vectors are perpendicular in $\mathbf{R}^{2019}$?

We can do the same thing with the notion of a vector, and this is an incredibly useful idea. First, let's just do this over $\mathbf{R}$.

DEFINITION (Vector space over $\mathbf{R}$). A vector space V over $\mathbf{R}$ is a set of vectors, such that the following hold for all $u, v, w \in V$:

- $u + v = v + u$
- $(u + v) + w = u + (v + w)$
- There is a vector 0 so that $v + 0 = 0 + v = v$. This is provably unique.
- For any v, there is $-v$ so $v + (-v) = 0$.

This tells us how to add vectors. We also need to know how to scale them. Let $r, s \in \mathbf{R}$. We have

- $r(sv) = (rs)v$
- $(r + s)v = rv + sv$
- $1v = v$

There is a point to all of this, other than listing out seemingly obvious statements. If we prove something about vector spaces over $\mathbf{R}$ algebraically, we have only used these axioms. That means if we can find any set V that satisfies these axioms, we immediately have all of the theorems we know about vector spaces over $\mathbf{R}$ for V, even if we know very little about V other than that these axioms hold. Let's give some examples to demonstrate how there can be spaces that are distinctly *not* geometric, but are vector spaces over $\mathbf{R}$:

EXAMPLE. Let $V = f(x)$, where $\deg f(x) \leq 2$ and f has real coefficients. It's clear that if we add these functions the degree can't increase, so we have closure. Scaling by real numbers also can't increase degree. It is pretty easy to verify all the axioms.

Example. Consider $\mathbf{C}$. Since $\mathbf{R} \subset \mathbf{C}$, we can multiply by real numbers. Obviously we can add complex numbers - the axioms hold again. Thus, $\mathbf{C}$ is a vector space over $\mathbf{R}$. It is 'two-dimensional', in the sense that we can write every element as $a + bi$ for $a, b \in \mathbf{R}$. Abstractly, we can treat it as $\mathbf{R}^2$ even though it has more structure, which is where we get the picture of $\mathbf{C}$ as a plane.

Similarly, $\mathbf{C}^n$ is a vector space over $\mathbf{R}$.

Example. Vector spaces don't have to have finite dimension! Consider $\mathbf{R}[t]$, the space of polynomials with real coefficients in a variable t. This is a vector space over $\mathbf{R}$, but is infinite dimensional. We mostly won't treat examples like this, because the theory of infinite dimensional vector spaces is tricky.

Example. We can actually get even bigger vector spaces. Consider the space

$$V = C^\infty(\mathbf{R})$$

of infinitely differentiable functions defined on $\mathbf{R}$. Members include polynomials, $e^x, \sin x, \cos x$. Non-examples include $\ln x$ (not defined for $x \leq 0$), $|x|$ (not differentiable at 0). An interesting example is $|x|^{k+1}$, which is differentiable k times but not $k+1$ times. We then write $|x|^{k+1} \in C^k(\mathbf{R})$. You can again check the vector space axioms for $\mathbf{R}$ on these and you will find that they hold, although it is now unclear how we could even write down an arbitrary element. Despite this, linear algebra will still work here!

Hopefully these examples made the usefulness of this approach clear - instead of restricting ourselves to boring old $\mathbf{R}^n$, we can now talk about linear algebra in spaces that are quite different but behave the same algebraically. As you will see as this course progresses, all of the theorems we prove will depend only on these algebraic axioms so we can apply everything we will learn in each of these examples equally well.

Definition. A subspace $V \subset W$ is a subset of a vector space that is closed under addition and scalar multiplication in W.

Example. Some examples:

- $\mathbf{R} \subset \mathbf{R}^2$.
- $\mathbf{R} \subset \mathbf{C}$, $\mathbf{Q} \subset \mathbf{Q}(\sqrt{2})$, etc.
- A plane through the origin in $\mathbf{R}^n$.
- $\Pi_{\hat{n}}$.
- $C^k(\mathbf{R}) \subset C^{k-1}(\mathbf{R})$.

Lecture 2: Linear maps

From now on we will talk about vector spaces without explicit reference to the field. If it helps, you can think about these all being over $\mathbf{R}$ but keep in mind everything we do works over a general field.

Definition. Let V, W be vector spaces over the same field. A linear map $V \to W$ is a function

$$T : V \to W$$

satisfying $T(v+w) = T(v) + T(w)$ and $T(cv) = cT(v)$. The set of linear maps $V \to W$ is denoted $\mathrm{Hom}(V, W)$.

Why this definition? We can write down several justifications.

- It's structure preserving. If we can write $v', w' \in W$ as $T(v) = v', T(w) = w'$, then we know how to add v', w': it is just $T(v+w)$. A similar result holds for scalars. Suppose we have a linear map T which is bijective, so that we can write every $w' \in W$ as $T(w)$. If we relabel w' as w, then the addition rules for w', v' are exactly the same as those for w, v. That is, V, W are basically the same - we just give the elements different names.

- A less abstract reason is that we want to look at maps that "stretch" space uniformly. This is apparent when we look at linear maps in $\mathbf{R}^n$, where we can draw pictures.

Example. Let's write down some simple examples.

- Consider $V = \mathbf{R}^n$. The map sending $v \mapsto 2v$ is a linear map.

- Consider $\mathbf{R} \subset \mathbf{C}$. We know $\mathbf{C}$ is a vector space over $\mathbf{R}$. Multiplication by any complex number satisfies the axioms for a linear map.

- Take $V = \mathbf{R}^2$. The map sending v to $R_\theta v$, where $R_\theta v$ is v rotated by θ is a linear map.

- Note that $\mathrm{Hom}(V, W)$ itself is a vector space over $\mathbf{R}$. The map $T \mapsto T'(T)$ is a linear map if T is a linear map.

- Linear maps tells us about local information of complicated functions. If we have a function $f(x_1, x_2, \ldots, x_n)$ around a point $p \in \mathbf{R}^n$ the function will behave very similarly to $T(v-p)$. To see this is a low-dimensional case, consider the tangent line to a function $f : \mathbf{R} \to \mathbf{R}$. If you've seen multivariable calculus, this is the Jacobian matrix.

In order to do explicit calculations in $\mathbf{R}^n$, we has to pick coordinates of some sort. How can we generalize this idea? Well, roughly what we want is to be able to pick a set of vectors $\{v_i\}$ so that every $v \in V$ can be written uniquely as a combination of these vectors. Specifically, we mean a *linear combination*:

DEFINITION. Let $S = \{v_1, \ldots, v_n\} \subset V$ be a set of distinct vectors. A *linear combination* of vectors in S is a sum of the form

$$w = \sum_i c_i v_i$$

where $c_i \in \mathbf{F}$.

We call a set of vectors $\{v_i\}$ *spanning* if every $v \in V$ can be written as a linear combination of $\{v_i\}$. This is a great property, because of the following:

OBSERVATION. Suppose we have a linear map, and we know the values Tv_i for a spanning set $\{v_i\}$. Since $T(\sum_i c_i v_i) = \sum_i c_i T v_i$ by linearity, this allows us to calculate the value of T at every point from a finite set of data.

It is for this reason that we want to understand better how spanning sets of vectors work - this has great potential to let us easily describe what any linear map does! Next, we need to deal with the uniqueness part. Intuitively, this should have something to do with picking "independent" vectors for each "dimension" of our vector space, since this is what we did in $\mathbf{R}^n$.

DEFINITION. A set of vectors $\{v_i\}$ is linearly independent if the only way to write $0 \in V$ as a linear combination of $\{v_i\}$ is to set all of the c_i equal to 0.

In the case of $\mathbf{R}^n$, we can see that linearly independent is something like saying we don't pick vectors from subspaces we already span, and being spanning says that we have enough vectors to reach everything. We would expect to be able to write everything uniquely, and this is in fact the case in general.

THEOREM. *Let $\{v_i\} \subset V$ be a set of vectors in V. If they are spanning and linearly independent, then they are a basis. That is, we can write every element of V uniquely as*

$$v = \sum_i c_i v_i.$$

Proof. First, we see that we can write every element of V as *some* linear combination, by definition of spanning. If they are linearly independent, we claim this makes the linear combination unique. Suppose we could write v in two ways, Then we would have

$$v = \sum_i c_i v_i = \sum_i \overline{c}_i v_i,$$

and subtracting we obtain

$$0 = \sum_i (\overline{c}_i - c_i) v_i$$

and by linear independence we get $\bar{c}_i = c_i$, and hence the combination was unique. □

We call a vector space finite dimensional if it is spanned by a finite set of vectors. In this case, a finite basis will exist.

THEOREM. *Let V be finite dimensional. Then V admits a finite basis.*

Proof. Let $\{v_i\}$ be a finite spanning set of V. The idea is that we can remove vectors until we get a basis of V. □

In this case, we have the following theorem that confirms our experience in $\mathbf{R}^n$ holds in general:

THEOREM. *Let $\{v_i\}$ be a finite basis for V. Then every basis of V has the same size.*

Proof. Let $\{v_i\}$ be a basis of V consisting of n vectors. Let $\{w_i\}$ be a basis consisting of m vectors.

Take $\{v_1, \ldots, v_n\}$. We can write $w_1 = \sum_{i \leq n} c_i v_i$, and in particular we can extract

$$v_n = \frac{1}{c_n}\left(w_1 - \sum_{i \leq n-1} c_i v_i\right)$$

Hence, we can replace v_n by w_1 and our set remains spanning because we can reach v_n. Continue the process; eventually, we replace all of the v_i with elements of $\{w_i\}$. Thus, $n \geq m$. If we had $m > n$, we would have a proper subset of $\{w_i\}$ that is spanning, and hence we would not get linear independence.

Similarly, $m \geq n$. Thus, $m = n$. □

This lets us make the following definition:

DEFINITION. Let V be finite dimensional. Then $\dim V$ is the size of any basis of V.

We will only be talking about finite dimensional vector spaces here. Some of the properties of vector spaces become more subtle when we go to the infinite dimensional case.

DEFINITION. We call two vector spaces V, W isomorphic if there exists a linear map $T : V \to W$ which has an inverse $T^{-1} : W \to V$. We write $V \simeq W$. To be an inverse, we need $T^{-1}T = TT^{-1}$ to be the identity.

REMARK. We call the linear map an isomorphism. An equivalent condition is being bijective. To be bijective, a function is *surjective* and *injective*. A surjective function $f : X \to Y$ has $y = f(x)$ for each $y \in Y$ for some $x \in X$ - it 'hits everything'. An injective function has no repeat values: if $f(x) = f(y)$, then $x = y$.

THEOREM. *A vector space V of dimension n over $\mathbf{R}$ is isomorphic to $\mathbf{R}^n$.*

Proof. Let V have a basis $\{v_1, \dots, v_n\}$. Then define T to be the linear map sending

$$T : \sum_i c_i v_i \mapsto \sum_i c_i e_i$$

where $e_i = (0, \dots, 1, \dots, 0)$ with a 1 in the ith index, the canonical basis for $\mathbf{R}^n$. This has an easy inverse:

$$T^{-1} : \sum_i c_i e_i \mapsto \sum_i c_i v_i.$$

Thus, $V \simeq \mathbf{R}^n$. □

THEOREM. *Let $T : V \to W$ be a bijective linear map. Then there is $T^{-1} : W \to V$ so $T^{-1}T = TT^{-1}$ is the identity map, I. In other words, a bijective linear map is an isomorphism.*

Proof. If T sends $v \in V$ to $Tv = w \in W$, we define T^{-1} so that it sends $w \mapsto v$. Clearly $T^{-1}T = TT^{-1}$ is the identity. We have

$$T^{-1}(w + w') = T^{-1}(Tv + Tv') = T^{-1}T(v + v'),$$

then by definition this is sent to $v + v' = T^{-1}w + T^{-1}w'$. Thus, T is linear. We can also check that $T^{-1}(cw) = T^{-1}(T(cv))$ which is cv by definition, or $cT^{-1}w$. □

LECTURE 3: LINEAR MAPS AS MATRICES

At this point, we have covered the basics of what vector spaces over $\mathbf{R}$ actually look like. Although abstractly they are isomorphic to $\mathbf{R}^n$, it is helpful to keep in mind that the objects we describe using the vector space structure don't have to resemble $\mathbf{R}^n$ (as we saw with some previous examples).

One thing you might have about before is the idea of a matrix. It's a very similar idea to a linear map, but there is a subtle (but important!) difference between the two.

Let $T : V \to W$ be a linear map. A *matrix* for T is an explicit representation of T with respect to chosen bases for V and W. Suppose that V has a basis $\{v_1, \dots, v_n\}$ and W a basis $\{w_1, \dots, w_m\}$. Suppose that we know the values $Tv_1, Tv_2, \dots, Tv_n$. If we want to calculate Tv, we just write

$$Tv = T\left(\sum_j c_j v_j\right) = \sum_j c_j Tv_j.$$

Just like our observation with spanning sets, we can calculate everything. The advantage of a basis is that every element has a unique representation in the basis, so it is no longer ambiguous which c_j to pick.

We can go a step further: $Tv_j \in W$, so it can be written as a sum $Tv_j = \sum_i T_{ij} w_i$. We can then describe Tv as

$$Tv = \sum_{i,j} c_j T_{ij} w_i.$$

This tells us what element of W we get, in the basis we picked for W. As a result, the $n \times m$ numbers T_{ij} completely describe T.

This is where we get the 'matrix' notation. Basically, we write T like this:

$$T = \begin{bmatrix} | & | & \dots & | & | \\ Tv_1 & Tv_2 & \dots & Tv_{n-1} & Tv_n \\ | & | & \dots & | & | \end{bmatrix}$$

In the columns, we write Tv_j as a column vector with m entries in the basis of W. The entry T_{ij} then appears in row i and column j. The way to read off what a linear map does from a matrix representation is simple: just look at each column, and we can read off where each basis vector is sent. If we know a vector's representation in our basis as a linear combination, we just do the same linear combination of the columns. The number of columns is $\dim V$ (the input space) and the number of rows is $\dim W$ (the output space).

Explicitly, applying T to a vector $\vec{c} = (c_1, \dots, c_n) \in V$, the ith entry of the result is $\sum_j c_j T_{ij}$ or $\vec{c} \cdot T_i$, where T_i is the vector for row i in the matrix.

This also lets us compose linear maps. Let's specialize to $T \in \mathrm{Hom}(V, V)$. If we have A, B matrices of linear maps, then $A \circ B$ has us apply B and then A. This sends $v_i \mapsto Bv_i$, or the ith column of B. Denote this $B^i = (b_{i1}, \ldots, b_{in})$. Applying A, we

$$ABv_i = A(B^i) = \sum_j (B^i \cdot A_j) v_j.$$

Thus, we can just take dot products of the ith column of B and the jth row of A to get the entry in column i and row j of AB.

EXAMPLE. Let's take our maps from before and write down matrices.

- Consider $V = \mathbf{R}^n$. The map sending $v \mapsto v$ is a linear map, called I. Its matrix is the following:

$$T = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

- Consider $V = \mathbf{R}^n$. The map sending $v \mapsto cv$ is a linear map, denoted cI. Its matrix is the following:

$$cI = \begin{bmatrix} c & & & \\ & c & & \\ & & \ddots & \\ & & & c \end{bmatrix}$$

- Consider $\mathbf{R} \subset \mathbf{C}$. We know $\mathbf{C}$ is a vector space over $\mathbf{R}$. Multiplication by any complex number satisfies the axioms for a linear map. We have

$$a + bi = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

- Take $V = \mathbf{R}^2$. The map sending v to $R_\theta v$, where $R_\theta v$ is v rotated by θ is a linear map. This is given by

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

REMARK. Note that all throughout this we have picked a *specific* basis for both V, W. There are many matrix representations of a linear map, we just pick a particular one:

$$T \in \mathrm{Hom}(V, W) \mapsto \text{Matrix of } T \text{ with respect to bases}$$

Additionally, different linear maps might have the same numbers T_{ij} in their matrix representations but just be in different bases.

However, the properties of the linear map itself should be independent of the basis. Properties like 'how much does it stretch?' or 'can I undo it?' should be coordinate-free since they make no reference to coordinates, and will be the same for many matrix representations of the linear map.

In terms of matrices, if A and B are the same linear map under different bases we will be able to write $A = PBP^{-1}$ for some matrix P. This is an important point, so make sure to remember it!

LEMMA. *Let A be an $n \times n$ matrix with an inverse. Then the columns of A form a basis for $V = \mathbf{R}^n$.*

Proof. First, we show that they are spanning. Let $A_i \in V$ denote the ith column (going from left to right). Take $v \in \mathbf{R}^n$, and consider $w = A^{-1}v$. Then if $w = \sum_i w_i e_i$,

$$v = Aw = w_i A_i.$$

Hence, we can write v in terms of the vectors A_i. Now suppose that there exists a nonzero linear combination of the A_i that is zero - letting the coefficients be $c_1, \ldots, c_n$ and setting $\vec{c} = (c_1, \ldots, c_n)$, there is a nonzero vector $\vec{c}$ so $A\vec{c} = \vec{0}$. Applying A^{-1}, $\vec{c} = \vec{0}$, a contradiction. □

OBSERVATION. Matrix multiplication is not actually commutative - that is, $AB \neq BA$ in general. To see this, consider two linear maps: A sends $(x, y) \mapsto (x+1, y)$ and the other, B, sends $(x, y) \mapsto (2x, 2y)$. Then $AB = (2x + 1, 2y)$, $BA = (2x + 2, 2y)$. These are different! However, it is associative. We have $ABC = (AB)C = A(BC)$ - this is because as linear maps, matrices are functions and multiply via composition. Function composition is associative.

LEMMA. *Inverses for a matrix A are unique.*

Proof. Suppose there are B, C so that $AB = BA = I, AC = CA = I$. Then

$$ABC = (AB)C = (BA)C = B(Ac) = B.$$

But $AB = I$, so $ABC = (AB)C = C$. Hence, $B = C$. □

Lecture 4: Image and Kernel

Last lecture, we saw how the matrix of a linear map allows us to be able to effectively compute what a linear map does to V. In this lecture, we will look at two objects associated to a linear map (not a specific matrix) known as the *image* and *kernel*.

Definition. Let $T : V \to W$ be a linear map. The image is the set $T(V) \subset W$. We denote this by $\operatorname{im} T$.

Definition. Let $T : V \to W$ be a linear map. The kernel is the set $K \subset V$ of vectors v so $T(v) = 0 \in W$. We denote this by $\ker T$.

Lemma. *Both* $\operatorname{im} T, \ker T$ *are subspaces of* W *and* V *respectively.*

Proof. Just check the axioms and use linearity! □

Intuitively, the kernel measures how "degenerate" T is by seeing how information about V we lose by sending things to zero. The image acts as the opposite of the kernel, giving us information about how much content T preserves from V. In the case of an isomorphism, the kernel will be trivial and the image will be W so we lose no information and retain all the information from V.

Example. Consider a map $\pi : V \to V$ which is not surjective so that $\pi^2 = \pi$. This implies that π is a 'projection' (or a fancier word: idempotent). It takes V and sends everything to a particular subspace ($\operatorname{im} \pi$), which it fixes.

Let's restrict our attention to $\operatorname{Hom}(V, V)$ again. Suppose we have a subspace W of V. We claim that the set

$$V/W := \{v + W : v \in V\}$$

where $v + W = \{v + w : w \in W\}$ has a vector space structure on it.

Lemma. *We can make* V/W *a vector space, withe the addition operation*

$$(v + W) + (v' + W) = (v + v') + W.$$

Proof. Not hard, just verify the axioms. □

Lemma. *We have* $\dim V/W = \dim V - \dim W$.

Proof. A bit trickier. If $(w_1, \ldots, w_m)$ is a basis for W, we can extend to a basis of V by adding $v_1, \ldots, v_{n-m}$.

Under the map $V \to V/W$, this becomes $(v_1 + W, \ldots, v_{n-m} + W)$. This will still be spanning, since it was spanning before. Suppose that

$$\sum_i c_i(v_i + W) = W.$$

This is equivalent to $(\sum_i c_i v_i) + W = W$, or $\sum_i c_i v_i \in W$. Then in V, we obtain

$$\sum_i c_i v_i - \sum_j \overline{c}_j w_j = 0$$

and by linear independence in V all of the coefficients are 0. Thus, we get linear independence. This is then a basis, so the dimension is $n - m$. □

THEOREM. *Let T be a linear map. Then*

$$V/\ker T \simeq \operatorname{im} T.$$

Proof. We already have a map

$$T : V \to \operatorname{im} T \subset W.$$

However, it does not have the right domain. Instead, consider the following diagram:

$$\begin{tikzcd} V \arrow[d, "\pi"] \arrow[dr, "T"] & \\ V/\ker T \arrow[r, "\overline{T}"] & \operatorname{im} T \end{tikzcd}$$

Here, the map $\overline{T}$ sends $v + \ker T \mapsto T(v)$. We claim that this is both well-defined and is actually an isomorphism.

To be well-defined, we need to show that if $v + \ker T = w + \ker T$ then $T(v) = T(w)$. Note that $v + \ker T = w + \ker T$ implies $v - w \in \ker T$ so

$$T(v) - T(w) = T(v - w) = 0$$

and hence they are equal. Thus, $\overline{T}$ is well-defined. It is easily checked to be linear.

Next, we claim it is bijective. Let $w \in \operatorname{im} T$. Then $w = T(v)$, and $\overline{T}(v + \ker T) = T(v) = w$. Thus, $\overline{T}$ is surjective. Suppose that $T(v + \ker T) = T(w + \ker T)$. Then $T(v) = T(w)$, so $T(v - w) = 0$ and $t := v - w \in \ker T$. But then $w + t = v$, so $w + \ker T \subset v + \ker T$. Similarly we get the other inclusion, so $\overline{T}$ is injective. It is then bijective, and hence an isomorphism. □

THEOREM. *We have* $\dim V = \dim \ker T + \dim \operatorname{im} T$.

This makes the intuitive explanation from before precise, telling us that the number of 'degrees of freedom' in the kernel are balanced out by those in the image.

LEMMA. *A map $T \in \operatorname{Hom}(V, V)$ is an isomorphism if and only if it is surjective. The same is true for being injective.*

Proof. A linear map is injective if and only if $\dim\ker T = 0$, and surjective if and only if $\dim\operatorname{im} T = \dim V$. Rank nullity shows that these conditions imply each other, and T is an isomorphism if and only if both are true. □

LEMMA. *Let* $\dim V = n, \dim W = m < n$. *Then if* $T : V \to W$ *is a linear map, we have* $\dim\ker T \geq n - m$.

Proof. We have $\dim\ker T + \dim\operatorname{im} T = n$. But $\dim\operatorname{im} T \leq m$, so $\dim\ker T \geq n - m$. □

DEFINITION. The *column rank* of a matrix is the dimension of the subspace spanned by its columns. The *row rank* is the dimension of the subspace spanned by its rows.

LEMMA. *The column rank of a matrix* A *is* $\dim\operatorname{im} A$.

Proof. If $w \in \operatorname{im} A$, we have $w = Av$ for some $v \in V$. If $v = \sum_i v_i e_i$, we get

$$Av = \sum_i v_i A e_i,$$

and since Ae_i are the columns of A we see w is in the span. Conversely, every vector spanned by $\{Ae_i\}$ is in the image, since $\sum_i c_i A e_i = A(\sum_i c_i e_i) \in \operatorname{im} A$. □

THEOREM (Warning: Tricky proof!). *The row rank of a matrix is the same as the column rank.*

Proof. The matrix $A : V \to W$ will be a $\dim W \times \dim V$ matrix. Recall that if V is a vector space, $\operatorname{Hom}(V, \mathbf{R})$ is also a vector space and we denote this by V^*. We can 'dualize' the situation $A : V \to W$ as follows:

$$V \xrightarrow{A} W \qquad\qquad \operatorname{Hom}(W, \mathbf{R}) \xrightarrow{A^*} \operatorname{Hom}(V, \mathbf{R})$$

We define A^* as sending a map $\varphi : W \to \mathbf{R}$ to $\varphi A : V \to W \to \mathbf{R}$ (this is matrix multiplication). How does this relate to the row rank? Well, we can write down φ as a $1 \times \dim W$ matrix since it sends $W \to \mathbf{R}$. Let's write down a matrix for A^*. It is a $\dim V \times \dim W$ matrix, since the input space has dimension $\dim W$ and the output space dimension $\dim V$. It sends e_i^* to $e_i^* A$, a $1 \times \dim V$ vector which is the ith row of A. Thus, as a map $W^* \to V^*$, the columns of A^* have become the rows! By our lemma, $\dim\operatorname{im} A^*$ is then the row rank. Hence, the statement that row rank equals column rank is that $(\operatorname{im} A)^* \simeq \operatorname{im} A^*$.

With this result in hand, we are nearly there. Given $A : V \to W$, this is also a surjective map $V \to \operatorname{im} A$. Dualizing gives $A^* : (\operatorname{im} A)^* \to V^*$. The kernel is now trivial. This is because $A^*\varphi = \varphi A = 0$ implies φ is 0 on its entire domain, and hence is the zero map. The rank-nullity theorem gives us an isomorphism

$$\frac{(\operatorname{im} A)^*}{\ker A^*|_{(\operatorname{im} A)^*}} \simeq \operatorname{im} A^*$$

but the kernel is just $\{0\}$ so we get $(\operatorname{im} A)^* \simeq \operatorname{im} A^*$, □

EXAMPLE. Let's go through an entire example and verify that each of the above theorems actually makes sense. Let $V = \mathbf{R}^3$, and consider a matrix projecting onto the plane $x + y + z = 0$. We will pick the matrix Π sending $\vec{x}$ to

$$\Pi(\vec{x}) = \vec{x} - \frac{\vec{x} \cdot (1,1,1)}{3}(1,1,1).$$

Geometrically, this is the orthogonal projection onto the plane. If $\vec{x}$ is in the plane, then the extra term goes away because the dot product with $(1,1,1)$ is zero and so $\vec{x}$ is fixed. Otherwise, taking the dot product of the result with $(1,1,1)$ gives us zero and hence the result is in the plane. It is not too hard to check that this is also a linear map (we need to check: $\Pi(c\vec{x}) = c\Pi(\vec{x})$, $\Pi(\vec{x} + \vec{y}) = \Pi(\vec{x}) + \Pi(\vec{y})$).

Let's write down the matrix. We get

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Because $\operatorname{im}\Pi$ is the plane $x + y + z = 0$, the dimension of the image is 2. We know the kernel should be $\langle(1,1,1)\rangle$, but we can do it algebraically. For (x, y, z) to be in the kernel, we need $2x - y - z = 0, -x + 2y - z = 0, -x - y + 2z = 0$. Subtracting the second equation from the first, $3x - 3y = 0$ so $x = y$. Similarly, $y = z$ and so $x = y = z$ - this works as a solution, so all solutions are $(x, y, z) = (c, c, c)$.

We have $\dim V = 3$ and so rank nullity holds: $3 = 2 + 1$. Now let's consider what the quotient space $V/\ker\Pi$ looks like. We can visualize it as a vector space where the individual vectors are now lines through the plane $x+y+z = 0$ that are perpendicular to the plane. We add two lines by adding all pairs of vectors on the lines and writing down a new line as the set of possible sums. It is uniquely identified by using representatives in $\operatorname{im}\Pi$ and adding these in the subspace $\operatorname{im}\Pi$ will be the same as adding the lines.

LECTURE 5: THE DETERMINANT

Let $T : \mathbf{R}^n \to \mathbf{R}^n$ be a linear map. Given a region $\Omega \subset \mathbf{R}^n$, we have

$$\operatorname{vol} T(\Omega) = C \operatorname{vol} \Omega.$$

Why is this the case? Imagine making Ω out of very small hypercubes. Note that T affects a cube's shape in the same way, no matter where it is placed in $\mathbf{R}^n$. This is because if the cube is centered at v, its vertices are $\{v + v_S : S \subset \{1, \dots, n\}\}$ so that v_S has 1 as an entry on elements of S and 0 otherwise. Applying T, by linearity we get a cube at $T(v)$ but the vertices are $\{T(v) + T(v_S)\}$ - relative to $T(v)$, this is the same for a cube anywhere! Hence, the volume of each of these cubes making up Ω gets scaled by some constant C. Touching cubes remain touching, so applying T to our mass of cubes gives us a mass of touching deformed cubes, each scaled in volume by C. Hence, $\operatorname{vol} \Omega$ scales by C. A picture of this is shown below:

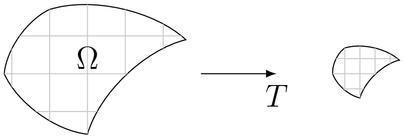


FIGURE 1. Applying T to a blob Ω.

This constant C should be independent of basis, because it is a geometric fact. You might have already seen formulas like

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc,$$

or some big expressions for larger matrices. Taking into account signed volume, these tell us the constant C. We call this number the determinant of T. We can actually write down an explicit formula.

DEFINITION. A *permutation* σ is a bijective map $\{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$. Here, bijective means for each i we can find $\sigma(j) = i$, and also that $\sigma(i) \neq \sigma(j)$ if $i \neq j$.

DEFINITION. The symmetric group S_n is the set of all permutations $\{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$. Because permutations are functions, we can compose the functions to get extra structure on this set. This is why we call it a 'group'.

DEFINITION. We have a map $\operatorname{sign} : S_n \to \{\pm 1\}$, which sends

$$\sigma \mapsto (-1)^{N(\sigma)}$$

where $N(\sigma)$ is the number of inversions of σ (i.e. pairs (i, j) where $i < j$ but $\sigma(i) > \sigma(j)$).

EXAMPLE. Consider the permutation in S_3 sending $1 \mapsto 1, 2 \mapsto 3, 3 \mapsto 2$. This has 1 inversion, and hence sign -1.

We can now state the usual definition of the determinant for the matrix of a map $M \in \operatorname{Hom}(V, W)$, which is given by

$$\det M = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \prod_{i \leq n} M_{i,\sigma(i)}.$$

At this point is seems like magic, but this number is actually the same for all matrix representations of our linear map. It is intrinsic to the linear map, and actually has nothing to do with a given basis.

This allows us to calculate it, but why this formula? Where does it come from? Let's start from a geometric perspective in $\mathbf{R}^n$, and carefully work our way up to this formula.

Matrix multiplication corresponds to composition of the respective linear maps. If we view these as scaling space, what this means is that the matrix AB scales by B and then by A, making it clear that the determinant should be multiplicative. Just follow what happens to $S : S \mapsto B(S) = S', S' \mapsto A(S')$. Because the determinant is the scaling fact for *any* subset, we see that

$$\text{vol}(S') = |\text{det}(B)|\text{vol}(S) \text{ and } \text{vol}(A(S')) = |\text{det}(A)|\text{vol}(S') = |\text{det}(A)\text{det}(B)|\text{vol}(S).$$

This is probably the best way to understand what a determinant means because you can get a visual picture and it helps you to understand matrix multiplication as repeated scaling.

How do we calculate the determinant? Well, it suffices to find out how it stretches a unit cube by our observations above. The idea is to turn the volume parallelepiped into an algebraic object which represents the volume. In this case, if the parallelepiped is defined by the vectors $v_1, \ldots, v_n$ we will denote its volume by $v_1 \wedge v_2 \wedge \ldots \wedge v_n$. We can make statements about relative volume, such as

$$v_1 \wedge v_2 \wedge \ldots \wedge v_n = C v_1' \wedge v_2' \wedge \ldots \wedge v_n'$$

This would mean that the parallelepiped defined by the v_i' has C times the volume of the parallelepiped defined by the v_i. We can write down the following rules about these volumes:

- (1) (Additivity) $\ldots \wedge v + w \wedge \ldots = \ldots \wedge v \wedge \ldots + \ldots \wedge w \wedge \ldots$
- (2) ($\mathbf{R}$-Linearity) $\ldots \wedge Cv_i \wedge \ldots \wedge v_j \wedge \ldots = \ldots \wedge v_i \wedge \ldots \wedge Cv_j \wedge \ldots$
- (3) (Alternating)$\ldots \wedge v \wedge \ldots \wedge v \wedge \ldots = 0$.
- (4) (Anticommutative) Swapping any two vectors inverts the sign.

From the perspective of representing volume, each of these relations has a very specific meaning. The first and second properties must be true because the signed volume is the volume of the parallelepiped P spanned by $v_1, v_2, \dots, v_{n-1}$ multiplied by length of the projection of v_n onto the $n-1$ dimensional hyperplane through P. For example, in $\mathbf{R}^2$ this is just base times height. The projection is a linear map, which is where relations (1) and (2) come from. This illustrated below:

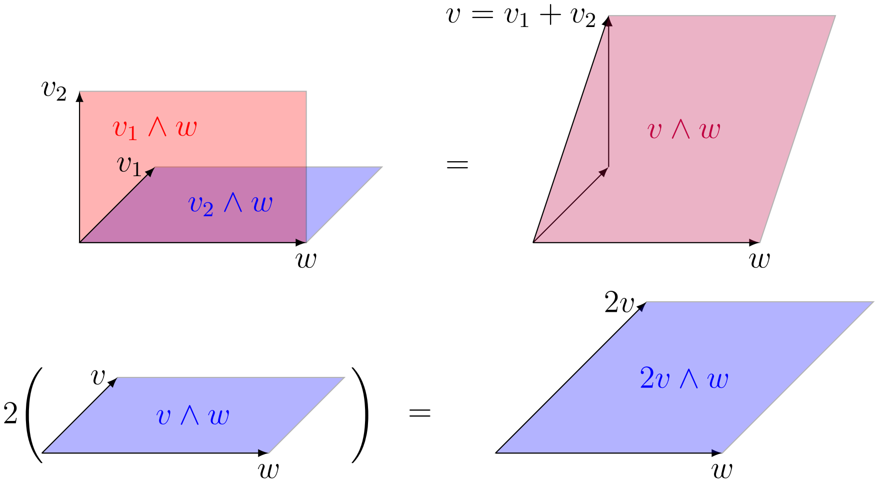
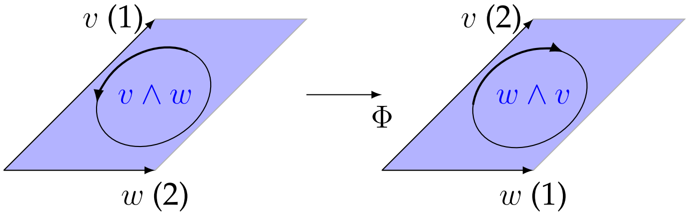


FIGURE 2. Axioms (1) and (2).

Here, we can see that the area is given by $\mathrm{base} \times \mathrm{height}$. The base is constant, so we look only at the height (i.e. the length of the projection onto w). The sum of the two heights of v_1, v_2 is the height of v, and hence the sum of the areas of the red and blue parallelograms equals the area of the purple one.

The next relation, (3), simply eliminates degenerate parallelepipeds. The final relation, (4), introduces the notion of a signed volume that reflects the orientation of the parallelepiped. Let's look at the case of $v \wedge w$ versus $w \wedge v$. On the surface, these would appear to define the same parallelogram. If we number the vector based on which appears first, the linear map Φ sending $w \mapsto v, v \mapsto w$ goes between the two cases.



In the case of a square, we can view Φ as just flipping the parallelogram (hence the name 'orientation'!). If we consider a directed loop γ within the parallelogram, applying Φ will actually reverse the direction! Thus, these two parallelograms have different orientations. To reflect the direction swap, we set $v \wedge w = -w \wedge v$. [2] This is illustrated below. Note that (4) implies (3), but we include both for the geometric interpretation.

It turns out these are enough to define the determinant. Since if we set Ω to be the parallelipiped defined by $v_1, \ldots, v_n$ we have $\mathrm{vol}(T\Omega) = \det T \,\mathrm{vol}(\Omega)$, We have already explained why it should be the case that

$$\frac{Tv_1 \wedge Tv_2 \wedge \ldots \wedge Tv_n}{v_1 \wedge v_2 \wedge \ldots \wedge v_n} = \det T.$$

Using the rules (1)-(4), we can compute the top quantity when the v_i are the standard basis in $\mathbf{R}^n$.

Pick a matrix M for T. This becomes a matter of simplifying $M(e_1) \wedge M(e_2) \wedge \ldots \wedge M(e_n)$ into the form $Ce_1 \wedge e_2 \wedge \ldots \wedge e_n$. We can expand this as

$$M(e_1) \wedge M(e_2) \wedge \ldots \wedge M(e_n) = \sum_j M_{1j} e_j \wedge \sum_j M_{2j} e_j \wedge \ldots \wedge \sum_j M_{nj} e_j$$

and we can then simplify using the relations for $\Lambda^n(V)$. The first and second are essentially the distributive property, and so we can proceed like we are multiplying and implement the other relations later. Since we cannot have repeated vectors (property (3)), every combination we pick upon expanding that is not sent to zero must be a permutation. Hence, upon expanding we will get

$$\sum_j M_{1j} e_j \wedge \sum_j M_{2j} e_j \wedge \ldots \wedge \sum_j M_{nj} e_j = \sum_{\sigma \in S_n} \left(\prod_{i=1}^{n} M_{i\sigma(i)} \right) e_{\sigma(1)} \wedge e_{\sigma(2)} \wedge \ldots \wedge e_{\sigma(n)}.$$

Noting that $e_{\sigma(1)} \wedge e_{\sigma(2)} \wedge \ldots \wedge e_{\sigma(n)} = \mathrm{sign}(\sigma) e_1 \wedge e_2 \wedge \ldots \wedge e_n$ by the anticommutative property, we immediately obtain the Liebniz formula for the determinant. From this formula, it is easy to derive most of the other common methods for computing the determinant.

Let's go back to the geometric picture in $\mathbf{R}^n$ and think about what the determinant actually tells us. Suppose $T \in \mathrm{Hom}(V, V)$ has $\det T = 0$. This means it sends every $\Omega \subset \mathbf{R}^n$ to a subset of volume zero. It is certainly not invertible in this case, since then there would exist a linear map sending $T(\Omega)$ of volume 0 to $\mathrm{vol}\,\Omega > 0$. But this doesn't exist, since $0 \times C = 0$ and linear maps scale volumes. Thus, we should expect the following:

[2] In general, this idea of orientation is captured by an idea called local homology. This is probably not accessible right now, but the first step to learning something is knowing it exists!

THEOREM. *Let $T \in \operatorname{Hom}(V, V)$. Then T is an isomorphism if and only if* $\det T \neq 0$.

Proof. We know the only if part already, but we will formalize it. Formalizing the above argument, suppose $\det T = 0$. Then since $T^{-1}T = \mathrm{id}$ and $\det \mathrm{id} = 1$, we would obtain

$$\det(T^{-1})\det(T) = 1$$

but then 0 would have an inverse, which is impossible. Hence, we can only have an inverse if $\det T \neq 0$.

Next, suppose that $\det T \neq 0$. We claim that T is surjective. Noting that $\operatorname{im} T$ is a subspace of V, if it was not the entire space then T would send $e_1, \dots, e_n$ to a set of n vectors in a subspace W of dimension at most $n-1$. But then they are not linearly independent. Considering the map in $\Lambda^n(V)$

$$v_1 \wedge v_2 \wedge \ldots \wedge v_n \mapsto T(v_1) \wedge T(v_2) \wedge \ldots \wedge T(v_n),$$

setting $v_i = e_i$ we get the result

$$e_1 \wedge e_2 \wedge \ldots \wedge e_{n-1} \wedge \sum_{i \leq n-1} c_i e_i.$$

Axiom (3) tells us this is zero. Hence $\det T = 0$, a contradiction. Thus, $\operatorname{im} T = V$. By the results of the previous section, it is an isomorphism. □

In fact, we can explicitly give the inverse in terms of the cofactor matrix:

$$T^{-1} = \frac{1}{\det T}[\det C_{ij}]_{ij},$$

where C_{ij} is the determinant of T when we remove the ith row and jth column.

LEMMA. *A left inverse is also a right inverse. That is, $AB = I$ if and only if $BA = I$.*

Proof. We will show $AB = I$ implies $BA = I$. Note that $\det(AB) = 1$, so $\det(A), \det(B)$ are both nonzero. Thus, both A and B are isomorphisms by the theorem above.

In particular, $B(AB)B^{-1} = BB^{-1} = I$, but $B(AB)B^{-1} = BA$ so we are done. □

Lecture 6: Eigenvalues and Eigenvectors

Last time, we defined the determinant and proved the Liebnitz formula

$$\det M = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \prod_{i \le n} M_{i\sigma(i)}.$$

There are some clear advantages to understanding determinants as we showed last time. For example, it allows us to determine whether or not a linear operator is invertible.

It also lets us solve a different problem. Given a linear map T, there are many different sorts of bases we could pick. However, we usually want to pick the easiest possible basis, so computations are easy.

The easiest possible thing we could expect is to get a matrix in 'diagonal' form

$$T = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

This form makes every sort of analysis easy - it's easy to multiply, it's easy to see that it's invertible, etc. Well, in such a basis $\{v_1, \dots v_n\}$, by our construction of the matrix we would need $Tv_i = \lambda_i v_i$. How do we find these vectors and numbers? Well, consider the linear map

$$T_i := T - \lambda_i \mathrm{id}.$$

On the subspace W spanned by v_i, this is zero. Hence, $\ker T \neq 0$ and T is not an isomorphism. Then we have

$$\det(T - \lambda_i \mathrm{id}) = 0.$$

We can set v_i to be some vector in the kernel! Hence, the task of determining the λ_i (if they exist) is equivalent to finding λ so that $\det(T - \lambda \mathrm{id}) = 0$. Using the Liebnitz formula, this is a degree n polynomial.

Definition. An *eigenvalue* is a solution λ to $\det(T - \lambda \mathrm{id}) = 0$.

Definition. An eigenvector for an eigenvalue λ is a vector v in $\ker(T - \lambda \mathrm{id})$. The set of all eigenvectors of λ is the eigenspace.

Lemma. *$\det T$ is the product of the eigenvalues of T.*

Proof. Practically by definition: $p_T(0) = \det(T)$. The constant term in a polynomial is the product of its roots. The roots of p_T are the eigenvalues. □

Theorem. *Suppose $\det(T - \lambda \mathrm{id})$ has n distinct nonzero roots over k. Then T can be written as a diagonal matrix like above. We call T diagonalizable, and the basis is known as the eigenbasis.*

Proof. The roots are nonzero, so in particular $\det(T) \neq 0$ so it is an isomorphism. We have n distinct λ_i so that $\det(T - \lambda_i \text{id}) = 0$. Each is not an isomorphism, so $\dim\ker(T - \lambda_i \text{id}) > 0$. Hence, we can pick some v_i in the kernel. These vectors must be distinct, since λ_i are distinct. Additionally, if $\sum_i c_i v_i = 0$, upon applying T we get

$$\sum_i c_i \lambda_i v_i = 0.$$

But also note that $\sum_i c_i \lambda_n v_i = 0$, so taking the difference of the two we get $\sum_{i<n} (\lambda_i - \lambda_n) v_i = 0$. Using induction, we see that $c_1 = 0$. Plugging this back in and using the same argument, $c_2 = 0$ and so on until we see all c_i are zero. Hence, the v_i are linearly independent. They also span because there are n of them, so they are in fact a basis. In this basis we get a matrix of the desired form.

Note that we also get $\dim\ker(T - \lambda_i \text{id}) = 1$ from this explicit representation, so we can only pick the v_i up to scalars. □

This polynomial $\det(T - \lambda \text{id})$ is important, and it is known as the *characteristic polynomial*. We denote it by $p_T(\lambda) := \det(T - \lambda \text{id})$. Because $\det$ depends only on T as a linear operator, p_T depends only on T as a linear operator. This means it is invariant under basis change.

Definition. The trace $\text{tr}(T)$ is the coefficient of λ^{n-1} in p_T. Equivalently, it is the sum of the eigenvalues. Since p_T is invariant under basis change, $\text{tr}(T)$ is as well.

Lemma. *We have*

$$\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B).$$

Proof. Pick matrices for A and B. We claim $\text{tr}T$ is the sum of the diagonal entries. To see this, consider $\det(A - \lambda \text{id})$ and the Liebnitz formula. To contribute to the λ^{n-1} term, the permutation σ needs to be the identity. Otherwise, you can only get at most λ^{n-2} since a permutation fixing $\geq n - 1$ elements is the identity. This term is $\prod_i (A_{ii} - \lambda)$, and so the coefficient of λ^{n-1} is $\sum_i A_{ii}$. The claim is easy from here, since the trace is independent of the matrix representation of an operator. □

Example. Remember that $\mathbf{C}$ is a vector space over $\mathbf{R}$. Picking the basis $\{1, i\}$, we write $a + bi$ as the following matrix:

$$a + bi = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

Let's make a few observations.

- We have $\det_{\mathbf{R}}(z) = |z|^2$.
- We have $z + \overline{z} = \text{tr}_{\mathbf{R}} z$

- ∘ Multiplication of complex numbers corresponds to multiplying their matrices (check this!).

In other words, we can say everything we need about complex numbers by representing them as matrices over $\mathbf{R}$. They are not really imaginary after all!

EXAMPLE. Consider the matrix

$$T = \begin{bmatrix} 1 & 4 & 1 \\ 0 & 6 & 4 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let's find all of the eigenvalues. We can immediately spot that 1 is an eigenvalue, since $Te_1 = e_1$. Since the sum of the eigenvalues is 8 and the product is 6, for the other two eigenvalues we get

$$\lambda_2 + \lambda_3 = 7, \lambda_2\lambda_3 = 6.$$

That is, they are roots of $x^2 - 7x + 6$. The roots of this are 1 and 6. Thus, we get $p_T(\lambda) = (\lambda - 1)^2(\lambda - 6)$.

We can also do this another way. Consider $\det(T - \lambda \mathrm{id})$. For every $\sigma \in S_n$, we get a zero in the product unless $\sigma = \mathrm{id}$. Hence, $\det(T - \lambda \mathrm{id}) = \prod_i (\lambda - T_{ii})$, which matches what we got.

EXAMPLE. Consider the matrix

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

Let's find the eigenvalues as well as the eigenvectors. We have

$$\det(T - \lambda \mathrm{id}) = \det \begin{bmatrix} 1 - \lambda & 2 \\ 3 & 4 - \lambda \end{bmatrix} = (1 - \lambda)(4 - \lambda) - 6.$$

Solving, we get $\lambda = \frac{5 \pm \sqrt{33}}{2}$. Explicitly solving $T(x, y) = \lambda(x, y)$, we get

$$v_1 = \left(\frac{1}{6}(-3 + \sqrt{33}), 1\right), v_2 = \left(\frac{1}{6}(-3 - \sqrt{33}), 1\right)$$

as our eigenvectors. Then span of these gives the eigenspaces. Written in this basis, T is diagonal.

EXAMPLE. The matrix

$$T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

cannot be made diagonal. This doesn't contradiction our theorem, because if you use the method above you'll see the eigenvalues are not distinct.

THEOREM (Cayley-Hamilton). *Let $k = \mathbf{C}$, $T \in \mathrm{Hom}(V, V)$. Then $p_T(T) = 0$, where 0 is the zero matrix.*

Proof. Suppose that T can be made diagonal. Then

$$p_T(T) = \begin{bmatrix} p_T(\lambda_1) & & & \\ & p_T(\lambda_2) & & \\ & & \ddots & \\ & & & p_T(\lambda_n) \end{bmatrix} = 0.$$

But this isn't possible for every matrix - just look at the example above! What is true is that given a random matrix, its characteristic polynomial will probably have distinct roots (a random polynomial almost certainly has distinct roots) and hence T will be diagonalizable. Hence, the theorem holds for almost all matrices. Noting that $\mathrm{Hom}(V, V)$ is a $\mathbf{C}$-vector space, if we identify it with $\mathbf{C}^n$ the diagonalizable matrices form a 'dense' subset, meaning there is a diagonalizable matrix abritrarily close to any given matrix.

Now the map

$$\varphi : T \mapsto p_T(T)$$

is a differentiable map from $\mathrm{Hom}(V, V))$ to itself. Hence, it is a continuous map, and is 0 on a dense subset of W. Given a matrix T, for any $\epsilon > 0$ there exists a ball of radius δ so that $|T - T'| < \delta$ and $|\varphi(T) - \varphi(T')| = |\varphi(T)| < \epsilon$. Note that $|\cdot|$ here is the norm on $\mathbf{C}^n$, $|v| := v \cdot v$. Hence, $\varphi(T)$ is arbitrarily close to 0 and must be 0. □

This works over $\mathbf{R}$, since $\mathbf{R} \subset \mathbf{C}$.

EXAMPLE. We can use this to find eigenvectors! Take the 3×3 matrix from the previous example with characteristic polynomial $p_T(\lambda) = (\lambda - 1)^2(\lambda - 6)$. By Cayley-Hamilton, for any $v \in \mathbf{R}^3$ we know that

$$(T - I)^2(T - 6I)v = 0,$$

as the result is the zero matrix. Suppose we want to find the eigenspace for 6. Given any $v \in V$, we know that $(T - I)^2 v \in \ker(T - 6I)$, or that it is in the desired eigenspace. So if we pick a vector at random, this sends us to the eigenspace we want. Since it is one dimensional, we have completely determined it.

In general, if T has distinct eigenvalues the linear map $\Pi : V \to \ker(T - \lambda_i I)$ given by

$$\Pi = \prod_{j \neq i} \frac{T - \lambda_j I}{\lambda_i - \lambda_j}$$

is a projection from V to the 1-dimensional eigenspace for λ_i.

EXAMPLE. There is a simple formula for the inverse of a 2×2 matrix, provided it exists:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

You can check this by multiplying, but we can actually derive it in a pain-free way if we use Cayley-Hamilton. Let A be our matrix. Then $A^2 - (\mathrm{tr} A)A + (\det A)I = 0$. Then

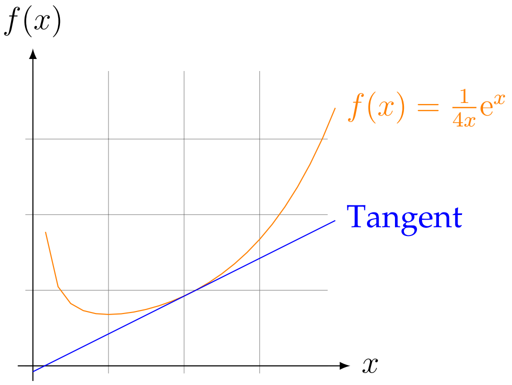
$$\frac{1}{\det A}\left((\mathrm{tr} A)A - A^2\right) = I.$$

Multiplying by A^{-1}, we get $A^{-1} = \frac{1}{\det A}\left((\mathrm{tr} A)I - A\right)$.

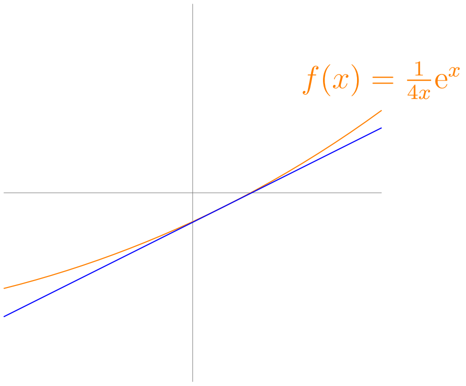
Lecture 7: Uses of Linear Algebra

We've seen some important theorems in linear algebra and a lot of linear maps, but why do people care so much about linear algebra? To answer this question, we need to see some ways that linear algebra can be incredibly useful. This lecture will be all examples.

EXAMPLE. Consider a function $f : \mathbf{R} \to \mathbf{R}$. We can make a graph of it like this:

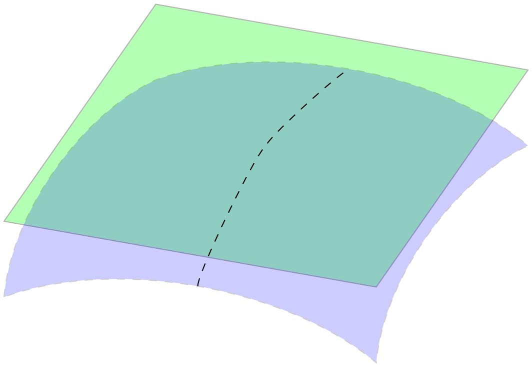


At each point, we can draw a tangent line which tells us how fast f is changing. If we were to zoom in on this point, we would see that the tangent line approximates the curve really well around a point:



In fact, very close to the point of tangency the line and the function are virtually indistinguishable. The line is a good linear approximation for the function at a point. We call the set of slopes of all these lines the derivative of a function. These lines are all of the best linear approximations of the functions at each point. This is true for most 'nice' functions. If we treat the point of tangency p as the origin, what this says is that f looks locally like a linear function in $\mathrm{Hom}(\mathbf{R}, \mathbf{R})$. These are 1×1 matrices, representing the maps $x \mapsto cx$. The value c is the slope of the line.

Now suppose we have a function $f : \mathbf{R}^n \to \mathbf{R}$, which we will write as $f(x_1, \ldots, x_n)$. Before, at a point we approximated with a line, but now we need a (hyper)plane:



This is again saying the same thing - around the point of tangency p, the function is very close to the plane. As before, the plane describes the rate of change of f at the point p. If we pretend that p is the origin, we can again phrase this in terms of linear maps. If we were to plot a linear map $\mathbf{R}^n \to \mathbf{R}$ in the same way as we plot f, we would get a plane through the origin. Hence, this is saying that around p the function f is approximated by some linear map in $\operatorname{Hom}(\mathbf{R}^n, \mathbf{R})$.

We can go a step further, but it becomes a bit trickier to draw pictures. Assuming that for a nice function $\mathbf{R}^n \to \mathbf{R}^m$ the 'rate of change' becomes constant locally, we can approximate f at a point $p \in \mathbf{R}^n$ with a linear map $df_p \in \operatorname{Hom}(\mathbf{R}^n, \mathbf{R}^m)$ so that $f(p + \vec{x}) \approx df_p(\vec{x})$.

If you have seen multivariable calculus, you know this linear as the Jacobian matrix. We can write it in terms of derivatives as

$$df_p = \left[\frac{\partial f_i}{\partial x_j}\right]_{ij},$$

where $f(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x}))$. These are incredibly useful because they allow us to generalize the notion of a derivative, and also can be very handy for changing variables in an integral of the form

$$I = \int_\Omega f(x_1, \ldots, x_n)\mathrm{d}x_1\mathrm{d}x_2 \ldots \mathrm{d}x_n,$$

since we might be able to pick variables that make integration easier over $\Omega \subset \mathbf{R}^n$.

EXAMPLE. A graph is a pair (V, E), where V is the set of vertices and E is the set of edges (which connect two vertices). They might look something like this:

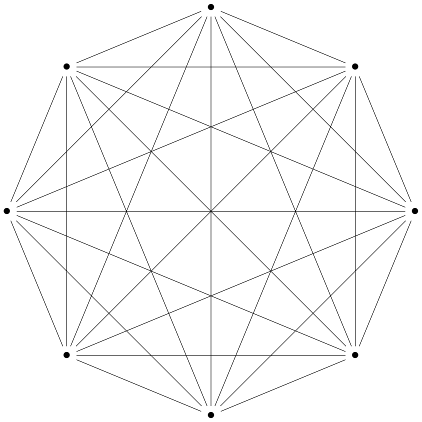


FIGURE 3. A complete graph

In the modern world, graphs are extremely useful as data structures and any understanding we can gain about graphs on a large scale is invaluable. Linear algebra actually allows us to do some of these things.

Let's look at the first version of PageRank. Each node represents a website, and the edges represent connections between them such as links. We want to assign a vector v describing the ranks of each page, so that a higher rank rage indicates that it is more important in the network.

Imagine walking from node to node, taking a random edge each time. It makes sense that you would end up at important nodes like Wikipedia more often than some random website. So, if we can measure the "stable" distribution of many random walkers along the graph, we can get some sense of how important websites are.

At each node i, we assign probabilities p_{ij} of visiting node j along a random link. Define the matrix

$$\Sigma := [p_{ij}]_{ij},$$

which we will call the random-walk matrix. Let v_0 denote some random distribution of random walkers living at each node, so that the sum of its components is 1. Then Σv_0 is the distribution after each walker picks a node, and $\Sigma^n v_0$ for $n \gg 0$ will reflect the distribution after a long time. The sum of its components will remain 1.

Almost certainly, the eigenvalues of Σ will be distinct. If this is the case, under the correct basis Σ is a diagonal matrix $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$. For Σ^n, the largest eigenvalue will dominate. In particular, our result will go towards a stable state which is the eigenvector of the largest eigenvalue. Normalizing this so that it is a unit vector, we can use this to rank the important of each node. There exist fast algorithms to calculate eigenvectors, so this is reasonable in practice even with a very large matrix.

EXAMPLE. Suppose we have some discrete function

$$x : \mathbf{Z} \to \mathbf{C}$$

which is periodic: that is, there exists some N so that $x[n + N] = x[n]$, so really all the information about x is contained in $x|_{[0,N-1]}$. This function could be a lot of things, but here we will pretend it is some digital signal and we want to retain important information about x while lowering the amount of information we have to store. For example, we want to remember things like peaks but maybe not a tiny bump.

Suppose our signal contains information about light. You're probably aware that light comes in different frequencies, which determine the color. But when we get light from a source, it's usually a lot of different frequencies of light all mixed up. There is a nice basis for these functions in terms of waves. In particular, we can use the basis functions $x_k[n] = e^{2\pi ikn/N}$. If you've seen complex numbers, basically what $x_k[n]$ does it rotate around a circle at faster and faster paces. The parameter k describes the frequency. At our precision, these are really all the frequencies we can make sense of. We want to write

$$x[n] = \sum_k a_k x_k[n].$$

Because these elements $x_k[n]$ form a basis, we can uniquely find the numbers a_k and figure out the frequencies that went into the signal.

This is tremendously useful! Here are just a few:

- Image compression: ignore unimportant frequencies, so that we store fewer numbers but get roughly the same function.
- Removal of noise in images.
- Fast multiplication using about $n \log n \log \log n$ operations multiply n-bit integers (need to use FFT algorithm).
- Solve PDEs.
- Spectral analysis.