The Elements of Linear Algebra
Volume One: Livre des Lignes Droites

Theron J Hitchman

Fall 2017

Contents

To the Student:

So. I am writing this book. (Let's not pretend I am done. You can see the state it is in.) There are dozens (hundreds?) of introductory linear algebra books, so it is pretty reasonable to ask why I am putting in the effort, and in the meantime, causing this much pain. I should explain.

I have taught linear algebra many times and I have liked some books, but never loved one. The closest match for what I wanted to teach is Strang's *Introduction to Linear Algebra*, and I am sure that people who have read that will see some influences here. But my students never seemed to connect with Prof. Strang's enthusiastic, stream-of-consciousness prose. And over time, I found that the things I need to emphasize for my students just don't match with that text, or any other.

In addition, most textbooks assume a certain class structure: lectures accompanied by weekly homework, with some exams. I don't want to run our course that way.

So this book is my solution. It is my attempt to make a thing which matches how I want our class to run.

Here is what you should expect. This book has the basics of linear algebra, done thoroughly. I want this to help you see why some very basic, important things are done the way they are. It is amazing how much of the subject of linear algebra can be done by focusing on small examples, that is, in small dimensions. (We'll learn about *dimension* later on.) We'll sort out other things through assignments and class discussions.

It is important to read this book actively. If you haven't learned how to read a math text before, there are some key ideas:

To the Student:

Time Mathematics is often technical and tricky. It takes time to absorb. Plan to give yourself lots of time to read and think. And don't be surprised if you have to read some section more than once. (This is not a novel. As much as I see it as a story, it won't sweep you away.)

Examples In the interest of brevity, I have streamlined the exposition. In particular, there are no examples. **The point is that you should make your own.** This is so important a skill that it is basically a mathematical super-power. Whenever you come across an idea, if you understand it or not, you should make some very explicit examples and consider them carefully.

Questions As part of your *active* engagement with the text, you will find things that don't quite make sense, yet. This is normal. The mathematician's best approach then is to (1) write down a specific question or two about the confusing bit, and (2) talk to other people about it. You are fortunate that you have an instructor and classmates to talk to. Make lists of questions and try to get them answered!

The real beauty in linear algebra is the tight set of connections between algebra and geometry. I hope you enjoy it.

The Equation of a Line in the Plane

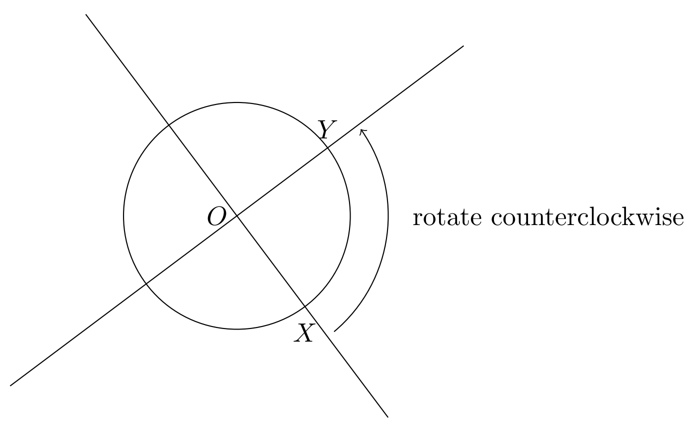
Let's recall the idea of *the plane* from classical geometry: the plane is like a flat sheet of drawing paper, which extends indefinitely in all directions without bound. It is the playground for lots of serious considerations from high school: points, lines, circles, triangles, rectangles, and various other doodles live in it. I say *in* it rather than *on* it, because all of those objects really have their existence inside the plane. If we were to say "on the plane" then one might think of them as sitting on top of the paper, where a light breeze might move them about. No, those things live inside the plane just as sure as you and I live inside the universe.

And while all of that is evocative and romantic, it doesn't make doing mathematics any easier. Our aim in this first chapter is to do some concrete mathematics: we want to figure out how to describe a single line in the plane very carefully. To do this, we will use tools that René Descartes taught us: coordinates. Better yet, we will use an update of the idea and introduce *vectors*. Our work has to rely on something, so at some points we will make use of geometry facts you learned in high school. But for all of the points and vectors, angles and dot products, we will go straight to the heart of a single important question:

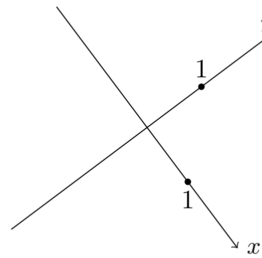
> **How can we clearly describe a single line in the plane with an equation?**

Points, Vectors, and Vectors

You have likely seen the idea of *Cartesian coordinates* on the plane before. To be clear, let's set things down carefully. In the plane, we choose a pair of perpendicular lines which meet at a point O. This special point is called the *origin*. Then, we choose a point X on one of the lines and draw the circle centered at O which passes through X. Note that this circle meets our two lines in two points each, four points total, one of which is the point X. Then, from X, we rotate around the circle by a quarter turn counterclockwise until we hit one of the points on the other line. This new point we will call Y. Are you drawing with me? Here is my picture so far.



We call the line OX the *x-axis* and the line containing Y the *y-axis*. Here comes the amazing part: we declare the circle we used to be of *unit size*, and make the lines OX and OY into number lines! The important part is that the point O should represent 0 on both number lines, and the points X and Y should each represent 1 on their lines. So, instead of marking things with O, X, and Y, we put down marks where X and Y are and label them with 1's, and add little arrows marked with x and y near the positive "ends" of the lines OX and OY, respectively.



Note that above I have done something a bit silly and let the picture just fall on the paper in an unusual way. I really mean unusual as "not usual." The usual way arranges the lines on the paper to match our expected horizontal (x) and vertical (y) directions. This isn't actually required, but it is what everyone always does. The typical picture looks like Figure 1.

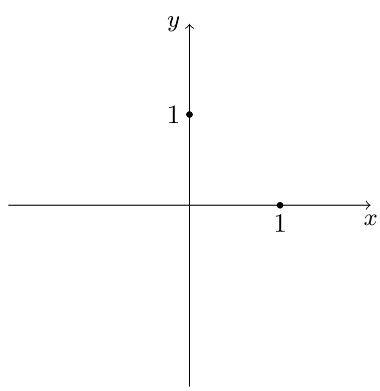


Figure 1: The Standard Cartesian Coordinate System

Now suppose we have some point in the plane, let's call it P. We can describe the location of P relative to our two lines in a simple way. First, we draw a line through P which is parallel to the y-axis and perpendicular to the x-axis. The foot of this perpendicular hits the x-axis at some point A. But this point A

is part of the number line OX, so it has an associated real number, which we will call a. So the point A is instead marked with the label a from this number line.

Similarly, we draw a line through P parallel to the x-axis and perpendicular to the y-axis. The foot of this perpendicular hits the y-axis at some point B, which is part of the number line OY. We denote the number associated to B by b. Again, the point is labeled with the number b from the number line.

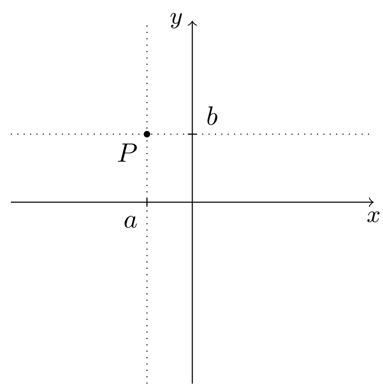


Figure 2: A point P and its coordinates (a, b).

So, to identify the point P, we can instead give the pair of numbers a and b. These numbers are called the *coordinates* of P. Of course, the order of the coordinates matters, so we make what we call an *ordered pair* (a, b) to keep things straight, where the x-coordinate comes first, and the y-coordinate comes second. Note that in Figure 2 the x-coordinate a is negative, but the y-coordinate b is postive.

This whole process is reversible, too. If we pick a pair of numbers c and d, in order, then we can find a point in the plane which corresponds, and we can do it unambiguously. Find the spot labeled c on the x-axis number line and construct a line perpendicular to the x-axis through this point. Similarly, find the spot labeled d on the y-axis and construct a line perpendicular to the y-axis through this point. The two lines you just drew will meet in exactly one point Q, and Q will have coordinates (c, d).

This setup of coordinates on the plane allows us to formalize a wonderful and useful idea from physics, too. Physicists use the concept of a *vector* to describe something (like the wind) which has both magnitude or size (like how fast the air is moving) and direction (which way the air is going). Usually, vectors are drawn as little arrows: the arrow has a direction, and it has a length which represents its magnitude. It is possible to draw vectors which have the same direction but different lengths, and vice versa.

We can use coordinates to describe vectors in the plane, too. Here's how: A physicist's vector v is some arrow in the plane. That arrow has an initial point P, called its *tail*, and a final point Q, called its *head*. We can write the coordinates of these points as $P = (a, b)$ and $Q = (c, d)$.

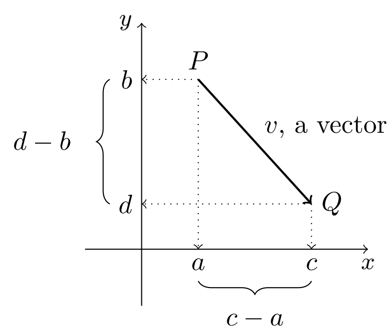


Figure 3: Coordinates for a physicist's vector v.

Then the coordinates of v are taken to be the numbers $c - a$ and $d - b$, which we interpret how much v acts in the directions parallel to the x-axis and the y-axis, respectively. Note that in Figure 3, the y-coordinate is negative, since $b > d$.

These coordinates have a hint of algebraic manipulation in them. Those subtractions line up almost like we could write $v = (c - a, d - b) = (c, d) - (a, b) = Q - P$. But v is a vector, and if we write it like $(c - a, d - b)$, it looks like the notation for a point. We should not do that because it could get confusing. Furthermore, that "equation" would mean that we are subtracting points and creating a vector, which is also weird. Still, there is something to it. We will return to this idea soon.

For now, let's focus on a bit of ambiguity in the physicist's idea of a vector. Where should that vector be? That is, given the coordinates of a vector in the plane, it is not clear where to draw it! I can slide a vector around the plane, and as long as I keep it parallel to the original, the coordinates won't change. So, unlike with the coordinates of a point, the coordinates of a vector do not uniquely specify the vector.

The mathematician's special fix is this: we simply declare all our vectors to have their initial points, their tails, at the origin O, of our coordinate system. That curtails some of the (admittedly useful) freedom in the physics notion, but it also lets us be more clear.

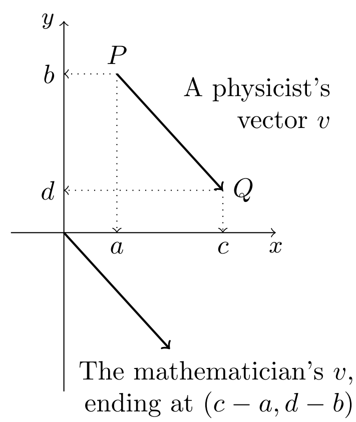


Figure 4: The physicist's vector vs. the mathematician's vector.

But it pays to keep in mind that the physicists conception of the vector V with coordinates $c-a$ and $d-b$ could be one of many different arrows, while the mathematician's vector v is the arrow from the point $O=(0,0)$ to the point $(c-a, d-b)$.

Now we have circled back around to a muddle. If a mathematician's vector is always based at O, we only need to specify where the head of the vector is... which is just a point. So, how is a vector supposed to be different from a point, again?

This confusion of three different, shifting, partially-overlapping interpretations causes some trouble to the new learner. Professionals tend to pass back and forth between these and use them flexibly to get results. Once you have gotten used to the ideas, you will, too. You should watch out for these instances where the words point and vector get interchanged. If they cause you trouble, remember that we have three different things, which are closely related.

For now, the simplest way to handle things is like this:

- Ignore the physicist's version of the word vector as much as possible.
- A point is a location in the plane, and represented by coordinates in the form of an ordered pair of numbers (a, b).
- A vector is an arrow based at the origin, which can be specified by the coordinates of its head. To keep this separate from the idea of a point, we will write it differently, with the numbers stacked vertically like this: $\left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$.
- Always remember that for each point in the plane, there is a unique mathematician's vector which corresponds, and vice versa.

With this in mind, we make our first official definition.

Definition 1. *A* 2*-vector is a vertical stack of* 2 *real numbers, like so:*

$$v = \begin{pmatrix} a \\ b \end{pmatrix}.$$

The collection of all such 2*-vectors is called* the plane*, and written with this notation:* $\mathbb{R}^2$*.*

The notation $\mathbb{R}^2$ is often read "arr-two," and many people use that language interchangeably with "the plane." Also, most of the time we will just say "vector," rather than "2-vector."

Vector Algebra

Let's return to that glimpse of subtraction we saw in Figure 3. We saw there that for points $P = (a, b)$ and $Q = (c, d)$, the vector v from P to Q has coordinates $c - a$ and $d - b$. This looks almost like we subtracted the points to get $Q - P = v$. Can we use that? The weird part is that it mixes up points and vectors. So, we will just change viewpoints, and instead think of P and Q as (mathematician's) vectors. To keep things clear, let's introduce new labels.

$$p = \begin{pmatrix} a \\ b \end{pmatrix}, \quad q = \begin{pmatrix} c \\ d \end{pmatrix}$$

If we put these together on the plane with the physicist's vector v from p to q and the mathematician's v we see a wonderful triangle, and an extra vector. So we see a way to talk about subtracting

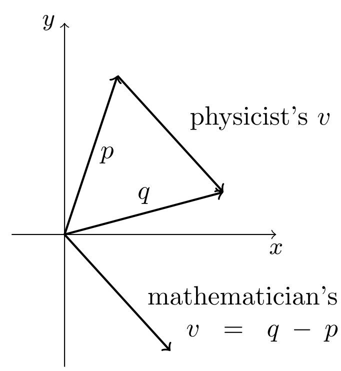


Figure 5: Subtraction of vectors

vectors: Given two vectors p and q as above, their *difference* is the vector

$$V = Q - P = \begin{pmatrix} c \\ d \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} c - a \\ d - b \end{pmatrix}.$$

Geometrically, we draw the arrow from p to q and then translate it down so that its tail is at the origin $O = (0, 0)$. Of course, the order of p and q in this operation matters. If we switch them, we get an arrow pointing in the opposite direction.

If we can subtract vectors, surely we can add vectors. How would that work? Algebraically, if $v = q - p$, then we expect $q = v + p$ by rearranging. That would mean

$$q = \begin{pmatrix} c \\ d \end{pmatrix} = v + p = \begin{pmatrix} c - a \\ d - b \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix},$$

which all fits. It looks like addition should be defined coordinate-by-coordinate.

Definition 2 (Addition of Vectors)**.** *Let $p = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ and $q = \left(\begin{smallmatrix} c \\ d \end{smallmatrix}\right)$ be two vectors in $\mathbb{R}^2$. Their sum is the vector*

$$p + q = \begin{pmatrix} a + c \\ b + d \end{pmatrix}.$$

Theorem 3. *Addition of vectors satisfies the same rules as addition of real numbers:*

- *when adding more than two vectors, it doesn't matter which operations you do first: $(p + q) + r = p + (q + r)$;*
- *one can add in either order $p + q = q + p$;*
- *the vector o corresponding to the origin O is a "zero" since $p + o = o + p = p$;*
- *for each vector p, there is an opposite vector $-p$ so that $p + (-p) = o$.*

Remember that you are supposed to read actively. You can draw all of these pictures and try out all of these things with specific examples that you invent. You should check these by making examples and working out the details. Can you also draw the pictures which go with your examples?

But what about subtracting geometrically? In Figure 5, I have a strong desire to complete the figure by joining the loose end of v to the rest of the figure. If we draw the arrow from the head of v to the head of q, we get this:

What should the label on the dashed vector in Figure 6 be? Just as the physicist's v and the mathematician's v are parallel,

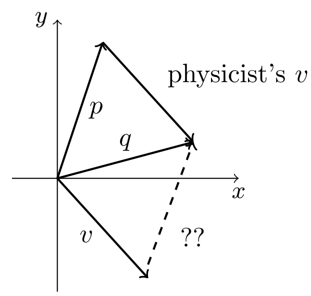


Figure 6: Subtraction of vectors

this new vector is parallel to the mathematician's vector p. So the dashed vector must be a physicist's version of p. Then we see $q = v + p$.

Now we know how to add geometrically: to add two vectors u and v, translate v until its tail is on the head of u, then draw a new vector $u + v$ as the vector from the tail of u to the head of this translated v. It just repurposes the structure of Figure 6.

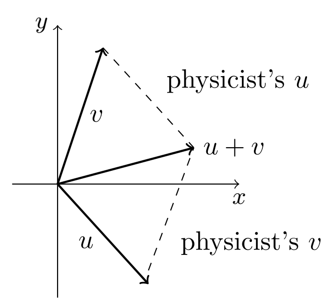


Figure 7: Geometric Addition of Vectors

This is sometimes called the *parallelogram rule* for addition of vectors.

There is another useful operation on vectors called *scalar multiplication*. The terminology comes from physics (again) where a *scalar* quantity is just a number, and not a vector. So "scalar multiplication" means to multiply a vector by a scalar.

Definition 4 (Scalar Multiplication for vectors)**.** *Let* $p = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ *be a vector in* $\mathbb{R}^2$*, and let* λ *be a real number. Then the scalar multiple* λp *is defined to be*

$$\lambda p = \begin{pmatrix} \lambda a \\ \lambda b \end{pmatrix}$$

If you have never seen the symbol λ before, it is an old Greek letter pronounced "lamb-duh." It is traditional to use it in linear algebra in lots of places. Welcome to the λ-club. Oh, there are others, too, like μ, which is pronounced "myoo."

Again, this operation has some important similarities to the familiar multiplication of numbers, but because it combines a scalar (a number) with a vector (not a number, exactly) to produce another vector (again, not a number) things are a little different.

Theorem 5. *Suppose that* p *and* q *are vectors, and* λ *and* μ *are numbers. Scalar multiplication has the following properties:*

- *Scalar multiplication distributes over vector addition:* $\lambda(p + q) = \lambda p + \lambda q$*;*
- *Scalar multiplication distributes over scalar addition:* $(\lambda + \mu)p = \lambda p + \mu q$*;*
- *Scalar multiplication and regular multiplication can be done in either order:* $\lambda(\mu p) = (\lambda\mu)p$*;*
- *if* $\lambda = 0$*, then* $\lambda p = 0p = o$ *is the zero vector.*
- *if* n *is a counting number, then* np *is the same thing as adding together* n *copies of* p*. In particular,* $1p = p$*.*

This Theorem, like the last one, just says that a bunch of natural things you expect to happen really do happen. When you study **Modern Algebra**, making lists of these kinds of properties will be really useful. So far, we have collected up the properties that describe a *vector space*.

What is the geometry of scalar multiplication? It corresponds to stretching (or shrinking) a vector, without changing its direction. Let λ be a non-zero number. Since

$$\lambda p = \lambda \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \lambda a \\ \lambda b \end{pmatrix},$$

we see that the ratio of the two coordinates of a vector doesn't change under scalar multiplication. This means that p and λp point in the same direction. One can see this by considering similar triangles with sides parallel to the x-axis, the y-axis, and p.

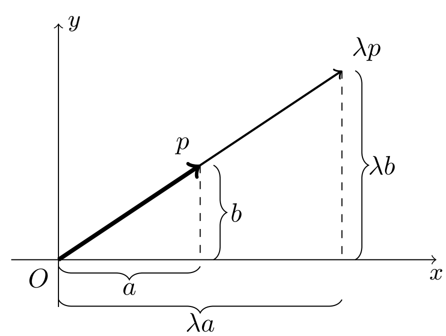


Figure 8: Similar triangles and scalar multiplication, $\lambda > 1$

The triangles in Figure 8 are similar: their corresponding horizontal and vertical sides are parallel, and those pairs of sides have a common ratio. We learn that p and λp lie in the same line.

Now we are getting somewhere! We want to understand lines in the plane, and we have just discovered how scalar multiplication relates to lines which pass through the origin, O.

By the way, this picture helps explain the terminology. The vector λp is a rescaled version of p. A *scalar* is a thing which *scales* vectors.

Lines as Parametric Objects

We see that for a non-zero vector p, and a non-zero number λ, the vectors p and λp lie on the same line through the origin, O. But this doesn't depend on which number λ we choose. So if we vary λ, we will get lots of different points on that line. The picture looks something like this one:

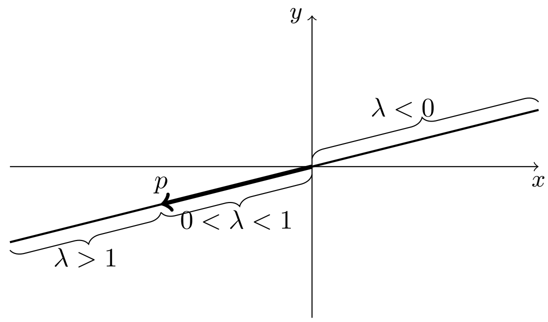


Figure 9: The points on a line: λp for different λ

This leads us to what is called a *parametric description* of the line. We think of some variable, say t, as a parameter. (I chose t here so that we think of it as "time.") As we change the value of t, the vectors tp trace out the line which goes through the origin and the point which is the head of the vector p.

Theorem 6 (parametric lines through the origin)**.** *Suppose that $P = (a, b)$ is some point in the plane. The the line which passes through the origin O and the point P consists of the heads of all the vectors*

$$tp = t \begin{pmatrix} a \\ b \end{pmatrix},$$

where t varies over all real numbers. That is, this line is the set of the heads of all scalar multiples of the vector p corresponding to P.

This, is fantastic. We can use simple vector algebra to describe any line through the origin. What about lines that are not through the origin? Suppose we just have two random points P and Q, neither of which is the origin, and we want to describe the line ℓ through P and Q?

Again, let's think of the points on this line ℓ as the heads of mathematics-style vectors. If we put in the vectors p and q which correspond to the points P and Q, we get a good picture like this one.

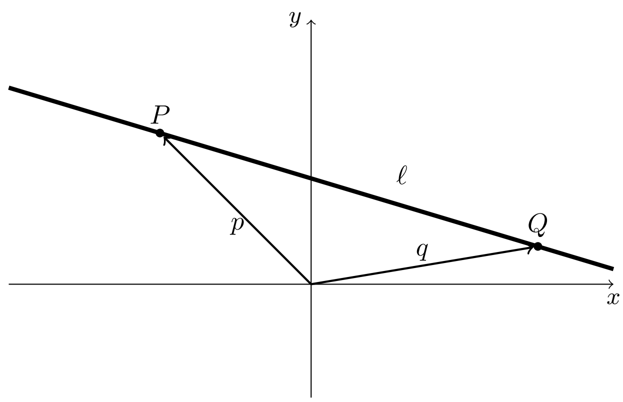


Figure 10: Working toward a parametric line, part 1

Each point in Figure 10 is the head of a vector from the origin to that point. In particular, the point P is the head of the vector p. So if we add $-p$ to every single one of these vectors, we will move the whole line ℓ in the direction of $-p$ and with the same distance as $-p$. That will make a new line, ℓ', and the point which comes from P will land on the origin.

Since q is one of the vectors with its head on the original line, the vector $q - p$ will have its head on the new line. But $q - p$ has its tail at the origin! So, our new line passes through the origin and the head of the vector $q-p$, and we are in a situation to apply Theorem 6.

By Theorem 6, the line ℓ' is described as all of the points which are heads of vectors of the form $t(q - p)$, where t is a parameter which is allowed to vary over all real numbers. But we get from ℓ' back to ℓ by simply adding the vector p back in. So, our original line ℓ, which goes through the heads of vectors p and q, is described

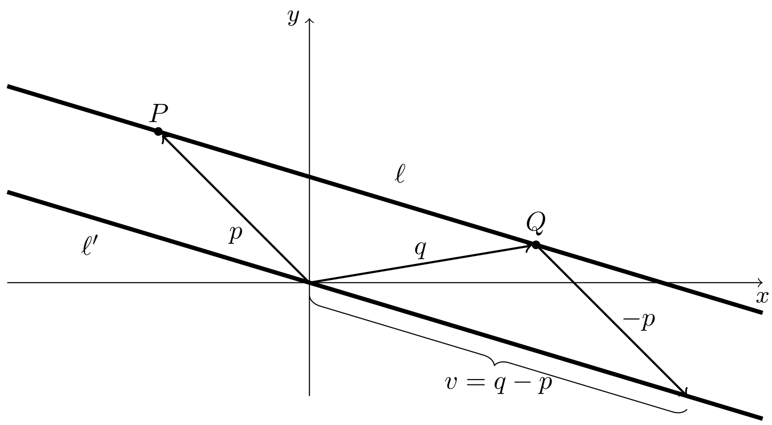


Figure 11: Working toward a parametric line, part 2

as the set heads of the vectors

$$p + t(q - p),$$

where t is a parameter which is allowed to vary over all real numbers. Now, let's sum up what we have learned.

Theorem 7 (parametric line)**.** *Let $P = (a, b)$ and $Q = (c, d)$ be two points in the plane. The line which passes through these two points can be described as the heads of all vectors of the form*

$$p + t(q - p) = \begin{pmatrix} a \\ b \end{pmatrix} + t \begin{pmatrix} c - a \\ d - b \end{pmatrix},$$

where t is a parameter which is allowed to vary over all real numbers.

Note that this theorem actually includes the previous one as a special case. If one of our points happens to be the origin, we simply use $P = O$, which corresponds to the zero vector, and this description collapses back into the one we found earlier. In either case, the vector $v = q - p$ is called a *direction vector* for the line.

The Equation of a Line in the Plane

As a bit of a palate cleanser, let's answer this: Suppose you are given a line described as in the last theorem. How would you sketch it in the plane? The simplest method is to choose two different values of t, use them to generate to points on the line, plot those points, and trace the line through them. Which values of t should you choose? It is often convenient to use $t = 0$ and $t = 1$.

In fact, a parametric description can (perhaps "should") be thought of as describing the motion of a vector or point in the plane. The parameter t should be thought of as *time*. As the time changes, the location of the vector (or point) in the plane changes. This creates a kind of function, which is sometimes written like so:

$$t \mapsto p + t(q - p)$$

The input time t is some real number, and the corresponding output is the vector $p + t(q - p)$ in $\mathbb{R}^2$. For any fixed t, our function gives us the "location of the moving point at time t." You can think of running through all of these different pictures, in sequence, as a movie showing the motion of our vector (or point). Then the line is like a time-lapse photograph which traces out the whole motion.

Note that at time $t = 0$, we have the location $0 \mapsto p + 0(q - p) = p$, and at time $t = 1$, we get the location $1 \mapsto p + 1(q - p) = q$. So it makes sense to think of our parametrized line as describing motion "from p to q." And the whole time, the direction of motion is exactly $v = q - p$: from p to q in the physicist's sense.

Lengths and Angles in the plane

What should we mean by the "length" of a vector u in $\mathbb{R}^2$? The most natural thing is to use the notion of the length of the line segment from the tail to the head — the regular length of the arrow involved. This natural thing is just what we want, but because we wish to generalize it later, it will be best to give it an unusual name: the *norm* of a vector.

How shall we understand it? The key comes from looking upon

Figure 8 with new eyes. There we see a vector in $\mathbb{R}^2$ as the hypotenuse of a right triangle having its legs parallel to the two coordinate axes.

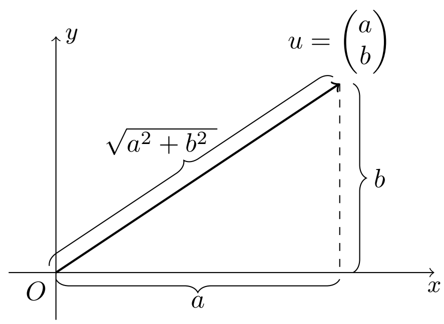


Figure 12: The Norm of a vector

These two legs then have lengths which are equal to our two coordinates, respectively. So we can compute the length of the arrow by the Pythagorean Theorem. Now that we understand what we want, we can make our definition.

Definition 8 (Norm of a vector)**.** *Let* $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ *be a vector in* $\mathbb{R}^2$*. The* norm *of* u *is the number*

$$\|u\| = \sqrt{a^2 + b^2}\ .$$

It is important to remember that the norm is a number, but never a negative number. Also, the only vector which has norm equal to zero is the zero vector. We shall often have the occasion to use norms of vectors as scalars in forming linear combinations.

As long as we have the concepts of norm and scalar in front of us, let us investigate how those interact. If λ is a scalar, then how do the norms a vector u and the rescaled λu compare? Let's write our vector in the form $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$. Then we can do the following straightforward computation:

$$\|\lambda u\| = \left\| \lambda \begin{pmatrix} a \\ b \end{pmatrix} \right\| = \left\| \begin{pmatrix} \lambda a \\ \lambda b \end{pmatrix} \right\| = \sqrt{(\lambda a)^2 + (\lambda b)^2} = |\lambda| \, \|u\|$$

This conforms to our basic ideas about scalar multiplication. If we rescale a vector, we are really just changing its length by that factor. (And then there is the worry over the sign of the scalar, because lengths must not be negative.)

In a terrible linguistic collision, the most useful role for the norm of a vector is in the process of *normalizing* that vector. Beware, mathematics has several words that are overused! Norm, normal, normalize, normalized and other forms of this are a prime example. To *normalize* a vector u, we multiply it by the scalar $\|u\|^{-1}$, and thus produce a new vector $u/\|u\|$ which points in the same direction as u, but has norm equal to 1. This is because we can use our observation just above about how norms interact with scalar multiplication to compute:

$$\left\| \|u\|^{-1} u \right\| = \|u\|^{-1} \|u\| = 1.$$

That last multiplication is just a multiplication of numbers. Note that all of that only makes sense as long as u is not the zero vector. If u is the zero vector, its norm is $\|u\| = 0$. Since it makes no sense to divide by zero, it is not possible to normalize the zero vector.

Just to make sure all of this stays confusing for learners, mathematicians call a vector with norm equal to 1 a *unit vector*.

To sum up: given a non-zero vector u, it is possible to normalize it, which means to replace it with the vector $u/\|u\|$, which is a unit vector.

Now that we are talking about proper geometry with length, can we find a way to measure the angle between two vectors? We can, if we use some trigonometry. We will use the following three facts:

1. Each point on the unit circle can be represented as a point of

the form $(\cos(\alpha), \sin(\alpha))$ for some angle α between 0 and 2π, so the corresponding vector can be written as $p = \binom{\cos(\alpha)}{\sin(\alpha)}$.

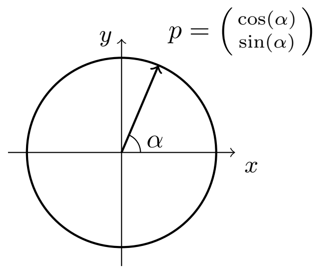


Figure 13: Trigonometry for points on the unit circle

2. The function $\theta \mapsto \cos(\theta)$ associates each angle in the interval $[0, \pi]$ to a unique real number in the interval $[-1, 1]$, and vice versa. For us, this means that it to measure an angle, we can get away with instead finding the cosine of that angle.

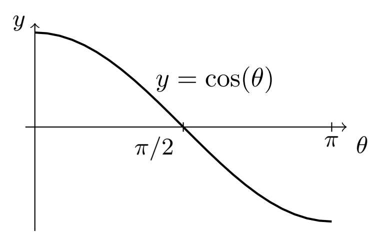


Figure 14: Part of the graph of $y = \cos(\theta)$

3. There is an identity on trigonometric functions that helps us deal with the the cosine of a difference of angles: If α and β are two angles, then

$$\cos(\beta - \alpha) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta).$$

So, let us consider the angle between two vectors, u and v, in $\mathbb{R}^2$. First, the angle between these vectors does not depend at all on their lengths. If we change one, or both, of the vectors by rescaling them, that will not change the directions involved, and hence will not change the angle we seek. We can get to a more uniform set up for our task by normalizing u and v. Thus we will instead consider the unit vectors $u/\|u\|$ and $v/\|v\|$, and look for the angle θ between them.

Since our vectors are unit vectors, we can represent them with trigonometric functions:

$$\frac{u}{\|u\|} = \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}, \qquad \frac{v}{\|v\|} = \begin{pmatrix} \cos(\beta) \\ \sin(\beta) \end{pmatrix}.$$

Let's suppose that the angle β that v makes with the x-axis is larger than the angle α that u makes with the x-axis. Then we want to find the angle $\theta = \beta - \alpha$.

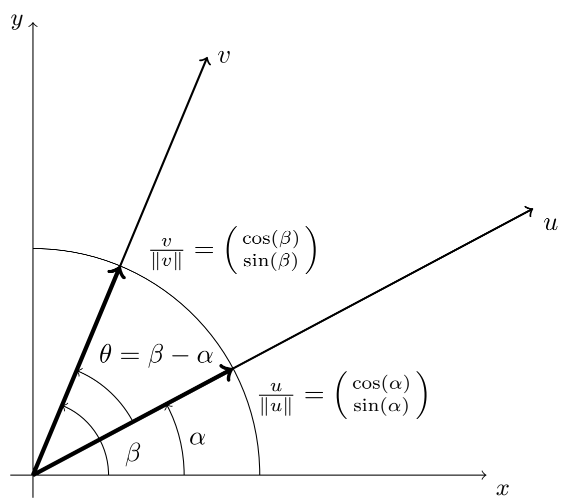


Figure 15: Finding the angle between two vectors.

Since cosine is such a nice function on $[0, \pi]$, it will do to first try to find $\cos(\theta)$. But then we can simply apply the angle difference formula for cosine to see

$$\cos(\theta) = \cos(\beta - \alpha) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta).$$

This is a great situation. To find $\cos(\theta)$, we can just form a funny combination of the components of the vectors $u/\|u\|$ and $v/\|v\|$.

That funny combination takes the form "product of first components plus product of second components." That is familiar! It looks a bit like what happens when we compute a norm, where we have a "square of the first component plus the square of the second component." Motivated by this happy coincidence, we make up a new object to help us organize the similarities:

Definition 9 (Dot Product)**.** *Let $u = \left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right)$ and $v = \left(\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right)$ be vectors in $\mathbb{R}^2$. The* dot product *of u and v is the number*

$$u \cdot v = u_1 v_1 + u_2 v_2.$$

The dot product is mysterious for new learners. You take this funny combination of the components, and somehow learn all kinds of information about lengths and angles. The dot product has many useful properties, which we gather together now.

Theorem 10. *Let u, v, and w be vectors in $\mathbb{R}^2$ and let λ and μ be scalars. Then*

- *The angle θ between u and v is given by*

$$\theta = \arccos\left(\frac{u \cdot v}{\|u\| \, \|v\|}\right);$$

- *The norm of u is $\|u\| = \sqrt{u \cdot u}$;*
- *The dot product is symmetric: $u \cdot v = v \cdot u$;*
- *The dot product distributes over linear combinations*

$$u \cdot (\lambda v + \mu w) = \lambda(u \cdot v) + \mu(u \cdot w)$$

As our study progresses, we will not have much use for measuring particular angles, but it will be very important to us to understand situations when two vectors make an angle of $\pi/2$ radians (i.e. $90°$, a right angle). It is common to call such vectors *perpendicular*. It is a fact that $\cos(\pi/2) = 0$, so that non-zero vectors u and v make an angle of $\pi/2$ when

$$0 = \cos(\pi/2) = \frac{u \cdot v}{\|u\| \, \|v\|}.$$

By clearing out the denominator of this fraction, we see that u and v make an angle of $\pi/2$ exactly when $u \cdot v = 0$. We will generalize this idea eventually, but it is convenient to start using the proper word now.

Definition 11 (Orthogonal vectors)**.** *We say that a pair of non-zero vectors u and v are* orthogonal *exactly when $u \cdot v = 0$.*

We have to be a little careful and include the caveat that u and v are nonzero, because of course the dot product of any vector with the zero vector is zero (the number, this time), but it does not feel right to say that the zero vector makes an angle of any type with any other vector.

Normal Vectors and Equations for Lines

Our discussion of the basic geometry of vectors feels like a digression. It has led us away from our main goal, which is to describe lines in the plane. But now we have enough information to circle back[1] and use the dot product and the concept of orthogonality to give a different way of describing lines. As before, we will start with the case of lines through the origin.

Suppose that we are given a vector $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$. What vectors are orthogonal to u? When we try to just sketch this, we get a line through the origin! It looks a lot like Figure 16.

Let's try to be more precise. Suppose we were lucky enough to figure out one single vector v_0 which is orthogonal to u. Then any

[1] This is a geometry joke. I am sorry. But not really.

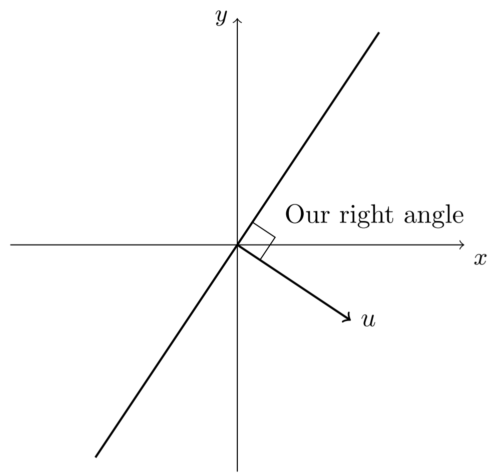


Figure 16: The line of vectors orthogonal to u.

scalar multiple tv_0 of v_0 will also be orthogonal to u, because

$$(tv_0) \cdot u = t(v_0 \cdot u) = t(0) = 0.$$

But the collection of vectors of the form tv_0 is the parametric line through the origin. So, the line we seek is the line through the origin with direction v_0.

How can we find such a v_0? Well, in $\mathbb{R}^2$, we can write $v_0 = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ and then use the orthogonality condition to see that x and y must satisfy

$$0 = u \cdot v_0 = \begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = ax + by.$$

After a little poking around, one can discover that one solution to the equation $ax + by = 0$ is $x = -b, y = a$. So we can choose $v_0 = \left(\begin{smallmatrix} -b \\ a \end{smallmatrix}\right)$. Of course, there are other solutions to this equation (a whole line's worth!), but we just need one and this one is convenient.

If we collect this sequence of observations, we get the following theorem.

Theorem 12 (Equation for a line through the origin)**.** *Let $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ be a non-zero vector. Then the set of vectors orthogonal to u is the line with parametric description*

$$t \mapsto t \begin{pmatrix} -b \\ a \end{pmatrix}.$$

Moreover, we can describe that line by an equation: The line of vectors orthogonal to u consists of the vectors $X = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ whose coordinates satisfy the equation

$$ax + by = 0.$$

On the other hand, for each line ℓ through the origin with direction vector $v = \left(\begin{smallmatrix} c \\ d \end{smallmatrix}\right)$, the vector $u = \left(\begin{smallmatrix} -d \\ c \end{smallmatrix}\right)$ is orthogonal to all of the vectors in the line ℓ. Furthermore, the line ℓ can be represented as the collection of vectors $X = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ whose coordinates satisfy the equation

$$-dx + cy = 0.$$

Definition 13 (Normal Vector to a line)**.** *A vector which is orthogonal to all of the direction vectors of a line is called a* normal vector *for that line.*

Note that there is a little looseness here in the descriptions. A line has many direction vectors, and also many normal vectors. But there is only so much leeway: all of the direction vectors are scalar multiples of each other, and all of the normal vectors are scalar multiples of each other.

And here we have introduced yet another use for some form of the word "normal." Be aware.

Now that we can write equations to help us describe lines through the origin, we would like to have a similar understanding for lines which do not pass through the origin. First, take any line ℓ. Then ℓ is parallel to a line ℓ' which passes through the origin. Using what we know above, pick out a normal vector $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ to the line ℓ'. Since ℓ and ℓ' are parallel, u is a normal vector for both. Then, pick out just one vector p which has its head on the line ℓ. (So the corresponding point P lies on ℓ.)

If we let some unknown, generic vector with its head on the line ℓ be represented by the variable vector $X = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$, we see that the physicist's $X - p$ lies entirely in the line ℓ. Hence the mathematician's $X - p$ is a direction vector for ℓ, and therefore orthogonal to the vector u.

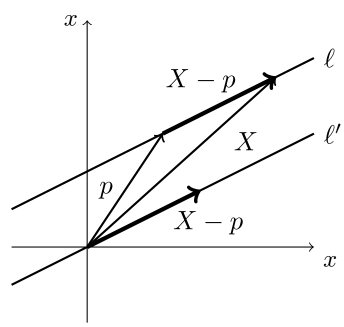


Figure 17: Sorting out the equation of a generic line.

Using the condition for orthogonality, we can derive an equation that the components x and y must satisfy in the following manner. First,

$$0 = u \cdot (X - p) = u \cdot X - u \cdot p;$$

therefore,

$$u \cdot X = u \cdot p.$$

But $c = u \cdot p$ is just a number. And we have already learned how to re-express $u \cdot X$ as $ax + by$. Rewriting each side of the last displayed equation, we obtain

$$ax + by = c.$$

That is the form of an equation for a generic line in the plane. This way of writing the equation of a line in the plane is called the *standard form* of the equation of a line. There are some important features to track in this form:

- The coefficients a and b of the variables are just the components of the normal vector, but this normal vector is chosen just by looking at a line through the origin. There are a great many lines which are all parallel to the same line through the origin, so they will all have the same set of coefficients.
- Changing which line in that family of parallel lines will involve changing the point P and the corresponding vector p. So even though a and b will not change, the number $c = u \cdot p$ will be different.
- In each family of parallel lines, the one with right hand side $c = 0$ is the one through the origin.
- Changing the coefficents a and b results in a different normal vector, and hence a different family of parallel lines.

Theorem 14 (Equation of a generic line in the plane)**.** *Suppose that ℓ is a line in the plane. Let P be a point on the line, with corresponding vector p, and let $u = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ be a normal vector for ℓ. Set $c = u \cdot p$.*

Then ℓ is the collection of points at the heads of the vectors $X = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ which satisfy the equation

$$ax + by = c.$$

Working with Lines Described by Equations

To make a graph of a line described by an equation, it is easiest to choose two points P and Q which lie on the line, and then draw the line passing through them. In most cases, the most convenient points are the *intercepts* where the line meets the coordinate axes. If the line is $ax + by = c$, then we can find the x-intercept by setting $y = 0$ and solving for $x = c/a$. Hence the point of interest is $P = (c/a, 0)$. Similarly, the y-intercept can be found by setting $x = 0$. The resulting point will be $Q = (0, c/b)$. Then it is no trouble to plot these two points and draw a line through them.

Now, suppose we have a line described parametrically. We can find the equation of that line by a process known as "eliminating the parameter." If the parametric form is

$$t \mapsto \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} + t \begin{pmatrix} v_0 \\ v_1 \end{pmatrix},$$

we take apart the two coordinates of the vector to form a pair of parametric equations

$$\left\{ \begin{array}{ccccc} x & = & p_0 & + & tv_0 \\ y & = & p_1 & + & tv_1 \end{array} \right.$$

Now, we multiply through the first equation by $-v_1$ and the second equation by v_0

$$\left\{ \begin{array}{ccccc} -v_1 x & = & -v_1 p_0 & - & tv_0 v_1 \\ v_0 y & = & v_0 p_1 & + & tv_0 v_1 \end{array} \right. ,$$

and then we add the two equations, thus eliminating the appearance of the parameter t, to obtain

$$-v_1 x + v_0 y = -v_1 p_0 + v_0 p_1.$$

This is the equation of the line.

Finally, suppose we have an equation of the line, and we wish to find a parametric representation for it. There is a simple process for doing so, with a neat little trick in the heart of it.

Let our equation be $ax+by=c$, and assume for simplicity that $a \neq 0$. Then, with a little bit of manipulation, we may rewrite the equation as

$$x = \frac{c}{a} - y\frac{b}{a}.$$

The trick is to treat the variable y as if it was our parameter. That is, just declare $y = t$. That gives us a pair of equations, which we can write as

$$\left\{ \begin{array}{ccccc} x & = & \frac{c}{a} & - & t\frac{b}{a} \\ y & = & & & t \end{array} \right. .$$

Now, written this way, it is not too hard to translate this into a vector parametric form for the line.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c/a \\ 0 \end{pmatrix} + t \begin{pmatrix} -b/a \\ 1 \end{pmatrix},$$

or

$$t \mapsto \begin{pmatrix} c/a \\ 0 \end{pmatrix} + t \begin{pmatrix} -b/a \\ 1 \end{pmatrix}.$$

If you like, you can rescale that direction vector, and instead write

$$t \mapsto \begin{pmatrix} c/a \\ 0 \end{pmatrix} + t \begin{pmatrix} -b \\ a \end{pmatrix}.$$

Later it will be important to us that the direction vector $\left(\begin{smallmatrix} -b \\ a \end{smallmatrix}\right)$ is a solution to the associated equation $ax + by = 0$.

The Equations of a Line in Space

Let us now think about the nature of space. Okay, that sounded goofy. I don't mean contemplating planets and nebulae and the enormity of *UY Scuti*. So far, all of our work has been in the plane, which is supposed to represent the ideal of a flat thing which is infinite in extent and "two-dimensional." Our next theater for exploration is a model for the space in which we live. Rather than having two dimensions, it has three, so we can talk about motions which are not just forward and back, and right and left, but also up and down.

Fortunately, much of the hard work we have done understanding vectors in $\mathbb{R}^2$ carries over quickly. There are ready analogies for point, vector, and dot product, and though the geometry of these gets more interesting, the algebra really doesn't. That will eventually be our lifeline: as the spaces we consider get more exotic, the geometries will get more complicated (and eventually almost impossible to visualize), but the corresponding algebra will always have the same level of complexity. That is why the subject is called "Linear Algebra" rather than "Linear Geometry."

Drawing pictures becomes much more challenging. One option is to study drawing and painting for a few years until representing figures from space on a plane becomes more natural. In fact, there is a strong connection between linear algebra and some of the science behind perspective drawing, which mathematicians call *projective geometry*. But as that is outside the plan of our current study, we will resort to images made with software.

Again, it will take some work to get there, but we are interested in the following question:

How can we clearly describe a single line in space with an equation?

Points, Vectors, and the Dot Product in $\mathbb{R}^3$

Just as we worked in the plane with points and vectors (both physicist's and mathematician's vectors) by using coordinates, we can do the same for the space around us. We have the same sort of muddled trichotomy to deal with, which should feel a little less mysterious by now. Anyway, a *point* in space is supposed to be a location; a *(physcist's) vector* is an arrow, having direction and magnitude; and a *(mathematician's) vector* is an arrow whose tail is at a special, pre-selected point called the *origin* and denoted by O.

There are many ways to attempt to set up coordinates in space, but experience has taught us that a good way mimics the set-up from the plane. There, in the plane, we used a pair of perpendicular lines through the origin as our backdrop, which we call *coordinate axes*. In space, we have more room to deal with, so instead we use three planes as reference, and we try to make them as perpendicular as we can.

How do we arrange that? The key fact we need is that when distinct planes meet, they do so along a line. Let's think through

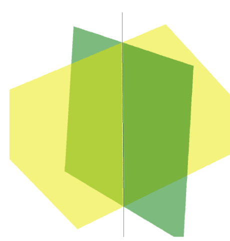


Figure 18: Two planes intersect along a line.

that a little. What does the picture look like near the intersection where two planes meet? A good way to understand it is to look "across the intersection." To this end, turn the picture in front of you until the line of intersection is exactly the “line of sight” pointing directly away from your eye. If you do this correctly, the line should essentially vanish, by collapsing down to a single point in your vision. If you prefer, rather than imagining that you turn the picture, you can imagine moving around in space until you are in the line, and then looking in the correct direction. Either way, you should get the same image: the line of intersection collapses to a point, and the two planes under consideration collapse to lines in your vision, as you are looking at things “edge on.”

This picture is now much simpler. We have turned the set-up of two planes which intersect along a line into a picture where two lines meet at a point. What we want is that these two lines should be perpendicular. In fact, they should mimic our usual coordinate axes picture for the plane.

So, now we are ready to set up our *coordinate planes and axes* for space. First, choose a point O to play the role of the origin. Then, through that point, choose three distinct planes $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$. Each pair of these planes meets along a line: call the intersection of plane $\mathcal{P}_i$ with plane $\mathcal{P}_j$ the line ℓ_{ij}. So, for example, ℓ_{12} is the line along which plane $\mathcal{P}_1$ meets plane $\mathcal{P}_2$.

The important condition we want is this: If you look down the line ℓ_{ij}, then the planes $\mathcal{P}_i$ and p_j appear to be a pair of perpendicular lines. We require this to be true for each of the three lines ℓ_{ij} simultaneously.

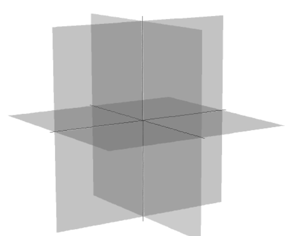


Figure 19: The Coordinate Planes and Axes

Now we are prepared to give some of these objects their traditional names. The planes $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ are called the yz-plane, the xz-plane, and the xy-plane, respectively. The line ℓ_{12} is called the z-axis, the line ℓ_{23} is called the x axis, and the line ℓ_{13} is called the y-axis.

Sometimes, instead of x, y, and z, we may use the variables x_1, x_2, and x_3. Then all of the names above change accordingly.

Just as in the plane, we can now introduce coordinates for points and vectors. We will just discuss how to do this for points, and you can see how to extend it to both interpretations of vectors in the same manner as before.

Given a point P in space, draw a plane parallel to the yz-plane which passes through P. This plane will meet the x-axis in exactly one point. We treat the x-axis as a number line, so this point has associated to it a number P_x, which we call the *x-coordinate of P*. Geometrically, the number P_x represents the "height" of P above the yz-plane. The words height and above are used in a relative sense, here.

Similarly, we can draw a plane parallel to the xz-plane which passes through P. This plane will intersect the number line of the y-axis, and we call the corresponding number P_y the *y-coordinate of P*. Finally, we can draw a plane parallel to the xy-plane through P, which will intersect the z-axis in a point, whose corresponding number P_z is called the *z-coordinate of P*. We then take these three numbers and bundle them together and represent the point P by the ordered triple $P = (P_x, P_y, P_z)$. Just as before, we will distinguish between points and vectors in our notation, and write vectors as vertical stacks of numbers. Here, such a stack has three parts. Motivated by our work in the plane, we make the following definition.

Definition 15. *A 3-vector is a vertical stack of three real numbers, like so:*

$$v = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}.$$

The collection of all such 3-vectors is called 3-space*, and written* *with this notation:* $\mathbb{R}^3$*.*

The notation $\mathbb{R}^3$ is often read as "arr-three," and many people use that language in place of "three-space." Also, most of the time we will just say "vector" when the context is clear. But it can be helpful to have the extra bit of information in situations where vectors of different sizes are involved.

The algebra of 3-vectors is handled in exactly the same way as the algebra of 2-vectors, except that there is another coordinate to track. All of the geometric connections between points, mathematician's vectors, and physicist's vectors work just as before, but there is "more space" to work with. Most people find visualization in three-space to be much harder than in the plane. This is completely normal, and you should expect to get used to this only if you work at it.

Definition 16 (Algebra in $\mathbb{R}^3$)**.** *Suppose that* u *and* v *are* 3-*vectors, and that* λ *and* μ *are scalars (real numbers). Then the* linear combination of u and v with weights λ and μ *is defined coordinate-by-coordinate:*

$$\lambda u + \mu v = \lambda \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} + \mu \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} \lambda u_x + \mu v_x \\ \lambda u_y + \mu v_y \\ \lambda u_z + \mu v_z \end{pmatrix}$$

It is, of course, possible to form linear combinations of larger collections of vectors, using the corresponding number of weights. You should also notice that addition of 3-vectors and scalar multiplication for 3-vectors are special cases of linear combination. (How would you do this? Check this statement!) The statements of Theorem 3 and Theorem 5 carry over unchanged to our new setting. The algebra of vectors behaves just as sensibly in $\mathbb{R}^3$ as in $\mathbb{R}^2$, even if the pictures are more challenging.

The geometry of $\mathbb{R}^3$ can be handled similarly, too, with dot products, norms, and angles.

Definition 17 (Dot Product, norm, and angles in $\mathbb{R}^3$)**.** *Let* u *and*

v be 3-vectors. The dot product *of u and v is the real number*

$$u \cdot v = \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} \cdot \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = u_x v_x + u_y v_y + u_z v_z,$$

the norm *of u is the non-negative real number*

$$\|u\| = \sqrt{u \cdot u} = \sqrt{u_x^2 + u_y^2 + u_z^2},$$

and the angle between u and v is defined to be

$$\theta = \arccos\left(\frac{u \cdot v}{\|u\| \, \|v\|}\right).$$

We say two vectors are orthogonal *when their dot product is* 0*. A vector is called a* unit vector *when its norm is equal to* 1*.*

Again, these behave just as well in $\mathbb{R}^3$ as in $\mathbb{R}^2$. It is possible to "normalize" a non-zero vector u to produce a unit vector $\|u\|^{-1} u$, and the statement of Theorem 10 works in $\mathbb{R}^3$.

Parametric lines

Given two points, P and Q, in $\mathbb{R}^3$, we can describe the line which passes through those points. This works just as in the case of the plane, using similar triangles for lines through the origin and then picking up those lines and moving them for the general case.

Theorem 18. *Let P, Q, and R be points in* 3*-space. Let p, q, and r be the* 3*-vectors with tails at the origin and heads at P, Q, and R, respectively.*

- *The line which passes through the origin O and the point R is described parametrically as the heads of all the vectors traced by the function*

$$t \mapsto tr.$$

– *The line which passes through the points P and Q is described parametrically as the heads of all of the vectors traced by the fuction*

$$t \mapsto p + t(q - p).$$

Note that the second statement collapses into the first if we choose $P = O$, since then p is the zero vector. In the second statement, things are arranged so that $t = 0$ gives the vector p and $t-1$ gives the vector q. We think of P as the "base point" and the motion shows travel in the direction of the direction vector $q - p$. Also, if we are just interested in describing the line, it does not matter which of the points P or Q we consider as base point: switching them just changes the direction vector to $p-q = -(q-p)$, but does not the line as a whole.

Planes through the Origin in $\mathbb{R}^3$ and their Normal Vectors

In the last chapter, we made productive use of the following question:

> What is the set of vectors in $\mathbb{R}^2$ which are orthogonal to a given vector u?

Our considerations led us an understanding of lines in the plane as objects described as solution sets to a linear equation. With hearts full of hope, we charge off in the same direction here.[2] Perhaps we can find some way to describe lines with equations.

The Set of Vectors Orthogonal to a Given Vector
So, we fix some vector $u = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, and try to figure out the form of those vectors which are orthogonal to u. Suppose that $v = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ is some variable vector. We will assume that v is orthogonal to u.

[2]This is a geometry joke. I am sorry. But not really.

This means that

$$0 = u \cdot v = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = ax + by + cz.$$

Thus the set of vectors in $\mathbb{R}^3$ which are orthogonal to u is exactly the collection of vectors whose coordinates satisfy the linear equation $ax + by + cz = 0$.

What is that shape, though? We must try to say something about it. To make our discussion easier, let us give it a name. We will call it $\mathcal{P}$. In standard mathematical notation, we write

$$\mathcal{P} = \left\{ v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \middle| \; ax + by + cz = 0 \right\}.$$

This notation is the way that a mathematician writes a "set," a collection of things. The curly braces tell you that you are using *set notation*, the stuff before the vertical bar sets up our notation, and the stuff after the vertical bar tells us what conditions describe our set. So, in this case, you should read that displayed equation as

> "$\mathcal{P}$ is the set of all 3-vectors v with components x, y, and z such that the equation $ax+by+cz = 0$ is true."

Our first observation is that this shape contains all the linear combinations of the vectors it has in it. To see this, suppose that v_1 and v_2 are two vectors which are orthogonal to u, and that λ_1 and λ_2 are two scalars, and consider the linear combination $\lambda_1 v_1 + \lambda_2 v_2$. The dot product of this vector with u is

$$u \cdot (\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 u \cdot v_1 + \lambda_2 u \cdot v_2 = \lambda_1(0) + \lambda_2(0) = 0.$$

Hence, the linear combination $\lambda_1 v_1 + \lambda_2 v_2$ is also orthogonal to u.

As a special case, the collection of scalar multiples of a single vector is a kind of linear combination, λv. But the set of multiples of a single vector is a line through the origin in the direction of that vector. This means that our set $\mathcal{P}$ has to contain all of the lines through pairs O, P, where P is a point in $\mathcal{P}$.

Let's sum up some of what we know, so we don't lose track.

Theorem 19. *Fix a 3-vector u. The set $\mathcal{P}$ of all vectors which are orthogonal to u has the following properties:*

- *If $u = \left(\begin{smallmatrix} a \\ b \\ c \end{smallmatrix}\right)$, then $\mathcal{P}$ consists of those vectors $v = \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right)$ whose coordinates satisfy the equation $ax + by + cz = 0$.*
- *For each vector p in $\mathcal{P}$, the entire line $t \mapsto tp$ lies in $\mathcal{P}$.*
- *If p and q are vectors in $\mathcal{P}$, then any linear combination of p and q also lies in $\mathcal{P}$.*

This is good, but it does not quite tell us what $\mathcal{P}$ is, yet. Maybe it is time to just find some examples of vectors which satisfy the defining equation. To make things smoother below, we assume that $a \neq 0$. This is not essential, but it will help a little to avoid dividing by 0.

We begin by looking for simple forms. Let's just specify that one of the variables is 0 and see what we get. For example, if we assume $x = 0$, the equation collapses to

$$0 = ax + by + cz = a(0) + by + cz = by + cz.$$

This can be satisfied by choosing $y = -c$ and $z = b$. So we get the vector

$$v_1 = \begin{pmatrix} 0 \\ -c \\ b \end{pmatrix}.$$

Similarly, we can search for a vector which has $y = 0$. The defining equation for $\mathcal{P}$ collapses to $ax + cz = 0$, which we can solve with $x = -c$ and $z = a$. This gives us the vector

$$v_2 = \begin{pmatrix} -c \\ 0 \\ a \end{pmatrix}.$$

Since we have searched for, and found, vectors by starting with $x = 0$ and $y = 0$, respectively, it makes sense to search for one with $z = 0$, too. So let's do that. If $z = 0$, the defining equation collapses to $ax + by = 0$, which has solution $x = -b$, $y = a$. So we obtain the vector

$$v_3 = \begin{pmatrix} -b \\ a \\ 0 \end{pmatrix}.$$

Given Theorem 19, we know that $\mathcal{P}$ contains at least the collection of all linear combinations of the three vectors v_1, v_2 and v_3. But note that we do not truly require all three of those vectors! This is because if we pick scalars carefully and form a linear combination of v_2 and v_3, we obtain

$$bv_2 - cv_3 + = b \begin{pmatrix} -c \\ 0 \\ a \end{pmatrix} - c \begin{pmatrix} -b \\ a \\ 0 \end{pmatrix} = \begin{pmatrix} -bc + bc \\ -ac \\ ab \end{pmatrix} = a \begin{pmatrix} 0 \\ -c \\ b \end{pmatrix} = av_1.$$

This means that $v_1 = \frac{b}{a} v_2 - \frac{c}{a} v_3$ is already a linear combination of v_1 and v_2. So we do not need to keep v_3. It is redundant in our current description. Maybe, just maybe, we are lucky that any vector in $\mathcal{P}$ is a linear combination of v_2 and v_3?

Let's try to approach this from a different direction.[3] We can rewrite the defining equation for $\mathcal{P}$ in the form

$$x = -\frac{b}{a} y - \frac{c}{a} z.$$

This means that if we know the values of y and z, then the value of x is completely determined. So, really, we only have two independent choices to make. If we choose $y = s$ and $z = t$ to be two numbers, then we have $x = \frac{-b}{a} s + \frac{-c}{a} t$. Which means that our vector in $\mathcal{P}$ has the form

$$\begin{pmatrix} \frac{-b}{a} s + \frac{-c}{a} t \\ s \\ t \end{pmatrix} = s \begin{pmatrix} \frac{-b}{a} \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} \frac{-c}{a} \\ 0 \\ 1 \end{pmatrix} = \frac{s}{a} v_2 + \frac{t}{a} v_3.$$

[3] That is a geometry joke. I am sorry. But not really.

This is the big result we want, though the shape is not, yet, clear. What we have found is that any vector which has coordinates which satisfy the equation for $\mathcal{P}$ must be a linear combination of v_2 and v_3.

What sort of shape is made up of the linear combinations of two vectors which point in different directions? Well, if we are allowed to move in the direction of one vector, then we generate a line through the origin. Now, from each of the points on this line, we are allowed to move in a different (fixed) direction. That makes a plane! This is the answer we sought: the set $\mathcal{P}$ is a plane through the origin, the one which contains the vectors v_2 and v_3.

By the way, we call u the *normal vector* to the plane $\mathcal{P}$.

Again, we should make a summary.

Theorem 20. *Fix a vector u in 3-space. The set $\mathcal{P}$ of all vectors orthogonal to u is a plane through the origin.*

Moreover, if we write $u = \left(\begin{smallmatrix} a \\ b \\ c \end{smallmatrix}\right)$, and assume $a \neq 0$, then

- *We can describe the plane $\mathcal{P}$ as the collection of all points $P = (x, y, z)$ which satisfy the equation*

$$ax + by + cz = 0.$$

- *We can also describe the plane parametrically as the heads of all of the vectors of the form*

$$s \begin{pmatrix} -b/a \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -c/a \\ 0 \\ 1 \end{pmatrix}.$$

Like before, we view this parametric description as a defining a kind of function.

$$(s, t) \mapsto s \begin{pmatrix} -b/a \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -c/a \\ 0 \\ 1 \end{pmatrix}.$$

Only this one has *two* inputs, s and t. The output for any particular choice of s, t is some vector in 3-space. If we make some sort of "multidimensional time lapse photograph," then the heads of all of the vectors we get trace out a plane in 3-space.

So, we set out to see if we could describe a line in 3-space, but we ended up working on describing planes instead. Let's try to finish that up, and we will return to describing lines in a bit.

The Plane Through the Origin and Two Given Points

In Theorem 20, we start with a vector u, and then find two ways to understand the plane $\mathcal{P}$ through the origin which consists of vectors orthogonal to u. Of the two ways, the equation came first, then the parametric description. And it was the parametric description that helped us realize the object was a plane. Now, let us try to do things in the opposite order.

Suppose that we have a plane $\mathcal{Q}$ in 3-space which passes through the origin. How can we determine it completely? First note that to determine a line we need two points, but to determine a plane we need three points. Since we already have the origin singled out, consider two more points P and Q. It is important that these three points are distinct, or we will lose the idea of the plane.

First, $\mathcal{Q}$ must contain the whole line through P and Q. Thus it must contain all of the points on the line with parametric description

$$t \mapsto p + t(q - p),$$

where p is the vector from O to P, and q is the vector from O to Q. Now, if we fix a particular $t = t_0$, we get a point X_0 which lies in $\mathcal{Q}$. This point X_0 is the head of the vector

$$x_0 = p + t_0(q - p) = (1 - t_0)p + t_0 q.$$

Since the plane $\mathcal{Q}$ contains O and X_0, it must contain the line through them, which is described parametrically by

$$s \mapsto s x_0 = s(1 - t_0)p + s t_0 q.$$

Notice that this time we use the variable s as parameter, because we have already used t above and we want to distinguish between the two choices. This is very nice, we have described lots of points on $\mathcal{Q}$ as the heads of vectors which are just linear combinations of p and q. For varying choices of s and t, we get may points on $\mathcal{Q}$.

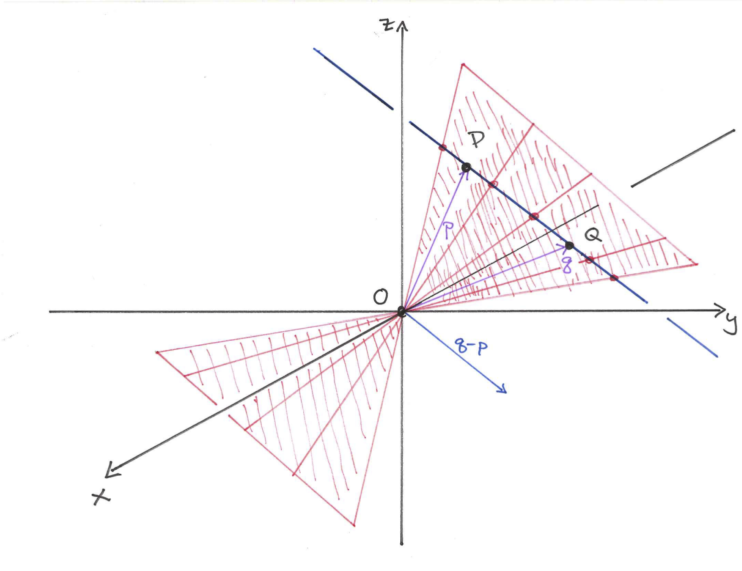


Figure 20: The line PQ and the sweep of lines through O and points X on this line.

A careful look at this situation shows that we have described almost all of the points on $\mathcal{Q}$. We have missed those on a single line: the line passing through the origin O which is parallel to the line through P and Q. That is, our exceptions are the multiples of $q-p$.

Another way to see this is to look at the two scalars involved in our description, $s(1-t)$ and st. There ratio has the form

$$\frac{st}{s(1-t)} = \frac{t}{1-t}.$$

But it is not too hard to plot the function $t \mapsto t/(1-t)$ and see that it misses the value -1, and only the value -1.

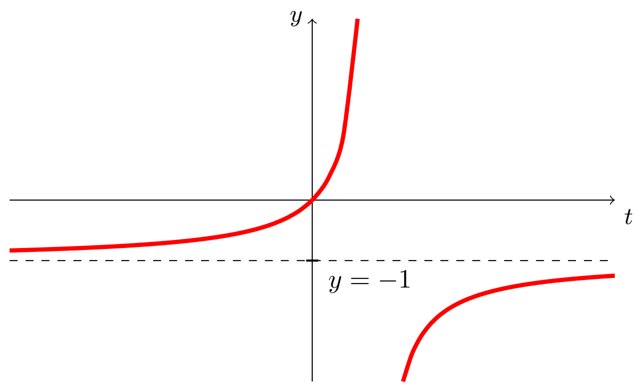


Figure 21: The graph of $y = t/(1-t)$.

This corresponds to missing vectors of the form $v = \lambda p - \lambda q = \lambda(q-p)$, which is exactly the line through the origin in the direction $q - p$.

Anyway, we can patch the plane back together, and add this line back in, by simply allowing *any* linear combination of the vectors p and q.

Theorem 21. *Let P and Q be two points in 3-space. If p and q are the vectors with tails at the origin O and heads at P and Q, respectively, then the plane through the three points O, P, and Q is*

$$\mathcal{Q} = \{v = sp + tq | s \text{ and } t \text{ are real numbers}\}.$$

Now that we have the fact that we can describe a plane through the origin parametrically as the collection of vectors which are linear combinations of two fixed vectors p and q, we want to find a way to turn that information into the equation which describes that plane. There are two readily available methods: (1) we can write out some equations and eliminate the parameters s and t; or, (2) we can find a normal vector to the plane and use that to write down the equation immediately. These two methods should end up in approximately the same place.

From a Parametric Plane to the Equation: Eliminating Parameters

First, let us try the method of eliminating parameters. We will take a variable vector

$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

which we assume to lie in our plane $\mathcal{Q}$. Our goal is to find an equation satisfied by x, y, and z. In order to make progress, we will need to have more detailed notation for p and q, so we write them out in coordinates as

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \quad \text{and} \quad q = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}.$$

Since v is in $\mathcal{Q}$, we can write

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = s \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + t \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}.$$

We then unbundle this vector equation into three separate equations, one for each coordinate.[4]

$$\begin{cases} x = p_1 s + q_1 t \\ y = p_2 s + q_2 t \\ z = p_3 s + q_3 t \end{cases}$$

Our first step is to eliminate s using the first two equations. We do this by multiplying the first equation by p_2 and the second by

[4] We are about to do a computation involving 11 different variables. This looks intimidating, but it isn't so bad. Just stay calm and do it. Choose a particular example and follow along by doing the steps for yourself.

$-p_1$ and then adding. We obtain

$$p_2x - p_1y = t(p_2q_1 - p_1q_2).$$

Similarly, we will eliminate s by using the second and third equations. We do this by multiplying the second equation by p_3 and the third by $-p_2$ and then adding. We obtain

$$p_3y - p_2z = t(p_3q_2 - p_2q_3).$$

This gives us a pair of equations relating x, y, and z to t. We have effectively eliminated s. So, now consider our new system of equations:

$$\left\{ \begin{array}{ccccc} p_2x & - & p_1y & = & t(p_2q_1 - p_1q_2) \\ p_3y & - & p_2z & = & t(p_3q_2 - p_2q_3) \end{array} \right.$$

It remains to eliminate t from these. The clearest way to do that is to multiply the first equation by $p_3q_2 - p_2q_3$ and the second equation by $-(p_2q_1 - p_1q_2)$ and then add. We do this to find

$$(p_3q_2 - p_2q_3)(p_2x - p_1y) - (p_2q_1 - p_1q_2)(p_3y - p_2z) = 0.$$

This is our equation! All of the s's and t's are gone, and we are done. Now, in practice, it is better to have things organized so that like terms are together. If we do a little bit of manipulation and cleaning up, we eventually get this:

$$\{p_2(p_3q_2 - p_2q_3)\}\, x + \{p_2(p_1q_3 - p_3q_1)\}\, y + \{p_2(p_2q_1 - p_1q_2)\}\, z = 0.$$

Note that there is a p_2 multiplied through the whole thing, so we should factor it out and get rid of it. That leaves us with

$$(p_3q_2 - p_2q_3)x + (p_1q_3 - p_3q_1)y + (p_2q_1 - p_1q_2)z = 0.$$

This is the equation of our plane $\mathcal{Q}$, written in a relatively convenient form.

Because it will be useful for comparison later, take note that we can read off of this expression that a normal vector to $\mathcal{Q}$ is[5]

$$n = \begin{pmatrix} p_3q_2 - p_2q_3 \\ p_1q_3 - p_3q_1 \\ p_2q_1 - p_1q_2 \end{pmatrix}.$$

From a Parametric Plane to the Equation: the Normal Vector

Now let us try the other method. We know that $\mathcal{Q}$ is the collection of linear combinations of p and q, and the normal vector to $\mathcal{Q}$ must be orthogonal to every one of those vectors. In particular, the normal vector has to be orthogonal to p and to q. We keep the same notation for components of p and q as above, and introduce notation for n as follows:[6]

$$n = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}.$$

The fact that n is orthogonal to both p and q leads us to this pair of equations

$$\left\{ \begin{array}{ccccccc} p_1n_1 & + & p_2n_2 & + & p_3n_3 & = & 0 \\ q_1n_1 & + & q_2n_2 & + & q_3n_3 & = & 0 \end{array} \right. .$$

We have to do a similar sort of elimination process here to figure out what the n_i's are in terms of the other variables. We are going to do this in a slightly different fashion than the above. It will feel less symmetrical than what we have done before, but it will be more efficient.

Our first step is to multiply the first equation through by $-q_1/p_1$ and then add it to the second equation. Then we'll take a look at

[5] This is almost, but not quite, the cross product of p and q. We are off by a sign.

[6] We are about to do a computation involving 9 different variables. This looks intimidating, but it isn't so bad. Just stay calm and do it. Choose a particular example and follow along by doing the steps for yourself.

our updated system.

$$\begin{cases} p_1 n_1 & + & p_2 n_2 & + & p_3 n_3 & = & 0 \\ & & \left(q_2 - \frac{p_2 q_1}{p_1}\right) n_2 & + & \left(q_3 - \frac{p_3 q_1}{p_1}\right) n_3 & = & 0 \end{cases}.$$

Next, we will multiply the second equation by the number

$$\frac{-p_2}{q_2 - \frac{p_2 q_1}{p_1}} = \frac{-p_1 p_2}{p_1 q_2 - p_2 q_1}$$

and add the result back into the first equation. After a lots of work with fractions, we see that this updates our system to read

$$\begin{cases} p_1 n_1 & & & + & \left(p_3 - \frac{p_1 q_3 - p_3 q_1}{p_1 q_2 - p_2 q_1} p_2\right) n_3 & = & 0 \\ & & \left(q_2 - \frac{p_2 q_1}{p_1}\right) n_2 & + & \left(q_3 - \frac{p_3 q_1}{p_1}\right) n_3 & = & 0 \end{cases}.$$

We are certainly not happy with this, yet. You should note that this is messy because we are working in complete generality. For any particular example this is not quite so frightening. To clean up some more, we can work out that n_3 term in the first equation. Then we obtain

$$\begin{cases} p_1 n_1 & & & + & p_1 \frac{p_3 q_2 - p_2 q_3}{p_1 q_2 - p_2 q_1} n_3 & = & 0 \\ & & \left(q_2 - \frac{p_2 q_1}{p_1}\right) n_2 & + & \left(q_3 - \frac{p_3 q_1}{p_1}\right) n_3 & = & 0 \end{cases}.$$

Finally, divide through the first equation by p_1 to get rid of a common factor, and divide through the second equation by the coefficient of n_2. We now have something that does not look overly scary.

$$\begin{cases} n_1 & & & + & \frac{p_3 q_2 - p_2 q_3}{p_1 q_2 - p_2 q_1} n_3 & = & 0 \\ & & n_2 & + & \frac{p_1 q_3 - p_3 q_1}{p_1 q_2 - p_2 q_1} n_3 & = & 0 \end{cases}.$$

Something wonderful has happened. Note that in each equation, only two of the variables n_i appear. In fact, if we just choose

n_3, then the equations tell us exactly what n_1 and n_2 should be. We can just write

$$n = \begin{pmatrix} \frac{p_2q_3 - p_3q_2}{p_1q_2 - p_2q_1} n_3 \\ \frac{p_3q_1 - p_1q_3}{p_1q_2 - p_2q_1} n_3 \\ n_3 \end{pmatrix} = n_3 \begin{pmatrix} \frac{p_2q_3 - p_3q_2}{p_1q_2 - p_2q_1} \\ \frac{p_3q_1 - p_1q_3}{p_1q_2 - p_2q_1} \\ 1 \end{pmatrix}$$

This is a good sign. Recall that a plane has many normal vectors. If we find one, we could always just rescale it to get another one. Since our work shows us this property, we are happy.

But for now, we just want to find any one single normal vector, so we have to make a choice. And it would be good to make a choice that helps simplify all of the messy fractions we have derived. We choose $n_3 = p_1q_2 - p_2q_1$ so that we can easily clear the denominator common to the first two coordinates. We substitute this expression in, and we have finally found that our preferred normal vector[7] is

$$n = \begin{pmatrix} p_2q_3 - p_3q_2 \\ p_3q_1 - p_1q_3 \\ p_1q_2 - p_2q_1 \end{pmatrix}.$$

We deduce that the equation for the plane must be

$$0 = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot \begin{pmatrix} p_2q_3 - p_3q_2 \\ p_3q_1 - p_1q_3 \\ p_1q_2 - p_2q_1 \end{pmatrix},$$

or, more simply,

$$(p_2q_3 - p_3q_2)x + (p_3q_1 - p_1q_3)y + (p_1q_2 - p_2q_1)z = 0.$$

This is our desired result.

We should compare our two versions of the work. Looking carefully, we see that the only differences in the end results are a change of sign. The normal vectors are the same, up to multiplying by the scalar -1, and the equations are the same up to multiplying through by -1.

[7] Some texts call this particular choice of n the *cross product* of p and q.

The Equations of a Line in Space

After all of that intense computation, it feels like a little bit of a miracle. But the miracle is just that we did things carefully and got through without arithmetic errors.

Theorem 22. *Let $\mathcal{Q}$ be a plane in $\mathbb{R}^3$ which passes through the origin O and two points $P = (p_1, p_2, p_3)$ and $Q = (q_1, q_2, q_3)$. Then any normal vector to $\mathcal{Q}$ is a scalar multiple of the vector*

$$n = \begin{pmatrix} p_2q_3 - p_3q_2 \\ p_3q_1 - p_1q_3 \\ p_1q_2 - p_2q_1 \end{pmatrix},$$

and the plane may be described as the set of points which satisfy the equation

$$(p_2q_3 - p_3q_2)x + (p_3q_1 - p_1q_3)y + (p_1q_2 - p_2q_1)z = 0,$$

or parametrically as the image of

$$(s,t) \mapsto s\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + t\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}.$$

General Planes in $\mathbb{R}^3$

So far, all of our hard work has led us to a deep understanding of planes in $\mathbb{R}^3$, but only those that pass through the origin O. But most of the planes in $\mathbb{R}^3$ do not contain the origin. We must figure out how to handle those, too.

The Parametric Form of a General Plane Through Three Given Points

In the case of lines in the plane, we used a simple method: Given a line not through the origin, find a parallel line through the origin. Describe that line by whatever method you want, and then translate back to get the description you want. We shall use this same method here.

Definition 23. *Let $\mathcal{P}$ and $\mathcal{Q}$ be planes in $\mathbb{R}^3$. We say that $\mathcal{P}$ and $\mathcal{Q}$ are* parallel *if they have no points in common.*

So, we begin with three points which define a plane $\mathcal{P}$ in $\mathbb{R}^3$. Call them P, Q, and R. As usual, we write p, q, and r for the mathematician's vectors which have their heads at these three points, respectively. Now the physicist's vectors $p-r$ and $q-r$ lie entirely in in $\mathcal{P}$. The corresponding mathematician's vectors lie in a plane $\mathcal{Q}$ which is parallel to $\mathcal{P}$. Also, the new plane $\mathcal{Q}$ contains the origin. In fact, these two planes differ exactly by a parallel translation by the vector r: for each vector v with its head in $\mathcal{P}$, the corresponding vector $v-r$ lies entirely in $\mathcal{Q}$.

Now, since the plane $\mathcal{Q}$ passes through the origin, we can give it one of our two types of descriptions, each of which will help us build the corresponding description for the plane $\mathcal{P}$. Let us take them in turn, starting with the parametric description.

The plane $\mathcal{Q}$ is the collection of vectors formed by linear combinations of $p-r$ and $q-r$.

$$s(p-r)+t(q-r).$$

Since $v-r$ lies in $\mathcal{Q}$, we see that

$$v-r=s(p-r)+t(q-r),$$

or

$$v=r+s(p-r)+t(q-r).$$

Since each of the vectors p, q, and r are known to us, it is straightforward to compute this explicitly. We get the parametric description

$$(s,t)\mapsto r+s(p-r)+t(q-r).$$

Or, in set notation,

$$\mathcal{Q}=\{r+s(p-r)+t(q-r)\mid s \text{ and } t \text{ are real numbers}\}.$$

FROM THE PARAMETRIC DESCRIPTION OF A GENERAL PLANE TO AN EQUATION

Next, let us consider the equation of $\mathcal{P}$. The plane $\mathcal{Q}$ is the set of points $w = (x, y, z)$ which are orthogonal to the normal vector n which is as described in Theorem 22. But $v - r$ is an element of $\mathcal{Q}$, so we see that our equation for $\mathcal{P}$ is just

$$n \cdot (v - r) = 0, \quad \text{or} \quad n \cdot v = n \cdot r.$$

The trick is that first you must compute n from $p - r$ and $q - r$.[8] If n has components a, b, and c, we get an equation of the form

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

Since the planes $\mathcal{P}$ and $\mathcal{Q}$ are parallel, they share the same normal vector. This is apparent in the structure of their equations: the coefficients of the variables represent the normal vector, and they are the same. But the number on the right-hand side, d, depends on exactly which plane you consider in the family of parallel planes with normal vector n.

If we put together all of this work with the things we learned above, we obtain this result.

Theorem 24. *Let $\mathcal{P}$ be the plane in $\mathbb{R}^3$ which contains the points $R = (r_1, r_2, r_3)$, $P = (p_1, p_2, p_3)$, and $Q = (q_1, q_2, q_3)$. Let p, q, and r be the vectors with their heads at these points, respectively. Then we have the following:*

- *$\mathcal{P}$ is parallel to a plane $\mathcal{Q}$ which passes through the origin. These two planes have a common normal vector*

$$n = \begin{pmatrix} a \\ b \\ c \end{pmatrix} := \begin{pmatrix} (p_2 - r_2)(q_3 - r_3) - (p_3 - r_3)(q_2 - r_2) \\ (p_3 - r_3)(q_1 - r_1) - (p_1 - r_1)(q_3 - r_3) \\ (p_1 - r_1)(q_2 - r_2) - (p_2 - r_2)(q_1 - r_1) \end{pmatrix}$$

[8] We did this before in two ways, each of which led us to the idea of the cross product.

– *$\mathcal{P}$ has the parametric description*

$$(s,t) \mapsto r + s(p-r) + t(q-r).$$

– *$\mathcal{P}$ has the equation*

$$ax + by + cz = ar_1 + br_2 + cr_3,$$

where a, b, c are the components of n as above.

It is worth noting that the actual techniques of our earlier work carry over just fine, too. We can pass back and forth between a parametric description of a plane and an equation for that plane.

From the Equation of a General Plane to a Parametric Description

Suppose you have the equation $ax + by + cz = d$. To find a parametric description of the plane, do the following.

1. Choose a nonzero coefficient. For simplicity, assume it is $a \neq 0$. Then solve for the corresponding variable:

$$x = \frac{d}{a} - \frac{b}{a}y - \frac{c}{a}z.$$

2. Choose parameter names, and introduce dummy equations for them. Say $y = s$ and $z = t$.
3. Substitute the new parameters in the equation, and set up a system of all of the equations you have:

$$\begin{cases} x &= \frac{d}{a} - \frac{b}{a}s - \frac{c}{a}t \\ y &= s \\ y &= t \end{cases}.$$

4. Rewrite the system as a vector equation.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{d}{a} \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -\frac{b}{a} \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -\frac{c}{a} \\ 0 \\ 1 \end{pmatrix}.$$

From a Parametric Description of a General Plane to an Equation

In the other direction, suppose that you have a parametric description and want to change it into an equation. The key is to unbundle the parametric description into a system of equations, and then steadily eliminate the variables.

If the system of equations is

$$\begin{cases} x &= a + bs + ct \\ y &= d + es + ft \\ z &= g + hs + it \end{cases},$$

Then we can use the first and second equation to eliminate s, and then the second and third equations to also eliminate s, to get two equations involving only x, y, z, and t, like so:[9]

$$\begin{cases} ex - by &= \alpha + \beta t \\ gy - ez &= \gamma + \delta t \end{cases},$$

Now we can similarly use these two equations to eliminate t and obtain an equation of the form

$$Ax + By + Cz = D.$$

Lines in $\mathbb{R}^3$ as Solutions to Systems of Equations

We had a major goal to describe lines in $\mathbb{R}^3$, but got distracted by the nature of planes in 3-space.

So, how can we find something like an equation for a line in $\mathbb{R}^3$? We have learned that a single equation relating three variables x, y, and z ends up describing a plane. This means that a single equation can't work!

[9] I ran out of symbols, so here are some new friends. The letters $\alpha, \beta, \gamma, \delta$ come from an old form of Greek, and are called alpha, beta, gamma, and delta.

The key idea is that a line is what you get when you intersect two planes. So, to describe a line, we need two equations, not one.

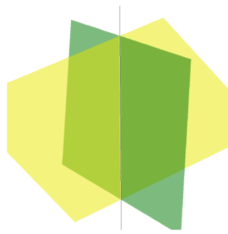


Figure 22: A Line in $\mathbb{R}^3$ is the intersection of two planes

So the implicit description of a line in $\mathbb{R}^3$ must take the form of a pair of linear equations. Each single equation represents a plane, in that the set of solutions of that one equation is a plane in space. The the line is the *set of common solutions* to the pair of equations, that is, the set of points which satisfy both of the equations, considered at the same time.

From a Parametric Line to a Pair of Equations

Suppose we are given a line described parametrically. $t \mapsto p + tv$. How can we find two equations which will together describe the line? Well, our line has a direction vector v. Choose two vectors $n_{\mathcal{P}}$ and $n_{\mathcal{Q}}$ which are orthogonal to v and are not scalar multiples of each other. There are lots and lots of pairs of vectors like this, just pick. It seems like a staggering amount of freedom, but we might as well enjoy it.

Now we consider the plane $\mathcal{P}$ with equation

$$n_{\mathcal{P}} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = n_{\mathcal{P}} \cdot p,$$

and the plane $\mathcal{Q}$ with equation

$$n_{\mathcal{Q}} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = n_{\mathcal{Q}} \cdot p.$$

Since the normal vectors $n_{\mathcal{P}}$ and $n_{\mathcal{Q}}$ are not scalar multiples of each other, these planes are not parallel. That means these planes intersect along a line. The important thing is that the line of intersection is exactly our line. To see this, we just check that the vectors $p + tv$ satisfy the two equations. It is straightforward to check this directly:

$$n_{\mathcal{P}} \cdot (p + tv) = n_{\mathcal{P}} \cdot p + tn_{\mathcal{P}} \cdot v = n_{\mathcal{P}} \cdot p + 0 = n_{\mathcal{P}} \cdot p,$$

so the vectors $p + tv$ have heads in $\mathcal{P}$, and

$$n_{\mathcal{Q}} \cdot (p + tv) = n_{\mathcal{Q}} \cdot p + tn_{\mathcal{Q}} \cdot v = n_{\mathcal{Q}} \cdot p + 0 = n_{\mathcal{Q}} \cdot p,$$

so the vectors $p + tv$ also have heads in $\mathcal{Q}$.

Theorem 25. *A line in $\mathbb{R}^3$ can be described as the common solution set for a system of equations representing two planes*

$$\left\{ \begin{array}{ccccccc} a_{11}x & + & a_{12}y & + & a_{13}z & = & b_1 \\ a_{21}x & + & a_{22}y & + & a_{23}z & = & b_2 \end{array} \right. ,$$

where the two normal vectors

$$n_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix}, \quad \textit{and} \quad n_2 = \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix}$$

are both orthogonal to the direction vector v of the line, but are not scalar multiples of each other.

There are many such pairs of planes which have the line as their intersection.

From a Pair of Equations to a Parametric Line

This shows us how we might go from a parametric description to a system of equations. How might we work in the other direction? Suppose we are given a pair of planes described by equations, and we wish to find a parametric description for the line which is their common solution set.

Now the process is exactly that like we used to find a normal vector to a plane through the origin. Just as in the few pages running up to Theorem 22, we have a pair of equations and we want a common solution. The idea is to replace the very general set of equations

$$\begin{cases} a_{11}x & + & a_{12}y & + & a_{13}z & = & b_1 \\ a_{21}x & + & a_{22}y & + & a_{23}z & = & b_2 \end{cases}$$

with a nicer, equivalent set of equations that looks like this

$$\begin{cases} x & & & + & \alpha z & = & d_1 \\ & & y & + & \beta z & = & d_2 \end{cases},$$

which we can rewrite in vector form as the parametric description

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ 0 \end{pmatrix} + t \begin{pmatrix} -\alpha \\ -\beta \\ 1 \end{pmatrix}.$$

This process is routine to do once learned, but it doesn't always work. There are two kinds of trouble that could happen.

The first is that some pairs of equations represent planes that actually don't intersect. The planes are parallel, and at some point we find an equation that doesn't actually make any sense, like $0 = 1$.

The second problem is that some pairs of equations are really the same plane written twice. In this case, the solution set is all of (both of) the planes, not a line. Then the process ends up turning one equation into something trivially true, usually $0 = 0$, and the rest collapses back on our technique for writing the parametric description for the plane.

Both of these situations require that the normal vectors to the planes are scalar multiples of each other, so with a little bit of care, you can learn to look for that before getting started on the time-consuming arithmetic.

Theorem 26. *There are three cases possible for the shape of the set of simultaneous solutions to a system of 2 equations in three unknowns.*

Generic Case: *The two equations represent a pair of planes which are not parallel. The solution set is a line in* $\mathbb{R}^3$.

Degenerate Case 1: *The two equations are different representations for the same plane. The solution set is that plane.*

Degenerate Case 2: *The two equations represent planes which are distinct, but parallel. There are no solutions because the planes do not intersect.*

Three Viewpoints & Five Questions

We have been motivated by questions that come up naturally when studying the basic geometry of vectors, lines, and planes in $\mathbb{R}^2$ and $\mathbb{R}^3$. In particular, we have found ways to describe the lines in $\mathbb{R}^2$ and $\mathbb{R}^3$ in two ways: *parametrically* as the images of functions; and *implicitly* as the solutions sets of equations.

We have also encountered questions that ask us to find intersections of lines and planes, which means we want to find points which simultaneously solve several different equations. In any particular case, you can usually find answers using ad hoc methods, as long as you are dealing with a smallish number of variables and a smallish number of equations. Our next goal is to find a systematic way of handling such questions.

To do so, we will engage in a particularly mathematical way of working: we will solve these problems by taking them to be special cases of problems which look harder. It is those harder problems we will solve in one go. Somehow the wider perspective helps us see what is important and get to the bottom of things.

This chapter will introduce the core questions of linear algebra, which we will spend the next few chapters answering. But first we will introduce our most important conceptual tools.

What are the three geometric viewpoints for understanding linear algebra?

Systems of Equations – Progress through Generalization

Let us consider an example problem.

Given two lines in the plane $\mathbb{R}^3$*, find their intersection, if it exists.*

Approach #1: parametric lines

Assume that the lines are given parametrically, as

$$t \mapsto \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right) + t \left(\begin{smallmatrix} c \\ d \end{smallmatrix}\right)$$

and

$$s \mapsto \left(\begin{smallmatrix} e \\ f \end{smallmatrix}\right) + s \left(\begin{smallmatrix} g \\ h \end{smallmatrix}\right).$$

Note that we must choose different letters for the two parameters in this task, because the lines are independent objects. We seek a point (x, y) which lies on both of these lines.

If such a point exists, then the fact that the point lies on the first line means there is some value of t so that $x = a + ct$ and $y = b + dt$. Similarly, the fact that the point lies on the second line means that there is some value of s so that $x = e + gs$ and $y = f + hs$. Taking all of those, bundling them together, and rearranging them, we get this collection of equations:

$$\begin{cases} x & & - & ct & & & = & a \\ & y & - & dt & & & = & b \\ x & & & & - & gs & = & e \\ & y & & & - & hs & = & f \end{cases}.$$

You might be wondering why we would bother to rewrite things in this way, but it will help us see the unity of our endeavors going forward.

A more serious objection to this approach is that it is *wasteful.* To solve the problem, we only need to find t, or just s, or just x and y. Our current collection of equations asks us to find all of those. Perhaps we can be more efficient.

Approach #2: parametric lines, again

Rather than take apart the vectors into x and y components, we leave things as vectors. We still observe that the point we want is on both lines, so it must be the head of a vector which can be written in two ways, one for each line. Thus we must have some pair of numbers s and t so that

$$\binom{a}{b} + t\binom{c}{d} = \binom{e}{f} + s\binom{g}{h}.$$

Again, we'll do a little bit of re-organizing, and write this in the form below:

$$t\binom{c}{d} + s\binom{-g}{-h} = \binom{e-a}{f-b}$$

That is much better. If we like we can unbundle these 2-vectors to write:

$$\left\{\begin{array}{ccccc} ct & - & gs & = & e-a \\ dt & - & hs & = & f-b \end{array}\right. .$$

Notice that this set of equations is what you get from the last approach if you eliminate the variables x and y. Still, this formulation asks us to find the two numbers s and t, when just one of them will do to solve the original problem.

Approach #3: lines given by equations

Suppose that each line is described by an equation, so that the points (x, y) on the first line are those that satisfy the equation $ax + by = c$, and the points on the second line are those that satisfy the equation $dx + ey = f$. Then we need to find x and y which satisfy both equations:

$$\left\{\begin{array}{ccccc} ax & + & by & = & c \\ dx & + & ey & = & f \end{array}\right. .$$

And this feels like the perfect fit. We need to find the numbers x and y, no more, no less, to solve this problem.

Of course, all three of these have the same information in them, but the set-up and arrangement makes a difference.

What the three set-ups have in common is that they all involve looking for a set of numbers which together are a solution for several different equations all at the same time, and the equations are of a rather simple type.

That is the key insight for our reformulation and generalization. We will try to solve many equations all at the same time, but we will restrict our attention to equations which are simple in form.

Definition 27. *Let m and n be counting numbers. A* system of m linear equations in n unknowns *has the form*

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{cases},$$

where the unknowns are the variables $x_1, \dots, x_n$ we are meant to find, and all of the other letters a_{ij} and b_j are scalars that we are supposed to know already. The numbers a_{ij} are commonly called coefficients.

A solution *of this system is a single set of values for the x_i's which makes all m of the equations true at the same time.*

Each of the three approaches we outlined above involved setting up a system of linear equations. The first had $m = 4$ equations in $n = 4$ unknowns, and the second and third each had $m = 2$ equations in $n = 2$ unknowns.

If we think carefully about how we handled the second approach of the three, we see that we should be able to reframe a system of linear equations as an equation involving vectors. But this time the vectors are "bigger" than what we have used so far.

Definition 28. *Let m be a counting number. We define an m-vector to be a vertical stack of m real numbers, like so:*

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}.$$

The individual entries u_i of u are called its components. *The collection of all n-vectors is called m*-space, *and denoted $\mathbb{R}^m$.*

We may form linear combinations *of m-vectors by doing scalar multiplication and addition in a component-by-component fashion.*

Similarly, the notions of dot product, norm, and angle extend in the expected way to m-vectors.

Definition 29. *A* linear combination of m-vectors equation with n terms *is an equation of the form*

$$x_1 u_1 + x_2 u_2 + \cdots + x_n u_n = v,$$

where all of the m-vectors u_i and v are known, but the scalars x_i are unknowns we seek.

A solution *to such an equation is a collection of scalars x_i which make the equation true.*

Theorem 30. *A system of m linear equations in n unknowns can be re-written as an equivalent linear combination of m-vectors equation with n terms which has the same solution set, and vice versa.*

Proof. We start with a generic system of linear equations

$$\left\{\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{array}\right. ,$$

and then use each equation as if it were a coordinate position

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

There are now n different m-vectors on the left-hand side, and we have the form of a linear combination of m-vectors equation with n terms.

Also, this process is reversible. We just redistribute the variables x_i in the scalar multiplication, and then remove the parentheses. □

A First Geometric Viewpoint: Hyperplanes

We have already seen that the set of solutions to a single linear equation $ax + by = c$ in $\mathbb{R}^2$ is the points which lie on a line. Similarly, the set of solutions of a single linear equation $ax+by+cz = d$ in $\mathbb{R}^3$ is the collection of points which lie on a plane.

Later, we will make a technical definition of the word *dimension* for our concept, but for now we think of dimension as a naïve concept of counting the number of independent directions of allowed motion. So, in the plane, with two dimensions, we get a one dimensional thing as solution to a linear equation. In 3-space, with three dimensions, we get a two dimensional thing as solution to a linear equation.

What should happen in $\mathbb{R}^n$, for $n > 3$? Well, we will work by analogy. We expect some sort of "plane-like" object as the solution to a single linear equation.

Definition 31. *The set of points in $\mathbb{R}^n$ which is described as the solution set to a single linear equation,*

$$\mathcal{H} = \{(x_1, \dots, x_n) | a_1x_1 + a_2x_2 + \cdots + a_nx_n = b\},$$

is called a hyperplane in $\mathbb{R}^n$.

Now, given a system of m linear equations in n unknowns, we think of each individual equation as defining a hyperplane in $\mathbb{R}^n$. If a point is going to satisfy all of the m equations at the same time, then it must lie on all of the hyperplanes simultaneously. That is, the point must lie in the common intersection of all m hyperplanes.

> **The Row Picture:** We interpret solving a system of m linear equations in n unknowns as finding the common intersection of m different hyperplanes in $\mathbb{R}^n$.

If we write our constraints instead in the form of a linear combination of m-vectors with n terms,

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

then the question comes out differently. Our fundamental objects here are m-vectors, that is, elements of $\mathbb{R}^m$. The goal is to figure out if, and how, we may write the vector on the right-hand side of the equation as a linear combination of the vectors

$$u_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix}, \quad u_2 = \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix}, \quad \ldots u_n = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix}.$$

Definition 32. *Let $\{u_1, \ldots, u_n\}$ be a set of vectors in $\mathbb{R}^m$. The* span *of this set is the collection*

$$\operatorname{Span}\{u_1, \ldots, u_n\} = \{a_1 u_1 + \ldots a_n u_n \mid a_1, \ldots, a_n \text{ are scalars}\}$$

of all the possible linear combinations of these vectors.

This new language has us recast the task of solving the linear combination equation this way.

> **The Column Picture:** We interpret solving a linear combination of m-vectors with n terms as asking if a given m-vector does or does not lie in the span of a collection of n other m vectors, and a solution is a set of coefficients which makes the given linear combination hit the prescribed vector.

Sometimes it can be useful to imagine a linear combination as a type of linkage which has extendible and contractible arms, but for which the angles are frozen. We are allowed to change the lengths of the arms, but not their relative directions. That analogy is imperfect, though, as it is hard to imagine the cases where the lengths of the arms become zero or even negative.

Matrices, a New Kind of Function

Comparing the written version of a system of linear equations with the written version of a linear combination equation, it appears that things get simpler as we go from the system to the vector equation. This is a bit of an illusion, of course, as all of the same information is there. Rather, we have bundled together each of the columns of coefficients into vectors, which abstracts the problem a bit. That is, it packages up several pieces of information (a whole column of independent coefficients) and into one new kind of object which we treat as a single unit (an m-vector).

The result is that our m equations in n-unknowns now looks like a *single* equation in n unknowns. We have used abstraction to compartmentalize things and make a conceptual simplification. This has so successfully reduced things, we should think about doing it again. What is left to abstract? Well, we combined across the columns, so it is only left to combine across the rows. So we introduce a concept that allows us to bundle together a row's worth of m-vectors.

Definition 33. *A* matrix with m rows and n columns*, or an* $m \times n$ matrix *for short, is a rectangular array of numbers a_{ij} enclosed in a large set of parentheses.*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Each individual number a_{ij} is called an entry *of the matrix. The numbers i and j are called* indices*. The first, i, is allowed to take a value between* 1 *and m and indicates which row the entry sits* *in. The second, j, is allowed to take a value between 1 and n and indicates which column the entry sits in. So the ij-entry is the number which sits at the intersection of row i and column j.*

Note that an $m \times n$ matrix must have mn entries. The whole matrix must be full! We are not allowed to leave any positions empty.

Sometimes, if things are otherwise clear, a matrix is written with just a generic entry label like so:

$$A = \begin{pmatrix} a_{ij} \end{pmatrix}$$

It is often useful to think of an $m \times n$ matrix as if it is a collection of n different column vectors, each of which is an m-vector. Similarly, it is sometimes useful to think of such a matrix as a collection of m different *row vectors*, written left-to-right, each of which is an n-vector.

This perspective with rows dominates the way matrices are read aloud. The convention is to specify the size of the matrix, and then read across the rows, top row first, bottom row last. For example the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is read aloud as "the two-by-two matrix a, b, c, d."

Now, to reorganize our linear combination equation, we will bundle the different columns into a matrix. But that leaves out the unknowns? Where should they go? Our approach is to define a new kind of multiplication.

Definition 34. *Let A be an $m \times n$ matrix, and let x be an n-vector. Then, if the n different columns of A are written as m-vectors $u_1, \dots, u_n$, we define the* matrix-vector product *of A and*

x to be the m-vector

$$\begin{aligned} Ax &= \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &= x_1 \begin{pmatrix} | \\ u_1 \\ | \end{pmatrix} + x_2 \begin{pmatrix} | \\ u_2 \\ | \end{pmatrix} + \dots + x_n \begin{pmatrix} | \\ u_n \\ | \end{pmatrix}. \end{aligned}$$

With this definition in place, we can rewrite our system of m linear equations in n unknowns, or equivalently, our linear combination of m-vectors equation with n terms, as a *matrix-vector equation*:

$$Ax = b$$

In this new version of the problem, we have the matrix A and the m-vector b, and we wish to find the n-vector x.

This new set-up is open to the criticism that we have just made up a bunch of new words in order to pack everything into much smaller notation. And that is exactly correct! Mathematicians love to do things like this. The true test is if the new words and notation help us think in a new and creative way about the problems we wish to solve. This particular instance has been effective, so we give up the objection.

How should we think of these matrix objects? Here are some perfectly acceptable ways to think about it:

- A matrix is just a convenient shorthand for a rectangular array of numbers.
- A matrix, like a vector, is a generalized sort of number that we can study for its own sake.
- A matrix is a sort of machine that operates on vectors. It somehow takes an n-vector as an input, and remixes things to produce an m-vector as output.

All of these are valid, and we will use them all. The first one is the most obvious and it is that one we will use to actually solve systems of linear equations. The second one is a mathematician's dream problem: make up something new and see what it is like! We shall come back to this later. For now, let us focus on the third approach.

Given an $m \times n$ matrix, we can pair it with any n vector, and produce an m vector. This is a kind of *function*. (We shall also use the words *transformation* or *mapping*, here.) This is like the ordinary kind of function from a calculus class, in that it is some sort of way to taking some inputs and turning them into outputs which is unambiguously repeatable. But the functions in calculus all have inputs *which are a single number* and outputs *which are a single number*. In our new situation, the matrix takes an input from $\mathbb{R}^n$ and gives an output in $\mathbb{R}^m$. Mathematicians like to write this kind of information down like this:

$$A : \mathbb{R}^n \to \mathbb{R}^m$$

We are supposed to read this as "A is a transformation of n-vectors into m-vectors."

In calculus, a very useful tool for studying functions is the *graph* of that function. Given a function $f : \mathbb{R} \to R$, you imagine the input numbers as living on a horizontal line, and the output numbers as living on a vertical line, and then plot all of the points $(x, f(x))$ to make the graph. This will not work in our situation. At least not if we want to make an actual picture. The trouble is that our set of inputs is n-dimensional, where $n > 1$, and our set of outputs is m-dimensional where $m > 1$. So, if we want to think about $\mathbb{R}^n$ horizontally and $\mathbb{R}^m$ vertically... We run out of room on a piece of paper.

So we will not use graphs very often. Instead, we will make *transformational pictures* like Figure 23.

In Figure 23, the set of possible inputs, called the *domain*, for our transformation is the $\mathbb{R}^n$ on the left, and the set of possible outputs, called the *target*, is the $\mathbb{R}^m$ on the right. The arrow labeled A is there to remind us that A is a kind of function which picks somehow maps things from the domain into things in the target.

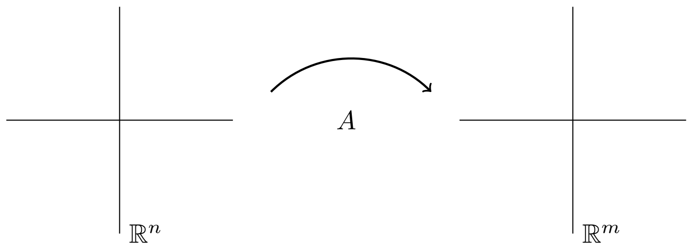


Figure 23: Matrix A as a transformation

The Transformation Picture: We interpret solving an $m \times n$ matrix equation $Ax = b$ as finding the input n-vector x which will get mapped by A onto the given output m-vector b. The solution is the vector x which we must find.

The Five Motivational Questions

Now we have three different viewpoints on what our equations might represent.

algebraic formulation	interpretation of solution
system of m linear equations in n unknowns	the common intersection of m hyperplanes in $\mathbb{R}^n$
linear combination of m-vectors with n terms	the coefficients realizing a vector in $\mathbb{R}^m$ as belonging to a span of n other vectors
matrix-vector equation with an $m \times n$ matrix	the input vectors from $\mathbb{R}^n$ which are mapped to a given output vector in $\mathbb{R}^m$

These three viewpoints are useful in different ways. Certain tasks are easier to think about with a particular model, in a way that often depends on the information you have at hand. So you will want to practice using all three of them, and passing from any one of these to the others, so that you are capable of working efficiently.

Whichever geometric model you use, there is still the overarching problem: solve the equation. But there are several facets to solving an equation! We will split the problem into these five parts, which will take us some time to study.

1. Is there an effective algorithm for finding solutions?
2. Does a given equation have a solution? Can you tell in advance, without computing the answer?
3. Suppose that a given equation has a solution, how many solutions does it have?
4. Suppose that a given equation has a solution, what shape is the set of solutions? Can it be reasonably understood and described?
5. If an equation has no solution, is there an approximate solution?

Solving Systems of Equations

We have seen that systems of linear equations come up naturally when studying the geometry of lines and planes in $\mathbb{R}^2$ and $\mathbb{R}^3$, and more generally when using coordinates and vectors in any $\mathbb{R}^n$. Experience has taught mathematicians that they come up often in many settings.

Our next goal is to answer the questions we listed at the end of the last chapter. We shall do this by taking a closer look with each of our three viewpoints, in turn. In this chapter, let us start with the row picture.

> **How can we find the solution to a system of m linear equations in n unknowns?**

This can be restated geometrically as follows.

> **How can we find the points of mutual intersection of m hyperplanes in $\mathbb{R}^n$?**

Elimination and Row Operations

A general principle for mathematical problem solving is to start with the smallest, easiest versions of your problem, and work towards more complicated versions slowly. That is how we will organize our work on systems.

One Equation in One Unknown

Consider the simplest case, a single linear equation in a single unknown. (This is $m = n = 1$.)

$$a_{11}x_1 = b_1$$

We are to find the variable x_1. Of course, if $a_{11} = 0$, we have to be careful. So, assume for now that $a_{11} \neq 0$. Then this has solution $x_1 = b_1/a_{11}$. Here, the equation has exactly one solution, and we have found it.

If $a_{11} = 0$, our equation takes the form $0x_1 = b_1$. So, if b_1 is not zero, this equation has no solutions at all. But if b_1 is equal to zero, our equation is just $0x_1 = 0$, which is always true. So *any* value of x_1 is a solution.

In any case, we consider the case of $m = n = 1$ to be completely understood. It is a small victory, but a victory nonetheless!

One Equation in Two Unknowns

Another simple case is one linear equation in two variables.

$$a_{11}x_1 + a_{12}x_2 = b_1$$

We are to find x_1 and x_2.

Assume for now that $a_{11} \neq 0$. Then we rearrange the equation to isolate x_1.

$$x_1 = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2$$

So, we can pick any x_2 we like, and then this equation tells us how to choose x_1. If we treat $x_2 = t$ as a parameter, we can write a solution in vector format as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}t \\ t \end{pmatrix} = \begin{pmatrix} b_1/a_{11} \\ 0 \end{pmatrix} + t\begin{pmatrix} -a_{12}/a_{11} \\ 1 \end{pmatrix}.$$

What happens if $a_{11} = 0$? The the equation has the form $0x_1 + a_{12}x_2 = b_1$. So x_1 can take on any value we want, but otherwise, the equation has the form we considered above: It is one equation in the variable x_2. So the analysis now follows the pattern of that situation. There can be exactly one value of x_2 that works, no values of x_2 that work, or every value of x_2 could work.

This pattern will continue. Every time we have a (system of) equations, we will see how to find a solution by reducing to a smaller version of our task (with fewer variables or fewer equations, or both). The number of sub-cases to consider rapidly gets large, and things get messy. So, for the time being, we will focus on the generic cases, and save consideration of the special cases for later when they can be handled uniformly.

Two Equations in Two Unknowns

The generic system of two equations in two unknowns has the form below.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & = & b_2 \end{cases}$$

We are to find x_1 and x_2.

What makes this complicated, relative to the cases we have already handled, is that both equations have both variables in them. We want to eliminate one of the variables from one of the equations to make things simpler. Let us try to eliminate the variable x_1 from the second equation.

The straightforward way to do this is to rearrange the first equation to isolate x_1, and then substitute the resulting expression into the second equation. So, we arrange the first equation to read

$$x_1 = (b_1 - a_{12}x_2)/a_{11},$$

and substitute the right-hand side into the second equation for x_1 to get

$$a_{21}\left(\frac{b_1 - a_{12}x_2}{a_{11}}\right) + a_{22}x_2 = b_2.$$

This is an equation with only the variable x_2 in it, but it is not arranged the way we usually write a linear equation. So, we reorganize this to read in the standard way, and we get

$$\left(a_{22} - \frac{a_{12}}{a_{11}}a_{21}\right)x_2 = b_2 - \frac{a_{12}}{a_{11}}b_1.$$

This is our new version of the second equation. Our system of equations now reads

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ & & \left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 & = & b_2 - \frac{a_{12}}{a_{11}}b_1 \end{cases}$$

This system is now one we can solve. Use the second equation to find x_2. Then substitute that value into the first equation to obtain an equation with just x_1 in it. Solve the resulting equation for x_1. Then write down the results.

Pause and notice what happened. The process was a little complicated, but we turned our original system

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & = & b_2 \end{cases}$$

into this one

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ & & \left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 & = & b_2 - \frac{a_{12}}{a_{11}}b_1 \end{cases}.$$

The result of all of our manipulations was to do this funny thing:

> Add $-a_{21}/a_{11}$ times the first equation to the second equation, and keep the result as the new second equation.

Try it. This has exactly the same results, and crucially, produces a zero for the coefficient of x_1 in the second equation.

That is easier to remember, and to say out loud, than the whole process of isolating, substituting, and regrouping. Since it deals with recombining the information in our rows, this is called a *row operation*.

We will develop three different types of row operations as part of our process. In fact, we have already seen *two*. Think back to the above work. Do you see another common thing we do when solving a system? Pause for a bit and try to figure it out. We will make a list below so you can check your guess.

Two Hyperplanes (lines) in $\mathbb{R}^2$

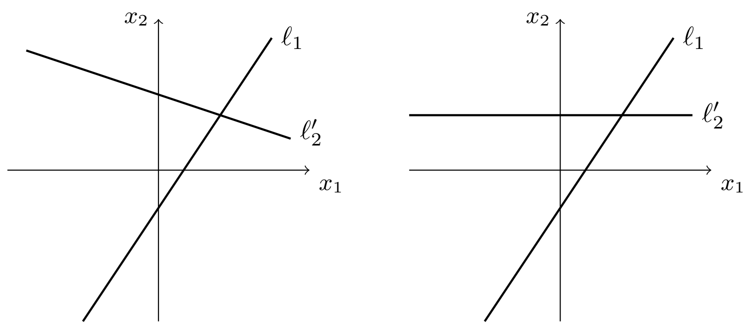
What is the geometric effect of our row operation? Recall that for the case $m = n = 2$, we are thinking about a row picture of two lines in $\mathbb{R}^2$. If our system is

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & = & b_2 \end{cases},$$

we will call ℓ_1 the line described by solutions to the first equation, and ℓ_2 the line described by solutions to the second equation. After applying the row operation, our system becomes something like this one:

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ & & a'_{22}x_2 & = & b'_2 \end{cases}.$$

This row picture has changed a little. It still has the line ℓ_1, but we have a new second equation. We call the line described by the second equation ℓ'_2. The important thing is that the new second equation does not have an x_1 in it. This means that ℓ'_2 is parallel to the x_1-axis.



(a) Before the Elimination Step (b) After the Elimination Step

Figure 24: Changing the Row Picture by a Row Operation

Recursion and Solving Three Equations in Three Unknowns

What about a larger system? Consider the case of $m = 3$ linear equations in $n = 3$ unknowns. With the standard type of notation we use, it looks like this.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{cases}$$

The row picture for this situation interprets solving the system as looking for the common intersection of three planes in $\mathbb{R}^3$. It will help our discussion to have names for these equations, so we call them r_1, r_2, and r_3, respectively. (The idea is that r_i is for "row i.")

Our goal is to use row operations to reduce this problem to a smaller one, namely the case where $m = n = 2$, which we have already considered. This is an example of *recursion*, which is a general principle of computation: You figure out how to handle your task by breaking it into successively smaller pieces of basically the same type of task, until the task at the end of the process is one you can handle. Then you zip back up the chain and put your solution together.

We have already done this a little. We showed how the $m = n = 2$ case can be reduced to (two applications of) the $m = n = 1$ case by a row operation. That $m = n = 1$ case, of the form $ax = b$ is the one we are confident we can do.

The Types of Row Operations

To be definite, it will help to state clearly what our set of row operations is.

Add a non-zero multiple of one row to another Add some scalar multiple of one row to another row, and use the result as a replacement for that latter row. That is, replace the rows r_i and r_j with the rows r_i and $\lambda r_i + r_j$, respectively.

Rescale a row by a nonzero number Multiply through an entire row by a number. That is, replace r_i by some λr_i.

Swap rows Interchange the positions of two rows in the order. That is take the rows r_i and r_j and replace them by rows r_j and r_i, respectively.

How to reduce the solution of a system of three linear equations in three unknowns to the solution of a system of two linear equations in two unknowns

We begin with the system in standard form.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{cases}$$

First, assume that a_{11} is not zero.

1. Eliminate the first term of the second equation using a row operation of the first type: add $-a_{21}/a_{11}$ times r_1 to r_2 and make this the new row two. In symbols, this is often denoted

$$\frac{-a_{21}}{a_{11}} \times r_1 + r_2 \to r_2.$$

2. Eliminate the first term of the third equation using a row operation of the first type: add $-a_{31}/a_{11}$ times r_1 to r_2 and make this the new row two. In symbols,

$$\frac{-a_{31}}{a_{11}} \times r_1 + r_3 \to r_3.$$

3. The system should now be in this form.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ & & a'_{22}x_2 & + & a'_{23}x_3 & = & b'_2 \\ & & a'_{32}x_2 & + & a'_{33}x_3 & = & b'_3 \end{cases}$$

Just the second and third equations together make a system of $m = 2$ linear equations in $n = 2$ unknowns, x_2, x_3. So, by our recursive structure, this is a system that we know how to solve.

Subroutine for $m = n = 2$ case

(a) Eliminate the coefficient of x_2 in the third equation using the first type of row operation: add $-a'_{32}/a'_{22}$ times r_2 to r_3 and make that the new r_3. In symbols:

$$\frac{-a_{32}}{a_{22}} \times r_2 + r_3 \to r_3$$

The third equation should now have the form $\alpha x_3 = \beta$.

(b) Use a rescaling by α^{-1} on r_3 to find the value of $x_3 = \beta/\alpha$.

(c) Substitute the value of x_3 into our current second equation and rearrange to get an equation of the form $\delta x_2 = \gamma$. Use a rescaling by δ^{-1} row operation on the result to find the value of $x_2 = \gamma/\delta$.

We now behave as if we have a solution x_2 and x_3 for the sub-system.

4. Substitute the (now known) values of x_2 and x_3 into the first row and rearrange, to obtain an equation of the form

$$a_{11}x_1 = b'_1.$$

Then use the row operation of rescaling to find the value $x_1 = b'_1/a_{11}$.

5. Put together the values of x_1, x_2, x_3 and report them.

That is it. That is the whole process in the generic case where $a_{11} \neq 0$. Notice that this is designed to use row operations sparingly, in an order to produce zeros that simplify things. For all of the complexity of this on a first reading, the process is straightforward and you should be able to do it after some practice.

It is worthwhile to notice that the process here is designed to reduce our original system to an equivalent one which has this form:

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ & & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ & & & & a_{33}x_3 & = & b_3 \end{cases}$$

Such a system is called an *upper triangular* system. The process of getting from the original to this triangular system is often called "Gaussian Elimination," or, sometimes, "Gauss-Jordan Elimination." (This honors two mathematicians, C.F. Gauss and Wilhelm Jordan, neither of whom was first to publish this method. Both used related methods on more complicated problems. There is decent evidence that this method predates either of these men by fifteen hundred years.) Strictly speaking, we have done the "forward pass" part of Gaussian Elimination. We will see the rest later. The process of successively substituting variables back into previous equations to determine the other variables is sometimes called "back-solving."

Note that Gaussian Elimination really does require the assumption that $a_{11} \neq 0$, since we divide by a_{11} in the first two steps.

Now assume that $a_{11} = 0$. But at least one of a_{21} or a_{31} is not zero. Then, the system has the form

$$\begin{cases} & & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{cases}$$

1. If $a_{21} \neq 0$, use the row swap operation to interchange the first row and the second. If $a_{21} = 0$ but $a_{31} \neq 0$, use the row swap operation to interchange the first row with the third row.
2. follow the steps 1-5 above to solve the system.

Again, we need a non-zero leading coefficient, because we will divide by that number. The special case where all of the coefficients of x_1 are equal to zero needs its own discussion. We will address this case next.

The Case of the Missing Variable

So, what happens when you are presented with a system of 3 linear equations in 3 unknowns, but one of the variables does not actually appear in the system? This may sound like an odd question. Why would we not just ignore that variable and pretend like we have a system of 3 linear equations in 2 unknowns? (That is *almost* what we want to do.) Why would we call it a system with 3 unknowns if it actually has fewer than 3 unknowns? It seems like an odd arrangement. Why would we even consider this?

Well, our recursive approach might force it on us. We could have a situation where our 3×3 system is the subsystem we devised from something larger by doing elimination steps. And it is possible that during the elimination steps we have eliminated more than just our target variable, so the subsystem we create is missing more variables that we expected.

In this case, we will end up considering a system where all of the leading coefficients are zero. That is, we really are looking at

$$\left\{ \begin{array}{ccccc} a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{32}x_2 & + & a_{33}x_n & = & b_3 \end{array} \right. ,$$

but considered as a system on the 3 variables x_1, x_2, x_3, even though the variable x_1 does not explicitly show up.

If we can find a triple $(x_1, x_2, x_3) = (t, u, v)$ which is a solution of this system, then we can find lots of other solutions simply by changing out t for some other number t'. Since the equations no longer feel the influence of x_1, it does not matter which choice we make for x_1.

So, our missing variable x_1 doesn't have any constraints on it. It is an example of what we will call a *free variable* for this system. So, as part of our complete solution, we must allow this variable to take on any value. The basic way to do this is to introduce a new variable name $x_1 = t$, like we did above, and include in our solution description the fact that t can be any number.

More about Free Variables

This problem where a variable goes missing from the equations can happen deeper down in the process, too. If we push through stubbornly and put the system in upper triangular form, we should end up with something like this:

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ & & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ & & & & a_{33}x_3 & = & b_3 \end{cases}.$$

We will get a free variable if any collection of the diagonal coefficients in our triangular system ends up being zeros.

To handle such situations, we do the following:

1. For each free variable x_i, introduce a parameter t_i and add an equation of the form $x_i = t_i$. You will find it useful to put this in the ith row; and then
2. Finish the process as usual, except make sure to write the final solution in a way that shows off the linear combination of vectors with the parameters as coefficients.

By way of example, let us consider the case of a triangular system where $a_{22} = a_{33} = b_3 = 0$. Because the last equation is of the form $0x_1 + 0x_2 + 0x_3 = 0$, which is always true, we will ignore it.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ & & & & a_{23}x_3 & = & b_2 \end{cases}.$$

In this case, we have that x_2 is a free variable. So we let $x_2 = t_2$ be a parameter, and rearrange like so:

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 & \\ & & x_2 & & & = & & t_2 \\ & & & & a_{23}x_3 & = & b_2 & \end{cases}.$$

Now, we do the back-solving: $x_2 = t_2$ and $x_3 = b_2/a_{23}$, and thus

$$\begin{cases} a_{11}x_1 & + & a_{12}t_2 & + & a_{13}\frac{b_2}{a_{23}} & = & b_1 & \\ & & x_2 & & & = & & t_2 \\ & & & & x_3 & = & \frac{b_2}{a_{23}} & \end{cases}.$$

Which can be rearranged to read:

$$\begin{cases} x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{13}\frac{b_2}{a_{23}}\right) - \frac{a_{12}}{a_{11}}t_2 \\ x_2 &= t_2 \\ x_3 &= \frac{b_2}{a_{23}} \end{cases}.$$

Finally, we rewrite this in vector format to describe the complete solution. Note that in this case, our set of solutions is a line in $\mathbb{R}^3$, described parametrically. Here, our solution set is $\mathcal{S}$

$$\mathcal{S} = \left\{ \begin{pmatrix} \frac{1}{a_{11}}\left(b_1 - a_{13}\frac{b_2}{a_{23}}\right) \\ 0 \\ \frac{b_2}{a_{23}} \end{pmatrix} + t_2 \begin{pmatrix} \frac{a_{12}}{a_{11}} \\ 1 \\ 0 \end{pmatrix} \middle| \quad t_2 \in \mathbb{R} \right\}$$

In fact, this process shows us a slightly different arrangement of the work we did to move from the description of a line in $\mathbb{R}^3$ as the intersection of two planes to a parametric description of that same line. Gaussian Elimination is just a generalization of those ideas, in a very real and legally binding sense.[10]

Inconsistent Systems

We constructed the last example by making enough explicit assumptions to force the third equation in our system to have the form $0 = 0$. Of course, elimination can produce zeros on the left hand side of an equation. It is designed to do that! Sometimes, you get more zero coefficients than you have a right to expect. But the right hand side of the equation is just a number, and these numbers are not directly used in our process. They just come along for the ride.

In our last example, we assumed that $b_3 = 0$, and so got the boring, always true equation $0 = 0$. Sometimes, the process of elimination will completely remove all of the variables on the left hand side, but will leave the number on the right hand side nonzero. That is, at the bottom of the triangular system, we will get

[10] I am not sorry about this. At all.

an equation that looks like this:

$$0x_1 + 0x_2 + 0x_3 = \beta,$$

where β is not zero. This is trouble. The equation $0 = \beta \neq 0$ is never true.

In such a case, we see that the system has no solutions whatsoever. If the work ever leads to an equation of the form where zero is supposed to be equal to a non-zero number, you can just stop. This is important enough that there is a word for this situation.

Definition 35. *A system of m linear equations in n unknowns is called* inconsistent *if it has no solutions.*

You should be able to design an example of an inconsistent system. Can you design one that is sneaky, so that it is hard to tell at first glance that the system is inconsistent?

The General System

The general system of m linear equations in n unknowns looks like this.

$$\left\{ \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{array} \right.$$

To describe the set of solutions, we use essentially the same recursive procedure outlined above. For the small cases, we plunged in to the work and dealt with oddities in a rather haphazard fashion. Since we have seen the kinds of things that can happen, we can now be more systematic.

Warning: What I am about to present is not exactly the standard Gaussian Elimination. It is a variant designed to make it easier to write down the solution in vector form. In particular, the introduction of new equations for parameters is non-standard.

Let us give a full description of the recursive procedure which is applicable in the general case. The standard way to do this is to describe the following:

- How to handle the possible set of "smallest cases" that cannot be reduced to something smaller; and
- How an instance of the problem can be reduced to a materially smaller version of the same problem.

The Smallest Case

The smallest case that cannot be reduced any further is of the form of a single equation in one unknown x_1. It will look like this

$$ax_1 = b.$$

If $a \neq 0$, the solution is $x_1 = b/a$, which can be computed by the rescaling row operation.

If $a = 0$, we have two cases to consider. If $b \neq 0$, we have an inconsistent system, and there is no solution. Stop the process and report "no solutions."

If $b = 0$, then we have the trivial equation $0 = 0$, and no constraints on x. This variable is free. So, introduce an equation with a parameter $x = t$ and note that t may be any real number.

The General Case

Now suppose that we have a system of m linear equations in n unknowns. If every coefficient in the first column is zero, then the first variable x_1 is free. Introduce a new equation $x_1 = t_1$ and set it aside as part of the required triangular system. Then move on to study the (original) system of m linear equations, but considered as having only $n - 1$ unknowns, $x_2, \ldots, x_n$.

Otherwise, at least one of the leading coefficients is non-zero. If it is a_{11}, proceed as below. If a_{11} is not zero, but some other a_{i1} is non-zero, choose the one with smallest index i. The perform a row swap operation which interchanges row 1 and row i.

Now we may assume that a_{11} is not zero. Perform the $m - 1$

row operations

$$\frac{-a_{k1}}{a_{11}}r_1 + r_k \mapsto r_k, \quad k = 2, \ldots, m,$$

to eliminate the occurrences of the variable x_1 in equations r_2 through r_m. Set aside the first equation as part of our triangular system. Move on to consider the system of $m-1$ linear equations in $n-1$ unknowns given by the current equations in rows 2 through m.

That is it. That is the whole thing. Some remarks are in order, however.

First, note that each step of the algorithm ensures that the current "first variable" is the leading term of some equation involving only variables with larger values of the index. Since this applies to every variable, we are sure to end up with an upper triangular system of n equations in n unknowns, with non-zero coefficients along the diagonal. The trade-off for this is that we may need to introduce parameters on the right hand side of these equations.

Second, the equations $0 = b \neq 0$ which make a system inconsistent tend to show up near the bottom of the system when doing this process. But they need not wait until the very last moment. If the system ever exhibits one of these, simply stop and report "no solutions."

Third, there are of course cases where it looks like the final equation has many variables in it. The process as outlined above still works! It is just that after we set aside the "last" equation, we really should consider the system of no equations on k variables. That sounds weird, so maybe it is better to think of padding the system with extra rows of $0x_k + 0x_{k+1} + \cdots + 0x_n = 0$. Then just apply the process to see that there are several free variables here.

Fourth, it is possible that the system under consideration has many equations in a single unknown. Again, the process still works. Use the elimination row operations, and check to see if the equations below the first one become trivial ($0 = 0$), or false. If all of those equations are trivial, proceed. If one is false, declare the system is inconsistent and report "no solutions."

Fifth, each major piece of this algorithm is designed to set aside one equation so we can move on to next consider a system with fewer equations, fewer variables, or both. This is why we can be sure the process will eventually stop. At some point, things have to get reduced to a single equation in a single variable.

Finally, after all of the tedious work of putting the system into triangular form, use back-solving to finish the computation. Report the answer in vector form, and organized in a way to highlight the structure of linear combinations with parameters as coefficients.

Subspaces of $\mathbb{R}^n$ and the Structure of Solving Equations

We have answered the principal question of linear algebra. We know how find the solution set for a system of linear equations, and we even have an algorithm for writing it down explicitly.

So we will add depth to our understanding, next. Part of this will involve looking more closely at the two geometric interpretations of a system: the column picture and the transformational picture. Another part will be to introduce and use the notion of a *subspace* to organize our thinking, and find the structure hiding in plain sight. The subspaces associated to a system/linear combination/matrix will give us the tools to answer all of the questions we set ourselves at the end of chapter 3, though maybe not all of them right away. We will encounter many new geometric concepts about vectors, including *spanning sets*, *linear dependence and linear independence*, and *basis*. Finally, we will see how the process of Gaussian Elimination can be refined a bit to get at the answers to these questions more quickly.

How can we understand the structure of a system of linear equations?

Notational Refresher

It will help us to have a uniform notation system for our work, so let us set one up. A system of m linear equations in n unknowns in standard form looks like this.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{cases}$$

We can rewrite it as a linear combination of n vectors in $\mathbb{R}^m$ like so:

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

It will be convenient to have a short-hand for the names of these vectors, so for each $j = 1, \dots, n$, call the jth column u_j, and call the vector on the right-hand side b.

$$u_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

Now we can write our system in this form

$$x_1 u_1 + \cdots + x_n u_n = b.$$

Finally, we can also bundle together those vectors u_j as the columns of an $m \times n$ matrix A, and put the unknowns into a vector x

$$A = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_n \\ | & | & & | \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

and then rewrite our system as a matrix-vector equation

$$Ax = b.$$

The Idea of a Subspace

The new organizing idea for our study is that of a subspace. Roughly speaking, a subspace of $\mathbb{R}^n$ is a collection of points in $\mathbb{R}^n$ that has the same basic structure of linear combinations.

Definition 36. *Let S be a collection of vectors in $\mathbb{R}^n$. We say that S is a* subspace *of $\mathbb{R}^n$ when the following conditions hold:*

- *The zero vector is an element of S.*
- *For any vectors w and v in S, and any scalars λ, μ, the linear combination $\lambda w + \mu v$ is also in S.*

We have already encountered some subspaces, but it helps to consider the most important classes of examples.

Spans as Subspaces

Suppose you have a collection of m-vectors $\{u_1, \dots, u_n\}$. Since these are all m-vectors, they all live in $\mathbb{R}^m$. The set

$$\operatorname{Span}\{u_1, \dots, u_n\} = \{x_1 u_1 + \cdots + x_n u_n | x_1, \dots, x_n \in \mathbb{R}\}$$

consisting of all linear combinations of the given vectors is a subspace. This is not difficult to check. First, if we choose all of the coefficients $x_i = 0$, we see that the zero vector is an element of $\operatorname{Span}\{u_1, \dots, u_n\}$. Then if we have two elements w and v in the span, and two numbers λ, μ, we need to check that the linear combination $\lambda w + \mu v$ is an element of $\operatorname{Span}\{u_1, \dots, u_n\}$. Since w is an element of the span, we can write w as a linear combination of the u_j's for some choice of coefficients s_j:

$$w = s_1 u_1 + s_2 u_2 + \cdots + s_n u_n.$$

Similarly, v is an element of the span, so we can write v as a linear combination of the u_j's for some choice of coefficients t_j:

$$v = t_1 u_1 + t_2 u_2 + \cdots + t_n u_n.$$

If we are careful and do a little rearranging, we learn that

$$\begin{aligned}\lambda w + \mu v &= \lambda\left(s_1 u_1 + \cdots + s_n u_n\right) + \mu\left(t_1 u_1 + \cdots + t_n u_n\right)\\ &= \left(\lambda s_1 + \mu t_1\right) u_1 + \cdots + \left(\lambda s_n + \mu t_n\right) u_n,\end{aligned}$$

so that the linear combination $\lambda w + \mu v$ is also an element of $\operatorname{Span}\{u_1, \ldots, u_n\}$. Since we have checked both parts of the definition, we are done. So we don't lose it, let us recap with a firm statement.

Theorem 37. *If $u_1, \ldots, u_n$ are all m-vectors, the set $\operatorname{Span}\{u_1, \ldots, u_n\}$ is a subspace of $\mathbb{R}^m$.*

Since we can write our systems of equations in the form of a linear combination of vectors equation, we see that subspaces built as spans will be relevant.

HOMOGENEOUS EQUATIONS DEFINING SUBSPACES

The other natural way to build subspaces has come up, but is hidden deeper in our work so far.

Definition 38. *A system of m linear equations in n unknowns is called* homogeneous *if the right hand side consists of all zeros. That is, we say a system is homogeneous if it has the form*

$$\left\{\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & 0\\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & 0\\ \vdots & & \vdots & & \ddots & & \vdots & = & \vdots\\ a_{m1}x_1 & + & a_{m2}x_2 & + & \ldots & + & a_{mn}x_n & = & 0\end{array}\right.$$

The fact that the right hand side is all zeros gives this system just a little extra help, and it is enough to make the solution set into a subspace of $\mathbb{R}^n$.

Theorem 39. *The solution set of a homogeneous system of m linear equations in n unknowns is a subspace of $\mathbb{R}^n$.*

Proof. There are two conditions to check. First, we must show that the zero vector is a solution to this system. But, if we set

$x_1 = \cdots = x_n = 0$, of course the left hand side of every equation collapses to zero. Since this system is homogeneous, the right hand side of every equation is zero. Thus the zero vector is a solution to this system.

Next, we must show that if w and v are two solutions, and λ and μ are scalars, then the linear combination $\lambda w + \mu v$ is also a solution. Write

$$w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \qquad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

For the moment, consider just the first equation. Since w is a solution of this system, its components must satisfy the first equation. That is, we must have that

$$a_{11}w_1 + a_{12}w_2 + \cdots + a_{1n}w_n = 0.$$

Similarly, since v is a solution of the system, its components also satisfy the first equation.

$$a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n = 0.$$

But now we may multiply through the first equation by λ and the second equation by μ and add to find (after some rearranging)

$$a_{11}(\lambda w_1 + \mu v_1) + a_{12}(\lambda w_2 + \mu v_2) + \cdots + a_{1n}(\lambda w_n + \mu v_n) = 0.$$

These entries are now just the components of the linear combination $\lambda w + \mu v$. So we deduce that $\lambda u + \mu v$ satisfies the first equation.

But there is nothing special about the first equation in that argument. We can do exactly the same thing for all m equations. We deduce that $\lambda w + \mu v$ is a solution of the homogeneous system of equations.

We have checked both parts of the definition, hence the solution set to our homogeneous system of equations is a subspace of $\mathbb{R}^n$.

$\square$

DESCRIBING A SPAN WITH EQUATIONS

It seems now that we have two methods for making subspaces: spans, and solution sets to homogeneous equations. But stop for a moment and reflect on the work we did to describe a solution set. If we start with a homogeneous system of linear equations and apply the Gaussian elimination routine, we eventually produce a parametric description for the solution set. Those always look like this:

$$\text{Soln} = \{d + t_1 v_1 + t_2 v_2 + \cdots + t_k v_k | t_i \in \mathbb{R}\}$$

Where does that d come from? It is always the set of constants on the right hand side left over after the elimination steps. But if we start with all zeros there, every row operation we do will leave all zeros there. So we must have that for a homogeneous system $d = 0$. That means our solution set has the form

$$\text{Soln} = \{t_1 v_1 + t_2 v_2 + \cdots + t_k v_k | t_i \in \mathbb{R}\},$$

which is just $\text{Span}\{v_1, \ldots, v_k\}$!

For a homogeneous system of equations, Gaussian elimination tells us how to turn our implicit description of the subspace of solutions into a parametric description as a span. So, the two types of examples are really just the same type, with different descriptions.

Naturally, we would like to know if we can pass the other way. Fortunately, the answer is "Yes."

Theorem 40. *Let $u_1, \ldots, u_n$ be a collection of vectors in $\mathbb{R}^m$. Then the subspace $\text{Span}\{u_1, \ldots, u_n\}$ can be realized as the solution set to a homogeneous system of linear equations.*

Proof. The proof is constructive, in that we will show how to explicitly calculate the equations which are required to make the homogeneous system.

A vector b lies in $\text{Span}\{u_1, \ldots, u_n\}$ exactly when it is possible to solve the linear combination of vectors equation

$$x_1 u_1 + x_2 u_2 + \ldots x_n u_n = b.$$

Equivalently, b lies in Span $\{u_1, \dots, u_n\}$ exactly when it is possible to solve the system of equations

$$\left\{\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{array}\right.$$

In this case, we don't actually care what the solution vector x looks like. Instead, we want to know what conditions the components b_i of b must satisfy for this to have a solution. So, think of the b_i's as the variable names for the space of possible elements of the span. (This is a copy of $\mathbb{R}^m$, since b is an m-vector.) We will try to find equations on the unknowns b_i.

The key is to apply the Gaussian elimination process to the equations, and just carry along the b_i's on the right hand side however it goes. Here, it is not necessary to do any step where you introduce a parameter, because all we care about is if the system of equations has a solution or not. We don't actually care to solve it.

But what is the test for solvability? The critical thing is that no equation can become inconsistent. So, do the Gaussian elimination process until the system is triangular. Then gather together all of the equations which have taken the form

$$0x_1 + \dots + 0x_n = \text{some combination of the } b_i\text{'s.}$$

The system is consistent exactly when each of these equations is trivially true, that is, when those combinations of b_i's on the right hand sides really do equal zero. This forms a homogeneous system of equations on the m unknowns $b_1, \dots, b_m$, and that system is one we desire.

If there are no equations of this special type, then we can conclude that there are no conditions required of the components of b, so any vector b will work. That is, Span $\{u_1, \dots, u_n\} = \mathbb{R}^m$ is the whole of $\mathbb{R}^m$. $\square$