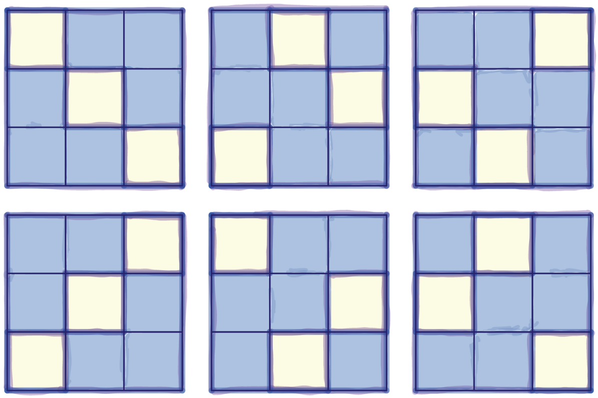


Notes on Linear Algebra

Dave Bayer

Dave Bayer
Department of Mathematics
Barnard College
Columbia University
New York, New York

Draft of Monday 28th September, 2020

A note about the type
The text is set in Caecilia LT Std, designed by Peter Matthias Noordzij. The sans serif font is Chaparral Pro, designed by Carol Twombly. The math font is Euler, designed by Hermann Zapf. The typewriter font is ITC Officina Serif Std, designed by Erik Spiekermann.

Contents

Contents i

Preface iii

1 Linear maps 1

1.1 Linear Maps 1

1.2 Affine Combinations 9

1.3 Matrix Multiplication 11

2 Systems of Equations 15

2.1 Row Reduction 15

2.2 The General Case 23

Preface

These are course notes to supplement my linear algebra lectures at Barnard College and Columbia University. I am writing them as I teach online during the Fall semester of 2020.

This is a draft of Monday 28th September, 2020. The most recent draft of this text may be downloaded from

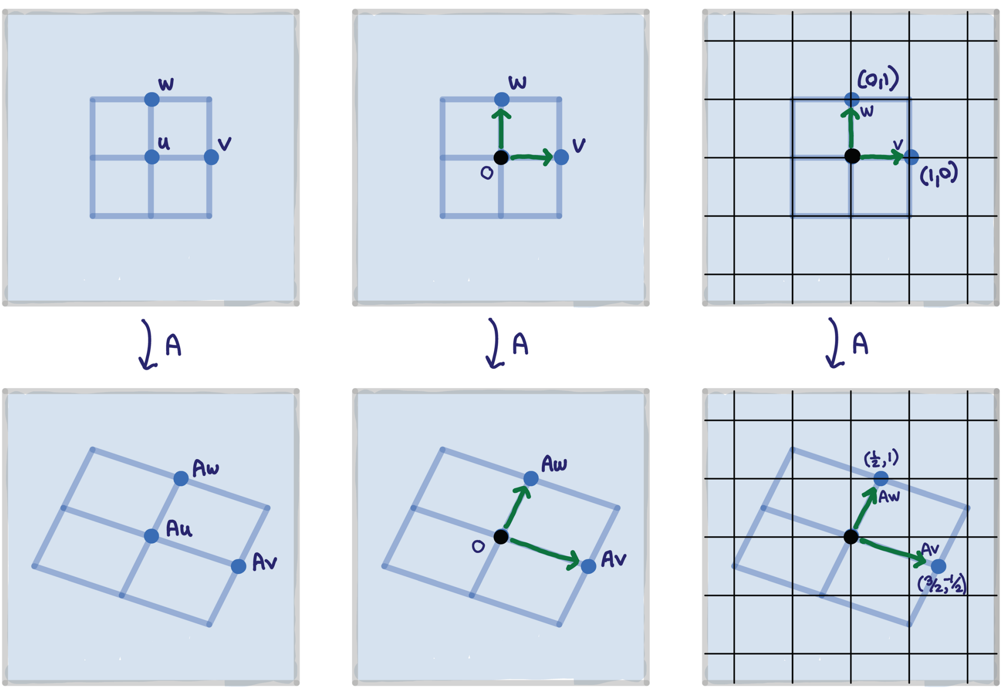
https://www.math.columbia.edu/~bayer/LinearAlgebra

DAVE BAYER
New York, NY
September 2020

Chapter 1

Linear maps

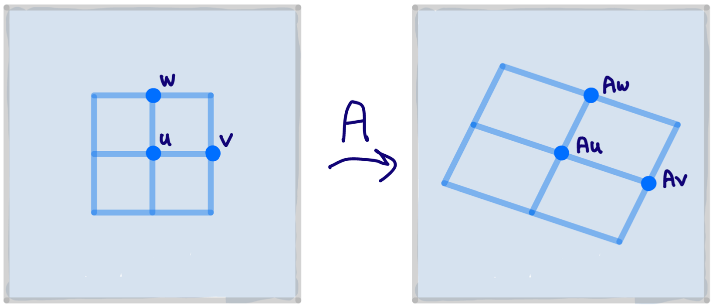
1.1 Linear Maps



Linear algebra is the study of maps that preserve relative positions, in spaces such as the plane.

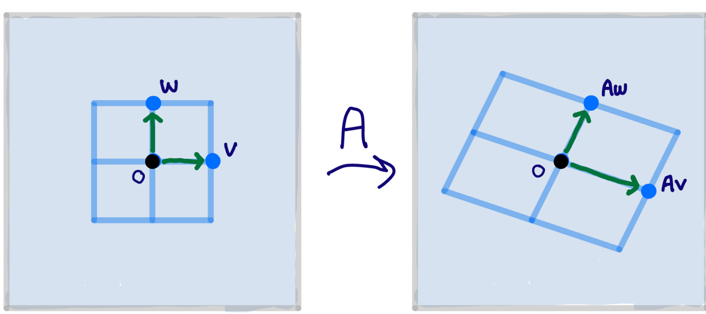
The above drawing shows three views of a linear map A from the plane to itself. The left view is like a blank sheet of paper. There are no coordinates, and no origin.

Three points determine a plane: For any three points in a higher dimensional space, if the points don't all lie on a line, then there is a unique plane passing through those points. A stool with three legs doesn't rock; its three legs determine a plane.



In this view, the three marked points u, v, w on our plane serve as a reference to find any other point on the plane, by relative position. Our map A has the freedom to take u, v, w to any three points in the image, but this data determines the image of every other point: For any point p in the domain of A, find its position relative to u, v, w. The map A will take p to the same position relative to the images of u, v, w. We have drawn four squares positioned around the points u, v, w. The map A takes these squares to parallelograms in the same positions relative to the images of u, v, w.

A map that preserves structure in this way is called an *affine map*. Affine maps preserve *affine combinations*; they are our tool for measuring this structure.

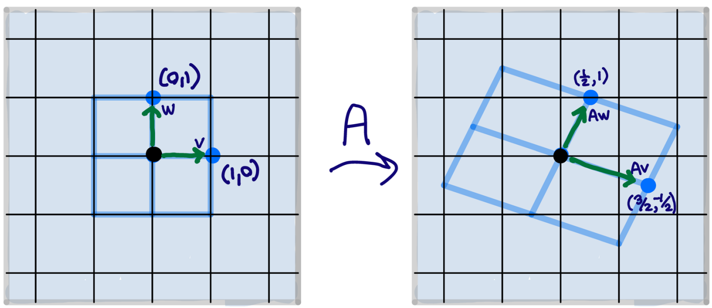


In this next view, we have chosen origins. An affine map that takes the domain origin to the image origin is called a *linear map*. Linear maps preserve *linear combinations*.

Origins complicate our geometric view, but simplify computations. Linear maps are a special case of affine maps, but this is not a significant restriction: For any affine map A, we can arbitrarily declare that some u and its image Au are origins, and we have a linear map.

We often naturally have such a pair. For example, when we study the derivative of a function F in multivariable calculus, we consider how F behaves near a point u and its image $F(u)$. In the limit F acts like a linear map, after we declare that u and $F(u)$ are origins.

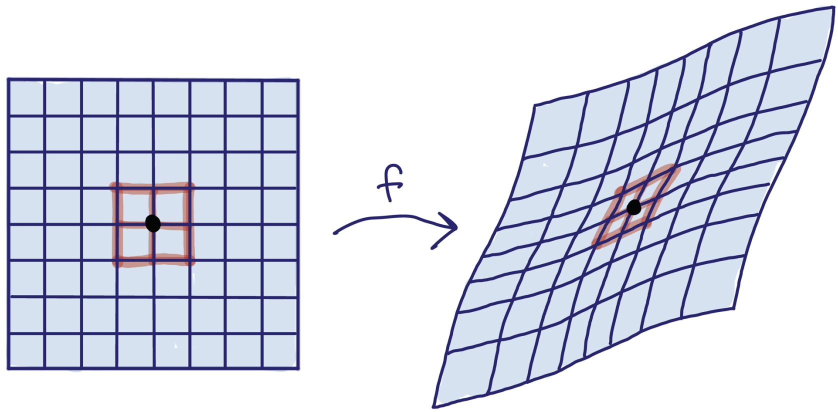
With the introduction of origins, we can use algebra to compute relative positions. For example, thinking of v and w as the *vector* motions from the origin to the points v and w, we can add these motions together to compute the remaining corner of their unit square. We can add the motions from the image origin to the points Av and Aw, to compute the remaining corner of their parallelogram.



In this final view, we have added coordinates. Coordinates allow us to do arithmetic to compute relative positions and work explicitly with linear maps. Shown are *Cartesian coordinates*, a standard choice relying on a horizontal x-axis and a vertical y-axis. Our choice of a coordinate system is however arbitrary.

One challenge in learning any subject is separating what is intrinsically there from what is the shadow of our own hand. Scientific progress is a process of throwing off the yokes of past misconceptions. One needs to develop technical abilities, but they alone no more determine success than physical conditioning alone makes an athlete. The key journey is psychological, questioning assumptions and trusting one's own perspective.

Struggle is inevitable, for this is a process of breaking habits. From this perspective, coordinate systems are a liberating playground. It is easy to see that a choice of coordinate system is arbitrary, and to reject the conventional choice of Cartesian coordinates. Any suitable coordinate system can be used to compute relative positions and work with maps, and standard coordinates are often uninformative. We will learn to choose coordinate systems that best reveal the structure of our linear maps.

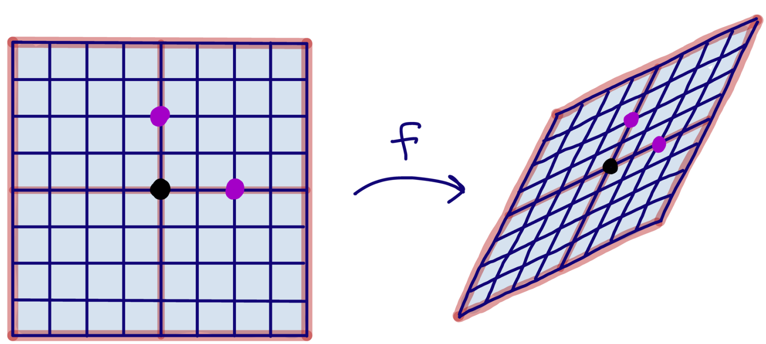


The derivative of a function at a point is a fundamental example of a linear map.

Most functions behave differently in different regions; their graphs will curve. A linear map, however, behaves uniformly.

A *smooth* function changes its behavior gradually, with no abrupt kinks or jumps. This means that the closer we zoom in to the region around a point, the more regular the behavior appears. For very close views, the behavior appears uniform. We recognize this uniformity from the example we have just studied: In the limit near a point, smooth functions preserve relative positions. By declaring this point and its image to be origins, we can use the language of linear maps to describe what we see.

In other words, the derivative answers the question: Which linear map looks everywhere like our function looks near a point?



In the first drawing, the straight coordinate lines in the domain map to curved lines in the image, because f is a nonlinear map. However, the four squares around the marked point map approximately to four parallelograms around the image of the marked point, because the behavior near this point is nearly uniform. In this second drawing, we have zoomed in on these four red squares, and subdivided. The three marked points in the domain maps to the three points shown in the image, and the rest of the map appears to fill in by relative position.

Meditate on this. Physicists like to say that children understand physics best. To be a good mathematician, your urge to play needs to survive your formal education. The physicist Albert Einstein imagined riding a beam of light. The topologist Bill Thurston imagined living inside geometric spaces. The challenge of mathematics is that its objects exist in our imagination.

You need to understand the spaces of linear algebra as if you can grab them by your hands. Twist them; how can they deform? A cardboard box is rigid, but sit on one with both ends open and it collapses. Our spaces are like that box: We can flexibly position reference points such as three points in the plane, but then the remaining points fall rigidly into place by relative position. Understand this, and one will appreciate linear combinations as an effective tool for working with relative position.

So far we have used the plane as an example of an n-dimensional real vector space. One could learn all of linear algebra, working only with such spaces. However, we want to apply linear algebra in other settings. Rather than working with a space of points, we could as easily work with a space of functions. This requires a more general language.

By *map* we often mean a function such as studied in calculus. We sometimes mean a more general transformation, such as the operation of taking the derivative of a function.

A *linear combination* is an expression $rv + sw$ for *scalars* r, s and *vectors* v, w. *Scalars* are often real numbers $r, s \in \mathbb{R}$. *Vectors* are often lists of real numbers $v, w \in \mathbb{R}^n$. However, we will use the language of maps, scalars, and vectors in any setting where we recognize the patterns of linear algebra.

A *linear map* is a map that preserves linear combinations:

$$\underbrace{f}_{\text{map}} : \underbrace{V \to W}_{\text{spaces}} \qquad f(\underbrace{rv+sw}_{\substack{\text{linear combination}\\ \text{first}}}) = r\,\underbrace{f(v)}_{} + s\,\underbrace{f(w)}_{\text{apply function first}}$$

There are two operations here: Taking a linear combination, and applying a map. For a linear map, we can carry out these operations in either order, with the same result.

Let $a, r, s, v, w \in \mathbb{R}$. Then multiplication by a is a linear map. We recognize this as the distributive law:

$$\underbrace{a : \mathbb{R} \to \mathbb{R}}_{\text{multiplication by } a} \qquad a(rv+sw) = rav + saw$$

distribute a

Differentiation is an example of a linear map. Using real numbers $r, s \in \mathbb{R}$ as scalars, and differentiable functions $f, g : \mathbb{R} \to \mathbb{R}$ as vectors, $\frac{d}{dx}$ is a linear operator. We can distribute $\frac{d}{dx}$ exactly as we multiplied by a:

$$\frac{d}{dx}\left(r f(x) + s g(x)\right) = r \frac{d}{dx} f(x) + s \frac{d}{dx} g(x)$$

Similarly, the definite integral is also a linear operator. We use these rules repeatedly in calculus:

$$\int_a^b \left(r f(x) + s g(x)\right) dx = r \int_a^b f(x)\,dx + s \int_a^b g(x)\,dx$$

An integral is the simplest example of a differential equation. Linear algebra is pervasive in the study of differential equations.

A map $f : \mathbb{R}^m \to \mathbb{R}^n$, all of whose terms are degree one, is a linear map. We confirm this for $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$f(x,y) = (ax+by,\ cx+dy)$$

$$(x,y) = r(p,q) + s(t,u) = (rp+st,\ rq+su)$$

$$f(r(p,q)+s(t,u)) \overset{?}{=} rf(p,q) + sf(t,u)$$

$$f(rp+st,\ rq+su) \overset{?}{=} r(ap+bq,\ cp+dq) + s(at+bu,\ ct+dq)$$

$$(a(rp+st)+b(rq+su),\ c(rp+st)+d(rq+su))$$

$$= (rap+rbq+sat+sbu,\ rcp+rdq+sct+sdq)$$

This verification was tedious; we can do better. This is the first of many times that we will be faced with a blizzard of terms. It gets worse; the formula for the determinant of a 70×70 matrix has more terms than the estimated number of elementary particles in the known universe.

We don't think about these terms individually. We organize them into patterns, and reason about these patterns.

There are many elementary particles in a lump of clay on a potter's wheel, yet potters easily manipulate the clay. Perhaps we should be surprised that in linear algebra we would even consider thinking about individual terms? As a potter's clay, linear algebra is remarkably amenable to organizing into patterns. There are only a few kinds of patterns to consider, and we see them everywhere.

$$
\begin{aligned}
f(x,y) &= (ax+by,\ cx+dy)\\
e_{11}(x,y) &= (x,\ 0)\\
e_{12}(x,y) &= (y,\ 0)\\
e_{21}(x,y) &= (0,\ x)\\
e_{22}(x,y) &= (0,\ y)\\
f &= ae_{11}+be_{12}+ce_{21}+de_{22}
\end{aligned}
$$

We can write a degree one map as a linear combination of basic building blocks. Here, we define four helper functions $e_{11}, e_{12}, e_{21}, e_{22}$. They each copy one input coordinate to one output coordinate, doing no arithmetic. Our function f is a linear combination of these helper functions.

It is much easier to verify that these helper functions are linear maps. All they do is copy a coordinate, so the only algebra is that of the linear combination itself, before or after the copying. We see that the order of events doesn't matter; this is the essence of linearity:

$$
\begin{aligned}
e_{11}(x,y) &= (x,\ 0)\\
e_{11}(r(p,q)+s(t,u)) &\stackrel{?}{=} re_{11}(p,q)+se_{11}(t,u)\\
e_{11}(rp+st,\ rq+su) &\stackrel{?}{=} rp + st\\
rp+st &= rp+st
\end{aligned}
$$

A linear combination of linear maps is a linear map:

$$f = ag + bh \qquad u = rv + sw$$

$$\begin{aligned} f(u) = f(rv+sw) &= ag(rv+sw) + bh(rv+sw) \\ &= arg(v) + asg(w) + brh(v) + bsh(w) \\ &= r(ag(v)+bh(v)) + s(ag(w)+bh(w)) \\ &= r f(v) + s f(w) \end{aligned}$$

This computation is a bit involved, but we only need to make it once. We now understand that any degree one map is linear, because it can be written as a linear combination of helper functions that are linear.

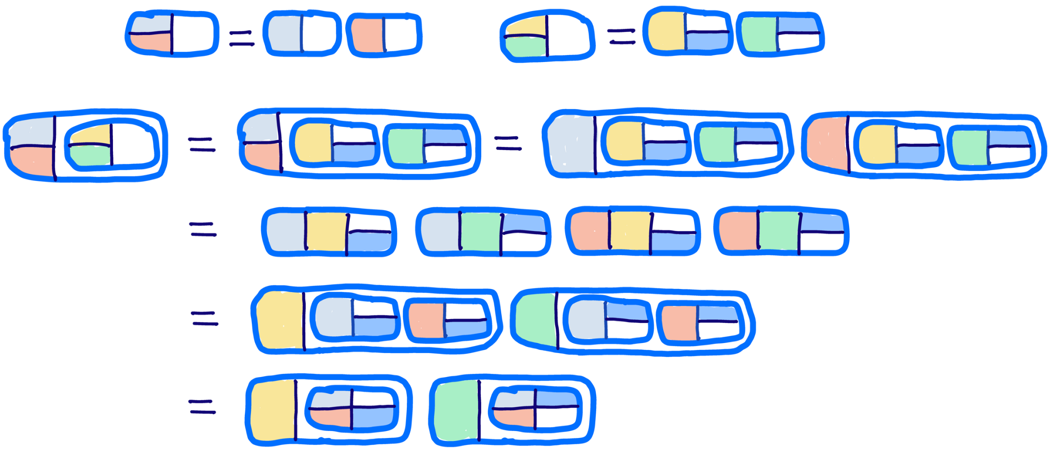
Our first computation for $f : \mathbb{R}^2 \to \mathbb{R}^2$ did not scale well. We could roughly see that it was lots of positioning, and the distributive law over and over again. This would probably work in higher dimensions, but are we sure? By isolating the positioning, and using linear combinations to break the problem into simple pieces, we have much greater confidence that we understand the general case.

Linearity means compatible with linear combinations. Proving linearity is also compatible with linear combinations? This aesthetic coherence is common in many branches of mathematics. Once one gets a feel for the fabric of a subject, one sees the fabric everywhere, including in reasoning about the subject itself. If you find yourself marveling over this, enjoy the moment.

This computation is a kind of Cat's Cradle, where we are passing one linear combination through another. Shouldn't this be easier to look at? Shouldn't this dance be beautiful?

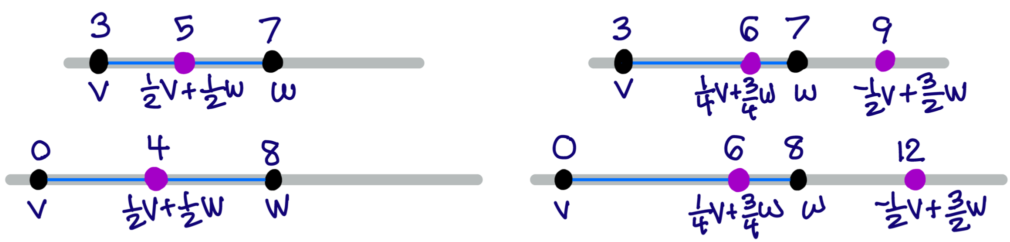
We have expressed this operation in conventional notation. It is important to master conventional notation, but do not make the mistake of believing that everyone thinks this way. Mathematical notation is a convention for formal communication. One doesn't confuse the idea of a computer algorithm with its expression in a particular programming language. One shouldn't make the same mistake here.

The following drawing attempts to capture the dance I see, as one linear combination passes through the other. The act of drawing this helps me to separate the dance from the notation. It's not clear that this drawing will help anyone else to see this dance; we each think in different ways. Instead, reflect on the original computation, and come up with your own notation that makes the dance clear?



1.2 Affine Combinations

The simplest examples of a linear combination are an average, and a weighted average:

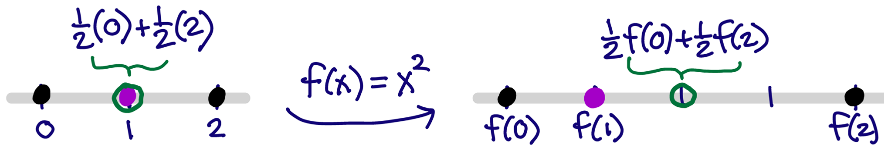


The scalars r, s of a weighted average $rv + sw$ add up to 1. These expressions are called *affine combinations*. A scalar can however be negative; this moves us outside of the interval between v and w.

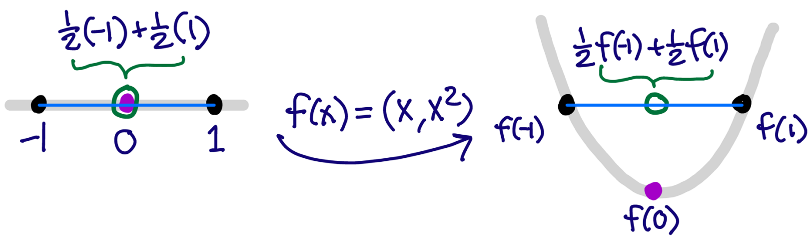
Observe that scale is arbitrary. The numbers shown are uniformly spaced, and we can check our averages using these numbers, but any uniformly spaced numbers will do. Affine combinations don't see absolute distance, they only see proportions relative to the two reference points v and w. We could uniformly stretch the line, and affine combinations wouldn't even notice. Before, we saw algebraically that multiplication by a number a was a linear map. Now, we see geometrically that stretching by a preserves weighted averages. This is the same thing.

Linear maps take the average of two points to the average of their images. An average is a linear combination; this is a special case of linear maps preserving linear combinations. This special case is enough to show us that for a linear map, the image of a line is both uniformly spaced and straight.

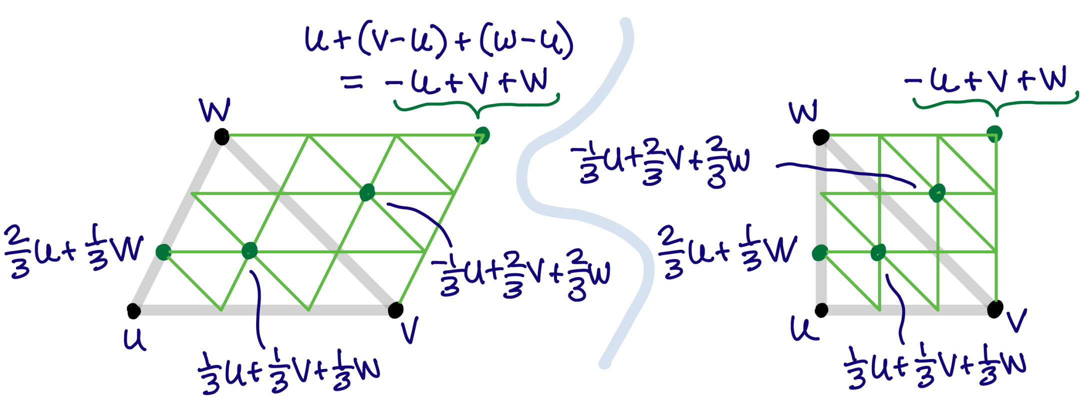
In woodworking we can use a ruler to check either of these features. Linear combinations play the role of a ruler here. For example, the function $f(x) = x^2$ is not linear; it doesn't map the average of 0 and 2 to the average of their images:



Rulers are straight. One can check whether a surface is flat by holding a ruler against it. Here, averages also show us that the image curve of the function $f(x) = (x, x^2)$ is not straight. This f is also not linear; it doesn't map the average of -1 and 1 to the average of their images:



Averages and weighted averages generalize to higher dimensions. Here these affine combinations are also known as *barycentric coordinates*. Again, our scalars add up to 1. In the plane, for example, we are breaking 1 into pieces to divvy up the influence of the three corners of a triangle. In other contexts one sees this idea called a *partition of unity*. Here, if one corner gets all of 1, we're at that corner. If two corners share parts of 1, we're along their side. If all three corners share parts of 1, then we're inside the triangle. The center of the triangle is found by equally dividing 1. Points outside the triangle are reached by letting scalars go negative:



Linear combinations do not see a difference between these two side-by-side drawings. They cannot detect angle. Just as they are oblivious to uniform stretching in a given direction, they cannot detect that different directions are being stretched differently. The plane is only rigid in the sense that position relative to the points u, v, w is rigid. Otherwise, we can warp the plane in ways that preserve these proportions, and linear combinations see the same plane.

Imagine picking up a coin. We know that it's the same coin, as we move it. Our mind separates what is intrinsic to the coin itself from artifacts of our perspective. In math we do the same, in many new settings. Here, what is intrinsic to the plane is what a linear combination sees. The differences in these two drawings are as transient as the differences we see when we twirl a coin with our fingers.

1.3 Matrix Multiplication

$$A\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}a\\b\end{bmatrix} \qquad A\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}c\\d\end{bmatrix} \qquad A = \begin{bmatrix}a & c\\b & d\end{bmatrix}$$

$$\begin{bmatrix}a & c\\b & d\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}ax+cy\\bx+dy\end{bmatrix} \qquad \begin{matrix}(a,c)\cdot(x,y)\\(b,d)\cdot(x,y)\end{matrix}$$

$$f(x,y) = (ax+cy,\, bx+dy) \qquad f : \mathbb{R}^2 \to \mathbb{R}^2$$

$$A\begin{bmatrix}x\\y\end{bmatrix} = A\left(x\begin{bmatrix}1\\0\end{bmatrix} + y\begin{bmatrix}0\\1\end{bmatrix}\right) = xA\begin{bmatrix}1\\0\end{bmatrix} + yA\begin{bmatrix}0\\1\end{bmatrix}$$

$$= x\begin{bmatrix}a\\b\end{bmatrix} + y\begin{bmatrix}c\\d\end{bmatrix} = \begin{bmatrix}ax+cy\\bx+dy\end{bmatrix} = \begin{bmatrix}a & c\\b & d\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}$$

$$\begin{bmatrix} r & t \\ s & u \end{bmatrix}\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} r & t \\ s & u \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} ra+tb \\ sa+ub \end{bmatrix}$$

$$\begin{bmatrix} r & t \\ s & u \end{bmatrix}\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} r & t \\ s & u \end{bmatrix}\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} rc+td \\ sc+ud \end{bmatrix}$$

$$\begin{bmatrix} r & t \\ s & u \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} ra+tb & rc+td \\ sa+ub & sc+ud \end{bmatrix}$$

$$\begin{bmatrix} r & t \\ & \end{bmatrix}\begin{bmatrix} a & \\ b & \end{bmatrix} = \begin{bmatrix} ra+tb & \\ & \end{bmatrix} \qquad (r,t)\cdot(a,b)$$

$$\begin{bmatrix} & \\ s & u \end{bmatrix}\begin{bmatrix} a & \\ b & \end{bmatrix} = \begin{bmatrix} & \\ sa+ub & \end{bmatrix} \qquad (s,u)\cdot(a,b)$$

$$\begin{bmatrix} r & t \\ & \end{bmatrix}\begin{bmatrix} & c \\ & d \end{bmatrix} = \begin{bmatrix} & rc+td \\ & \end{bmatrix} \qquad (r,t)\cdot(c,d)$$

$$\begin{bmatrix} & \\ s & u \end{bmatrix}\begin{bmatrix} & c \\ & d \end{bmatrix} = \begin{bmatrix} & \\ & sc+ud \end{bmatrix} \qquad (s,u)\cdot(c,d)$$

$$f(x,y) = (ax+cy, bx+dy)$$
$$g(x,y) = (rx+ty, sx+uy)$$

$$\begin{aligned} g(f(x,y)) &= g(ax+cy, bx+dy) \\ &= \big(r(ax+cy)+t(bx+dy), s(ax+cy)+u(bx+dy)\big) \\ &= \big((ra+tb)x+(rc+td)y, (sa+ub)x+(sc+ud)y\big) \\ &= \begin{bmatrix} ra+tb & rc+td \\ sa+ub & sc+ud \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r & t \\ s & u \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \end{aligned}$$

Chapter 2

Systems of Equations

2.1 Row Reduction

The first computation one masters in linear algebra is solving a system of equations.

$$\left\{\begin{matrix} 2x+y=1 \\ x+2y=0 \end{matrix}\right\} \qquad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \left[\begin{array}{cc|c} 2 & 1 & 1 \\ 1 & 2 & 0 \end{array}\right]$$

On the left is a system of linear equations in two variables. In the middle, we rewrite this system as a matrix equation. On the right is an *augmented matrix*, isolating the numbers. Our goal is to describe the set of solution points (x, y) for which these equations hold.

There is a standard algorithm for transforming such systems into a form where the solutions become evident: *Gaussian elimination*, also known as *row reduction*. We carry out a sequence of *elementary row operations*, each of which simplifies the system of equations.

$$\begin{matrix} 1* \\ -2* \end{matrix}\left\{\begin{matrix} 2x+y=1 \\ x+2y=0 \end{matrix}\right\} \qquad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \left[\begin{array}{cc|c} 2 & 1 & 1 \\ 1 & 2 & 0 \end{array}\right]$$

$$\cancel{0x} - 3y = 1$$

For example, if we subtract twice the second equation from the first, we obtain a new equation that also holds on any solution to the original system. Why? The new equation is a linear combination of the original equations; we are handling both sides of each equation the same way. If substituting a point (x, y) gives the same value on both sides of each original equation, then it will give the same value on both sides of the new equation. Same numbers in, same numbers out.

We can now leave out the first equation, because we can recover it by adding twice the second equation back to the new equation. In other words, this operation is reversible. We can replace the first equation by this new equation. We have seen that any solution from before this operation is a solution after this operation. Because this operation is reversible, we can reverse this reasoning: Any solution from after this operation is also a solution before this operation. Thus the solution sets are the same before and after. We have found a way to transform the system into an simpler system that is *equivalent* to the original system: It has the same set of solutions.

$$\left\{\begin{aligned} -3y &= 1 \\ x+2y &= 0 \end{aligned}\right\} \qquad \begin{bmatrix} 0 & -3 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \left[\begin{array}{cc|c} 0 & -3 & 1 \\ 1 & 2 & 0 \end{array}\right]$$

$$\textcircled{1} \leftarrow \textcircled{1} - 2\textcircled{2}$$

We use $\textcircled{1} \leftarrow \textcircled{1} - 2\textcircled{2}$ as shorthand for "*subtract twice the second row from the first*". The above shows our system in all three notations, after replacing the first equation by the new equation.

$$\left\{\begin{aligned} x+2y &= 0 \\ -3y &= 1 \end{aligned}\right\} \qquad \begin{bmatrix} 1 & 2 \\ 0 & -3 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & -3 & 1 \end{array}\right]$$

$$\textcircled{1} \leftrightarrows \textcircled{2}$$

We use $\textcircled{1} \leftrightarrows \textcircled{2}$ as shorthand for "*swap the first and second rows*". The new system is again equivalent to the old system; we have changed the order of presentation but not the content. This operation is also reversible: swap again.

$$\left\{\begin{aligned} x+2y &= 0 \\ y &= -\tfrac{1}{3} \end{aligned}\right\} \qquad \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{3} \end{bmatrix} \qquad \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} \end{array}\right]$$

$$\textcircled{2} \leftarrow -\tfrac{1}{3}\textcircled{2}$$

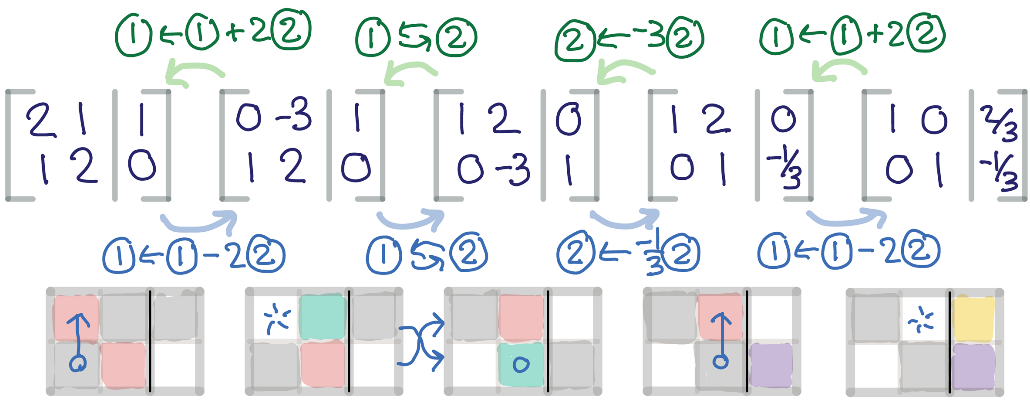
We use $\textcircled{2} \leftarrow -\frac{1}{3}\textcircled{2}$ as shorthand for "*multiply the second row by* $-1/3$". The new system is again equivalent to the old system; rescaling an equation does not change where equality holds, as long as we don't multiply by 0. This operation is also reversible: multiply the second row by -3.

$$\begin{cases} x = 2/3 \\ y = -1/3 \end{cases} \qquad \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}2/3\\ -1/3\end{bmatrix} \qquad \left[\begin{array}{cc|c}1 & 0 & 2/3\\ 0 & 1 & -1/3\end{array}\right]$$

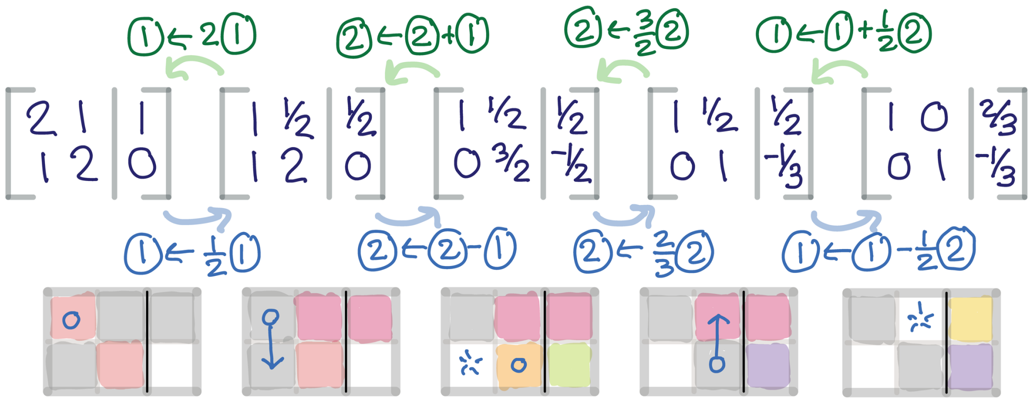
①←①−2②

Finally, we again apply the operation ①←①−2②. Now, on the left we see a system of equations telling us that the solution set is the single point $(x, y) = (2/3, -1/3)$. The notations in the middle and on the right are saying the same thing.

The notation on the right is far less work to write out, and a more realistic portrayal of what a computer actually stores when it carries out these operations. We need to train our minds to see the left when we look at the right. Eventually, like learning a foreign language, we directly see meaning on the right. This takes practice.



The above summarizes what we have done, using augmented matrix notation. The blue steps left to right reduce our system to an identity matrix. The green steps right to left return to the original system, showing that our operations are reversible. The abstract sequence at the bottom is one artist's rendition of what we did, substituting colors for numbers to avoid being distracted by the arithmetic, as if we were playing a game on our phone. Our strategy is independent of most of the arithmetic, and it is good to understand the game plan without worrying about the numbers. What do you see just before your mind lapses into chaos as you fall asleep? You can learn to pay attention deeper into the dive. Draw it. You're likely to derive more satisfaction from sketching this process for yourself than by deciphering my attempt. Take my example as a model that this is reasonable behavior. And send me your sketches.

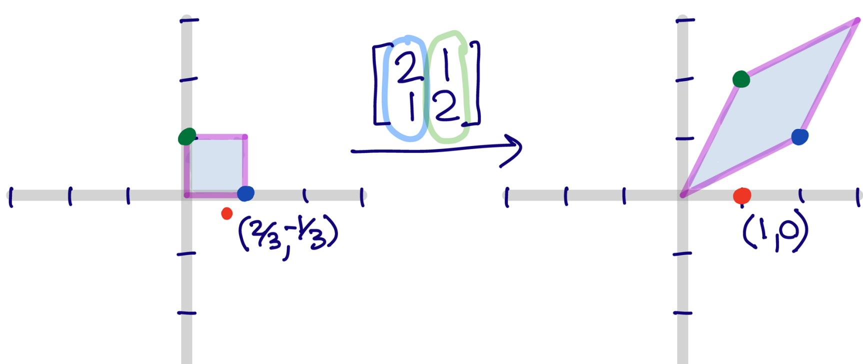


This is a game where we try to reduce a matrix to the identity matrix. There are many ways to proceed. The above shows a different route to the solution. It also happens to take four steps. Because we want to end up with a one in the upper left corner, it seems reasonable to make that entry a one at the start. However, by introducing fractions earlier than necessary, the remaining steps are harder to carry out.

It is often better to work ten problems ten times, than a hundred problems once. One has to explore choices, to learn how to make good choices.

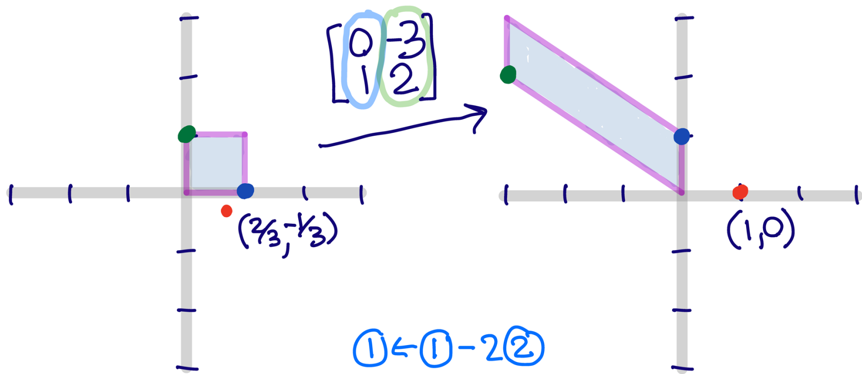
We would like to understand elementary row operations geometrically. Each augmented matrix is shorthand for a matrix equation. Each matrix equation can be visualized as a map between two planes. We would like to look at each of these maps in sequence, as if watching a movie. Can we make sense out of what we see?

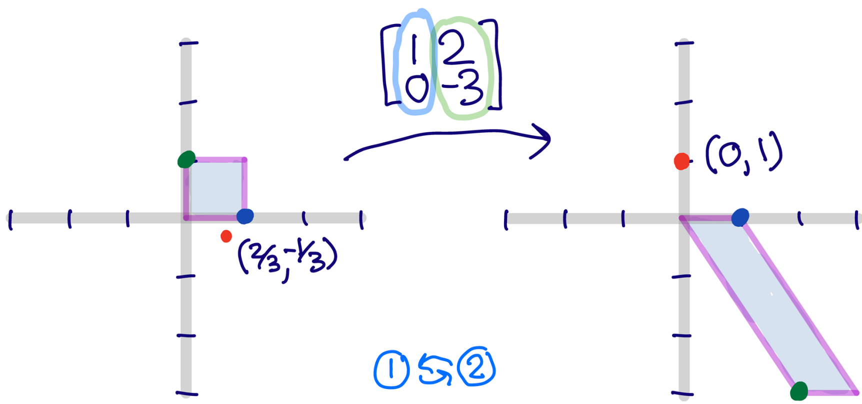
First, let's review how we draw our original matrix equation.

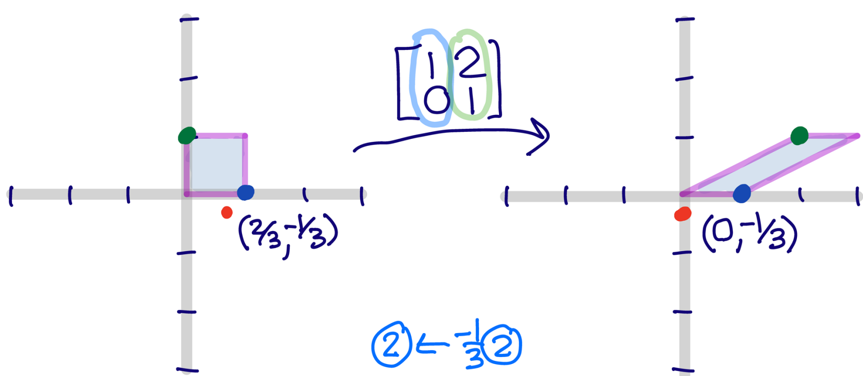


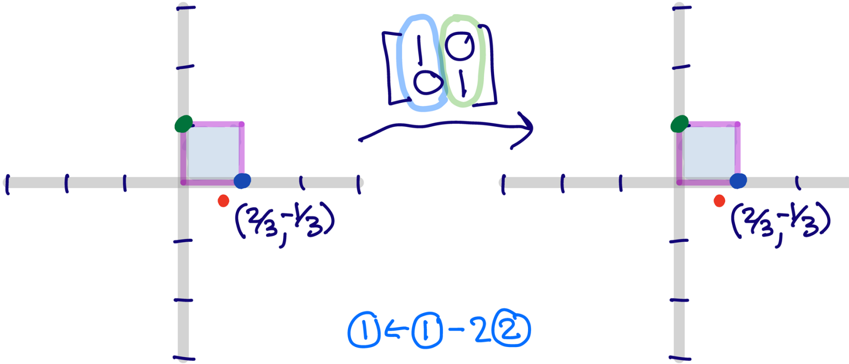
This drawing shows our matrix mapping its domain to its image. The unit square gets stretched to a parallelogram. We want to determine which point in the domain maps to the red point in the image.

We return to our first route to a solution; the numbers are easier. Each of the following drawings is labeled by the elementary row operation that got us here, and illustrates the matrix equation after this step.









This last matrix equation can be solved by inspection, but how did we get here? The domain for each of our drawings is the same, but the image flips around as we row reduce. Each elementary row operation is turning the preceding image into the image we see, but how? These transformations look like linear maps. But what linear maps? Can we think of an elementary row operation as a linear map?

This is how mathematical research often proceeds. We gain experience by working many examples, and we notice a pattern that we recognize. Here, the elementary row operations look like linear maps. We make an adventurous guess that they are linear maps, and set about to figure out the explanation. This is not unlike wishing a professor would go faster in class, and guessing what's next. Sometimes you end up guessing what *should* be next.

To *reify* is make something abstract more concrete or real. Here, we can reify an elementary row operation as left multiplication by an *elementary matrix*.

Here is the first step in our row reduction:

$$\begin{bmatrix}1&0\\0&1\end{bmatrix} \underset{\textcircled{1}\leftarrow\textcircled{1}-2\textcircled{2}}{\longrightarrow} \begin{bmatrix}1&-2\\0&1\end{bmatrix} \qquad \begin{bmatrix}1&-2\\0&1\end{bmatrix}\left[\begin{array}{cc|c}2&1&1\\1&2&0\end{array}\right] \underset{\textcircled{1}\leftarrow\textcircled{1}-2\textcircled{2}}{=} \left[\begin{array}{cc|c}0&-3&1\\1&2&0\end{array}\right]$$

The identity matrix is like a blank slate, or the wax cylinder that Thomas Edison first used to reify sound. By itself, multiplication by the identity matrix has no effect. We can record an elementary row operation by carrying it out on the identity matrix, and play it back by multiplying by the resulting elementary matrix.

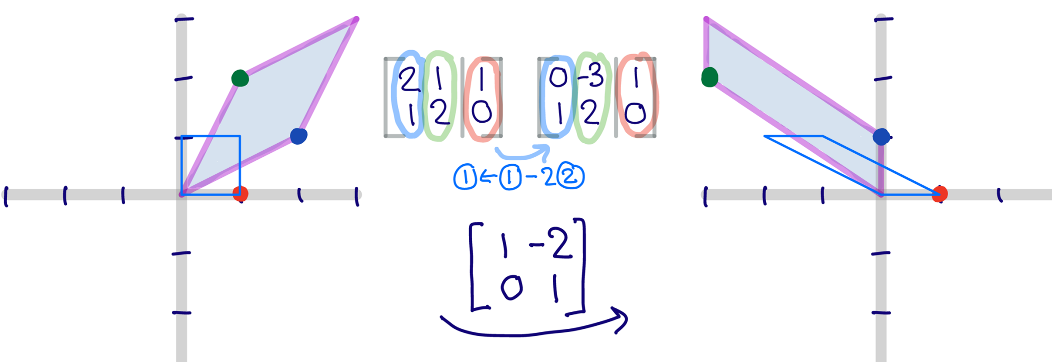
Here are the remaining steps in our row reduction:

$$\begin{bmatrix}1&0\\0&1\end{bmatrix} \xrightarrow[\text{①} \leftrightarrows \text{②}]{} \begin{bmatrix}0&1\\1&0\end{bmatrix} \qquad \begin{bmatrix}0&1\\1&0\end{bmatrix}\left[\begin{array}{cc|c}0&-3&1\\1&2&0\end{array}\right] = \left[\begin{array}{cc|c}1&2&0\\0&-3&1\end{array}\right] \quad \text{①} \leftrightarrows \text{②}$$

$$\begin{bmatrix}1&0\\0&1\end{bmatrix} \xrightarrow[\text{②} \leftarrow -\frac{1}{3}\text{②}]{} \begin{bmatrix}1&0\\0&-\frac{1}{3}\end{bmatrix} \qquad \begin{bmatrix}1&0\\0&-\frac{1}{3}\end{bmatrix}\left[\begin{array}{cc|c}1&2&0\\0&-3&1\end{array}\right] = \left[\begin{array}{cc|c}1&2&0\\0&1&-\frac{1}{3}\end{array}\right] \quad \text{②} \leftarrow -\tfrac{1}{3}\text{②}$$

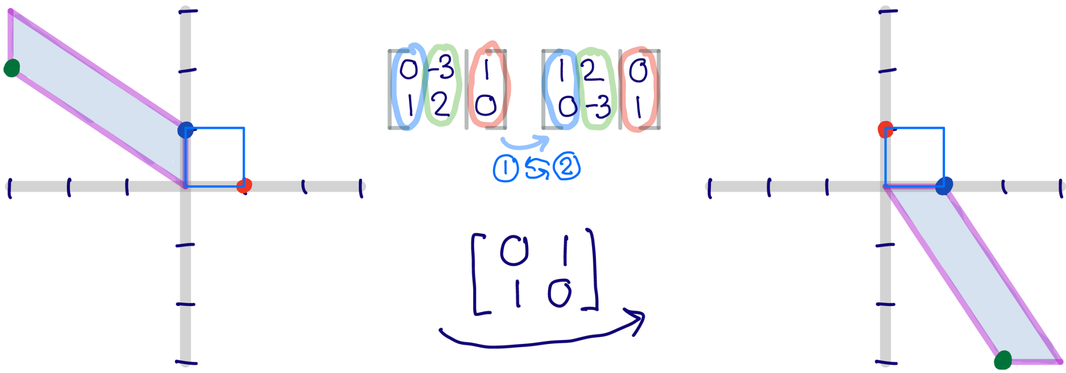
$$\begin{bmatrix}1&0\\0&1\end{bmatrix} \xrightarrow[\text{①} \leftarrow \text{①} - 2\text{②}]{} \begin{bmatrix}1&-2\\0&1\end{bmatrix} \qquad \begin{bmatrix}1&-2\\0&1\end{bmatrix}\left[\begin{array}{cc|c}1&2&0\\0&1&-\frac{1}{3}\end{array}\right] = \left[\begin{array}{cc|c}1&0&\frac{2}{3}\\0&1&-\frac{1}{3}\end{array}\right] \quad \text{①} \leftarrow \text{①} - 2\text{②}$$

These elementary matrices not only transform each augmented matrix into the next, they transform each drawing into the next. As we saw, the domains don't change, so we focus on the images. Here is the first elementary matrix, transforming our original image into the image for the next step:

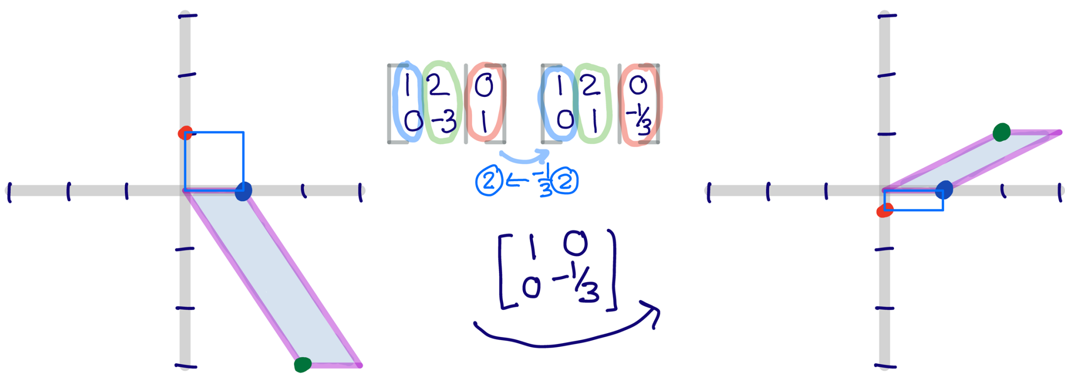


We have added a blue unit square on the left, that transforms into the blue parallelogram on the right. We can confirm that this parallelogram is drawn correctly, by comparing its corners with the columns of the elementary matrix. The rest of the picture matches the columns in the augmented matrices.

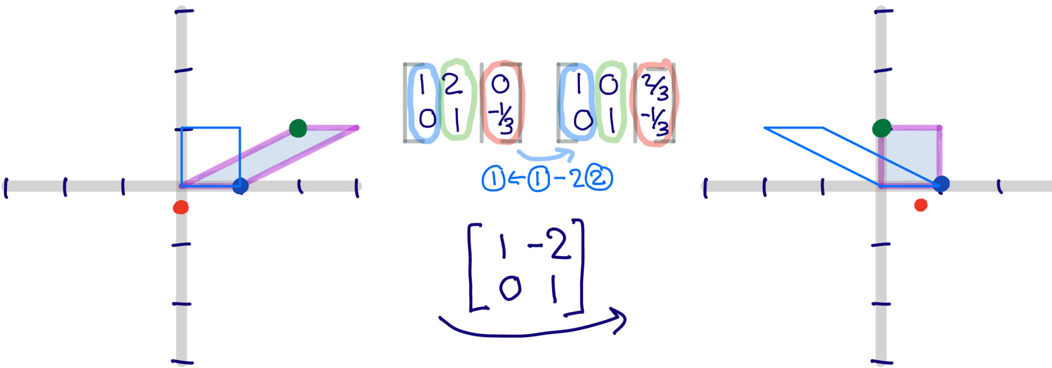
Our second operation swaps the x-axis and y-axis. The entire drawing flips over:



Our third operation stretches the y-axis by a factor of $-1/3$. This shortens the y-axis, and flips its direction:



Finally, we again shear the upper half plane to the left, further as we rise up the y-axis. The lower half plane shears right, moving the red point:



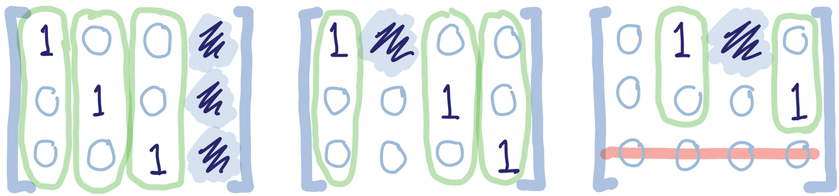
It will rarely be possible to visualize row reduction with this level of detail. Nevertheless, something like this is always going on. We can think of each step as applying a matrix. We can think of each step as changing image coordinates. These are useful lessons.

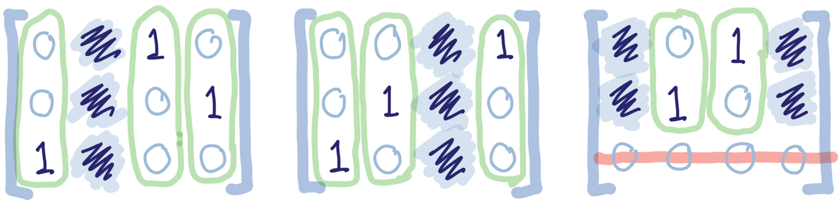
For example, sometimes we look at a matrix, and we want to draw conclusions about the domain or the image. It is tempting to simplify the matrix by row reduction. We can then draw valid conclusions about the domain, but we have completely scrambled the image. It would be a mistake to draw conclusions about the image without taking this scrambling into account. If all you retain from the above sequence of drawings is that row reduction scrambles the image, if this is part of your experience, then you won't make that mistake.

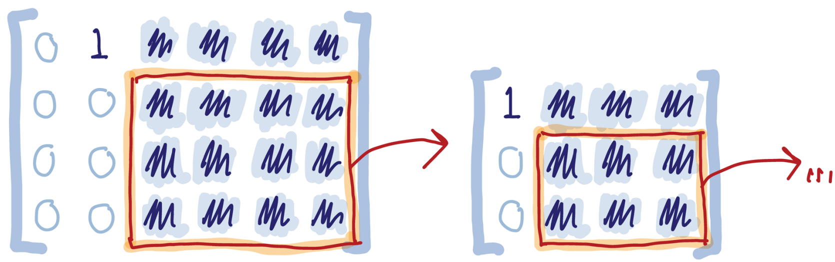
2.2 The General Case

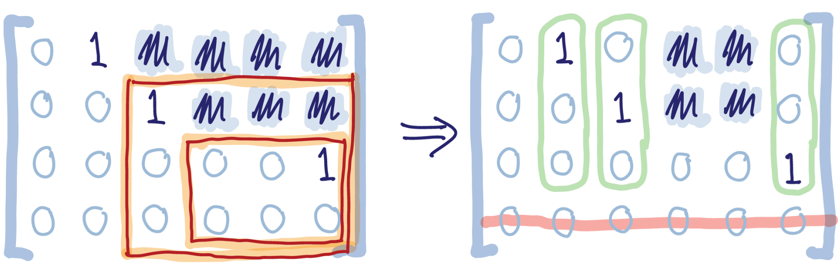
Elementary row operations:

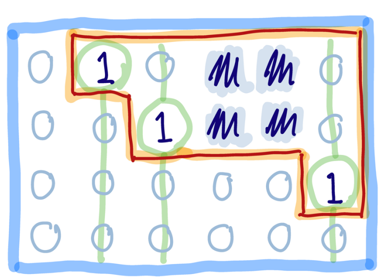
Forward	reverse	
$1 \rightleftarrows 2$	$1 \rightleftarrows 2$	swap two rows
$1 \leftarrow r\,1$	$1 \leftarrow \frac{1}{r}\,1$	rescale a row $(r \neq 0)$
$1 \leftarrow 1 + r\,2$	$1 \leftarrow 1 - r\,2$	add a multiple of another row











$$\begin{bmatrix} 0 & & 1 & 0 \\ 0 & & 0 & 1 \\ 1 & & 0 & 0 \end{bmatrix} : \quad \begin{bmatrix} 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \\ 1 & c & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w + \begin{bmatrix} a \\ b \\ c \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} y + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} z = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} f \qquad\qquad + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} d + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} e = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$$

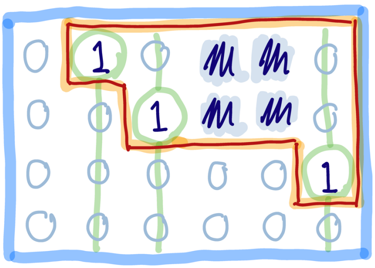
$$\begin{bmatrix} 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \\ 1 & c & 0 & 0 \end{bmatrix} \begin{bmatrix} f \\ 0 \\ d \\ e \end{bmatrix} = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$$

$$\begin{bmatrix} 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \\ 1 & c & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} c + \begin{bmatrix} a \\ b \\ c \end{bmatrix} (-1) + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} a + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \\ 1 & c & 0 & 0 \end{bmatrix} \begin{bmatrix} c \\ -1 \\ a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} f \\ 0 \\ d \\ e \end{bmatrix} + \begin{bmatrix} c \\ -1 \\ a \\ b \end{bmatrix} t$$



:

$$\begin{bmatrix} 0 & 1 & 0 & 2 & 4 & 0 \\ 0 & 0 & 1 & 3 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 7 \\ 8 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 2 & 4 & 0 \\ 0 & 0 & 1 & 3 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 \\ 6 \\ 7 \\ 0 \\ 0 \\ 8 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & -4 \\ 0 & -3 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} r \\ s \\ t \end{bmatrix} \right) = \begin{bmatrix} 6 \\ 7 \\ 8 \\ 0 \end{bmatrix} + 0$$

$$\begin{bmatrix} u \\ v \\ w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ 7 \\ 0 \\ 0 \\ 8 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & -4 \\ 0 & -3 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} r \\ s \\ t \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 2 & 4 & 0 \\ 0 & 0 & 1 & 3 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & -4 \\ 0 & -3 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = 0$$