LINEAR ALGEBRA: SUPPLEMENTARY NOTES

PETE L. CLARK

Contents

0. Some Motivating Examples

Linear algebra is the study of linear systems, matrices, vector spaces and linear transformations. As with most higher mathematics courses, it will take time to present and appreciate these new concepts and objects. Rather than proceeding in a strict logical (linear?) order, I want to begin with some motivational examples: some problems we can solve using linear algebra.

0.1. **Example 1: Stoichiometry.**

In chemical reactions, adding one compound to another may yield one or more new compounds using the component atoms. We write equations to describe these reactions, but they need to be *balanced*. For example, consider

$$N_2 + H_2 \mapsto NH_3.$$

This reaction cannot happen as written: on the left hand side there are 2 nitrogen atoms and 2 hydrogen atoms, whereas on the right hand side there is only 1 nitrogen atom and there are 3 hydrogen atoms. We need to balance the equation by supplying positive whole number coefficients to make the number of atoms of each element on the left hand side equal the number of atoms of each element on the right hand side: say

$$?N_2 + ?H_2 \mapsto ?NH_3.$$

Notice that we have twice as many nitrogen atoms on the left hand side as the right hand side, so why not multiply the N_2 by 1 and the NH_3 by 2:

$$1N_2 + ?H_2 \mapsto 2NH_3.$$

This balances the nitrogen. What about the hydrogen? Well, no problem: we have 6 on the right so we need six on the left, so the last "?" should be 3:

$$N_2 + 3H_2 \implies 2NH_3.$$

Is it always so easy? No, it isn't, as you know if you've taken chemistry. Here is another example, taken from an online chemistry guide to balancing equations.

$$?\,\mathrm{Zn_3\,Sb_2} + ?H_2O \mapsto ?\,\mathrm{Zn}(OH)_2 + ?\,\mathrm{Sb}\,H_3.$$

Suppose we start by balancing the O:

$$?\,\mathrm{Zn_3}\,Sb_2 + 2H_2O \mapsto 1\,\mathrm{Zn}(OH)_2 + ?\,\mathrm{Sb}\,H_3.$$

And suppose we continue by balancing the Zn (zinc): we get stuck, because whatever positive integer we take for the coefficient of $\mathrm{Zn_3\,Sb_2}$ we'll get at least 3 zinc atoms on the left and we only have one on the right. What do we do??

The handout instructs us to start again with an atom that only appears once on each side of the equation, say Sb.[1] This would lead us to

$$1\,\mathrm{Zn_3\,Sb_2} + ?H_2O \mapsto ?\,\mathrm{Zn}(OH)_2 + 2\,\mathrm{Sb}\,H_3.$$

Maybe we try the Zn next? We have to put a coefficient of 3 on the right, getting

$$1\,\mathrm{Zn_3\,Sb_2} + ?H_2O \mapsto 3\,\mathrm{Zn}(OH)_2 + 2\,\mathrm{Sb}\,H_3.$$

Finally, look at the H and the O. If we put a 6 on the left, it works out:

$$1Zn_3Sb_2 + 6H_2O \mapsto 3Zn(OH)_2 + 2SbH_3.$$

Well, that was fortunate. The same handout includes the following hints:

1. **Start with** an atom that appears only once on each side of the equation.
2. **Continue choosing** atoms needing only one new coefficient to become balanced.
3. Balance polyatomic ions as a group.[2]
4. Save hydrogen and oxygen atoms for last.
5. **If you get totally stuck**, try doubling (or tripling, etc.) all the coefficients already entered.

Oh, dear. The last bit suggests that we are being given less than a complete recipe. This stoichiometry business looks like more of an art than a science.

Instead, let's be scientific about it. We only need a small change: instead of repeated question marks, let's write variables for the coefficients, say

$$aZn_3Sb_2 + bH_2O \mapsto cZn(OH)_2 + dSbH_3.$$

Now equating the total number of instances of Zinc on both sides gives

$$3a = c.$$

Equating the total number of instances of Sb on both sides gives

$$2a = d.$$

Equating the total number of instances of H on both sides gives

$$2b = 2c + 3d.$$

Equating the total number of instances of O on both sides gives

$$b = 2c.$$

So we get a system of four equations in four unknowns. If only we knew how to solve such things. Well, we'll learn! Note though that this is a pretty *sparse* system of equations. Many of the coefficients are 0:

$$3a + 0b - c + 0d = 0.$$

$$2a + 0b + 0c - d = 0.$$

$$0a + 2b - 2c - 3d = 0.$$

$$0a + b - 2c + 0d = 0.$$

[1] Chemical trivia question: what element is denoted Sb?!?

[2] I'm not sure exactly what that means, but I wanted to give you all the advice.

Maybe including the zero coefficients looks fastidious. But when we get serious about solving linear systems, we'll see that recopying the variables over and over again is unnecessary and even slightly distracting. So long as we keep the zeros as placeholders, we can just take the coefficients and put them in a rectangular array – a **matrix**:

$$\left[\begin{array}{ccccc} 3 & 0 & -1 & 0 & 0 \\ 2 & 0 & 0 & -1 & 0 \\ 0 & 2 & -2 & -3 & 0 \\ 0 & 1 & -2 & 0 & 0 \end{array}\right].$$

Note that the final column consists entirely of zeros: this is characteristic of **homogeneous linear systems**. Such systems always have a solution: take all variables equal to zero! Here, as usual, we are looking for solutions *other than* the all zero solution.

In fact the zeros in the matrix are not just placeholders but welcome guests. The more zero coefficients in the corresponding matrix, the easier it is to solve the linear system. In this case solving the system certainly doesn't require any special knowledge or ingenuity: two of the variables are simply being given to us in terms of a. Suppose for the sake of argument that $a = 1$. Then we get $c = 3$ and $d = 2$, and using this information we get

$$2b = 2c + 3d = 2 \cdot 3 + 3 \cdot 2 = 12,$$

so

$$b = 6.$$

And now we have one more equation involving b. Luckily it is *consistent* with what we already know:

$$b = 2c = 2 \cdot 3 = 6.$$

Thus

$$(a, b, c, d) = (1, 6, 3, 2)$$

is a solution to the system...exactly the solution we found by hand above. It is not the only solution: no matter what a is we can solve uniquely for b, c and d. In fact we can do this simply by leaving a as is: we get

$$(a, b, c, d) = (a, 6a, 3a, 2a).$$

Notice that this amounts to taking our previous solution and just multiplying it through by a. However the solution with $a = 1$ is the one that the chemists want: the entries need to be positive integers, and we don't want redundancy: mathematically speaking, we don't want all of a, b, c, d to be divisible by any common factor greater than 1.

This simple mathematical analysis is very illuminating. Here are some key points:

I. The entire task is being reduced to solving a system of linear equations. If we know how to do that systematically, balancing equations has no fear for us.

II. We have in fact been given some good advice about how to solve linear systems. In particular, whenever a certain atom appears exactly once on each side, we'll get an equation of the form $\alpha a = \beta b$, where a and b are the variables we're

trying to solve for any α and β are positive integers. This tells us that $b = \frac{\beta}{\alpha}a$, i.e., we've eliminated one of the variables from the system of equations, making it that much easier to solve.

III. It seems to be an implicit assumption that the system is close to having a unique solution: namely it has a unique solution if we require the variables to be positive integers without a common factor. This is much less clear, even if for those who have some knowledge in the solution of linear systems. Note for instance that we have four equations in four unknowns. As we will see later, "most of the time" this type of homogeneous system has only the all zero solution, so our stoichiometric system is somewhat atypical. Neither is it clear that we will always have the same number of variables as equations. In fact, the inspiration to motivate linear systems through stoichiometry came from the course text [SA], which does so on p. 66. However, their example leads to a system of three equations in four unknowns, which as we will learn later, *always* has a solution apart from the all zero solution.

Could we in fact *prove* that the solutions to these stoichiometric systems always have a unique solution in positive integers with no common factor? Or are there chemical reactions that are "stoichiometrically impossible"? This is an interesting question which we'll come back to later.

0.2. **Partial Fractions Decomposition.**

Suppose I want to find an antiderivative of the function $\frac{2x+3}{x^3+x}$. In second semester calculus we learn to do this via the method of **partial fractions**, namely we posit an algebraic identity of the form

$$\frac{x^2+2x+3}{x^3+x} = \frac{A}{x} + \frac{Bx+C}{x^2+1}$$

and try to solve it for real numbers A, B, C. How is this done? Well, if we multiply both sides by x^3+x to clear denominators we get

$$x^2+2x+3 = A(x^2+1) + (Bx+C)x = (A+B)x^2 + Cx + A.$$

Now the polynomial on the left will certainly be equal to the polynomial on the right if they are equal coefficient by coefficient (in fact this is the only way for two polynomials with real numbers as coefficients to be equal, as we will probably have occasion to recall later on), so it is enough to enforce

$$A = 3,$$
$$C = 2,$$
$$A + B = 1.$$

Again we get a linear system to solve! (This time the system is **inhomogeneous**: the right hand sides of the equations are not all zero.) And again it's an easier system than the general case, in this case very easy: clearly $A = 3$ and $C = 2$, which tells us that $B = 1 - A = -2$, thus the desired identity is

$$\frac{x^2+2x+3}{x^3+x} = \frac{3}{x} + \frac{-2x+2}{x^2+1},$$

so

$$\int \frac{x^2+2x+3}{x^3+x} = \int \frac{3}{x} - \int \frac{2x}{x^2+1} + 2\int \frac{1}{x^2+1}$$

$$= 3 \log x - \log(x^2 + 1) + 2 \arctan x + c.$$

Of course this was a relatively benign example: in general, to integrate a proper rational function $\frac{P(x)}{Q(x)}$ when the denominator Q is a polynomial of degree n (i.e., the highest power of x which appears is x^n), then this method gives us an $n \times n$ system of linear equations to solve. It is not always the case that one can solve the system so easily: sometimes there is still nontrivial work to do. In fact, the class of rational functions you are asked to integrate in second semester calculus is limited to those for which solving the corresponding linear systems is sufficiently easy to do without knowledge of the methods of linear algebra.

Again though there is also a theoretical question here: how do we know that the linear system we set up to do a partial fractions decomposition will always have a unique solution? This is the type of question that linear algebra can answer.

0.3. **Polynomial Interpolation.**

It is a well-known adage that "two points determine a line". What this means is that given any two distinct points P_1 and P_2 in the plane, there is exactly one line passing through both of them. If we dismiss vertical lines as being a not especially fascinating degenerate case, then the line is the graph of a function $f : \mathbb{R} \to \mathbb{R}$ and the problem is one of **interpolation**: suppose we have x-coordinates $x_1 < x_2$ and numbers y_1, y_2, and we want to find the unique linear function $\ell = mx + b$ which passes through the points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. What do we do?

Really, we just do it: brute force algebra will work nicely. Namely, writing out the equations $\ell(x_1) = y_1$ and $\ell(x_2) = y_2$ we get

$$y_1 = mx_1 + b,$$

$$y_2 = mx_2 + b.$$

Again this gives us a linear system: this time the two unknowns are m and b. We can solve it, for instance, by subtracting the equatations, giving

$$y_1 - y_2 = m(x_1 - x_2),$$

so – perhaps we could have gotten here faster! –

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

(The denominator cannot be zero since we have assumed $x_1 < x_2$.) Then

$$b = y_1 - mx_1 = y_1 - \left(\frac{y_1 - y_2}{x_1 - x_2}\right) x_1.$$

Now suppose that we are given three points

$$P_1 = (x_1, y_1),\ P_2 = (x_2, y_2),\ P_3 = (x_3, y_3).$$

Of course there will usually be no line that passes through all three points: if at first there is a line passing through all three, then change the value of y_3! However this time there is a unique **quadratic function**

$$f(x) = ax^2 + bx + c$$

such that $f(x_1) = y_1$, $f(x_2) = y_2$, $f(x_3) = y_3$. By plugging everything in we get a linear system:

$$y_1 = f(x_1) = ax_1^2 + bx_1 + c$$

$$y_2 = f(x_2) = ax_2^2 + bx_2 + c$$

$$y_3 = f(x_3) = ax_3^2 + bx_3 + c.$$

(Note that these are not purely linear equations, but they are linear in the unknown variables a, b and c.) For particular $x_1 < x_2 < x_3$ and y_1, y_2, y_3, we can try to solve the linear system: if there is indeed a unique solution, we will find it. However this time we do not have a *sparse* system of three equations and three unknowns: we really have to do some work to solve it. It is another matter entirely to explain why there is always exactly one solution (a, b, c). In fact there are some general theorems along these lines, for instance.

Theorem 0.1. *(Lagrange Interpolation) Let n be a positive integer, let $x_1 < \ldots < x_n < x_{n+1}$ be real numbers, and let $y_1, \ldots, y_n, y_{n+1}$ be real numbers. Then there is exactly one polynomial $P(x)$ of the form $P(x) = a_n x^n + \ldots + a_1 x + a_0$ – i.e., of degree at most n – such that $P(x_1) = y_1$, $P(x_2) = y_2$, ..., $P(x_{n+1}) = y_{n+1}$.*

We will explain how to prove this theorem later on in the course. For now let me notice that there is another polynomial interpolation theorem which is even more familiar. Namely, given an n times differentiable function f defined on an interval containing $c \in \mathbb{R}$, there is a unique polynomial function $T_n(x)$ of degree at most n such that: for all $0 \leq i \leq n$, the ith derivative of T_n at c is equal to the ith derivative of f at c:

$$T_n^{(i)}(c) = f^{(i)}(c).$$

Namely, T_n must be the **degree n Taylor polynomial of f**,

$$T_n(x) = \sum_{i=0}^{n} \frac{f^{(i)}(c)}{i!}(x - c)^i.$$

This **Taylor Interpolation Theorem** can be (and is, say in Math 3100) proved without using linear algebra. In fact one can give a linear-algebra free proof of Lagrange Interpolation: see e.g. [HC, § 12.5].

But what if we want to interpolate *between* Lagrange Interpolation and Taylor Interpretation? For instance, suppose I have a function f, and I want a polynomial $P(x)$ which matches the value of the function and the first two derivatives at 1, the value of the function at 3 and the value of the function and the first three derivatives at 7. If you are patient enough to write all this out you will see that this amounts to $3 + 1 + 4 = 8$ different linear equations on the coefficients of an unknown polynomial. Since a degree n polynomial has $n + 1$ different coefficients, it is plausible that to do this we should look for a polynomial of degree (at most) 8. The **Hermite Interpolation Theorem** says that one can always interpolate in this way by a polynomial of degree at most n, as long as $n + 1$ is at least as large as "the number of conditions" we are imposing. It is a very satisfying generalization of both Lagrange and Taylor interpolation, and in contrast to the above I *only* know how to prove this result using linear algebra. We will do so later in the course.

0.4. **Fibonacci Numbers.**

There is a very famous and ubiquitous sequence of positive integers defined by $F_1 = F_2 = 1$ and for all $n \geq 3$, $F_n = F_{n-1} = F_{n-2}$. In other words,

$$F_3 = F_1 + F_2 = 1 + 1 = 2,$$

$$F_4 = F_2 + F_3 = 1 + 2 = 3,$$

$$F_5 = F_3 + F_4 = 2 + 3 = 5,$$

$$F_6 = F_4 + F_5 = 3 + 5 = 8,$$

and so forth. There are all kinds of amazing identities surrounding the Fibonacci numbers. Here are three:

Theorem 0.2. *(Cassini Identity) For all positive integers n,*

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n.$$

Theorem 0.3. *(Addition Formula) For all positive integers m and n,*

$$F_{m+n} = F_m F_{n+1} + F_{m-1}F_n.$$

Theorem 0.4. *(Binet's Formula) Let $\varphi = \frac{1+\sqrt{5}}{2}$ and $\overline{\varphi} = \frac{1-\sqrt{5}}{2}$. Then for all positive integers n,*

$$F_n = \frac{\varphi^n - \overline{\varphi}^n}{\sqrt{5}}.$$

It is in fact possible to prove all of these identities by induction on n. I have done so when teaching induction in Math 3200. But the kryptonite of mathematical induction is that it does not give you any help with the (often much more difficult) task of *coming up with* the statements you are trying to prove. It turns out that one can not only prove but also *discover* these identities by using the algebraic properties of a certain matrix

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

In particular, as we will see, one can *multiply* any two square matrices of the same size to get another square matrix of the same size. In particular, one can take powers M^k of any square matrix. Then the key to all three of the above identities is the following matrix identity.

Theorem 0.5. *For all positive integers n, we have*

$$M^n = \begin{bmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{bmatrix}.$$

0.5. **Some Differential Equations.**

One of the most important areas of both pure and applied mathematics is differential equations. For instance, Newton's Second Law is $F = ma = mx''$. If the force is given as a function of the position, then we get a second order differential equation whose solutions tell us how the body moves when subjected to the given force: this is mathematical magic. For instance, in the case of a mass suspended on a spring, Hooke's Law says that upon distending the mass a distance of x units

from the equilibrium, the spring pulls back in the opposite direction and with a force which is simply proportional to x: $F = -kx$. This leads to

$$-kx(t) = mx''(t),$$

or

$$x'' = \frac{-m}{k}x.$$

Let us suppose for the sake of simplicity that $m = k$, so we get the equation

$$x'' = -x.$$

This has two fundamental solutions $x_1(t) = \cos t$, $x_2(t) = \sin t$, and the general solution is obtained by linearly combinbing them:

$$x(t) = C_1 \cos t + C_2 \sin t.$$

Note that the differential equaiton $x'' = -x$ implies that

$$x'''' = (x'')'' = (-x)'' = -x'' = -(-x) = x.$$

Thus we are looking for functions which are equal to their own fourth derivative. This larger space has four fundamental solutions

$$e^t, e^{-t}, \cos t, \sin t$$

and the general solution is a linear combination of them. However e^t and e^{-t} are "degenerate" solutions in that they satisfy $x'' = x$ rather than the desired $x'' = -x$.

It turns out that we are very lucky to be able to write down nondegenerate solutions to $x'''' = x$ by pure thought as we did above. Suppose for instance we want a function $x = x(t)$ which is equal to its own third derivative: $x''' = x$. Again there is the "degenerate solution" $x(t) = e^t$, which satisfies $x' = x$. What about a nondegenerate solution? The methods of linear algebra will help with this.

1. Row Operations and Row Echelon Form

We work over the real numbers (but everything we say would be valid for matrices with coefficients in any **field** of scalars). For positive integers m, n we denote by $M_{m,n}$ the set of $m \times n$ matrices.

Two matrices $A, B \in M_{m,n}$ are **row equivalent**, and write $A \sim B$, if we can get from A to B by performing a finite sequence of elementary row operations. If $A \sim B$, then for any column vectors $x = (x_1, \ldots, x_n)$ and $d = (d_1, \ldots, d_m)$, we have $Ax = d$ if and only if $Bx = d$. This is just matrix notation for the fact that the elementary row operations preserve the solution set of a linear system.

Row equivalence is indeed an equivalence relation on $M_{m,n}$, hence it partitions $M_{m,n}$ into equivalence classes. One can motivate the main result of this note by asking for a canonical way of choosing one matrix from each equivalence class.

For $M \in M_{m,n}$, an entry m_{ij} is a **leading entry** if, reading from left to right, it is nonzero and is the first nonzero entry in its row. Thus a row has a leading entry if and only if it is not a *zero row*, and every row has at most one leading entry. Thus the number of leading entries of M is at most m, the number of rows.

A matrix $A \in M_{m,n}$ is in **row echelon form** if

(REF1) Every zero row of A lies below every nonzero row of A, and
(REF2) Every leading entry occurs to the right of every leading entry in the rows above it.

Exercise 2.1:
a) Show that (REF2) is equivalent to: if $1 \leq i_1 < i_2 \leq m$ and $1 \leq j_1, j_2 \leq n$, $a_{i_1 j_1}$ is a leading entry and $a_{i_2 j_2} \neq 0$, then $j_2 > j_1$.
b) Show that (REF2) implies that every entry lying directly below a leading entry is 0. More precisely, if a_{ij} is a leading entry, then for all $i \leq i' \leq m$, $a_{i'j} = 0$.

A matrix $A \in M_{m,n}$ is in **reduced row echelon form** (or **rref**) if it is in row echelon form and moreover

(RREF1) Every entry lying directly above a leading entry is 0 and
(RREF2) Every leading entry is equal to 1.

Proposition 1.1. *Every $A \in M_{m,n}$ is row equivalent to a rref matrix.*

Proof. The proof is constructive: that is, we give an explicit procedure.
Step 1: We use elementary row operations to put $A \in M_{m,n}$ in row echelon form. We begin by looking at the first column.
Case 1: If every entry of the first column is 0, we move on to the $m \times (n-1)$ submatrix A' obtained by removing the first column of A. Any row operations that put A' in row echelon form will also put A in row echelon form (moreover if a matrix has a zero column, then so does any row equivalent matrix).
Case 2: Suppose that some entry of the first column is nonzero.
Case 2a: Suppose $a_{11} \neq 0$. Then by using the type (III) row operation, for all $2 \leq i \leq m$, we multiply row 1 by $\frac{-a_{m1}}{a_{11}}$ and add it to row i, thus making the $(i,1)$ entry equal to zero. Thus we end up with a matrix with $a_{1,1}$ nonzero (thus a leading entry) and $a_{i,1} = 0$ for all $2 \leq i \leq m$. We then proceed inward to the $(m-1) \times (n-1)$ submatrix A' formed by removing the first row and column of A, observing that any sequence of row operations that puts A' in row echelon form does the same for our matrix.[3]
Case 2b: Suppose that $a_{11} = 0$. Since we are in Case 2, there is some i such that $a_{i1} \neq 0$. For the sake of definiteness[4] take the smallest such i and perform the type (I) row operation switching the first and ith rows. This places us back in Case 2a.
We now have a smaller – either $m \times (n-1)$ or $(m-1) \times (n-1)$ – matrix to put in row echelon form, so we can apply the above procedure to this matrix. (In other words, the algorithm is *recursive*: we do something very simple and then allow the algorithm to call on itself for smaller parameter values.)
Step 2: We have now replaced A by a row equivalent matrix which in row echelon form. We may easily go further and put in reduced row echelon form. First, in each row containing a leading entry a_{ij}, we use the type (III) row operation to make

[3]This is straightforward to check but not, I think, immediately obvious. It is also very important...so please do check it.

[4]We do actually want to give an algorithm. An algorithm is not allowed to "make choices": it must do the same thing every time.

all the entries in that column *above* a_{ij} equal to zero just as we did for the entries below to get to row echelon form. (It is worth thinking about why this process necessarily preserves row echelon form: e.g. how do we know we don't produce any zero rows lying above nonzero rows by doing this?) Finally, for every row containing a leading entry a_{ij} we use the type (II) row operation to multiply the ith row by $\frac{1}{a_{ij}}$, which makes the leading entry equal to 1 (and does not change which entries are zero or nonzero so preserves everything we've done so far). □

Exercise 2.2: a) Suppose $A \in M_{m,n}$ has entries in the rational numbers $\mathbb{Q}$ – i.e., numbers of the form $\frac{a}{b}$ with $a, b \in \mathbb{Z}$ and $b \neq 0$. Show that our algorithm produces a row echelon form and then a reduced row echelon form with entries in $\mathbb{Q}$.
b) Suppose $A \in M_{m,n}$ has entries in $\mathbb{Z}$. Show that our algorithm *does not* in general produce a row echelon form or a reduced row echelon form with entries in $\mathbb{Z}$.
c) Show however that a modified algorithm will take any A with entries in $\mathbb{Z}$ and yield a row echelon form with entries in $\mathbb{Z}$. (Hint: when you want to divide by something, multiply a different row by that thing instead.) In fact, show that we can start with any A with entries in $\mathbb{Q}$ and find a row equivalent matrix in row echelon form with entries in $\mathbb{Z}$.
d) Show that if A has entries in $M_{m,n}$, then a modified algorithm will yield a row echelon form which satisfies (RREF1).[5]

Let A be a matrix in row echelon form. Then the variables corresponding to the columns which contain leading entries are called **pivot variables**, whereas the variables corresponding to the other columns are called **free variables**.

Theorem 1.2. *The reduced row echelon form is unique. More precisely, for each $A \in M_{m,n}$, there is exactly one matrix $R \in M_{m,n}$ with $A \sim R$ and R in reduced row echelon form.*

Proof. We follow T. Yuster [Y] by inducting on n, the number of columns.

Base Case ($n = 1$): Suppose A has only one column. If A is the all zero matrix, it is row equivalent only to itself and is in reduced row echelon form. Every nonzero matrix with one column has a nonzero entry, and all such matrices have reduced row echelon form the column vector $(1, 0, \ldots, 0)$ and no other row echelon form.

Induction Step: Suppose now that $n > 1$, that the result holds for all $m \times n$ matrices, and let $A \in M_{m,n+1}$. For any $M \in M_{m,n+1}$, we let $M' \in M_{m,n}$ be obtained by removing the last column from M. Let B and C be reduced row echelon forms of A. Here is the key observation: the matrices B' and C' are in reduced row echelon form and row equivalent to A'.
By induction, we have $B' = C'$. In other words, we know that the reduced row echelon matrices B and C are equal except possibly in the last column. Seeking a contradiction we suppose that their last columns are not equal: i.e., there is some $1 \leq i \leq m$ such that $b_{i,n+1} \neq c_{i,n+1}$. Now let $x = (x_1, \ldots, x_{n+1})$ be any vector with $Bx = 0$, i.e., a solution of the associated homogeneous linear system. Because B and C are row equivalent, x is also a solution to the homogeneous system $Cx = 0$. It follows that $(B - C)x = 0$. Since the matrix $B - C$ is zero except in its last column, performing the multiplication of the ith row of $B - C$ by x simply gives $(b_{i,n+1} - c_{i,n+1})x_{n+1} = 0$. Since $b_{i,n+1} \neq c_{i,n+1}$ we deduce that $x_{n+1} = 0$. Thus

[5]Perhaps we should call this **almost reduced row echelon form?**

x_{n+1} is not a free variable for either B or C, so in each of these matrices the last column must contain a leading entry of 1 and have all the other entries 0. Moreover, in both B and C the 1 must lie in the first zero row of B' and C'. Thus $B = C$. □

The uniqueness of reduced row echelon form has several important consequences. For now we point the following one.

Corollary 1.3. *Let $A \in M_{m,n}$, and let B and C be row equivalent matrices each in row echelon form. Then the pivot variables with respect to the matrix B are the same as the pivot variables with respect to the matrix C.*

Proof. We gave an algorithm to take the matrix B and put it in *reduced* row echelon form. At every step this algorithm preserves the positions of the leading entries, so it preserves pivot variables. Thus the pivot variables with respect to B are the same as the pivot variables for some reduced row echelon form matrix R_B which is row equivalent to A. Similarly, the pivot variables with respect to C are the same as the pivot variables for some reduced row echelon form matrix R_C which is row equivalent to A. But by Theorem 1.2, $R_B = R_C$, and thus the pivot variables with respect to B are the same as the pivot variables with respect o C. □

This allows us to make the following important definition. For a matrix $A \in M_{m,n}$, we define the **rank of A** to be the number of pivot variables in any row echelon form of A and the **nullity of A** to be the number of free variables in any row echelon form of A. Corollary 1.3 ensures that this is "well-defined", i.e., independent of the row echelon form chosen. The following result follows immediately but, when translated into other contexts, is in fact important and quite powerful.

Theorem 1.4. *(Rank-Nullity Theorem, Version 1) For any $A \in M_{m,n}$ we have*

$$\operatorname{rank}(A) + \operatorname{nullity}(A) = n.$$

Exercise 2.3: Prove the Rank-Nullity Theorem. (Hint: it's very easy!)

Exercise 2.4: Find all rref matrices $R \in M_{2,2}$. (Hint: what are the possibilities for the first column?)

We denote by I_n the $n \times n$ matrix with (i,i) entry equal to 1 for all 1 and other entries 0. Notice that I_n is in reduced row echelon form, and every column contains a leading entry. Thus it has n pivot variables and 0 free variables, so it has rank n. Conversely, it is easy to see that I_n is the only reduced row echelon form matrix in $M_{n,n}$ in which every column contains a leading entry.

Theorem 1.5. *For $A \in M_{n,n}$, the following are equivalent:*
(i) $\operatorname{rref}(A) = I_n$.
(ii) For all $b \in \mathbb{R}^n$, there is a unique $x \in \mathbb{R}^n$ such that $Ax = b$.
(iii) For all $x \in \mathbb{R}^n$, if $Ax = 0$, then $x = 0$.
(iv) $\operatorname{rank} A = n$.

Proof. (i) $\implies$ (ii): If $\operatorname{rref} A = I_n$, then for any $b \in \mathbb{R}^n$, to solve $Ax = b$ we put the augmented matrix $[A \mid b]$ in reduced row echelon form, getting $[\operatorname{rref}(A) \mid b'] = [I_n \mid b']$. (Here b' is whatever get by starting with b and doing the row reduction process.) In terms of equations this reads $I_n x = b'$, i.e., $x = b'$. So there is always a unique solution.

(ii) $\implies$ (iii): Let $b = 0$ can always solve the homogeneous system $Ax = 0$ by taking $x = 0$. Since we are assuming the solution is *unique*, we must not have any other solutions: if $Ax = 0$, then $x = 0$.
(iii) $\implies$ (iv): The number of parameters of the solution space to $Ax = 0$ is equal to the number of free variables. So if we have only the trivial solution to $Ax = 0$, we have no free variables and therefore all n variables are pivot variables: $\operatorname{rank} A = n$.
(iv) $\implies$ (i): Since $\operatorname{rank} A = n$ and n is the number of columns, $\operatorname{rref}(A)$ has n leading entries. For an $n \times n$ matrix in reduced row echelon form to have n leading entries, it must be I_n. (If you are doubtful, imagine the leading entry in the first row occurs anywhere to the right of the first column: you'll quickly see that you cannot then get $n - 1$ further leading entries. Now move on to the second row; to be in row echelon form the leading entry cannot be any sooner than the second column; if it were later, then we cannot get $n - 2$ further leading entries. And so forth.) □

2. Matrices and Linear Transformations

2.1. **Review of Composition of Functions.**

Let $f : X \to Y$ and $g : Y \to Z$. We may **compose** the functions to get

$$g \circ f : X \to Z,\ x \mapsto g(f(x)).$$

Note the somewhat confusing fact that $g \circ f$ means "first perform f, then g". This can be traced to the fact that we evaluate functions *on the right*, i.e., we write $f(x)$. If we were willing to evaluate functions *on the left* – i.e., to write $(x)f$ instead – then composition would behave less confusingly. I have seen function evaluaiton written on the left in some old textbooks, but not within the last thirty years or so (and I don't advocate it myself). It seems that we are stuck with things the way they are.

For any set A, let X denote the set of all functions $f : A \to A$. Then composition of $f, g \in X$ is always defined and gives a binary composition law on X. The composition of functions on a fixed set is surely the most important example of a binary composition law on a set, and many other important laws reduce to it.

Example 2.1. *Given $f, g : A \to A$, we need* not *have $g \circ f = f \circ g$. That is, composition of functions is not generally* ***commutative****. One learns some form of this in elementary school when one is taught the order of operations. In general, the order in which procedures are performed may affect the outcome! For a simple example, suppose $f, g : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ and $g(x) = x + 1$. Then*

$$g \circ f : x \mapsto g(f(x)) = g(x^2) = x^2 + 1,$$

while

$$f \circ g : x \mapsto f(g(x)) = f(x + 1) = (x + 1)^2 = x^2 + 2x + 1.$$

Since for all $x \neq 0$, $x^2 + 2x + 1 > x^2 + 1$, $g \circ f \neq f \circ g$.

Exercise 3.1: Our above example of non-commutativity of function composition used the infinite set of real numbers. What is the *smallest* set A which admits functions $f, g : A \to A$ such that $g \circ f \neq f \circ g$? E.g. can you find such functions with $A = \{1, 2, 3, 4\}$? What about with a smaller set A?

Theorem 2.2. *Composition of functions is associative (when defined). That is, if $f : X \to Y$, $g : Y \to Z$ and $h : Z \to W$, then*

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

Proof. Sometimes the way to show that two things are equal is simply to write out both of them and see that we get the same thing. This is one of those times. The function $(h \circ g) \circ f$ is the function which takes

$$x \mapsto ((h \circ g) \circ f)(x) = (h \circ g)(f(x)) = h(g(f(x)),$$

while the function $h \circ (g \circ f)$ is the function which takes

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

No problem! □

For any set X, we denote the 1_X the function which maps each $x \in X$ to itself. This is called the **identity function**.

Portions of the following result ought to be familiar from a previous course, but we will provide a complete proof anyway.

Proposition 2.3. *Let X and Y be nonempty sets; consider a function $f : X \to Y$.*
a) If X is nonempty, the following are equivalent:
(i) There is a function $g : Y \to X$ such that $g \circ f = 1_X$.
*(ii) f is **injective**: for all $x_1, x_2 \in X$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.*
b) The following are equivalent:
(i) There is a function $g : Y \to X$ such that $f \circ g = 1_Y$.
*(ii) f is **surjective**: for all $y \in Y$, there is $x \in X$ such that $f(x) = y$.*
c) The following are equivalent:
(i) There is a function $g : Y \to X$ such that $g \circ f = 1_X$ and $f \circ g = 1_Y$.
*(ii) f is **bijective**.*

Proof. a) Suppose that $g \circ f = 1_X$, and let $x_1, x_2 \in X$ be such that $f(x_1) = f(x_2)$. Applying g we get $x_1 = g(f(x_1)) = g(f(x_2)) = x_2$. So f is injective. Conversely, suppose that f is injective. We have several choices for $g : Y \to X$. For each $y \in Y$ which is of the form $f(x)$ for some $x \in X$, we put $g(y) = x$: this makes sense because, since f is injective, if $y = f(x)$, then it is of this form for exactly one $x \in X$. Fix an element $x_0 \in X$ (here we use $X \neq \varnothing$); if y is *not* of the form $f(x)$ for any $x \in X$, we put $g(y) = x_0$. The point is that this latter definition doesn't matter: for all $x \in X$ we have $g(f(x)) = x$, which is what we wanted.
b) Suppose that $f \circ g = 1_Y$, and let $y \in Y$. Then $f(g(y)) = y$. Thus every element of y is mapped to by some element of X: f is surjective. Conversely, suppose that f is surjective. Then for each $y \in Y$ there is at least one $x \in X$ with $f(x) = y$. We choose *any* such x and put $g(y) = x$. Then for all $y \in Y$, $f(g(y)) = f(x) = y$.
c) We simply combine parts a) and b). □

Exercise 3.2 (for nullologists[6] only.)
a) Observe that the nonemptiness of Y was never used.
b) If $X = \varnothing$, show that a(ii) always holds, whereas a(i) holds iff $Y = \varnothing$.
c) Show that part c) holds even if $X = \varnothing$.

[6]Nullology: the study of the empty set. C.f. the sound of one hand clapping.

Example 2.4. *Let X be the set of all functions $f : \mathbb{R} \to \mathbb{R}$, with binary composition law the usual composition of functions. In this case there are certainly functions which are injective but not surjective – e.g. $f(x) = \arctan x$ – as well as functions which are surjective but not injective – e.g. $g(x) = x \sin x$. Thus it is certainly possible for an element to have a left inverse but no right inverse, or conversely.*

2.2. **Linear Transformations.**

A **linear transformation** $L : \mathbb{R}^n \to \mathbb{R}^m$ is a function which satisfies:

(LT1) For all $v, w \in \mathbb{R}^n$, $L(v + w) = L(v) + L(w)$.
(LT2) For all $\alpha \in \mathbb{R}$ and $v \in \mathbb{R}^n$, $L(\alpha v) = \alpha L(v)$.

Proposition 2.5. *Let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a function.*
a) If L is a linear transformation and $k \geq 2$, then for $v_1, \ldots, v_k \in \mathbb{R}^n$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, we have

$$L(\alpha_1 v_1 + \ldots + \alpha_n v_n) = \alpha_1 L(v_1) + \ldots + \alpha_n L(v_n).$$

*b) (**One-Step Linear Transformation Test**): Suppose that for all $\alpha \in \mathbb{R}$ and $v, w \in \mathbb{R}^n$, $L(\alpha v + w) = \alpha L(v) + L(w)$. Then L is a linear transformation.*

Proof. a) This is a clasic example of a "binary property" extending immediately to an n-ary property: c.f. [Cl-I, § 7]. To get a formal proof, we go by induction on k. Base Case ($k = 2$): Applying (LT1) and then (LT2) twice we get

$$L(\alpha_1 v_1 + \alpha_2 v_2) = L(\alpha_1 v_1) + L(\alpha_2 v_2) = \alpha_1 L(v_1) + \alpha_2 L(v_2).$$

Induction Step: Suppose the result holds for $k \geq 2$. Then

$$L(\alpha_1 v_1 + \ldots + \alpha_{k+1} v_{k+1}) = L((\alpha_1 v_1 + \ldots + \alpha_k v_k) + \alpha_{k+1} v_{k+1})$$

$$= L(\alpha_1 v_1 + \ldots + \alpha_k v_k) + L(\alpha_{k+1} v_{k+1}) \stackrel{\text{IH}}{=} \alpha_1 L(v_1) + \ldots + \alpha_k L(v_k) + \alpha_{k+1} L(v_{k+1}).$$

b) Taking $\alpha = 1$ we get (LT1). Taking $w = 0$ we get (LT2). □

Remark: Don't take Proposition 2.5b) too seriously. Really it is just a way of collecting two easy things together so that we can call it one easy thing. In fact I think it saves more space in writing than it does time in thinking, so although I will use it below when proving that maps are linear transformations, when you are asked to think about whether a map is a linear transformation you may as well think in terms of (LT1) and (LT2) separately.

Lemma 2.6. *If $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then $L(0) = 0$.*

Proof. We have $L(0) = L(0 + 0) = L(0) + L(0)$. Subtracting the vector $L(0)$ from both sides yields $L(0) = 0$. □

Remark: Our statement of Lemma 2.6 is slightly sloppy: the zero on the left hand side is the zero vector in $\mathbb{R}^n$, whereas the zero vector on the right hand side is the zero vector in $\mathbb{R}^m$. In principle we should distinguish them notationally, perhaps by writing 0_m and 0_n. But in practice this adds complication without clarity.

In a way our definition of linear transformation is overly abstract and fancy. I claim that a linear transformation is really just a vector of linear functions with

zero constant terms. In other words, for $1 \leq i \leq m$ and $1 \leq j \leq n$, let a_{ij} be a real number. We define the function $L : \mathbb{R}^n \to \mathbb{R}^m$ by

$$L(x_1, \ldots, x_n) = (a_{11}x_1 + \ldots + a_{1n}x_n, \ldots, a_{m1}x_1 + \ldots + a_{mn}x_n). \tag{1}$$

Proposition 2.7. *The function L defined by (1) is a linear transformation.*

Proof. Let $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ and $\alpha \in \mathbb{R}$. Using the One-Step Linear Transformation Test,

$$L(\alpha x + y) = L(\alpha x_1 + y_1, \ldots, \alpha x_n + y_n)$$

$$= (a_{11}(\alpha x_1 + y_1) + \ldots + a_{1n}(\alpha x_1 + y_1), \ldots, a_{m1}(\alpha x_1 + y_1) + \ldots + a_{mn}(\alpha x_n + y_n))$$

$$= \alpha((a_{11}x_1 + \ldots + a_{1n}x_n, \ldots, a_{m1}x_1 + \ldots + a_{mn}x_n))$$

$$+(a_{11}y_1 + \ldots + a_{1n}y_n, \ldots, a_{m1}y_1 + \ldots + a_{mn}y_n)$$

$$= \alpha L(x) + L(y).$$

□

The converse also holds: every linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ is of the form (1) for some numbers a_{ij}. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. First, like any function with values in a product, we may write $L = (L_1(x), \ldots, L_m(x))$. Then L is a linear transformation if and only if each component function $L_i : \mathbb{R}^n \to \mathbb{R}$ is a linear transformation. (We have tried to provide complete details for this basic but important material, but we cannot think of any way to write this claim out that is any more convincing than if you just think about it for a moment. Please do so.) Thus we have reduced to the case $m = 1$ and must show that any linear transformation $L : \mathbb{R}^n \to \mathbb{R}$ is of the form

$$L(x) = a_1x_1 + \ldots + a_nx_n.$$

To this end, let $e_j = (0, 0, \ldots, 1, \ldots, 0)$ be the vector with a 1 in the jth component and all other components zero. Then

$$L(x) = L(x_1e_1 + \ldots + x_ne_n) = x_1L(e_1) + \ldots + x_nL(e_n).$$

Since $L(e_1), \ldots, L(e_n)$ are just real numbers, we may call them $a_1, \ldots, a_n$, and then

$$L(x) = a_1x_1 + \ldots + a_nx_n.$$

Thus a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ amounts precisely to a vector of m linear functions with zero constant terms.

There is another way to view an arbitrary linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$. For $1 \leq i \leq m$ and $1 \leq j \leq n$, we let a_{ij} be the ith component of $L(e_i)$, so that

$$L(x_1, \ldots, x_n) = (a_{11}x_1 + \ldots + a_{1n}x_n, a_{21}x_1 + \ldots + a_{2n}x_n, \ldots, a_{mn}x_1 + \ldots + a_{mn}x_n)$$

as above. Let $A \in M_{m,n}$ be the matrix with $(A)_{ij} = a_{ij}$. Then

$$L(x_1, \ldots, x_n) = Ax,$$

where in the above equation we regard x as an $n \times 1$ column vector. In summary:

Theorem 2.8. *For any $m, n \in \mathbb{Z}^+$, the following are equivalent:*
(i) A vector of m linear expressions in $x_1, \ldots, x_n$ with zero constant terms.
(ii) A linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$.
(iii) A matrix $A \in M_{m,n}$.

Let us now consider composition of linear transformations. If $L_1 : \mathbb{R}^n \to \mathbb{R}^m$ and $L_2 : \mathbb{R}^p \to \mathbb{R}^n$, then the composition

$$L_1 \circ L_2 : \mathbb{R}^p \to \mathbb{R}^m, x \mapsto L_1(L_2(x))$$

is defined.

Lemma 2.9. *If $L_1 : \mathbb{R}^n \to \mathbb{R}^m$ and $L_2 : \mathbb{R}^p \to \mathbb{R}^n$ are linear transformations, then their composition $L_1 \circ L_2 : \mathbb{R}^p \to \mathbb{R}^m$ is a linear transformation.*

Proof. Let me give away a secret: if a map is indeed a linear transformation, checking that it is is almost always trivial. We use the One-Step Linear Transformation Test and follow our noses: if $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, then

$$(L_1 \circ L_2)(\alpha x + y) = L_1(L_2(\alpha x + y)) = L_1(\alpha L_2(x) + L_2(y))$$

$$= L_1(\alpha L_2(x)) + L_1(L_2(y)) = \alpha L_1(L_2(x)) + L_1(L_2(y)) = \alpha(L_1 \circ L_2)(x) + (L_1 \circ L_2)(y).$$

□

Lemma 2.10. *For a linear $L : \mathbb{R}^n \to \mathbb{R}^m$, the following are equivalent:*
(i) L is injective.
(ii) If $L(v) = 0$, then $v = 0$.

Proof. (i) $\implies$ (ii): By Lemma 2.6 $L(0) = 0$, so if $L(v) = 0$ then $0 = L(v) = L(0)$. By injectivity, $v = 0$.
(ii) $\implies$ (i): Suppose $v, w \in \mathbb{R}^n$ are such that $L(v) = L(w)$. Then $0 = L(v) - L(w) = L(v - w)$. By hypothesis, this implies $v - w = 0$, so $v = w$. □

Lemma 2.11. *Let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a bijective linear transformation, so L admits an inverse function $g : \mathbb{R}^m \to \mathbb{R}^n$: that is, for all $x \in \mathbb{R}^n$, $g(L(x)) = x$, and for all $y \in \mathbb{R}^m$, $L(g(y)) = y$. Then g is a linear transformation.*

Proof. Note first that L is injective: if $x_1, x_2 \in \mathbb{R}^n$, then $L(x_1) = L(x_2)$ implies $x_1 = g(L(x_1)) = g(L(x_20) = x_2$. Thus, let $y_1, y_2 \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. To show that $g(\alpha y_1 + y_2) = \alpha g(y_1) + g(y_2)$, since L is injective, it suffices to show this equality after applying L, i.e., that $L(g(\alpha y_1 + y_2)) = L(\alpha g(y_1) + g(y_2))$. But since L is linear we have

$$L(g(\alpha y_1 + y_2)) = \alpha y_1 + y_2 = \alpha L(g(y_1)) + L(g(y_2)) = L(\alpha g(y_1) + g(y_2)).$$

□

On the matrix side we can represent L_1 as multiplication by a matrix $M_1 \in M_{m,n}$ and L_2 as multiplication by a matrix $M_2 \in M_{n,p}$, and then $L_1 \circ L_2$, being a linear transformation from $\mathbb{R}^p$ to $\mathbb{R}^m$, must be represented by some matrix M_3 in $M_{m,p}$. A natural question to ask is how the entries of this "composite matrix" M_3 depend on the entries of M_1 and M_2. The answer to this question will lead us to define the fundamental operation of **matrix multiplication**.

Let $M_1 = (a_{ij})$ and $M_2 = (b_{jk})$. Then for $x = (x_1, \ldots, x_p)$,

$$(L_1 \circ L_2)(x) = M_1(M_2 x) = M_1 \begin{pmatrix} b_{11}x_1 + \ldots + b_{1p}x_p \\ \vdots \\ b_{n1}x_1 + \ldots + b_{np}x_p \end{pmatrix} =$$

$$\begin{pmatrix} a_{11}(b_{11}x_1 + \ldots + b_{1p}x_p) + a_{12}(b_{21}x_1 + \ldots + b_{2p}x_p) + \ldots + a_{1n}(b_{n1}x_1 + \ldots + b_{np}x_p) \\ \vdots \\ a_{m1}((b_{11}x_1 + \ldots + b_{1p}x_p) + a_{m2}(b_{21}x_1 + \ldots + b_{2p}x_p) + \ldots + a_{mn}(b_{mn}x_1 + \ldots + b_{np}x_p) \end{pmatrix}$$

$$= \begin{pmatrix} (a_{11}b_{11} + a_{12}b_{21} + \ldots + a_{1n}b_{n1})x_1 + \ldots + (a_{11}b_{1p} + \ldots + a_{1n}b_{np})x_p \\ \vdots \\ (a_{m1}b_{11} + \ldots + a_{mn}b_{1p})x_1 + \ldots + (a_{m1}b_{1p} + \ldots + a_{mn}b_{np})x_p \end{pmatrix} = M_3 x,$$

where M_3 is the $m \times p$ matrix whose (i, j) entry is the dot product

$$(a_{i1}, \ldots, a_{in}) \cdot (b_{1j}, \ldots, b_{nj}),$$

i.e., the dot product of the **ith row** of M_1 with the **jth column** of M_2.

This motivates the following definition: if $A \in M_{m,n}$ and $B \in M_{n,p}$, we define the **matrix product** $AB \in M_{m,p}$ to be the matrix with (i, j) entry the dot product of the ith row of A with the jth column of B. Notice that in order for this definition to make sense we need these vector to have the same number of components; the number of components of the first vector is the number of columns of A, and the number of components of the second vector is the number of rows of B: in our setup both are equal to n, so this makes sense.

Exercise 3.3: Let $A \in M_{m,n}$ and $B \in M_{n,p}$.
a) Suppose that for some $1 \leq i \leq m$, the ith row of A is zero. Show that the ith row of AB is zero.
b) Suppose that for some $1 \leq j \leq n$, the jth column of B is zero. Show that the jth column of AB is zero.

2.3. **Matrix Products as Linear Combinations of Rows and Columns.**

There is an alternate interpretation of the matrix product AB that is often underutilized. Namely, we claim that the rows of AB are linear combinations of the rows of B and that the columns of AB are linear combinations of the columns of A. To see this it suffices to work one row or column at a time, so first consider what happens if A is a $1 \times n$ matrix – i.e., a **row vector** – and B is an $n \times p$ matrix, so AB is a $1 \times p$ row vector. We may write $A = (x_1, \ldots, x_n) = x_1 e_1 + \ldots + x_n e_n$, with $e_i = (0, \ldots, 1, \ldots, 0)$ as usual. Now observe that the matrix product $e_i B$ is simply the ith row of B. It follows that $AB = x_1(e_1 B) + \ldots + x_n(e_n B)$ is a linear combination of the rows of B, as claimed. In the general case A and AB will each have m different rows, which simply means that we get m (possibly) different linear combinations of the rows of B.

The corresponding interpretation of the columns of AB as linear combinations of the columns of A is similar but more familiar. In particular, the fact that Ae_j returns the jth column of A is a key insight in analyzing the linear transformation $L_A : \mathbb{R}^m \to \mathbb{R}^n$ given by $x \mapsto Ax$.

2.4. **Fundamentals of Matrix Algebra.**

We have now defined addition and multiplication operations on matrices, so we have some kind of algebraic structure on them. Especially, things work out best if we restrict to **square matrices**, i.e., when the number of rows equals the number of columns, for then if $A, B \in M_{n,n}$, their product AB is defined and is again an element of $M_{n,n}$. In other words, matrix multiplication gives a **binary composition law** on $M_{n,n}$.

2.4.1. *Identity Elements.*

Let $(X, \cdot)$ be a set endowed with a binary operation. An **identity element** is an element $1 \in X$ such that for all $x \in X$, $1 \cdot x = x \cdot 1 = x$.

Lemma 2.12. *A set endowed with a binary operation $(X, \cdot)$ can have at most one identity element.*

Proof. Suppose 1 and $1'$ are both identity elements. Then

$$1 = 1 \cdot 1' = 1'.$$

□

For $n \in \mathbb{Z}^+$, we define the **identity matrix** I_n to be the matrix with (i, i) entry 1 for all i and all other entries 0.

Exercise 3.4: a) Let $A \in M_{m,n}$. Show that $AI_n = A$.
b) Let $B \in M_{n,p}$. Show that $I_n B = B$.
c) Deduce that I_n is the unique identity element for $(M_{n,n}, \cdot)$.

2.4.2. *Absorbing Elements.*

Exercise 3.5: Let $0_{m,n}$ denote the $m \times n$ matrix consisting entirely of zeros.
a) Let $A \in M_{k,m}$. Show that $A0_{m,n} = 0_{k,n}$.
b) Let $B \in M_{n,p}$. Show that $0_{m,n}B = 0_{m,p}$.
c) Let $0 = 0_{n,n}$. Deduce that for all $A \in M_{n,n}$, $A0 = 0A = 0$.

Exercise 3.6: Let $(X, \cdot)$ be a set endowed with a binary operation. An element $Z \in X$ is **absorbing** if for all $A \in X$, $ZA = AZ = Z$.
a) Show that there can be at most one absorbing element.
b) Deduce that the zero matrix is the unique element $Z \in M_{n,n}$ such that for all $A \in M_{n,n}$, $ZA = AZ = Z$.

2.4.3. *Commutativity.*

A binary operation $\cdot$ on a set X is **commutative** if $xy = yx$ for all $x, y \in X$.

Multiplication in $M_{1,1}$ is the usual multiplication of real numbers, so it is of course commutative. However, the situation for $n \geq 2$ is quite different.

Example 2.13. *Let* $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ *and* $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. *Then*

$$AB = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = B,$$

while

$$BA = 0.$$

Thus matrix multiplication is not commutative in general.

Exercise 3.7: For $1 \leq i, j \leq n$, let E_{ij} be the matrix with all zero entries except for a 1 in the (i, j) entry.
a) Let $A \in M_{n,n}$. Show that AE_{ij} has every column zero except for the jth column, and the jth column is the ith column of A.
b) Let $A \in M_{n,n}$. Show that $E_{ij}A$ has every row zero except for the ith row, and the ith row is the jth row of A.

Exercise 3.8:
a) Let $\alpha \in \mathbb{R}$. Show that αI_n is the matrix with (i, i) entry equal to α and (i, j) entry equal to 0 for all $i \neq j$. Such matrices are called **scalar matrices**.
b) Let A be a scalar matrix. Show that A **commutes with** every $B \in M_{n,n}$: for all $B \in M_{n,n}$, $AB = BA$.

Exercise 3.9: Let $A \in M_{n,n}$ be such that A commutes with every $B \in M_{n,n}$.
a) Fix $1 \leq i \leq n$. Use $AE_{ii} = E_{ii}A$ to show that if $j \neq i$, then $a_{ij} = a_{ji} = 0$.
b) Fix $1 \leq j \leq n$. Use $AE_{1j} = E_{1j}A$ to show that $a_{11} = a_{jj}$.
c) Deduce that A is a scalar matrix.

Define the **center** of $M_{n,n}$ to be the set of all matrices $A \in M_{n,n}$ which commute with every $B \in M_{n,n}$. The preceding exercises show that the center of $M_{n,n}$ is precisely the set of all scalar matrices.

2.4.4. *Associativity.*

Proposition 2.14. *Matrix multiplication is associative: if* $A \in M_{m,n}$, $B \in M_{n,p}$ *and* $C \in M_{p,q}$, *then*

$$(AB)C = A(BC).$$

Proof. First Proof: Because matrix multiplication corresponds to composition of linear maps, and function composition is always asssocative, matrix multiplication must be associative.
Second Proof: The above argument is certainly the best explanation of why matrix multiplication is associative. On the other hand it should be possible to show directly that $(AB)C = A(BC)$, and it is. A little quiet contemplation shows that for $1 \leq i \leq m$ and $1 \leq l \leq q$, the (i, l) entry of *both* $(AB)C$ and $A(BC)$ is

$$\sum_{1 \leq j \leq n,\ 1 \leq k \leq p} a_{ij} b_{jk} c_{kl}.$$

□

Because we have shown matrix multiplication is associative, we may freely drop the parentheses, writing ABC for $(AB)C = A(BC)$. Moreover, it is a general fact

about binary operations that as soon as we have associativity then we have a "generalized associativity": any n-fold product $A_1 \cdots A_n$ is well-defined independent of the parentheses. See [Wa01] for a nice treatment of this.

Exercise 3.10: Let $M_1, \ldots, M_n$ be matrices, of dimensions such that for $1 \leq i \leq n-1$, each product $M_i M_{i+1}$ is defined. Write down an explicit expression for the general entry of $M_1 \cdots M_n$.

2.4.5. *Inverses.*

Let $(X, \cdot)$ be a set endowed with a binary operation, and possessing an identity element 1. Let $A \in X$.

An element $A_L \in X$ is a **left inverse** to A if $A_L A = 1$.

An element $A_R \in X$ is a **right inverse** to A if $A A_R = 1$.

Let $A, B \in X$. Sometimes we will say that B is a **one-sided inverse** to A if either $AB = 1$ or $BA = 1$. Note that this is symmetric: if B is a one-sided inverse to A, then A is a one-sided inverse to B.

An element $B \in X$ is an **inverse** to A if $AB = BA = 1$.

An element $A \in X$ is **invertible** if it has an invere.

Proposition 2.15. *Let $(X, \cdot)$ be a set endowed with an associative binary composition law and possessing an identity element 1. Then any element $A \in X$ has at most one inverse: if B and C are both inverses to A, then $B = C$.*

Proof.

$$C = 1 \cdot C = (BA)C = BAC = B(AC) = B \cdot 1 = B.$$

□

Because an element has at most one inverse, it makes sense to denote the inverse of an (invertible!) element $A \in (X, \cdot)$ by A^{-1}.

Exercise 3.11: Suppose that $B = A^{-1}$. Show that $A = B^{-1}$. (We sometimes say "A and B are mutually inverse.")

Exercise 3.12: Consider $f : \mathbb{R} \to [0, \infty)$ by $f(x) = x^2$.
a) Show that f is surjective but not injective, so by Theorem 2.1 it has a right inverse g – i.e., $f \circ g = 1_{[0,\infty)}$ – but no left inverse.
b) Find one right inverse to f.
c) Find *all* right inverses to f. Conclude in particular that an element can have more than one right inverse.

Exercise 3.13: Consider $f : [0, \infty) \to \mathbb{R}$ by $f(x) = x$. (Note that this is not the identity function, since the domain and codomain are different!)
a) Show that f is injective but not surjective, so by Theorem 2.1 it has a left inverse g – i.e., $g \circ f = 1_{[0,\infty)}$ – but no right inverse.

b) Find one left inverse to f.
c) Find *all* left inverses to f. Conclude in particular that an element can have more than one left inverse.

Exercise 3.14: Addition on $M_{m,n}$ is also a binary composition law.
a) Show that there is an identity element for matrix addition: what is it?
b) Show that every matrix $A \in M_{m,n}$ has an *additive inverse* matrix: what is it?

Remark: Very often in mathematics we will have a set endowed with two binary operations, called $+$ and $\cdot$. (There are whole courses on this...) In this case, A^{-1} refers to the *multiplicative inverse* rather than the additive inverse. That goes in particular for $M_{n,n}$.

Proposition 2.16. *(Shoes 'n' Socks) Let $(X, \cdot)$ be a set endowed with an associative binary composition law and possessing an identity element 1. Let $A_1, \ldots, A_n \in X$ be invertible elements. Then the product $A_1 \cdots A_n$ is invertible and*

$$(A_1 \cdots A_n)^{-1} = A_n^{-1} \cdots A_1^{-1}.$$

Proof. Consider

$$(A_n^{-1} \cdots A_1^{-1})(A_1 \cdots A_n).$$

Working our way from the inside out we cancel $A_1^{-1}A_1$, then cancel $A_2^{-1}A_2$, and so forth, finally cancelling $A_n^{-1}A_n$ to get 1. And much the same goes for

$$(A_1 \cdots A_n)(A_n^{-1} \cdots A_1^{-1}) :$$

we first cancel $A_n A_n^{-1}$, then $A_{n-1}A_{n-1}^{-1}$, and so forth, finally cancelling $A_1 A_1^{-1}$ to get 1. □

We can recapture the spirit of the statement and proof of Proposition 2.16 as follows: think of an invertible element as a process which can be reversed. (This is especially reasonable when the elements of X are functions and the binary operation is composition, because one often thinks of a function as being a procedure which takes an input, does something to it, and returns an output.) Any process which is obtained by performing several reversible processes can itself be reversed: however, to do so we must reverse the individual processes in reverse order. If that sounds like a mouthful, consider: barring a catastrophe, putting on your socks is a reversible process, as is putting on your shoes. In the morning we put on our socks first and then our shoes. In the evening we undo this composite procedure by undoing the individual components, but in order to do so we must now deal with our shoes first and our socks second.

Exercise 3.15: Let $A \in M_{n,n}$. Suppose that A is invertible. Show that A *cannot* have a zero row or column.

Let's reconsider the all-important example where $(X, \cdot)$ is the set of all functions from a set A to itself, and the binary operation is composition of functions. In this setting, Theorem 2.1 says that if f has a left inverse f_L and a right inverse f_R, then it is injective and surjective, hence bijective, hence it has an inverse function. This turns out to be a general fact about composition laws.

Proposition 2.17. *Let $(X, \cdot)$ be a set endowed with an associative binary composition law and possessing an identity element 1. Let $A \in X$ have a left inverse A_L and a right inverse A_R. Then $A_L = A_R$ is the inverse of A.*

Proof. We have

$$A_R = 1 \cdot A_R = (A_L A)A_R = A_L A A_R = A_L(AA_R) = A_L \cdot 1 = A_L.$$

It follows immediately that $A_R A = AA_R = 1$, so $A_R = A_L$ is the inverse of A. □

Here is a related result.

Proposition 2.18. *Let $(X, \cdot)$ be a set endowed with an associative binary composition law and possessing an identity element 1. Let $A, B \in X$ be such that $BA = 1$. If* either *of A or B is invertible then A and B are mutually inverse.*

Proof. Suppose first that B is invertible. Multiplying both sides of $BA = 1$ on the left by B^{-1}, we get $A = 1 \cdot A = B^{-1}BA = B^{-1}1 = B^{-1}$. Next suppose that A is invertible. Multiplying both sides of $BA = 1$ on the right by A^{-1}, we get $B = B \cdot 1 = BAA^{-1} = A^{-1}$. □

Although matrix multiplication is an instance of function composition, it is an especially simple instance which behaves better than the general case. It turns out that a matrix $A \in M_{n,n}$ which has a left inverse must also have a right inverse and thus be invertible (and similarly, if A has a right inverse it must also have a left inverse and thus be invertible). It turns out to be difficult to prove this directly, however. The right thing to do is to take a more ambitious approach by trying to give a *characterization* of invertible matrices in terms of row reduction.

Theorem 2.19. *For $A \in M_{n,n}$, the following are equivalent:*
(i) $\operatorname{rref} A = I_n$.
(ii) $\operatorname{rank} A = n$.
(iii′): The equation $Ax = 0$ has only the trivial solution $x = 0$.
(iii′) $L_A : \mathbb{R}^n \to \mathbb{R}^n$ is injective.
(iv) For all $b \in \mathbb{R}^n$, the equation $Ax = b$ has a unique solution.
(iv′) $L_A : \mathbb{R}^n \to \mathbb{R}^n$ is bijective.
(v) A is invertible: there is $B \in M_{n,n}$ with $AB = BA = I_n$.

Proof. Step 1: Recall that we already know (i) $\iff$ (ii) $\iff$ (iii) $\iff$ (iv): this is Theorem 1.5. The equivalence of (iii) and (iii′) is Lemma 2.1, and the equivalence of (iv) and (iv′) is immediate from the definitions: to say that L_A is bijective is exactly to say that for every $b \in \mathbb{R}^n$, there is a unique $x \in \mathbb{R}^n$ such that $Ax = L_A(x) = b$.
Step 2: It remains to show that any one of the conditions other than (v) implies (v) and that (v) implies any one of the other conditions. We will show that (iv′) implies (v) and that (v) implies (i).
(iv′) $\implies$ (v): Since $L_A : \mathbb{R}^n \to \mathbb{R}^n$ is bijective, by Theorem 2.1 it has an inverse function, i.e., there is $g : \mathbb{R}^n \to \mathbb{R}^n$ with $g \circ L_A = L_A \circ g = 1$. By Lemma 2.11 g is a linear transformation, hence by Theorem 2.8 $g = L_B$ for some matrix B. Expressing that L_A and L_B are mutually inverse functions in terms of matrices we get precisely that $BA = I_n = AB$.
(v) $\implies$ (i): If $BA = I_n$ then $L_B \circ L_A = 1_{\mathbb{R}^\times}$, so by Theorem 2.1, L_A is injective. □

The logic of the proof yields a further result. We could have stated it as an equivalent condition in Theorem 2.19, but we decided to optimize our exposition so as to make the proof of that key result as simple and clean as possible.

Corollary 2.20. *Let $A, B \in M_{n,n}$. If $BA = I_n$, then $AB = I_n$.*

Proof. As in the proof of Theorem 2.19, $BA = I_n$ implies that L_A is injective. By Theorem 2.19, A is invertible. According to Proposition 2.18, its one-sided inverse B must actually be its inverse, hence also $AB = I_n$. □

This result can be restated as follows: a square matrix which has a one-sided inverse must be invertible, and any one-sided inverse is in fact the inverse.

Up until this point our discussion of inverses of matrices has been purely theoretical. We now know that we can tell whether a square matrix is invertible by putting it in reduced row echelon form, but if $\operatorname{rref}(A) = I_n$, how do we go about finding the inverse? Again by row reduction!

Theorem 2.21. *If $\operatorname{rref}(A) = I_n$, then $\operatorname{rref}[A \mid I_n] = \left[I_n \mid A^{-1}\right]$.*

Proof. Write B for A^{-1}, so we have the equation $AB = I_n$. Focusing in on the jth columns of both sides of this equation gives us a system of matrix equations $Ab_j = e_j$ which we are trying to solve for the vectors $b_1, \ldots, b_n$. To solve this system we proceed in the usual way: write down $[A \mid e_j]$ and put in reduced row echelon form. Since A is invertible, $\operatorname{rref} A = I_n$, the solution vector b_j is *unique*, and thus the reduced row echelon form is $[I_n \mid b_j]$. Since the columns of a matrix function quite independently under row reduction, nothing stops us from writing down the wider augmented matrix $[A \mid I_n]$. The same row operations that put A in reduced row echelon form convert the jth column on the right to the unique solution vector b_j, so $\operatorname{rref}[A \mid I_n] = \left[I_n \mid A^{-1}\right]$. □

2.5. **Elementary Matrices.**

Fix a positive integer m.

For $1 \leq i \neq j \leq m$, a **type I elementary matrix** $S_{i,j}$ is the $m \times m$ matrix which results from interchanging the ith and jth rows of I_m.

For $1 \leq i \leq m$, a **type II elementary matrix** $M_{i,j}(\alpha)$ is the $m \times m$ matrix which results from multiplying every entry in a single row of I_m by some nonzero $\alpha \in \mathbb{R}$.

For $1 \leq i \neq j \leq m$ and nonzero $\alpha \in \mathbb{R}$, a **type III elementary matrix** $T_{i,j}(\alpha)$ is the $m \times m$ matrix which differs from I_m precisely in having its (j, i) entry equal to α.

Exercise 3.16: Show that every elementary matrix is invertible. In fact, show that the inverse of any elementary matrix is again an elementary matrix of the same type.

In all cases an elementary matrix is the matrix you get by performing an elementary row operation on the identity matrix I_m. In fact a little more is true: in each case the above data is sufficient to describe an elementary row operation on any $m \times n$ matrix.

Proposition 2.22. *Let $A \in M_{m,n}$. We perform one elementary row operation R on A to get a new matrix B. Then $B = EA$, where E is the corresponding elementary matrix.*

Exercise 3.17: Prove Proposition 2.22.

The upshot is that we can track row reduction as a (finite!) sequence of premultiplications – i.e., multiplications on the left – by elementary matrices. Suppose that these elementary matrices are called $E_1, \ldots, E_r$. Then we get the matrix equation

$$\operatorname{rref}(A) = E_r \cdots E_1 A. \tag{2}$$

Use of elementary matrices is not essential, but often allows one to make cleaner arguments involving matrices rather than row operations.

Example 2.23. *We will use elementary matrices to give a second proof of the important fact that $A \in M_{n,n}$ is invertible if and only if $\operatorname{rref}(A) = I_n$. First suppose $\operatorname{rref}(A) = I_n$. Then (2) reads*

$$I_n = E_r \cdots E_1 A.$$

Each elementary matrix E_i is invertible, so by Shoes 'n' Socks so is $E_r \cdots E_1$, and the inverse is $E_1^{-1} \cdots E_r^{-1}$. Premultiplying both sides by this product, we get

$$E_1^{-1} \cdots E_r^{-1} = A.$$

As a product of invertible elements, A is itself invertible: to be explicit about it, we can apply Shoes 'n' Socks again to get

$$A^{-1} = E_r \cdots E_1.$$

Here is another example of a result that can be proved via row reduction considerations but is a little cleaner via elementary matrices. This time the result is a new one: we could have included it among the equivalent conditions of Theorem 2.19 but we had enough conditions to deal with at one time.

Theorem 2.24. *Let $A \in M_{n,n}$. The equivalent conditions (i) through (v) of Theorem 2.19 are also equivalent to each of the following:*
(vi) There are no constraints: for all $b \in \mathbb{R}^n$, the linear equation $Ax = b$ is consistent, i.e., has at least one solution.
(vi′) The linear transformation $L_A : \mathbb{R}^n \to \mathbb{R}^n$ is surjective.

Proof. Step 1: Conditions (vi) and (vi′) are equivalent to each other: indeed, unpacking the definition of the surjectivity of L_A, one gets precisely condition (vi).
Step 2: Certainly conditions (iv) implies condition (vi) (or, equally true: certainly condition (iv′) implies condition (vi′)).
Step 3: We suppose condition (vi). It suffices to show that rref A has no zero rows, for then $\operatorname{rank} A = n$: condition (ii) of Theorem 2.19. Suppose on the contrary that the last row of rref $A = 0$. Start with $b \in \mathbb{R}^n$, and let

$$\operatorname{rref}[A \mid b] = [\operatorname{rref}(A) \mid b']. \tag{3}$$

In turn let b' be $(b_1, \ldots, b_n)$. If $b_n \neq 0$, then the last equation of the augmented rref matrix reads $0 = b_n$, a contradiction. So in particular we get an inconsistent system if $b'_n = e_n = (0, \ldots, 0, 1)$. What we need to argue for now is that there is some choice of $b \in \mathbb{R}^n$ such that the b' defined by (3) turns out to be e_n. In other

words, we need to work the row reduction process backwards. This is exactly what elementary matrices make clean and easy: there are elementary matrices such that

$$E_n \cdots E_1 A = \operatorname{rref}(A),$$

and then for any $b \in \mathbb{R}^n$, performing the same row operations gives

$$E_n \cdots E_1 b = b'.$$

But $E_1, \ldots, E_n$ are all invertible, so we have

$$b = (E_n \cdots E_1)^{-1} b' = E_1^{-1} \cdots E_n^{-1} b'.$$

Taking $b' = e_n$ this shows exactly what coefficient vector b we need to start with to get $\operatorname{rref}[A \mid b] = [\operatorname{rref}(A) \mid e_n]$. Summing up: if $\operatorname{rref} A$ had a zero row, there is some b such that $Ax = b$ is inconsistent. This is precisely the contrapositive of what we wanted to show, so we're done. □

2.6. **Diagonal Matrices.**

A matrix $A \in M_{n,n}$ is **diagonal** if for all $i \neq j$, $a_{ij} = 0$. In other words the only nonzero entries lie along the **main diagonal** $a_{11}, a_{22}, \ldots, a_{nn}$.

Since we only need n numbers to specify an $n \times n$ diagonal matrix, we don't need double indexing, and we will denote the diagonal matrix with diagonal entries $a_1, \ldots, a_n$ as $\Delta(a_1, \ldots, a_n)$.

Example 2.25. *a)* $\Delta(1, \ldots, 1)$ *is the identity matrix* I_n.
b) For any $\alpha \in \mathbb{R}$, $\Delta(\alpha, \ldots, \alpha)$ *is a scalar matrix.*

The algebra of diagonal matrices is much simpler than that of arbitary matrices.

Exercise 3.18: Let $A = \Delta(a_1, \ldots, a_n)$, $B = \Delta(b_1, \ldots, b_n)$ be diagonal matrices.
a) Show that $AB = \Delta(a_1 b_1, \ldots, a_n b_n)$.
b) Deduce that $AB = BA$: diagonal matrices commute.

Proposition 2.26. *Consider a diagonal matrix* $A = \Delta(a_1, \ldots, a_n)$.
a) The following are equivalent:
(i) A *is invertible.*
(ii) $a_1, \ldots, a_n$ *are all nonzero.*
b) If the equivalent conditions of part a) hold, then $A^{-1} = \Delta(a_1^{-1}, \ldots, a_n^{-1})$.

Proof. Step 1: To show (i) $\implies$ (ii) in part a) we will verify the contrapositive: not (ii) $\implies$ not (i). The negation of (ii) is that for some i, $a_i = 0$. If this is the case then the ith row of A is zero, so $\operatorname{rref} A \neq I_n$ and A is not invertible.
Step 2: Suppose $a_1, \ldots, a_n$ are all nonzero. By Exercise 3.18a),

$$\Delta(a_1, \ldots, a_n)\Delta(a_1^{-1}, \ldots, a_n^{-1}) = \Delta(1, \ldots, 1) = I_n.$$

This shows that $\Delta(a_1, \ldots, a_n)$ is invertible and that its inverse is $\Delta(a_1^{-1}, \ldots, a_n^{-1})$. Hence we get (ii) $\implies$ (i) in part a) and also the result of part b): we're done. □

2.7. **Triangular Matrices.**

A matrix $A \in M_{n,n}$ is **upper triangular** if for all $i > j$, $a_{ij} = 0$.

A matrix $A \in M_{n,n}$ is **lower triangular** if for all $i < j$, $a_{ij} = 0$.

A matrix $A \in M_{n,n}$ is diagonal if and only if it is both upper and lower triangular. Above we introduced the idea that diagonal matrices are ideally simple and easy to work with that we wish that every matrix could be diagonal. In a precise sense that we have not yet encountered – that of **similarity** – it is not possible to make every matrix diagonal, but (at least if we are able to use complex numbers as scalars, as we eventually will!) in this same sense it will turn out that every matrix is similar to an upper triangular matrix.

For an $n \times n$ matrix the entries $a_{11}, \ldots, a_{nn}$ are said to lie on the **main diagonal**. Thus $A \in M_{n,n}$ is upper triangular if all the entries lying *below* the main diagonal are zero, is lower triangular if all the entries lying *above* the main diagonal are zero, and is diagonal if all the entries are zero except possibly those on the main diagonal.

Exercise 3.19: Let $A \in M_{n,n}$.
a) Suppose A is in row echelon form. Show that A is upper triangular.
b) Suppose A is upper triangular. Must A be in row echelon form?

Proposition 2.27. *Let $A, B \in M_{n,n}$.*
a) If A and B are both upper triangular, then AB is upper triangular.
b) If A and B are both lower triangular, then AB is lower triangular.

Proof. We have, as always, that

$$(AB)_{ij} = \sum_{j=1}^{n} a_{ik} b_{kj}. \tag{4}$$

a) Suppose A and B are upper triangular and $i > j$. If $i > k$ we have $a_{ik} = 0$ since A is upper triangular. On the other hand, if $k > j$ we have $b_{kj} = 0$ since B is upper triangular. But no matter what k is, at least one of these two conditions must hold, because if both fail we have $j \leq k \leq i$, contradicting our assumption that $i > j$. Thus each term in the sum of (4) is the product of two numbers, at least one of which is zero. So $(AB)_{ij} = 0$.
b) I leave this to you: you'll know that you've understood the proof of part a) if you can supply the proof of part b). □

Exercise 3.20: a) Let $A, B \in M_{n,n}$ be upper triangular. Show that for all $1 \leq i \leq n$, $(AB)_{ii} = a_{ii} b_{ii}$.
b) Does the conclusion of part a) hold if instead of being upper triangular, A and B are both lower triangular?

Exercise 3.21: Suppose $A \in M_{n,n}$ is upper triangular with $a_{ii} \neq 0$ for all $1 \leq i \leq n$. Show that A is in row echelon form.

Proposition 2.28. *For an upper triangular $A \in M_{n,n}$, the following are equivalent:*
(i) For all $1 \leq i \leq n$, $a_{ii} \neq 0$.
(ii) A is invertible.

Proof. (i) $\implies$ (ii): By Exercise 3.21, A is in row echelon form. Since every row has the leading entry a_{ii}, A has rank n and is thus (by Theorem 2.19) invertible.
(ii) $\implies$ (i): We will work our way from the bottom right to the top left. First, we must have $a_{nn} \neq 0$, since otherwise the last row is zero and A cannot be invertible. Since $a_{nn} \neq 0$, we zero out the entries of the last column. Because a_{nn} is the only nonzero entry in the last row, this process does not any of the entries of the other column, so in particular does not change the diagonal entries: it is enough to prove that the diagonal entries are all nonzero. Now if $a_{n-1,n-1}$ were equal to zero, the $(n-1)$st row would be zero, again contradicting invertibility. Now we can zero out the other entries of the $(n-1)$st column without disturbing any of the diagonal entries. And so forth: we get that $a_{nn} \neq 0$, $a_{n-1n-1} \neq 0$,..., $a_{11} \neq 0$. □

Proposition 2.29. *Let $A \in M_{n,n}$ be an invertible matrix. Then:*
a) If A is upper triangular, then A^{-1} is upper triangular.
b) If A is lower triangular, then A^{-1} is lower triangular.

Proof. a) If A is upper triangular and invertible, then by Proposition 2.28, all the entries along the main diagonal are nonzero. Thus to put it in reduced row echelon form we only need to perform type (II) row operations to make the leading entries 1 and type (III) row operations of the form of adding a multiple of row j to row i with $i < j$ to zero out the entries above each leading entry. The elementary matrices E_i corresponding to each of these row operations is upper triangular, and

$$I_n = \mathrm{rref}(A) = E_r \cdots E_1 A,$$

so $A^{-1} = E_r \cdots E_1$ is a product of upper triangular matrices and hence, by Proposition 2.27, upper triangular.
b) The argument is very similar to that of part a) and we leave it to the reader. □

Exercise 3.22: Prove part b) of Theorem 2.29. (One thing to convince yourself of is the fact that an invertible lower triangular matrix is generally *not* in row echelon form is not a problem.)

2.8. **The Transpose Matrix.**

Before we leave the realm of basic matrix algebra for more exalted terrain, we want to discuss one last operation on matrices. In a way this last operation is the easiest – certainly you could explain it to anyone, regardless of their mathematical background – but it is a little less clear why this operation should be important in linear algebra. The latter question, unfortunately, will only be fully addressed later in the course.

Let $A \in M_{m,n}$. We define the **transpose matrix** $A^T \in M_{n,m}$ as the matrix whose (i,j) entry is a_{ji}. In other words, the rows of A become the columns of A^T, and vice versa. One can also think of taking the transpose as reflecting the entries of A across the main diagonal.

Proposition 2.30. *Let $A, B \in M_{m,n}$, $C \in M_{n,p}$ and $\alpha \in \mathbb{R}$. Then:*
a) $(A^T)^T = A$.
b) $(\alpha A)^T = \alpha A^T$.
c) $(A+B)^T = A^T + B^T$.

d) $(AC)^T = C^T A^T$.
e) If A *is invertible, so is* A^T *and* $(A^T)^{-1} = (A^{-1})^T$.

Proof. a) To get from a matrix to its transpose we interchange the rows and the columns. Doing this twice gets us back to the original matrix.
b) Both $(\alpha A)^T$ and αA^T have (i,j) entry αa_{ji}.
c) Both $(A+B)^T$ and $A^T + B^T$ have (i,j) entry $a_{ji} + b_{ji}$.
d) The (k,i) entry of $C^T A^T$ is the dot product of the kth row of C^T with the ith column of A^T. This is also the dot product of the ith row of A with the jth column of C, hence it is the (k,i) entry of AC and thus the (i,k) entry of $(AC)^T$.
e) Using part d), we have $(A^{-1})^T A^T = (AA^{-1})^T = I_n^T = I_n$. □

In the above proof we used that the identity matrix I_n has the property that $I_n^T = I_n$. A matrix which has this property must be square. This turns out to be an interesting and important class of square matrices: we say $A \in M_{n,n}$ is **symmetric** if $A^T = A$. We also say that $A \in M_{n,n}$ is **skew symmetric** if $A^T = -A$.

Every symmetric matrix can be built as follows: we will in all $1+2+\ldots+n = \frac{n(n+1)}{2}$ entries lying on our above the main diagonal arbitrarily. Then the symmetry condition tells us that each entry a_{ij} for $i < j$ is equal to the corresponding entry a_{ji}. In particular, in a natural sense that we will later make precise, the set $S_{n,n}$ of all $n \times n$ symmetric matrices can be paramterized in terms of $\frac{n(n+1)}{2}$ parameters.

Exercise 3.22: Give a similar description of an $n \times n$ skew symmetric matrix. How many parameters does it take to specify such a matrix?

Exercise 3.23: a) Let $A \in M_{n,n}$. Suppose $A = A^T = -A$. Show that $A = 0$.
b) Let $A \in M_{n,n}$. Show that $A + A^T$ is symmetric and $A - A^T$ is skew symmetric.
c) Let $A \in M_{n,n}$. Show that there are *unique* matrices $A_s, A_{\text{ss}} \in M_{n,n}$ such that: A_s is symmetric, A_{ss} is skew symmetric, and $A = A_s + A_{\text{ss}}$.

One merit of the transpose matrix is to give us a formalism between switching between "row vectors" – i.e., elements of $M_{1,n}$ – and "column vectors" – i.e., elements of $M_{m,1}$: namely, take the transpose. We reiterate our standard convention that when viewing a vector $v \in \mathbb{R}^n$ as a matrix we view it as a column vector, not a row vector. With this convention, we can reinterpret the dot product as itself being a matrix multiplication:

$$\forall v, w \in \mathbb{R}^n, v \cdot w = v^T w.$$

Proposition 2.31. *Let* $A \in M_{m,n}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. *Then*

$$(Ax) \cdot y = x \cdot (A^T y).$$

Proof. Much of the work in this result is appreciating that both sides are well-defined, even though the left hand side is a dot product of vectors in $\mathbb{R}^m$ and the right hand side is a dot product of vector in $\mathbb{R}^n$. Once you agree that both expressions are well-defined, we can move on to the proof:

$$(Ax) \cdot y = (Ax)^T y = x^T A^T y = x^T (A^T y) = x \cdot (A^T y).$$

□

3. Subspaces, Bases and Dimension

3.1. **Subspaces.**

A subset V of $\mathbb{R}^n$ is a **linear subspace** if all of the following hold:

(SS1) $0 \in V$.
(SS2) For all $v, w \in V$, $v + w \in V$.
(SS3) For all $v \in V$ and all $\alpha \in \mathbb{R}$, $\alpha v \in V$.

Remark 3.1. *Since we will not be considering any other kind of subspace of $\mathbb{R}^n$ in this course, we will omit the "linear" from "linear subspace"...but you should be aware that subspaces of various sorts are ubiquitous throughout mathematics.*

We claim that in the presence of (SS2) and (SS3), we could replace (SS1) with

(SS1′) $V \neq \varnothing$ (i.e., V is nonempty).

Indeed, it is clear that (SS1) $\implies$ (SS1′). Conversely, suppose (SS1′) holds: thus there is some element $v \in V$. By (SS3), $0 \cdot v = 0 \in V$.

We prefer to give the axioms for a subspace in this form because (i) in practice it is certainly no harder to check (SS1) than (SS1′) and (ii) the formulation (SS1′) is so innocuous that it is easy to forget.

Example: The subset $\{0\}$ is a subspace of $\mathbb{R}^n$: it may seem too obvious to be worth mentioning, but this is important for bookkeeping purposes. Yet more obviously, $\mathbb{R}^n$ is a subspace of itself.

Example: Let ℓ be a line in $\mathbb{R}^n$. We claim that ℓ is a subspace iff ℓ passes through the origin. Indeed the condition $0 \in \ell$ is necessary by (SS1). Conversely, if $0 \in \ell$ then we may express ℓ parametrically as $\{tx \mid t \in \mathbb{R}\}$ for some nonzero vector $x \in \mathbb{R}^n$. We can now easily check (SS1) and (SS2): if $v, w \in \ell$ then $v = t_1 x$ and $w = t_2 x$ for some $t_1, t_2 \in \mathbb{R}$. Then

$$v + w = t_1 x + t_2 x = (t_1 + t_2)x$$

lies in ℓ. Similarly but yet more easily, if $v = t_1 x$ and $\alpha \in \mathbb{R}$, then $\alpha v = (\alpha t_1)x \in \ell$.

Exercise: Try to convince yourself that the subspaces of $\mathbb{R}^2$ are precisely: $\{0\}$, the lines through the origin, and $\mathbb{R}^2$. (Suggestion: the key is to believe that as soon as a subspace V of $\mathbb{R}^2$ contains two vectors v and w which do not lie on the same line, we must have $V = \mathbb{R}^2$. You can try to prove this now if you like, but we will carefully prove this and more general facts later on. The point of this exercise is to acquire intuition that it should be true.)

Example: Let P be a plane in $\mathbb{R}^3$. We claim that P is a subspace iff P passes through the origin. Again the condition is certainly necessary. We can show the converse in several different ways, depending upon how we choose to represent P: recall that we know at least two ways to do so. Suppose first that we use the *parametric expression* of P: namely there are vectors x and y – which are not scalar

multiples of one another – such that

$$P = \{sx + ty \mid s, t \in \mathbb{R}\}.$$

Now let $v = s_1x + t_1y$ and $w = s_2x + t_2y$ be vectors in P. Then

$$v + w = s_1x + t_1y + s_2x + t_2y = (s_1 + s_2)x + (t_1 + t_2)y \in P.$$

Similarly but more easily, if $v = sx + ty \in P$ and $\alpha \in \mathbb{R}$, then

$$\alpha v = \alpha(sx + ty) = (\alpha s)x + (\alpha t)y \in P.$$

Thus P is a subspace.

Example: Let's revisit the previous example of a plane $P \subset \mathbb{R}^3$ passing through the origin, but this time we represent P using a normal vector $n = (a, b, c)$, namely

$$P = \{x = (x_1, x_2, x_3) \in \mathbb{R}^3 \mid 0 = n \cdot x = ax_1 + bx_2 + cx_3 = 0\}.$$

Once again we have rigged things so that $0 \in P$. Further, if $x = (x_1, x_2, x_3), y = (y_1, y_2, y_3) \in P$ then

$$ax_1 + bx_2 + cx_3 = ay_1 + by_2 + cy_3 = 0$$

and thus

$$a(x_1+y_1)+b(x_2+y_2)+c(x_3+y_3) = (ax_1+bx_2+cx_3)+(ay_1+by_2+cy_3) = 0+0 = 0,$$

so $x + y \in P$. Finally, if $x = (x_1, x_2, x_3) \in P$ and $\alpha \in \mathbb{R}$, then $ax_1 + bx_2 + cx_3 = 0$, hence

$$0 = \alpha(ax_1 + bx_2 + cx_3) = (\alpha a)x_1 + (\alpha b)x_2 + (\alpha c)x_3 = 0.$$

Exercise: I claim the subspaces of $\mathbb{R}^3$ are: $\{0\}$, the lines through the origin, the planes through the origin, and $\mathbb{R}^3$. Try to convince yourself that this is true.

The above two techniques of showing that a plane through the origin in $\mathbb{R}^3$ is a subpace each generalizes in a different way.

Exercise: Recall that for any $n \geq 2$, a plane P in $\mathbb{R}^n$ is a subset of the form $\{sx + ty + z \mid s, t \in \mathbb{R}\}$, where $x, y, z \in \mathbb{R}^n$ and x, y are not scalar multiples of each other. Show that P is a subspace if and only if it passes through the origin, i.e., $0 \in P$.

Exercise: Recall that for any $n \geq 2$, a **hyperplane** H in $\mathbb{R}^n$ is a subset of the form $\{x \in \mathbb{R}^n \mid n \cdot x = c\}$ for any nonzero vector n and $c \in \mathbb{R}$. Show that a hyperplane is a subspace of $\mathbb{R}^n$ if and only if it passes through the origin (if and only if $c = 0$).

Exercise: Try to convince yourself that the subspaces of $\mathbb{R}^4$ are: $\{0\}$, lines through the origin, planes through the origin, hyperplanes through the origin, and $\mathbb{R}^4$.

Exercise: Show there are more subspaces of $\mathbb{R}^5$ than just: $\{0\}$, lines through the origin, planes through the origin, hyperplanes through the origin, and $\mathbb{R}^5$. (Hint: we are missing "three-dimensional subspaces", whatever that means. But try to write one down.)

We should also be sure to give some examples of subsets of $\mathbb{R}^n$ which are *not* subspaces. Of course any subset which does not contain the origin is such an example. Having said that, we may as well consider subsets which have this property.

Non-Example: Let $Q = \{(x, y) \in \mathbb{R}^2 \mid x, y \geq 0\}$ be the first quadrant in $\mathbb{R}^2$. Then $0 \in Q$, so Q satisfies (SS1). Moreover, it is easy to see that if $v, w \in Q$, so is $v+w$. However, Q does not satisfy (SS3): $(1, 1) \in Q$ but $(-1, -1) = -1 \cdot (1, 1) \notin Q$. So Q is not a subspace. (However it is rather close to being a subspace in the sense that it satisfies (SS1), (SS2) and (SS3) for all $\alpha \geq 0$. Such subsets of $\mathbb{R}^n$ are called **cones**. They too show up throughout mathematics.)

Non-Example: Let $S = \{(x, y) \in \mathbb{R}^2 \mid x = 0 \text{ or } y = 0\}$. Then S satisfies (SS1) and (SS3) but not (SS2): both $e_1 = (1, 0)$ and $e_2 = (0, 1)$ lie in S, but $(1, 1) = e_1 + e_2 \notin S$.

Notice that the previous nonexample is precisely the union of two lines through the origin. We conclude that **the union of two subspaces of $\mathbb{R}^n$ need not be a subspace**. The following exercise purues this phenomenon more closely.

Exercise: Let V, W be subspaces of $\mathbb{R}^n$.
a) Suppose that $V \subset W$ or $W \subset V$. Show that $V \cup W$ is a subspace.
(Hint: this is a triviality.)
b) Suppose that $V \cup W$ is a subspace. Show that either $V \subset W$ or $W \subset V$. (This is not. Suggestion: work by contradiction and suppose that neither $V \subset W$ nor $W \subset V$. Thus there is $x \in V \setminus W$ and $y \in W \setminus V$. Since $V \cup W$ is assumed to be a subspace, we must have $x + y \in V \cup W$. Deduce a contradiction.)

In particular, $\mathbb{R}^n$ is never the union of two proper subspaces. One can carry this argument further, but we will wait until we know a bit more about subspaces.

Subspaces behave much better with respect to intersection.

Proposition 3.2. *Let $V, W \subset \mathbb{R}^n$ be subspaces. Then $V \cap W$ is a subspace of $\mathbb{R}^n$.*

Proof. The key idea is that no ideas are necessary (!!): we just follow our nose and check the properties. First, since $0 \in V$ and $0 \in W$, $0 \in V \cap W$. Second, let $x, y \in V \cap W$. Thus $x, y \in V$ and $x, y \in W$. Since V is a subspace, $x + y \in V$; since W is a subspace, $x + y \in W$. Thus $x + y \in V \cap W$. Finally, let $x \in V \cap W$ and $\alpha \in \mathbb{R}$. Since V is a subspace, $\alpha x \in V$; since W is a subspace, $\alpha x \in W$. Thus $\alpha x \in V \cap W$. We're done! □

Exercise: a) Let $V_1, \ldots, V_k$ be subspaces of $\mathbb{R}^n$. Show that the common intersection $V_1 \cap \ldots \cap V_k$ is a subspace of $\mathbb{R}^n$.
b) Suppose that I is a nonempty set and that for each $i \in I$ we are given a subspace V_i of $\mathbb{R}^n$. (Thus we have $\{V_i\}_{i \in I}$, an **indexed family** of subspaces.) Show that the common intersection $\bigcap_{i \in I} V_i$ – i.e., the set of $x \in \mathbb{R}^n$ which lie in V_i for all i – is a subspace of $\mathbb{R}^n$.

3.2. **Universal Examples of Subspaces.**

In this section we give further examples of subspaces. However, in contrast to the examples that we gave before, each of these examples will turn out to be a *universal example*: that is, in fact every subspace of $\mathbb{R}^n$ will attained as each of our examples.

Example 3.3. *Let $A \in M_{m,n}$. Then the* ***null space***

$$N(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$$

is a subspace of $\mathbb{R}^n$. As usual, this is easy to check: since $A0 = 0$, $0 \in N(A)$. If $x, y \in N(A)$ then $Ax = Ay = 0$, so $A(x + y) = Ax + Ay = 0$. Finally, if $x \in N(A)$ and $\alpha \in \mathbb{R}$ then $A(\alpha x) = \alpha Ax = \alpha \cdot 0 = 0$.

Example 3.4. *Let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then the* ***kernel***

$$\operatorname{Ker} L = \{x \in \mathbb{R}^n \mid Lx = 0\}$$

is a subspace of $\mathbb{R}^n$. Indeed, $L0 = 0$, so $0 \in \operatorname{Ker} L$. If $x, y \in \operatorname{Ker} L$, then $Lx = Ly = 0$, so $L(x+y) = 0$. Finally, if $x \in \operatorname{Ker} L$ and $\alpha \in \mathbb{R}$ then $L(\alpha x) = \alpha L(x) = \alpha \cdot 0 = 0$.

Remark 3.5. *Since every linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ is of the form $x \mapsto Ax$ for a unique matrix $A \in M_{m,n}$, Examples 3.3 and 3.4 are really the same example: the null space of a matrix A is the kernel of the corresponding linear transformation L_A, and conversely. Nevertheless both terms are commonly used, so we wanted to introduce them both.*

3.3. **Reducing Spanning Sets to Bases.**

Theorem 3.6. *Let $S = \{v_1, \ldots, v_k\}$ be a finite set of vectors in $\mathbb{R}^n$, and let $V = \operatorname{span}\{v, \ldots, v_k\}$. Then there is a subset $T \subset S$ such that T is a basis for V. In brief,* ***every finite spanning set can be reduced to a basis****.*

Proof. The idea is simple: we write out the vectors in our spanning set $v_1, \ldots, v_k$ and work from left to right: whenever we get a vector v_{i+1} which is a linear combination of the previous vectors $v_1, \ldots, v_i$, then we may remove it from S without changing the span. (In particular, we remove v_1 if and only if it is the zero vector.) We are left with a subset $T \subset S$ of vectors which still spans V, and for which none of which can be written as a linear combination of the previous vectors, hence it is linearly independent and also spans V, so it is a basis for V. □

The previous proof could hardly have been simpler, but maybe it seems a bit too theoretical for its own good: if I give you actual, numerical vectors, e.g.

$$S = \{(1, 3, 5, 7), (-2, 3, 0, 4), (1, 1, 1, 1), (x, x, x, x)\},$$

how do we actually find a subset T which is a basis for $\operatorname{span} S$?

Algorithm: Given $S = \{v_1, \ldots, v_k\}$ be a finite subset of $\mathbb{R}^n$. Let $M \in M_{n,k}$ be the matrix with columns $v_1, \ldots, v_k$. Put M in reduced row echelon form. Let $\mathfrak{t} \subset \{1, \ldots, k\}$ be the set of indices i such that the ith column of $\operatorname{rref} M$ contains a leading entry. Then $T = \{v_i \mid i \in \mathfrak{t}\}$ is a basis for $\operatorname{span} S$.

For $1 \leq i \leq k$, let M_i be the matrix obtained by taking the first i columns of M. The algorithm works because:
v_i is a linear combination of $v_1, \ldots, v_{i-1}$
$\iff$ there are $\alpha_1, \ldots, \alpha_i \in \mathbb{R}$ with $\alpha_i \neq 0$ such that $\alpha_1 v_1 + \ldots + \alpha_i v_i = 0$

$\iff$ in rref M_i, the ith column is a free variable.
Since elementary row operations work independently on the columns, we can test this all at once by looking at rref M.

3.4. **Enlarging Linearly Independent Sets to Bases.**

For a finite set S, $\#S$ denotes the number of elements of S; e.g. $\#\{2,4,6,8,10\} = 5$.

To be honest, for stylistic/aesthetic reasons I would prefer to defer the proof of the following result until we discuss the **Exchange Lemma**, at which point it comes for free. But giving the proof now will enable us to quickly establish a result which is the natural complement to the result of the last section.

Lemma 3.7. *Let $S \subset \mathbb{R}^n$ be linearly independent. Then S is finite, and $\#S \leq n$.*

Proof. It is enough to show that any set of $n+1$ vectors in $\mathbb{R}^n$ is linearly dependent. But recall that any set $\{v_1, \ldots, v_k\}$ of vectors in $\mathbb{R}^n$ is linearly dependent if and only if the matrix $M \in M_{n,k}$ with $v_1, \ldots, v_k$ as its columns has a nontrivial null space. But if $k = n+1$, then $M \in M_{n,n+1}$, i.e., M has more columns than rows, so must have at least one free variable and thus a nontrivial null space. □

Theorem 3.8. *Let V be a subspace of $\mathbb{R}^n$, and let $S \subset V$ be a linearly independent set. Then there is a finite subset $T \supset S$ which is a basis for V.*

Proof. Again the idea is very simple: S is a linearly independent subset of V, so it is *not* a basis precisely if there is some vector $v_1 \in V$ which is not in the span of S. If so, $S_1 = S \cup \{v_1\} \subset V$ is still linearly independent. If this larger set S_1 does not span V then there is some $v_2 \in V$ which is not in the span of S, so $S_2 = S_1 \cup \{v_2\} = S \cup \{v_1, v_2\}$ is linearly independent. And so on. This process stops precisely when we get a basis for V. And it must stop eventually. Indeed, it must stop after at most n steps: otherwise we get a linearly independent subset of $\mathbb{R}^n$ consisting of more than n vectors, contradicting the previous result. □

Corollary 3.9. *Every subspace V of $\mathbb{R}^n$ has a finite basis.*

Proof. Apply Theorem 3.8 to $S = \varnothing$. □

It is not quite as clear how to make Theorem 3.8 concrete as we did for Theorem 3.6. One difference is that in Theorem 3.6 what we aer given is completely concrete: a finite set of vectors in $\mathbb{R}^n$. In Theorem 3.8, we are instead given a subspace $V \subset \mathbb{R}^n$, an object which, as we have seen and will continue to see, is a bit more abstract and can be concretely realized in several different ways. But let us suppose for isnetance that $S = \{v_1, \ldots, v_k\}$ and that V is given to us as the span of a finite set of vectors:

$$V = \operatorname{span} w_1, \ldots, w_\ell.$$

Now we can be very concrete about how to enlarge S to a finite basis for V: we use the algorithm given as the concrete form of Theorem 3.6! We start with the ordered list $v_1, \ldots, v_k, w_1, \ldots, w_\ell$, which is certainly a spanning set of V. Then we *reduce* this ordered list to a basis as we did before: we form the matrix $M \in M_{n,k+\ell}$ with columns $v_1, \ldots, v_k, w_1, \ldots, w_\ell$, and we put in reduced row echelon form. Then we keep the vectors formed by the columns of M such that the corresponding columns of rref M have leading entries. Because $v_1, \ldots, v_k$ is linearly independent, we will necessarily keep all of the first k columns, so we have expanded S to a basis of V.

3.5. **The Exchange Lemma.**

In the previous two sections we have established a kind of "duality" between spanning sets and linearly independent sets: spanning sets have one of the two properties of a basis but are in general "too large" to have the other property (linear independence). However, any finite spanning set can be reduced so as to gain the linear independence property without losing the spanning property. Dually, linearly independent sets have the other of the two properties of a basis but are in general "too small" to have the first property (spanning). However, any linearly independent set (and all linearly independent sets in $\mathbb{R}^n$ are finite, indeed have at most n elements) can be enlarged so as to gain the spanning property without losing the linear independence property.

There is however a further "largeness" property of spanning sets that we have yet to establish and a further "smallness" property of linearly independent sets that we have as yet seen only in a weak form. Namely, we know that every linearly independent set of vectors in $\mathbb{R}^n$ has at most n elements. That's nice, but we want more. For instance, suppose that I have a linearly independent subset S not only of $\mathbb{R}^3$ but of some plane P in $\mathbb{R}^3$. We would then like to say not only that $\#S \leq 3$ but that in fact $\#S \leq 2$: it seems geometrically clear that a set of three linearly independent vectors in $\mathbb{R}^3$ should not lie in any plane but rather span all of $\mathbb{R}^3$. We feel this way because we think of a plane as a **two-dimensional object**: it is given as the span of two linearly independent vectors. However, so far as we've shown so far, a plane in $\mathbb{R}^3$ might also be the span of a linearly independent set of three vectors, and a plane in $\mathbb{R}^{17}$ might also be the span of a linearly independent set of up to 17 vectors.

In crisper terms, so far as we know, a plane $P \subset \mathbb{R}^n$ is defined as a subspace with a basis $\mathcal{B}$ with $\#\mathcal{B} = 2$, but how do we know that it does not have a different basis $\mathcal{B}'$ – recall that every nonzero subspace has infinitely many bases – with $\#\mathcal{B}' = 3$? We hope that this cannot happen. In fact, to have a notion of "dimension" of a subspace V of $\mathbb{R}^n$, what we need is that any two bases of V have the same number of elements.

The following elegant result will allow us to show this and more.

Lemma 3.10. *(Steinitz Exchange Lemma) Let V be a subspace of $\mathbb{R}^n$. Let $(v_1, \ldots, v_m)$ be a linearly independent sequence of vectors in V, and let $(w_1, \ldots, w_n)$ be a sequence of vectors in V with $V = \operatorname{span} w_1, \ldots, w_n$. Then $m \leq n$, and – after reordering the w_i's, if necessary – we have $V = \operatorname{span} v_1, \ldots, v_m, w_{m+1}, \ldots, w_n$.*

The Steinitz Exchange Lemma is an archetypical example of a "lemma": namely a result that is a little too technical to stand alone on its own right, but does the lion's share of the work of the sexier theorem that is used to prove. And indeed, although the proof of Lemma 3.10 is not so bad, it is a little technical, so before we give it let's see the remarkable consequences that it has. It *immediately* implies the following fundamental result.

Theorem 3.11. *Let $V \subset \mathbb{R}^n$ be a subspace.*
a) Let $S \subset V$ be a finite linearly independent subset, and let $T \subset V$ be a finite spanning set. Then $\#S \leq \#T$.

b) Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two bases for V. Then both $\mathcal{B}_1$ and $\mathcal{B}_2$ are finite sets, and

$$\#\mathcal{B}_1 = \#\mathcal{B}_2.$$

Well, just to be sure: part a) of Theorem 3.11 is precisely what the conclusion $m(= \#S) \leq n(= \#T)$ of the Exchange Lemma is telling us. As for part b): first, we get a nice new proof that any linearly independent subset $S = \mathbb{R}^n$ has at most n elements, by applying the Exchange Lemma with $T = \{e_1, \ldots, e_n\}$. Second, a basis of V is precisely a subset of V which is both linearly independent and spans V, so if $\mathcal{B}_1$ and $\mathcal{B}_2$ are bases of V, we can apply the Exchange Lemma with linearly independent subset $\mathcal{B}_1$ and spanning set $\mathcal{B}_2$ to get

$$\#\mathcal{B}_1 \leq \#\mathcal{B}_2$$

and then we can turn things around, applying the Exchange Lemma with linearly independent subset $\mathcal{B}_2$ and spanning set $\mathcal{B}_1$ to get

$$\#\mathcal{B}_2 \leq \#\mathcal{B}_1.$$

We conclude $\#\mathcal{B}_1 = \#\mathcal{B}_2$.

Now that we are fully invested, we turn to the *Proof of the Steinitz Exchange Lemma*. We will show in fact that for any $1 \leq i \leq m$, then $i \leq n$, and after reordering the w's if necessary, we have

$$\operatorname{span} v_1, \ldots, v_i, w_{i+1}, w_n = V.$$

Taking $i = m$ gives us the result we want. The point of this setup is precisely to allow us to proceed by induction on i.

Base Case: $i = 1$. Sure, $1 \leq n$. So we just need to "exchange" v_1 for one of the w's. We do this in two steps: first we simply put v_1 into our list: $v_1, w_1, \ldots, w_n$. But now we have too many w's: we need to take one out and still get a spanning set. This is really what we want to show: if we can do that, then we just reorder the remaining $n-1$ w's and call them $w_2, \ldots, w_n$. For this: $w_1, \ldots, w_n$ is a spanning set for V and $v_1 \in V$, so we can write $v_1 = \alpha_1 w_1 + \ldots + \alpha_n w_n$ for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Moreover, not all the α_i's can be zero: if so, $v_1 = 0$, but v_1 was an element of a linearly independent sequence, so it can't be 0. Suppose for instance that $\alpha_j \neq 0$. Then we can write w_j as a linear combination of the other w's and v:

$$\alpha_j w_j = v - \alpha_1 w_1 - \ldots - \alpha_{j-1} w_{j-1} - \alpha_{j+1} w_{j+1} - \ldots - \alpha_n w_n,$$

so

$$w_i = \frac{1}{\alpha_j} v_1 - \frac{\alpha_1}{\alpha_j} w_1 - \ldots - \frac{\alpha_{i-1}}{\alpha_j} w_{j-1} - \frac{\alpha_{j+1}}{\alpha_j} w_{j+1} - \ldots - \frac{\alpha_n}{\alpha_j} w_n.$$

Thus indeed can remove w_j without changing the span, getting the n element spanning sequence $v, w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n$. The bit about reordering is just that we will change the indices on the w's around so as to write this as $v, w_2, \ldots, w_n$.

Induction Step: Let $1 \leq i < m$, and suppose (inductively) that $i \leq n$ and after reordering we have a spanning sequence $v_1, \ldots, v_i, w_{i+1}, \ldots, w_n$. We need to show that $i + 1 \leq n$ and that we can exchange v_{i+1} for one of the w's. First: since $i \leq n$, we need to rule out the possibility that $i = n$ (if so, $i < n$, so $i + 1 \leq n$). If $i = n$, then this means we have already exchanged out all the w's, so $v_1, \ldots, v_i$ is a spanning set for V. But since $i < m$, $i + 1 \leq m$, so we have another vector v_{i+1} in V, which must then be a linear cominbation of $v_1, \ldots, v_i$, contradicting the assumed linear independence of $v_1, \ldots, v_m$. Having negotiated that slightly tricky

part, the rest of the argument is the same as the base case: first add in the next v vector, getting a spanning sequence $v_1, \ldots, v_i, v_{i+1}, w_{i+1}, \ldots, w_n$. Now we need to remove one of the w's. Since we already had a spanning sequence, there are $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$v_{i+1} = \alpha_1 v_1 + \ldots + \alpha_i v_i + \alpha_{i+1} w_{i+1} + \ldots + \alpha_n w_n.$$

Moreover, it cannot be that $\alpha_{i+1}, \ldots, \alpha_n$ are all 0: if so, we would have

$$v_{i+1} = \alpha_1 v_1 + \ldots + \alpha_i v_i,$$

and again this contradicts the linear independence of $v_1, \ldots, v_m$. Thus there is a j with $i+1 \leq j \leq n$ such that $\alpha_j \neq 0$, so as above we can write

$$w_j = \frac{1}{\alpha_j} v_{i+1} - \frac{\alpha_1}{\alpha_j} v_1 - \ldots - \frac{\alpha_i}{\alpha_j} v_i - \frac{\alpha_{i+1}}{\alpha_j} w_{i+1} - \ldots - \frac{\alpha_{j-1}}{\alpha_j} w_{j-1} - \frac{\alpha_{j+1}}{\alpha_j} w_{j+1} - \ldots - \frac{\alpha_n}{\alpha_j} w_n.$$

This shows that we can remove w_j and still get a spanning sequence

$$v_1, \ldots, v_{i+1}, w_{i+1}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_n.$$

Changing the indices on the w's, we write this as

$$v_1, \ldots, v_{i+1}, w_{i+2}, \ldots, w_n,$$

and we're done.

Remark: I don't know as much about the historical development of linear algebra as I should. There is the following complicating factor: some time in the early 20th century, the algebraic parts of mathematics became much more "abstract" following work of the golden gods Hilbert, Noether and Artin, among others. This abstract approach to mathematics has proven to be both very powerful and in many ways simpler than the previous, more numerical/concrete approach. Of course it is difficult for students to grasp at first: in a way, the difficulties you have grappling with mathematical abstraction recapitulate those of many contemporaries of those aforementioned golden gods: famously, a very eminent but "old-fashioned" mathematician Paul Gordan was alarmed at the way Hilbert proved vastly more general results about certain "rings of invariants" than Gordan had over the course of a long career. The gist of it is very similar to the dichotomy between showing that a subspace of $\mathbb{R}^n$ has a finite spanning set by a theoretical argument versus giving an algorithmic procedure for actually producing such a finite set. (Compare especially the proof of Corollary 3.9 to the more explicit procedure which is given just afterwards.) Hilbert was one of the first mathematicians to realize that it can be *much easier* to prove that something like a finite spanning set exists than to give an explicit recipe for writing one down, and he exploited this brilliantly. Gordan remarked: "This is not mathematics; this is theology." History has proven that Gordan was wrong: the abstract approach is most certainly mathematics. Proponents of the abstract approach were similarly disdainful of Gordan: "Er war ein Algorithmiker," wrote Max Noether (Emmy Noether's father, and a great mathematician in his own right...though probably not quite as good as his daughter) in Gordan's obituary. Nowadays – and especially with the ubiquitousness of modern computers – mathematics well understands that "algorithmikers" (i.e., algorithm-makers) can be leading mathematicians too. In our golden age both the abstract and the concrete approaches are extremely important.

The change in the approach to mathematics that took place in the early 20th century makes it hard for contemporary mathematicians like me to understand exactly how mathematicians in the 19th century and previously understood linear algebra. From a superficial perspective it often looks like deeper, more recondite topics were studied before earlier, more basic topics, e.g. determinants before matrices.

Steinitz's Exchange Lemma was developed by Ernst Steinitz, a German mathematician who lived from 1871 to 1928. (This was not the same Steinitz who was the world's first chess champion, although the next world chess champion, Emanuel Lasker, was in fact a student of Emmy Noether who did some important work.) Like much of Steinitz's work, it seems to have been somewhat neglected in his own time, but contemporary mathematics has firmly embraced it. In particular, in my opinion more core linear algebra content resides in the Steinitz Exchange Lemma than any other single result in this course.

3.6. **The Dimension of a Subspace.**

Let V be a subspace of $\mathbb{R}^n$. By Theorem 3.11, V has a finite basis; every basis of V is finite; and any two finite bases of V have the same number of elements. We may therefore define the **dimension** of V to be the number of elements in any basis for V.

Proposition 3.12. *Let $V \subset W \subset \mathbb{R}^n$ be subspaces. If $\dim V = \dim W$, then $V = W$.*

Proof. Let $\mathcal{B}_V = \{v_1, \ldots, v_d\}$ be a basis for V. Then $\mathcal{B}_V$ is a linearly independent subset of W, so by Theorem 3.8 there is a basis $\mathcal{B}_W$ for W containing $\mathcal{B}_V$. Since $\mathcal{B}_V \subset \mathcal{B}_W$ are finite sets with the same cardinality, $\mathcal{B}_V = \mathcal{B}_W$, and thus

$$V = \operatorname{span} \mathcal{B}_V = \operatorname{span} \mathcal{B}_W = W.$$

□

Example 3.13. *a) The zero-subspace has dimension 0: its only basis is $\varnothing$.*
b) $\mathbb{R}^n$ has dimension n, since $e_1, \ldots, e_n$ is a basis.
c) A line in $\mathbb{R}^n$ (passing through the origin) can be formally defined as a 1-dimensional subspace of $\mathbb{R}^n$.
d) A plane in $\mathbb{R}^n$ (passing through the origin) can be formally defined as a 2-dimensional subspace of $\mathbb{R}^n$.

Proposition 3.14. *Let $A \in M_{m,n}$. Then the dimension of the null space of A is precisely the number of free variables in* rref A. *Thus*

$$\dim \operatorname{nullity} A + \operatorname{rank} A = n.$$

Example 3.15. *Recall that a **hyperplane** in $\mathbb{R}^n$ is a subspace of the form*

$$H = \{x \in \mathbb{R}^n \mid x \cdot n = 0\}$$

for some nonzero vector n. We claim that the hyperplanes are precisely the $(n-1)$-dimensional subspaces of $\mathbb{R}^n$.
Step 1: Let H be a hyperplane. Then H is the null space of the $1 \times n$ matrix A with n as its row. Since $n \neq 0$, M has rank one, and nullity $A = n - 1$.

3.7. **Dimensions of Intersections; Independent Subspaces.**

Theorem 3.16. *Let V and W be subspaces of $\mathbb{R}^n$. Then*

$$\dim V + W = \dim V + \dim W - \dim V \cap \dim W.$$

Proof. Let $r = \dim V \cap W$, $k = \dim V$, $\ell = \dim W$. Let $v_1, \ldots, v_r$ be a basis for $V \cap W$, and extend it to a basis $v_1, \ldots, v_r, u_{r+1}, \ldots, u_k$ for V and again to a basis $v_1, \ldots, v_r, w_{k+1}, \ldots, w_\ell$ for W. We claim that

$$\mathcal{B} = \{v_1, \ldots, v_r, u_{r+1}, \ldots, u_k, w_{r+1}, \ldots, w_\ell\}$$

is a basis for $V + W$. If so, then indeed

$$\dim V + W = \dim V + \dim W - \dim V \cap W$$

and the result follows. Since $\mathcal{B}$ is obtained as the union of spanning sets for V and W, it is a spanning set for $V + W$, so it remains to show that $\mathcal{B}$ is linearly independent. Let $\alpha_1, \ldots, \alpha_\ell \in \mathbb{R}$ be such that

$$\alpha_1 v_1 + \ldots + \alpha_r v_r + \alpha_{r+1} u_{r+1} + \ldots + \alpha_k u_k + \alpha_{k+1} w_{k+1} + \ldots + \alpha_\ell w_\ell = 0.$$

We rewrite this as

$$z = \alpha_1 v_1 + \ldots + \alpha_r v_r + \alpha_{r+1} u_{r+1} + \ldots + \alpha_k u_k = -\alpha_{k+1} w_{k+1} - \ldots - \alpha_\ell w_\ell.$$

The left hand side lies V and the right hand side lies in W, so $z \in V \cap W$. Since $v_1, \ldots, v_r, u_{r+1}, \ldots, u_k$ is linearly independent, this implies $\alpha_{r+1} = \ldots = \alpha_k = 0$. Since $v_1, \ldots, v_w, w_1, \ldots, w_\ell$ is linearly independent, this implies $\alpha_1 = \ldots = \alpha_r = \alpha_{k+1} = \ldots = \alpha_\ell = 0$. □

Theorem 3.17. *For subspaces V and W of $\mathbb{R}^n$, the following are equivalent:*
(i) $\dim V + W = \dim V + \dim W$.
(ii) $V \cap W = \{0\}$.
(iii) If $v \in V \backslash \{0\}$ and $w \in W \backslash \{0\}$, then the sequence (v, w) is linearly independent.
(iv) If L_1 is a linearly independent list in V and L_2 is a linearly independent list in W, then (L_1, L_2) is a linearly independent list.
(v) If $\mathcal{B}_1$ is an ordered basis for V and $\mathcal{B}_2$ is an ordered basis for W, then $(\mathcal{B}_1, \mathcal{B}_2)$ is an ordered basis for $V + W$.

Proof. (i) $\iff$ (ii) follows from Theorem 3.16: since $\dim V + \dim W - \dim V + W = \dim V \cap W$, we have $\dim V + \dim W = \dim V + W$ if and only if $\dim V \cap W = 0$.
(ii) $\iff$ (iii): If v is a nonzero vector in V and w is a nonzero vector in W, then (v, w) is linearly dependent if and only if $w = \alpha v$ for some nonzero $\alpha \in \mathbb{R}$. If this happens, then w is a nonzero vector in $V \cap W$, so $\dim V \cap W \geq 1$. Conversely, if $\dim V \cap W \geq 1$, then taking any nonzero $v \in V \cap W$ we get a linearly dependent sequence (v, v) with $v \in V$ and $v \in W$.
(ii) $\implies$ (iv): Let $L_1 = (v_1, \ldots, v_k)$ and $L_2 = (w_1, \ldots, w_\ell)$, and suppose $\alpha_1, \ldots, \alpha_{k+\ell}$ are real numbers such that

$$\alpha_1 v_1 + \ldots + \alpha_k v_k + \alpha_{k+1} w_1 + \ldots + \alpha_{k+\ell} w_\ell = 0.$$

Equivalently,

$$\alpha_1 v_1 + \ldots + \alpha_k v_k = -\alpha_{k+1} w_1 - \ldots - \alpha_{k+\ell} w_\ell.$$

If the left hand side is zero, then since L_1 is linearly independent, $\alpha_1 = \ldots = \alpha_k = 0$, and then also the right hand side is zero and by linear independence of L_2, $\alpha_{k+1} = \ldots = \alpha_{k+\ell} = 0$. Similarly if the right hand side is zero. The only

other possibility is that both sides are nonzero, and thus we get a nonzero vector in $V \cap W$, contradicting (ii).
(iv) $\implies$ (v): By applying (iv) we get that $(\mathcal{B}_1, \mathcal{B}_2)$ is linearly independent. And we always get a spanning set for $V + W$ by taking the union of a spanning set of V and a spanning set of W.
(v) $\implies$ (i): By contraposition: if $V \cap W$ were nonempty, then any nonzero vector would be part of an ordered basis $\mathcal{B}_1$ for V and an ordered basis $\mathcal{B}_2$ for W and then $(\mathcal{B}_1, \mathcal{B}_2)$ is a linearly dependent list, contradicting our assumption. □

We describe the equivalent conditions of Theorem 3.17 by saying that V and W are **independent subspaces**, and in this case we sometimes write $V \oplus W$ instead of $V + W$. This notation is analogous to writing $S \coprod T$ for the union of two sets which are known to be disjoint.

3.8. **Rank Revisited.**

For any matrix $A \in M_{m,n}$, we defined $\operatorname{rank} A$ to be the number of leading entries in $\operatorname{rref} A$ (or in any row echelon form of A). And we have seen the uses of that definition in solving systems of linear equations and elsewhere. However, one must admit that our definition of rank is not the most graceful one: it is a number that we associate to a matrix after performing a certain algorithm (Gaussian reduction) on it. In this section we pursue other, more intrinsic definitions of rank, with applications to an improved Rank Nullity Theorem and the important result that $\operatorname{rank} A = \operatorname{rank} A^T$ for all A.

Recall also that the **row space** $R(A)$ of $A \in M_{m,n}$ is the subspace of $\mathbb{R}^n$ spanned by the row of A. The elementary row operations change the rows but not the row space: one way to see this is to think of an elementary row operation as premultiplication by an elementary matrix: $A \mapsto EA$. In general, the rows of BA are linear combinations of the rows of A – c.f. § 2.3 – so this shows that the row space of EA is contained in the row space of A. But E is invertible, and applying the same argument with E^{-1} and EA in place of E and A gives that the row space of $A = E^{-1}(EA)$ is contained in the row space of EA, so the row spaces of A and EA are equal.

Perhaps the argument of the preceding paragraph is too slick for its own good. The reader may prefer a more hands-on approach:

Exercise: Check more concretely that the row space of EA is equal to the row space of A for every elementary matrix by considering each of the three elementary row operations separately, and for each one, explicitly writing each row of EA as a linear combination of the rows of A, and explicitly writing each row of A as a linear combination of the rows of EA.

Anyway, since there are elementary matrices $E_1, \ldots, E_r$ such that

$$\operatorname{rref}(A) = E_r \cdots E_1 A,$$

we deduce the following.

Proposition 3.18. *For any $A \in M_{m,n}$, the row space of A is equal to the row space of* rref A.

Along with the row space of $A \in M_{m,n}$ we can also consider its **column space** $C(A)$, the subspace of $\mathbb{R}^m$ spanned by the columns of A. But beware:

Warning: Elementary row operations *need not* preserve the column space!

Here is an example meant to convince you that this is not even close to being true: take $A \in M_{m,1}$, i.e., consisting of a single column. Then so long as A is not the zero matrix, rref $A = e_1 = (1, \ldots, 0)$. Thus the column space of rref A is the line spanned by e_1. But the matrix A we started with could be given by any nonzero vector, so its span need not be the span of e_1.

Back to the row space and the rank. Here is the first important result.

Theorem 3.19. *For any matrix $A \in M_{m,n}$,* rank A *is the dimension of the row space of A.*

Proof. Since A and rref A have the same rank (by definition) and the same row space (by Proposition 3.18), it is enough to show that the number of leading entries in rref A is equal to rank rref A. In fact I claim that a basis for rref A is simply obtained by taking the nonzero rows. Indeed, the nonzero rows certainly span the row space: the zero rows contribute nothing to the span. Moreover they are linearly independent because each row contains a leading entry: we can be sure that a finite set $v_1, \ldots, v_k$ of vectors is linearly independent if, as we move from left to right, each v_i has a nonzero entry in a coordinate where all the vectors to the left of it have zero entries (this observation is formalized in the following exercise and you are asked to prove it: it's not hard!). So indeed the nonzero rows of rref A form a basis for the row space. Since the number of nonzero rows of rref A is precisely rank A, we're done. □

Exercise: Prove the **Eyeball Criterion for Linear Independence**: Let $v_1, \ldots, v_k$ be vectors in $\mathbb{R}^n$. Suppose that for all $1 \leq i \leq k$, there is some coordinate (i.e, $1 \leq j \leq n$) such that v_i is nonzero in the jth coordinate, but all $v_{i'}$ with $i' < i$ are zero in the jth coordinate. Show that $\{v_1, \ldots, v_k\}$ is a linearly independent set.

Thus we can think of row reduction as taking the hard-to-see property of linear independence of a set of vectors and reworking it until it is visible to the naked eye. Note also that this gives us a method for testing a finite list of vectors $v_1, \ldots, v_k$ for linear independence: make a matrix with these vectors as *row vectors* and row reduce: if the rref has rank k, the vectors are linearly independent; otherwise they are linearly dependent. Is this interesting? We already know how to row reduce matrices to test for linear independence. But, wait – this is a **different test** than the one we've seen before: according to the definition of linear independence, $v_1, \ldots, v_k$ are linearly independent if and only if we when we make a matrix with these vectors as *column vectors*, then the rref has no free variables. That's not the same test, and we have begun to uncover the deep relationship between the rows and columns of a matrix.

Let's push it farther: recall that the **nullity** $\operatorname{nullity} A$ of $A \in M_{m,n}$ is the number of free variables in $\operatorname{rref} A$. Since every variable is either a free variable or a pivot variable (i.e., a column containing a leading entry), we get the Rank-Nullity Theorem:

$$\forall A \in M_{m,n},\ \operatorname{rank} A + \operatorname{nullity} A = n.$$

Let $v_1, \ldots, v_m$ be vectors in $\mathbb{R}^n$; let $A \in M_{m,n}$ be the matrix with ith row v_i. Then:

$$\begin{aligned} &\operatorname{rank} A = m \\ \iff &v_1, \ldots, v_m \text{ is linearly independent} \\ \iff &\operatorname{nullity}(A^T) = 0 \\ \iff &\operatorname{rank} A^T = m. \end{aligned}$$

Thus $A \in M_{m,n}$ has rank m iff A^T has rank m.

Let's go even farther: let $A \in M_{m,n}$ have rank k. Then there are some k rows of A which are linearly independent: let $B \in M_{k,n}$ be the matrix formed by these rows. By what we just said, $\operatorname{rank} B^T = k$. But sine B^T simply consists of some of the rows of A^T, we must have

$$\operatorname{rank} A = k = \operatorname{rank} B = \operatorname{rank} B^T \leq \operatorname{rank} A^T.$$

So for *any* matrix, $\operatorname{rank} A \leq \operatorname{rank} A^T$. Applying this inequality to A^T we get $\operatorname{rank} A^T \leq \operatorname{rank}(A^T)^T = A$. Thus, we've proven:

Theorem 3.20. *For any matrix $A \in M_{m,n}$,*

$$\operatorname{rank} A = \operatorname{rank} A^T.$$

Equivalently,

$$\dim R(A) = \dim C(A).$$

The rank of A^T is often called the **column rank** of A (as seems reasonable). Thus Theorem 3.20 is often abbreviated as **row rank equals column rank**. This is actually one of the least obvious of the fundamental theorems relating dimensions of subspaces, because the row space $R(A)$ and the column space $C(A)$ of A are usually not in any sense "the same" subspace: indeed, $R(A)$ is a subspace of $\mathbb{R}^n$ and $C(A)$ is a subspace of $\mathbb{R}^m$. Nevertheless these subspaces of *different Euclidean spaces* always have the same dimension. That's quite a deep result!

We want to give yet another interpretation of $\operatorname{rank} A$. Namely, consider the associated linear transformation $L_A : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$.

Theorem 3.21. *For any $A \in M_{m,n}$, $\operatorname{rank} A = \dim \operatorname{Image} L_A$.*

Proof. The proof will use the (very easy) fact that the null space of A is the kernel of L_A, so that by Rank-Nullity, $\dim \operatorname{Ker} L_A + \operatorname{rank} A = n$ and thus

$$\operatorname{rank} A = n - \dim \operatorname{Ker} L_A. \tag{5}$$

Really what we will show is

$$\dim \operatorname{Image} L_A = n - \dim \operatorname{Ker} L_A. \tag{6}$$

And of course, combining (5) and (6) gives $\operatorname{rank} A = \dim \operatorname{Image} L_A$. To show (6) we argue as follows: let $k = \dim \operatorname{Ker} L_A$, so there is a basis $v_1, \ldots, v_k$ of $\operatorname{Ker} L_A$. Like any basis for a subspace, we can extend this to a basis for all of

$\mathbb{R}^n$, say $v_1, \ldots, v_k, v_{k+1}, \ldots, v_n$. Now I claim that $L(v_{k+1}), \ldots, L(v_n)$ is a basis for $\operatorname{Image} L_A = L_A(\mathbb{R}^n)$: if so, we're done, because we have $n-k$ vectors in our basis, so $\dim \operatorname{Image} L_A = n - k = n - \dim \operatorname{Ker} L_A$. Now $\operatorname{Image} L_A$ is spanned by $L(v_1), \ldots, L(v_n)$, and since $L(v_1) = \ldots = L(v_k) = 0$, it is certainly also spanned by $L(v_{k+1}), \ldots, L(v_n)$. So it remains to check the linear independence: suppose we have $\alpha_{k+1}, \ldots, \alpha_n$ such that

$$\alpha_{k+1} L(v_{k+1}) + \ldots + \alpha_n L(v_n) = 0.$$

Then

$$0 = \alpha_{k+1} L(v_{k+1}) + \ldots + \alpha_n L(v_n) = L(\alpha_{k+1} v_{k+1} + \ldots + \alpha_n v_n),$$

so $\alpha_{k+1} v_{k+1} + \ldots + \alpha_n v_n \in \operatorname{Ker} L_A$. Since $\operatorname{Ker} L_A$ is spanned by $v_1, \ldots, v_k$, there are $\alpha_1, \ldots, \alpha_k$ such that

$$\alpha_{k+1} v_{k+1} + \ldots + \alpha_n v_n = \alpha_1 v_1 + \ldots + \alpha_k v_k,$$

or

$$(-\alpha_1) v_1 + \ldots + (-\alpha_k) v_k + \alpha_{k+1} v_{k+1} + \ldots + \alpha_n v_n.$$

Since $v_1, \ldots, v_n$ is linearly independent this means $-\alpha_1 = \ldots = -\alpha_k = \alpha_{k+1} = \ldots = \alpha_n = 0$; this shows that $L(v_{k+1}), \ldots, L(v_n)$ is linearly independent. □

Here is an equivalent statement.

Theorem 3.22. *(Dimension Theorem) For any linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$, we have*

$$\dim \operatorname{Ker} L + \dim \operatorname{Image} L = \dim \mathbb{R}^n.$$

Exercise: Deduce Theorem 3.22 from Theorem 3.21.

Really the Rank-Nullity Theorem and the Dimension Theorem are *the same theorem*; the former is couched in matrix language, the latter in the language of linear transformations. This is a common theme in linear algebra: many if not most results can be expressed either way. This is useful because linear transformations are more conceptual and thus ultimately more useful for theorems and proofs, whereas it is critical to be able to phrase things in terms of matrices to do calculations.

As an example of this duality, let us use Theorem 3.21 (and Proposition 3.18) to give another proof that row rank equals column rank. Namely, for $A \in M_{m,n}$,

$$\dim R(A) = \operatorname{rank} A = \dim \operatorname{Image} L_A = \dim \operatorname{span} L_A(e_1), \ldots, L_A(e_n) = \operatorname{span} A(e_1), \ldots, A(e_n).$$

But $A(e_1), \ldots, A(e_n)$ are precisely the columns of A, so

$$\dim \operatorname{span} A(e_1), \ldots, A(e_n) = \dim C(A) = \operatorname{rank} A^T.$$

3.9. **Sylvester's Law of Nullity.**

Theorem 3.23. *(Sylvester's Law of Nullity) Let $A \in M_{m,n}$ and $B \in M_{n,p}$. Then*

$$\operatorname{nullity} AB \leq \operatorname{nullity} A + \operatorname{nullity} B. \tag{7}$$

Proof. If $Bv = 0$ then $ABv = A0 = 0$, so $\operatorname{nullity} B \subset \operatorname{nullity} AB$. Let $v_1, \ldots, v_a$ be a basis for B, and extend it to a basis $v_1, \ldots, v_a, w_1, \ldots, w_b$ for AB. We claim that $B(w_1), \ldots, B(w_b)$ is linearly independent. The argument for this is essentially

the same one used in the proof of Theorem 3.21: let $\alpha_1, \ldots, \alpha_b \in \mathbb{R}$ be such that $\alpha_1 B(w_1) + \ldots + \alpha_b B(w_b) = 0$. Then

$$0 = \alpha_1 B(w_1) + \ldots + \alpha_a B(w_b) = B(\alpha_1 w_1 + \ldots + \alpha_b w_b),$$

so $\alpha_1 w_1 + \ldots + \alpha_b w_b \in B$. Thus there are $\beta_1, \ldots, \beta_a \in \mathbb{R}$ such that

$$\alpha_1 w_1 + \ldots + \alpha_b w_b = \beta_1 v_1 + \ldots + \beta_a v_a,$$

and by linear independence of $v_1, \ldots, v_a, w_1, \ldots, w_b$, this gives $0 = \alpha_1 = \cdots = \alpha_b (= \beta_1 = \ldots = \beta_a$, though we don't need this). Since $B(w_1), \ldots, B(w_b) \subset A$, this gives

$$\text{nullity } A \geq b,$$

and thus

$$\text{nullity } A + \text{nullity } B \geq b + \text{nullity } B = b + a = \text{nullity } AB.$$

□

Exercise: Maintain the notation of Theorem 3.23. Use the Rank-Nullity Theorem to deduce the following additional inequalities.
a) (**Sylvester's Rank Inequality**) $\text{rank } AB \geq \text{rank } A + \text{rank } B \leq \text{rank } AB + n$.
b) $\text{rank } B \leq \text{rank } AB + \text{nullity } A$.

Corollary 3.24. *a) Let $A_1, \ldots, A_N$ be matrices such that the product $A_1 \cdots A_N$ is defined. Then*

$$\text{nullity } A_1 \cdots A_N \leq \sum_{i=1}^{N} \text{nullity } A_i.$$

b) Let $A_1, \ldots, A_N \in M_{n,n}$ be such that $A_1 \cdots A_N = 0$. Then $\sum_{i=1}^{N} \text{nullity } A_i \geq n$.

Exercise: Prove Corollary 3.24.

4. Some Linear Transformations

4.1. Permutations.

4.2. Projections.

Let V, W be subspaces of $\mathbb{R}^n$ with $V \oplus W = \mathbb{R}^n$. Recall this means: $V + W = \mathbb{R}^n$ and $V \cap W = \{0\}$. We define a linear transformation $\pi_{V,W}$, the **projection onto V with respect to W**, as follows: the complementarity of V and W means precisely that every $x \in \mathbb{R}^n$ can be written uniquely as $x = v + w$ for $v \in v$, $w \in W$. Then we define

$$\pi_{V,W}(x) = \pi_{V,W}(v + w) = v.$$

To be slightly informal (but without sacrificing any mathematical accuracy), the map $\pi_{V,W}$ resolves a vector into the sum of its V component and its W component, keeps the V component and kills the W-component.

Proposition 4.1. *The projection map $\pi_{V,W} : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation.*

Proof. As usual, this is straightforward to show. Namely, let $x_1, x_2 \in \mathbb{R}^n$ and write $x_1 = v_1 + w_1$, $x_2 = v_2 + w_2$ with $v_i \in V$ and $w_i \in W$ for $i = i, 2$. Then $x_1 + x_2 = (v_1 + v_2) + (w_1 + w_2)$. It follows that

$$\pi_{V,W}(x_1 + x_2) = v_1 + v_2 = \pi_{V,W}(x_1) + \pi_{V,W}(x_2).$$

Similarly, for any $\alpha \in \mathbb{R}$, if $x = v + w$ then

$$\alpha x = \alpha v + \alpha w,$$

so

$$\pi_{V,W}(\alpha x) = \pi_{V,W}(\alpha v + \alpha w) = \alpha v = \alpha \pi_{V,W}(v).$$

□

Example 4.2. *Let V be the span of e_1 in $\mathbb{R}^2$, i.e., the line $y = 0$. Let w be any vector in $\mathbb{R}^2 \setminus V$, and put $W = \langle w \rangle$. Then $V \cap W = \{0\}$, hence*

$$\dim V + W = \dim V + \dim W \setminus \dim V \cap W = 1 + 1 - 0 = 2,$$

and thus $V + W = \mathbb{R}^2 = V \oplus W$. Let's find the standard matrix representation of $\pi_{V,W}$, i.e., the 2×2 matrix with columns $\pi_{V,W}(e_1)$ and $\pi_{V,W}(e_2)$. First,

$$\pi_{V,W}(e_1) = e_1.$$

To find $\pi_{V,W}(e_2)$, the key idea is to express e_2 as a linear combination of e_1 and w. For this, write $w = (x, y)$; since $w \notin \langle e_1 \rangle$, we must have $y \neq 0$. Then

$$w = xe_1 + ye_2,$$

and we solve for e_2, getting

$$e_2 = \frac{1}{y}w - \frac{x}{y}e_1$$

and thus

$$\pi_{V,W}(e_2) = \frac{1}{y}\pi_{V,W}(w) - \frac{x}{y}\pi_{V,W}(e_1) = 0 - \frac{x}{y} = \frac{-x}{y}.$$

So the standard matrix is

$$A = \begin{bmatrix} 1 & \frac{-x}{y} \\ 0 & 0 \end{bmatrix}.$$

Some things to notice about this matrix are: (i) it has a row of zeros so is not invertible. This is not surprising, since the linear transformation had a nontrivial kernel: it killed w. We also see that A takes the simplest form if $x = 0$, in other words if $W = \langle (0, y) \rangle = \langle e_2 \rangle$ is just the y-axis. In this special case we have

$$A = A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

a diagonal matrix. We get this case by taking w to be perpendicular to $v = e_1$.

There is a nice way of thinking about projection operators in terms of splitting a basis. Namely, let $\mathcal{B} = \{v_1, \ldots, v_n\}$ be a basis for $\mathbb{R}^n$. We split it in two, namely we choose subsets $\mathcal{B}_1, \mathcal{B}_2 \subset \mathcal{B}$ such that $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B}, \mathcal{B}_1 \cap \mathcal{B}_2 = \varnothing$. Then $\mathcal{B}_1$ and $\mathcal{B}_2$ are both subsets of a linearly independent set, hence linearly independent, hence $\mathcal{B}_1$ is a basis for $V_1 = \operatorname{span} \mathcal{B}_1$ and $\mathcal{B}_2$ is a basis for $V_2 = \operatorname{span} \mathcal{B}_2$.

Exercise: With notation above, show that $\mathbb{R}^n = V_1 \oplus V_2$.

Thus we can define a projection operator π_{V_1,V_2} associated to the splitting of the basis. Moreover, like any linear transformation, π_{V_1,V_2} is uniquely specified by what it does to the elements of any basis, so let's see what it does to $\mathcal{B}$. Well, we have split $\mathcal{B}$ into $\mathcal{B}_1$ and $\mathcal{B}_2$, and π_{V_1,V_2} keeps the elements of $\mathcal{B}_1$ and kills the elements of $\mathcal{B}_\in$:

$$\forall b_1 \in \mathcal{B}_\infty,\ \pi_{V_1,V_2}(b_1) = b_1,$$

$$\forall b_2 \in \mathcal{B}_\in,\ \pi_{V_1,V_2}(b_2) = 0.$$

Next we observe that any projection has the property that doing it and then doing it again is the same as doing it once:

$$\pi_{V,W} \circ \pi_{V,W} = \pi_{V,W}. \tag{8}$$

Indeed, for any $x = v + w \in \mathbb{R}^n$,

$$(\pi_{V,W} \circ \pi_{V,W})(x) = \pi_{V,W}(\pi_{V,W}(v+w)) = \pi_{V,W}(v) = v = \pi_{V,W}(x).$$

Let A be the standard matrix of $\pi_{V,W}$. Since composition of linear transformations corresponds to multiplication of matrices, the matrix version of (9) is

$$A^2 = A. \tag{9}$$

A matrix $A \in M_{n,n}$ satisfying (9) is called **idempotent**. The equation $A^2 = A$ looks a little silly from the perspective of high school algebra: if A is a real number it has precisely the solutions $A = 0$ and $A = 1$. Equivalently, these are the only idempotent 1×1 matrices. However, for any $n \geq 2$ there are many, many more solutions $A \in M_{n,n}$: namely the standard matrix associated to any projection operator.

Exercise: Show that if A is idempotent, so is $1 - A$.

Exercise: a) It follows from our discussion so far that if $A = \begin{bmatrix} 1 & b \\ 0 & 0 \end{bmatrix}$ for any $b \in \mathbb{R}$, then $A^2 = A$. Check this directly.
b) Write down an idempotent matrix $A \in M_{3,3}$ different from 0 and 1.
c) Write down a nondiagonal idempotent matrix $A \in M_{3,3}$.

That the standard matrix of a projection operator is idempotent is an instance of geometry (linear transformations) governing algebra (matrix operations). It seems a bit more interesting that the converse is also true.

Theorem 4.3. *Let $A \in M_{n,n}$ be an idempotent matrix: $A^2 = A$. Then:*
a) We have $\mathrm{Ker}(1 - A) = \mathrm{Image}(A)$ *and* $\mathrm{Ker}\, A = \mathrm{Image}(1 - A)$.
b) We have $\mathbb{R}^n = \mathrm{Image}(A) \oplus \mathrm{Image}(1 - A)$.
c) A is the standard matrix of the projection operator $\pi_{\mathrm{Image}(A),\mathrm{Image}(1-A)}$.

Proof. a) If $v \in \mathrm{Ker}(1 - A)$, then $0 = (1 - A)v = v - Av$, so $Av = v$, so $v \in \mathrm{Image}(A)$. Conversely, if $v \in \mathrm{Image}(A)$ then $v = Aw$ for some $w \in \mathbb{R}^n$, and then $(1 - A)v = (1 - A)(Aw) = Aw - A^2w = Aw - Aw = 0$. This shows

$$\mathrm{Ker}(1 - A) = \mathrm{Image}(A).$$

The equality $\mathrm{Ker}\, A = \mathrm{Image}(1-A)$ can be shown similarly, or by applying the above argument with $1 - A$ in place of A, which is valid since $1 - A$ is also idempotent by Exercise X.X.
b) Suppose $v \in \mathrm{Image}(A) \cap \mathrm{Image}(1 - A) = \mathrm{Image}(A) \cap \mathrm{Ker}(A)$. Then $v = Aw$ for some $w \in \mathbb{R}^n$ and also $Av = 0$, so

$$0 = Av = A(Aw) = A^2w = Aw = v.$$

This shows that $\mathrm{Image}(A) \cap \mathrm{Image}(1 - A) = 0$. To see that $\mathrm{Image}(A) + \mathrm{Image}(1 - A) = \mathbb{R}^n$ is even easier: we may write any $x \in \mathbb{R}^n$ as

$$x = Ax + (1 - A)x.$$

c) It is enough to see that if $x \in \text{Image}(A)$ then $Ax = x$ and if $x \in \text{Image}(1 - A)$ then $Ax = 0$. Really we've done this already, but nce again: if $x = Aw$ for some $w \in \mathbb{R}^n$ then $Ax = A(Aw) = A^2w = Aw = x$, and if $x = (1 - A)w$,then $Ax = A(1 - A)w = (A - A^2)w = 0$. □

4.3. **Reflections.** The above discussion of projections is easily modified so as to apply to another class of operators, the **reflections**. We begin in the same way, with a decomposition of $\mathbb{R}^n$ into complementary subspaces $\mathbb{R}^n = V_1 \oplus V_2$ and choices of bases $\mathcal{B}_1$ of V_1 and $\mathcal{B}_2$ of V_2, so that

$$\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$$

is a basis for $\mathbb{R}^n$. Whereas to get π_{V_1,V_2} we *kept* the elements of $\mathcal{B}_1$ and *killed* the elements of $\mathcal{B}_2$, we now define a **reflection** τ_{V_1,V_2} by:

$$\forall b_1 \in \mathcal{B}_1,\ \tau(b_1) = b_1,$$

$$\forall b_2 \in \mathcal{B}_2,\ \tau(b_2) = -b_2.$$

That is, instead of killing the basis elements of V_2, we **flip** them.
In more algebraic terms, we write any x in $\mathbb{R}^n$ uniquely as $v_1 + v_2$ with $v_1 \in V_1$ and $v_2 \in V_2$ and put

$$\tau_{V_1,V_2}(x) = \tau_{V_1,V_2}(v_1 + v_2) = v_1 - v_2.$$

Now let A be the standard matrix of the reflection τ_{V_1,V_2}. Above we found that the geometry of projection was faithfully recorded in the simple algebraic equation $A^2 = A$, so it is natural to ask whether the same kind of thing will hold for reflections. The answer is yes.

Proposition 4.4. *a) Let $\mathbb{R}^n = V_1 \oplus V_2$, and let τ_{V_1,V_2} be the corresponding reflection operator. Then*

$$\tau_{V_1,V_2} \circ \tau_{V_1,V_2} = 1_{\mathbb{R}^\times}. \tag{10}$$

Equivalently, if A is the standard matrix of τ_{V_1,V_2}, then

$$A^2 = 1. \tag{11}$$

b) Conversely, if $A \in M_{n,n}$ is such that $A^2 = 1$, then A is the standard matrix of a projection operator π_{V_1,V_2}, with

$$V_1 = \{x \in \mathbb{R}^n \mid Ax = x\}$$

and

$$V_2 = \{x \in \mathbb{R}^n \mid Ax = -x\}.$$

Proof. a) For $x \in \mathbb{R}^n$, write $x = v_1 + v_2$ with $v_1 \in V_1$ and $v_2 \in V_2$. Then

$$(\tau_{V_1,V_2} \circ \tau_{V_1,V_2})(x) = \tau_{V_1,V_2}(\tau_{V_1,V_2}(v_1 + v_2)) = \tau_{V_1,V_2}(\tau_{V_1,V_2}(v_1) + \tau_{V_1,V_2}(v_2))$$

$$= \tau_{V_1,V_2}(v_1 - v_2) = \tau_{V_1,V_2}(v_1) - \tau_{V_1,V_2}(v_2) = v_1 - (-v_2) = v_1 + v_2 = x.$$

This shows (10); since composition of linear operators corresponds to multiplication of matrices, (11) follows immediately.
b) Notice first that we defined V_1 to be the set of vectors "kept" (or, in more common parlance, "fixed") by A and V_2 to be the set of vectors "flipped" by A. A little thought shows that if A is the standard matrix of a reflection operator then V_1 and V_2 have to be as we defined them, and what is left to show is that $\mathbb{R}^n = V_1 \oplus V_2$.

If $x \in V_1 \cap V_2$, then $Ax = x$ and also $Ax = -x$, so $x = Ax = -x$ and thus $2x = 0$ so $x = 0$. Now let $x \in \mathbb{R}^n$. Here's a trick: put $y = x + Ax$ and $z = x - Ax$. Then

$$Ay = A(x + Ax) = Ax + A^2x = Ax + x = y,$$

so $y \in V_1$. Also

$$Az = A(x - Ax) = Ax - A^2x = Ax - x = -(x - Ax) = -z,$$

so $z \in V_2$. Finally,

$$x = \frac{1}{2}y + \frac{1}{2}z \in V_1 + V_2.$$

□

Remark: In more advanced linear algebra one works with vector spaces not only over $\mathbb{R}$ but over an arbitrary *field* of scalars. *Most* of linear algebra by its nature carries over to this general context with no changes in the statements of proofs, but there are some exceptions. Here we fundamentally used that $2 = 1 + 1 \neq 0$. If we took as our scalars the field $\mathbb{F}_2$ of two elements (or any field containing it), then we would in fact have $1 + 1 = 0$ and the above argument breaks down. In fact the result becomes false: there are matrices with $A^2 = 1$ which do not correspond to projection operators. The simplest such example is $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

5. Orthogonality

5.1. Orthogonal Matrices.

Lemma 5.1. *(Polarization Identity) For all $v, w \in \mathbb{R}^n$ we have*

$$v \cdot w = \frac{||v + w||^2 - ||v||^2 - ||w||^2}{2}. \tag{12}$$

Exercise: Prove Lemma 5.1.
(Hint: no need to think: just calculate the right hand side.)

Notation: For $1 \leq i, j \leq n$, we set $\delta_{i,j}$ to be 1 if $i = j$ and 0 otherwise.

We say that $v_1, \ldots, v_n \in \mathbb{R}^n$ forms an **orthonormal basis** if for all $1 \leq i, j \leq n$, $v_i \cdot v_j = \delta_{i,j}$. This is just a compact way of saying that we have an orthogonal basis of unit vectors.

Theorem 5.2. *For a matrix $A \in M_{n,n}$, the following are equivalent:*
(i) For all $v \in \mathbb{R}^n$, $||Av|| = ||v||$.
(ii) For all $v, w \in \mathbb{R}^n$, $Av \cdot Aw = v \cdot w$.
(iii) For every orthonormal ordered basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$, $Av_1, \ldots, Av_n$ is an orthonormal ordered basis.
(iv) We have that $Ae_1, \ldots, Ae_n$ is an orthonormal ordered basis for $\mathbb{R}^n$.
(v) $A^TA = AA^T = 1$.

Proof. (i) $\implies$ (ii): The follows from the polarization identity:

$$Av \cdot Aw = \frac{||Av + Aw||^2 - ||Av||^2 - ||Aw||^2}{2} = \frac{||v + w||^2 - ||v||^2 - ||w||^2}{2} = v \cdot w.$$

(ii) $\implies$ (iii): This is immediate: for all $1 \leq i, j \leq n$, we have

$$Av_i \cdot Av_j = v_i \cdot v_j = \delta_{i,j}.$$

(iii) $\implies$ (iv): Since $e_1, \ldots, e_n$ is an orthonormal ordered basis of $\mathbb{R}^n$, this is a special case of (iii).
(iv) $\implies$ (v): First recall that for any $A, B \in M_n$, if $AB = 1$ then also $BA = 1$. So it is enough to assume (iv) and show that $A^T A = 1$. The (i, j) entry of AA^T is the dot product of the ith row of A^T with the jth column of A, which is the dot product of the ith and jth columns of A, which is $Ae_i \cdot Ae_j$. Since the (i, j) entry of the identity matrix is δ_{ij}, this shows $A^T A = 1$.
(v) $\implies$ (i): It's equivalent to show that for all $v \in \mathbb{R}^n$, $Av \cdot Av = v \cdot v$. For this,

$$Av \cdot Av = (Av)^T Av = v^T A^T Av = v^T v = v \cdot v.$$

□

A matrix which satisfies the equivalent properties of Theorem 5.2 is called an **orthogonal matrix**. we denote by O_n the set of all $n \times n$ orthogonal matrices.

Exercise: Show that $A \in M_n$ is orthogonal $\iff$ A is invertible and $A^T = A^{-1}$.

Exercise: For $\theta \in \mathbb{R}$, show that

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

is orthogonal.

Exercise: Show that any permutation matrix is orthogonal.

Lemma 5.3. *If $A, B \in O_n$, $A^{-1} \in O_n$ and $AB \in O_n$.*

Proof. If $A \in O_n$, then $A^T = A^{-1}$. Thus $(A^{-1})^T = (A^T)^T = A = (A^{-1})^{-1}$, so $A^{-1} \in O_n$. Similarly, if $B \in O_n$ then $B^T = B^{-1}$, so

$$(AB)^T = B^T A^T = B^{-1} A^{-1} = (AB)^{-1}.$$

□

6. Determinants

7. Eigenvectors and Diagonalization

7.1. **Diagonalization.**

7.2. **Eigenvectors, Eigenvalues and Eigenspaces.**

A vector $v \in \mathbb{R}^n$ is an **eigenvector for a linear transformation** $L : \mathbb{R}^n \to \mathbb{R}^n$ if

(EV1) $v \neq 0$, and
(EV2) There is $\lambda \in \mathbb{R}$ such that $L(v) = \lambda v$.

Thus an eigenvector is a nonzero v such that $L(v)$ is a scalar multiple of v.

If v is an eigenvector, then the scalar λ such that $L(v) = \lambda v$ is unique: if $L(v) = \lambda_1 v = \lambda_2 v$, then $(\lambda_1 - \lambda_2)(v) = 0$. Since $v \neq 0$, this forces $\lambda_1 = \lambda_2$. This scalar is called the **eigenvalue** of v. Moreover, a scalar $\lambda \in \mathbb{R}$ is called an **eigenvalue for L** if there is some eigenvector v with eigenvalue λ.

Example 7.1. *Let D be a diagonal matrix with diagonal entries $d_1, \ldots, d_n$. Then each of the standard basis vectors $e_1, \ldots, e_n$, and the eigenvalues are (respectively) $d_1, \ldots, d_n$. In particular there is a basis – the standard basis! – of $\mathbb{R}^n$ consisting of eigenvectors of D.*

Remark 7.2. *Although the zero* vector *is not allowed to be an eigenvector, the zero* scalar *is allowed to be an eigenvalue, and this is an important case: 0 is an eigenvalue for L if and only if there is $0 \neq v$ such that $L(v) = 0v = 0$. Thus the eigenvectors with eigenvalue 0 are precisely the nonzero vectors in the kernel (or null space) of T, and L has 0 has an eigenvalue if and only if it is singular.*

For any $\lambda \in \mathbb{R}$ we define the λ-**eigenspace**

$$V_\lambda = \{v \in \mathbb{R}^n \mid L(v) = \lambda v\}.$$

In other words, V_λ consists of the eigenvectors for v with eigenvalue λ (if any) along with the zero vector. We also define the **geometric multiplicity** of λ as $\dim V_\lambda$.[7]

The following exercise is very easy but all-important: it tells us that computing eigenspaces is a special case of our favorite linear algebraic computation.

Exercise: Show that for all $\lambda \in \mathbb{R}$, V_λ is the null space of $\lambda I_n - A$.

Example 7.3. *We return to the case of a diagonal matrix $D \in M_{n,n}$ with diagonal entries $d_1, \ldots, d_n$. Earlier we saw that the standard basis vectors $e_1, \ldots, e_n$ are eigenvectors, with corresponding eigenvalues $d_1, \ldots, d_n$. Now we want to go further by computing all the eigenspaces. First, suppose $v = (x_1, \ldots, x_n)$ is an eigenvector for D. Then there is $\lambda \in \mathbb{R}$ such that*

$$\lambda v = (\lambda x_1, \ldots, \lambda x_n) = Dv = (d_1 x_1, \ldots, d_n x_n).$$

Thus for all $1 \leq i \leq n$, we have $\lambda x_i = d_i x_i$, so if $x_i \neq 0$ then $\lambda = d_i$. By definition of eigenvectors, $v \neq 0$ hence at least one x_i is nonzero, so $\lambda = d_i$. This shows that the only eigenvalues of D are the diagonal entries. Moreover, if for $1 \leq i \neq j \leq n$ we have both $x_i \neq 0$ and $x_j \neq 0$, then $d_i = \lambda = d_j$. In other words, if several components of v are nonzero, then the corresponding diagonal entries must all be equal; conversely when this happens we do indeed have $Dv = \lambda v$. This shows:

$$V_{d_i} = \operatorname{span}_{1 \leq j \leq n}\{e_j \mid d_j = d_i\}.$$

The dimension of V_{d_i} is the number of $1 \leq j \leq n$ such that $d_j = d_i$.
*(**Subexample**: For instance, if*

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

then $V_1 = \operatorname{span} e_1, e_3$, $V_2 = \operatorname{span} e_2$, $v_3 = \operatorname{span} e_4$.)
Further, the eigenspaces are independent and the sum of their dimensions is n.

[7] Yes, this is a lot to swallow. Our pedagogical strategy here is to put all the basic definitions in one place for easy reference, and then explore the consequences of these definitions in a more leisurely manner.

Exercise: As the previous example indicates, cleaner bookkeeping arises for diagonal matrices if we assume that repeated diagonal entries occur in blocks of consecutive terms (unlike the subexample above, in which the two 1's occur nonconsecutively). Show that any diagonal matrix is similar to a diagonal matrix with this property.

Proposition 7.4. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation.*
a) For any $\lambda \in \mathbb{R}$, V_λ is a subspace of $\mathbb{R}^n$.
b) $V_\lambda \supsetneq \{0\} \iff \lambda$ is an eigenvalue of T.

Proof. It is tempting to leave this as an exercise – it is quite straightforward – but because of its importance to our narrative we prefer to give a complete proof.
a) As usual, this is easy: $0 \in V_\lambda$. Further, if $v, w \in V_\lambda$ and $\alpha \in \mathbb{R}$, then

$$L(\alpha v + w) = \alpha L(v) + L(w) = \alpha\lambda v + \lambda w = \lambda(\alpha v + w),$$

so $\alpha v + w \in V_\lambda$. By the One-Step Linear Transformation Test, V_λ is a subspace.
b) According to our definitions, λ is an eigenvalue for T if and only if there is a nonzero vector $v \in \mathbb{R}^n$ with $T(v) = \lambda v$; this occurs if and only if $V_\lambda \neq \{0\}$. □

Proposition 7.5. *Let $L : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation, and let $\lambda_1, \ldots, \lambda_n$ be distinct real numbers. Then the eigenspaces $V_{\lambda_1}, \ldots, V_{\lambda_n}$ are independent:*

$$V_{\lambda_1} + \ldots + V_{\lambda_n} = V_{\lambda_1} \oplus \ldots \oplus V_{\lambda_n}.$$

Proof. According to X.X it suffices to show: for any nonzero vectors $v_i \in V_{\lambda_i}$, the set $\{v_1, \ldots, v_n\}$ is linearly independent. Suppose not; then, after reordering the vectors there is some $2 \leq k \leq n$ such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_k v_k = 0 \tag{13}$$

with *every* $\alpha_k \neq 0$ (we cannot have $k = 1$ because each v_i is nonzero) and among all such relations we may choose one with k as small as possible. If we can conjure up a similar linear dependence relation among $k-1$ vectors, we get a contradiction and we'll be done. Well, the big idea is to apply L to (13), getting

$$\alpha_1\lambda_1 v_1 + \alpha_2\lambda_2 v_2 + \ldots + \alpha_k\lambda_k v_k = 0. \tag{14}$$

Multiplying (13) by λ_1 and subtracting what we get from (14), we obtain

$$\alpha_2(\lambda_2 - \lambda_1)v_2 + \alpha_3(\lambda_3 - \lambda_1)v_3 + \ldots + \alpha_k(\lambda_n - \lambda_1)v_k = 0. \tag{15}$$

Since the λ_i's are distinct, for all $2 \leq i \leq k$, $\lambda_i - \alpha_1 \neq 0$, and thus (15) is a linear dependence relation with all nonzero coefficients but with $k-1$ terms instead of k terms: contradiction. □

Exercise: Give a much shorter proof of Proposition 7.5 when $n = 2$.

Corollary 7.6. *A linear transformation $L : \mathbb{R}^n \to \mathbb{R}^n$ has at most n eigenvalues.*

Proof. If we had $n+1$ eigenvalues $\lambda_1, \ldots, \lambda_{n+1}$, then $V_{\lambda_1} + \ldots + V_{\lambda_{n+1}}$ would be a subspace of $\mathbb{R}^n$ of dimension $\sum_{i=1}^{n+1} \dim V_{\lambda_i} \geq \sum_{i=1}^{n+1} 1 = n+1$. □

Why do we care about all this eigenstuff anyway?? Because of the following result.

Theorem 7.7. *Let $A \in M_{n,n}$, and let $P \in M_{n,n}$ be invertible.*
a) The following are equivalent:
(i) $P^{-1}AP$ is diagonal.
(ii) The columns of P are eigenvectors for L_A.
b) The following are equivalent:

(i) A is diagonalizable.
(ii) There is a basis $\{v_1, \ldots, v_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors for A.
(iii) There is an eigenvalue for A, and if $\lambda_1, \ldots, \lambda_k$ are the eigenvalues, then

$$\mathbb{R}^n = V_{\lambda_1} \oplus \ldots \oplus V_{\lambda_k}.$$

Proof. a) (i) $\implies$ (ii): Suppose that $P^{-1}AP = D$, where D is diagonal with diagonal entries $\lambda_1, \ldots, \lambda_n$. Let v_i be the ith column of P. Then for all $1 \leq i \leq n$,

$$\lambda_i e_i = De_i = P^1APe_i = P^{-1}Av_i.$$

Multiplying on the left by P gives

$$\lambda_i v_i = \lambda_i P(e_i) = P(\lambda_i e_i) = PP^{-1}Av_i = Av_i,$$

so $v_i \in V_{\lambda_i}$. Since P is invertible, each v_i is nonzero, so v_i is an eigenvector for L_A.
(ii) $\implies$ (i): Suppose that for all $1 \leq i \leq n$, there is $\lambda_i \in \mathbb{R}$ such that $Av_i = \lambda_i v_i$, and let $\mathcal{B} = (v_1, \ldots, v_n)$. Then the change of basis formula gives

$$A_{L,\mathcal{B}} = P^{-1}AP.$$

Moreover, since $Av_i = \lambda_i v_i$, the ith column of $A_{L,\mathcal{B}}$ is $\lambda_i e_i$: thus $A_{L,\mathcal{B}}$ is a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$.
b) The equivalence of (i) and (ii) follows immediately from part a). Moreover, if we have a basis of $\mathbb{R}^n$ consisting of eigenvectors for L, then breaking it apart into subsets $\mathcal{B}_i$ consisting of eigenvectors for the eigenvalue λ_i gives a direct sum decomposition $\mathbb{R}^n = V_{\lambda_1} \oplus \ldots \oplus V_{\lambda_k}$. And conversely: given such a direct sum decomposition, we take $\mathcal{B}_i$ to a basis for V_{λ_i}, and then $\mathcal{B} = \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_k$ is a basis for $\mathbb{R}^n$ consisting of eigenvectors. □

Exercise: a) Show that if $A^2 = A$, then A is diagonalizable.
b) Show that if $A^2 = 1$, then A is diagonalizable.

Recall that the determinant of a matrix is a **similarity invariant**: in concrete terms, this means that if $A, B \in M_{n,n}$ are similar – i.e., $B = PAP^{-1}$ for some invertible P – then $\det A = \det B$. But there is also a richer perspective: as we have seen, similarity of matrices is an equivalence relation on $M_{n,n}$, and thus it partitions $M_{n,n}$ into equivalence classes. However, if I give you two matrices $A, B \in M_{n,n}$, it is usually not so easy to tell whether they lie in the same equivalence class: in principle we would have to try conjugating A by *every* invertible matrix P to see whether we get B, but there are infinitely many such matrices so it is not clear that this can always be done in a practical manner. This is common in higher mathematics: for an interesting equivalence relation $\sim$ on a set X, it is not always clear how to check in practice whether two objects are equivalent. One strategy is to find an **invariant** of the equivalence relation. One can think of this as a function $f : X \to Y$ defined on the entire set such that if $x_1 \sim x_2$ then $f(x_1) = f(x_2)$. Then, if we have two objects x_1, x_2 such that $f(x_1) \neq f(x_2)$, we knw that they can't be equivalent. The determinant is such a function: if $\det A \neq \det B$, then A and B cannot be similiar. Unfortunately the converse does not hold, as we have seen. Thus we want further such invariants. If we are lucky, then eventually we will find a **complete set of invariants**, such that if all the invariants of x_1 and x_2 agree, then indeed $x_1 \sim x_2$.

In the case of similarity of matrices this can indeed be done, but unfortunately

the end of this story lies beyond the end of this course. But here are some further important invarints.

Theorem 7.8. *Let $A, P \in M_{n,n}$, and suppose P is invertible.*
a) For $v \in \mathbb{R}^n$, if $Av = \lambda v$, then

$$(P^{-1}AP)(P^{-1}v) = \lambda(P^{-1}v).$$

b) For all $\lambda \in \mathbb{R}$, let V_λ be the λ-eigenspace for A, and let W_λ be the λ-eigenspace for $P^{-1}AP$. Then

$$P^{-1}V_\lambda = W_\lambda.$$

c) For all $\lambda \in \mathbb{R}$, the geometric multiplicity of λ for A is equal to the geometric multiplicity of λ for $P^{-1}AP$.

Proof. a) This is such a straightforward computation that we leave it as an exercise to the reader in order to keep her awake.
b) Part a) says precisely that if v is an eigenvector for A, then $P^{-1}v$ is an eigenvector for $P^{-1}AP$. In other words, $P^{-1}V_\lambda \subset W_\lambda$. Conversely, if $w \in W_\lambda$ then $P^{-1}APw = \lambda w$, so $A(Pw) = P(\lambda w) = \lambda(Pw)$ and thus $Pw \in V_\lambda$: thus $PW_\lambda \subset V_\lambda$; applying P^{-1} to both sides gives $W_\lambda \subset P^{-1}V_\lambda$. Thus $P^{-1}V_\lambda = W_\lambda$.
c) Let $(v_1, \ldots, v_k)$ be an ordered basis for V_λ. Then $(P^{-1}(v_1), \ldots, P^{-1}(v_k))$ is an ordered basis for $P^{-1}V_\lambda = W_\lambda$, so $\dim V_\lambda = \dim W_\lambda$. □

Thus similar matrices have the same eigenvalues and the same geometric multiplicities. Thus for all $\lambda \in \mathbb{R}$, the geometric multiplicity of λ (which will be 0 if λ is not an eigenvalue) is a similarity invariant.

Exercise: Note that Theorem 7.8 does not say that similar matrices have *equal* λ-eigenspaces. Give an explicit example of similar matrices with distinct λ-eigenspaces for some λ. (Suggestion: any matrix which is diagonalizable but not diagonal will give rise to an example.)

Exercise: a) Let $A \in M_{n,n}$, and let $N \in \mathbb{Z}^+$. If v is an eigenvector for A with eigenvalue λ, show that v is an eigenvector for A^N with eigenvalue λ^N.
b) Recall that a matrix $A \in M_{n,n}$ is **nilpotent** if $A^N = 0$ for some $N \in \mathbb{Z}^+$. Show that a nilpotent matrix has precisely one eigenvalue: 0.
c) In the setting of part a), show that it is possible that the Nth powers of the eigenvalues of A do not give all of the eigenvalues of A^N. For instance, exhibit an A which has no eigenvalues but A^2 does.

Corollary 7.9. *If $A \in M_{n,n}$ has n eigenvalues, then it is diagonalizable.*

Proof. As in the proof of Corollary 7.9, each eigenvalue contributes at least one dimension to the sum of the eigenspaces, hence if we have n eigenvalues then the sum of the eigenspaces is all of $\mathbb{R}^n$. □

It is part of Theorem 7.7 that in order for a matrix $A \in M_{n,n}$ to be diagonalizable it must have at least one eigenvalue. This also follows from Example X.X and Exercise X.X: similar matrices have the same eigenvalues, and diagonal matrices have their diagonal entries as eigenvalues, hence at least one. For any $n \geq 1$, an $n \times n$ matrix may well have only a single eigenvalue and still be diagonalizable.

If $A \in M_{n,n}$ has at least one and fewer than n eigenvalues, whether it is diagonalizable or not depends upon the geometric multiplicities: for A to be diagonalizable, it is necessary and sufficient that the geometric multiplicities sum to n. Thus, the fewer eigenvalues we have, the larger each of their geometric multiplicities must be in order for the matrix to be diagonalizable. Here is the extreme case:

Example 7.10. *Let $\lambda \in \mathbb{R}$. The scalar matrix λI_n has λ as an eigenvalue. In fact, $\mathbb{R}^n = V_\lambda$; in particular λ is the only eigenvalue. Conversely, this property is evidently characteristic of scalar matrices: if $\mathbb{R}^n = V_\lambda$ for some λ, this precisely means that $Av = \lambda v$ for all $v \in \mathbb{R}^n$. In particular this holds for $v = e_1, \ldots, e_n$, so $A = \lambda I_n$ is a scalar matrix.*

Thus a matrix $A \in M_{n,n}$ with exactly one eigenvalue λ is diagonalizable if and only if it is the scalar matrix λI_n. In particular, a nondiagonal matrix with at most one eigenvalue cannot be diagonalizable.

Exercise: Let $A \in M_{n,n}$ be a singular matrix with a single eigenvalue. Show that A is diagonalizable if and only if $A = 0$.

Example 7.11. *Let $R : \mathbb{R}^2 \to \mathbb{R}^2$ be rotation through an angle $\theta \in [0, 2\pi)$. If $\theta \neq 0, \pi$, then there are no eigenvectors: rotating a nonzero vector through an angle of θ changes the line that it lies on. (Rotating a vector through an angle of 0 fixes it, and rotating a vector through an angle of π scales it by -1.) Having no eigenvectors –equivalently, having no eigenvalues) – R is* not *diagonalizable.*

Exercise: For any even positive integer n, construct a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^n$ with no eigenvectors.

7.3. **The Characteristic Polynomial.**

So far we have developed just the beginning of the theory of eigenvectorss and diagonalization: there are many (many!) results which give necessary and/or sufficient conditions for a linear transformation (or a matrix) to admit a basis of eigenvectors. However, before we press on in this direction we should first address a computational issue.

Question 7.12. *Let $A \in M_{n,n}$. How do we compute the eigenvalues of A?*

Notice that computing the eigenvalues is the key to computing the eigenspaces. Indeed, recall Exercise X.X: for any $\lambda \in \mathbb{R}$, the eigenspace V_λ is simply the null space of $\lambda I_n - A$, so we can compute it via row reduction. However, there are of course infinitely many real numbers, so we can't simply compute the null spacess of all $\lambda I_n - A$. In some cases one can successfully guess (or know) some good candidates for λ. But in general this would be difficult.

Example 7.13. *Let $M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$. We claim that $\lambda = \frac{1 \pm \sqrt{5}}{2}$ are eigenvalues. Indeed,*

$$(\frac{1+\sqrt{5}}{2})I_2 - M = \begin{bmatrix} \frac{1+\sqrt{5}}{2} - 1 & -1 \\ -1 & \frac{1+\sqrt{5}}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{5}-1}{2} & -1 \\ -1 & \frac{\sqrt{5}+1}{2} \end{bmatrix}.$$

Multiplying the first row by $\frac{2}{\sqrt{5}-1}$ *and adding it to the second row, we get*

$$\begin{bmatrix} \frac{\sqrt{5}-1}{2} & -1 \\ 0 & \frac{\sqrt{5}+1}{2} - \frac{2}{\sqrt{5}-1} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{5}-1}{2} & -1 \\ 0 & 0 \end{bmatrix}.$$

So the null space is nontrivial and $v_1 = (\frac{2}{\sqrt{5}-1}, 1) = (\frac{\sqrt{5}+1}{2}, 1)$ *is an eigenvector. Similarly,*

$$(\frac{1-\sqrt{5}}{2})I_2 - M = \begin{bmatrix} \frac{-1-\sqrt{5}}{2} & -1 \\ -1 & \frac{1-\sqrt{5}}{2} \end{bmatrix}.$$

Multiplying the first row by $\frac{2}{-1-\sqrt{5}}$ *and adding it to the second row, we get*

$$\begin{bmatrix} \frac{-1-\sqrt{5}}{2} & -1 \\ 0 & \frac{2}{1+\sqrt{5}} + \frac{1-\sqrt{5}}{2} \end{bmatrix} = \begin{bmatrix} \frac{-1-\sqrt{5}}{2} & -1 \\ 0 & 0 \end{bmatrix}.$$

Again the null space is nontrivial, and $v_2 = (\frac{1-\sqrt{5}}{2}, 1)$ *is an eigenvector.*

Probably few of us would have guessed taking $\lambda = \frac{1\pm\sqrt{5}}{2}$ without some extra insight/information. We need a technique for computing eigenvalues.

Now determinants come to our rescue: for any $A \in M_{n,n}$ and $\lambda \in \mathbb{R}$, we have:
λ is an eigenvalue $\iff$
$\mathrm{Ker}(\lambda I_n - A) \supsetneq \{0\}$ $\iff$
$\lambda I_n - A$ is singular $\iff$
$\det(\lambda I_n - A) = 0$.

Thus the eigenvalues are precisely the real numbers λ such that $\det(\lambda I_n - A) = 0$. This is useful because of the following result.

Proposition 7.14. *For any* $A \in M_{n,n}$ *and* $\lambda \in \mathbb{R}$, $\det(\lambda I_n - A)$ *is a monic*[8] *polynomial of degree* n *with real coefficients.*

Proof. The (i,j) entry of $\lambda I_n - A$ is $\lambda\delta(i,j) - a_{ij}$: when $i = j$ this is a linear polynomial in λ; otherwise it is a real number. Because the determinant of any matrix is a certain polynomial expression involving the matrix entries, $\det \lambda I_n - A$ is certainly a polynomial in λ. More precisely each of the $n!$ terms in the determinant is, up to ± 1, obtained by multiplying a choice of one entry from each row and column of the matrix, hence each term is a product of n factors each of which is either a constant or a linear polynomial, so each term is a polynomial of degree at most n. In order to get a degree n polynomial we must have a factor of λ every time, and this happens precisely when we choose the diagonal entries (or, if you like, the identity permutation): this term contributes $(\lambda - a_{11}) \cdots (\lambda - a_{nn})$, which is a monic degree n polynomial in λ. If we add a monic polynomial of degree n to a polynomial of smaller degree, the leading term cannot change so we get another monic polynomial of degree λ. □

We define the **characteristic polynomial of A**

$$\chi_A(t) = \det(tI_n - A).$$

[8]A polynomial is **monic** if its highest order term has leading coefficient 1.

(Why did we switch from λ to t? This is a fastidiousness on my part: I want to distinguish between a polynomial and the numbers that we plug into it.) The characteristic polynomial $\chi_A(t)$ is precious to us because (i) we know how to compute it, just by computing the determinant (e.g. by row reduction to upper triangular form) and (ii) as we have seen, the eigenvalues of A are precisely the **real roots** of $\chi_A(t)$, i.e., the real numbers λ such that $\chi_A(\lambda) = 0$.

Exercise: Let $A \in M_{n,n}$.
a) Suppose each a_{ij} is a rational number. Show that all the coefficients of $\chi_A(t)$ are rational numbers.
b) Supose each a_{ij} is an integer. Show that all the coefficients of $\chi_A(t)$ are integers.

In fact the characteristic polynomial is the final tool we need to get a complete computational method for determining whether $A \in M_{n,n}$ is diagonalizable and if so finding a matrix P such that $P^{-1}AP$ is diagonal.

Step 1: We compute the characteristic polynomial $\chi_A(t)$ and find all the real roots $\lambda_1, \ldots, \lambda_k$. These are the eigenvalues of A.

Step 2: For each eigenvalue λ, we compute the eigenspace $V_\lambda = \mathrm{Ker}(\lambda I_n - A)$.

Step 3: A is diagonalizable iff $\sum_{i=1}^k \dim V_\lambda = n$. If so, we find a basis $\mathcal{B}_i$ for each V_{λ_i}, and let P be the matrix with columns the elements of the $\mathcal{B}_i$'s.

Example 7.15. *(Fibonacci Matrices)*

The above example indicates a certain complication when it comes to the eigenvalues of a matrix. Although we have tried not to call attention to it, although we are in theory working with the real numbers as our "scalars", in practice most of our matrices have had rational numbers as entries. Gaussian reduction applied to a matrix with rational entries will yield a rref matrix also with rational entries, and thus a basis of the null space consisting of vectors with rational entries can always be found. Similarly orthogonalization does not introduce irrational numbers (ortho*normalization* does, but only square roots). This explains why most of our calculations have involved only rational numbers. It would be great if the characteristic polynomial of a matrix with rational entries necessarily had only rational numbers as roots, because then we can find all of them easily using the following high school result.

Theorem 7.16. *(Rational Roots Theorem) Let $P(t) = a_n t^n + \ldots + a_1 t + a_0$ be a polynomial with integer coefficients a_i. Then the only possible nonzero rational numbers r such that $P(r) = 0$ are of the form $\pm\frac{c}{d}$ where c and d are nonzero integers, $d \neq 0$, a_0 is divisible by c and a_n is divisible by d.*

As Example X.X shows, the characteristic polynomial can have rational (even integral) coefficients but still have irrational numbers as roots. In fact a polynomial with integral coefficents will have irrational roots most of the time: consider for instance the case of a quadratic polynomial $P(t) = t^2 + bt + c$, in which the quadratic formula gives us a square root. You have probably solved enough quadratic equations in your time to understand that if you choose integers b and c at random, the discriminant $b^2 - 4ac$ is very unlikely to be a perfect square, so the roots are

very likely to be irrational numbers.

Exercise: a) Let $a_0, a_1, a_2 \in \mathbb{R}$, and let $A = \begin{bmatrix} 0 & 0 & -a_0 \\ 1 & 0 & -a_1 \\ 0 & 1 & -a_2 \end{bmatrix}$. Show:

$$\chi_A(t) = (t - a_0)(t - a_1)(t - a_2).$$

b) Let n be a positive integer, let $a_0, \ldots, a_{n-1} \in \mathbb{R}$, and let $p(t) = t^n + a_{n-1}t^{n-1} + \ldots + a_1 t + a_0$, i.e., an arbitrary monic degree n polynomial. Let

$$A_p = \begin{bmatrix} 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & \ldots & 0 & -a_1 \\ 0 & 1 & \ldots & 0 & -a_2 \\ \vdots & & & & \\ 0 & 0 & \ldots & 1 & -a_{n-1} \end{bmatrix}.$$

Show that $\chi_{A_p}(t) = p(t)$. (A_p is called the **companion matrix** of p.)

Exercise: Let A be the companion matrix of t^n.
a) Show that A is nilpotent.
b) Show that $A^n = 0$.
c) Show that $A^{n-1} \neq 0$.

Exercise: Let $A \in M_{n,n}$ be nilpotent.
a) Let $k \geq 1$ be such that $A^k \neq 0$. Thus there is $v \in \mathbb{R}^n$ such that $A^k v \neq 0$. Show that there is $\ell \geq 0$ such that $A^{k+\ell} v \neq 0$ and $A^{k+\ell+1} v = 0$.
b) Show that $A^\ell v \in \operatorname{Ker} A^{k+1} \setminus \operatorname{Ker} A^k$.
c) Deduce that $A^n = 0$.

Proposition 7.17. *Let $A, P \in M_{n,n}$ with P invertible. Then*

$$\chi_{P^{-1}AP}(t) = \chi_A(t).$$

In other words, the characteristic polynomial is a similarity invariant.

Proof. We use the fact that scalar matrices commute with all matrices:

$$\begin{aligned} \chi_{P^{-1}AP}(t) &= \det(tI_n - P^{-1}AP) = \det(tP^{-1}I_nP - P^{-1}AP) \\ &= \det(P^{-1}(tI_n - A)P) = \det(P^{-1}) \det(tI_n - A) \det P \\ &= \chi_A(t) \det(P)^{-1} \det P = \chi_A(t). \end{aligned}$$

□

Example 7.18. *Let $A \in M_{n,n}$ be (upper or lower) triangular. Then $tI_n - A$ is also triangular, so its determinant is the product of the diagonal entries:*

$$\chi_A(t) = (t - a_{11}) \cdots (t - a_{nn}).$$

*In particular $\chi_A(t)$ is **split**: that is, it factors into a product of linear polynomials. Not every polynomial is split: e.g. $t^2 + 1$ is not, hence neither is any polynomial of the form $(t^2 + 1)g(t)$ for a nonzero polynomial $g(t)$.*

Theorem 7.19. *Let $A \in M_{n,n}$. If A is diagonalizable – or even **triangularizable**, i.e., similar to a triangular matrix – then $\chi_A(t)$ is split.*

Proof. This follows immediately from the two previous results. □

Perhaps the previous result looks a bit abstruse. In fact it is given here to help us out: i.e., to tell us that in certain situations diagonalization is hopeless so we need not compute the eigenspaces.

Example 7.20. *Let $A \in M_{3,3}$ have characteristic polynomial $t(t^2+1)$. (Recall that by Exercise X.X there is at least one $n \times n$ matrix with any given monic degree n polynomial as its characteristic polynomial.) Then the only eigenvalue of A is $\lambda = 0$, so the only way that A could be diagonalizable is if it is the zero matrix. But the zero matrix has characteristic polynomial t^3. So A is not diagonalizable.*

Let $A \in M_{4,4}$ have characteristic polynomial $(t-1)(t-2)(t^2+1)$. Then A has two eigenvalues, $\lambda = 1$ and $\lambda = 2$. The argument of the preceding paragraph does not apply. But it would be a waste of time to compute the eigenspaces V_1 and V_2: because $\chi_A(t)$ is not split, the matrix cannot be diagonalizable.

Having a split characteristic polynomial is not enough for a matrix to be diagonalizable: take a nonzero nilpotent matrix. However, the following weaker result is true, which we leave as a challenging exercise.

Theorem 7.21. *For $A \in M_{n,n}$, the following are equivalent:*
a) The characteristic polynomial $\chi_A(t)$ is split.
b) A is similar to an upper triangular matrix.
c) A is similar to a lower triangular matrix.

Exercise: Prove Theorem 7.21. (Suggestion: to show e.g. that a) $\implies$ b), work by induction. Since the characteristic polynomial is split, there is an eigenvector v_1; complete that to an ordered basis $v_1, \ldots, v_n$ and consider the matrix of A with respect to this basis: it is in block form $\begin{bmatrix} \lambda & * \\ 0 & A' \end{bmatrix}$, where $A' \in M_{n-1,n-1}$. Relate $\chi_{A'}(t)$ to $\chi_A(t)$ and apply the induction hypothesis.)

We now want to identify a further situation in which diagonalizability is hopeless. Let D be a diagonal matrix with diagonal entries $d_1, \ldots, d_n$. As a special case of Example X.X, we know that $\chi_D(t) = (t-d_1)\cdots(t-d_n)$. Comparing with Example X.X we see that not only is $\chi_D(t)$ split with roots precisely the eigenvalues of D, but that the **multiplicity** of each root λ – i.e., the number of occurrences of the linear factor $t-\lambda$ – is equal to $\dim V_\lambda$, i.e., to the geometric multiplicity of λ. This motivates the following definition.

For an eigenvalue λ of $A \in M_{n,n}$, the **algebraic multiplicity** of λ is equal to its multiplicity as a root of the characteristic polynomial $\chi_A(t)$. Since $\chi_A(t)$ is a similarity invariant, so are the algebraic multiplicities of the eigenvalues. Moreover we deduce the following result.

Proposition 7.22. *Let $A \in M_{n,n}$. Then A is diagonalizable if and only if $\chi_A(t)$ is split and for all eigenvalues λ, the geometric multiplicity of λ is equal to the algebraic multiplicity of λ.*

Proof. We have seen that if A is diagonal, then $\chi_A(t)$ is split and the algebraic and geometric multiplicities coincide. Since all of these are similarity invariants, this is a necessary condition for diagonalizability. Conversely, if $\chi_A(t)$ is split then

it factors as a product of n (not necessarily distinct) linear factors, and if all of the geometric multiplicities are equal to the algebraic multiplicities, then the sum of the geometric multiplicities is equal to the sum of the algebraic multiplicities, which (since $\chi_A(t)$ is split!) is equal to n. Thus A is diagonalizable. □

Example 7.23. *Suppose that $A \in M_{4,4}$ has characteristic polynomial $\chi_A(t) = (t-1)^2(t-3)(t-4)$. Since $\chi_A(t)$ is split, A may or may not be diagonalizable: we need to do some computations. We begin by computing V_1 and its dimension. If $\dim V_1 = 1$, then the geometric multiplicity of $\lambda = 1$ is less than its algebraic multiplicity, so there is no need to compute V_3 and V_4: A is not diagonalizable. Conversely if $\dim V_2 = 2$ then since $\dim V_3 \geq 1$ and $\dim V_4 \geq 1$ we must have $\dim V_2 = 2$, $\dim V_3 = \dim V_4 = 1$ and thus $\dim V_2 + \dim V_3 + \dim V_4 = 4 = \dim \mathbb{R}^4$, so A is diagonalizable.*

In our discussion we have saved the following result for last: it seems more technical and less useful than the others.

Theorem 7.24. *The geometric multiplicitiy is always less than or equal to the algebraic multiplicity. More precisely: let λ be an eigenvalue for $A \in M_{n,n}$. Then $\dim V_\lambda$ is less than or equal to the multiplicity of λ as a root of $\chi_A(t)$.*

Proof. Let $v_1, \ldots, v_k$ be a basis for V_λ and extend it to a basis $v_1, \ldots, v_n$ for $\mathbb{R}^n$. The matrix of A with respect to this basis has the block form

$$B = \begin{bmatrix} \lambda I_k & * \\ 0 & A' \end{bmatrix},$$

where $A' \in M_{n-k,n-k}$. Since B is similar to A, we have

$$\chi_A(t) = \chi_B(t) = \det(tI_n - B).$$

In order to compute this determinant we need only row reduce to get an upper triangular matrix. We can do so by performing row operations on the last $n-k$ rows only, so as to make the lower right corner the identity matrix; doing so we acquire a factor of $\chi_{A'}(t)$, so that

$$\chi_A(t) = \chi_{A'}(t) \begin{bmatrix} (t-\lambda)I_k & * \\ 0 & I_{n-k} \end{bmatrix} = (t-\lambda)^k \chi_{A'}(t).$$

So the multiplicity of λ as a root of $\chi_A(t)$ is at least k. □

7.4. **An Alternative to the Characteristic Polynomial.**

We denote by $\mathbb{R}[t]$ the set of all polynomials $p(t) = a_n t^n + \ldots + a_1 t + a_0$ with real coefficients. Polynomials are ubiquitous and flexible algebraic objects (much like vector spaces and matrices). We can think of them on one hand as formal expressions which can be added and multiplied. On the other hand, we can "plug things into them". E.g. in calculus a polynomial is usually thought of as a function $\mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R} \mapsto p(x)$. But we can also plug in an $n \times n$ matrix, with the convention that the constant term a_0 denotes the scalar matrix $a_0 I_n$. Thus e.g. if $p(t) = t^2 + t + 1$, then

$$p(A) = A^2 + A + I_n.$$

We say that a matrix A **satisfies** a polynomial $p(t)$ if $p(A) = 0$.

Exercise: Let $A, B \in M_{n,n}$ be similar matrices, and let $p(t)$ be a polynomial. Suppose A satisfies $p(t)$. Show that also B satisfies $p(t)$.

Theorem 7.25. *Let $A \in M_{n,n}$. Then there is a monic polynomial $p(t)$ of degree at most n^2 such that $p(A) = 0$.*

Proof. We identify $M_{n,n}$ with $\mathbb{R}^{n^2}$, e.g. by sending the matrix $A = (a_{ij})$ to the vector $(a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{nn})$. In this way we consider the powers $I = A^0$, A^1, A^2, ... as elements of $\mathbb{R}^{n^2}$. Because $\dim \mathbb{R}^{n^2} = n^2$, there must be some nontrivial linear relation among $I_n == A^0, \ldots, A^{n^2}$. Arguing in the usual manner there is a $k \leq n^2$ such that $A^0, \ldots, A^{k-1}$ are linearly independent and

$$A^k = c_{k-1}A^{k-1} + \ldots + c_1 A + C_0 I_n.$$

Then if $p(t) = t^k - c_{k-1}t^{k-1} - \ldots - c_1 t - c_0$, we have $p(A) = 0$. □

In the above proof we found a monic polynomial $p(t)$ of *least degree* k such that $p(A) = 0$. In fact this polynomial is unique. This may seem surprising at first, but the argument is simple: suppose $q(t)$ is another monic polynomial of minimal degree such that $q(A) = 0$. Put $r(t) = p(t) - q(t)$. Then $r(A) = p(A) - q(A) = 0 - 0 = 0$. Since $p(t)$ and $q(t)$ are both monic of the same degree, the highest order terms cancel out and r has smaller degree. If r is not the zero polynomial then we may write

$$r(t) = d_\ell t^\ell + \ldots + d_1 t + d_0$$

and $d_\ell \neq 0$. Then $\frac{1}{d_\ell} r(t)$ is a monic polynomial of degree $\ell < k$ and $\frac{1}{d_\ell} r(A) = \frac{1}{d_\ell} 0 = 0$: this contradicts the minimality of $p(t)$. Thus it must be that $r(t)$ is the zero polynomial: i.e., $p(t) = q(t)$.

We call this unique monic polynomial of least degree satisfied by A the **minimal polynomial of A** and write it as $m_A(t)$.

Proposition 7.26. *Let $A \in M_{n,n}$, and let $p(t)$ be a polynomial satisfied by A. Then $m_A(t)$ divides $p(t)$: there is a polynomial $q(t)$ such that $p(t) = m_A(t)q(t)$.*

Proof. We use polynomial division with remainder: there are polynomials $q(t)$ and $r(t)$ such that

$$p(t) = m_A(t)q(t) + r(t)$$

and $\deg r < \deg m_A$. Now plug in A:

$$0 = p(A) = m_A(A)q(A) + r(A) = 0 \cdot q(A) + r(A) = r(A).$$

Thus $r(t)$ is a polynomial of smaller degree than $m_A(t)$ satisfied by A; as we saw above, this means that $r(t)$ is the zero polynomial and thus $p(t) = m_A(t)q(t)$. □

Proposition 7.27. *Let $A \in M_{n,n}$, and let $p(t)$ be a monic polynomial satisfied by A. Then for every eigenvalue λ of A, $p(\lambda) = 0$.*

Proof. Let v be an eigenvector for λ, i.e., a nonzero vector in $\mathbb{R}^n$ such that $Av = \lambda v$. Then since $A^k v = \lambda^k v$, adding these up we find that

$$0v = p(A)v = p(\lambda)v.$$

Since $v \neq 0$, we must have $p(\lambda) = 0$.

□

A monic polynomial $p(t)$ is **squarefree split** if it is a product of *distinct* linear factors. Thus e.g. $t^2+t = t(t+1)$ is squarefree split and $t^3+t^2 = t^2(t+1)$ is split but not squarefree split.

Exercise: Let $f(t)$ and $g(t)$ be monic polynomials. If $f(t)$ divides $g(t)$ and $g(t)$ is squarefree split, show that $f(t)$ is squarefree split.

Theorem 7.28. *For $A \in M_{n,n}$, the following are equivalent:*
(i) A is diagonalizable.
(ii) There is a squarefree split polynomial $p(t)$ such that $p(A) = 0$.
(iii) The minimal polynomial $m_A(t)$ is squarefree split.

Proof. (i) $\implies$ (ii): By Exercise X.X, the set of polynomials satisfied by a matrix is a similarity invariant, so we may as well assume that A is diagonal. Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of A (i.e., the distinct real numbers which comprise the diagonal entries of A). Let $p(t) = (t-\lambda_1)\cdots(t-\lambda_k)$, and observe that $p(t)$ is squarefree split. We claim that A satisfies $p(t)$. Indeed, $p(A) = (A-\lambda_1 I_n)\cdots(A-\lambda_k I_n)$ is a diaonal matrix with ith diagonal entry equal to $(a_{ii}-\lambda_1)\cdots(a_{ii}-\lambda_k) = 0$ since each $a_i i$ is equal to one of the λ's.
(ii) $\implies$ (iii): Since $p(A) = 0$ and $m_A(t)$ is the minimal polynomial, we have $m_A(t)$ divides $p(t)$, so by Exercise X.X, $m_A(t)$ is squarefree split.
(iii) $\implies$ (ii) is immediate.
(ii) $\implies$ (i): Let $p(t) = (t-r_1)\cdots(t-r_k)$ with $r_1, \ldots, r_k$ distinct be a squarefree split polynomial satisfied by A: thus

$$(A-r_1 I_n)\cdots(A-r_k I_n) = 0.$$

By Corollary 3.24, we have

$$\sum_{i=1}^{k} \dim V_{r_i} = \sum_{i=1}^{k} \text{nullity}(A-r_i I_n) \geq n,$$

so A is diagonalizable by Theorem 7.7. □

Theorem 7.28 seems in many ways more insightful than the characterization of diagonalization in terms of algebraic and geometric multiplicities. For one thing, if A has rational entries, then $m_A(t)$ has rational coefficients and its computation requires only row operations with matrices with rational entries. If our primary goal is to determine whether A is diaonalizable then computin $m_A(t)$ and determining whether it is squarefree split is faster and more straightforward than computing eigenspaces.

The result also gives us insight into some of our previous results. For instance, we saw that every (not necessarily orthogonal) projection and reflection was diagonalizable. From our present persepctive, this is immediate: a matrix A is a projection if $A^2 = A$, i.e., if A satisfies the squarefree split polynomial $t^2 - t = t(t-1)$. Similarly, A is a reflection if $A^2 = I_n$, i.e., if A satisfies the squarefree split polynomial $t^2 - 1 = (t+1)(t-1)$.

There is a natural question that our discussion so far has (somewhat nontraditionally) avoided. Namely, we have defined two polynomials attached to A, the minimal polynomial $m_A(T)$ and the characteristic polynomial $\chi_A(t)$. How are they

related?

Here is one case: if A is diagonalizable, then the characteristic polynomial is split, so is of the form $\prod_{i=1}^k (t-\lambda_i)^{r_i}$. In the proof of Theorem 7.28 we showed that the squarefree part of this polynomial, namely $p(t) = \prod_{i=1}^k (t-\lambda-i)$, is satisfied by A. Thus the minimal polynomial $m_A(T)$ divides $\prod_{i=1}^k (t-\lambda_i)$. Looking back at that calculation, we see that we need each factor $t-\lambda_i$ to kill all the diagonal entries of $p(A)$, so that $m_A(t) = \prod_{i=1}^k (t-\lambda_i)$. In particular $m_A(t)$ divides $\chi_A(t)$: equivalently, A satisfies its characteritic polynomial $\chi_A(t)$ in this case.

Although the proof of the following result is beyond the scope of this course, we cannot resist giving the statement.

Theorem 7.29. *(Cayley-Hamilton) Let $A \in M_{n,n}$.*
a) A satisfies its characteristic polynomial: $\chi_A(A) = 0$. Equivalently, the minimal polynomial divides the characteristic polynomial.
b) The minimal polynomial and the characteristic polynomial have the same irreducible factors (although the characteristic polynomial may have them with larger multiplicities).

Example 7.30. *By pure brute force we will verify the Cayley-Hamilon Theorem for $n = 2$. Namely, let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then as we know, $\chi_A(t) = t^2-(a+d)t+(ad-bc)$. We compute*

$$\chi_A(A) = A^2 - (a+d)A + (ad-bc)I_2$$

$$= \begin{bmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2 \end{bmatrix} - \begin{bmatrix} a^2+da & ab+bd \\ ac+cd & ad+d^2 \end{bmatrix} + \begin{bmatrix} ad-bc & 0 \\ 0 & ad-bc \end{bmatrix}$$

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

Exercise: By pure brute force (but more of it!), verify the Cayley-Hamilton Theorem for $n = 3$.

Suppose that $\chi_A(t)$ is split. Then the irreducible factors of the characteristic polynomial are precisely $t-\lambda$ as λ runs through the eigenvalues of A. By part a) of the Cayley-Hamilton Theorem, the minimal polynomial $m_A(t)$ is also split, and by Proposition 7.27 (and the Root Factor Theorem: if $p(\lambda) = 0$ then $(t-\lambda)$ divides $p(t)$) for every eigenvalue λ, $t-\lambda$ divides $m_A(t)$. Thus in this case part b) follows from part a).

From our perspective, the new content of part b) of the Cayley-Hamilton Theorem lies in the case where the characteristic polynomial has irreducible quadratic factors, e.g. t^2+1. In these notes we have for better or worse insisted on working over the real numbers as a scalar field. If instead we admitted complex numbers as scalars then the "splitness" issue would evaporate and we would only need part a) of the Cayley-Hamilton Theorem.

Corollary 7.31. *For $A \in M_{n,n}$, the minimal polynomial $m_A(t)$ has degree at most n.*

Exercise: Prove Corollary 7.31.

Some history: The Cayley-Hamilton Theorem was stated by Arthur Cayley in 1858. Arthur Cayley (1821-1895) was the greatest English mathematician since Newton. In fact, in the long years in between Newton's death (1727) and Cayley's ascendancy (circa 1860), English mathematics lay remarkably fallow. The high esteem in which British pure mathematics has been held for the last century or so is probably due more to Cayley than anyone else. William Rowan Hamilton (1805-1865) was a leading Irish mathematician who deeply studied rotations in three-dimensional space and invented quaternions, among other things. By modern standards of pure mathematics, the following seems somewhere between amusing and scandalous: neither Cayley nor Hamilton even attempted a proof of the Cayley-Hamilton Theorem in the general case! In 1858 Cayley checked the $n = 2$ case – as we did in Example 7.30 – and in the $n = 3$ case (as we assigned as an exercise). On the basis of these calculations he was quite confident of the general case. Five years earlier, Hamilton had checked the result for rotations in $\mathbb{R}^3$. On this basis the result is named after them! It seems that the first proof of the general case was given by Georg Frobenius (a German mathematician and one of the true founders of modern algebra) in 1878.

Many proofs have since been given. However, elementary proofs of the Cayley-Hamilton theorem tend not to be very insightful or rewarding. We gave a proof in the diagonalizable case. Once one reaches a certain level of algebraic sophistication it is possible to explain by "pure thought" why the general case follows from this. Alternately, my notes [Cl-IS] (which are pitched at the advanced undergraduate / beginning graduate level) contain a proof of the Cayley-Hamilton Theorem built around the fact that any matrix is similar to a block diagonal matrix each of whose blocks is a companion matrix.

7.5. **The Spectral Theorem.**

Let $A \in M_{n,n}$. So far we have studied the question of whether there is an invertible $P \in M_{n,n}$ such that $P^{-1}AP$ is diagonal. One interpretation of this is that the columns of B form a new coordinate system for $\mathbb{R}^n$ with respect to which the linear transformation has a very simple structure.

However, not all coordinate systems in $\mathbb{R}^n$ are created equal. As we saw, the standard basis has a property that most other bases lack: it is an orthonormal basis, and this explains why if $v = a_1e_1 + \ldots + a_ne_n$, then the coefficient a_i is simply $v \cdot e_i$. Although from an algebraic perspective it is certainly very helpful to have any basis of eigenvectors, from a geometric perspective it would be more natural to have an orthonormal basis of eigenvectors. This motivates the following definition.

A matrix $A \in M_{n,n}$ is **orthogonally diagonalizable** if there is an orthogonal matrix P such that $P^{-1}AP$ is diagonal.

Exercise: Show that for $A \in M_{n,n}$ the following are equivalent.
(i) A is orthogonally diagonalizable.
(ii) A admits an orthonormal basis of eigenvectors.

(iii) A admits an orthogonal basis of eigenvectors.

It is perhaps not immediately clear that orthogonal diagonalizability is really a stronger condition than mere diagonalizability. Up until this point, our take on orthonormal bases is that they are nice but nothing special: if you have a basis and want an orthonormal basis, no problem: apply the Gram-Schmidt process. However, the Gram-Schmidt process usually does not preserve eigenvectors, and indeed it is not hard to come up with examples of matrices which admit bases for eigenvectors but no orthonormal bases. Indeed, consider $A \in M_{2,2}$ which has distinct eigenvalues λ_1, λ_2. Then the eigenspaces V_{λ_1}, V_{λ_2} are both lines in the plane. These lines might be orthogonal to each other and they might not: if not, there is clearly no orthogonal basis of eigenvectors.

Example 7.32. *Let* $A = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$. *The characteristic polynomial is* $\chi_A(t) = t^2 + t$, *so the eigenvalues are* $\lambda = 0, l$. *Solving for the nullspaces, we find* $V_0 = \text{span}(1,1)$ *and* $V_1 = \text{span}(1,0)$. *The vectors* $(1,0)$ *and* $(1,1)$ *are simply not perpendicular to each other, so – although there is a basis of eigenvectors and* A *is diagonalizable – there is on orthogonal basis of eigenvectors, so* A *is not orthgonally diagonalizable. Notice that* A *is a projection operator:* $A^2 = A$, *and for all such operators* $\mathbb{R}^n = V_1 \oplus V_0$. *However it is not an* orthogonal *projection: this means precisely that* V_0 *and* V_1 *are not orthogonal subspaces.*

Exercise: Let $A \in M_{n,n}$ be a projection: $A^2 = A$, so that $\mathbb{R}^n = V_1 \oplus V_0$ and A is diagonalizable. Show that A is orthogonally diagonalizable if and only if $V_1 \perp V_0$, i.e., if and only if A is an orthogonal projection.

I claim that in fact I can tell immediately upon looking at $\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$ that it is not orthogonally diagonalizable. Why is that? Because of the following result.

Proposition 7.33. *Let* $A \in M_{n,n}$. *If* A *is orthogonally diagonalizable then* A *is symmetric:* $A^T = A$.

Proof. Suppose that there is an orthogonal matrix P such that $PAP^{-1} = D$ is diagonal. Then $A = P^{-1}DP$, so

$$A^T = (P^{-1}DP)^T = P^T D^T (P^{-1})^T = P^{-1}D(P^T)^T = P^-1DP = A.$$

□

Exercise: Show that if A is symmetric and P is orthogonal, then $P^{-1}AP$ is symmetric.

Well, that was easy. In general in mathematics when you learn a result of the form $A \implies B$, you should immediately inquire about the converse. Sometimes the proof that you gave of $A \implies B$ can be easily turned around to give a proof of $B \implies A$. Not always, of course: often enough the converse is false.[9] There is a third possibility: sometimes the converse is also true, but the proof of $B \implies A$ has nothing to do with the proof of $B \implies A$. Sometimes the innocent question "Is the converse also true?" leads us to some deep results. That is the case here.

[9]All squares are rectangles, but not all rectangles are squares.

Theorem 7.34. *(Spectral Theorem) Every symmetric matrix $A \in M_{n,n}(\mathbb{R})$ is orthogonally diagonalizable.*

We will prove the Spectral Theorem in the next section: it will take some doing. For now we record some consequences.

7.6. **Proof of the Spectral Theorem.**

Let $A \in M_{n,n}$, and let $V \subset \mathbb{R}^n$ be a subspace. We say V is **invariant under A** if $A(V) \subset V$: that is, if for every $v \in V$, $Av \in V$.

In the following, when we say "invariant subspace" we mean with respect to some fixed matrix $A \in M_{n,n}$.

Exercise: Let $V_1, \ldots, V_k \subset \mathbb{R}^n$ be invariant subspaces
a) Show that $V_1 + \ldots + V_k$ is invariant.
b) Show that $\bigcap_{i=1}^k V_i$ is invariant.

Exercise: Show that for any $\lambda \in \mathbb{R}$, the eigenspace V_λ is an invariant subspace.

Example 7.35. *Let v be an eigenvector for A, with eigenvalue λ, and let $V = \operatorname{span} v$. Then every element w of V is of the form $w = \alpha v$ for some $\alpha \in \mathbb{R}$, so $Aw = A(\alpha v) = \alpha Av = \alpha\lambda v \in V$. Thus V is a one-dimensional invariant subspace. Conversely, let $V = \operatorname{span} v$ be a one-dimensional invariant subspace. Then $Av \in V$, and every element of V is of the form αv for some $\alpha \in \mathbb{R}$, so $Av = \alpha v$ and v is an eigenvector. We deduce that the one-dimensional invariant subspaces are precisely the lines spanned by eigenvectors.*

Example 7.36. *Suppose A is nonsingular, and $V \subset \mathbb{R}^n$ is invariant under A. Then it is also invariant under A^{-1}. Indeeed, since A is injective, $\dim A(V) = \dim V$, and since $A(V) \subset V$ is an inclusion of subspaces of the same dimension, we must have equality: $A(V) = V$. Thus every $v \in V$ can be written (uniquely) in the form $v = Av'$ for some $v' \in V$. Then*

$$A^{-1}v = A^{-1}(Av') = v' \in V.$$

Proposition 7.37. *Let $A \in M_{n,n}$ be either symmetric or orthogonal. Then:*
a) If $V \subset \mathbb{R}^n$ is invariant under A, then also $V^\perp$ is invariant under A.
b) Let $\lambda_1 \neq \lambda_2$ be eigenvalues of A. Then $V_{\lambda_1} \perp V_{\lambda_2}$.

Proof. Recall our convention is that vectors can be viewed as matrices consisting of a single column, and with this convention the dot product $v \cdot w$ is an instance of the matrix product $v^T w$. In particular then for all $v, w \in \mathbb{R}^n$ and any matrix $M \in M_{n,n}$,

$$Av \cdot w = (Av)^T w = v^T A^T w = v \cdot (A^T w).$$

a) Let $w \in V^\perp$. We need to show that $Aw \in V^\perp$, or in other words to show that $Aw \cdot v = 0$ for all $v \in V$. Suppose first that A is symmetric; then

$$Aw \cdot v = w \cdot (A^T v) = w \cdot Av = 0$$

since $Av \in V$ and $w \in V^\perp$. Similarly, if A is orthogonal then we use X.X: V is also A^{-1} invariant, so $a^{-1}v \in V$ and thus

$$Aw \cdot v = w \cdot (A^T v) = w \cdot A^{-1}v = 0.$$

b) Let $v \in V_{\lambda_1}$ and $w \in V_{\lambda_2}$. Suppose first that $A^T = A$. Then

$$\lambda_1(v \cdot w) = (\lambda_1 v) \cdot w = Av \cdot w = v \cdot A^T w = v \cdot Aw = v \cdot (\lambda_2 w) = \lambda_2(v \cdot w).$$

Thus

$$(\lambda_1 - \lambda_2)(v \cdot w) = 0,$$

and since $\lambda_1 \neq \lambda_2$, we conclude $v = w$. Now suppose that $A^T = A^{-1}$; by X.X, we may assume $\lambda_1 = 1$ and $\lambda_2 = -1$; in particular each eigenvalue is equal to its reciprocal. Thus

$$v \cdot w = Av \cdot w = v \cdot A^T w = v \cdot A^{-1} w == v \cdot (-w) = -v \cdot w.$$

Thus $2(v \cdot w) = 0$ and $v \cdot w = 0$. □

In light of Proposition X.Xb), in order to prove the Spectral Theorem it is enough to show that every symmetric matrix has a basis of eigenvalues.

Lemma 7.38. *Let $A \in M_{n,n}(\mathbb{R})$ be symmetric, and let $\alpha, \beta \in \mathbb{R}$ be such that $\alpha < 4\beta$. Then $A^2 + \alpha A + \beta I_n$ is invertible.*

Proof. We claim that for all $0 \neq v \in \mathbb{R}^n$,

$$(A^2 + \alpha A + \beta I_n)v \cdot v > 0.$$

If so then the null space of $A^2 + \alpha A + \beta I_n$ is $\{0\}$ and thus $A^2 + \alpha A + \beta I_n$ is invertible. Step 1: Recall the Cauchy-Schwarz inequality: for $v, w \in \mathbb{R}^n$,

$$|v \cdot w| \leq ||v||||w||.$$

Thus for any $\alpha \in \mathbb{R}$,

$$-\alpha(v \cdot w) \leq |-\alpha||v \cdot w| = |\alpha||v \cdot w| \leq |\alpha|||v||||w||;$$

multiplying through by -1 gives

$$\alpha(v \cdot w) \geq -|\alpha|||v||||w||.$$

Using this we find

$$\begin{aligned}
(A^2 + \alpha A + \beta I_n)v \cdot v &= (A^2 v) \cdot v + \beta(v \cdot v) \\
&= Av \cdot A^T v + \alpha(Av \cdot v) + \beta||v||^2 \\
&= Av \cdot aV + \alpha(Av \cdot v) + \beta||v||^2 \\
&= ||Av||^2 + \alpha(Av \cdot v) + \beta||v||^2 \\
&\geq ||Av||^2 - |\alpha|||Av||||v|| + \beta||v||^2 \\
&= \left(||Av|| - \frac{|\alpha|||v||}{2}\right)^2 + \left(\beta - \frac{\alpha^2}{4}\right)||v||^2 > 0.
\end{aligned}$$

□

Proposition 7.39. *Let $A \in M_{n,n}(\mathbb{R})$ be symmetric. Then A has an eigenvalue.*

Proof. Let $0 \neq v \in \mathbb{R}^n$. Then the set $\{v, Av, \ldots, A^n v\}$ consists of $n+1$ vectors in $\mathbb{R}^n$, so they are linearly dependent: there are $a_0, \ldots, a_n \in \mathbb{R}$, not all zero, such that

$$a_n A^n v + \ldots + a_1 Av + a_0 v = 0.$$

There is $N \leq n$ such that $a_N \neq 0$ and

$$a_N a^N v + \ldots + a_1 Av + a_0 v = 0.$$

Let $p(t) = a_N t^N + \ldots + a_1 t + a_0$, so $p(A)v = 0$. We factor $p(t)$ as

$$p(t) = a_N(t^2 + \alpha_1 t + \beta_1) \cdots (t^2 + \alpha_s t + \beta_s) \cdot (t - \lambda_1) \cdots (t - \lambda_r),$$

where the quadratic polynomials $t^2 + \alpha_i t + \beta_i$ have no real roots – equivalently by the quadratic formula, $\alpha_i^2 < 4\beta_i$ for all i. Since

$$0 = p(A)v = a_N(A^2 + \alpha_1 A + \beta_1 I_n) \cdots (A^2 + \ldots + \alpha_s A + \beta_s I_n)(A - \lambda_1 I_n) \cdots (A - \lambda_r I_n)v.$$

By Lemma X.X each matrix $A^2 + \alpha_i A + \beta_i I_n$ is invertible, so multiplying by their inverses gives

$$0 = (A - \lambda_1 I_n) \cdots (A - \lambda_r I_n)v.$$

If $(A - \lambda_r I_n)v = 0$, then λ_r is an eigenvalue. If not, then $v' = (A - \lambda_r I_n)v \neq 0$, so if $(A\lambda_{r-1} I_n)v' = 0$ then λ_{r-1} is an eigenvalue. And so forth: since the product is zero, at some point multiplying by $(A - \lambda_i I_n)$ must convert a nonzero vector to the 0 vector, so one of $\lambda_1, \ldots, \lambda_r$ is an eigenvalue for A. □

Proof of the Spectral Theorem: We go by induction on n, the case $n = 1$ being trivial. So suppose $n \geq 2$ and every symmetric matrix $B \in M_{n-1,n-1}(\mathbb{R})$ is diagonalizable. Let $A \in M_{n,n}(\mathbb{R})$ be a symmetric matrix. By X.X, there is an eigenvector v for A, say with eigenvalue λ. By rescaling v we may choose v to be a unit vector. Let $v, v_2, \ldots, v_n$ be an orthonormal basis of $\mathbb{R}^n$ (extend v to a basis then apply Gram-Schmidt to get an orthonormal basis). Let P be the (orthogonal!) matrix with columns $v, v_2, \ldots, v_n$, and let $A' = P^{-1}AP$. Since P is orthogonal, by Exercise X.X, A' is again symmetric. The matrix A' is block diagonal, with upper left 1×1 block λ and bottom right $n-1 \times n-1$ block B, say. Since A' is symmetric, so is B. By induction, B is orthogonally diagonalizable: there is an orthogonal matrix $Q_1 \in M_{n-1,n-1}$ such that $Q_1^{-1} B Q_1$ is diagonal. Thus if Q is the block matrix $1 \oplus Q_1$, then Q is orthogonal and

$$Q^{-1}A'Q = Q^{-1}P^{-1}APQ = (PQ)^{-1}A(PQ)$$

is diagonal. Since P and Q are orthogonal, so is PQ, so A is orthogonally diagonalizable.

References

[A] S. Axler, *Linear algebra done right*. Second edition. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1997.

[Ax95] S. Axler, *Down with determinants!* Amer. Math. Monthly 102 (1995), no. 2, 139-154.

[BB03] K. Boulabiar and G. Buskes, *After the determinants are down: a criterion for invertibility*. Amer. Math. Monthly 110 (2003), no. 8, 737-741.

[Bu73] M.D. Burrow, *The minimal polynomial of a linear transformation*. Amer. Math. Monthly 80 (1973), 1129-1131.

[Cl12] P.L. Clark, *Covering numbers in linear algebra*. Amer. Math. Monthly 119 (2012), 65-67.

[Cl-I] P.L. Clark, *Lecture notes on mathematical induction*. `http://www.math.uga.edu/~pete/3200induction.pdf`

[Cl-IS] P.L. Clark, *Linear algebra: invariant subspaces*. `http://math.uga.edu/~pete/invariant_subspaces.pdf`

[De03] H. Derksen, *The fundamental theorem of algebra and linear algebra*. Amer. Math. Monthly 110 (2003), 620-623.
[Gi91] G.T. Gilbert, *Positive definite matrices and Sylvester's criterion*. Amer. Math. Monthly 98 (1991), 44-46.
[HC] P.L. Clark, *Honors Calculus./..*
[Hu02] C. Huneke, *The Friendship Theorem*. Amer. Math. Monthly 109 (2002), 192–194.
[Ja73] H.G. Jacob, *Another proof of the rational decomposition theorem*. Amer. Math. Monthly 80 (1973), 1131-1134.
[PP07] B. Palais and R. Palais, *Euler's fixed point theorem: The axis of a rotation*. J. Fixed Point Theory Appl. 2 (2007), 215-220.
[SA] T. Shifrin and M. Adams, *Linear Algebra: A Geometric Approach*.
[St93] G. Strang, *The fundamental theorem of linear algebra*. Amer. Math. Monthly 100 (1993), 848-855.
[Wa01] W.P. Wardlaw, *A Generalized General Associative Law*. Mathematics Magazine 74 (2001), 230-233.
[Wa80] W. Watkins, *Similarity of matrices*. Amer. Math. Monthly 87 (1980), 300.
[Y] T. Yuster, *The Reduced Row Echelon Form of a Matrix is Unique: A Simple Proof*. Math. Mag. 57 (1984), 93-94.

Department of Mathematics, Boyd Graduate Studies Research Center, University of Georgia, Athens, GA 30602-7403, USA
E-mail address: pete@math.uga.edu