A first course on linear algebra

Ricardo Souza

October 18, 2019

Contents

Chapter 1

Basic set theory

1.1 Naive set theory

1.1.1 Axioms and you

Most, if not all, concepts in mathematics are phrased in the language of set theory: Geometric figures are just collections of points, transformations between two different objects are the collections of all the transitional states inbetween etc.

Hence, it makes sense to give some more formal foothold when studying any area of maths by beginning with some basic set theory.

But then, why *naive*?

Well, formal mathematics (that is, all contemporary and modern mathematics for more than a hundred years) is based on what we like to call *axioms* - you can think of them as the "rules of the game", in some sense.

Let me give you all an example of a well-accepted axiom of Euclidean geometry:

Axiom

Given any two distinct points, there is one, and only one, line through them.

Some people say it's "something that you can't prove", but it's not exactly that - axioms are either things that you don't *want* to prove, and just want to assume as truth (maybe because it is, indeed, impossible to prove it) or things that are, in some vague sense, "natural" or "self-evident".

Either way, the correct mindset to approach axioms is to think of them as the building blocks with which you build maths - just like atoms are the building blocks of matter -: by combining different axioms in different ways you get different results - the so called "theorems".

That's what maths is all about: Working with axioms and already proven theorems to prove new theorems. It's kinda like a game of scrabble, where the axioms are not only the blocks you (and everyone else) has in their hands, but also the rules of the game and the game board, and the theorems are the words you can make - subject to the rules of the game, the pieces and the board.

Hence, *naive* set theory is called so not because it is a theory of naive sets, but because it's a theory that's not properly formalized, and relies heavily on intuition and common sense.

In proper, axiomatic, set theory you'd have to define what is, and what isn't, a set. In naive set theory, however, we can just hand-wave it and say

Naive axiom(?)

Any collection of things is a set.

Now, as to *why* this isn't formal, it's due to the fact that it leads to a logical contradiction - a paradox. We're gonna show this contradiction in what follows, but if it doesn't interest you (you filthy, you) you can just skip the next section. It's fine, I won't judge you (actually, I will).

1.1.2 Russell's Paradox

Imagine that every random collection of random things is a set. Then it is only natural to consider the collection of all sets. But, since it is a collection (duh) it is also a set. But since it is the collection of all sets, it is an element of itself.

That's weird, ok? Try thinking of any sets - I'll give you plenty of time, don't worry - that are like that: they contain themselves as elements. You can't, right?

While that's not a contradiction, per se, it *really is weird*.

So let us consider N the collection of all non-weird sets - that is, the collection of all sets that do not contain themselves as elements.

Now, one naturally asks the question: Is N itself a weird set? That is, $N \notin N$?

Well, I don't know. But if it was, then, by definition, all elements of N are non-weird sets, so N, being an element of N, would have to be a non-weird set - that is, $N \in N$. So... If we assume that $N \notin N$ we can logically infer that $N \in N$...

Okay, maybe we made a mistake all along by assuming that $N \notin N$! Yeah, that must be the case! Clearly, N can't be a weird set!... But then, since N isn't weird, it must be an element of N (since N contains *all* non-weird sets)... So it is weird. So if we assume $N \in N$ we can logically deduce that $N \notin N$.

We have just proven that $N \in N$ and $N \notin N$ are *logically equivalent*. But by the **Principle of Non-Contradiction** (something can't be simultaneously both true and false) those two can't be equivalent!

So, by assuming that there is a set containing all sets we can logically derive a contradiction - that, my friends, is the definition of a paradox.

This is the famous **Russell's Paradox** and it applies in broader contexts - it basically means that, from a logical POV, self-references are *kinda weird and you shouldn't actually do that*.

For instance, if you put as an axiom that "anything that can be stated can be proven", then you could ask "can I prove that there is something that cannot be proven?" and the answer would have to be *yes*, since you said (by axiom) that everything had to be provable. But that's a contradiction - by forcing everything to have a proof you have proven that you cannot prove everything.

This was proposed by philosopher-mathematician Bertrand Russell to show that maths really does need a formal framework to work with - otherwise we might be working in a system where contradictions arise (as we have seen).

There is, however, a solution to this. We have a set of axioms for set theory called the Zermello-Frankel axioms, which are a list of axioms that do not generate that kind of contradiction. It is, however, *impossible* to prove whether it does or doesn't generate *any* paradox (this is due to a bunch of hard maths/philosophy that is waaaaay out of the scope of this text).

Just know that if you ever see ZFC anywhere you can rest safe because you're working with a (relatively) safe set of axioms.

1.2 Basic results and properties of sets

1.2.1 Set equality

As we have previously stated, a *set* is a collection of objects. We will usually denote a set by a capital letter (not always), such as X, A or B.

Since we cannot (as seen in the previous section) consider "the set of all sets", fix any set X. Now, X might be any set - numbers, birds, colours, the numerous of ways you can insult someone's mum etc.

When we have an object that is in that set we say that it is an **element** of that set, and usually denote it by a non-capital letter (once again, not always). In symbols, if we want to say that a is an element of X we would write that as $a \in X$ - which should be read as "a is an element of X", "a is in X" or even "X contains a as an element".

Example(s)

Let E be the collection of all even integers. So $2 \in E$ and $28 \in E$, but $5 \notin E$ ("5 isn't in E", or "5 isn't an even integer") and $dog \notin E$ (because *dog* is **not** an even integer). Actually, you can see this as a formal proof of the well known fact that all dogs are *odd*.

You can, however, take all the elements of a set and ask if they satisfy a certain condition.

Example(s)

Following up on the previous example, let ϕ denote the proposition "*can be written in english with only three letters*". Now we can consider the *subset* of E formed by all elements of E that also satisfy ϕ (if $x \in E$ is such an element, we simply write $\phi(x)$ to denote "x satifies ϕ"). This is written as follows:

$$E_\phi := \{x \in E \mid \phi(x)\}.$$

Let us break this down bit-by-bit:

- The symbol E_ϕ is non-standard notation that we're introducing here to mean "*the set E subject to the condition ϕ*";
- The symbol $:=$ means "*equals, by definition*". This can be used in two distinct ways: During a logical regression, we can use this symbol to justify one step by saying "this thing that I'm claiming is true, is actually true by definition"; or we can use it to define new terms - we're basically saying "the LHS is a new symbol whose meaning I'm defining to be the RHS" - kinda like attributing a value to a variable.

  In this text we're **always** going to use this symbol with the second meaning - so in the preceding expression the $:=$ means "I'm defining E_ϕ to mean $\{x \in E \mid \phi(x)\}$".
- The brackets, in mathematics, almost always denote a *set*, and always are presented with the following structure: $\{A \mid B\}$.

The A part is *what kind of elements does this set have*. In the example above, $x \in E$ means that the elements we're working with are even integers.

The B part is *which condition these elements are subject to*. In the example above, $\phi(x)$ means that the elements of this set must satisfy ϕ.

Now that that's out of the way, what is E_ϕ? What are *the even integers that can be written in english using only three letters*? There are only three such numbers: **two**, **six** and **ten**. So we write $E_\phi = \{2, 6, 10\}$.

Definition 1.2.1.1. *Two sets A and B are said to be **equal** if they have the same elements. This means that every element of A is an element of B, and every element of B is an element of A.*
In this case we write $A = B$.

Let us give some examples of equalities.

Example(s)

- Let A be the set of all animals that are wooly, fluffy and go *baa*, and let B be the set of all sheep. Clearly $A = B$.
- Let A be the set of roots of the polynomial $x^2 - x$ and let $B = \{0, 1\}$. It is an easy exercise to see that these two sets are the same.
- However, $A = \mathbb{N}$ the set of all natural numbers, and $B = \mathbb{Z}^{\geq 0}$ the set of non-negative integers, are **not** equal sets. You can see this in any proper course of number/set theory, but the elements of $\mathbb{Z}$ are always signed: -2, $+6$, $+1$ etc. (aside from 0), whereas the elements of $\mathbb{N}$ are **not** signed: 1, 6 etc. So $1 \notin \mathbb{Z}$ and $+1 \notin \mathbb{N}$, and therefore $A \neq B$.

Remark 1.2.1.2

In mathematics, a *definition* is the term we use to "assign" a new value to a certain term. In the definition above, we assigned a meaning to the phrase "two sets are equal".
Please be aware that this text will be filled with definitions of this kind, so take your time to get accostumed to them.

Notice, however, that we can sort of "relax" the conditions of the preceding definition. For instance, consider the following case:

Example(s)

Let $A = \mathbb{N}$ the set of all natural numbers and $B = E$ the set of all even natural numbers. Notice that $A \neq B$ - for instance, 3 is in A, but not in B - so they can't be equal.
On the other hand, notice that it is impossible to produce such a counterexample starting from B: No matter which element you choose in B it will always be a natural number, of course, and therefore it will also be an element of A.

So these two sets, although not-equal, are not *entirely* different.

Definition 1.2.1.3. *Let A and B be two sets such that every element of B is also an element of A. In this case, we say that A* ***contains*** *B* ***as a subset*** *- or more simply that B* ***is a subset of*** *A, which we'll denote in symbols by $B \subseteq A$.*

Example(s)

- In the preceding example, we see that $B \subseteq A$.
- Take any set A, and let $B = A$. We then ask the question: Is B a subset of A? Well, by definition, $B \subseteq A$ if, and only if, every element of B is also an element of A... But this is trivially true - since $B = A$!

  This gives us some insight on our first result:

Proposition 1.2.1.4. *For any set A we have that $A \subseteq A$.*

Proof

We want to show that every element $a \in A$ is also an element of A. But that's trivial. The result follows. □

Remark 1.2.1.5

In mathematics, a *proof* of a proposition/lemma/theorem/corolary is nothing more than a logical reasoning explaining why what we said is true. Proofs are to mathematics as scientific experiments are to sciences. This is what mathematicians do and work with all their lives. One could argue that maths is the science of reasoning and arguing.

Now we have our first non-trivial result:

Proposition 1.2.1.6. *Let A and B be two sets. Then $A = B$ if, and only if, $A \subseteq B$ and $B \subseteq A$.*

Proof

Assume that $A = B$. We want to show that $A \subseteq B$ and $B \subseteq A$, but this is trivial in light of the preceding proposition.

Assume now that $A \subseteq B$ and $B \subseteq A$. We want to show that $A = B$ - that is, every element of A is an element of B, and every element of B is an element of A.
Notice, however, that the phrase "every element of A is an element of B" is the definition of the symbol $A \subseteq B$, and the phrase "every element of B is an element of A" is the definition of the symbol $B \subseteq A$ - both of which we are assuming to be true.
Therefore, we have just proven that $A = B$, as stated, which finishes the proof. □

Remark 1.2.1.7

In mathematics, an *if, and only if,* statement is the equivalent of a logical equivalence. Basically, whenever we say "*this* holds if, and only if, *that* holds" what that means is that *this* and *that* are equivalent: *this* is true precisely when *that* is true, and *this* is false precisely when *that* is also false.
Without going too much into propositional logic, we usually write "a if, and only if, b" in symbols as $a \iff b$, which is logically equivalent to saying that "a being true is sufficient for us to prove that b is also true" and "b being true is sufficient for us to prove that a is also true". In symbols we would write these, respectively, as $a \implies b$ and $b \implies a$ - which should be read as "a implies b" and "b implies a", respectively.
That's what we did in the preceding proposition: If a="$A = B$" and b="$A \subseteq B$ and $B \subseteq A$", we proved that assuming a we can conclude b, and that assuming b we can conclude a - that is, we proved that a implies b and b implies a - which is logically equivalent to proving that a and b are equivalent.

This proposition is the most common tool used by mathematicians to prove that two sets are equal: We simply prove that each one contains the other - therefore, they must be equal.

Example(s)

Let A be the set of roots of the polynomial $x^2 - x$ - that is, the set of numbers r such that $r^2 - r = 0$ - and $B = \{0, 1\}$. We claim that $A = B$.
First, let us show that $B \subseteq A$ - that is, both 0 and 1 are roots of $x^2 - x$. This is done by a simple verification:

$$0^2 - 0 = 0 - 0 = 0 \qquad \text{and} \qquad 1^2 - 1 = 1 - 1 = 0$$

so they are, indeed, roots of $x^2 - x$ - and therefore, $B \subseteq A$.
Now, to prove that $A \subseteq B$ we need to show that those are the only two possible roots.
To do that, let r be any root of $x^2 - x$ - that is, $r^2 - r = 0$. But then, $r^2 = r$, by adding r on both sides, and we see that $r = 0$ is indeed a solution to this equation ($0^2 = 0$). So if we assume that $r \neq 0$ we can divide both sides by r and get $\frac{r^2}{r} = \frac{r}{r}$ which is the same as $r = 1$, which was a unique solution being $r = 1$.
Hence we have proven that any root r of $x^2 - x$ is either 0 or 1, and therefore $A \subseteq B$.
Finally, since $A \subseteq B$ and $B \subseteq A$ we can finally say that $A = B$, as we had previously stated.

Definition 1.2.1.8. *We say that A is a **proper subset** of B if A is a subset of B, but B isn't a subset of A. In this case we use the symbol $A \subset B$.*

Example(s)

Consider $A = \mathbb{N}$ the set of natural numbers, and $B = E$ the set of even natural numbers. We clearly have $B \subseteq A$ and $A \not\subseteq B$, so we can see that B is a *proper* subset of A - that is,

$B \subset A$.

Finally, we can use all that we've done so far to construct a very special set - the empty set.

Example(s)

Let $\mathbb{N}$ be the set of natural numbers and let ϕ be the proposition "is not a natural number". For instance, $\phi(\text{car})$ is just "car is not a natural number", which is true.
Now we can do just as we did before and consider

$$\mathbb{N}_\phi := \{n \in \mathbb{N} \mid \phi(n)\}$$

that is, the set of all natural numbers which are not natural numbers.
What **is** this set? Is there any natural number that isn't a natural number? Of course not! So this is a set *which has no elements*.
Take now $\mathbb{Z}$ the set of all integers and let ψ be the proposition "is not an integer". We can then define, once more,

$$\mathbb{Z}_\psi := \{n \in \mathbb{Z} \mid \psi(n)\}$$

that is, the set of all integers which aren't integers.
This set is, once again, empty.
This begets the question: $\mathbb{N}_\phi = \mathbb{Z}_\psi$ - that is, are two empty sets always equal?

Definition 1.2.1.9. *Given any set X we call the **empty set defined by** X to be the set of all elements of X which aren't elements of X, denoted by $\varnothing_X$.*

Theorem 1.2.1.10. *Given any two sets A and B, then $\varnothing_A = \varnothing_B$.*

Proof

If they were different, then there would either be some element of $\varnothing_A$ which is not in $\varnothing_B$, or some element of $\varnothing_B$ which is not in $\varnothing_A$. But both of these are impossible, since both sets are empty.
So they can't be different, and, therefore, $\varnothing_A = \varnothing_B$ □

Corollary 1.2.1.11. *For any set A, its empty set $\varnothing_A$ is uniquely determined.*

Corollary 1.2.1.12. *There a unique empty set.*

Remark 1.2.1.13

In mathematics, a *corolary* is a result that follows immediately from something that came before it - sometimes even foregoing a proof because of how immediate this conclusion is.

Definition 1.2.1.14. *We're going to define the **unique empty set** to be the empty set of any set, which will be denoted in symbols by $\varnothing$.*

Proposition 1.2.1.15. *For any set A we have that $\varnothing \subseteq A$. Furthermore, we have that $A \subseteq \varnothing$ if, and only if, $A = \varnothing$.*

Proof

If $\varnothing \not\subseteq A$, then there'd be some element in $\varnothing$ that was not in A. But $\varnothing$ is empty, therefore $\varnothing \not\subseteq A$ is false, and hence $\varnothing \subseteq A$.

For the second statement, we clearly have $A \subseteq \varnothing$ if $A = \varnothing$, by definition of set equality. But if we assume that $A \subseteq \varnothing$, we can now use the first statement of this proof, which proves that $\varnothing \subseteq A$, to conclude, by definition of set equality, that $A = \varnothing$, and this finishes the proof. □

1.2.2 Unions and intersections

Now that we have a basic understanding of sets and subsets, we're going to build new sets from existing ones.

Definition 1.2.2.1. *Let A and B be two sets. The **union** of A and B is another set - denoted by $A \cup B$ defined by the following properties:*

(a) $A \cup B$ contains both A and B as subsets;

(b) Any other set C that contains both A and B as subsets also contains $A \cup B$ as a subset.

First things first, let us show that this definition makes sense - that is, that given two sets, their union is a unique set:

Lemma 1.2.2.2. *Let A and B be two sets, and C and D be two sets satisfying the above definition. Then $C = D$.*

Proof

Since C and D are unions of A and B, they contain both of them as subsets (item (a)). Now, since C satisfies (a) and D satisfies (b), we get that $D \subseteq C$. Similarly, since D satisfies (a) and C satisfies (b), we get that $C \subseteq D$.
It follows that $C = D$, and so the union of two sets is indeed well-defined, □

With that out of the way, let us show some examples to build some intuition:

Example(s)

Let A be the set of all dogs and B be the set of all cats. Let C be the set of all animals. We ask: $C = A \cup B$?
Certainly, C satisfies (a) (since all dogs and all cats are animals), but does it satisfy (b)?
Well, certainly not! Because the set D of all mammals also contains A and B, but it clearly doesn't contain C (because not every animal is a mammal - for instance, there are birds).
Now we ask: Ok, since C is not the union of A and B, maybe D is?
Well, no, because we can consider E - the set of all mammal quadrupeds - and see that it contains both A and B as susbets, but not D.
And so on, and so forth...

How can we make sure that we don't get an endless regression - that is, we're always inching closer to the result, but never truly getting there?

Well, in formal set theory, for instance ZFC, you can always use your axioms to guarantee the existence of such a set. Here, however, we're going to have to appeal to intuition:

Proposition 1.2.2.3. *Given two sets A and B and any set C containing both A and B, their union is precisely the subset of C given by the proposition $\phi =$ "is in any one of the sets A or B".*

Proof

First, we'll show that $A \subset C_\phi$ and $B \subseteq C_\phi$.
To do that we'll just use the definition: Take $a \in A$ (resp. $b \in B$). Since $A \subseteq C$ (resp. $B \subseteq C$) we have that $a \in C$ (resp. $b \in C$). We then ask: is $\phi(a)$ (resp. $\phi(b)$) true? Well, it trivially is - $\phi(x)$ is true if, and only if x is in A or B - and a (resp. b) certainly is. Therefore, for any $a \in A$ (resp. $b \in B$) we can conclude $\phi(a)$ (resp. $\phi(b)$) - and therefore, $a \in C_\phi$ (resp. $b \in C_\phi$). This shows that $A \subseteq C_\phi$ (resp. $B \subseteq C_\phi$) - and therefore, C_ϕ satifies item (a) of the definition of set union.
Now, take any set D such that D contains both A and B as subsets. If we show that D also contains C_ϕ as a subset, we'll have shown that C_ϕ satisfies the definition of union - and therefore it must be the union.
To do that, take any $x \in C_\phi$. Then, by definition, $\phi(x)$ is true - that is, $x \in A$ or $x \in B$. But since both $A \subseteq D$ and $B \subseteq D$ hold, it doesn't matter if $x \in A$ or $x \in B$ is true - as long as one of them is true, we can conclude that $x \in D$. And since this holds for any $x \in C_\phi$, we have just shown that $C_\phi \subseteq D$.
Since the D we chose was general, the result follows. □

This is important: We now have a way to construct the union of two sets - just take any set containing both of them and restrict it to be only the elements from the original sets.

That, however, requires the existence of some set containing both of them - and that's where ZFC comes in: There's an axiom that states that there always exists a set containing any amount of other sets.

Since we're foregoing axioms here, we're going to provide a "construction" that should be enough for most purposes:

Example(s)

Following up on the previous example, we can now see that $A \cup B$ is any one of "the set of all animals which are cats or dogs", "the set of all mammals which are cats or dogs" or "the set of all mammal quadrupeds which are cats or dogs" - any one of those work, by what we've already proven.
We could, however, give a more explicit construction: $A \cup B$ is just the set of all cats and dogs.

Example(s)

Another, even more constructive example: Let $A = \{a, b, c\}$ and $B = \{c, d, e, f\}$. Then $A \cup B = \{a, b, c, d, e, f\}$ (prove it using the definition if you're not convinced).

Finally, let's end our discussions on the union with the following alternative characterization of it:

Lemma 1.2.2.4. *Let A and B be sets. Then $x \in A \cup B$ if, and only if, $x \in A$ or $x \in B$.*

Proof

One side of this proof is trivial and follows from the definition of set union.
Let us prove then that $x \in A \cup B$ implies $x \in A$ or $x \in B$.
Define $N = \{x \in A \cup B \mid x \notin A \text{ and } x \notin B\}$ the collection of all elements of $A \cup B$ which are in neither A nor B - which is, by definition, a subset of $A \cup B$.
We can now define $U = \{x \in A \cup B \mid x \notin N\}$ the collection of all elements of $A \cup B$ which are not in N - which is, by definition, a subset of $A \cup B$.
We claim that U contains A and B as subsets. This is easy to see: Take y in either A or B (doesn't matter which). Since $A \cup B$ contains both of them, $y \in A \cup B$. But since y came from either A or B, it cannot be in N (by definition of N) - so it must be in U (by definition of U). It follows that both A and B are contained in U.
But this is a conundrum, because $A \cup B$ is *contained* in every set that contains A and B (by definition of set union) - in particular, since U contains A and B this means that U *also contains* $A \cup B$.
This shows that $U = A \cup B$, and therefore $N = \varnothing$.
Finally, to show that $x \in A \cup B$ implies $x \in A$ or $x \in B$, take any $x \in A \cup B$ and notice, by what we've done, that this is the same as saying that $x \in U$. But, then again, this is the same as saying that $x \notin N$ - that is x is in A or B, just as stated. This finishes the proof. □

Going in the opposite direction of unions, there is the concept of intersections. If unions take two sets to build a bigger one, interceptions take two sets to build a smaller one:

Definition 1.2.2.5. *Let A and B be two sets. The **intersection** of A and B is another set - denoted by $A \cap B$ defined by the following properties:*

(a) $A \cap B$ is contained in both A and B as a subset;

(b) Any other set C that is contained both A and B as a subset is also contained $A \cap B$ as a subset.

Remark 1.2.2.6

Notice that the two definitions are basically the same, just changing, in some sense, the "order" of the inclusions $\subseteq$.

Now, let us proceed to prove essentially the same results for intersections as we did for unions:

Lemma 1.2.2.7. *Let A and B be two sets, and C and D be two sets satisfying the above definition. Then $C = D$.*

Proof

Since C and D are intersections of A and B, they are contained in both of them as subsets (item (a)). Now, since C satisfies (a) and D satisfies (b), we get that $C \subseteq D$. Similarly, since D satisfies (a) and C satisfies (b), we get that $D \subseteq C$.
It follows that $C = D$, and so the intersection of two sets is indeed well-defined, □

Contrary to unions, however, we cannot refine intersections. We can, however, still give a construction of the intersection:

Lemma 1.2.2.8. *Let A and B be sets. Then $x \in A \cap B$ if, and only if, $x \in A$ and $x \in B$.*

Proof

One side of this proof is trivial and follows from the definition of set intersection.
Let us prove then that $x \in A$ and $x \in B$ implies $x \in A \cap B$.
Define $N = \{x \in A \text{ and } x \in B \mid x \notin A \cap B\}$ the collection of all elements which are, at once, in both A and B, but not in $A \cap B$ - which is, by definition, a subset of both A and B.
We can now define $I = \{x \in A \text{ and } x \in B \mid x \notin N\}$ the collection of all elements of both A and B which are not in N - which is, by definition, a subset of both A and B. This implies, by definition of set intersection, that $I \subseteq A \cap B$.
We claim now that I contains $A \cap B$ as a subset. This is easy to see: Take any $y \in A \cap B$. Since $A \cap B \subseteq A$ and $A \cap B \subseteq B$, we see that $y \in A$ and $y \in B$. So this y is an element of both A and B which is in $A \cap B$ - which is the definition of an element of I. This shows that $y \in I$.
But this is a conundrum, because $A \cap B$ *contains* every set that is contained in both A and B (by definition of set intersection).
This shows that $I = A \cap B$, and therefore $N = \varnothing$.
Finally, to show that $x \in A$ and $x \in B$ implies $x \in A \cap B$, take any $x \in A$ and $x \in B$ and notice, by what we've done, that this is the same as saying that $x \notin N$ (since it is empty). But, then again, this is the same as saying that $x \in I$ - that is x is in $A \cap B$, just as stated.
This finishes the proof. □

Finally, let's do some examples:

Example(s)

- Let $A = \{1, 2, 3, 4\}$ and $B = \{1, 3, 5, 7, 9\}$. Then, $A \cap B = \{1, 3\}$.
- If A is the set of all even integers, and B is the set of all odd integers, then $A \cap B = \varnothing$.
- If A is the set of all cats and B is the set of all brown animals, then $A \cap B$ is the set of all brown cats.

As a final topic on this section, let us consider another construction.

Definition 1.2.2.9. *Given any set X we denote the **set of all subsets of** X by $\mathcal{P}(X)$ (or 2^X) and call it the **power set** of X.*

Remark 1.2.2.10

Note that, at this point, we have not defined products and sums of sets - even less exponents. So, for now, the symbol 2^X is just that - a symbol. It has no meaning resembling the powering

of real numbers.
We will, however, as this text progresses, show two reasons why this notation makes sense, and we'll expand it to be able to take any set to the power of any other set.

Okay, before anything else, let us do some examples:

Example(s)

Let $A = \{1, 2, 3\}$. What is $\mathcal{P}(A)$? Well, by definition it is the set of all subsets of A. Well then - what are the subsets of A?
We can list a few: $\varnothing$, A, $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$. But are there any others?
Well, assume $B \subseteq A$. Then we can ask if B has any elements or not. If it doesn't, great!, because $B = \varnothing$, which we've already accounted for.
If it does, we can ask if it contains 1. And then, we can ask if it contains 2 and 3. And depending on those answers we can pinpoint B exactly, and see that it is, indeed, in the list above (e.g., if it contains 1 and 2, but not 3, then $B = \{1, 2\}$, which is on the list above).
At this point, it is easy to see that

$$\mathcal{P}(A) = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, A\}.$$

The preceding reasoning, however, gives us our first insight into how to understand the symbol 2^A: Making a subset of A is the same as asking each element of A if it is, or not, in there.
Imagine the elements of A are cards in a deck and you want to make a hand. Making a hand is the same as going through the deck, card by card, and choosing which cards you want to keep, or not.
Since every card has two options (to be, or not to be), the amount of hands is precisely 2 to the number of cards.
Note that in this particular example, $2^A = \mathcal{P}(A)$ has precisely $2^3 = 8$ elements, while A has precisely 3 elements.

Now that we're talking about power sets, we can define one of the most important concepts of set theory:

Definition 1.2.2.11. *Let X be a set and 2^X its power set. Given any $A \in 2^X$, we define its* ***complement*** *to be the set denoted by $X \setminus A$, which is given by*

$$X \setminus A := \{x \in X \mid x \notin A\}.$$

That is, the complement of a set is the collection of all elements that *do not* belong to that set.

Example(s)

Following up on the previous example, let $B = \{1\}$. Then what is $A \setminus B$? Well, by definition, it's the collection of all elements of A that are not in B - that is, 2 and 3, so $A \setminus B = \{2, 3\}$.
Call $C = A \setminus B$. What is, then, $A \setminus C$? Once again, by definition, it's the set of all elements

of A which are not in C - that is, 1, so $A \setminus C = B$.

And finally, just before wrapping up this section, let us give one final definition and example:

Definition 1.2.2.12. *Let A and B be any two sets. We define $A \setminus B$ to be equal to $(A \cup B) \setminus B$ - that is, $A \setminus B$ is the complement of B in $A \cup B$.*

Example(s)

Let $A = \{a, b, c, d, e, f, g, h, i, j\}$ and $B = \{a, e, i, o, u\}$. Then $A \setminus B$ is, by definition, the set of all elements of $A \cup B$ which are not in B. So writing $A \cup B = \{a, b, c, d, e, f, g, h, i, j, o, u\}$ we see that $A \setminus B$ is just $\{b, c, d, f, g, h, j\}$.
Similarly, $B \setminus A$ is the set of all elements of $A \cup B$ which are not in A - that is, $B \setminus A = \{o, u\}$.

To really wrap up this section, then, we're gonna make a list of properties for the things we've just described. You're welcome to try to prove them, although most of them are really trivial (that is, they follow immediately from the definitions or a quick observation).

Proposition 1.2.2.13. *Let A, B, C be any three subsets of a given, fixed, set X. Then the following properties always hold:*

(1) $A \cup B = B \cup A$;

(2) $A \cup (B \cup C) = (A \cup B) \cup C$;

(3) $A \cup \varnothing = A$;

(4) $A \cup X = X$;

(5) $A \cup A = A$;

(6) $A \cap B = B \cap A$;

(7) $A \cap (B \cap C) = (A \cap B) \cap C$;

(8) $A \cap \varnothing = \varnothing$;

(9) $A \cap X = A$;

(10) $A \cap A = A$;

(11) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;

(12) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;

(13) $A \cup B = A$ *if, and only if,* $B \subseteq A$;

(14) $A \cap B = B$ *if, and only if,* $B \subseteq A$;

(15) $A \subseteq B$ *implies* $A \setminus B = \varnothing$;

(16) $A \cap B = \varnothing$ *implies* $A \setminus B = A$ *and* $B \setminus A = B$;

(17) $X = (X \setminus A) \cup A$;

(18) $(A \setminus B) \cup (B \setminus A) \cup (A \cap B) = A \cup B$;

(19) $(A \setminus B) \cap (B \setminus A) = \varnothing$;

(20) $X \setminus A \in 2^X$;

(21) $X \setminus (A \cap B) = (X \setminus A) \cup (X \setminus B)$;

(22) $X \setminus (A \cup B) = (X \setminus A) \cap (X \setminus B)$;

(23) $A \setminus (A \setminus B) = B$;

(24) $A \setminus (A \cap B) = A \setminus B$.

Remark 1.2.2.14

In the preceding proposition, as well as in maths as a whole, we usually save the parenthesis to mean "this should be done first". For instance, $A \cup (B \cup C)$ means "the union of A and the union of B and C", whereas $(A \cup B) \cup C$ means "the union of the union of A and B and C".

1.3 About functions, relations and cardinality

1.3.1 Functions

Understanding functions is, basically, the most important thing in all of mathematics - and that's not an overstatement. Even if you forego set theory, the concept of a function still makes sense and it's still at the center of any mathematical discussion.

Since this is a naive introduction to set theory, we're not gonna bother with certain technicalities and simply define:

Definition 1.3.1.1. *Let A and B be two sets. A formula ϕ is said to be of* ***function type*** *(or a* ***function****) from A to B if for any $a \in A$ there's a unique $b \in B$ such that $\phi(a, b)$.*

In that case, we will write that as $\phi(a) = b$ and say that b ***is the image of*** *a* ***under*** *ϕ.*

Example(s)

Let $A = B = \mathbb{N}$ the set of natural numbers, and let $\phi(x, y) =$ "y is the square of x". Then ϕ is clearly a function: for any $a \in A$, there is a unique $b \in B$ such that $\phi(a, b)$, and that b is precisely a^2. So we write this as $\phi(a) = a^2$.
Now, define $\psi(x, y) = \phi(y, x)$. Is ψ also a function? The answer is no: Indeed, for any $a \in A$, there exists, at most, one $b \in B$ such that $\psi(a, b)$. But the thing is - there are some a for which there is no b! For instance, for $a = 3$, there is no b such that $\psi(3, b)$. So ψ can't be a function.

Definition 1.3.1.2. *Let A, B be sets and ϕ be a function from A to B. We will call A the* ***domain*** *of the function and B its* ***codomain****, somtimes written as $A = \mathrm{Dom}(\phi)$ and $B = \mathrm{Cod}(\phi)$.*

In this case, we wil also use the notation $\phi : A \to B$ or $A \xrightarrow{\phi} B$ to say that "ϕ is a function whose domain is A and whose codomain is B".

Definition 1.3.1.3. *Two functions $f, g : A \to B$ between the same two sets are said to be* ***equal*** *if $f(a) = g(a)$ for all $a \in A$. That is, $f(a, g(a))$ and $g(a, f(a))$ hold for all $a \in A$.*

Example(s)

Let $A = B = \mathbb{R}$ the set of real numbers, and let $f, g : A \to B$ be functions defined by $f(x) = \sqrt{x^2}$ and $g(x) = \begin{cases} x, \text{ if } x \geq 0 \\ -x, \text{ otherwise.} \end{cases}$
We claim that $f = g$.
To see that, take any real number, $x \in \mathbb{R}$. Now, if $x \geq 0$, then $g(x) = x$. Furthermore, $f(x) = \sqrt{x^2} = x$, so $f(x) = g(x)$. On the other hand, if $x < 0$, we have $g(x) = -x$ and $f(x) = \sqrt{x^2} = \sqrt{(-x)^2}$ and since $x < 0$, $-x$ must be greater than 0, so $f(x)$ is simply $-x$.
Therefore, $f(x) = g(x)$ for all $x \in \mathbb{R}$ (since any real number is either negative or non-negative), and we see that $f = g$, as stated.

Definition 1.3.1.4. *If $f : A \to B$ is a function such that $f(a) = b$ for some $a \in A$ and $b \in B$, we then say that f **takes** a **to** b, which will be written as $a \mapsto b$.*

Example(s)

The functions f, g of the previous example can be rewritten as

$$\begin{aligned} f : A &\to B \\ a &\mapsto \sqrt{a^2} \end{aligned}$$

and

$$\begin{aligned} g : A &\to B \\ a &\mapsto \begin{cases} a, \text{ if } a \geq 0 \\ -a, \text{otherwise.} \end{cases} \end{aligned}$$

Before we move forward, a couple of important definitions:

Definition 1.3.1.5. *Given any function $f : A \to B$, the set of all elements of B which are image of some element of A under f will be called the **image of** A **under** f (or just the image of f) and denoted by $f(A)$ (or $\mathrm{Im}(f)$).*

*Analogously, given any $X \subseteq A$, we denote the set of all elements of B which are image of some element of X under f by **image of** f **when restricted to** X, and denote it by $f(X)$.*

Proposition 1.3.1.6. *For any function $f : A \to B$, and any $X \subseteq A$, $f(X) \subseteq B$.*

Proof

Trivial, by the definition of image of a function. □

Definition 1.3.1.7. *Given any function $f : A \to B$ and any point $b \in f(A)$, we define the **inverse image of** b **under** f to be the set $f^{-1}(b) := \{a \in A \mid f(a) = b\}$ of all points in A whose image under f is precisely b.*

*Analogously, given any $Y \subseteq f(A)$, we define the **inverse image of** Y **under** f to be the set $f^{-1}(X) := \{a \in A \mid f(a) \in Y\}$ of all points in A whose image under f is in Y.*

Proposition 1.3.1.8. *For any function $f : A \to B$, and any $Y \in f(A)$, $f^{-1}(Y) \subseteq A$.*

Proof

Trivial, by the definition of inverse image of a function. □

Finally we can start working with some very important classes of functions: Injections, surjections and bijections.

Definition 1.3.1.9. *A function $f : A \to B$ is called an **injection** if $f(a) = f(a')$ implies $a = a'$.*

Remark 1.3.1.10

This is logically equivalent to saying that a function is an injection if diferent points of the domain have different images in the codomain.

Example(s)

Let $f, g : \{1, 2, 3\} \to \{a, b, c, d\}$ be defined by: $f(1) = a, f(2) = b, f(3) = c$ and $g(1) = g(2) = g(3) = d$. Then f is injective and g is clearly not injective.

Let, now, $h : \mathbb{R} \to \mathbb{R}$ be defined by $h(x) = x^2$. Is h injective?
Well, suppose x and x' are such that $h(x) = h(x')$. This means that $x^2 = x'^2$. Taking square roots on both sides we get that $|x| = |x'|$ - which can be further simplified to mean $x = \pm x'$. In other words, we see that if two points have the same square, then they must differ only by a sign. That's good and all, but also shows us that two numbers that differ by a sign have the same image under h - and therefore h cannot be injective.
For instance, $2 \neq -2$, but $h(2) = h(-2) = 4$.

Definition 1.3.1.11. *A function $f : A \to B$ is called a **surjection** if for any $b \in B$ there is some $a \in A$ such that $f(a) = b$.*

Example(s)

Following up on the previous example, neither f nor g are surjections: $d \in \{a, b, c, d\}$ isn't in the image of any point over both f and g.
Let us then define $f' : \{a, b, c, d\} \to \{1, 2, 3\}$ by putting $f'(a) = 1, f'(b) = 2, f'(c) = 3, f'(d) = 3$. Now, f' is indeed a surjection.

Notice that h too isn't a surjection: -1 isn't the image of any real number under h. However, if we define $h' : \mathbb{R} \to \mathbb{R}^{\geq 0}$, where $\mathbb{R}^{\geq 0}$ is the set of all non-negative real numbers, by putting $h'(x) := h(x)$, we see that h' is, now, a surjection.

Remark 1.3.1.12

Note that, in the example above, we defined h' by putting $h'(x) := h(x)$. Does that mean that $h' = h$?
The answer is **no**: By the definition of function equality, for two functions to be equal they must have the same domain and codomain.
This is a very important distinction, and one that most mathematicians and students rarely pay attenttion to.

Finally, we can define:

Definition 1.3.1.13. *A function $f : A \to B$ is called a **bijection** if it is both an injection and a surjection.*

Example(s)

None of the previous examples are bijections, so we have to come up with new examples.
Let $f : \{a, b, c\} \to \{1, 2, 3\}$ be defined by $f(a) = 1, f(b) = 2, f(c) = 3$. Then f is both injective and surjective, and, therefore, a bijection by definition.

Let $g : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be defined by $g(x) = x^2$. Then g is both injective (since there's only one sign on the domain) and surjective (since that are no negatives on the codomain), and, therefore, bijective by definition.

Let $h : \{a, b, c\} \to \{a, b, c\}$ be defined by $h(a) = a, h(b) = c, h(c) = b$. Is h a bijection? Well, it clearly is both injective and surjective, so it has to be by definition.
Notice that bijections don't have to abide by our expectations (such is life).

Lemma 1.3.1.14. *If $f : A \to B$ is injective, then there is a bijection $g : A \to f(A)$.*

Proof

Let $g : A \to f(A)$ be defined by $g(a) := f(a)$ for all $a \in A$.

- g is injective:

  To see that, take $a, a' \in A$ such that $g(a) = g(a')$. Then, by definition, this implies $f(a) = f(a')$, and since f is injective, this in turn implies $a = a'$, so g is injective.

- g is surjective:

  To see that, take any $b \in f(A)$. By definition of image, there exists some $a \in A$ such that $b = f(a)$. But now, by defintion of g, this means that $b = g(a)$.

  We have just shown that every point in the codomain of g is the image of some point in the domain of g under g - this means that g is surjective.

Since g is both injective and surjective, it is, by definition, a bijection, which ends the proof. □

We can use this lemma to easily determine whether a function is, or isn't, an injection.

Example(s)

Let $f : \{a, b, c\} \to \{1, 2\}$ be defined by $f(a) = 1, f(b) = 2, f(c) = 1$. Is f injective?
Well, f takes two different points (a and c) to the same point (1), so it can't be injective.

Actually - is it possible for there to be an injective function from $\{a, b, c\}$ to $\{1, 2\}$?
Let's try making one: First, we choose an image for a - it can be either 1 or 2 - doesn't matter which. Now, to choose an image for b we can't choose the same point as we chose for a - otherwise f won't be injective. So b's image is now uniquely determined: the only point left in $\{1, 2\}$ after we take out $f(a)$. Finally, when we try to choose an image for c, it can't

be $f(a)$, nor can in be $f(b)$ (otherwise, f wouldn't be injective). But $\{1,2\} = \{f(a), f(b)\}$ - that is, if $f(c)$ can't be $f(a)$ and it can't be $f(b)$, then there's **nothing** that it can be!
But, on the other hand, since f is a function, we **have to** take c somewhere. This means that we **have to** repeat either $f(a)$ or $f(b)$.
This shows that there are no injective functions from $\{a,b,c\}$ to $\{1,2\}$.

Lemma 1.3.1.15. *If $f : A \to B$ is surjective, then there is a bijection $g : f(A) \to B$.*

Proof

Let $g : f(A) \to B$ be defined by $g(b) := b$ for all $b \in f(A)$.

- g is surjective:

  To see that, take any $b \in B$. Since f is surjective, for each point in B there is at least one point in A which is its inverse image under f - in particular, there is some $a \in A$ such that $f(a) = b$. But this means that $b \in f(A)$, by definition of image of f. Now, since $b \in f(A)$, we see that $g(b) = b$ and, therefore, g is surjective.

- g is injective:

  To see that, take any two points $b, b' \in f(A)$ such that $g(b) = g(b')$. But, by definition of g, this is the same as saying $g = g'$ - therefore g is injective.

Since g is both injective and surjective, it is, by definition, a bijection, which ends the proof. □

Analogously to injections, this lemma gives us a clear cut method for distinguishing surjections:

Example(s)

Let $f : \{1,2\} \to \{a,b,c\}$ be given by $f(1) = a$ and $f(2) = b$. Clearly, then, f isn't surjective, because there is one point in its codomain (c) which is not the image of any point of the domain under f.

And then we ask: Can there ever be a surjective function from $\{1,2\}$ to $\{a,b,c\}$?
Once again, let's try building one: First, we choose $f(1)$. It can be anything, so choose anything. Now to choose $f(2)$, there's also no restrictions, but remember that we're trying to make a function that "covers" $\{a,b,c\}$ with guys from $\{1,2\}$, so even though we could put $f(2) := f(1)$, it makes sense to choose $f(2)$ to be anything aside from $f(1)$... And we're done.
Notice, however, that no matter **how** we do that choice, there'll always be some point left in $\{a,b,c\}$. Therefore, there can be no surjections from $\{1,2\}$ to $\{a,b,c\}$.

These last two examples give us a nice intuition of what injections and surjections measure: Injections measure how much "smaller" the domain is, when compared to the codomain, and surjections measure how much "bigger" the codomain is, when compared to the domain.

This allows us to consider one final example:

Example(s)

Let $f : \{a, b, c\} \to \{1, 2, 3\}$ be a function. Can f be a bijection?
Let's try: First, we choose any of $\{1, 2, 3\}$ to be $f(a)$. Now, since we want f to be a bijection, it needs to be injective and surjective, so we can't choose $f(b) = f(a)$, so choose $f(b)$ to be any of $\{1, 2, 3\} \setminus \{f(a)\}$. Again, by the same reasoning, choose $f(c)$ to be any of $\{1, 2, 3\} \setminus \{f(a), f(b)\}$ - which isn't really a choice, since there's only one point left.
And we're done! By construction, $f(a) \neq f(b)$, $f(a) \neq f(c)$ and $f(b) \neq f(c)$ (so f is injective) and all of $\{1, 2, 3\}$ have inverse images.
This is a strong intuition that we want to build at this point:
Bijections between two sets tell us if they have the same amount of points. In many ways, then, bijections can be thought of as a relabeling of your set - or even, in some cases, as the *definitive and improved* notion of set equality.
And it makes sense - why should the sets $\{a, b, c\}$ and $\{1, 2, 3\}$ be treated as being different? You might argue that $1 + 2 = 3$, but $a + b$ doesn't even make sense - but the point here is that even $1 + 2$ doesn't make sense. There's no operations being taken into consideration, nothing. Just sets with elements. The only information we have is that "$\{a, b, c\}$ is a set with three distinct things inside it" and that "$\{1, 2, 3\}$ is a set with three distinct things inside it". *What* those things are doesn't really matter to us from a set-theoretical POV. What matters is that there are some things.

To expand in that idea - that bijections are the new equality - we're gonna start a more technical subsection.

The reader is encouraged to **not** skip this section, although I don't own you, so you do you. This subsection will have many proofs, so it's good for practicing your proofs, but not only that - the reasoning employed here is central to understanding what's behind many of the most intricate results in linear algebra.

1.3.2 Bijection as an equality

Definition 1.3.2.1. *Given any two functions $f : A \to B$ and $g : B \to C$, we call the function $g \circ f : A \to C$ defined by $(g \circ f)(a) := g(f(a))$ the **composition** of f and g.*

Definition 1.3.2.2. *Given any set X, we call the function $\mathrm{id}_X : X \to X$ defined by $\mathrm{id}_X(x) := x$ the **identity function** of X.*

Definition 1.3.2.3. *Given a function $f : A \to B$, we say that f **is an isomorphism** if there is some function $g : B \to A$ such that $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$. In this case, we say that g is an **inverse** for f.*

Proposition 1.3.2.4. *Function composition is associative - that is, if $A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D$, then $h \circ (g \circ f) = (h \circ g) \circ f$.*

Proof

Take $a \in A$. Then,

$$\begin{aligned}(h \circ (g \circ f))(a) &= h((g \circ f)(a)) \\ &= h(g(f(a))) \\ &= (h \circ g)(f(a)) = ((h \circ g) \circ f)(a)\end{aligned}$$

and therefore $(h \circ (g \circ f))(a) = ((h \circ g) \circ f)(a)$ for any $a \in A$ which, by the definition of function equality, implies that $h \circ (g \circ f) = (h \circ g) \circ f$. □

Proposition 1.3.2.5. *Given any function $f : A \to B$, we have that $f = \mathrm{id}_B \circ f = f \circ \mathrm{id}_A$.*

Proof

Take any $a \in A$. Then:

$$(\mathrm{id}_B \circ f)(a) = \mathrm{id}_B(f(a)) = f(a) = f(\mathrm{id}(a)) = (f \circ \mathrm{id}_A)(a)$$

which implies, by the definition of function equality, that $\mathrm{id}_B \circ f = f = f \circ \mathrm{id}_A$. □

Proposition 1.3.2.6. *Let $f : A \to B$ be an isomorphism, and let $g, h : B \to A$ be two inverses for f. Then $g = h$.*

Proof

This follows from the two preceding propositions and the definition of isomorphism:

$$g = g \circ \mathrm{id}_B = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_A \circ h = h.$$

□

Definition 1.3.2.7. *Given an isomorphism f, we will denote its (unique!) inverse by f^{-1}.*

Definition 1.3.2.8. *A function $f : A \to B$ is called a **monomorphism** if given any other two functions $g, h : C \to A$, we have that $f \circ g = f \circ h$ implies $g = h$.*

Example(s)

Let $f : \{1, 2, 3\} \to \{a, b, c, d\}$ be defined by $f(1) = a, f(2) = b, f(3) = c$. We claim that f is a monomorphism.
To see that, take any $g, h : C \to \{1, 2, 3\}$ such that $f \circ g = f \circ h$. In particular, for any $x \in C$ we have that $f(g(x)) = f(h(x))$. Well, this means that $f(g(x))$ is either a, b or c. In either case, we know precisely who $g(x)$ is (for instance, if $f(g(x))) = b$, then $g(x) = 2$, since 2 is the only point which is taken to b via f).
But since $f(g(x)) = f(h(x))$, there's a unique $u \in \{1, 2, 3\}$ such that $y = g(x) = h(x)$. In particular, $g(x) = h(x)$.
This shows that $g = h$, and, therefore, f is a monomorphism.

Theorem 1.3.2.9. *A function is a monomorphism if, and only if, it is an injection.*

Proof

Assume that $f : A \to B$ is injective. Then given $g, h : C \to A$ such that $f \circ g = f \circ h$ we want to show that $g = h$. Since f is injective, $f(g(c)) = f(h(c))$ implies $g(c) = h(c)$, for all $c \in C$. It follows, then, that $g = h$ and f is monic.

Conversely, if $f : A \to B$ is monic, define $g, h : \{c\} \to A$ by putting $g(c) = a$ and $h(c) = a'$ for two $a \neq a' \in A$ fixed. Now, since f is monic, by assumption, and since $g \neq h$, we have that $f \circ g \neq f \circ h$ (otherwise we would have f monic, $f \circ g = f \circ h$ and $g \neq h$ all being true, which is impossible). But then:

$$f(a') = f(h(c)) = f \circ h(c) \neq f \circ g(c) = f(g(c)) = f(a)$$

that is, $a \neq a'$ assures us that $f(a) \neq f(a')$, and so f is injective.
Notice that we used the fact that there are two distinct points in A: a and a'. If, however, A has only one point it is even simpler: Any function from a set with a single point has to be injective - in particular, monomorphisms whose domain are a single point are injective.
This finishes the proof. □

Lemma 1.3.2.10. *Every isomorphism is a monomorphism.*

Proof

Let $f : A \to B$ be an isomorphism and $g, h : C \to A$ any two functions such that $f \circ g = f \circ h$. Then, since f is an isomorphism, there is a unique inverse $f^{-1} : B \to A$ such that $\mathrm{id}_A = f^{-1} \circ f$

and $\mathrm{id}_B = f \circ f^{-1}$. Therefore:

$$g = \mathrm{id}_A \circ g = (f^{-1} \circ f) \circ g = f^{-1} \circ (f \circ g) = f^{-1} \circ (f \circ h) = (f^{-1} \circ f) \circ h = \mathrm{id}_A \circ h = h$$

and we see that f is monic. □

Definition 1.3.2.11. *A function $f : A \to B$ is called an **epimorphism** if given any other two functions $g, h : B \to C$, we have that $g \circ f = h \circ f$ implies $g = h$.*

Example(s)

Let $f : \{a, b, c\} \to \{1, 2\}$ be defined by $f(a) = f(b) = 1$ and $f(c) = 2$. We claim that f is epic.
To see that, take any two functions $g, h : \{1, 2\} \to C$ such that $g \circ f = h \circ f$. We want to show that $g = h$ - that is, for any $x \in \{1, 2\}$, we have that $g(x) = h(x)$.
But since f is surjective (check!), there is some $y \in \{a, b, c\}$ such that, then $x = f(y)$, so

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

and therefore we see that $g = h$, which proves that f is indeed an epimorphism.

Theorem 1.3.2.12. *A function $f : A \to B$ is an epimorphism if, and only if, it is a surjection.*

Proof

First, let us assume that $f : A \to B$ is surjective. Then, given $g, h : B \to C$ such that$g \circ f = h \circ f$ we wish to show that $g = h$. We can just proceed as above: Proving that $g = h$ is the same as proving that for all $x \in B$, we have that $g(x) = h(x)$, but since f is surjective, by assumption, we have that there is some $y \in A$ such that $x = f(y)$. It follows then that

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

and therefore $h = g$, which shows that f is epic.
Conversely, assume that f is epic, and let $C = \{c, c'\}$. Now pick a point $b \in B$ and define $g, h : B \to C$ by putting $g(x) = c$ for all $x \in B$, $h(b) = c'$ and $h(x) = c$ if $x \neq b$.
Now, $g(b) \neq h(b)$, so $g \neq h$. Since f is epic, we must then have that $g \circ f \neq h \circ f$, by definition of epimorphism. This means that there is some $a \in A$ such that $g(f(ay)) \neq h(f(a))$.
Since g takes everyone to c, the only possible value of $h(f(a))$ that could be different from that is is $h(f(a)) = c'$, but the only element of B that is taken to c' by h is b - this means that $f(a) = b$.
We have just proven that given any $b \in B$ there is some $a \in A$ such that $f(a) = b$ - that is, f is surjective, which finishes the proof. □

Lemma 1.3.2.13. *Every isomorphism is an epimorphism.*

Proof

Let $f : A \to B$ be an isomorphism and $g, h : B \to C$ two functions such that $g \circ f = h \circ f$. Since f is an isomorphism, it has an inverse $f^{-1} : B \to A$. Therefore:

$$g = g \circ \mathrm{id}_B = g \circ (f \circ f^{-1}) = (g \circ f) \circ f^{-1} = (h \circ f) \circ f^{-1} = h \circ (f \circ f^{-1}) = h \circ \mathrm{id}_B = h$$

and we see that $g = h$, which proves that f is an epimorphism. □

Definition 1.3.2.14. *A function $f : A \to B$ is called a **bimorphism** if it is a mono-epimorphism.*

Clearly, by what we've already shown, bijections and bimorphisms are the same thing. However, we can do one better than that:

Lemma 1.3.2.15. *Every monomorphism $f : A \to B$ has a left-inverse, that is, a function $g : B \to A$ such that $\mathrm{id}_A = g \circ f$, and every function which has a left-inverse is a monomorphism.*

Proof

Let $f : A \to B$ be a monomorphism and consider $f(A)$. Since f is monic, by theorem 1.3.2.9 we see that f is injective, which means that for all $b \in f(A)$ we have that $f^{-1}(b)$ is a single point in A.
We then define $g : B \to A$ by

$$g(b) = \begin{cases} f^{-1}(b), & \text{if } b \in f(A) \\ a, & \text{otherwise,} \end{cases}$$

where $a \in A$ is any (literally any) element of A.
We claim that this g is a left-inverse for f. Indeed, for any $x \in A$ we have

$$(g \circ f)(x) = g(f(x)) = f^{-1}(f(x)) = x = \mathrm{id}_A(x)$$

since $f(x) \in f(A)$ for all $x \in A$. Therefore, we have shown that for all x we have $(g \circ f)(x) = x = \mathrm{id}_A(x)$ - which implies, by the definition of function equality, that $g \circ f = \mathrm{id}_A$.

Take now $f : A \to B$ a function that has a left-inverse $g : B \to A$ - that is, $\mathrm{id}_A = g \circ f$. Now take two functions $h, j : C \to A$ such that $f \circ h = f \circ j$. We want to show that $h = j$ (and therefore, f is monic).
Since $f \circ h = f \circ j$, we can compose g on the left on both sides of the equation to obtain $g \circ (f \circ h) = g \circ (f \circ j)$, which, by proposition 1.3.2.4, is the same as $(g \circ f) \circ h = (g \circ f) \circ j$, and since g is a left-inverse to f, we can further affirm that this is the same as $\mathrm{id}_A \circ h = \mathrm{id}_A \circ j$. Finally, by the definition of id_A, we see that this implies $h = j$ - and therefore f is monic, as stated.
This finishes the proof. □

Lemma 1.3.2.16. *Every epimorphism $f : A \to B$ has a right-inverse, that is, a function $g : B \to A$ such that $\mathrm{id}_B = f \circ g$, and every function which has a right-inverse is an epimorphism.*

Proof

Let $f : A \to B$ be an epimorphism, and consider $f(A)$ its image. By theorem 1.3.2.12, we know that g is a surjection. Since g is a surjection, then, for every $b \in B$ the set $f^{-1}(b)$ is well defined (by definition of surjection).
Now, choose $a_b \in f^{-1}(b)$ for each $b \in B$ (here the index is simply so we know where it came from), and consider the function $g : B \to A$ taking each b to the a_b we chose above.
This is clearly a function (check!), and so we can do, for evrey $b \in B$:

$$(f \circ g)(b) = f(g(b)) = f(a_b) = b = \mathrm{id}_B(b)$$

and therefore $f \circ g$ and id_B are equal in every point - which means that they're equal, and g is a right-inverse for f, as stated.

Take now $f : A \to B$ a function with a right-inverse $g : B \to A$ - that is, $f \circ g = \mathrm{id}_B$. Now take two functions $h, j : B \to C$ such that $h \circ f = j \circ f$. We want to show that $h = j$ (and, therefore, f is epic).
Since $h \circ f = j \circ f$, we can compose g on the right on both sides of the equation to obtain $(h \circ f) \circ g = (j \circ f) \circ g$, which, by proposition 1.3.2.4, is the same as $h \circ (f \circ g) = j \circ (f \circ g)$, and since g is a right-inverse to f, we can further affirm that this is the same as $h \circ \mathrm{id}_B = j \circ \mathrm{id}_B$. Finally, by the definition of id_B, we see that this implies $h = j$ - and therefore f is epic, as stated.
This finishes the proof. □

Lemma 1.3.2.17. *If a function $f : A \to B$ is such that $g, h : B \to A$ are a left- and a right-inverse, respectively, then $g = h$, f is an isomorphism and g is its inverse.*

Proof

It follows trivially by the following computation:

$$g = g \circ \mathrm{id}_B = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_A \circ h = h,$$

which shows at once that $g = h$. This means that $\mathrm{id}_A = g \circ f$ and $\mathrm{id}_B = f \circ g$ - and therefore g is an inverse to f, which shows that f is an isomorphism, as stated. □

Theorem 1.3.2.18. *A function f is an isomorphism if, and only if, it is a bimorphism.*

Proof

In light of lemmas 1.3.2.10 and 1.3.2.13, we see that every isomorphism is monic and epic and, therefore, a bimorphism.
Conversely, by lemmas 1.3.2.15 and 1.3.2.16 we see that any bimorphism has both a left- and a right-inverse. But now, lemma 1.3.2.17 shows us that since every bimorphism has a left- and a right-inverse, it must be an isomorphism, which ends the proof. □

Corollary 1.3.2.19. *Every bijection has a unique inverse.*

And now, finally, to end this section, some technical results that appear all the time in mathematics.

Lemma 1.3.2.20. *$A \xrightarrow{f} B \xrightarrow{g} C$ be two functions. Then the following hold:*

(a) If both f and g are monic, then so is $g \circ f$;

(b) If both f and g are epic, then so is $g \circ f$;

(c) If both f and g are iso, then so is $g \circ f$;

(d) If $g \circ f$ is monic, then so is f;

(e) If $g \circ f$ is epic, then so is g;

(f) If $g \circ f$ and f are iso, then so is g;

(g) If $g \circ f$ and g are iso, then so is f.

Proof

(a) Assume both f and g are monic, and let $f' : B \to A$ and $g' : C \to B$ be their respective left-inverses. We claim that $f' \circ g'$ is a left-inverse to $g \circ f$. Indeed:

$$(f' \circ g') \circ (g \circ f) = f' \circ (g' \circ g) \circ f = f' \circ \mathrm{id}_B \circ f = f' \circ f = \mathrm{id}_A$$

so $g \circ f$ is monic.

(b) Assume both f and g are epic, and let $f' : B \to A$ and $g' : C \to B$ be their respective right-inverses. We claim that $f' \circ g'$ is a right-inverse to $g \circ f$. Indeed:

$$(g \circ f) \circ (f' \circ g') = g \circ (f \circ f') \circ g' = g \circ \mathrm{id}_B \circ g' = g \circ g' = \mathrm{id}_C$$

so $g \circ f$ is epic.

(c) Follows immediately from (a) and (b).

(d) If $g \circ f$ is monic, by lemma 1.3.2.15 we see that there is some function $h : C \to A$ that is a left-inverse to $g \circ f$ - that is, $\mathrm{id}_A = h \circ (g \circ f)$. But now, by proposition 1.3.2.4, we

see that $h \circ (g \circ f) = (h \circ g) \circ f$ and, therefore, $\mathrm{id}_A = (h \circ g) \circ f$ and we see that $h \circ g$ is a left-inverse for f. Now the converse of lemma 1.3.2.15 tells us that since f has a left-inverse, it must be monic.

(e) If $g \circ f$ is epic, by lemma 1.3.2.16 we see that there is some function $h : B \to C$ that is a right-inverse to $g \circ f$ - that is, $\mathrm{id}_B = (g \circ f) \circ h$. But now, by proposition 1.3.2.4, we see that $(g \circ f) \circ h = g \circ (f \circ h)$ and, therefore, $\mathrm{id}_B = g \circ (f \circ h)$ and we see that $f \circ h$ is a right-inverse to g. Now, the converse of lemma 1.3.2.16 tells us that since g has a right-inverse, it must be epic.

(f) If f is iso, then so is f^{-1} (its inverse is precisely f). Therefore, the equality $g \circ f = g \circ f$ yields the equality $(g \circ f) \circ f^{-1} = g$ by composing f^{-1} to the right on both sides of the equality. Now we use item (c) to conclude that since g is the composition of two isomorphisms, it is also an isomorphism.

(g) If g is iso, then so is g^{-1} (its inverse is precisely g). Therefore, the equality $g \circ f = g \circ f$ yields the equality $g^{-1} \circ (g \circ f) = f$ by composing g^{-1} to the left on both sides of the equality. Now we use item (c) to conclude that since f is the composition of two isomorphisms, it is also an isomorphism.

This ends the proof. □

And finally we end this section with a definition.

Definition 1.3.2.21. *Two sets are said to have* ***the same cardinality*** *if they are isomorphic - that is, if there is an isomorphism between them. If A and B have the same cardinality, we will represent that in symbols by $\#A = \#B$.*

Remark 1.3.2.22

Notice that for finite sets, $\#A$ is precisely the formalization of the intuitive notion of "number of elements of A".

1.3.3 Universal properties

Now that we have dealt with functions and their properties, we can use them to define new sets.

Before that, though, let's have a quick talk about *universal properties*.

A *universal property* is a way of defining something by saying it's somewhat singular in the universe. For instance, when defining the union and intersection of two sets, we didn't use the classical definition, but, instead, used a universal property to define those sets.

Here's the advantage of working with universal properties:

Definition 1.3.3.1. *We say that a set X is **initial** if there is a unique function from X to any other set.*

*Similarly, we say that X is **terminal** if there is a unique function from any other set to X.*

Meta-theorem

All universal properties can be coded in terms of initial/terminal objects on a specific class of sets.

For instance, the union of A and B is the *initial* set in the class of all sets containing A and B. Analogously, the intersection of A and B is the *terminal* set in the class of all sets contained in A and B.

Meta-theorem

All sets defined by universal properties are uniquely defined (up to isomorphism).

This means that if X is initial/terminal regarding a certain class of sets, then it is the unique set in that class that is initial/terminal (not counting sets that are isomorphic to X).

Proof

Let X and Y be two sets which are initial regarding a certain class of sets. Since X is initial, there's a unique function $!_Y : X \to Y$. Since Y is initial, there's a unique function $!_X : X \to Y$. This gives us, by composition, a function $!_X \circ !_Y : X \to X$ and a function $!_Y \circ !_X : Y \to Y$.

Now remember that for any set, there's always its identity map. So we have $\mathrm{id}_X : X \to X$ and $\mathrm{id}_Y : Y \to Y$.

But since X is initial, there's a unique function from X to itself. Since there's always an identity function, that function must be the unique map. But we've just shown that $!_X \circ !_Y$ is also a map from X to itself. It follows then that $\mathrm{id}_X = !_X \circ !_Y$.

Arguing similarly for Y we can show that $\mathrm{id}_Y = !_Y \circ !_X$, and, therefore, $!_Y$ is an inverse for $!_X$, and hence they are isomorphisms.

The proof for the terminal case is identical and left as an exercise to the reader. □

This is another reason why isomorphism is a better notion of set equality - because sets defined by universal properties are unique, up to isomorphism.

Finally, before defining new sets using universal properties, let us prove a couple of interesting results:

Lemma 1.3.3.2. *The set $\varnothing$ is the initial set of all sets.*

Proof

There clearly is only one function from $\varnothing$ to any other set.
To see this, think of what would have to go wrong for there to be two different functions, f and g: We'd have to have one element x of $\varnothing$ such that $f(x) \neq g(x)$. But $\varnothing$ doesn't have any elements, so any two functions defined on it must be equal. □

Lemma 1.3.3.3. *The set $\{a\}$ is the terminal set of all sets.*

Proof

There clearly is only one function from any set to $\{a\}$: The function sending all elements of your domain to a. □

Corollary 1.3.3.4. *There is only one function from any singleton (i.e. a set with a single element) to $\{a\}$.*

Proof

Follows trivially by the preceding lemma. □

Corollary 1.3.3.5. *Any two singletons are isomorphic.*

Proof

It is also trivial to prove that any other singleton is also a terminal object. Therefore, it must be isomorphic to $\{a\}$.
Take, then, two singletons $*$ and $\bullet$, and do:

$$* \leftrightarrow \{a\} \leftrightarrow \bullet,$$

where the $\leftrightarrow$ denote the unique isomorphism between the two sets. This is a composition of isomorphisms and, therefore, an isomorphism. □

Corollary 1.3.3.6. *All terminal objects are singletons.*

Proof

Take T any terminal object.
By the preceding lemma, it must be isomorphic to a singleton $\{a\}$. This means that there is a bijection $f : \{a\} \to T$. This means, by lemma 1.3.1.14 and lemma 1.3.1.15, that the image of this isomorphism is isomorphic to $\{a\}$ (since isomorphisms are injective), and the image is also equal to T (since isomorphisms are surjective).
But since the image is a singleton ($\operatorname{Im} f = \{f(a)\}$), we have that T is a singleton as well, which ends the proof. □

Remark 1.3.3.7

For reasons that will become clearer further ahead, for here onwards we're gonna denote the unique initial set by 0 and the unique terminal set by 1 (sometimes by $*$ to avoid misconceptions and misunderstandings).

1.3.4 Multiplying sets

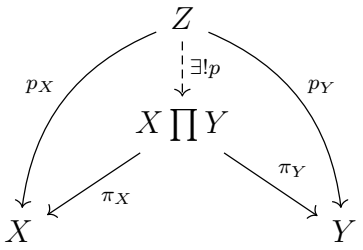
Now that this is done, let us define new sets using universal properties:

Definition 1.3.4.1. *Let X and Y be two sets. We define the* ***product of*** X ***and*** Y *to be the set $X \prod Y$ which is terminal in the class of sets with functions to both X and Y - this means that:*

i. There are functions $\pi_X : X \prod Y \to X$ and $\pi_Y : X \prod Y \to Y$;

ii. If Z is some set with functions $p_X : Z \to X$ and $p_Y : Z \to Y$, then there is a unique function $p : Z \to X \prod Y$ such that $p_X = \pi_X \circ p$ and $p_Y = \pi_Y \circ p$.

We usually denote this saying that the following diagram commutes:



This means that no matter which path we take on the diagram, the end result should be the same.

Now, this may seem very abstract and weird at first. But I assure you that you already know what that is.

Definition 1.3.4.2. *Let X and Y be two sets. We define the* ***cartesian product of*** X ***and*** Y *to be the set $X \times Y$ of all ordered pairs (x, y) such that $x \in X$ and $y \in Y$.*

This is a well-known definition.

Example(s)

Let $A = \{a, b, c\}$ and $B = \{1, 2\}$. Then $A \times B = \{(a, 1), (a, 2), (b, 1), (b, 2), (c, 1), (c, 2)\}$ is the cartesian product of A and B.
Let $C = \{$yellow pants, brown pants, shorts$\}$ and $D = \{$crop top, black shirt, sweater, jacket$\}$. Then $C \times D$ is the set of all possible combinations of pants types (C) and shirt types (D) in your closet.

Now, since the two previous definitions have such similar names they must be related in some way, right? Well...

Proposition 1.3.4.3. *For any two sets X and Y, their cartesian product $X \times Y$ is the product $X \prod Y$.*

Proof

We have to show two things: (i.) There are functions $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ and; (ii.) It is terminal with that property.

To show (i.), let us define π_X and π_Y as follows: $\pi_X(x, y) := x$ and $\pi_Y(x, y) := y$ for all $(x, y) \in X \times Y$. These are clearly functions from $X \times Y$ to both X and Y, so (i.) is done.

Now take any other set Z with functions $p_X : Z \to X$ and $p_Y : Z \to Y$. We want to build a function $p : Z \to X \times Y$ such that $p_X = \pi_X \circ p$ and $p_Y = \pi_Y \circ p$, and show that it is the unique function with that property.
Well, take $z \in Z$ and follow it around: If we apply p_X to z we get $p_X(z) \in X$, and, similarly, applying p_Y we get $p_Y(z) \in Y$. Since $p_X(z)$ is in X and $p_Y(z)$ is in Y, by definition of cartesian product we have that $(p_X(z), p_Y(z))$ is in $X \times Y$.
Now, clearly we have that $\pi_X(p_X(z), p_Y(z)) = p_X(z)$ and $\pi_Y(p_X(z), p_Y(z)) = p_Y(z)$ (this is precisely how we defined π_X and π_Y above). So it is obvious what we must define $p : Z \to X \times Y$ to be:

$$p(z) := (p_X(z), p_Y(z))$$

for all $z \in Z$.
This is clearly a function from Z to $X \times Y$ and, by definition, $p_X = \pi_X \circ p$ and $p_Y = \pi_Y \circ p$.

To finish this, we need to show that this p is unique. Well, suppose there is another function, $q : Z \to X \times Y$ such that $p_X = \pi_X \circ q$ and $p_Y = \pi_Y \circ q$. But then, for every $z \in Z$ we'd have that $p_X(z) = \pi_X(q(z))$ and $p_Y(z) = \pi_Y(q(z))$.
This means that $q(z)$ is a point in $X \times Y$ whose X-coordinate is $p_X(z)$ (this is what the equation $p_X(z) = \pi_X(q(z))$ tells us), and whose Y-coordinate is p_Y (this is what the equation $p_Y(z) = \pi_Y(q(z))$ tells us).
Therefore, $q(z) = (p_X(z), p_Y(z))$ and the RHS is just $p(z)$, by definition. So we have $q(z) = p(z)$ for all $z \in Z$ - which implies, by the definition of function equality, that $q = p$.
It follows then that the p we've defined is the unique function with that property, so $X \times Y$ is indeed the product of X and Y, which ends the proof. □

What's the advantage of defining via universal property instead of just outright using the classical definition? Well...

It's easy to prove that, using the classical definition, we can take any finite number of sets $\{A_i\}_{i \leq n}$ and take their product $A_1 \times A_2 \times \cdots \times A_n$ to be the iterated product:

A_1 and A_2 are well-defined, so $A_1 \times A_2$ is well-defined.
But now, $A_1 \times A_2$ and A_3 are well-defined, so $(A_1 \times A_2) \times A_3$ is well-defined.
And so on, up to A_n.

But try doing that for infinitely many sets. Heck, try doing that for an uncountable amount of sets.

It's not easy to see how to even *define* such an operation if the set of indices isn't, say, ordered.

However, using the universal property, we can define the product of *any* amount of sets:

Definition 1.3.4.4. *Let $\{A_i\}_{i\in I}$ be a collection of sets indexed by another set I (which can be infinite or finite, countable or uncountable, doesn't matter, as long as it's a set). We define $\prod_{i\in I} A_i$ to be the set given by:*

i. There is a function $\pi_n : \prod_{i\in I} A_i \to A_n$ for each $n \in I$;

ii. If Z is a set with a function $p_n : Z \to A_n$ for each $n \in I$, then there's a unique function $p : Z \to \prod_{i\in I} A_i$ such that $p_n = \pi_n \circ p$ for each $n \in I$.

Which is just the *same* definition we used for the product of two sets, but generalized for *any* amount of sets.

This is another great reason to prefer definitions via universal properties instead of explicit ones.

Definition 1.3.4.5. *Given three sets A, B and C with functions $f : A \to B$ and $g : A \to C$, the unique function from A to $B \times C$ induced by the definition of product will be called the **product map of** f **and** g and denoted by $f \times g : A \to B \times C$.*

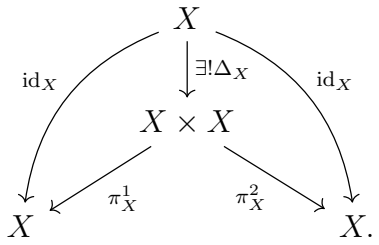
By definition, $(f \times g)(a) := (f(a), g(a))$ for any $a \in A$.

Example(s)

Let $\mathbb{R}$ be the set of real numbers, $f, g : \mathbb{R} \to \mathbb{R}$ be defined by $f := \mathrm{id}_{\mathbb{R}}$ and $g(x) := x^2$ for any $x \in \mathbb{R}$. Then the product map $f \times g : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ is the map $x \mapsto (x, x^2)$ for any $x \in \mathbb{R}$.

Definition 1.3.4.6. *The product map of id_X and id_X, for any set X will be called the **diagonal map of** X and denoted by $\Delta_X : X \to X \times X$.*

It is, as in the above definition, the unique map commuting the diagram



Example(s)

Let $A = \{a, b, c\}$. Let's calculate Δ_A.
By definition, we must have $\mathrm{id}_A = \pi^1_A \circ \Delta_A$ and $\mathrm{id}_A = \pi^1_A \circ \Delta_A$. So take any $a \in A$.
We know that, from the first equation, we must have $a = (\pi^1_A \circ \Delta_A)(a) = \pi^1_A(\Delta_A(a))$, so the first coordinate of $\Delta_A(a)$ must be a.
Similarly, the second equation gives us $a = (\pi^2_A \circ \Delta_A)(a) = \pi^2_A(\Delta_A(a))$ and so, the second coordinate of $\Delta_A(a)$ must also be a.

Since $\Delta_A(a) = (\pi_A^1(\Delta_A(a)), \pi_A^2(\Delta_A(a)))$, we get that $\Delta_A(a) = (a, a)$ for any $a \in A$.
This is why it's called the *diagonal* map.

Remark 1.3.4.7

From here onwards, the symbol $\cong$ will mean "is isomorphic to" - so $A \cong B$ should be read as "A is isomorphic to B".

Let us then prove some nice properties of products:

Lemma 1.3.4.8. *For any three sets A, B and C the following hold:*

(a) $A \times B \cong B \times A$ *(but not equal);*

(b) $A \times (B \times C) \cong (A \times B) \times C$ *(but not equal);*

(c) $A \times 0 = 0$*;*

(d) $A \times 1 \cong A$ *(but not equal).*

Proof

(a) Since $A \times B$ is the product of A and B, there are functions $\pi_A : A \times B \to A$ and $\pi_B : A \times B \to B$. Similarly, since $B \times A$ is the product of B and A, there are functions $\pi'_B : B \times A \to B$ and $\pi'_A : B \times A \to A$.

Now, $A \times B$ has a function to B and a function to A, so, by definition of product, there's a unique function $\phi : A \times B \to B \times A$ such that $\pi_A = \pi'_A \circ \phi$ and $\pi_B = \pi'_B \circ \phi$. Analogously, since $B \times A$ has functions to A and B, there's a unique function $\psi : B \times A \to A \times B$ such that $\pi'_A = \pi_A \circ \psi$ and $\pi'_B = \pi_B \circ \psi$.

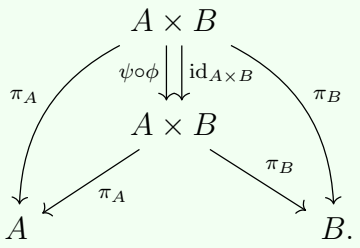
As done previously, we can consider the functions $\mathrm{id}_{A\times B}$ and $\psi \circ \phi$, both from $A \times B$ to itself. Notice that $\pi_A = \pi_A \circ \mathrm{id}_{A\times B}$ and $\pi_B = \pi_B \circ \mathrm{id}_{A\times B}$. Notice also that

$$\pi_A = \pi'_A \circ \phi = (\pi_A \circ \psi) \circ \phi = \pi_A \circ (\psi \circ \phi)$$

and

$$\pi_B = \pi'_B \circ \phi = (\pi_B \circ \psi) \circ \phi = \pi_B \circ (\psi \circ \phi),$$

so both $\mathrm{id}_{A\times B}$ and $\psi \circ \phi$ commute the diagram



But by definition of product, there's a unique function from any set (including $A \times B$ itself) to $A \times B$ for which that holds. Hence, these two functions must be the same - that is, $\psi \circ \phi = \mathrm{id}_{A \times B}$.

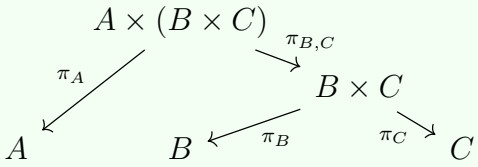
Arguing analogously, we can prove that $\mathrm{id}_{B \times A}$ and $\phi \circ \psi$ are two functions from $B \times A$ to itself which commute the corresponding diagram and, once again, by definition of product, they must be the same - that is, $\phi \circ \psi = \mathrm{id}_{B \times A}$.

Hence, ϕ and ψ are inverses and, therefore, isomorphisms.

(b) We'll show that both $A \times (B \times C)$ and $(A \times B) \times C$ are isomorphic to $A \times B \times C$, so they must be isomorphic (since composition of isomorphisms is isomorphism).

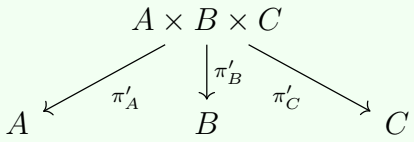
Actually, we'll only show one of these and leave the other one as an exercise to you, reader.

Let



be the functions from $A \times (B \times C)$ to each one of A, B and C (which exist by definition of all the products involved).

Similarly, let



be the functions from $A \times B \times C$ to each one of A, B and C (which exist by definition of product).

Now, by definition of $A \times B \times C$, there's a unique $\phi : A \times (B \times C) \to A \times B \times C$ such that $\pi_A = \pi'_A \circ \phi$, $\pi_B \circ \pi_{B,C} = \pi'_B \circ \phi$ and $\pi_C \circ \pi_{B,C} = \pi'_C \circ \phi$.

But, by definition of $B \times C$, there's a unique $\psi_{B,C} : A \times B \times C \to B \times C$ such that $\pi'_B = \pi_B \circ \psi_{B,C}$ and $\pi'_C = \pi_C \circ \psi_{B,C}$.

Finally, by definition of $A \times (B \times C)$, the above $\psi_{B,C}$ together with π'_A show that there's a unique $\psi : A \times B \times C \to A \times (B \times C)$ such that $\pi'_A = \pi_A \circ \psi$ and $\psi_{B,C} = \pi_{B,C} \circ \psi$.

Now, it's easy to see that using all of the equations above we get that

$$\begin{aligned} \pi'_A \circ (\phi \circ \psi) &= (\pi'_A \circ \phi) \circ \psi \\ &= \pi_A \circ \psi = \pi'_A \end{aligned}$$

$$\begin{aligned}\pi'_B \circ (\phi \circ \psi) &= (\pi'_B \circ \phi) \circ \psi \\ &= (\pi_B \circ \pi_{B,C}) \circ \psi \\ &= \pi_B \circ (\pi_{B,C} \circ \psi) = \pi_B \circ \psi_{B,C} = \pi'_B\end{aligned}$$

$$\begin{aligned}\pi'_C \circ (\phi \circ \psi) &= (\pi'_C \circ \phi) \circ \psi \\ &= (\pi_C \circ \pi_{B,C}) \circ \psi \\ &= \pi_C \circ (\pi_{B,C} \circ \psi) = \pi_C \circ \psi_{B,C} = \pi'_C\end{aligned}$$

and, like before, this shows that $\phi \circ \psi = \mathrm{id}_{A\times B\times C}$ - since $\mathrm{id}_{A\times B\times C}$ is the unique function satisfying those three equalities.

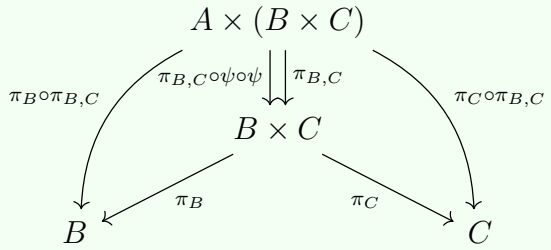
Similarly:

$$\begin{aligned}\pi_A \circ (\psi \circ \phi) &= (\pi_A \circ \psi) \circ \phi \\ &= \pi'_A \circ \phi = \pi_A\end{aligned}$$

$$\begin{aligned}(\pi_B \circ \pi_{B,C}) \circ (\psi \circ \phi) &= \pi_B \circ (\pi_{B,C} \circ (\psi \circ \phi)) \\ &= \pi_B \circ ((\pi_{B,C} \circ \psi) \circ \phi) \\ &= \pi_B \circ (\psi_{B,C} \circ \phi) \\ &= (\pi_B \circ \psi_{B,C}) \circ \phi \\ &= \pi'_B \circ \phi = \pi_B \circ \pi_{B,C}\end{aligned}$$

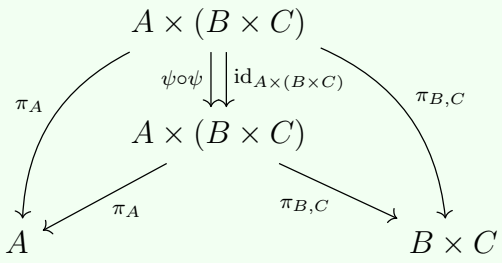
$$\begin{aligned}(\pi_C \circ \pi_{B,C}) \circ (\psi \circ \phi) &= \pi_C \circ (\pi_{B,C} \circ (\psi \circ \phi)) \\ &= \pi_C \circ ((\pi_{B,C} \circ \psi) \circ \phi) \\ &= \pi_C \circ (\psi_{B,C} \circ \phi) \\ &= (\pi_C \circ \psi_{B,C}) \circ \phi \\ &= \pi'_C \circ \phi = \pi_C \circ \pi_{B,C}\end{aligned}$$

and now the last two equalities show that since $\pi_{B,C}$ and $\pi_{B,C} \circ \psi \circ \phi$ are two maps from $A \times (B \times C)$ to $B \times C$ commuting the diagram



and since, by definition of product, that map must be unique, it follows that $\pi_{B,C} = \pi_{B,C} \circ \psi \circ \phi$.

But this, on the other hand, shows that both $\mathrm{id}_{A\times(B\times C)}$ and $\psi\circ\psi$ commute the diagram



which, once again, implies that $\psi\circ\phi = \mathrm{id}_{A\times(B\times C)}$, by definition of product.

This shows that ψ and ϕ are mutually inverse and, therefore, isomorphisms.

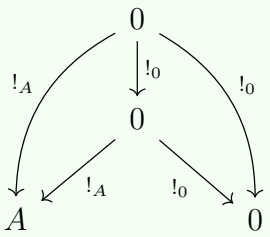
(c) This is equivalent to proving that if there is a function between some set Z and 0, then $Z = 0$.

To see this, consider such a function $f : Z \to 0$. If $Z \neq 0$, then there's at least one point $z \in Z$ - and, therefore, $f(z) \in \mathrm{Im}(f) \subseteq 0$. But 0 is empty, so it has no elements, but we've just showed that if $Z \neq 0$, then $f(z) \in 0$, which is a contradiction. It follows that $Z \neq 0$ is false, and, so, $Z = 0$.

Now, we're going to prove that 0 is the product of A and 0 using the definition:

First, see that since 0 is initial, there's a unique function $!_A : 0 \to A$ and a unique function $!_0 : 0 \to 0$, but since id_0 always exists, we must have $!_0 = \mathrm{id}_0$.

Now let B be a set with functions $f : B \to A$ and $g : B \to 0$. Then $B = 0$ (since g is a function to 0). It follows, then, that $f = !_A$ and $g = !_0$. Clearly, then, the function $!_0 : 0 \to 0$ commutes the diagram



for $!_A$ and $!_A\circ!_0$ are two functions from 0 to A, so they must be equal (since 0 is initial), and similarly for $!_0\circ!_0$ and $!_0$.

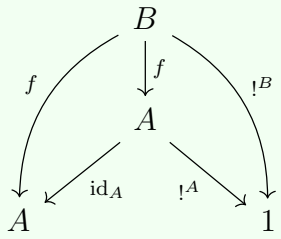
We have just proven that 0 satisfies the universal property defining $A \times 0$, so we must have $0 \cong A \times 0$, but the only set isomorphic to 0 is 0, so we get $0 = A \times 0$.

(d) Similarly, to prove this result we'll use that A satisfies the definition of product of A and 1:

First, notice that there's a function $\mathrm{id}_A : A \to A$ and a unique function $!^A : A \to 1$ (since 1 is terminal).

Now, suppose B is a set with functions $f : B \to A$ and $g : B \to 1$. Since 1 is terminal, $g :=!^B$ the unique function from B to 1.

It is now easy to see that $f : B \to A$ commutes the diagram



for $f = \mathrm{id}_A \circ f$, trivially, and $!^B$ and $!^A \circ f$ are two functions from B to 1, so they must be equal (since 1 is terminal).

We've just shown that A satisfies the universal property which defines $A \times 1$, so, since universal properties define sets uniquely up to isomorphism, we get $A \cong A \times 1$.

□

Remark 1.3.4.9

From here onwards, whenever we multiply a set by itself, we'll take inspiration on numbers and denote the product of n copies of any set A by A^n. Note that this makes sense because of the item (b) above.

If you think about it, there's no immediate reason why these four statements should be true. Yet, the fact that they are true makes it way more reasonable for our naming it the "product" of two sets: Because it looks just like number multiplication.

1.3.5 Adding sets?

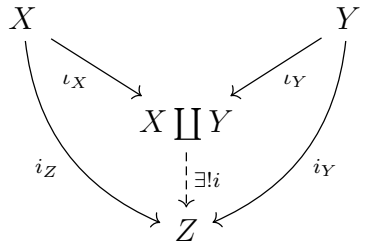
Well, now that we have multiplication, you know what to do next, right?

Definition 1.3.5.1. *Let X and Y be two sets. We define the **coproduct of X and Y** to be the set $X \coprod Y$ which is initial in the class of sets with functions from both X and Y - this means that:*

i. There are functions $\iota_X : X \to X \coprod Y$ and $\iota_Y : Y \to X \coprod Y$;

ii. If Z is some set with functions $i_X : X \to Z$ and $i_Y : Y \to Z$, then there is a unique function $i : X \coprod Y \to Z$ such that $i_X = i \circ \iota_X$ and $i_Y = i \circ \iota_Y$.

We usually denote this saying that the following diagram commutes:



Once again, this all sounds too abstract. What's this so called "coproduct"? Well, it might not seem obvious at first...

Definition 1.3.5.2. *Let A and B be two sets. We define the **disjoin union** of A and B to be the set $A \sqcup B$ of all elements in either A or B, but disregarding equalities.*

This is till too abstract, so lets give an example:

Example(s)

Consider the sets $A = \{a, b\}$ and $B = \{1, 2, 3\}$. Now, since $A \cap B = \varnothing$, we see that the set $A \cup B$ is just taking all elements of A and B, and putting them all in the same box: $A \cup B = \{a, b, 1, 2, 3\}$.
If, however, we take the sets $C = \{a, b, c, d, e\}$ and $D = \{a, e, i, o, u\}$, we see that $C \cup D = \{a, b, c, d, e, i, o, u\}$ which is **not** the same as just taking every element in C and D and putting them all in the same box. That happens because, contrary to the first case where $A \cap B = \varnothing$, in this case we have $C \cap D = \{a, e\} \neq \varnothing$.
In other words, there is at least one $x \in C$ and one $y \in D$ such that $x = y$.
We can, however, fix that problem by "coloring" the elements.

Imagine that C was a large blue paint can, and D was a giant red paint can. Now, putting the elements from C and D together, we can see that the a and the e that came from C are blue, whereas the corresponding elements from D are red.
This is the idea behind disjoint unions:
C, as a set, is clearly isomorphic to the set $C_C := \{a_C, b_C, c_C, d_C, e_C\}$, which has "the same elements, but painted in the color C". Similarly, the set D is clearly isomorphic to the set

$D_D := \{a_D, e_D, i_D, o_D, u_D\}$, which has "the same elements, but painted in the color D".
Now, instead of computing $C \cup D$, we compute

$$C_C \cup D_D = \{a_C, b_C, c_C, d_C, e_C, a_D, e_D, i_D, o_D, u_D\},$$

which is precisely what we would get by taking all elements of C and D and putting them in a box, disregarding equalities.
So we put $C \sqcup D := C_C \cup D_D$, and call it the **disjoint union of C and D**.

This final argument in the example suggests the following result:

Lemma 1.3.5.3. *For any two sets X and Y we have that $X \sqcup Y \cong (X \times \{0\}) \cup (Y \times \{1\})$.*

Proof

It follows directly from the example above and from the fact that we can denote any element in $X \times \{0\}$ (which is of the form $(x, 0)$, by definition) as x_0, just as a shorthand notation (and similarly for $Y \times \{1\}$), so we can just write $X_0 := X \times \{0\}$ and $Y_1 := Y \times \{1\}$..

Define $f : X \sqcup Y \to X_0 \cup Y_1$ by putting

$$f(a) := \begin{cases} a_0, & \text{if } a \in X \\ a_1, & \text{if } a \in Y. \end{cases}$$

- **f is injective:**

  Take $a, b \in X \sqcup Y$ such that $f(a) = f(b)$. We have four different possibilities:

  1. If $a \in X$ and $b \in X$, we see that $a_0 = f(a) = f(b) = b_0$ and hence $a = b$.
  2. If $a \in X$ and $b \in Y$, we see that $a_0 = f(a) = f(b) = b_1$, which is absurd because the second coordinate of b_1 is 1, and the second coordinate of a_X is 0, and $0 \neq 1$, so $b_1 \neq a_0$. So this case cannot happen.
  3. Similarly, if $a \in Y$ and $b \in X$ we'd get that same contradiction, so this case also cannot happen.
  4. Finally, if $a \in Y$ and $b \in Y$, just like the first case we can see that $a_1 = b_1$, and hence $a = b$.

  Since in all cases, the only possible situation which doesn't lead to a contradiction is $a = b$, we see that f is injective.

- **f is surjective:**

  Take $c \in X_0 \cup Y_1$. We want to show that there's some $a \in X \sqcup Y$ such that $f(a) = c$.

  Since $X_0 \cap Y_1 = \varnothing$, c must lie in either X_0 or Y_1 - that is, either $c = x_0$ for some $a \in X$ or $c = y_1$ for some $y \in Y$. But this means that c is either $f(x)$, for some $x \in X$ or $f(y)$, for some $y \in Y$.

No matter which one of these hold, there's always one $a \in X \sqcup Y$ such that $f(a) = c$, for any $c \in X_0 \cup Y_1$. This proves that f is surjective.

Since f is both injective and surjective, it is a bijection and, therefore, an isomorphism. This ends the proof. □

Now, an important question at this point is *WHY?*. I mean, why do we need this new notion of union, when the old one suited us just fine? Well, let me present you some results to convince you on that:

Proposition 1.3.5.4. *For any two sets X and Y, their disjoint union $X \sqcup Y$ is the coproduct $X \coprod Y$.*

Proof

We have to show two things:

(i.) There are functions $\iota_X : X \to X \sqcup Y$ and $\iota_Y : Y \to X \sqcup Y$.

This can easily be seen by considering the lemma 1.3.5.3, and so we can put $\iota_X(x) := x_0$ and $\iota_Y(y) := y_1$ for any $x \in X$ and any $y \in Y$.

(ii.) Let Z be any set with functions $i_X : X \to Z$ and $i_Y : Y \to Z$. We want to show that there's a unique function $i : X \sqcup Y \to Z$ such that $i_X = i \circ \iota_X$ and $i_Y = i \circ \iota_Y$.

This is easy: Define $i : X \sqcup Y \to Z$ by putting

$$i(a) := \begin{cases} i_X(x), & \text{if } a = x_0 \\ i_Y(y), & \text{if } a = y_1. \end{cases}$$

Clearly then we have:

$$(i \circ \iota_X)(x) = i(x_0) = i_X(x)$$

and

$$(i \circ \iota_Y)(y) = i(y_1) = i_Y(y),$$

so our i satisfies the equalities.

Now to prove that it is unique: Suppose there's some $j : X \sqcup Y \to Z$ that also satisfies the equalities. In particular, we have that $(j \circ \iota_X)(x) = (i \circ \iota_X)(x)$ and $(j \circ \iota_Y)(y) = (i \circ \iota_Y)(y)$.

But the RHS of the first equation evaluates to $i(x_0)$, by definition, and the RHS of the second equation evaluates to $i(y_1)$, also by definition.

But, on the other hand, the LHS evaluates to $j(x_0)$ and $j(y_1)$, respectively, again by definition.

Finally, since every element of $X \sqcup Y$ is either in X_0 or in Y_1, it follows that the functions j and i are equal for all elements of X_0 and Y_1 - and, therefore, equal for all of $X \sqcup Y$ - which implies, by definition of equality, $j = i$. This shows that there's a unique function satisfying the equalities

Finally, (i.) and (ii.) together show that $X \sqcup Y$ satisfies the universal property defining the coproduct of X and Y. This finishes the proof. □

Just like with products, we can use coproducts to generalize disjoint unions to infinitely many sets:

Definition 1.3.5.5. *Let $\{A_i\}_{i\in I}$ be a collection of sets indexed by another set I (which can be infinite or finite, countable or uncountable, doesn't matter, as long as it's a set). We define $\coprod_{i\in I} A_i$ to be the set given by:*

i. There is a function $\iota_n : A_n \to \coprod_{i\in I} A_i$ for each $n \in I$;

ii. If Z is a set with a function $i_n : A_n \to Z$ for each $n \in I$, then there's a unique function $i : \coprod_{i\in I} A_i \to Z$ such that $i_n = i \circ \iota_n$ for each $n \in I$.

And we can also get a similar concept to the diagonal map:

Definition 1.3.5.6. *Given three sets A, B and C with functions $f : A \to C$ and $g : B \to C$, the unique function from $A \sqcup B$ to C induced by the definition of coproduct will be called the **coproduct map of f and g** and denoted by $f \sqcup g : A \sqcup B \to C$.*

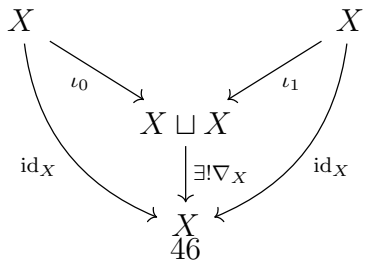
By definition, $(f \sqcup g)(x) := \begin{cases} f(x), & \text{if } x \in A \\ g(x), & \text{if } x \in B. \end{cases}$

Example(s)

Let $\mathbb{R}$ be the set of real numbers, $f, g : \mathbb{R} \to \mathbb{R}$ be defined by $f := \mathrm{id}_{\mathbb{R}}$ and $g(x) := x^2$ for any $x \in \mathbb{R}$. Then the coproduct map $f \sqcup g : \mathbb{R} \sqcup \mathbb{R} \to \mathbb{R}$ is the map $x \mapsto \begin{cases} x, & \text{if } x \in \mathbb{R}_0 \\ x^2, & \text{if } x \in \mathbb{R}_1 \end{cases}$ for any $x \in \mathbb{R} \sqcup \mathbb{R}$.

Definition 1.3.5.7. *The coproduct map of id_X and id_X, for any set X will be called the **fold map of X** and denoted by $\nabla_X : X \sqcup X \to X$.*

It is, as in the above definition, the unique map commuting the diagram



Example(s)

Let $A = \{a, b, c\}$. Let's calculate ∇_A.
By definition, we must have $\mathrm{id}_A = \nabla_A \circ \iota_0$ and $\mathrm{id}_A = \nabla_A \circ \iota_1$. So take any $a \in A$.
We know that, from the first equation, we must have $a = (\nabla_A \circ \iota_0)(a) = \nabla_A(\iota_0(a)) = \nabla_A(a_0)$.
Similarly, the second equation gives us $a = (\nabla_A \circ \iota_1)(a) = \nabla_A(\iota_1(a)) = \nabla_A(a_1)$.
So ∇_A is the function that ignores color. It's basically the colorblind function - in its eyes, x_0 and x_1 are just x.
It is, in some sense, folding x_0 and x_1 on top of eachother - gluing them together. Hence why it's called the *fold* map.

Now, finally, we'll state similar results to the ones in the product section:

Lemma 1.3.5.8. *For any three sets A, B and C the following hold:*

(a) $A \sqcup B \cong B \sqcup A$ *(but not equal);*

(b) $A \sqcup (B \sqcup C) \cong (A \sqcup B) \sqcup C$ *(but not equal);*

(c) $A \sqcup 0 \cong A$ *(but not equal).*

Proof

(a) It is basically the same proof as in the case of products, so we'll leave it to the reader.

(b) This follows from the fact that union is associative, from $X \sqcup Y \cong X_0 \cup Y_1$ and from the fact that all sigletons are isomorphic.

Specifically, write $A \sqcup (B \sqcup C)$ as $A_0 \cup (B \sqcup C)_1$, and writing $B \sqcup C \cong B_0 \cup C_1$, we can further rewrite $A \sqcup (B \sqcup C)$ as $A_0 \cup B_{0,1} \cup C_{1,1}$ - which, upon closer inspection, is just $A_0 \cup B_1 \cup C_2$.

Similarly, we can write $(A \sqcup B) \sqcup C$ as $(A \sqcup B)_0 \cup C_1$, and writing $A \sqcup B \cong A_0 \cup B_1$, we can further rewrite $(A \sqcup B) \sqcup C$ as $A_{0,0} \cup B_{1,0} \cup C_1$ - which, upon closer inspection, is just $A_0 \cup B_1 \cup C_2$.

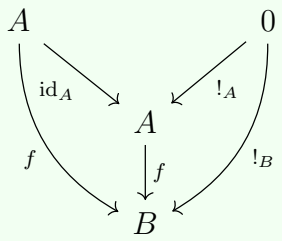
The result now follows trivially.

(c) This is proven by showing that A satisfies the universal property defining $A \sqcup 0$ - and is, therefore, isomorphic to it.

First, notice that there are functions $\mathrm{id}_A : A \to A$ and $!_A : 0 \to A$.

Now, let B be a set with functions $f : A \to B$ and $g : 0 \to B$. Since 0 is initial, $g = !_B$ the unique function from 0 to B. So it's easy to see that $f : A \to B$ makes it so that the

following diagram commutes



so A satisfies the universal property defining $A \sqcup 0$ and is, therefore, isomorphic to it, just as stated.

An alternative proof could be given:

Since $A \sqcup 0 \cong A_0 \cup 0_1$, we can use lemma 1.3.4.8(c) to see that 0_1, which is just fancy notation for $0 \times \{1\}$, is just 0 (since it is 0 times a set). So $A \sqcup 0 \cong A_0 \cup 0$, and we can now use proposition 1.2.2.13(3) and see that $A_0 \cup 0 = A_0$, which is clearly isomorphic to A. This shows that $A \sqcup 0 \cong A$, as stated.

This proves the result. □

If you think about it, there's no immediate reason why these three statements should be true. Yet, the fact that they are true makes it way more reasonable for our naming it the "coproduct" of two sets: Because it looks just like number addition.

The next subsection will be all about that: How to deal with sets as numbers, and numbers as sets.

1.3.6 Everything is a set!

The motivation for this section couldn't be simpler and more direct: We want to give a proper, formal definition of *numbers*.

The reasoning for this is simple: What is a number? What is the number 2, for instance, and why does it not equal, say, 4?

Through the history of mathematics, this has been tried to solve in many ways. For instance, Bertrand Russell (the same one from Russell's Paradox) proposed that the number 2, for instance, should be defined as being the collection of all collections with two elements. Of course, the way I put it here isn't very proper, or formal, but the general ideal is there: You take all collections in the universe, and group them together into mini-collections following a simple rule: Two collections belong in the same mini-collection if, and only if, they have the same amount of elements. And then you say that a *number* is each one of those mini-collections.

Now, we can agree or disagree that this is a good definition, but it sure as hell is a very interesting definition, at least. Think about it - if you go to any dictionary, it'll say that a number is a descriptive of quantity, and a quantity is the number of elements - which is a circular definition. Russell managed to sidestep that problem using the clever idea of grouping all collections with the same amount of elements together.

But then, we can ask: How can you count elements in a set without using numbers? Or, maybe a simpler problem: How to know if two sets have the same amount of elements without using numbers?

What Russell proposed as a solution to that problem is precisely what we're going to use: Bijections.

For Russell, two sets would be said to have the same amount of elements if you could make pairings of elements on both sets such that every element on the first set had a unique match on the second set, and that there was no elements left on either set after this process.

This idea is so simple, yet so brilliant, that even though it's not what we usually think of as numbers today, as mathematicians, it's still a very rich idea on the concept of numbers - so much so that it has been used as an alternative to the classic teaching method for introducing the concept of numbers to small children.

What we're going to do, however, is a tad different. We're going to use the so called **Peano's Axioms**, which is a list of axioms (which will be three in this text) that is widely accepted in mathematics as being the "best formalization of the set of natural numbers".

One of the reasons, and I could argue the *best* reason, why it is considered as such, comes from the fact that it is *categorical*, that is, it admits a unique model (up to isomorphisms, of course).

Basically, what that means is: Suppose I give you a list of axioms. What guarantee is there that there is something that is able to satisfy all those axioms at the same time? For instance, the list

- All of its elements are sets;
- Every set is in there;
- There are no sets in there.

admits no *model* - that is, there's nothing in the universe that satisfies all these three properties at the same time (you can actually take any two of these and it would work, but with all three at once, it's impossible).

Hence the name model - it's a way of *modelling* your list of conditions - taking it from something abstract and *modelling* it as something concrete.

Now, it's a known fact in mathematical logic that Peano's Axioms (which we're going to present) do admit a model (which we're going to construct), and that it is the *only* model (up to set isomorphism) for those axioms.

Without further ado, then, here we go.

Definition 1.3.6.1 (Peano's Axioms)**.** *Let N be a set satisfying*

(PA1) There is a point $0 \in N$ called zero*.*

(PA2) There is an injective function $s : N \to N$ called successor *such that $0 \notin \mathrm{Im}(s)$;*

(PA3) If $X \subseteq N$ is a set containing 0 such that for every $x \in X$ we have that $s(x) \in X$, then $X = N$.

Then N is called the ***set of natural numbers****.*

Now, there is no way, at first glance, to tell if these axioms do, or do not, induce a paradox.

The best way to do that is to produce a model for it - for if a list of axioms can be modelled, then it is consistent (i.e., it has no paradoxes).

So, we're going to do just that: Produce a model for it.

Definition 1.3.6.2 (Zero)**.** *We'll call the set $\varnothing$ the* ***zero set*** *and denote it as 0.*

Definition 1.3.6.3 (Successor)**.** *For any set X we define it's* ***successor*** *to be the set $s(X) := X \cup \{X\}$.*

Now, let's see what is this successor all about.

Example(s)

Let $A = \{a, b\}$. What is $s(A)$? Well, by definition, $s(A) = A \cup \{A\} = \{a, b, \{a, b\}\}$.
Similarly, if $B = \{1, 2, 3, 4\}$, then $s(B) = \{1, 2, 3, 4, \{1, 2, 3, 4\}\}$.
Notice that in both of these cases, the successor of a set has exactly one more element than the set itself - just like in the naturals, the successor of n is precisely $n + 1$.

What is then $s(0)$? Well, once again, by definition, $s(0) = 0 \cup \{0\}$, but we know that $0 \cup A = A$ for any set A! So $s(0) = \{0\}$ - a singleton!
Let's call $1 := s(0) = \{0\}$.
Now we can ask what is $s(1)$, and it will be $1 \cup \{1\} = \{0\} \cup \{1\} = \{0, 1\}$.
Let's call $2 := s(1) = \{0, 1\}$...
You can see where we're going with this, right?

Definition 1.3.6.4. *Let S denote any set containing 0 and closed under taking successors (that is, $x \in S$ implies $s(x) \in S$). We define $\mathbb{N}$ to be the set given by*

$$\mathbb{N} := \bigcap_{N \in \mathcal{N}} N$$

where $\mathcal{N} := \{X \in \mathcal{P}(S) \mid x \in X \implies s(x) \in X\}$.

In some sense, $\mathbb{N}$, as defined above, is "the smallest set containing zero and all its successors". We can finally model Peano's Axioms:

Theorem 1.3.6.5. *The set $\mathbb{N}$ satisfies all of (PA1), (PA2) and (PA3).*

Proof

By definition, $\mathbb{N}$ satisfies (PA1) and (PA2).
To prove (PA3), first notice that if we take $X \subseteq \mathbb{N}$ containing 0 and closed by successors, then, by definition of $\mathbb{N}$, we see that $\mathbb{N} \subseteq X$ (since $\mathbb{N}$ is the smallest set with those properties). Since we have both $X \subseteq \mathbb{N}$ and $\mathbb{N} \subseteq X$, we can conclude that $X = \mathbb{N}$, and so (PA3) holds.
This shows that $\mathbb{N}$ is a model for Peano's Axioms, as we wanted. □

With this, we see that Peano's Axioms are consistent (i.e., don't derive any logical contradictions) and, therefore, we can work with them.

So, from now on, every natural number n should be thought of as the set $\{0, 1, 2, 3, \cdots, n-1\}$.

Since, however, our main goal here is to get to Linear Algebra, we're not going to spend time defining the addition, subtraction, multiplication and division operations on $\mathbb{N}$, and we're going to assume that the reader is sufficiently familiar with them.

We're also not going to construct the integers, rationals or real numbers, not because it would be unfeasible, but because it doesn't suit the text.

Now that we have numbers, we can finally begin our study on cardinality, which will be central to our studies when we eventually get to linear algebra.

Definition 1.3.6.6. *Let X be a set. We say that X is **finite** if there is some $n \in \mathbb{N}$ such that $X \cong n$. In this case, we say that X **has** n **elements**, and denote it with the symbol $\#X = n$.*

Example(s)

Consider $A = \{a, b\}$. Then A is finite, because $A \cong 2$. To see that, consider the function $f : A \to 2$ taking $a \mapsto 0$ and $b \mapsto 1$. It clearly is bijective, by definition, so it is an isomorphism.

Conversely, the set $\mathbb{N}$ isn't finite. To see that, suppose it was - that is, there is some $n \in \mathbb{N}$ such that $n \cong \mathbb{N}$. This would imply that n satisfies Peano's Axioms - in particular, n would be closed by successors.
But $n = \{0, 1, 2, 3, \cdots, n-1\}$, and, clearly, for all $x \in n$, aside from $x = n - 1$, we see that

$s(x) \in n$. But if $n \cong \mathbb{N}$, we would have to have that $s(n-1) \in n$. But $s(n-1) = n$. So we would be saying that $n \in n$, which, by our previous discussion on Russell's Paradox, **cannot** happen.
Therefore, if $\mathbb{N} \cong n$, for some $n \in \mathbb{N}$, we would get a paradox akin to Russell's Paradox. Since we don't want that, we can't have $\mathbb{N} \cong n$, no matter which $n \in \mathbb{N}$ we choose.
It follows that $\mathbb{N}$ is not finite.

Proposition 1.3.6.7. *For all $n \in \mathbb{N}$, if $f : n \to n$ is injective, then it is automatically surjective.*

Proof

If $n = 0$, it is true (since the only function between 0 and 0 is the identity, which is a bijection).
Suppose it is true for some $n \in \mathbb{N}$. We will prove that it holds true also for $n+1$.
Let $f : n+1 \to n+1$ be an injective function, and consider $f^{-1}(n)$.

- If there is no $x \in n$ such that $f(x) = n$, we can define $f' : n \to n$ by putting $f'(x) := f(x)$ and see that f' is injective:

  If $f'(x) = f'(y)$ for some $x, y \in n$, then, by definition, $f(x) = f'(x) = f'(y) = f(y)$ and therefore $f(x) = f(y)$, but since f is injective, we have that this implies $x = y$.

  Now, we have that $f' : n \to n$ is injective and, therefore, by our assumption that the result holds for any injection from n to n, we see that f' is surjective.

  Now we ask: Can $f(n)$ be in n? Well, if it was, there'd be some $m \in n$ such that $f'(m) = f(n)$, since f' is injective. But $f'(m) = f(m)$, by definition of f'. So we would have $f(m) = f'(m) = f(n)$ which would imply, since f is injective by hypothesis, that $m = n$ - which is absurd, because $m \in n$ and we've shown many times before that $X \notin X$ for any set X.

  This means that $f(n) \notin n$ and, therefore, $f(n)$ must be the only element left in $n+1$ - that is n, from which it follows that f is surjective.

- If, however, there is some $x \in n$ such that $f(x) = n$, then we know that this x is unique since f is injective. Let's call this element k, so $f(k) = n$. This allows us, then, to define a new function $g : n+1 \to n+1$ as such: $g(k) := f(n)$, $g(n) := f(k)$, $g(m) := f(m)$ for all $m \in n$.

  Now g is an injection for which there's no $x \in n$ such that $g(x) = n$ (by construction). But now g is just as in the previous case, so it must be surjective. But then, f must also be surjective, since they have the same images.

So we've shown that the set of all n such that the proposition holds is a subset of $\mathbb{N}$ which contains 0 and is closed under taking successors. It follows from (PA3) that this set is the whole $\mathbb{N}$, so the proposition follows. □

Proposition 1.3.6.8. *For all $n \in \mathbb{N}$, if $f : n \to n$ is surjective, then it is automatically injective.*

Proof

If $f : n \to n$ is surjective, it has a right-inverse $g : n \to n$ such that $\mathrm{id}_n = f \circ g$.
Since id_n is an isomorphism, we see that g is injective. But now, the preceding proposition tells us that g is also an isomorphism.
Finally, lemma 1.3.2.20(f) tells us that since $\mathrm{id}_n = f \circ g$ and both id_n and g are iso, we have that f is also iso - and hence, injective.
This finishes the proof. □

Corollary 1.3.6.9. *For any $f : X \to X$ with X finite, the following are equivalent:*

1. *f is injective;*
2. *f is surjective;*
3. *f is bijective.*

Corollary 1.3.6.10. *For all $n, m \in \mathbb{N}$, we have that $n \in m$ if, and only if, there's an injection $f : n \to m$ which is not a surjection.*

Proof

If $n \in m$, the function $i_n : n \to m$ defined by $i_n(x) := x$ for all $x \in n$ is clearly injective (by definition), but it is not surjective: $n \in m$, but $n \notin \mathrm{Im}(f)$.

Conversely, assume that $n \notin m$. Then either $m \in n$ or $n = m$ (by definition of $\mathbb{N}$).

- If $m \in n$, then, by the first case, we see that there's an injection $g : m \to n$ that is not surjective.

  Now, if there was an injection $f : n \to m$, then their composite $g \circ f : n \to n$ would be an injection too and so, by the preceding lemma, $g \circ f$ would be a bijection, which, by lemma 1.3.2.20, implies that g would be a surjection, which is a contradiction.

  This shows that there cannot be such an f.

- If $m = n$, then every injection is surjective, by the preceding proposition.

This finishes the proof. □

Corollary 1.3.6.11. *For all $n, m \in \mathbb{N}$, we have that $n \in m$ if, and only if, there's a surjection $f : m \to n$ which is not an injection.*

Lemma 1.3.6.12. *For any $n \in \mathbb{N}$, we have that $\mathbb{N} \cong \mathbb{N} \setminus \{n\}$.*

Proof

Consider the function:

$$f : \mathbb{N} \to \mathbb{N} \setminus \{n\}$$

$$f(x) := \begin{cases} x, & \text{if } x \in n \\ x+1, & \text{otherwise.} \end{cases}$$

- f is surjective:

  Take any $x \in \mathbb{N} \setminus \{n\}$. If $x \in n$, then clearly $x = f(x)$. If $x \notin n$, then x is the successor of some y, and, clearly, $x = f(y)$.

  Either way, we see that $x \in \mathrm{Im}(f)$, so f is surjective.

- f is injective:

  Take $x, y \in \mathbb{N}$ such that $f(x) = f(y)$.

  If $f(x) \in n$, then, by definition, $x \in n$ and, therefore, $x = f(x) = f(y) = y$.

  If $f(x) \notin n$, then, by definition, $x \notin n$ and, therefore, $x + 1 = f(x) = f(y) = y + 1$ and since s is injective, by (PA2), we have that $s(x) = x + 1 = y + 1 = s(y)$ implies $x = y$.

  These two together show that f is indeed injective.

Since f is injective and surjective, it is bijective, as we wished to show. □

Proposition 1.3.6.13. *A set is finite if, and only if, there's no surjection from it to* $\mathbb{N}$.

Proof

Let X be finite. For all intents and purposes we can consider that, up to isomorphism, $X = n$, for some $n \in \mathbb{N}$.

- If $n = 0$, there's clearly no surjection from n to $\mathbb{N}$.
- Assume the result is true for some $n \in \mathbb{N}$ - that is, there's no surjection from n to $\mathbb{N}$. We're gonna prove the result for $n + 1$.

  Let $f : n+1 \to \mathbb{N}$ be a surjection. In that case, we get a surjection $f' : n \to \mathbb{N} \setminus \{f(n)\}$ by putting $f'(x) := f(x)$ for all $x \in n$.

  But the preceding lemma tells us that $\mathbb{N} \cong \mathbb{N} \setminus \{f(n)\}$, so f' is a surjection from n to $\mathbb{N}$ - which is absurd, by hypothesis.

  It follows then that f cannot be a surjection, just as we wished to show.

We have just shown that "the set of all $n \in \mathbb{N}$ for which there's no surjection to $\mathbb{N}$" contains 0 and is closed by successors, so, by (AP3), it must be $\mathbb{N}$ itself.

Conversely, suppose X is not finite, that is, for all $n \in \mathbb{N}$ we have $X \ncong n$.

Now, for each $Y \subseteq X$, choose an element $x_Y \in Y$.
We define $f : \mathbb{N} \to X$ by putting

$$f(n) := \begin{cases} x_X, & \text{if } n = 0; \\ x_{X \setminus \{f(0), f(1), \cdots, f(n-1)\}}, & \text{otherwise.} \end{cases}$$

Clearly, for all n, we have that $X \setminus \{f(0), f(1), \cdots, f(n-1)\} \neq \varnothing$ because, otherwise, X would be finite.
This function is then clearly injective by definition (since, for each $n \in \mathbb{N}$, $f(n)$ is chosen from a set that doesn't contain any of $f(0)$ up to $f(n-1)$). Therefore, it has a left-inverse $g : X \to \mathbb{N}$ such that $\mathrm{id}_{\mathbb{N}} = g \circ f$.
But $\mathrm{id}_{\mathbb{N}}$ is clearly a bijection - and hence a surjection. It follows that g is also a surjection from X to $\mathbb{N}$, just as we wanted.

This ends the proof. □

Corollary 1.3.6.14. *A set is finite if, and only if, there's no injection from* $\mathbb{N}$ *to it.*

Definition 1.3.6.15. *A set is* ***infinite*** *if it is not finite.*

Corollary 1.3.6.16. *Every infinite set contains a subset that is isomorphic to* $\mathbb{N}$*.*

1.3.7 Relations, order and quotients

The next sections should be way less technical, so that's a plus.

It would be great to have a way to describe when two sets are related in mathematical terms.

Example(s)

Let $A = \{1, 2\}$ and $B = \{2, 4\}$. Then all elements of B are precisely the double of some element of A. So A and B are related by the following relation: "B contains the doubles of the elements of A".
But if you think about it, there are other relations you could make using those two sets: "B contains one of the elements of A", "B contains the quadruple of an element of A", "B contains some element of A added with 3", "the elements of B are greater than the elements of A".
You can literally make as many relations as you want to.

Definition 1.3.7.1. *Let A and B be sets. We define the* ***set of (binary) relations between*** A ***and*** B *to be the set $\mathcal{P}(A \times B)$.*

In other words, a ***(binary) relation between*** A ***and*** B *is just a subset of $A \times B$.*

Example(s)

Continuing the example above, we have the following relations, in order:

- $R_1 := \{(1, 2), (2, 4)\} \subseteq A \times B$;
- $R_2 := \{(2, 2)\} \subseteq A \times B$;
- $R_3 := \{(1, 4)\} \subseteq A \times B$;
- $R_4 := \{(1, 4)\} \subseteq A \times B$;
- $R_5 := \{(1, 2), (1, 4), (2, 4)\} \subseteq A \times B$.

And we can see that relations R_3 and R_4 are the same relation, just phrased differently.

Definition 1.3.7.2. *Given a relation $R \subseteq A \times B$, we say that $a \in A$* ***is related to*** $b \in B$ ***via*** R *if $(a, b) \in R$.*

In symbols, we write that as $a \,\mathrm{R}\, b$.

Example(s)

Still on that example, take for instance R_1. Then we have $1 \,\mathrm{R}_1\, 2$ (this means "2 is the double of 1", or "1 is related to 2 via R_1") and $2 \,\mathrm{R}_1\, 4$.
Similarly, we have $1 \,\mathrm{R}_5\, 2$ (this means "5 is greater than 1", or "1 is related to 5 via R_5"), $1 \,\mathrm{R}_5\, 4$ and $2 \,\mathrm{R}_5\, 4$ on R_5.

Proposition 1.3.7.3. *A function $f : A \to B$ is just a relation $R := \{(a, b) \in A \times B \mid b = f(a)\}$.*

Proposition 1.3.7.4. *A relation $R \subseteq A \times B$ is a function if, and only if, for each $a \in A$ there's a unique $b \in B$ such that $a \,\mathrm{R}\, b$.*

These two propositions follow trivially from the definition of function and relation.

Example(s)

Let $\mathbb{N}$ be the set of natural numbers. Remember that this set has a natural order $\leq$ given by "$n \leq m$ if, and only if, $n \in m$ or $n = m$", which is just the usual order we're used to: $0 \leq 1 \leq 2 \leq 3 \leq 4 \leq \cdots$.
Notice that this order is a *relation* on $\mathbb{N} \times \mathbb{N}$: We can think of the symbol $n \leq m$ as meaning "n is related to m via $\leq$".
In other words, we can think of $\leq$ as being the set

$$\leq = \{(0,0), (0,1), (0,2), \cdots, (1,1), (1,2), (1,3), \cdots, (2,2), (2,3), \cdots\}$$

or equivalently, but more specifically, the set

$$\leq = \{(n, m) \in \mathbb{N} \times \mathbb{N} \mid n \in m \text{ or } n = m\}.$$

Now this relation is special: You can check that it satisfies three special conditions:

- (Reflexivity) For all $n \in \mathbb{N}$ we have that $n \leq n$;
- (Antisymmetry) If both of $n \leq m$ and $m \leq n$ hold, for some $n, m \in \mathbb{N}$, then $n = m$;
- (Transitivity) If $n \leq m$ and $m \leq l$, for some $n, m, l \in \mathbb{N}$, then $n \leq l$.

Let's check:

- It is clearly reflexive, since, by definition, $n \leq m$ includes the case $n = m$;
- Assume $n \leq m$ and $m \leq n$ both hold at once, for some $n, m \in \mathbb{N}$. Now, we have four cases to consider:

  1. $n \in m$ and $m \in n$.

     This case cannot happen, because we would have injections $f : n \to m$ and $g : m \to n$, whose compositions $f \circ g : m \to m$ and $g \circ f : n \to n$ would be injections from m to itself, and from n to itself, but we've already seen that those are always bijective.

     Hence, we'd get that both g and f are also bijective, which is absurd, because we'd be able to show that $n = m$ while $n \in m$, and we know this cannot be true (a set cannot be an element of itself);

  2. $n \in m$ and $n = m$.

     This case also cannot happen, by what we've just said;

3. $n = m$ and $m \in n$.
   Same as the previous case;
4. $n = m$ and $m = n$.

Since the only case that doesn't let to a contradiction is (4), we have that the only possible way that we can have both of $n \leq m$ and $m \leq n$ at once is if $n = m$. This shows that $\leq$ is antisymmetric.

- If $n \leq m$ and $m \leq l$, for some $n, m, l \in \mathbb{N}$, we have, once again, four cases to consider:

  1. $n \in m$ and $m \in l$.
     This is fine, since $m \in l$ implies $m \subseteq l$, by definition. This, coupled with $n \in m$ shows us that $n \in m \subseteq l$ and hence $n \in l$, so $n \leq l$;
  2. $n \in m$ and $m = l$.
     Again this is fine, since $n \in m = l$ implies $n \in l$, and hence $n \leq l$;
  3. $n = m$ and $m \in l$.
     Once again, fine: $n = m \in l$ implies $n \in l$, and thus $n \leq l$;
  4. $n = m$ and $m = l$.
     The simplest one: $n = m = l$, so $n = l$, and thus $n \leq l$

  In each of the four cases, the only logical conclusion is $n \leq l$, so we have that $\leq$ is transitive.

Definition 1.3.7.5. *Let A be any set with a relation $R \subseteq A \times A$. We say that R is a **partial order** if R is reflexive, antisymmetric and transitive.*

Now we will justify the terminology *partial*:

Example(s)

Let X be any set, and consider $\mathcal{P}(X)$ its power set.
We claim that the relation $Y \subseteq Z$ is a partial order on $\mathcal{P}(X)$.

- Clearly, for all $Y \in \mathcal{P}(X)$ we have that $Y \subseteq Y$, since Y is a set, so $\subseteq$ is reflexive;
- Assume that for some $Y, Z \in \mathcal{P}(X)$ we have both $Y \subseteq Z$ and $Z \subseteq Y$. Then, by definition of set equality, $Y = Z$, so $\subseteq$ is antisymmetric;
- Assume that for some $Y, Z, W \in \mathcal{P}(X)$ we have both $Y \subseteq Z$ and $Z \subseteq W$. Then every element of Y is an element of Z (by the first part) and every element of Z is an element of W (by the second part). This implies that every element of Y is an element of W, and, therefore, $Y \subseteq W$, and we see that $\subseteq$ is transitive.

This is all pretty standard.

Consider, then, $X = \{a, b, c\}$. Then $\mathcal{P}(X) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, X\}$.
We ask: is $\{b\}$ related to $\{a, b\}$ via $\subseteq$? It clearly is!
Now, is $\{b\}$ related to $\{a, c\}$ via $\subseteq$? It clearly **isn't**!
So contrary to the order $\leq$ in $\mathbb{N}$, the order $\subseteq$ in $\mathcal{P}(X)$ doesn't allow us to compare any two elements. This is why it's called a *partial* order.

Definition 1.3.7.6. *Let A be a set with a partial order $\leq$. We call the ordered pair $(A, \leq)$ a **poset** (**p**artially-**o**rdered **set**).*

Sometimes, we consider the order to be implicit and simply say that A is a poset.

Definition 1.3.7.7. *Let $(A, \leq)$ be a poset. We say that $\leq$ is a **total order** or a **linear order** if for all $a, b \in A$ we have that either or both of $a \leq b$ and $b \leq a$ hold.*

Lemma 1.3.7.8. *The set $\mathbb{N}$ with the usual order is a totally-ordered set.*

Definition 1.3.7.9. *Let $(A, \leq)$ be a poset and $X \subseteq A$. We define the following symbols:*

- *We say that x **is the maximum of** X if every other element of X is smaller than x, that is,*
$$\max X := \{x \in X \mid \forall y \in X, y \leq x\};$$
- *We say that x **is the minimum of** X if x is smaller than every other element of X, that is,*
$$\min X := \{x \in X \mid \forall y \in X, x \leq y\};$$
- *We say that x **is a maximal element of** X if x is not smaller than any other element of X;*
- *We say that x **is a minimal element of** X if there is no element of X which is smaller than x.*

Example(s)

Let X be as in the previous example, and consider $\mathcal{P}(X)$ ordered by set inclusion.
Then $\mathcal{P}(X)$ has a maximum: X; and a minimum: $\varnothing$.
Notice that they are also respectively maximal and minimal elements.
Consider however $\mathcal{P}(X) \setminus \{\varnothing, X\}$, that is, $\{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}$.
Does this set have a minimum or a maximum? Well, suppose $\{a\}$ was a minimum. Then it'd have to be smaller than everyone - including $\{b, c\}$. But $\{a\} \not\subseteq \{b, c\}$, so $\{a\}$ cannot be a minimum (and $\{b, c\}$ cannot be a maximum). Similarly, we can show that this set doesn't have a minimum or a maximum.
However, it *does* have minimals and maximals: For instance, $\{a\}$ is minimal - since there's no element that's smaller than $\{a\}$. Similarly, $\{b, c\}$ is maximal, since there's no element that's greater than it.

Lemma 1.3.7.10. *If X is a poset with maximum (resp. minimum), then that maximum (resp. minimum) is also maximal (resp. minimal).*

Proof

Let $x = \max X$. Then for all other $y \in X$ we have $y \leq x$. In particular, if $y \neq x$ we have that $x \not\leq y$, so x is a maximal element of X. □

Definition 1.3.7.11. *Given a set A, an order $\leq$ in A is said to be a **well-order** if for every non-empty $X \subseteq A$ we have that $\min X$ exists.*

*In this case, we say that A is a **well-ordered** set.*

Theorem 1.3.7.12 (Well-ordering Principle)**.** *The set $\mathbb{N}$ with the usual order is well-ordered.*

Proof

Assume there is some set $B \subset \mathbb{N}$ which doesn't have a minimum. Consider $B' := \mathbb{N} \setminus B$.

- $0 \in B'$.

  This is trivial, cause, otherwise $0 \in B$, but then $0 = \min B$. Since we're assuming B doesn't have a minimum, then $0 \notin B$ and therefore $0 \in B'$.

- Assume $m \in B'$ for all $m \leq n$. We're going to show that $n + 1 \in B'$.

  If $n + 1$ was in B, then, since every $m \leq n$ is not in B, by assumption, we'd have that $n+1$ is a minimum of B. But B doesn't have a minimum, so $n+1 \notin B$ and, therefore, $n + 1 \in B'$.

But then B' is a subset of $\mathbb{N}$ which contains 0 and all its successors - therefore it must be $\mathbb{N}$ (by (PA3)).
But $B' = \mathbb{N} \setminus B$. So $B = \mathbb{N} \setminus B' = \mathbb{N} \setminus \mathbb{N} = \varnothing$ and, therefore, the only subset of $\mathbb{N}$ which doesn't have a minimum is $\varnothing$ - this proves that $\mathbb{N}$ is well-ordered, as we wished to show. □

Let us now present a new kind of relation:

Proposition 1.3.7.13. *For any set X, "$x \sim y$ if, and only if, $x = y$" is a relation.*

Example(s)

Take X any set, and consider the relation above: $x \sim y$ if, and only if, $x = y$.

- First, we see that $x \sim x$ for all $x \in X$, since every element equals itself.
- Now, assume that $x \sim y$ for some $x, y \in X$. But this means that $x = y$, and therefore $y = x$, so $y \sim x$.
- Finally, if $x \sim y$ and $y \sim z$ for some $x, y, z \in X$, then this means that $x = y$ and $y = z$, so $x = y = z$ implies $x = z$. This means that $x \sim z$.

The first and the last items above say that $\sim$ is reflexive and transitive, respectively. The

second one, however, is a new property - one that orders, in general, don't satisfy. It's called *symmetry.*

Definition 1.3.7.14. *Given a set X with a relation R, we say that R **is symmetric** if $x \,\mathrm{R}\, y$ if, and only if, $y \,\mathrm{R}\, x$.*

So, in other words, a relation is symmetric if it doesn't care about order.

Example(s)

The relation "a is a son of b" is **not** symmetric: a being b's son doesn't imply that b is a's son.
However, the relation "a is in the same class as b" is symmetric: a being in the same class as b is the same as b being in the same class as a.
Similarly, the relation "X has the same amount of elements as Y" is symmetric.

Definition 1.3.7.15. *Given a set X, a relation $\sim$ is said to be an **equivalence** if it is reflexive, symmetric and transitive.*

Lemma 1.3.7.16. *In any set X, element equality is an equivalence relation on X.*

Lemma 1.3.7.17. *Given any set X, the relation "$Y \sim Z$ if, and only if, $Y \cong Z$" is an equivalence relation on $\mathcal{P}(X)$.*

Proof

- $\sim$ is reflexive:

  Clearly, any $Y \in \mathcal{P}(X)$ is isomorphic to itself via the identity function $\mathrm{id}_Y : Y \to Y$;

- $\sim$ is symmetric:

  Let $Y \sim Z$ - that is, there is a bijection $f : Y \to Z$. But we know that bijections are isomorphisms, so there is an inverse $f^{-1} : Z \to Y$ which is also an isomorphism - and hence a bijection. So there is a bijection from Z to Y, and so $Z \sim Y$;

- $\sim$ is transitive:

  This follows trivially from the fact that composition of isomorphisms is, again, an isomorphism.

Since $\sim$ is reflexive, symmetric and transitive it is, by definition, an equivalence.
This ends the proof. □

This is the final, and best reason why we should think of isomorphisms as being the true idea of set equality: Because it is an equivalence relation that tells us much more interesting information than set equality does.

Finally, to end this section we will use relations to build new sets.

Example(s)

Let $\mathbb{Z}$ be the set of integers, and define the following relation:

$$x \sim y \iff \exists k \in \mathbb{Z} \text{ such that } x - y = 2k.$$

Let us show that this is an equivalence:

- $\sim$ is reflexive:

  For all $x \in \mathbb{Z}$, $x - x = 0 = 2 \cdot 0$ and $0 \in \mathbb{Z}$, so $x \sim x$ and $\sim$ is reflexive.

- $\sim$ is symmetric:

  Let $x \sim y$ for some $x, y \in \mathbb{Z}$. Then there exists some $k \in \mathbb{Z}$ such that $x - y = 2k$, by definition of $\sim$. But then,

  $$y - x = -(x - y) = -(2k) = 2(-k),$$

  and since $k \in \mathbb{Z}$ we get that $-k \in \mathbb{Z}$, so $y \sim x$ and $\sim$ is symmetric.

- $\sim$ is transitive:

  Let $x \sim y$ and $y \sim z$, for some $x, y, z \in \mathbb{Z}$. Then, by definition of $\sim$, there exist some $k, l \in \mathbb{Z}$ such that $x - y = 2k$ and $y - z = 2l$.

  But then,

  $$x - z = x - y + y - z = 2k - 2l = 2(k - l)$$

  and since $k, l \in \mathbb{Z}$, we see that $k - l \in \mathbb{Z}$, so $x - z = 2(k - l)$ tells us that $x \sim z$, so $\sim$ is transitive.

So we see that $\sim$ is indeed an equivalence relation on $\mathbb{Z}$.
Now, this is our first weird equivalence. Let's see who is equivalent to whom in this new "equality":
For instance, who is equivalent to 1? Well, if $x \sim 1$, then $x - 1 = 2k$ for some $k \in \mathbb{Z}$, and so $x = 2k + 1$. But this means that x is *odd* (this is the definition of an odd number: one more than a multiple of 2).
Conversely, if $x = 2l + 1$ for some $l \in \mathbb{Z}$ is an odd number, we see that $x - 1 = (2l + 1) - 1 = 2l$ and so $x \sim 1$.
So if we denote by $[1]$ the **class of every element in $\mathbb{Z}$ which is $\sim$-equivalent to 1**, we see that

$$[1] = \{x \in \mathbb{Z} \mid x \text{ is odd}\}.$$

But since we've proven that $\sim$ is an equivalence, if we take any other odd number - say, 3, we see that $[3] = [1]$, for the following reasoning:

- 3 is odd, so $3 \sim 1$.
- If $x \sim 3$, for some x, then since $\sim$ is transitive this implies $x \sim 1$.

- This, in turn, implies that x is odd.

Conversely - if x is odd, then $x \sim 1$, and since $3 \sim 1$ and $\sim$ is transitive, we get $x \sim 3$.
So everyone who is $\sim$-equivalent to 3 is also $\sim$-equivalent to 1, and vice-versa.
This means that we don't need to check any more odd numbers to fully understand $\sim$.

Let us check then the even numbers. For instance, 0. What's $[0]$ - that is, who is $\sim$-equivalent to 0?
Well, if $x \sim 0$, then $x - 0 = 2k$ for some $k \in \mathbb{Z}$, but $x - 0 = x$, so we get that $x = 2k$. This means that x is even.
Conversely, if x is even, then $x = 2l$ for some $l \in \mathbb{Z}$ (this is the definition of an even number: it is a multiple of 2). So $x - 0$ is just $2l - 0$ which is simply $2l$, so all even numbers are $\sim$-equivalent to 0.
This means that

$$[0] = \{x \in \mathbb{Z} \mid x \text{ is even}\}.$$

And doing the same reasoning we did above, we can see that $[2k] = [0]$ for any $l \in \mathbb{Z}$.
Therefore, under this new sense of "equality", there are only two different elements in $\mathbb{Z}$: 0 and 1. Everyone else is either $\sim$-equal to 0 or 1.

Definition 1.3.7.18. *Given any set X with an equivalence $\sim$, we define the **quotient set of** X **over** $\sim$ to be the set $X/\sim$ of all equivalence classes of elements of X under the equivalence $\sim$.*

Example(s)

Continuing the previous example, $\mathbb{Z}/\sim$ is just the set $\{[0], [1]\}$.

To finish this section, we'll prove two important results.

Theorem 1.3.7.19 (Cantor-Bernstein-Schroeder)**.** *If $f : A \to B$ and $g : B \to A$ are both injective, then there's an isomorphism $\phi : A \to B$.*

Proof

For each $x \in A \sqcup B$ define:

- $s_0^x := x$;
- $s_{n+1}^x := (f \sqcup g)(s_n^x)$;
- $s_{-(n+1)}^x := (f \sqcup g)^{-1}(s_{-n}^x)$, if s_{-n}^x is well-defined.

Note that for each $x \in A \sqcup B$ we know nothing about s_n^x being well-defined for negative n.
So, for each $x \in A \sqcup B$, we can create a new set:

$$S^x := \{y \in A \sqcup B \mid y = s_n^x \text{for some } n \in \mathbb{Z}\}$$

that is, S^x is the collection of all s_n^x that are well-defined.
Now, it's easy to see that every element of $A \sqcup B$ is in some S^x - so we get

$$A \sqcup B = \bigcup_{x \in A \sqcup B} S^x$$

(we're not claiming that this union is disjoint, there could be $x \neq y$ but such that $S^x = S^y$. We're not excluding that possibility).
Indeed, assume that's the case - that there's some $x \neq y$ such that $s_n^x = s_m^y$, for some $n, m \in \mathbb{Z}$. Then, clearly, $s_{n+l}^x = s_{m+l}^y$ for all $l \in \mathbb{Z}$, so $S^y = S^x$.
So now the proof consists of showing not that $A \cong B$, but that for each $x \in A \sqcup B$, we have $(S^x \cap A) \cong (S^x \cap B)$ - that is, the parts of A and B inside each of S^x are isomorphic. This suffices, because we can then extend this isomorphism to the whole A and B.

- If s_n^x is well-defined for all $n \in \mathbb{Z}$, then this means that $f' : S^x \cap A \to S^x \cap B$ is surjective, and therefore a bijection (since it's already injective).
- If $s_n^x \in A$ is the smallest well-defined, then, once again, $f' : S^x \cap A \to S^x \cap B$ is surjective, and therefore a bijection (since it's already injective).
- If $s_n^x \in B$ is the smallest well-defined, then $g' : S^x \cap B \to S^x \cap A$ is surjective, and, therefore, a bijection (since it's already injective).

Either way, we see that for all $x \in A \sqcup B$ we have $(S^x \cap A) \cong (S^x \cap B)$, so $A \cong B$, and the result follows. □

Theorem 1.3.7.20. *Let X be a set. Then the relation "$Y \leq Z$ if, and only if, there is an injection $f : Y \to Z$" is an order in $\mathcal{P}(X)/\cong$.*

Proof

- $\leq$ is reflexive:

  Let $Y \in \mathcal{P}(X)$. Then the identity map is injective, so $Y \leq Y$, and $\leq$ is reflexive.

- $\leq$ is antisymmetric:

  Let $Y \leq Z$ and $Z \leq Y$, for some $Y, Z \in \mathcal{P}(X)$. Then, by definition of $\leq$, there are injections $f : Y \to Z$ and $g : Z \to Y$. Now, by theorem 1.3.7.19, we see that $Y \cong Z$, and so $\leq$ is antisymmetric.

- $\leq$ is transitive:

  If $Y \leq Z$ and $Z \leq W$, for some $Y, Z, W \in \mathcal{P}(X)$, then there are injections $f : Y \to Z$ and $g : Z \to W$. But since composition of injections is, again, an injection, we see that $g \circ f : Y \to W$ is injective, and, therefore, $Y \leq W$, so $\leq$ is transitive.

Since $\leq$ is reflexive, antisymmetric and transitive, it is an order, by definition. This ends the proof. $\square$

Corollary 1.3.7.21. *Let X be a set. Then $Y \leq Z$ in $\mathcal{P}(X)/\cong$ if, and only if, there is a surjection $g : Z \to Y$.*

Proof

Follows trivially from the previous theorem and the fact that every injection has a left-inverse which is surjective. $\square$

We can finally proceed to the final section.

1.3.8 About sets of functions

Now, to finish up our brief talk on set theory, we're gonna talk about sets of functions.

Definition 1.3.8.1. *Let A and B be two sets. We denote the* ***set of functions from A to B*** *by* $\mathrm{Hom}(A, B)$.

Example(s)

Let $A = \{1, 2\}$ and $B = \{a, b, c\}$. Then, the only possible functions $A \to B$ are:

- $f_1 = \{(1, a), (2, a)\}$
- $f_2 = \{(1, a), (2, b)\}$
- $f_3 = \{(1, a), (2, c)\}$
- $f_4 = \{(1, b), (2, a)\}$
- $f_5 = \{(1, b), (2, b)\}$
- $f_6 = \{(1, b), (2, c)\}$
- $f_7 = \{(1, c), (2, a)\}$
- $f_8 = \{(1, c), (2, b)\}$
- $f_9 = \{(1, c), (2, c)\}$

so $\mathrm{Hom}(A, B) \cong 9$.
Notice that $A \cong 2$ and $B \cong 3$.

Analogously, the only possible functions $B \to A$ are:

- $g_1 = \{(a, 1), (b, 1), (c, 1)\}$
- $g_2 = \{(a, 1), (b, 1), (c, 2)\}$
- $g_3 = \{(a, 1), (b, 2), (c, 1)\}$
- $g_4 = \{(a, 1), (b, 2), (c, 2)\}$
- $g_5 = \{(a, 2), (b, 1), (c, 1)\}$
- $g_6 = \{(a, 2), (b, 1), (c, 2)\}$
- $g_7 = \{(a, 2), (b, 2), (c, 1)\}$
- $g_8 = \{(a, 2), (b, 2), (c, 2)\}$

so $\mathrm{Hom}(B, A) \cong 8$.
We can readily see that $\mathrm{Hom}(A, B) \cong \#B^{\#A}$ and $\mathrm{Hom}(B, A) \cong \#A^{\#B}$.

Lemma 1.3.8.2. *For any set X, we have that $\mathcal{P}(X) \cong \mathrm{Hom}(X, 2)$.*

Proof

Let $Y \in \mathcal{P}(X)$ and consider the function $f_Y : X \to 2$ defined by

$$f_Y(x) := \begin{cases} 1 & \text{if } x \in Y \\ 0, & \text{otherwise.} \end{cases}$$

This defines a unique function $f : \mathcal{P}(X) \to \mathrm{Hom}(X, 2)$ given by $f(Y) := f_Y$.

- f is injective:

  Let $Y, Z \in \mathcal{P}(X)$ be such that $f(Y) = f(Z)$. But this means that the functions $f_Y, f_Z : X \to 2$ are the same. But by definition of function equality, this means that $f_Y(x) = f_Z(x)$ for all $x \in X$.

In particular, $1 = f_Y(y)$. But this implies $1 = f_Y(y) = f_Z(y)$, so $y \in Z$ (by definition of f_Z). This shows that every element of Y is also an element of Z - that is, $Y \subseteq Z$.

Analogously, for all $z \in Z$ we have $f_Z(z) = 1$ and so $f_Y(z) = f_Z(z) = 1$ which implies that $z \in Y$ (by definition of f_Y), and hence every element of Z is also an element of Y - that is, $Z \subseteq Y$.

These two together imply that $Y = Z$ and so f is injective.

- f is surjective:

  Let $g : X \to 2$ be a function. Then, $g^{-1}(1) \subseteq X$. Call it X'. We claim that $f(X') = g$.

  Indeed, given any $x \in X$ we want to show that $f_{X'}(x) = g(x)$.

  If $x \in X'$, then:

  $$f_{X'}(x) = 1 = g(x)$$

  where the first equality holds by definition of $f_{X'}$ and the second equality holds by definition of X' (i.e., it's the inverse image of 1 under g - which means that if we apply g to X' we always get 1).

  If $x \notin X'$, the:

  $$f_{X'}(x) = 0 = g(x)$$

  where the two equalities hold by exactly the same reason as above.

  This shows that $f_{X'} = g$, and so $g \in \mathrm{Im}(f)$, which means that f is surjective.

Since f is both injective and surjective, we see that it is bijective - and hence an isomorphism. This ends the proof. □

Definition 1.3.8.3. *Given A and B, we're going to denote the set $\mathrm{Hom}(A, B)$ by B^A and call it the **exponential set**.*

Remark 1.3.8.4

This finally justifies the notation 2^X for the power set of X. It is, as we've shown, a set of functions: Any subset of X can be seen as a function which takes all its elements to 1, and everyone else to 0.

Theorem 1.3.8.5. *For any set X there is a bijection $\mathrm{Hom}(1, X) \cong X$.*

Proof

Define $f : \mathrm{Hom}(1, X) \to X$ like this: Given any $g : 1 \to X$, since 1 is a singleton, $g(0) \in X$ is also a singleton. So we define $f(g) := g(0) \in X$.

- f is injective:

Take two functions $g, h : 1 \to X$ such that $f(g) = f(h)$. But this means, by definition of f that $g(0) = h(0)$, and since $1 = \{0\}$ this means that g and h are equal for all elements of 1 - the definition of function equality.

This shows that $g = h$ and so f is injective.

- f is surjective:

  Take any element $x \in X$. Define the function $g_x : 1 \to X$ by putting $g_x(0) := x$. Then clearly, $f(g_x) = x$, by definition of both f and g_x.

  This shows that f is surjective.

Since f is both injective and surjective, it is a bijection, which ends the proof. □

Remark 1.3.8.6

This is one remarkable result: It tells us that if, somehow, we could define functions without knowing what elements are, we would be able to completely recover the elements of any set simply by looking at functions from 1 to that set.

Now we can do arithmetic with finite sets:

Lemma 1.3.8.7. *For any $n, m \in \mathbb{N}$ the following hold:*

- $n \sqcup m \cong n + m$;
- $n \times m \cong nm$;
- $\mathrm{Hom}(m, n) \cong n^m$.

Theorem 1.3.8.8 (Sets are cartesian-closed)**.** *Let X, Y, Z be sets. Then we have the following set isomorphism:*

$$\mathrm{Hom}(X \times Y, Z) \cong \mathrm{Hom}(X, \mathrm{Hom}(Y, Z)).$$

Proof

Fix, first and foremost, X, Y, Z to be any sets.
Now we define $f : \mathrm{Hom}(X \times Y, Z) \to \mathrm{Hom}(X, \mathrm{Hom}(Y, Z))$ as such: For any function $g : X \times Y \to Z$ we define $f(g)$ to be the function that takes any $x \in X$ to the function $(f(g))(x)$ that takes any $y \in Y$ to $((f(g))(x))(y) := g(x, y)$.
Define, now, $f' : \mathrm{Hom}(X, \mathrm{Hom}(Y, Z)) \to \mathrm{Hom}(X \times Y, Z)$ as such: For any function $g : X \to \mathrm{Hom}(Y, Z)$, we define $f'(g)$ to be the function that takes any $(x, y) \in X \times Y$ to $(f'(g))(x, y) := (g(x))(y)$.
We claim that $f' = f^{-1}$ and, thus, that f is an isomorphism.
Take any $g : X \to \mathrm{Hom}(Y, Z)$, any $x \in X$ and any $y \in Y$. We can see that

$$\begin{aligned}(((f\circ f')(g))(x))(y) &= (f(f'(g))(x))(y)\\ &= (f'(g))(x,y) && \text{(by definition of } f)\\ &= g(x)(y) && \text{(by definition of } f'),\end{aligned}$$

so, as functions, $((f\circ f')(g))(x) = g(x)$ (since they are equal for every $y\in Y$), and so $(f\circ f')(g) = g$ (since they are equal for every $x\in X$), and so $f\circ f' = \mathrm{id}_{\mathrm{Hom}(X,\mathrm{Hom}(Y,Z))}$ (since they are equal for every $g\in \mathrm{Hom}(X,\mathrm{Hom}(Y,Z))$).

Conversely, we have, for any $g: X\times Y\to Z$ and any $(x,y)\in X\times Y$ that

$$\begin{aligned}((f'\circ f)(g))(x,y) &= (f'(f(g)))(x,y)\\ &= ((f(g))(x))(y) && \text{(by definition of } f')\\ &= g(x,y) && \text{(by definition of } f)\end{aligned}$$

so, as functions, $(f'\circ f)(g) = g$ (since they are equal for every $(x,y)\in X\times Y$), and so $f'\circ f = \mathrm{id}_{\mathrm{Hom}(X\times Y,Z)}$ (since they are equal for every $g\in\mathrm{Hom}(X\times Y\to Z)$).

This shows that $f' = f^{-1}$ and hence f is an isomorphism.
The proves the result. □

Remark 1.3.8.9

This result is basically saying that there's a one-to-one correspondence between functions with two inputs and functions with one input whose input is another function with one input.

We will show, at some point, that an equivalent result also holds for linear transformations - and that will be of much use for us further ahead.

Proposition 1.3.8.10. *Any function $f: X\to Y$ induces unique functions from $\mathrm{Hom}(A,X)$ to $\mathrm{Hom}(A,Y)$ and from $\mathrm{Hom}(Y,A)$ to $\mathrm{Hom}(X,A)$ for any set A.*

Proof

Define $\phi_f: \mathrm{Hom}(A,X)\to\mathrm{Hom}(A,Y)$ by putting $\phi_f(g) := f\circ g$ for any $g\in\mathrm{Hom}(A,X)$ and $\psi^f: \mathrm{Hom}(Y,A)\to\mathrm{Hom}(X,A)$ by putting $\psi^f(h) := h\circ f$.
Clearly ϕ and ψ are functions, so the result follows. □

Definition 1.3.8.11. *Given any set A and any function $f: X\to Y$, the unique functions induced by f will be denoted by $\mathrm{Hom}(A,f): \mathrm{Hom}(A,X)\to\mathrm{Hom}(A,Y)$ and $\mathrm{Hom}(f,A): \mathrm{Hom}(Y,A)\to\mathrm{Hom}(X,A)$.*

Lemma 1.3.8.12. *If $X\xrightarrow{f} Y\xrightarrow{g} Z$, then, for any A we have that $\mathrm{Hom}(A,g\circ f) = \mathrm{Hom}(A,g)\circ\mathrm{Hom}(A,f)$ and $\mathrm{Hom}(g\circ f,A) = \mathrm{Hom}(f,A)\circ\mathrm{Hom}(g,A)$.*

Proof

Take any $h \in \mathrm{Hom}(A, X)$. Then:

$$\begin{aligned}\mathrm{Hom}(A, g \circ f)(h) &= (g \circ f) \circ h \\ &= g \circ (f \circ h) \\ &= \mathrm{Hom}(A, g)(f \circ h) \\ &= \mathrm{Hom}(A, g)(\mathrm{Hom}(A, f)(h)) = (\mathrm{Hom}(A, g) \circ \mathrm{Hom}(A, f))(h),\end{aligned}$$

and so $\mathrm{Hom}(A, g \circ f) = \mathrm{Hom}(A, g) \circ \mathrm{Hom}(A, f)$.
Similarly, given any $h \in \mathrm{Hom}(Z, A)$ we have:

$$\begin{aligned}\mathrm{Hom}(g \circ f, A)(h) &= h \circ (g \circ f) \\ &= (h \circ g) \circ f \\ &= \mathrm{Hom}(f, A)(h \circ g) \\ &= \mathrm{Hom}(f, A)(\mathrm{Hom}(g, A)(h)) = (\mathrm{Hom}(f, A) \circ \mathrm{Hom}(g, A))(h),\end{aligned}$$

and so $\mathrm{Hom}(g \circ f, A) = \mathrm{Hom}(f, A) \circ \mathrm{Hom}(g, A)$. □

Lemma 1.3.8.13. *For any sets A, B, we have that $\mathrm{Hom}(A, \mathrm{id}_B) = \mathrm{id}_{\mathrm{Hom}(A,B)} = \mathrm{Hom}(\mathrm{id}_A, B)$.*

Proof

This follows trivially by inspection:

$$\mathrm{Hom}(A, id_B)(g) = \mathrm{id}_B \circ g = g \circ \mathrm{id}_A = \mathrm{Hom}(\mathrm{id}_A, B)(g)$$

for all $g \in \mathrm{Hom}(A, B)$. □

Finally, the last result of this section - the Yoneda Lemma:

Definition 1.3.8.14. *Let X, Y be sets. A **natural transformation** $\phi : \mathrm{Hom}(-, X) \to \mathrm{Hom}(-, Y)$ is a family of functions $\phi := \{\phi_A : \mathrm{Hom}(A, X) \to \mathrm{Hom}(A, Y)\}_A$, where A ranges over all sets, such that*

$$\begin{array}{ccc} \mathrm{Hom}(B, X) & \xrightarrow{\phi_B} & \mathrm{Hom}(B, Y) \\ {\scriptstyle \mathrm{Hom}(g,X)}\downarrow & & \downarrow{\scriptstyle \mathrm{Hom}(g,Y)} \\ \mathrm{Hom}(A, X) & \xrightarrow[\phi_A]{} & \mathrm{Hom}(A, Y) \end{array}$$

commutes for all $g : A \to B$.

*If each ϕ_A in the family is an isomorphism, then we say that ϕ is a **natural isomorphism**.*

Theorem 1.3.8.15 (Yoneda's Lemma)**.** *Let X and Y be any sets. Then there is a bijective correspondence between natural transformations $\phi : \mathrm{Hom}(-, X) \to \mathrm{Hom}(-, Y)$, and functions $f : X \to Y$.*

Proof

Take $g : A \to B$ any function.
Since the diagram commutes, by hypothesis, $\mathrm{Hom}(g, Y) \circ \phi_B = \phi_A \circ \mathrm{Hom}(g, X)$ holds for any choice of A, B and g.
So if we take $g : A \to X$, by applying the equality above to id_X we see that

$$(\mathrm{Hom}(g, Y) \circ \phi_X)(\mathrm{id}_X) = \mathrm{Hom}(g, Y)(\phi_X(\mathrm{id}_X)) = (\phi_X(\mathrm{id}_X)) \circ g$$

and

$$(\phi_A \circ \mathrm{Hom}(g, X))(\mathrm{id}_X) = \phi_A(\mathrm{Hom}(g, X)(\mathrm{id}_X)) = \phi_A(\mathrm{id}_X \circ g) = \phi_A(g)$$

must be equal - in other words, $\phi_X(\mathrm{id}_X) \circ g = \phi_A(g)$.
But this shows us that $\phi_A : \mathrm{Hom}(A, X) \to \mathrm{Hom}(A, Y)$ is uniquely defined by $\phi_A = \mathrm{Hom}(A, \phi_X(\mathrm{id}_X))$.
In other words, the correspondence

$$\{\phi_A : \mathrm{Hom}(A, X) \to \mathrm{Hom}(A, Y)\}_A \mapsto (\phi_X(\mathrm{id}_X) : X \to Y)$$

$$(f : X \to Y) \mapsto \{\mathrm{Hom}(A, f) : \mathrm{Hom}(A, X) \to \mathrm{Hom}(A, Y)\}_A$$

is bijective.
This finishes the proof. □

Definition 1.3.8.16. *Let X, Y be sets. A **natural transformation** $\phi : \mathrm{Hom}(X, -) \to \mathrm{Hom}(Y, -)$ is a family of functions $\phi := \{\phi_A : \mathrm{Hom}(X, A) \to \mathrm{Hom}(Y, A)\}_A$, where A ranges over all sets, such that*

$$\begin{array}{ccc} \mathrm{Hom}(X, A) & \xrightarrow{\phi_A} & \mathrm{Hom}(Y, A) \\ {\scriptstyle \mathrm{Hom}(X,g)}\downarrow & & \downarrow{\scriptstyle \mathrm{Hom}(Y,g)} \\ \mathrm{Hom}(X, B) & \xrightarrow[\phi_B]{} & \mathrm{Hom}(Y, B) \end{array}$$

commutes for all $g : A \to B$.

*If each ϕ_A in the family is an isomorphism, then we say that ϕ is a **natural isomorphism**.*

Corollary 1.3.8.17 (Yoneda's Lemma)**.** *Let X and Y be any sets. Then there is a bijective correspondence between natural transformations $\psi : \mathrm{Hom}(X, -) \to \mathrm{Hom}(Y, -)$, and functions $f : Y \to X$.*

Proof

It's essentially the same proof as the theorem's, so it'll be left as an exercise to the reader. □

Corollary 1.3.8.18. *For any pair of sets X, Y the following are equivalent:*

(i.) There is a natural isomorphism $\phi : \mathrm{Hom}(-, X) \to \mathrm{Hom}(-, Y)$;

(ii.) There is a natural isomorphism $\psi : \mathrm{Hom}(Y, -) \to \mathrm{Hom}(X, -)$*;*

(iii.) $X \cong Y$*.*

Proof

We will only prove (i.) if, and only if, (iii.). The case (ii.) if, and only if, (iii.) is similar and will be left as an exercise to the reader.

Let $\phi := \{\phi_A : \mathrm{Hom}(A, X) \to \mathrm{Hom}(A, Y)\}_A$ be a natural isomorphism. Analogously, let $\phi^{-1} := \{\phi_A^{-1} : \mathrm{Hom}(A, Y) \to \mathrm{Hom}(A, X)\}_A$ be the inverse natural isomorphism.
In light of Yoneda's Lemma, we see that ϕ and ϕ^{-1} determine unique functions $f_\phi : X \to Y$ and $f_{\phi^{-1}} : Y \to X$ given by $f_\phi := \phi_X(\mathrm{id}_X)$ and $f_{\phi^{-1}} := \phi_Y^{-1}(\mathrm{id}_Y)$, respectively, such that $\phi_A = \mathrm{Hom}(A, f_\phi)$ and $\phi_A^{-1} = \mathrm{Hom}(A, f_{\phi^{-1}})$ for every set A.
We claim that they are mutually inverse and, therefore, $X \cong Y$.
To see this, let us compute:

$$\begin{aligned}\mathrm{Hom}(X, \mathrm{id}_X) &= \mathrm{id}_{\mathrm{Hom}(X,X)} \\ &= \phi_X^{-1} \circ \phi_X \\ &= \mathrm{Hom}(X, f_{\phi^{-1}}) \circ \mathrm{Hom}(X, f_\phi) \\ &= \mathrm{Hom}(X, f_{\phi^{-1}} \circ f_\phi)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Hom}(Y, \mathrm{id}_Y) &= \mathrm{id}_{\mathrm{Hom}(Y,Y)} \\ &= \phi_Y \circ \phi_Y^{-1} \\ &= \mathrm{Hom}(Y, f_\phi) \circ \mathrm{Hom}(Y, f_{\phi^{-1}}) \\ &= \mathrm{Hom}(Y, f_\phi \circ f_{\phi^{-1}})\end{aligned}$$

together tells us that f_ϕ and $f_{\phi^{-1}}$ are mutually inverse and, therefore, isomorphisms.
This proves $X \cong Y$.

Conversely, if $f : X \to Y$ is an isomorphism, then clearly $\phi := \{\phi_A := \mathrm{Hom}(A, f)\}_A$ is a natural isomorphism.
We have thus shown that (i.) if, and only if, (iii.), which ends the proof. □

Summing it all up, what Yoneda's Lemma tells us is that the only possible natural transformations are of the form $\mathrm{Hom}(-, f) : \mathrm{Hom}(-, X) \to \mathrm{Hom}(-, Y)$ or $\mathrm{Hom}(f, -) : \mathrm{Hom}(Y, -) \to \mathrm{Hom}(X, -)$, for some $f : X \to Y$.

This is very useful in Linear Algebra where we use that two vector spaces are isomorphic if, and only if, there is a natural isomorphism of sets of linear transformations - basically the preceding corollary, but applied to vector spaces.

Theorem 1.3.8.19. *The isomorphism in theorem 1.3.8.8 is natural for all three entries - that is, we have three natural isomorphisms:*

$$\phi_{X,Z} : \mathrm{Hom}(X \times -, Z) \to \mathrm{Hom}(X, \mathrm{Hom}(-, Z))$$

$$\phi_X^Y : \mathrm{Hom}(X \times Y, -) \to \mathrm{Hom}(X, \mathrm{Hom}(Y, -))$$

$$\phi_Z^Y : \mathrm{Hom}(- \times Y, Z) \to \mathrm{Hom}(-, \mathrm{Hom}(Y, Z))$$

We will not prove this result, since it follows trivially by inspection. We will, once more, leave it as an exercise to the reader.

With this, we have proven most of the stuff that we'll need to study Linear Algebra.
So, without further ado, let us begin.

Chapter 2

Planar Linear Algebra

2.1 Introduction

2.1.1 Operations

To start working with vector spaces we first need to understand that the perspective is going to change a bit from the previous chapter. We're leaving the domain of **set theory** and jumping right in the domain of **algebra**.

Algebra is the domain of mathematics that deals with operations and their properties.

Definition 2.1.1.1. *Given any set X, a **(binary) operation** on X is a function $f : X \times X \to X$.*

Example(s)

Let $\mathbb{N}$ be the set of natural numbers, as before. We have a few operations here:

$$f, g, h : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$$

$$(n, m) \mapsto f(n, m) := n + m$$
$$(n, m) \mapsto g(n, m) := nm$$
$$(n, m) \mapsto h(n, m) := n^m$$

and some of these operations have some properties that the others don't.
For instance, all three functions satisfy the following property:

- Let ϕ be an operation on X. There is some $n_e \in X$ such that $\phi(n, n_e) = n$ for all $n \in X$.

In the case of f, if we choose $n_e := 0$, we see that $f(n, 0) = n + 0 = n$, no matter which $n \in \mathbb{N}$ we chose, so f satisfies the property above.
In the case of g, if we choose $n_e := 1$, we see that $g(n, 1) = n \cdot 1 = n$, no matter which $n \in \mathbb{N}$ we chose, so g satisfies the property above.

Finally, in the case of h, if we choose $n_e := 1$, we see that $h(n, 1) = n^1 = n$, no matter which $n \in \mathbb{N}$ we chose, so h satisfies the property above.

Next up is the property:

- Let ϕ be an operation on X. There is some $n_e \in X$ such that $\phi(n_e, n) = n$ for all $n \in X$.

What can we say about f, g, h in this case? Well, it's easy to see that for f and g it still holds true - and it does so for the same value of n_e as before.
However, for h it fails. For instance, is there some number $x \in \mathbb{N}$ such that $h(x, 2) = 2$? Well, by definition of h we would need to have $x^2 = 2$ and so $x = \sqrt{2}$ which is not in $\mathbb{N}$ - this tells us that there's no such $x \in \mathbb{N}$. It follows that this property fails for h.

Summing up all of these together, we get the following property:

- (Identity element) Let ϕ be an operation on X. There is some $n_e \in X$ such that $\phi(n, n_e) = n = \phi(n_e, n)$ for all $n \in X$.

And we see that f and g have what's called an *identity element* - it's an element n_e such that if you fix it in any input of your operation, then your operation is just the identity function.

Consider now the following property:

- (Associativity) Let ϕ be an operation on X. Then, for all $n, m, l \in X$ we have that $\phi(\phi(n, m), l) = \phi(n, \phi(m, l))$.

In the case of f we can check

$$f(f(n, m), l) = f(n + m, l) = (n + m) + l = n + (m + l) = f(n, m + l) = f(n, f(m, l))$$

and see that f is associative.
In the case of g we can check

$$g(g(n, m), l) = g(nm, l) = (nm)l = n(ml) = g(n, ml) = g(n, g(m, l))$$

and see that g is associative.
However, for h, once again, this property fails: For instance, let us compare $h(h(2, 2), 3)$ and $h(2, h(2, 3))$:

$$h(h(2, 2), 3) = h(2^2, 3) = (2^2)^3 = 4^3 = 64$$

$$h(2, h(2, 3)) = h(2, 2^3) = 2^{(2^3)} = 2^8 = 256$$

so they are clearly different, and h is not associative.

One more:

- (Commutativity) Let ϕ be an operation on X. Then, for all $n, m \in X$ we have that $\phi(n, m) = \phi(m, n)$.

In the case of f we can easily see that $f(n, m) = n + m = m + n = f(m, n)$.
Similarly for g, we see that $g(n, m) = nm = mn = g(m, n)$.
But, once again, $h(2, 3) = 8 \neq 9 = h(3, 2)$, so h is not commutative.

These are the most common operations in $\mathbb{N}$ and some of their properties. Now, let us show something that is **not** an operation:
Consider the functions

$$f', g' : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$$
$$(n, m) \mapsto f'(n, m) := n - m$$
$$(n, m) \mapsto g'(n, m) := n/m.$$

Notice that I've just lied to you - these are **not** functions. To see that, take f' and apply it on $(3, 1)$. By definition of function, $f'(3, 1)$ should lie on $\mathbb{N}$, the codomain of f'. But, by definition of f', we see that $f'(3, 1) = 3 - 1 = -2$, which is **not** in $\mathbb{N}$.
Similarly, g' isn't a function for the same reason: It should take, for instance, $(1, 2)$ to a natural number - but it doesn't. It takes $(1, 2)$ to $g'(1, 2) = 1/2$ which, once more, is not a natural number.
However, for *some* specific values of the input, f' and g' really have outputs in $\mathbb{N}$. For that reason, they are called **partial operations** and, sadly, won't be studied in this text, since we're mostly concerned with proper operations.
If, however, you'd like to learn more about partial operations, you should click here or Google for "groupoid" - which is precisely the mathematical notion of a set with an associative partial operation.

Definition 2.1.1.2. *Given an operation $\phi : X \times X \to X$ we will say that*

- *(Identity element) ϕ admits an **identity element** if there is some $e \in X$ such that $\phi(x, e) = x = \phi(e, x)$ for all $x \in X$. In this case, e is called an **identity element**;*
- *(Associativity) ϕ is **associative** if $\phi(x, \phi(y, z)) = \phi(\phi(x, y), z)$ for all $x, y, z \in X$;*
- *(Commutative) ϕ is **commutative** if $\phi(x, y) = \phi(y, x)$ for all $x, y \in X$;*
- *(Inverse element) ϕ admits **inverse elements** if for all $x \in X$ there is some $y \in X$ such that $\phi(x, y) = e = \phi(y, x)$ for some identity element $e \in X$.*

Definition 2.1.1.3. *Let X be a set with two operations, $f, g : X \times X \to X$. We say that f **distributes over g on the left** (resp. **on the right**) if*

$$f(x, g(y, z)) = g(f(x, y), f(x, z))$$

(resp.

$$f(g(x, y), z) = g(f(x, z), f(y, z)))$$

for all $x, y, z \in X$.

If f *distributes over* g *on both sides, we simply say that* f ***distributes over*** g.

Example(s)

Following up on the previous example, we see that g (the multiplication) distributes over f (the addition):

$$g(n, f(m,l)) = g(n, m+l) = n(m+l) = nm + nl = f(nm, nl) = f(g(n,m), g(n,l))$$

$$g(f(n,m), l) = g(n+m, l) = (n+m)l = nl + ml = f(nl, ml) = f(g(n,l), g(m,l))$$

but f doesn't distribute (on either side!) over g:

$$1 + (1 \cdot 1) = 1 + 1 = 2 \neq 4 = 2 \cdot 2 = (1+1) \cdot (1+1)$$

$$(1 \cdot 1) + 1 = 1 + 1 = 2 \neq 4 = 2 \cdot 2 = (1+1) \cdot (1+1)$$

All this talk now brings us to a very specific definition:

Definition 2.1.1.4. *Let* X *be a set with two operatios* $A, M : X \times X \to X$. *We will say that* (X, A, M) *is a* ***field*** *if*

(1) A *is associative;*

(2) A *is commutative;*

(3) A *has an identity element;*

(4) A *has inverses;*

(5) M *is associative;*

(6) M *is commutative;*

(7) M *has an identity element;*

(8) M *has inverses (excluding the additive identities);*

(9) M *distributes over* A.

In this case, we call A *and* M, *respectively, the field's* ***addition*** *and* ***multiplication*** *operations, and denote them simply by* $x + y := A(x, y)$ *and* $xy := M(x, y)$ *for all* $(x, y) \in X \times X$.

Proposition 2.1.1.5. *The set* $\mathbb{R}$ *of real numbers with the usual addition and multiplication is a field.*

Proof

This is immediate, since for every $x, y, z \in \mathbb{R}$ we have:

(1) $x+(y+z) = (x+y)+z$;

(2) $x + y = y + x$;

(3) $x + 0 = 0 + x = x$;

(4) $x+(-x) = (-x)+x = 0$;

(5) $x(yz) = (xy)z$;

(6) $xy = yx$;

(7) $x \cdot 1 = 1 \cdot x = x$;

(8) $xx^{-1} = x^{-1}x = 1$ if $x \neq 0$;

(9) $x(y+z) = xy + xz$ and $(x+y)z = xz + yz$.

□

Example(s)

Notice, however, that the sets $\mathbb{N}$ and $\mathbb{Z}$, of the naturals and integers, respectively, are **not** fields: $\mathbb{N}$ doesn't have either additive or multiplicative inverses (so it fails properties (4) and (8)), and $\mathbb{N}$ doesn't have multiplicative inverses (so it fails property (8)).
On the other hand, it's easy to see that $\mathbb{Q}$, the set of rational numbers, is indeed a field. It is actually constructed to be, in some sense, "the smallest field which extends $\mathbb{Z}/\mathbb{N}$".
Finally, the set $\mathbb{C}$ of complex numbers is also a field if you define the inverse of $z = x + iy$ to be $z^{-1} := \dfrac{x - iy}{x^2 + y^2}$. Indeed:

$$zz^{-1} = (x + iy)\left(\frac{x - iy}{x^2 + y^2}\right) = \frac{x^2 + y^2}{x^2 + y^2} = 1$$

so it is indeed an inverse for z.

Let us show some properties of fields:

Lemma 2.1.1.6. *Let $(k, +, \cdot)$ be a field. Then the following hold:*

(a) There's a unique additive identity;

(b) For each $x \in k$, there's a unique additive inverse;

(c) There's a unique multiplicative identity;

(d) For each $x \in k$, there's a unique multiplicative inverse;

(e) Let 0 be an additive identity of k. Then $0x = 0$ for all $x \in k$.

(f) Let 1 be a multiplicative identity of k. Then $-x = (-1)x$, where $(-1) + 1 = 0$.

Proof

(a) Let 0 and $0'$ be two additive identities of k. Then

$$0 = 0 + 0' = 0'$$

where the leftmost equality holds since $0'$ is additive identity, and the rightmost equality holds since 0 is additive identity, and so $0 = 0'$.

(b) Given $x \in k$, let x' and x'' be two additive inverses to x. Then

$$x' = x' + 0 = x' + (x + x'') = (x' + x) + x'' = 0 + x'' = x''$$

and so $x' = x''$.

(c) Let 1 and $1'$ be two multiplicative identities of k. Then

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

where the leftmost equality holds since $1'$ is a multiplicative identity, and the rightmost equality holds since 1 is a multiplicative identity, so $1 = 1'$.

(d) Given $x \in k$, let x'' and x'' be two multiplicative inverses to x. Then

$$x' = x' \cdot 1 = x'(xx'') = (x'x)x'' = 1 \cdot x'' = x''$$

and so $x' = x''$.

(e) Given $x \in k$, we have that

$$0x = (0 + 0)x = 0x + 0x$$

since k is a field and 0 is the additive identity. Let $y \in k$ be the additive inverse of $0x$.

Then, since the above is true, we can see that $(0x) + y = (0x + 0x) + y$ is also true. But the LHS is just 0, since y is the additive inverse of $0x$, and the RHS is just $(0x+0x)+y = 0x + (0x + y) = 0x + 0 = 0x$, so the above equation evaluates to $0 = 0x$.

(f) Given $x \in k$, we have that $0x = (1 + (-1))x$ since $1 + (-1) = 0$. But now, by the distributive property of fields we see that $0x = (1 + (-1))x = (1)x + (-1)x$.

But $0x = 0$ and $1x = x$, so this is just $0 = x + (-1)x$. Since additive inverses are unique, we see that $(-1)x = -x$.

□

This result basically tells us that every field is "similar" to $\mathbb{R}$, in some sense.

Remark 2.1.1.7

The reason why we require that the multiplication has inverses for all elements *except for 0* is precisely because of item (e) above. Since $0x = 0$ for all x, if we could have some 0^{-1}, then $0 = 00^{-1} = 1$ so we would have $0 = 1$.
But since $1x = x$ for all x, this would imply that $x = 1x = 0x = 0$, so **every element of the field would have to be 0 for it to be consistent**.
In other words, the only set that satisfies all the properties of a field and also has a multiplicative inverse to 0 is the set $\{0\}$.
In fact:

Proposition 2.1.1.8. *The set $1 = \{0\}$ with addition and multiplication being equal and given by $0 + 0 = 0 \cdot 0 = 0$ is a field. Its multiplicative and additive identities are 0, who is also the inverse of 0.*

Proof

There's literally nothing to prove. □

Finally, to end this section, let us give some examples of fields that aren't 1, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$.

Example(s)

Let $p \in \mathbb{N}$ be a prime number (that is, there are only two ways to write $p = nm$: $n = p, m = 1$ and $n = 1, m = p$). Consider the set $p \in \mathbb{N}$ - that is, $p = \{0, 1, 2, \cdots, p-1\}$. We will give a field structure to p as follows:
For any $x, y \in p$, define:

- $x + y$ is the remainder of the division of $x + y$ in $\mathbb{N}$ by p;
- xy is the remainder of the division of xy in $\mathbb{N}$ by p.

We claim that p with those two operations is a field, which will be denoted by either $\mathbb{Z}_p$, $\mathbb{Z}/p\mathbb{Z}$ or $\mathbb{F}_p$.
For instance, let us do some computations with $p = 3$.
In this case, $p = \{0, 1, 2\}$, and so we have the following tables of operations:

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

and

$\times$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

It is, then, readily seen that 0 and 1 are, respectively, the additive and multiplicative identities of $\mathbb{F}_3$.
We can also see that $1 + 2 = 0$ so 1 and 2 are additive inverses to each other. Similarly, we see that $1 \cdot 1 = 1 = 2 \cdot 2$ so both 1 and 2 are multiplicative inverses to themselves.
This shows that $\mathbb{F}_3$ is a field.
Building similar tables of operations we can prove that any $\mathbb{F}_p$ is a field.
Let us now show the necessity of p being a prime.

Let $4 = \{0, 1, 2, 3\}$. Let's try building the same operations:

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

and

$\times$	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

but this shows that 2 doesn't have any multiplicative inverses: $2 \cdot 0 = 0$, $2 \cdot 1 = 2$, $2 \cdot 2 = 0$ and $2 \cdot 3 = 2$.
But by definition of a field, the only element that has no multiplicative inverse is 0. But clearly $2 \neq 0$ (since $1 + 2 \neq 1$), so $\mathbb{Z}/4\mathbb{Z}$ cannot be a field.
This happens precisely because 4 can be written as $4 = nm$ in *three* different ways: $4 = 4 \cdot 1 = 1 \cdot 4 = 2 \cdot 2$.
Since this isn't supposed to be a course on field theory, we won't go into much detail on how to prove that $\mathbb{Z}/n\mathbb{Z}$ is a field if, and only if, n is prime.

2.1.2 The plane $\mathbb{R}^2$

Let us start this section with the set that will be the focus of most, if not all, of this chapter: $\mathbb{R}^2$.

By definition, $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the set of ordered pairs of real numbers.

Definition 2.1.2.1. *We're going to define the* ***addition*** $A : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ *to be given by* $A((x,y),(z,w)) := (x+z, y+w)$ *for any* $(x,y),(z,w) \in \mathbb{R}^2$.

Proposition 2.1.2.2. *The addition* A *we've just defined satisfies the following properties:*

(i.) A *is associative;*

(ii.) A *is commutative;*

(iii.) A *admits an identity element;*

(iv.) A *admits inverses.*

Proof

Choose any three elements $(a,b),(c,d),(e,f) \in \mathbb{R}^2$. Then:

(i.)

$$\begin{aligned} A\left(A\left((a,b),(c,d)\right),(e,f)\right) &= A((a+c,b+d),(e,f)) \\ &= ((a+c)+e,(b+d)+f) \\ &= (a+(c+e),b+(d+f)) \\ &= A((a,b),(c+e,d+f)) = A((a,b),A((c,d),(e,f))), \end{aligned}$$

so A is associative;

(ii.)

$$A((a,b),(c,d)) = (a+c,b+d) = (c+a,d+b) = A((c,d),(a,b))$$

so A is commutative;

(iii.) $A((a,b),(0,0)) = (a+0,b+0) = (a,b) = (0+a,0+b) = A((0,0),(a,b))$ so A has identity $(0,0)$;

(iv.) $A((a,b),(-a,-b)) = (a-a,b-b) = (0,0) = (-a+a,-b+b) = A((-a,-b),(a,b))$ so A has inverses.

□

Remark 2.1.2.3

From now on, we're gonna denote $A((a,b),(c,d))$ by $(a,b)+(c,d)$ for any $(a,b),(c,d) \in \mathbb{R}^2$ since, by the preceding proposition, it behaves well-enough like number addition.

And it seems there's not much else we can do with $\mathbb{R}^2$ for now.
To proceed with our studies, then, we're gonna need a new approach.

Definition 2.1.2.4. *We'll denote E_* the **pointed Euclidean plane**. That is, E is the set of all points in the Euclidean plane, and $*$ is a distinguished point.*

Example(s)

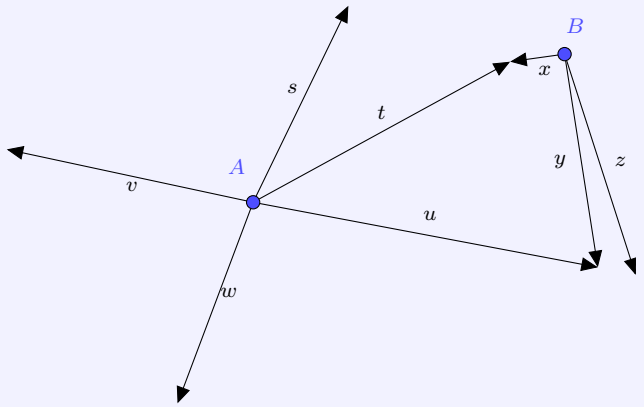
For instance,

we can think of the set E_A (whose elements are all points in the plane, including A and B, but distinguishing A) and the set E_B (whose elements are all points in the plane, including A and B, but distinguishing B).

Definition 2.1.2.5. *Given a pointed Euclidean plane E_*, we define a **vector in** E_* to be any directed segment starting in $*$.*

Example(s)

Continuing the above example,



we see that s, t, u, v, w are vectors in E_A, but not in E_B, whereas x, y, z are vectors in E_B, but not in E_A.

Lemma 2.1.2.6. *Let V_A denote the set of all vectors in E_A. Then $E_A \cong V_A$.*

Proof

Consider the function $t : V_A \to E_A$ that takes any vector $v \in V_A$ to its endpoint $t(v)$, and takes the null vector to A.

- t is injective:

  If $t(v) = t(u)$ for some $v, u \in V_A$, then v and u have the same endpoints. Since they also have the same starting points (by definition), they are equal - hence, $v = u$ and t is injective.

- t is surjective:

  Let $P \in E_A$ be a point. Consider the directed segment $\overrightarrow{AP}$. It is, by definition, a vector in V_A whose endpoint is P, and, hence, $t(\overrightarrow{AP}) = P$, and so t is surjective.

Definition 2.1.2.7. *Given any vector $v \in V_A$, we define its **magnitude** or **size** or **norm** to be $\|v\| := |AP|$, where $v = \overrightarrow{AP}$.*

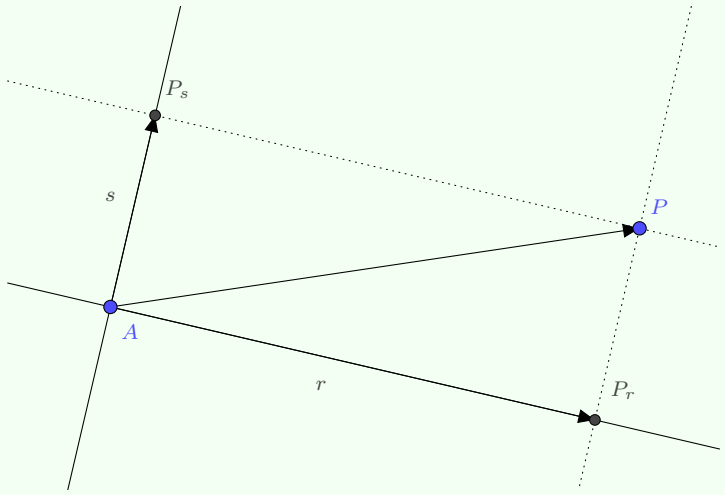
Since t is both injective and surjective, it is a bijection. This proves the result. □

Let now $V_A^{r,s}$ be the following set: Take any two perpendicular lines r and s through A. Then, $V_A^{r,s}$ will be the set $(V_A \cap r) \times (V_A \cap s)$ - that is, the set of all pairs vectors of E_A such that the first vector lies entirely in r and the second vector lies entirely in s.

Lemma 2.1.2.8. *For any E_A, we have that $V_A \cong V_A^{r,s}$.*

Proof

Let $f : V_A \to V_A^{r,s}$ be the following function: For any vector $\overrightarrow{AP}$ in V_A, we can consider the parallel to r through P, r_P and the parallel to s through P, s_P.
Since $r \parallel r_P$, we have that $r \cap r_P = \varnothing$, and similarly we have that $s \cap s_P = \varnothing$.
But since $r \nparallel s_P$ and they are both lines, we have that $r \cap s_P$ is a single point - let's call it P_r. Similarly, we see that $s \cap r_P$ is a single point - let's call it P_s.
But now, $P_r \in r$, by definition. So $\overrightarrow{AP_r}$ is a vector which lies entirely in r. Similarly, $P_s \in s$ so $\overrightarrow{AP_s}$ is a vector which lies entirely in s.
We then define $f(\overrightarrow{AP}) := (\overrightarrow{AP_r}, \overrightarrow{AP_s})$.



- f is injective:

  Let v, u be vectors in V_A such that $f(v) = f(u)$. This means that $v = \overrightarrow{AP}$ and $u = \overrightarrow{AQ}$ for some uniquely determined points P, Q in E_A. Now, by definition, $f(v) = f(\overrightarrow{AP}) = (\overrightarrow{AP_r}, \overrightarrow{AP_s})$ and $f(u) = f(\overrightarrow{AQ}) = (\overrightarrow{AQ_r}, \overrightarrow{AQ_s})$.

  But $f(v) = f(u)$ implies $\overrightarrow{AP_r} = \overrightarrow{AQ_r}$ and $\overrightarrow{AP_s} = \overrightarrow{AQ_s}$, by definition of set product.

  But then this means that $P_r = Q_r$ and $P_s = Q_s$.

  This means that both P and Q belong to both of r_P and s_P. But r_P and s_P are lines, so their intersection has, at most, one point. This means that $P = Q$, and so f is injective.

- f is surjective:

Take any $(v,u) \in V_A^{r,s}$. Then, there are uniquely determined $P,Q \in E_A$ such that $v = \overrightarrow{AP}$ and $u = \overrightarrow{AQ}$.

Now, consider the lines r_Q and s_P defined, respectively, to be the lines parallel to r through Q, and parallel to s through P.

Since $s_P \parallel s \nparallel r \parallel r_Q$, we see that $s_P \nparallel r_Q$ and so $s_P \cap r_Q$ is a single point - T.

It is, then, easy tos see that $f(\overrightarrow{AT}) = (v,u)$, so f is surjective.

It follows then that f is indeed a bijection, which ends the proof. □

Finally, we can prove the result we've all been waiting for:

Theorem 2.1.2.9. *Let E_A be a pointed Euclidean plane. Then $V_A^{r,s} \cong \mathbb{R}^2$.*

Proof

Take any two points $R \in r$ and $P \in p$, respectively, both different from A.
Now, given any $v \in r$ a vector lying entirely in r, we say that v is **positive** if the endpoint of v lies in the same side of the semiplane defined by s as R, **zero** if $v = \overrightarrow{AA}$ the null vector, and **negative** otherwise.
Similarly, given $u \in s$ a vector lying entirely in s, we say that u is **positive** if the endpoint of u lies in the same side of the samiplane defined by r as S, **zero** if $u = \overrightarrow{AA}$ the null vector, and **negative** otherwise.
This can be seen as a function $\operatorname{sgn} : V_A^{r,s} \to \{0,1,-1\}$.
That said, we're gonna define a function $f : V_A^{r,s} \to \mathbb{R}^2$ by putting

$$f(v,u) := (\operatorname{sgn}(v)\|v\|, \operatorname{sgn}(u)\|u\|).$$

We claim that f is the bijection we're looking for.

- f is injective:

  Let $(v,u),(v',u') \in V_A^{r,s}$ be two elements such that $f(v,u) = f(v',u')$. This means that $\operatorname{sgn}(v)\|v\| = \operatorname{sgn}(v')\|v'\|$ and $\operatorname{sgn}(u)\|u\| = \operatorname{sgn}(u')\|u'\|$. Since $\|-\|$ is always a non-negative number, we see that $\operatorname{sgn}(v)\|v\| = \operatorname{sgn}(v')\|v'\|$ if, and only if, $\operatorname{sgn}(v) = \operatorname{sgn}(v')$, and similarly for $\operatorname{sgn}(u) = \operatorname{sgn}(u')$.

  Now this implies, together with the equations $\operatorname{sgn}(v)\|v\| = \operatorname{sgn}(v')\|v'\|$ and $\operatorname{sgn}(u)\|u\| = \operatorname{sgn}(u')\|u'\|$, that $\|v\| = \|v'\|$ and $\|u\| = \|u'\|$. But since the norm is entirely determined by the endpoint (since the starting point is fixed, being A), we see that $\|v\| = \|v'\|$ if, and only if, v and v' have the same endpoints. Similarly, we see that u and u' have the same endpoints.

  Finally, since v and v' have the same endpoints they must be equal, and the same goes for u and u'.

  It follows that $(v,u) = (v',u')$ and so f is indeed injective.

- f is surjective:

  Take any $(x, y) \in \mathbb{R}^2$. Now draw two circles centered in A: one with radius $|x|$ and one with radius $|y|$. Call these circles C_x and C_y, resp.

  Since C_x is a circle and r is a line which contains a point inside the circle, we know that $C_x \cap r$ is precisely two points - R_1, R_2. Similarly, we know that $C_y \cap s$ is precisely two points - S_1, S_2.

  Now, we take $\overrightarrow{x}$ to be the vector ending at either R_1 or R_2 such that $\text{sgn}(\overrightarrow{x}) = \text{sgn}(x)$. Similarly, we define $\overrightarrow{y}$ to be the vector ending at either S_1 or S_2 such that $\text{sgn}(\overrightarrow{y}) = \text{sgn}(y)$.

  Now, by construction, $(\overrightarrow{x}, \overrightarrow{y}) \in V_A^{r,s}$ is such that $\text{sgn}(\overrightarrow{x})\|\overrightarrow{x}\| = x$ and $\text{sgn}(\overrightarrow{y})\|\overrightarrow{y}\| = y$, so $f(\overrightarrow{x}, \overrightarrow{y}) = (x, y)$ and we see that f is surjective.

Finally, we can conclude that f is the bijection we were looking for. This ends the proof of the theorem. □

Corollary 2.1.2.10. *By composition of isomorphisms we have:*

$$E_A \cong V_A \cong V_A^{r,s} \cong \mathbb{R}^2,$$

so the elements of $\mathbb{R}^2$ *can be thought of as vectors in the pointed Euclidean plane* E_A*.*

Lemma 2.1.2.11. *In the above bijection* $f : V_A^{r,s} \to \mathbb{R}^2$*, the image of* r *is the set*

$$\mathbb{R} \times \{0\} := \{(x, y) \in \mathbb{R}^2 \mid y = 0\}$$

and the image of s *is the set*

$$\{0\} \times \mathbb{R} := \{(x, y) \in \mathbb{R}^2 \mid x = 0\}.$$

Proof

Take $(v, \overrightarrow{AA}) \in V_A^{r,s}$ any vector lying entirely on r. Then, by definition,

$$f(v, \overrightarrow{AA}) = (\text{sgn}(v)\|v\|, \text{sgn}(\overrightarrow{AA})\|\overrightarrow{AA}\|),$$

but both $\text{sgn}(\overrightarrow{AA})$ and $\|\overrightarrow{AA}\|$ equal 0. So $f(v, \overrightarrow{AA}) = (\text{sgn}(v)\|v\|, 0)$.
This shows that $f(r) \subseteq \mathbb{R} \times \{0\}$.
Conversely, take any $(x, 0) \in \mathbb{R} \times \{0\}$. It is clearly the image of some vector pair $(\overrightarrow{x}, \overrightarrow{AA})$, and so $\mathbb{R} \times \{0\} \subseteq f(r)$.
We can argue analogously and show that $f(s) = \{0\} \times \mathbb{R}$.
This ends the proof. □

Definition 2.1.2.12. *We will denote the sets* $\mathbb{R} \times \{0\}$ *and* $\{0\} \times \mathbb{R}$ *the* X***-axis*** *and the* Y***-axis****, respectively. Sometimes we'll also denote them, respectively, by* $\mathbb{X}$ *and* $\mathbb{Y}$*.*

Remark 2.1.2.13

Notice that, as sets, both the X- and the Y-axis are in bijection with $\mathbb{R}$, and with a line (the X-axis is in bijection with r and the Y-axis is in bijection with s).
Therefore, $\mathbb{R}$ is in bijection with a line. So it makes sense to think of the set of real numbers as a *line*, and call it the **real line**.

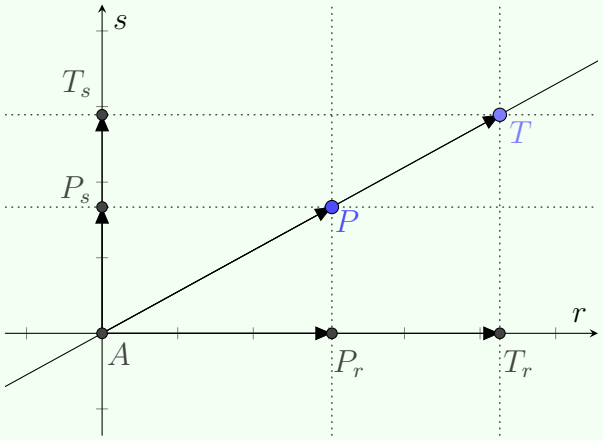
Finally, to end this section, one last result:

Lemma 2.1.2.14. *Take $t \subseteq E_A$ any line through A. Then there is some point $(x_t, y_t) \in \mathbb{R}^2$ such that $f(t) = \{(x, y) \in \mathbb{R}^2 \mid x = \lambda x_t \text{ and } y = \lambda y_t, \text{ for some } \lambda \in \mathbb{R}\}$ where f is the bijection $f : E_A \to \mathbb{R}^2$ defined in the previous lemma.*

Proof

Let $\mathbb{R}(x_t, y_t)$ denote the set $\{(x, y) \in \mathbb{R}^2 \mid x = \lambda x_t \text{ and } y = \lambda y_t, \text{ for some } \lambda \in \mathbb{R}\}$.
Draw $\mathcal{C}(A, 1)$ the circle of radius 1 centered at A. Since t contains a point in the inside of the circle, $t \cap \mathcal{C}(A, 1) = 2$. Choose any of those two points and call it P.
Now, let us define $(x_t, y_t) := f(\overrightarrow{AP_r}, \overrightarrow{AP_s})$.
Now take any point $T \in t$, and look at the vectors $\overrightarrow{AT_r}$ and $\overrightarrow{AT_s}$. Now, by using similar triangles we see that

$$\frac{\operatorname{sgn}(\overrightarrow{AT_r})\|\overrightarrow{AT_r}\|}{x_t} = \frac{\operatorname{sgn}(\overrightarrow{AT_s})\|\overrightarrow{AT_s}\|}{y_t} = \frac{\operatorname{sgn}(\overrightarrow{AT})\|\overrightarrow{AT}\|}{\operatorname{sgn}(\overrightarrow{AP})\|\overrightarrow{AP}\|}$$



So we pick λ_T to be any of those (since they're all equal).
Clearly, then, by definition of λ_T, we have that

$$f(\overrightarrow{AT_r}, \overrightarrow{AT_s}) = (\operatorname{sgn}(\overrightarrow{AT_r})\|AT_r\|, \operatorname{sgn}(\overrightarrow{AT_s})\|AT_s\|) = (\lambda_T x_t, \lambda_T y_t).$$

Since this holds true for any $T \in t$, we have that $f(t) \subseteq \mathbb{R}(x_t, y_t)$.

Conversely, take some $(x_t, y_t) \in \mathbb{R}^2$ fixed.

Now, let P be the inverse image of (x_t, y_t) under f (it can be done uniquely since f is bijective), and let t be the line through A that also passes through P (it is also unique, since a line is uniquely determined by two distinct points).
By the same reasoning as before, if we now take the inverse image of $(\lambda x_t, \lambda y_t)$ for any real $\lambda \in \mathbb{R}$ to be some P^λ, we see that the triangles AP_rP_s and $AP_r^\lambda P_s^\lambda$ are similar, since

$$\frac{\overline{AP_r}}{\overline{AP_r^\lambda}} = \frac{\overline{AP_s}}{\overline{AP_s^\lambda}} = \frac{\overline{P_sP_r}}{\overline{P_s^\lambda P_r^\lambda}} = \lambda,$$

so P^λ lies in t.
This shows that $\mathbb{R}(x_t, y_t) \subseteq f(t)$.
Finally, we can conclude that $f(t) = \mathbb{R}(x_t, y_t)$, which ends the proof. □

This shows that the subsets of $\mathbb{R}^2$ where each element is obtained simply by multiplying both coordinates of a fixed vector by a fixed number are **lines**.

With this we can finally define:

Definition 2.1.2.15. *We're gonna define the* ***multiplication of a point in*** $\mathbb{R}^2$ ***by a real number*** *to be as such:*

$$\lambda(x, y) := (\lambda x, \lambda y)$$

for any $\lambda \in \mathbb{R}$ *and* $(x, y) \in \mathbb{R}^2$.

Remark 2.1.2.16

By the preceding theorem, this multiplication is simply scaling the vector $(x, y) \in \mathbb{R}^2$ in a straight line to the center to become the same size as λ (notice that if $\lambda < 0$, this means that the vector will also change sign).
For this reason, this operation is more often than not called the **scalar multiplication** or **multiplication by a scalar**.

Corollary 2.1.2.17. *Every line through the point* $(0, 0)$ *in* $\mathbb{R}^2$ *is just the subset of* $\mathbb{R}^2$ *formed by all scalar multiples of a fixed vector.*

Definition 2.1.2.18. *Given any vector* $v \in \mathbb{R}^2$ *we're gonna denote the* ***line through zero containing*** v *to be the set* $\mathbb{R}v$ *given by*

$$\mathbb{R}v := \{u \in \mathbb{R}^2 \mid \exists \lambda \in \mathbb{R} \text{ such that } u = \lambda v\},$$

that is, the set of all scalar multiples of v.

And to wrap things up, some properties of scalar multiplication:

Proposition 2.1.2.19. *Let* $v, u \in \mathbb{R}^2$ *and* $\lambda, \mu \in \mathbb{R}$. *Then, the following hold:*

(1) (Associative) $(\mu\lambda)v = \mu(\lambda v)$;

(2) (Commutative) $\lambda v = v\lambda$*;*

(3) (Identity element) There is some $\epsilon \in \mathbb{R}$ *such that* $\epsilon v = v\epsilon = v$*;*

(4) (Scalar product distributes over vector sum) $\lambda(v+u) = \lambda v + \lambda u$ *and* $(v+u)\mu = v\mu + u\mu$*;*

(5) (Scalar product distributes over scalar sum) $(\mu + \lambda)v = \mu v + \lambda v$*.*

Proof

Write, once and for all, $v = (v_1, v_2)$ and $u = (u_1, u_2)$.

(1) This follows from the fact that real number multiplication is associative:

$$\begin{aligned}(\mu\lambda)v &= (\mu\lambda)(v_1, v_2)\\ &= ((\mu\lambda)v_1, (\mu\lambda)v_2)\\ &= (\mu(\lambda v_1), \mu(\lambda v_2))\\ &= \mu(\lambda v_1, \lambda v_2)\\ &= \mu(\lambda(v_1, v_2)) = \mu(\lambda v)\end{aligned}$$

which holds true since μ, λ, v_1, v_2 are real numbers.

(2) This follows directly from the fact that real number multiplication is commutative:

$$\begin{aligned}\lambda v &= \lambda(v_1, v_2)\\ &= (\lambda v_1, \lambda v_2)\\ &= (v_1\lambda, v_2\lambda)\\ &= (v_1, v_2)\lambda = v\lambda\end{aligned}$$

which holds true since λ, v_1, v_2 are real numbers.

(3) This follows from the fact that real number multiplication has an identity:

$$1v = 1(v_1, v_2) = (1v_1, 1v_2) = (v_1, v_2) = v.$$

(4) This follows from the fact that real number multiplication distributes over real number sums:

$$\begin{aligned}\lambda(v+u) &= \lambda((v_1, v_2) + (u_1, u_2))\\ &= \lambda(v_1 + u_1, v_2 + u_2)\\ &= (\lambda(v_1 + u_1), \lambda(v_2 + u_2))\\ &= (\lambda v_1 + \lambda u_1, \lambda v_2 + \lambda u_2)\\ &= (\lambda v_1, \lambda v_2) + (\lambda u_1 + \lambda u2)\\ &= \lambda(v_1, v_2) + \lambda(u_1, u_2) = \lambda v + \lambda u\end{aligned}$$

which holds true since $\lambda, v_1, v_2, u_1, u_2$ are real numbers.

Distribution on the right-side holds since we've already proven (2), so $\lambda(v+u) = (v+u)\lambda$ and $\lambda v + \lambda u = v\lambda + u\lambda$.

(5) Similar to (4):

$$\begin{aligned}(\mu+\lambda)v &= (\mu+\lambda)(v_1, v_2)\\ &= ((\mu+\lambda)v_1, (\mu+\lambda)v_2)\\ &= (\mu v_1 + \lambda v_1, \mu v_2 + \lambda v_2)\\ &= (\mu v_1, \mu v_2) + (\lambda v_1, \lambda v_2)\\ &= \mu(v_1, v_2) + \lambda(v_1, v_2) = \mu v + \lambda v\end{aligned}$$

which holds true since μ, λ, v_1, v_2 are real numbers.

Distribution on the right follows, again, from (2).

□

So from here onwards, vectors, points and elements of $\mathbb{R}^2$ will be taken as being the same thing, since everything we did in this section was to show that $\mathbb{R}^2$ is a great model of the pointed Euclidean plane.

To end this section, then, let us define:

Definition 2.1.2.20. *We will call the point $(0,0) \in \mathbb{R}^2$ of the **origin** or the **zero** point of $\mathbb{R}^2$, and denote it simply by 0 when there's no ambiguity.*

Proposition 2.1.2.21. *For any vector $v \in \mathbb{R}^2$ we have that $0v = 0$.*

Proof

Take any $v \in \mathbb{R}^2$. Now, by what we've already shown, we have:

$$0v = (0+0)v = 0v + 0v$$

and so, by subtracting $0v$ on both sides, we get $0v = 0$, just as we wanted. □

2.2 $\mathbb{R}^2$ as a set of vectors

2.2.1 The shape of $\mathbb{R}^2$

So, now that we know what $\mathbb{R}^2$ is, we're gonna study how functions interact with it.

Example(s)

Consider the two following functions $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ given by $f(x, y) := (x, x^2)$ and $g(x, y) := (2x - y, x + y)$ respectively.
It's not hard to see that the first function takes the line $y = 0$ (that is, the X-axis) into the parabola $y = x^2$. In particular, notice that this function doesn't preserve *any* lines through zero.
Conversely, we claim that g does preserve *every* line through zero. To see this, we will show that g preserves scalar multiplication. That suffices, by definition of lines through zero:
Indeed, if r is the line of all scalar multiples of a vector v, then any vector $w \in r$ is of the form λv for some $\lambda \in \mathbb{R}$. Let $g(v) = v' \in \mathbb{R}^2$ be the image of v under g. If we can prove that g preserves scalar multiplication, then $g(w) = g(\lambda v) = \lambda g(v) = \lambda v'$, and so every point in the line of the multiples of v goes to the line of the multiples of v'.
So it remains to show that g preserves scalar multiplication. But that's easy: Take any $(x, y) \in \mathbb{R}^2$ and any $\lambda \in \mathbb{R}$. Then:

$$\begin{aligned} g(\lambda(x, y)) &= g(\lambda x, \lambda y) \\ &= (2(\lambda x) - (\lambda y), (\lambda x) + (\lambda y)) \\ &= (2\lambda x - \lambda y, \lambda x + \lambda y) \\ &= (\lambda(2x - y), \lambda(x + y)) \\ &= \lambda(2x - y, x + y) = \lambda g(x, y) \end{aligned}$$

and so we see that g preserves scalar multiplication - and, therefore, preserves any line through the origin.

There's only one slight issue: Indeed, for a function to preserve lines it suffices for it to preserve scalar multiplication, but there's no telling what happens to the origin during this process. Think about it: Just because every line becomes another line after applying a certain function, that doesn't mean that every point must stand still.
For instance, consider the function $h : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $h(x, y) := (x, x)$ if $y = 0$ and $h(x, y) := (y, y)$ otherwise. Then:

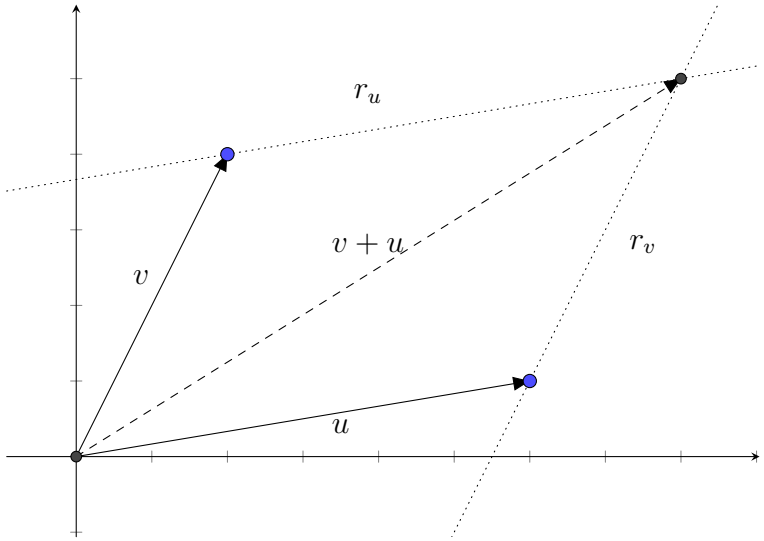
$$h(\lambda(x, 0)) = h(\lambda x, \lambda \cdot 0) = (\lambda x, \lambda x) = \lambda(x, x) = \lambda h(x, 0)$$

$$h(\lambda(x, y)) = h(\lambda x, \lambda y) = (\lambda y, \lambda y) = \lambda(y, y) = \lambda h(x, y)$$

and so h preserves scalar multiplication - and hence preserves lines. However, as we'll see further ahead, this function is still not good enough - for one simple reason: It might preserve lines, but it doesn't preserve parallelograms.

Proposition 2.2.1.1. *Given any two vectors $v, u \in \mathbb{R}^2$, consider the following construction:*

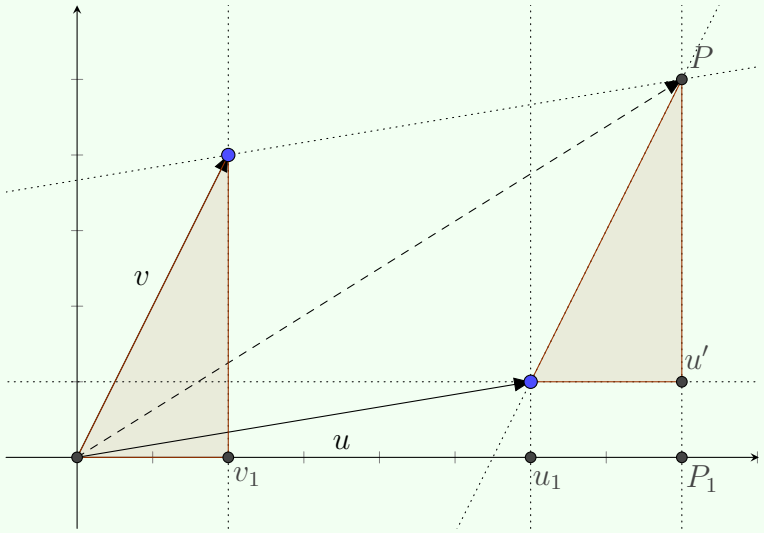
1. *Draw the two vectors $v, u \in \mathbb{R}^2$;*
2. *Take r_v the line parallel to v and containing the endpoint of u (there's a unique such line);*
3. *Take r_u the line parallel to u and containing the endpoint of v (there's a unique such line).*



Then, $v + u = r_u \cap r_v$.

Proof

First, how do we know that $r_v \cap r_u$ is non-empty? Well, by construction, $r_v \parallel v$ and $r_u \parallel u$. But v and u meet at 0, so r_v and r_u must also meet.
Second, since r_v and r_u are lines, they have, at most, a single meet point. Since we already know they have a non-empty intersection, then we know that they meet at a single point. Call this point $P \in \mathbb{R}^2$.
Consider now the following triangles:



First, notice that segments u_1P_1 and uu' are congruent (since $uu'P_1u_1$ is a rectangle, by construction).
Now we prove that the segments $0v_1$ and uu' are congruent. To do that, we use the triangles $0v_1v$ and $uu'P$. We claim they are congruent: Indeed, both are right triangles (on v_1 and u', respectively), the angles $v0v_1$ and Puu' are the same (since the segment Pu is parallel to $v0$, by construction) and the segments $0v$ and Pu are congruent (once again, by construction). This implies that the triangles are congruent and, therefore, vv_1 is congruent to Pu', and $0v_1$ is congruent to uu'.
Now, it's easy to see that the segment $0P_1$ is just the concatenation of the segments $0u_1$ and u_1P_1, so the length of $0P_1$ is the sum of the lengths of $0u_1$ and u_1P_1.
But the length of $0u_1$ is, by definition, u_1, and the length of u_1P_1 is the length of uu', which, as we've shown, is the length of $0v_1$ - which, once again by definition, is just v_1.
So the length of $0P_1$ - which is just P_1 - is the sum of u_1 and v_1 - that is, $P_1 = u_1 + v_1$.

We can proceed analogously on the Y-axis and show that $P_2 = u_2 + v_2$, which finally shows that $P = v + u$, just as stated. □

Corollary 2.2.1.2. *Any line $r \subseteq \mathbb{R}^2$ is of the form $r = v + r'$ where v is a fixed vector and r' is a line through zero which is parallel to r.*

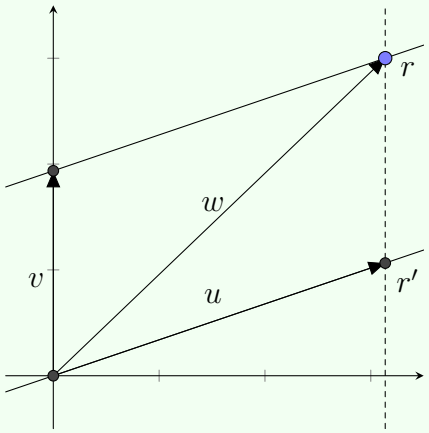
Proof

There are two possible cases:

1. $r \cap \mathbb{Y} \neq \varnothing$.

   In this case, take $v := r \cap \mathbb{Y}$ and draw $r' \parallel r$ through 0.

   Now given any $w \in r$, we can take a line parallel to v through w, and since v cuts r', this new line will also cut r'. Call this new point u.



Clearly, then, by construction, we see that $w = v + u$.

2. $\underline{r \cap \mathbb{Y} = \varnothing}$.

   In this case, just take $v := r \cap \mathbb{X}$. We know this point exists because $\mathbb{X} \perp \mathbb{Y}$ and $r \parallel \mathbb{Y}$ together imply $r \perp \mathbb{X}$, and so $r \cap \mathbb{X} \neq \varnothing$.

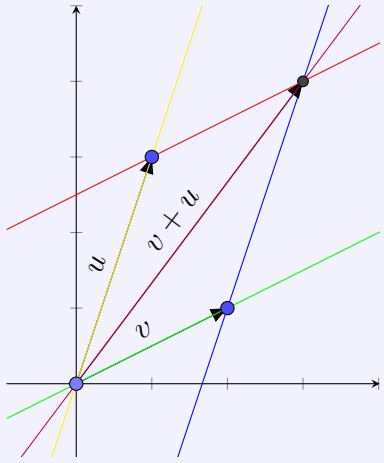
   Now we can just proceed as in the previous case. We'll skip the details of the proof.

This shows that every point in r can be written as $v + r'$, which ends the proof. □

Corollary 2.2.1.3. *Every line in* $\mathbb{R}^2$ *is of the form* $v + \mathbb{R}u$ *for some* $v, u \in \mathbb{R}^2$.

Example(s)

Consider the vectors $v = (2, 1)$, $u = (1, 3)$. Now, we can construct various lines: $\mathbb{R}v$, $\mathbb{R}u$, $v + \mathbb{R}u$, $u + \mathbb{R}v$ and $\mathbb{R}(v + u)$, for instance.
By the previous discussion, we know that $v + \mathbb{R}u$ and $u + \mathbb{R}v$ don't pass through zero, and all the other lines do.
Here's a sketch of these lines:



with the following coloring: $\mathbb{R}v$, $\mathbb{R}u$, $\mathbb{R}v + u$, $v + \mathbb{R}u$ and $\mathbb{R}(v + u)$.

Example(s)

Now let's resume the discussion from the previous example: Just to refresh, we're considering the function $h : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$h(x, y) := \begin{cases} (x, x), & \text{if } y = 0 \\ (y, y), & \text{otherwise} \end{cases}$$

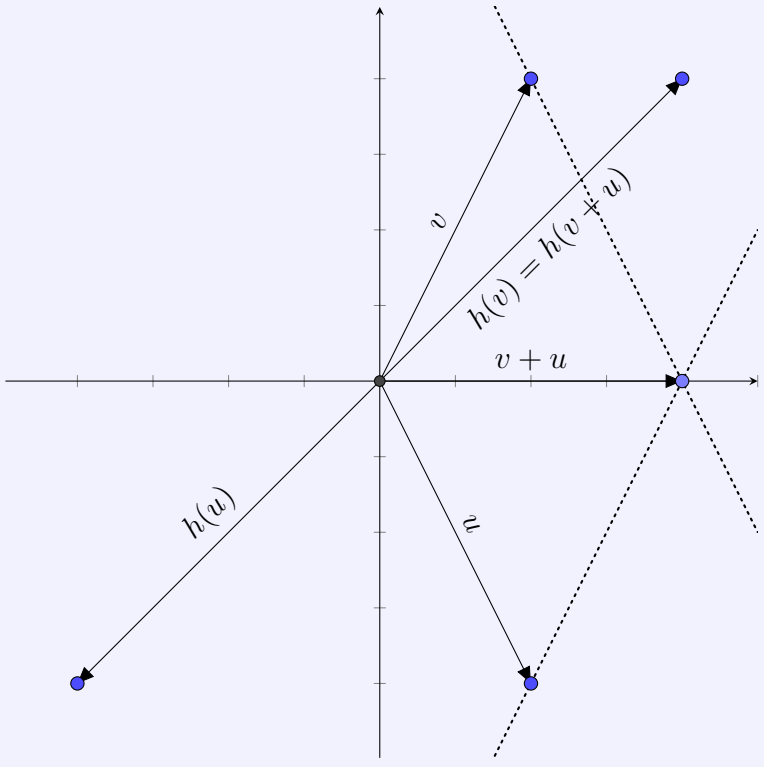
and we've seen that this function does preserve lines through 0. But what does it do with parallelograms?
For instance, take the line points $v = (2, 4)$ and $u = (2, -4)$. Then, by what we've already shown, the point $v + u = (4, 0)$ is precisely the remaining vertex of the parallelogram determined by v and u. But look what happens if we try applying the same rule to h:

$$h(v + u) = h(4, 0) = (4, 4)$$

but

$$h(v) + h(u) = h(2, 4) + h(2, -4) = (4, 4) + (-4, -4) = (0, 0)$$

so h didn't preserve our parallelogram!



This is why h is bad - it is distorting our drawings. We want our functions not only to preserve lines, but also to preserve *parallelograms*. This is the motivation for the following definition, which, I might add, is the single most important definition in this whole chapter.

Definition 2.2.1.4. *A function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is said to be **linear** if for every $v, u \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$ we have:*

- $f(v+u) = f(v) + f(u)$;
- $f(\lambda v) = \lambda f(v)$.

We'll denote the set of all linear functions in $\mathbb{R}^2$ by $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$.

Remark 2.2.1.5

This notation is, and should be, somewhat enigmatic at this point. Just trust us that it's going to make sense further on.
Much further on.

This is precisely the concept we've just built in the example: The first item tells us that a linear function preserves parallelograms, and the second item tells us that a linear function preserves lines.

Example(s)

Finishing up the motivating example, we claim that the function $g : \mathbb{R}^2 \to R^2$ we defined before as being $g(x, y) = (2x - y, x + y)$ is a linear function.
We've already proven, when we first introduced it, that g preserves scalar multiplication. Let us then prove that it preserves summation:

$$\begin{aligned} g((a,b)+(c,d)) &= g(a+c, b+d) \\ &= (2(a+c)-(b+d), (a+c)+(b+d)) \\ &= (2a+2c-b-d, a+c+b+d) \\ &= ((2a-b)+(2c-d), (a+b)+(c+d)) \\ &= (2a-b, a+b)+(2c-d, c+d) = g(a,b)+g(c,d) \end{aligned}$$

for any $(a, b), (c, d) \in \mathbb{R}^2$, and so g preserves summation and scalar multiplication - and is, therefore, a linear transformation.

Example(s)

Let us give some more examples of functions. Let $f, g, h : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $f(x, y) = (x, y)$, $g(x, y) = (y, x)$ and $h(x, y) = (x + y - 4, y)$. Which ones of these are linear? Well,

let's test: For this, fix, once and for all, $(a,b),(c,d) \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$. Then:

$$f((a,b)+(c,d)) = f((a+c,b+d)) = (a+c,b+d) = (a,b)+(c,d) = f(a,b)+f(b,d)$$

and

$$f(\lambda(a,b)) = f(\lambda a, \lambda b) = (\lambda a, \lambda b) = \lambda(a,b) = \lambda f(a,b)$$

so f is indeed linear.

$$g((a,b)+(c,d)) = g(a+c,b+d) = (b+d,a+c) = (b,a)+(d,c) = g(a,b)+g(c,d)$$

and

$$g(\lambda(a,b)) = g(\lambda a, \lambda b) = (\lambda b, \lambda a) = \lambda(b,a) = \lambda g(a,b)$$

so g is indeed linear.
h, however, isn't linear. It actually fails both checks: To see that, assume h *was* linear. Then $h(\lambda(a,b)) = \lambda h(a,b)$ for all $\lambda \in \mathbb{R}$ and $(a,b) \in \mathbb{R}^2$.
In particular, $h(3(1,1)) = h(3,3) = (3+3-4,3) = (2,3)$, while $3h(1,1) = 3(1+1-4,1) = 3(-2,1) = (-6,3)$ and clearly $2 \neq -6$ so they can't be equal.
Similarly, $h((1,1)+(1,1)) = h(2,2) = (2+2-4,2) = (0,2)$, while $h(1,1)+h(1,1) = (1+1-4,1)+(1+1-4,1) = (-4,2)$ and clearly $0 \neq -4$ so they can't be equal.

It would be great then if we could tell at a glance whether a function is linear or not.

Lemma 2.2.1.6. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function. Then f is linear if, and only if, there are real numbers $a,b,c,d \in \mathbb{R}$ such that $f(x,y) = (ax+cy, bx+dy)$ for all $(x,y) \in \mathbb{R}^2$.*

Proof

Let f be linear. We want to show that there exists some real numbers $a,b,c,d \in \mathbb{R}$ such that $f(x,y) = (ax+cy, bx+dy)$.
We can start off by seeing that $(x,y) = x(1,0)+y(0,1)$ and so, since f is linear, we get

$$f(x,y) = f(x(1,0)+y(0,1)) = f(x(1,0))+f(y(0,1)) = xf(1,0)+yf(0,1).$$

But now, $f(1,0), f(0,1) \in \mathbb{R}^2$ and so there exist some real numbers $a,b,c,d \in \mathbb{R}$ such that $f(1,0) = (a,b)$ and $f(0,1) = (c,d)$. This allows us to go back and proceed with calculations:

$$xf(1,0)+yf(0,1) = x(a,b)+y(c,d) = (xa,xb)+(yc,yd) = (xa+yc, xb+yd)$$

and finally we see that, since real number multiplication is commutative, we can conclude that $f(x,y) = (ax+cy, bx+dy)$, as previously stated.

Conversely, let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function defined by $f(x,y) = (ax+cy, bx+dy)$ for all $(x,y) \in \mathbb{R}^2$, for some $a,b,c,d \in \mathbb{R}$ fixed. We claim that such an f is linear.

- $$\begin{aligned} f((x,y)+(x',y')) &= f(x+x', y+y') \\ &= (a(x+x')+c(y+y'), b(x+x')+d(y+y')) \\ &= (ax+ax'+cy+cy', bx+bx'+dy+dy') \\ &= ((ax+cy)+(ax'+cy'), (bx+dy)+(bx'+dy')) \\ &= (ax+cy, bx+dy)+(ax'+cy', bx'+dy') = f(x,y)+f(x',y') \end{aligned}$$

  so f preserves sums.

- $$\begin{aligned} f(\lambda(x,y)) &= f(\lambda x, \lambda y) \\ &= (a(\lambda x)+c(\lambda y), b(\lambda x)+d(\lambda y)) \\ &= (a\lambda x+c\lambda y, b\lambda x+d\lambda y) \\ &= (\lambda(ax+by), \lambda(bx+dy)) \\ &= \lambda(ax+cy, bx+dy) = \lambda f(x,y) \end{aligned}$$

  so f preserves scalar multiplication.

It follows that f is linear, by definition, which ends the proof. □

This proof, however, far from only telling us that all linear functions are of such-and-such form, give us a very powerful tool for dealing with linear functions and vector spaces as a whole.

Re-analyze the proof above. Where did the a, b, c, d come from? They are precisely the images of $(1,0)$ and $(0,1)$ under f. What we've shown, then, to some extent, is that the image of any point under a linear function is entirely determined by the images of $(1,0)$ and $(0,1)$.

Theorem 2.2.1.7. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Then f is uniquely determined by $f(1,0)$ and $f(0,1)$.*

Proof

Basically the same as above: Let's compute the image of (x,y):

$$f(x,y) = f(x(1,0)+y(0,1)) = xf(1,0)+yf(0,1)$$

and so if we call $f(1,0) = v$ and $f(0,1) = u$, we see that for any $(x,y) \in \mathbb{R}^2$, we have that $f(x,y) = xv+yu$.
The result now follows. □

This has the following interesting consequence:

Corollary 2.2.1.8. *Let $f : \{(1,0),(0,1)\} \to \mathbb{R}^2$ be any function. Then there is a unique linear function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\phi(x,y) := xf(1,0)+yf(0,1)$.*

Proof

All that remains is to show that ϕ is indeed linear, but even that is pointless: Since f takes both $(1,0)$ and $(0,1)$ into $\mathbb{R}^2$, let $(a,b) = f(1,0)$ and $(c,d) = f(0,1)$. Then we readily see that $\phi(x,y)$ simply becomes

$$\phi(x,y) = (ax+cy, bx+dy)$$

which we've just shown is a linear function. □

This shows that the vectors $(1,0)$ and $(0,1)$ are very special. If you know what to do with them, you know what to do with literally everyone else.

Definition 2.2.1.9. *We'll denote the vectors* $(1,0)$ *and* $(0,1)$ *by* e_1 *and* e_2*, respectively.*

Definition 2.2.1.10. *A finite set* $X = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^2$ *is called a* ***base of*** $\mathbb{R}^2$ *or a* ***basic set*** *if*

$$\mathrm{Hom}(X, \mathbb{R}^2) \cong \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$$

that is, given any $(a,b) \in \mathbb{R}^2$ *there is a unique choice of* $\lambda_1, \lambda_2, \cdots, \lambda_n \in R$ *such that* $(a,b) = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$*, and therefore* $f(a,b) = \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n)$*.*

Remark 2.2.1.11

It follows trivially from this definition, taking the linear function $\mathrm{id}_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}^2$, that given a base $X = \{x_1, x_2, \cdots, x_n\}$, then for any point $v \in \mathbb{R}^2$ there is a unique choice of $\lambda_1, \lambda_2, \cdots, \lambda_n \in \mathbb{R}$ such that $v = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$.
In other words, every vector can be written uniquely as a linear combination of a fixed base.

Example(s)

We claim that $\{(1,1),(1,-1)\}$ is a base.
To see this, take any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$.
We want to write $f(x,y) = \lambda_1 f(1,1) + \lambda_2 f(1,-1)$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$. We claim that $\lambda_1 := \dfrac{x+y}{2}$ and $\lambda_2 := \dfrac{x-y}{2}$ are the ones we're looking for.
Indeed:

$$\begin{aligned}
\lambda_1 f(1,1) + \lambda_2 f(1,-1) &= \frac{x+y}{2} f(1,1) + \frac{x-y}{2} f(1,-1) \\
&= f\left(\frac{x+y}{2}, \frac{x+y}{2}\right) + f\left(\frac{x-y}{2}, \frac{y-x}{2}\right) \\
&= f\left(\frac{x+y}{2} + \frac{x-y}{2}, \frac{x+y}{2} + \frac{y-x}{2}\right) = f(x,y)
\end{aligned}$$

Assume now that $\lambda_1', \lambda_2' \in \mathbb{R}$ are other such numbers - that is, $f(x,y) = \lambda_1' f(1,1) + \lambda_2' f(1,-1)$.
We need to show that $\lambda_1 = \lambda_1'$ and $\lambda_2 = \lambda_2'$. But:

$$\begin{aligned} f(x,y) &= \lambda_1' f(1,1) + \lambda_2' f(1,-1) \\ &= f(\lambda_1', \lambda_1') + f(\lambda_2', -\lambda_2') \\ &= f(\lambda_1' + \lambda_2', \lambda_1' - \lambda_2') \end{aligned}$$

for all linear functions. In particular, for the identity function we see that

$$(x,y) = (\lambda_1' + \lambda_2', \lambda_1' - \lambda_2')$$

and so $x = \lambda_1' + \lambda_2'$ and $y = \lambda_1' - \lambda_2'$.
Solving this, we see that $\lambda_1' = \dfrac{x+y}{2} = \lambda_1$ and $\lambda_2' = \dfrac{x-y}{2} = \lambda_2$.
This proves that λ_1, λ_2 are indeed unique.
It follows that $\{(1,1),(1,-1)\}$ is a base.

Example(s)

For instance, the vector $(2,3)$ can be written in the base $\{(1,1),(1,-1)\}$ uniquely as $(2,3) = \dfrac{5}{2}(1,1) + \dfrac{-1}{2}(1,-1)$.
Indeed:

$$\frac{5}{2}(1,1) + \frac{-1}{2}(1,-1) = \left(\frac{5}{2}, \frac{5}{2}\right) + \left(\frac{-1}{2}, \frac{1}{2}\right) = \left(\frac{4}{2}, \frac{6}{2}\right) = (2,3).$$

Similarly, $(1,0)$ written in the base $\{(1,1),(1,-1)\}$ is just $(1,0) = \dfrac{1}{2}(1,1) + \dfrac{1}{2}(1,-1)$.

The corollary above is telling us that the set $\{e_1, e_2\}$ is a base of $\mathbb{R}^2$.

Definition 2.2.1.12. *The set $E := \{e_1, e_2\}$ will be called the **canonical base** of $\mathbb{R}^2$.*

Example(s)

Let $f : E \to \mathbb{R}^2$ be defined by $f(e_1) := (3, -\pi)$ and $f(e_2) := (4,0)$. Then there's a unique linear function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $\phi(x,y) := xf(e_1) + yf(e_2)$. Let us compute some images of ϕ:

$$\phi(1,1) = 1f(e_1) + 1f(e_2) = 1(3,-\pi) + 1(4,0) = (3,-\pi) + (4,0) = (7,-\pi)$$

$$\phi(1,0) = 1f(e_1) + 0f(e_2) = 1(3,-\pi) + 0(4,0) = (3,-\pi) = f(e_1)$$

$$\phi(4,5) = 4f(e_1) + 5f(e_2) = 4(3,-\pi) + 5(4,0) = (12,-4\pi) + (20,0) = (32,-4\pi)$$

Note that this is the same as defining $\phi(x,y) = (3x + 4y, -\pi x)$.

So for now, what you need to keep in mind regarding bases is that they are important precisely because of this property.

This will allow us to prove many results about vector spaces by simply stating a parallel result in terms of the base, and then using some previously proven set-theory result.

In order to get to those results, however, we'll need to keep on advancing.

2.2.2 Subspaces

Example(s)

Let $r \subseteq \mathbb{R}^2$ be the line $\mathbb{R}(2,3)$. That is, $v \in r$ if, and only if, $v = \lambda(2,3)$ for some $\lambda \in \mathbb{R}$.
Take then $v, u \in r$. Then, by what we've just said, there are some $\lambda, \mu \in \mathbb{R}$ such that $v = \lambda(2,3)$ and $u = \mu(2,3)$.
We can then ask the question: Is $v+u$ also in r? Well, $v+u = \lambda(2,3)+\mu(2,3) = (\lambda+\mu)(2,3)$, and since both λ, μ are real numbers, so is $\lambda+\mu$. This shows that if $v, u \in r$ then so is $v+u \in r$..
Similarly we can ask: Is $\alpha v \in r$ for any $\alpha \in \mathbb{R}$? Once again, $\alpha v = \alpha(\lambda(2,3)) = (\alpha\lambda)(2,3)$ and since both α and λ are real numbers, so is $\alpha\lambda$. This shows that if $v \in r$, then so is $\alpha v \in r$ for any $\alpha \in \mathbb{R}$.
What this tells us is that lines through zero are **closed** under addition and scaling - that is, if we add two points on the same line, then they remain on that line, and if we scale any point on a line, it also remains there.

Proposition 2.2.2.1. *Let $v \in \mathbb{R}^2$ be any vector, and consider the line $r := \mathbb{R}v$. Take now any $u \in r$. Then $r = \mathbb{R}u$.*

Proof

Let $u \in \mathbb{R}v$ - that is, $u = \mu v$ for some $\mu \in \mathbb{R}$.
Take now $w \in \mathbb{R}v$ - that is, $w = \omega v$ for some $\omega \in \mathbb{R}$. We want to show that $w = \omega' u$, for some $\omega' \in \mathbb{R}$, and therefore $w \in \mathbb{R}u$.

This is simple: Take $\omega' := \frac{\omega}{\mu} \in \mathbb{R}$, which we know is a real number, since both ω and μ are real numbers.
Now

$$\omega' u = \omega'(\mu v) = (\omega'\mu)v = \left(\frac{\omega}{\mu}\mu\right) v = \omega v = w$$

shows that $w \in \mathbb{R}u$.
We've just shown that every point in $\mathbb{R}v$ is also in $\mathbb{R}u$. We can proceed analogously and prove that every point in $\mathbb{R}u$ is also in $\mathbb{R}v$.
By definition of set equality, it follows that $\mathbb{R}v = \mathbb{R}u$, which ends the proof. □

Definition 2.2.2.2. *If $v, u \in \mathbb{R}^2$ are such that $\mathbb{R}v = \mathbb{R}u$ then we say that v **and** u **are parallel vectors**, which shall be denoted as $v \parallel u$.*

Corollary 2.2.2.3. *Two vectors $v, u \in \mathbb{R}^2$ are parallel if, and only if, there's some real number $\lambda \in \mathbb{R}$ such that $v = \lambda u$.*

Example(s)

Let $v = (1, 1)$ and $u = (-3, -3)$. We claim that $v \parallel u$.
That's easy to see, since $u = -3v$.
Now this shows us immediately that the lines $\mathbb{R}v$ and $\mathbb{R}u$ are the same line: Take any $w \in \mathbb{R}v$.
That means that $w = \omega(1, 1) = (\omega, \omega)$ for some $\omega \in \mathbb{R}$.
But now, taking $\omega' := \dfrac{\omega}{-3}$, we see that

$$\omega' u = \left(\frac{\omega}{-3}\right) u = \omega \left(\frac{u}{-3}\right) = \omega v = w$$

and we see that $w \in \mathbb{R}u$, just as stated.

We can now ask the following question: We know that lines through zero are closed under addition and scaling. Are there any other subsets of $\mathbb{R}^2$ that are like that? If so, what are they? If not, then why?

Let us give a name to that property, because it's a mouthful:

Definition 2.2.2.4. *Let $X \subseteq \mathbb{R}^2$ be a set satisfying*

- *For all $x, y \in X$, $x + y \in X$;*
- *For all $x \in X$ and $\lambda \in \mathbb{R}$, $\lambda x \in X$.*

*Then X will be called a **subspace of** $\mathbb{R}^2$ which will be denoted as $X \leq \mathbb{R}^2$.*

Example(s)

Clearly, since lines through zero are the motivating example for this definition, we know that for every vector $v \in \mathbb{R}^2$, the line $\mathbb{R}v$ is a subspace of $\mathbb{R}^2$ - that is, $\mathbb{R}v \leq \mathbb{R}^2$.
But, if you think about it, $\mathbb{R}^2 \subseteq \mathbb{R}^2$ is a subset of $\mathbb{R}^2$ which is also closed under additions and scalar multiplications (duh). So $\mathbb{R}^2 \leq \mathbb{R}^2$.

Conversely, **any** line which doesn't contain zero cannot be a subspace: To see this, take any two vectors $v, u \in \mathbb{R}^2$ and consider the line $\mathbb{R}v + u$, which does not contain zero.
Take now $w \in \mathbb{R}v + u$ and any other real number - for instance, 0.
If $\mathbb{R}v+u$ was a subspace, it would be closed under scalar multiplication, but $0w = 0 \notin \mathbb{R}v+u$, so it cannot be a subspace.

What else then can be a subspace?

This example gives us a very nice idea for how to approach subspaces:

Proposition 2.2.2.5. *Let $X \subseteq \mathbb{R}^2$ be a subspace. Then $0 \in X$.*

Proof

This can be seen in many ways.
Take $x \in X$. Since X is subspace it is closed under scalar multiplication, so $-x \in X$.
But since X is a subspace it is also closed under addition, so $x + (-x) \in X$.
But $x + (-x) = 0$ so $0 \in X$, as claimed. □

Now let's try building new subspaces from existing ones:

Lemma 2.2.2.6. *If $X, Y \subseteq \mathbb{R}^2$ are two subspaces, then $X \cap Y \subseteq \mathbb{R}^2$ is also a subspace.*

Proof

This is easy to see:

- Take $v, u \in X \cap Y$. This means that $v, u \in X$ and $v, u \in Y$, by definition of intersection. But since $v, u \in X$ and X is a subspace, then $v + u \in X$. Similarly, since $v, u \in Y$ and Y is a subspace, then $v + u \in Y$.

  Finally, using once more the definition of intersection, we see that $v + u \in X$ and $v + u \in Y$ implies $v + u \in X \cap Y$, and so $X \cap Y$ is closed under addition.

- Take $v \in X \cap Y$ and $\lambda \in \mathbb{R}$. Once again, $v \in X \cap Y$ implies $v \in X$ and $v \in Y$, and since both are subspaces, this implies $\lambda v \in X$ and $\lambda v \in Y$.

  Finally, using once more the definition of intersection, we see that $\lambda v \in X$ and $\lambda v \in Y$ implies $\lambda v \in X \cap Y$, and so $X \cap Y$ is closed under scalar multiplication.

This shows that $X \cap Y$ is a subspace, which ends the proof. □

With this we can already infer an important result:

Corollary 2.2.2.7. *The set containing only the origin (called the **zero** set $0 := \{(0, 0)\}$) is a subspace.*

We could check this directly, but that would imply that we even considered it a possible candidate for subspace before. Following this result, however, since lines through zero are subspaces, and since any two of those lines meet precisely at zero, we know, for free, that zero is also a subspace.

Now, I know what you're thinking: If intersection of subspaces is a subspace, then surely the union of subspaces is as well, right?
Well...

Example(s)

Consider the lines $\mathbb{X}$ (that is, $\mathbb{R}e_1$) and $\mathbb{Y}$ (that is, $\mathbb{R}e_2$). Then $\mathbb{X} \cup \mathbb{Y}$ is just the two axis together.
It clearly is closed under scalar multiplication, but, sadly, it is not closed under addition. This is very easy to see:

For instance, $e_1 \in \mathbb{X}$ and $e_2 \in \mathbb{Y}$, so $e_1, e_2 \in \mathbb{X} \cup \mathbb{Y}$, by definition of set union.
If $\mathbb{X} \cup \mathbb{Y}$ was a subspace, then $e_1 + e_2 \in \mathbb{X} \cup \mathbb{Y}$, but $e_1 + e_2 = (1,1)$ and we know that $(1,1) \notin \mathbb{X}$ and $(1,1) \notin \mathbb{Y}$ and, therefore, $(1,1) \notin \mathbb{X} \cup \mathbb{Y}$.
This shows that, in general, union of subspaces is **not** a subspace.

With this we see that the set-theoretical notion of union is *just not good enough* for vector spaces. We need a "vectorial" notion of union:

Definition 2.2.2.8. *Let $X, Y \subseteq \mathbb{R}^2$ be two subspaces. We define the **sum of X and Y** to be the subspace $X + Y$ given by:*

- *$X + Y$ contains both X and Y;*
- *Any other subspace Z that contains both X and Y also contains $X + Y$.*

Proposition 2.2.2.9. *Given any two subspaces $X, Y \subseteq \mathbb{R}^2$, then their sum can be uniquely expressed as*

$$X + Y := \{v \in \mathbb{R}^2 \mid \exists x \in X, \exists y \in Y \text{ such that } v = x + y\},$$

that is, $X + Y$ is the set of all sums of all elements in both X and Y.

Proof

This proof consists of two parts: First, we need to show that this set, let's call it S, is indeed a subspace, and then, second, we need to prove that it satisfies the two defining properties of the sum of X and Y.
So, just like we said, define

$$S := \{v \in \mathbb{R}^2 \mid \exists x \in X, \exists y \in Y : v = x + y\}.$$

Let us show that $S \leq \mathbb{R}^2$.
To do that, take any $v, v' \in S$ and $\lambda \in \mathbb{R}$. Then, by definition, there are some $x, x' \in X$ and some $y, y' \in Y$ such that $v = x + y$ and $v' = x' + y'$. With this, we can compute:

$$v + v' = (x + y) + (x' + y') = (x + x') + (y + y')$$

and since both X and Y are subspaces, $x + x' \in X$ and $y + y' \in Y$. This shows that $v + v'$ is the sum of $x + x' \in X$ and $y + y' \in Y$, and hence $v + v' \in S$.
Similarly,

$$\lambda v = \lambda(x + y) = \lambda x + \lambda y$$

and, once more, since X and Y are subspaces, $\lambda x \in X$ and $\lambda y \in Y$, and, so, λv is the sum of $\lambda x \in X$ and $\lambda y \in Y$. This shows that $\lambda v \in S$.
This shows that S is indeed a subspace.

Finally, let us show that it satisfies the definition of sum:

- First, it is easy to see that it contains both X and Y: Since X and Y are subspaces, they contain 0. So for every $x \in X$, $x + 0 \in S$, by definition of S. But $x + 0 = x \in X$. So $X \subseteq S$.

  Similarly, for every $y \in Y$, $0 + y \in S$, by definition of S. But $0 + y = y \in Y$, so $Y \subseteq S$.

- Let Z be another subspace that contains both X and Y.

  Take $v \in S$. By definition of S, there are some $x \in X$ and some $y \in Y$ such that $v = x + y$. But since Z is a subspace that contains X and Y, this means that $x \in Z$ and $y \in Z$ - and so $v = x + y \in Z$.

  We have just shown that any element $v \in S$ is also in Z, which means that $S \subseteq Z$.

This proves that S satisfies the definition of sum and, therefore, $S = X + Y$, just as we wanted to show. □

Example(s)

Let's use this to build new subspaces!
Take the line $\mathbb{X}$ and the line $\mathbb{Y}$, once again. What is $\mathbb{X} + \mathbb{Y}$? Well, by what we just did, it's the set of all sums of points in $\mathbb{X}$ and $\mathbb{Y}$...
But this is the whole $\mathbb{R}^2$!
To see this, take any $v \in \mathbb{R}^2$. Then $v = (x, y)$, for some $x, y \in \mathbb{R}$, and we can rewrite this as $xe_1 + ye_2$. But since $\mathbb{X}$ and $\mathbb{Y}$ are subspaces, $xe_1 \in \mathbb{X}$ and $ye_2 \in \mathbb{Y}$, so $v \in \mathbb{X} + \mathbb{Y}$.
This shows that $\mathbb{X} + \mathbb{Y} \subseteq \mathbb{R}^2 \subseteq \mathbb{X} + \mathbb{Y}$. That can only mean one thing: $\mathbb{X} + \mathbb{Y} = \mathbb{R}^2$.

What this example shows us is another weird behavior of subspaces: They grow in huge steps at a time:

Lemma 2.2.2.10. *For any two different lines through zero $r, s \subseteq \mathbb{R}^2$, we have that $r + s = \mathbb{R}^2$.*

Proof

It's similar to the example's: Let $r = \mathbb{R}v$ and $s = \mathbb{R}u$ for any two non-parallel vectors $v, u \in \mathbb{R}^2$ (if they were parallel, by our previous discussions, $r = s$ and we don't want that). We'll show that $r + s$ contains both e_1 and e_2. This suffices, because we've already shown that $\mathbb{X} + \mathbb{Y} = \mathbb{R}^2$.
To do that, write $v = (v_1, v_2)$ and $u = (u_1, u_2)$. Now, defining $\mu_1 := \frac{v_2}{u_2} \in \mathbb{R}$ we see that

$$\mu_1 u = \frac{v_2}{u_2}(u_1, u_2) = \left(\frac{v_2}{u_2}u_1, \frac{v_2}{u_2}u_2\right) = \left(\frac{v_2}{u_2}u_1, v_2\right)$$

and so

$$v - \mu_1 u = (v_1, v_2) - \left(\frac{v_2}{u_2}u_1, v_2\right) = \left(v_1 - \frac{v_2}{u_2}u_1, 0\right),$$

that is, $v - \mu_1 u \in \mathbb{R}e_1$. So we can just scale it up or down to obtain e_1. To do that, let us

rewrite it a little bit:

$$v - \mu_1 u = \left(v_1 - \frac{v_2}{u_2}u_1, 0\right) = \left(\frac{v_1u_2 - v_2u_1}{u_2}, 0\right)$$

and see that if we multiply it by $\mu_2 := \dfrac{u_2}{v_1u_2 - v_2u_1}$ we get

$$\mu_2(v - \mu_1 u) = \frac{u_2}{v_1u_2 - v_2u_1}\left(\frac{v_1u_2 - v_2u_1}{u_2}, 0\right) = (1, 0) = e_1.$$

Now we can just argue: Since $v \in r$ and $\mu_1 u \in s$, we get that $v - \mu_1 u \in r + s$. But since $r + s$ is a subspace, we see that $\mu_2(v - \mu_1 u) \in r + s$, which shows that e_1 (and all its scalar multiples) is also in $r + s$. This proves that $\mathbb{X} \subseteq r + s$.

Let us now show that $e_2 \in r + s$: It's essentially the same construction:
Define $\lambda_1 := \dfrac{u_1}{v_1} \in \mathbb{R}$, which allows us to do $\lambda_1 v = \left(u_1, \dfrac{u_1}{v_1}v_2\right)$ and so $\lambda_1 v - u = \left(0, \dfrac{u_1}{v_1}v_2 - u_2\right)$.
Now we rewrite that as $\lambda_1 v - u = \left(0, \dfrac{u_1v_2 - u_2v_1}{v_1}\right)$to see that by taking $\lambda_2 := \dfrac{v_1}{u_1v_2 - u_2v_1}$ we can do $\lambda_2(\lambda_1 v - u) = e_2$.
Now, since $\lambda_1 v \in r$ and $u \in s$, we see that $\lambda_1 v - u \in r + s$, and so, since it is a subspace, $\lambda_2(\lambda_1 v - u) \in r + s$. This shows that e_2 (and all its scalar multiples) is also in $r + s$, and so, $\mathbb{Y} \subseteq r + s$.

Finally, this shows that $r + s$ contains both $\mathbb{X}$ and $\mathbb{Y}$, and so, by, definition of sum, it contains $\mathbb{X} + \mathbb{Y}$. But we already know $\mathbb{X} + \mathbb{Y} = \mathbb{R}^2$. So we get $r + s \subseteq \mathbb{R}^2 \subseteq r + s$, which finally shows that $r + s = \mathbb{R}^2$.
This finishes the proof. □

Finally, we can prove:

Theorem 2.2.2.11. *Let $X \subseteq \mathbb{R}^2$ be a subspace. Then X is either 0, a line through zero, or $\mathbb{R}^2$.*

Proof

We already know that every subspace contains 0.
If $X \setminus \{0\} = \varnothing$, then $X = 0$.

If not, then there's some $x \in X$ with $x \neq 0$. But since X is a subspace, every multiple of x is also in X. In particular, $\mathbb{R}x \subseteq X$. This shows that X contains a line through zero.
If $X \setminus \mathbb{R}x = \varnothing$, then $X = \mathbb{R}x$ and X is a line through zero.

If not, then there's some $y \in X \setminus \mathbb{R}x$ with $y \neq 0$. But since X is a subspace, every multiple

of y is also in X. In particular, $\mathbb{R}y \subseteq X$. Also, the fact that we chose y outside of the line $\mathbb{R}x$ implies that $y \nparallel x$. This means that $\mathbb{R}x \neq \mathbb{R}y$.
But this means that X contains both $\mathbb{R}x$ and $\mathbb{R}y$, and so, by definition of sum, contains $\mathbb{R}x + \mathbb{R}y$. But we know that $\mathbb{R}x + \mathbb{R}y = \mathbb{R}^2$.
This shows that $X \subseteq \mathbb{R}^2 \subseteq X$, and thus $X = \mathbb{R}^2$.
This ends the proof. □

So now we know that there aren't many choices of subspaces for $\mathbb{R}^2$: There's only the origin, lines through the origin and the whole set.

2.2.3 Spanning sets and linear dependency

This new section that we're starting right now doesn't really make sense for $\mathbb{R}^2$ and would be better-suited for studying general vector spaces. However, since it is usually taught in any introductory text on the subject, I guess I'll have to comply.

Definition 2.2.3.1. *Let $v \in \mathbb{R}^2$ be any vector. We define the **subspace spanned by** v to be the subspace* $\text{span}\{v\}$ *(sometimes $\langle v \rangle$) defined by*

$$\text{span}\{v\} := \mathbb{R}v.$$

*Analogously, given any finite collection of vectors $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ we define the **subspace spanned by** $v_1, v_2, \cdots, v_n$ to be the subspace* $\text{span}\{v_1, v_2, \cdots, v_n\}$ *(sometimes $\langle v_1, v_2, \cdots, v_n \rangle$) defined by*

$$\text{span}\{v_1, v_2, \cdots, v_n\} := \mathbb{R}v_1 + \mathbb{R}v_2 + \cdots + \mathbb{R}v_n.$$

Why is this a stupid definition at this point? Well...

As we've already seen, $\mathbb{R}^2$ only has three kinds of subspaces: zero, lines through zero and $\mathbb{R}^2$ itself. Since the subspace spanned by a vector (or a collection of vectors) is a subspace (duh) it has to be one of those three.

This means that the whole study of spanning and subspaces comes down to classifying what kinds of collections of vectors span each kind of subspace.

Two of them are trivial:

Proposition 2.2.3.2. *Take $v \in \mathbb{R}^2$. Then* $\text{span}\{v\} = 0$ *if, and only if, $v = 0$.*

Proof

If $v = 0$ then clearly, $\lambda v = 0$ for all $\lambda \in \mathbb{R}$, so $\text{span}\{v\} = 0$.
Conversely, if $\text{span}\{v\} = 0$, since $v \in \text{span}\{v\}$ (by taking $\lambda = 1$) we can conclude that $v = 0$. □

Proposition 2.2.3.3. *Take $v \in \mathbb{R}^2$. Then* $\text{span}\{v\}$ *is a line if, and only if, $v \neq 0$.*

Proof

If $\text{span}\{v\}$ is a line, then $v \neq 0$ by the preceding proposition.
Conversely, if $v \neq 0$, then $\text{span}\{v\} = \mathbb{R}v$, by definition. So it is a line. □

There are still some cases left to classify, however:

- Is there some collection $\{v_1, v_2, \cdots, v_n\}$ which spans 0?
- Is there some collection $\{v_1, v_2, \cdots, v_n\}$ which spans a line?
- Is there some collection $\{v_1, v_2, \cdots, v_n\}$ which spans $\mathbb{R}^2$?

And you might be tempted to say that any collection of two non-null vectors spans $\mathbb{R}^2$. But that's not the case:

Example(s)

Let $v = (1, 1)$, $u = (-3, -3)$. We claim that $\text{span}\{v, u\} = \text{span}\{v\} = \mathbb{R}v$.
To see that, take any $w \in \text{span}\{v, u\}$ - that is, there are two real numbers $\lambda, \mu \in \mathbb{R}$ such that $w = \lambda v + \mu u$. In other words,

$$w = \lambda(1, 1) + \mu(-3, -3) = (\lambda, \lambda) + (-3\mu, -3\mu) = (\lambda - 3\mu, \lambda - 3\mu)$$

and we see that $w = (\lambda - 3\mu)v$, and so $w \in \text{span}\{v\}$. This shows $\text{span}\{v, u\} \subseteq \text{span}\{v\}$.
Conversely, take $w \in \text{span}\{v\}$ - that is, there is some $\lambda \in \mathbb{R}$ such that $w = \lambda v$. But that, $w = \lambda v + 0u$ shows that $w \in \text{span}\{v, w\}$. Therefore, $\text{span}\{v\} \subseteq \text{span}\{v, u\}$.
We have shown that $\text{span}\{v\} = \text{span}\{v, u\}$, even though $v \neq u$.

This kind of situation gives rise to an awful and useless definition, but very helpful when dealing with real-life applications of linear algebra:

Definition 2.2.3.4. *Let $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ be a collection of vectors in $\mathbb{R}^2$. We say that this collection is **linearly dependent** if one of the vectors in the collection is spanned by the other vectors.*

In symbols, there's some $i \leq n$ and some real numbers $\lambda_1, \lambda_2, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_n$ such that

$$v_i = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_{i-1} v_{i-1} + \lambda_{i+1} v_{i+1} + \cdots + \lambda_n v_n.$$

*Analogously, we say that the collection is **linearly independent** if it's not linearly dependent.*

See what I said about this area being useless? This mostly appears when doing linear algebra the **wrong way** - that is, by being focused on finding solutions to linear systems and doing matrix stuffs.

Why is it stupid? Well, because we already have a better language to describe the same things!

Lemma 2.2.3.5. *Take $v, u \in \mathbb{R}^2$ any two vectors. Then $\text{span}\{v, u\} = \mathbb{R}v + \mathbb{R}u$.*

Proof

There's literally nothing to be proven here. They are equal by definition. □

Corollary 2.2.3.6. *Let $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ be any collection of vectors. Then $\text{span}\{v_1, v_2, \cdots, v_n\} = \mathbb{R}v_1 + \mathbb{R}v_2 + \cdots + \mathbb{R}v_n$.*

Lemma 2.2.3.7. *Let $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ be any collection of vectors. Then it is linearly dependent if, and only if, there is some $i \leq n$ such that*

$$\mathbb{R}v_i \subseteq \mathbb{R}v_1 + \mathbb{R}v_2 + \cdots + \mathbb{R}v_{i-1} + \mathbb{R}v_{i+1} + \cdots + \mathbb{R}v_n.$$

Proof

Once again, there's nothing to be done here, since they are equal by definition. □

Corollary 2.2.3.8. *Two vectors $v, u \in \mathbb{R}^2$ are linearly independent if, and only if, $v \nparallel u$.*

But we've already proven that if we take any two non-parallel vectors then they span the whole $\mathbb{R}^2$. That means that if we take any three non-parallel vectors, then any one of them is already spanned by the other two. From this we can conclude two results:

Proposition 2.2.3.9. *Let $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ be any collection of vectors. If it spans $\mathbb{R}^2$, then $n \geq 2$.*

Proposition 2.2.3.10. *Let $\{v_1, v_2, \cdots, v_n\} \subseteq \mathbb{R}^2$ be any collection of vectors. If it is linearly independent, then $n \leq 2$.*

Finally, there's one last thing to prove here:

Theorem 2.2.3.11. *Take $X \subseteq \mathbb{R}^2$ any subset. Then the following conditions are equivalent:*

(a) X is a base;

(b) X spans $\mathbb{R}^2$ and has 2 elements;

(c) X spans $\mathbb{R}^2$ and is linearly independent;

(d) X has 2 elements and is linearly independent.

Proof

The two preceding propositions give us, for free, that (c) implies both (b) and (d): If X is linearly independent it has at most 2 elements, and if it spans it has at least 2 elements. Since it satisfies both, it must have precisely 2 elements.

- (b) implies (c):

  Let $X = \{x_1, x_2\}$ since it only has two elements. Assume they're not linearly independent. Then, $x_1 = \lambda x_2$ for some real number λ, and so $\mathbb{R}x_1 + \mathbb{R}x_2 = \mathbb{R}x_1$, so the sum is a line.

  However, we're assuming that X spans $\mathbb{R}^2$ - that is, $\mathbb{R}^2 = \mathbb{R}X = \mathbb{R}x_1 + \mathbb{R}x_2$. So we would have that $\mathbb{R}^2 = \mathbb{R}x_1$ - in other words, that the whole plane is just a line, which is clearly absurd.

  Therefore, X has to be linearly independent.

- (d) implies (c):

  Let $X = \{x_1, x_2\}$ since it has only two elements. Since they're linearly independent, we see that $x_1 \nparallel x_2$ and so $\mathbb{R}x_1 + \mathbb{R}x_2 = \mathbb{R}^2$, X spans $\mathbb{R}^2$.

This shows that (b), (c) and (d) are equivalent.
Finally, let us prove that (a) is equivalent to the other three:

- (a) implies (c):

  Let X be a base, and take the linear function $\mathrm{id}_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}^2$. Since X is a base, we know that for all $v \in \mathbb{R}^2$, $\mathrm{id}_{\mathbb{R}^2}(v) = \lambda_1 \mathrm{id}_{\mathbb{R}^2}(x_1) + \lambda_2 \mathrm{id}_{\mathbb{R}^2}(x_2) + \cdots + \lambda_n \mathrm{id}_{\mathbb{R}^2}(x_n)$.

  But $\mathrm{id}_{\mathbb{R}^2}$ is the identity function, so this becomes

  $$v = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$$

  which tells us that any $v \in \mathbb{R}^2$ is spanned by X - in other words, X spans $\mathbb{R}^2$.

  On the other hand, assume $x_i \parallel x_j$ for some $i \neq j$. Then, $x_j = \mu_{i,j} x_j$, by definition. But, $x_i = 1x_i$, so there's two ways of writing x_i as a linear combination of the base X. This cannot happen, by definition of base. So $x_i \nparallel x_j$, no matter which $i \neq j$ we start with.

  This means that we can take any two (say, x_1, x_2) and the others will be spanned by those two. But if $n > 2$ that's also a problem: For instance, take x_3. Since it is spanned by x_1, x_2, there are $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $x_3 = \lambda_1 x_1 + \lambda_2 x_2$. But we already know that the only possible way to write x_3 in terms of the base is $x_3 = 1x_3$ (by the same reasoning as above). So if there were more than two elements in X we'd have a contradiction.

  This allows us to finally conclude that $X = \{x_1, x_2\}$ and so it is linearly independent and spans.

- (c) implies (a):

  Let $X = \{x_1, x_2\}$ be a linearly independent spanning set for $\mathbb{R}^2$ (we already know it has 2 elements) and any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Then, since X spans $\mathbb{R}^2$ and is linearly independent, we see that any vector $v \in \mathbb{R}^2$ is of the form $v = \lambda_1 x_1 + \lambda_2 x_2$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$.

  But this tells us that $f(v) = f(\lambda_1 x_1 + \lambda_2 x_2)$ and since f is linear, this is simply $f(v) = \lambda_1 f(x_1) + \lambda_2 f(x_2)$.

  This tells us that X is a base, since any linear function is completely determined by X.

This shows that (a) and (c) are equivalent, and since (c) is already equivalent to (b) and (d) this suffices and ends the proof. □

2.2.4 Back to linearity

Now that we have the very needed tools of subspaces and spanning sets, let's move forward with our discussions on linear functions.

Proposition 2.2.4.1. *For any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$, $\operatorname{Im} f$ is a subspace.*

Proof

Take $v, u \in \operatorname{Im} f$. That means that $v = f(v')$ and $u = f(u')$ for some $v', u' \in \mathbb{R}^2$. Then

$$f(u' + v') = f(u') + f(v') = u + v$$

and therefore $u + v$ is also in the image of f (since it is the image of $u' + v'$ under f).
Similarly, given any scalar $\lambda \in \mathbb{R}$ we have that

$$f(\lambda v') = \lambda f(v') = \lambda v$$

and so λv is also in the image of f (since it is the image of $\lambda v'$ under f).
Since $\operatorname{Im} f$ is closed under sums and scalar multiplications, it must be a subspace. □

Corollary 2.2.4.2. *For any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$ its image is either zero, a line through zero or the whole set.*

Example(s)

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function given by $f(x, y) = (2x - y, 5y)$. What is, then, $\operatorname{Im} f$?
Well, it cannot be zero, since $f(0, 1) = (2 \cdot 0 - 1, 5 \cdot 1) = (-1, 5) \neq 0$.
So it can only be the line $\mathbb{R}(-1, 5)$ (since we already know $(-1, 5)$ is in the image) or the whole set $\mathbb{R}^2$.
But it's easy to see that $f(1, 1) = (2 \cdot 1 - 1, 5 \cdot 1) = (1, 5)$ and $(1, 5) \notin \mathbb{R}(-1, 5)$ so there are two non-parallel vectors in the image. It follows that the image of f has the be the whole set $\mathbb{R}^2$.

Consider now the linear function $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by $g(x, y) = (x - y, x - y)$. What is the image of g?
Once again, it can't be zero: $g(1, 0) = (1 - 0, 1 - 0) = (1, 1) \neq 0$.
Therefore $\operatorname{Im} g$ is either the line $\mathbb{R}(1, 1)$ or $\mathbb{R}^2$.
Now suppose $\operatorname{Im} g = \mathbb{R}^2$. That means that, for instance, $(1, 0) \in \operatorname{Im} g$. But this would mean that there's some $(a, b) \in \mathbb{R}^2$ such that $g(a, b) = (1, 0)$, by definition of g.
But doing computations it's easy to convince ourselves that is impossible: If it were possible, we'd have $g(a, b) = (a - b, a - b) = (1, 0)$ and so $a - b = 1$ and $a - b = 0$, which is a contradiction. A number can't equal both 1 and 0 at the same time!
Therefore, $(1, 0) \notin \mathbb{R}^2$. This already tells us that $\operatorname{Im} g$ is a line (because it's either a line or the whole plane).

Taking one step further, it's not hard to convince yourself that a point (x, y) lies in the image of g if, and only if, $x = y$.
That is, the image is **not** the whole set $\mathbb{R}^2$, but only the line $\mathbb{R}(1, 1)$.

We'll now make a somewhat arbitrary definition, but one that's going to make a **lot** of sense the more of you think about it.

Definition 2.2.4.3. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be any linear function. We define its **kernel** to be the set* $\operatorname{Ker} f$ *given by*

$$\operatorname{Ker} f := \{v \in \mathbb{R}^2 \mid f(v) = 0\}.$$

In other words, the kernel of a linear function is the set of points that is killed by it.

Example(s)

Expanding on the example above, let's compute $\operatorname{Ker} f$ and $\operatorname{Ker} g$:
Take $(a, b) \in \mathbb{R}^2$ and assume $f(a, b) = 0$. What can we say about a and b? Well, by definition of f, $f(a, b) = (2a - b, 5b)$ so if that equals 0 we must have $2a - b = 0$ and $5b = 0$.
The second equation alone tells us that $b = 0$. Now armed with that, the first equation becomes $2a = 0$ and so $a = 0$ too.
This means that for a point (a, b) to go to 0 under f, we must have $a = b = 0$. That means that $\operatorname{Ker} f = 0$.

Analogously, take $(a, b) \in \mathbb{R}^2$ and assume that $g(a, b) = 0$. But this is just $(a - b, a - b) = 0$, which is a single equation: $a = b$.
Therefore, the only way for a point $(a, b) \in \mathbb{R}^2$ to be taken to 0 by g is for a and b to be equal - for instance, $(1, 1)$ or $(-3, -3)$. This means that $\operatorname{Ker} g = \mathbb{R}(1, 1)$.

Lemma 2.2.4.4. *For any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$, we have that* $\operatorname{Ker} f$ *is a subspace.*

Proof

Take $v, u \in \operatorname{Ker} f$. This means that $f(v) = f(u) = 0$, by definition of kernel. But then, since f is linear, we have:

$$f(v + u) = f(v) + f(u) = 0 + 0 = 0$$

and so $v + u \in \operatorname{Ker} f$ since its image under f is 0.
Similarly, given any $\lambda \in \mathbb{R}$ we can easily see that

$$f(\lambda v) = \lambda f(v) = \lambda \cdot 0 = 0$$

and so $\lambda v \in \operatorname{Ker} f$ since its image under f is 0.
This shows that $\operatorname{Ker} f$ is closed under addition and scalar multiplication, and is, therefore, a subspace. □

Corollary 2.2.4.5. *For any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$, we have that* $\operatorname{Ker} f$ *is either zero, a line through zero or the whole plane* $\mathbb{R}^2$.

Now we're gonna present one of the main reasons why kernels are important: They tell us when functions are injective:

Proposition 2.2.4.6. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Then f is injective if, and only if,* $\operatorname{Ker} f = 0$.

Proof

Assume f is injective and take $v \in \operatorname{Ker} f$. This means that $f(v) = 0$. But linear functions always take 0 to 0 - that is, $f(0) = 0$.
Since f is injective, we can't have two different points in the domain with the same image under f. This means that $v = 0$. With this, we can prove that any point in the kernel is just 0, so $\operatorname{Ker} f = 0$.

Conversely, assume $\operatorname{Ker} f = 0$ and take two points $v, u \in \mathbb{R}^2$ such that $f(v) = f(u)$. We want to show that $v = u$.
But f is linear, so $f(v) = f(u)$ implies $f(v) - f(u) = 0$ - that is, $f(v - u) = 0$. This shows that $v - u \in \operatorname{Ker} f$.
But we're assuming that $\operatorname{Ker} f = 0$. This means that $v - u = 0$, and therefore $v = u$ and f is injective. □

So from now on, instead of checking directly whether or not a function is injective, we can just use kernels.

Example(s)

One final word about the preceding examples: We've already shown that $\operatorname{Ker} f = 0$ and $\operatorname{Ker} g = \mathbb{R}(1, 1)$. This shows, then, that f is injective and g is not injective.

Finally, we're ready to work with isomorphisms.

Definition 2.2.4.7. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function. We'll say that f* ***is a linear isomorphism*** *(or just an isomorphism) if it is a set isomorphism (that is, a bijection) which is also linear.*

Lemma 2.2.4.8. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear isomorphism. Then $f^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ is also a linear isomorphism.*

Proof

We already know that if f is an isomorphism, then f^{-1} is also an isomorphism, since they're mutually inverse to each other.
It suffices to prove, then, that f^{-1} is linear when f is linear.
Take $v, u \in \mathbb{R}^2$. Since f is an isomorphism, it is also a surjection. This means that $v = f(v')$

and $u = f(u')$ for some $v', u' \in \mathbb{R}^2$. Then:

$$f^{-1}(v+u) = f^{-1}(f(v')+f(u')) = f^{-1}(f(v'+u')) = (f^{-1}\circ f)(v'+u') = v'+u' = f^{-1}(v)+f^{-1}(u)$$

so f^{-1} preserves sums.
Analogously, for any scalar $\lambda \in \mathbb{R}^2$ we have

$$f^{-1}(\lambda v) = f^{-1}(\lambda f(v')) = f^{-1}(f(\lambda v')) = (f^{-1} \circ f)(\lambda v') = \lambda v' = \lambda f^{-1}(v)$$

and so f^{-1} preserves scalar multiplication.
It follows that f^{-1} is indeed linear, just as we stated, which finishes the proof. □

Proposition 2.2.4.9. *Let $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ be linear functions. Then both of $f \circ g$ and $g \circ f$ are also linear functions.*

Proof

We'll only show one of them, the other is identical.
Take $v, u \in \mathbb{R}^2$. Then:

$$(f \circ g)(v + u) = f(g(v + u)) = f(g(v) + g(u)) = f(g(v)) + f(g(u)) = (f \circ g)(v) + (f \circ g)(u)$$

so $f \circ g$ preserves sums.
Analogously, taking $\lambda \in \mathbb{R}$:

$$(f \circ g)(\lambda v) = f(g(\lambda v)) = f(\lambda g(v)) = \lambda f(g(v)) = \lambda(f \circ g)(v)$$

so $f \circ g$ preserves scalar multiplication.
It follows then that $f \circ g$ is linear, which ends the proof. □

Corollary 2.2.4.10. *Let $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ be two linear isomorphisms. Then both $f \circ g$ and $g \circ f$ are also linear isomorphisms.*

And finally, a surprising result:

Theorem 2.2.4.11. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Then the following are equivalent:*

(a) f is a linear isomorphism;

(b) f is injective;

(c) f is surjective.

Proof

Clearly (a) implies both (b) and (c), by definition.

- (b) implies (a):

  Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear injection and choose $X = \{x_1, x_2\} \subseteq \mathbb{R}^2$ base. Since f is injective, $f(x_1) \neq f(x_2)$.

  We claim that $f(x_1) \nparallel f(x_2)$. This would suffice, because then we'd have to non-parallel vectors in the image and, therefore, by the previous section, this is sufficient to span any other vector - that is, f would be surjective (and since it's already injective, it would be a bijection).

  To see then that $f(x_1) \nparallel f(x_2)$, assume they **are** parallel - that is, $f(x_1) = \lambda f(x_2)$ for some $\lambda \in \mathbb{R}$. But $f(\lambda x_2) = \lambda f(x_2)$ as well. Since f is injective, $f(x_1) = f(\lambda x_2)$ must then imply $x_1 = \lambda x_2$ - that is, $x_1 \parallel x_2$. But we've already proven that no two vectors in a base can be parallel.

  This is a contradiction! Therefore, we **cannot** have that $f(x_1) \parallel f(x_2)$, and the result follows by our previous discussion.

- (c) implies (a):

  Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear surjection and choose $v, u \in \mathbb{R}^2$ such that $v \nparallel u$. Since f is surjective, there's some $v', u' \in \mathbb{R}^2$ such that $f(v') = v$ and $f(u') = u$. Notice that this implies $v' \nparallel u'$ (cause otherwise $v' = \lambda u'$ would imply $v = \lambda u$, which cannot happen since we're taking $v \nparallel u$).

  So $\{v, u\}$ is a base of $\mathbb{R}^2$. This means that any other vector $w \in \mathbb{R}^2$ can be written in a unique way as $w = \lambda_1 v + \lambda_2 u$.

  In particular, 0 can always be written as $0 = 0v + 0u$, so that must be its unique way.

  Take then $z \in \operatorname{Ker} f$. Then $f(z) = 0$, by definition of kernel. But since $\{v', u'\}$ is also a base, we know, by definition of base, that we can write $z = \mu_1 v' + \mu_2 u'$ in a unique way.

  Now we can apply f to see that $f(z) = \mu_1 f(v') + \mu_2 f(u')$. But $f(z) = 0$, $f(v') = v$ and $f(u') = u$, which tells us that $0 = \mu_1 v + \mu_2 u$ is another way to write 0 in terms of the base $\{v, u\}$. But by definition of base there's a unique way to do so - which, as we've seen above, is $0 = 0v + 0u$.

  Therefore, $\mu_1 = \mu_2 = 0$. But this tells us that $z = \mu_1 v' + \mu_2 u1$ now becomes $z = 0v' + 0u' = 0$ - that is, $z = 0$.

  So we took any point $z \in \operatorname{Ker} f$ and showed that $z = 0$. This implies that $\operatorname{Ker} f = 0$ and so f is injective.

  Since f is already surjective, we see that f is a bijection, which finishes the argument.

Finally, since (a) is equivalent to both (b) and (c), it follows that (b) and (c) are also equivalent amongst themselves.
This ends the proof. □

Remark 2.2.4.12

If you were unhappy with the way we defined base previously, try proving this theorem we've just proven using the classical definition of a base (i.e. a linearly independent spanning set). Good luck. You're gonna need it.

2.3 About the geometry of $\mathbb{R}^2$

2.3.1 Matrices

In this section our aim is to study some more geometrical properties that vectors in $\mathbb{R}^2$ inherit simply because $\mathbb{R}^2$ is, as we've seen, an Euclidean plane.

Definition 2.3.1.1. *Given any two natural numbers $n, m \in \mathbb{N}$, we define a $n \times m$ **real matrix** to be a table with n rows and m columns wherein each entry is a real number.*

We denote the set of all $n \times m$ real matrices by $M_{n\times m}(\mathbb{R})$.

Example(s)

$$\begin{pmatrix} 1 & 2 & -1 \\ 0 & 0 & 9 \end{pmatrix}$$

is a 2×3 real matrix,

$$\begin{pmatrix} 1 \\ \pi \\ \pi \end{pmatrix}$$

is a 3×1 real matrix,

$$\begin{pmatrix} 27 \end{pmatrix}$$

is a 1×1 real matrix.

Proposition 2.3.1.2. *The set of 2×1 real matrices is in bijection with $\mathbb{R}^2$ - that is, $M_{2\times 1}(\mathbb{R}) \cong \mathbb{R}^2$.*

Proof

Let $\psi : M_{2\times 1}(\mathbb{R}) \to \mathbb{R}^2$ be given by $\psi \begin{pmatrix} x \\ y \end{pmatrix} := (x, y)$. Then ψ is clearly a bijection. □

Definition 2.3.1.3. *For any $X, Y \in M_{2\times 1}(\mathbb{R})$ and any $\lambda \in \mathbb{R}$ we define $X + Y$ and λX by putting*

$$X + Y := \psi^{-1}(\psi_X + \psi_Y)$$

$$\lambda X := \psi^{-1}(\lambda \psi_X)$$

where $\psi_X = \psi(X)$ and $\psi_Y = \psi(Y)$.

Lemma 2.3.1.4. *Let $X, Y, Z \in M_{2\times 1}(\mathbb{R})$ and $\lambda, \mu \in R$. Then:*

- $(X + Y) + Z = X + (Y + Z)$
- $X + Y = Y + X$;
- *There exists a matrix $0 \in M_{2\times 1}(\mathbb{R})$ such that $X + 0 = 0 + X = X$;*

- *There exists a matrix* $-X \in M_{2\times 1}(\mathbb{R})$ *such that* $X + (-X) = (-X) + X = 0$
- $\lambda(\mu X) = (\lambda\mu)X$*;*
- $\lambda X = X\lambda$*;*
- $1 \in \mathbb{R}$ *is such that* $1X = X1 = X$*;*
- $\lambda(X + Y) = \lambda X + \lambda Y$*;*
- $(\lambda + \mu)X = \lambda X + \mu X$

Proof

- $$\begin{aligned}(X+Y)+Z &= \psi^{-1}(\psi_X+\psi_Y)+Z\\ &= \psi^{-1}(\psi(\psi^{-1}(\psi_X+\psi_Y))+\psi_Z)\\ &= \psi^{-1}((\psi_X+\psi_Y)+\psi_Z)\\ &= \psi^{-1}(\psi_X+(\psi_Y+\psi_Z))\\ &= \psi^{-1}(\psi_X+\psi(\psi^{-1}(\psi_Y+\psi_Z)))\\ &= X+\psi^{-1}(\psi_Y+\psi_Z) = X+(Y+Z)\end{aligned}$$

  so addition is associative;

- $$\begin{aligned}X+Y &= \psi^{-1}(\psi_X+\psi_Y)\\ &= \psi^{-1}(\psi_Y+\psi_X) = Y+X\end{aligned}$$

  so addition is commutative;

- Let $0 := \psi^{-1}(0,0)$. Then $\psi(0) = (0,0)$, so

  $$X + 0 = \psi^{-1}(\psi_X + \psi(0)) = \psi^{-1}(\psi_X) = X$$

  so 0 is the identity element of the addition;

- Let $-X := \psi^{-1}(-\psi_X)$. Then

  $$\begin{aligned}\psi_X + \psi(-X) &= \psi_X + \psi(\psi^{-1}(-\psi_X))\\ &= \psi_X - \psi_X = (0,0) = \psi(0),\end{aligned}$$

  so

  $$\begin{aligned}X + (-X) &= \psi^{-1}(\psi_X + \psi(-X))\\ &= \psi^{-1}(\psi(0)) = 0\end{aligned}$$

  and $-X$ is the additive inverse of X;

- $$\begin{aligned}\lambda(\mu X) &= \lambda(\psi^{-1}(\mu\psi_X)) \\ &= \psi^{-1}(\lambda\psi(\psi^{-1}(\mu\psi_X))) \\ &= \psi^{-1}(\lambda(\mu\psi_X)) \\ &= \psi^{-1}((\lambda\mu)\psi_X) = (\lambda\mu)X\end{aligned}$$

  so scalar multiplication is associative;

- $\lambda X = \psi^{-1}(\lambda\psi_X) = \psi^{-1}(\psi_X\lambda) = X\lambda$ so scalar multiplication is commutative;

- $1X = \psi^{-1}(1\psi_X) = \psi^{-1}(\psi_X) = X$ so 1 is the identity of the scalar multiplication;

- $$\begin{aligned}\lambda(X+Y) &= \lambda\psi^{-1}(\psi_X + \psi_Y) \\ &= \psi^{-1}(\lambda\psi(\psi^{-1}(\psi_X + \psi_Y))) \\ &= \psi^{-1}(\lambda(\psi_X + \psi_Y)) \\ &= \psi^{-1}(\lambda\psi_X + \lambda\psi_Y) \\ &= \psi^{-1}(\psi(\psi^{-1}(\lambda\psi_X) + \psi(\psi^{-1}(\lambda\psi_Y)))) \\ &= \psi^{-1}(\lambda\psi_X) + \psi^{-1}(\lambda\psi_Y) \\ &= \lambda X + \lambda Y\end{aligned}$$

  so scalar multiplication distributes over addition;

- $$\begin{aligned}(\lambda+\mu)(X) &= \psi^{-1}((\lambda+\mu)\psi_X) \\ &= \psi^{-1}(\lambda\psi_X + \mu\psi_X) \\ &= \psi^{-1}(\psi(\psi^{-1}(\lambda\psi_X)) + \psi(\psi^{-1}(\mu\psi_X))) \\ &= \psi^{-1}(\lambda\psi_X) + \psi^{-1}(\mu\psi_X) \\ &= \lambda X + \mu X\end{aligned}$$

  so scalar multiplication distributes over real number addition.

This finishes the proof. □

Don't get too worried, what we did is more intuitive than it might seem at first:

Since we already know how to add vectors in $\mathbb{R}^2$ and we know that $M_{2\times1}(\mathbb{R}) \cong \mathbb{R}^2$, we can define $\begin{pmatrix}a\\b\end{pmatrix} + \begin{pmatrix}c\\d\end{pmatrix}$ by first taking each one of them to the corresponding vectors (a,b) and (c,d), then adding those up

$$(a,b) + (c,d) = (a+c, b+d)$$

and then bringing the result back to matrices, and putting that as the result of the matrix addition:

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

Similarly, in order to define $\lambda \begin{pmatrix} a \\ b \end{pmatrix}$ we, once again, first take $\begin{pmatrix} a \\ b \end{pmatrix}$ to (a, b), then we multiply that by λ:

$$\lambda(a, b) = (\lambda a, \lambda b)$$

and then bring that back to matrices and put that as the result of the scalar multiplication:

$$\lambda \begin{pmatrix} a \\ b \end{pmatrix} := \begin{pmatrix} \lambda a \\ \lambda b \end{pmatrix}.$$

For this very reason, we can think of the elements of $\mathbb{R}^2$ not only as pairs of real numbers, not only as points in a plane, not only as vectors, but also as 2×1 matrices.

Now that we can think of vectors as matrices, it's no surprise to also be able to think of linear maps as matrices.

Proposition 2.3.1.5. *There is a bijection* $M_{2\times 2}(\mathbb{R}) \cong \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$.

Proof

Let $E = \{e_2 e_2\}$ be the canonical base of $\mathbb{R}^2$. Define $\phi : M_{2\times 2}(\mathbb{R}) \to \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$ by putting, for each $A = \begin{pmatrix} a_1 & a_3 \\ a_2 & a_4 \end{pmatrix}$, $\phi(A)(e_1) := (a_1, a_2)$ and $\phi(A)(e_2) := (a_3, a_4)$.
We'll denote $\phi(A)$ simply by ϕ_A. Then the above equalities become $\phi_A(e_1) = (a_1, a_2)$ and $\phi_A(e_2) = (a_3, a_4)$ - that is, if c_1 and c_2 are, respectively, the first and second columns of A, then $\phi_A(e_1) = \psi(c_1)$ and $\phi_A(e_2) = \psi(c_2)$.

- ϕ is injective:
  Assume that $A, B \in M_{2\times 2}(\mathbb{R})$ are two matrices such that $\phi_A = \phi_B$. This means, in particular, that $\phi_A(e_1) = \phi_B(e_1)$ and $\phi_A(e_2) = \phi_B(e_2)$. But this implies that the first and second columns of A and B are equal. Since they only have two columns, this implies $A = B$, and so ϕ is injective.

- ϕ is surjective:
  Take any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Then by computing $f(e_1) = (a, b)$ and $f(e_2) = (c, d)$ we can then define $A^f := \begin{pmatrix} a & c \\ b & d \end{pmatrix}$. Clearly, then, $\phi_{A^f} = f$, so $f \in \mathrm{Im}(\phi)$ and so ϕ is surjective.

This shows that ϕ is bijective, and so we have proven the result. □

So we can think of 2×2 matrices as linear transformations, and vice-versa.

Since we used a result like this to introduce addition and scalar multiplication to 2×1 matrices, you can guess what's coming next, right?

Proposition 2.3.1.6. *Let $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ be two linear functions and $\lambda \in \mathbb{R}$ be any scalar. Then the functions $f + g, f \circ g, \lambda f : \mathbb{R}^2 \to \mathbb{R}^2$ defined, respectively, as $(f + g)(v) := f(v) + g(v)$, $(f \circ g)(v) := f(g(v))$ and $(\lambda f)(v) := \lambda f(v)$ are linear functions.*

Proof

Choose, once and for all, $v, u \in \mathbb{R}^2$ and $\mu \in \mathbb{R}$. Then:

- $f + g$ is linear:

$$\begin{aligned}(f+g)(v+u) &= f(v+u) + g(v+u)\\ &= f(v) + f(u) + g(v) + g(u)\\ &= f(v) + g(v) + f(u) + g(u) = (f+g)(v) + (f+g)(u)\end{aligned}$$

and

$$\begin{aligned}(f+g)(\mu v) &= f(\mu v) + g(\mu v)\\ &= \mu f(v) + \mu g(v)\\ &= \mu(f(v) + g(v)) = \mu(f+g)(v)\end{aligned}$$

so $f + g$ is linear.

- $f \circ g$ is linear:

$$\begin{aligned}(f \circ g)(v+u) &= f(g(v+u))\\ &= f(g(v) + g(u))\\ &= f(g(v)) + f(g(u)) = (f \circ g)(v) + (f \circ g)(u)\end{aligned}$$

and

$$\begin{aligned}(f \circ g)(\mu v) &= f(g(\mu v))\\ &= f(\mu g(v))\\ &= \mu f(g(v)) = \mu(f \circ g)(v)\end{aligned}$$

so $f \circ g$ is linear.

- λf is linear:

$$\begin{aligned}(\lambda f)(v+u) &= f\lambda(v+u)\\ &= \lambda(f(v) + f(u))\\ &= \lambda f(v) + \lambda f(u) = (\lambda f)(v) + (\lambda f)(u)\end{aligned}$$

and

$$\begin{aligned}(\lambda f)(\mu v) &= \lambda f(\mu v)\\ &= \lambda\mu f(v)\\ &= \mu\lambda f(v) = \mu(\lambda f)(v)\end{aligned}$$

so λf is linear.

This ends the proof. □

Definition 2.3.1.7. *The bijection $\phi : M_{2\times 2}(\mathbb{R}) \to \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$ induces, for all $X, Y \in M_{2\times 2}(\mathbb{R})$ and $\lambda \in \mathbb{R}$, $X + Y$, XY and λX by putting*

$$X + Y := \phi^{-1}(\phi_X + \phi_Y)$$

$$XY := \phi^{-1}(\phi_X \circ \phi_Y)$$

$$\lambda X := \phi^{-1}(\lambda\phi_X),$$

where $\phi_X = \phi(X)$ and $\phi_Y = \phi(Y)$.

Now, it would be great to be able to calculate each of those matrices. Well, it turns out to be easier than it seems:

Proposition 2.3.1.8. *Let $X, Y \in M_{2\times 2}(\mathbb{R})$ and $\lambda \in \mathbb{R}$. If we write $X = \begin{pmatrix} x_1 & x_3 \\ x_2 & x_4 \end{pmatrix}$ and $Y = \begin{pmatrix} y_1 & y_3 \\ y_2 & y_4 \end{pmatrix}$ then the following equalities hold:*

$$X + Y = \begin{pmatrix} x_1 + y_1 & x_3 + y_3 \\ x_2 + y_2 & x_4 + y_4 \end{pmatrix}$$

$$XY = \begin{pmatrix} x_1y_1 + x_3y_2 & x_1y_3 + x_3y_4 \\ x_2y_1 + x_4y_2 & x_2y_3 + x_4y_4 \end{pmatrix}$$

$$\lambda X = \begin{pmatrix} \lambda x_1 & \lambda x_3 \\ \lambda x_2 & \lambda x_4 \end{pmatrix}.$$

Proof

Let $E = \{e_1, e_2\}$ be the canonical base for $\mathbb{R}^2$. Then $\phi_X(e_1) = (x_1, x_2)$, $\phi_X(e_2) = (x_3, x_4)$, $\phi_y(e_1) = (y_1, y_2)$ and $\phi_Y(e_2) = (y_3, y_4)$.
It follows then that

$$(\phi_X + \phi_Y)(e_1) = \phi_X(e_1) + \phi_Y(e_1) = (x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$$

$$(\phi_X + \phi_Y)(e_2) = \phi_X(e_2) + \phi_Y(e_2) = (x_3, x_4) + (y_3, y_4) = (x_3 + y_3, x_4 + y_4)$$

so

$$X + Y = \begin{pmatrix} x_1 + y_1 & x_3 + y_3 \\ x_2 + y_2 & x_4 + y_4 \end{pmatrix}.$$

Similarly,

$$\begin{aligned} \phi_X(\phi_Y(e_1)) &= \phi_X(y_1, y_2) \\ &= \phi_X(y_1 e_1 + y_2 e_2) \\ &= y_1 \phi_X(e_1) + y_2 \phi_X(e_2) \\ &= y_1(x_1, x_2) + y_2(x_3, x_4) \\ &= (x_1 y_1, x_2 y_1) + (x_3 y_2, x_4 y_2) = (x_1 y_1 + x_3 y_2, x_2 y_1 + x_4 y_2) \end{aligned}$$

$$\begin{aligned} \phi_X(\phi_Y(e_2)) &= \phi_X(y_3, y_4) \\ &= \phi_X(y_3 e_1 + y_4 e_2) \\ &= y_3 \phi_X(e_1) + y_4 \phi_X(e_2) \\ &= y_3(x_1, x_2) + y_4(x_3, x_4) \\ &= (x_1 y_3, x_2 y_3) + (x_3 y_4, x_4 y_4) = (x_1 y_3 + x_3 y_4, x_2 y_3 + x_4 y_4) \end{aligned}$$

so

$$XY = \begin{pmatrix} x_1 y_1 + x_3 y_2 & x_1 y_3 + x_3 y_4 \\ x_2 y_1 + x_4 y_2 & x_2 y_3 + x_4 y_4 \end{pmatrix}.$$

Finally,

$$(\lambda \phi_X)(e_1) = \lambda \phi_X(e_1) = \lambda(x_1, x_2) = (\lambda x_1, \lambda x_2)$$

$$(\lambda \phi_X)(e_2) = \lambda \phi_X(e_2) = \lambda(x_3, x_4) = (\lambda x_3, \lambda x_4)$$

so

$$\lambda X = \begin{pmatrix} \lambda x_1 & \lambda x_3 \\ \lambda x_2 & \lambda x_4 \end{pmatrix}.$$

This ends the proof. □

This is the best explanation why matrix multiplication is so weird. Simply put, it's because it's the worst possible way to write function composition. But, all in all, it's just that - function composition.

From this, the next result now follows:

Lemma 2.3.1.9. *There is a unique matrix $I \in M_{2\times 2}(\mathbb{R})$ such that for any other matrix $A \in M_{2\times 2}(\mathbb{R})$ we have that $IA = AI = A$.*

Proof

Take $I := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. Then:

$$IA = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} 1 \cdot a + 0 \cdot c & 1 \cdot b + 0 \cdot d \\ 0 \cdot a + 1 \cdot c & 0 \cdot b + 1 \cdot d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A$$

and

$$AI = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a \cdot 1 + b \cdot 0 & a \cdot 0 + b \cdot 1 \\ c \cdot 1 + d \cdot 0 & c \cdot 0 + d \cdot 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A.$$

Finally, to prove uniqueness, assume $I' \in M_{2\times 2}(\mathbb{R})$ also satisfies $I'A = AI' = A$ for all $A \in M_{2\times 2}(\mathbb{R})$. Then:

$$I = II' = I'I = I'$$

so $I = I'$ and we see that I is indeed the unique matrix satisfying that property. □

Definition 2.3.1.10. *The matrix $I \in M_{2\times 2}(\mathbb{R})$ is called the **identity matrix**.*

Proposition 2.3.1.11. *The identity matrix is just the matrix representation of the identity function. That is, $\phi_I = \mathrm{id}_{\mathbb{R}^2}$.*

Proof

This is trivial: By definition, $\phi_I(e_1)$ is the first column of I, and $\phi_I(e_2)$ is the second columns of I.
But the first column of I is just $(1, 0) = e_1$, and the second columns of I is just $(0, 1) = e_2$, so $\phi_I(e_1) = e_1$ and $\phi_I(e_2) = e_2$.
This shows that $\phi_I = \mathrm{id}_{\mathbb{R}^2}$, which is what we wanted. □

Finally, we can prove one last result.

Definition 2.3.1.12. *We define the **evaluation map** $ev : \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2) \times \mathbb{R}^2 \to \mathbb{R}^2$ to be the map given by $ev(f, v) := f(v)$.*

Definition 2.3.1.13. *The evaluation map induces another evaluation map $M_{2\times 2}(\mathbb{R}) \times M_{2\times 1}(\mathbb{R}) \to M_{2\times 1}(\mathbb{R})$ given by*

$$AX := \psi^{-1}(\phi_A(\psi_X)).$$

Lemma 2.3.1.14. *For any $A \in M_{2\times 2}(\mathbb{R})$ and any $X \in M_{2\times 1}(\mathbb{R})$ we have that*

$$AX = \begin{pmatrix} a_1x_1 + a_3x_2 \\ a_2x_1 + a_4x_2 \end{pmatrix}$$

where $A = \begin{pmatrix} a_1 & a_3 \\ a_2 & a_4 \end{pmatrix}$ and $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

Proof

Let, once more, $E = \{e_1, e_2\}$ be the canonical base of $\mathbb{R}^2$. Then $\phi_A(e_1) = (a_1, a_2)$ and $\phi_A(e_2) = (a_3, a_4)$.
Therefore,

$$\begin{aligned}\phi_A(\psi_X) &= \phi_A(x_1, x_2)\\ &= \phi_A(x_1e_1 + x_2e_2)\\ &= x_1\phi_A(e_1) + x_2\phi_A(e_2)\\ &= x_1(a_1, a_2) + x_2(a_3, a_4)\\ &= (a_1x_1, a_2x_1) + (a_3x_2, a_4x_2) = (a_1x_1 + a_3x_2, a_2x_1 + a_4x_2)\end{aligned}$$

and we see that

$$AX = \begin{pmatrix} a_1x_1 + a_3x_2 \\ a_2x_1 + a_4x_2 \end{pmatrix}$$

which proves the result. □

Corollary 2.3.1.15. *For any $X \in M_{2\times 1}(\mathbb{R})$ we have that $IX = X$.*

2.3.2 The transpose

Proposition 2.3.2.1. *There is a bijection between the set of* 1×2 *real matrices and* $\mathbb{R}^2$ *- that is,* $M_{1\times 2}(\mathbb{R}) \cong \mathbb{R}^2$.

Proof

Consider the function $\varphi : M_{1\times 2}(\mathbb{R}) \to \mathbb{R}^2$ given by $\varphi\begin{pmatrix} x & y \end{pmatrix} = (x, y)$.
This is clearly a bijection. □

Definition 2.3.2.2. *For any* $X, Y \in M_{1\times 2}(\mathbb{R})$ *and any* $\lambda \in \mathbb{R}$ *we define* $X+Y$ *and* λX *by putting*

$$X + Y := \varphi^{-1}(\varphi_X + \varphi_Y)$$

$$\lambda X := \varphi^{-1}(\lambda\varphi_X)$$

where $\varphi_X = \varphi(X)$ *and* $\varphi_Y = \varphi(Y)$.

Lemma 2.3.2.3. *Let* $X, Y, Z \in M_{1\times 2}(\mathbb{R})$ *and* $\lambda, \mu \in \mathbb{R}$. *Then:*

- $(X + Y) + Z = X + (Y + Z)$
- $X + Y = Y + X$;
- *There exists a matrix* $0 \in M_{1\times 2}(\mathbb{R})$ *such that* $X + 0 = 0 + X = X$;
- *There exists a matrix* $-X \in M_{1\times 2}(\mathbb{R})$ *such that* $X + (-X) = (-X) + X = 0$
- $\lambda(\mu X) = (\lambda\mu)X$;
- $\lambda X = X\lambda$;
- $1 \in \mathbb{R}$ *is such that* $1X = X1 = X$;
- $\lambda(X + Y) = \lambda X + \lambda Y$;
- $(\lambda + \mu)X = \lambda X + \mu X$

We'll refrain from doing this proof since it is essentially the same proof as in the case for 2×1 matrices.

Since we already see matrices as linear transformations we would like to teach matrices how to act on 1×2 matrices then.

To do that, however, we need to create matrices in a new way:

Proposition 2.3.2.4. *The function* $\tau : M_{2\times 2}(\mathbb{R}) \to \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2)$ *given by* $\tau(A)(e_1) := (a_1, a_2)$ *and* $\tau(A)(e_2) := (a_3, a_4)$, *where* $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$, *is a bijection.*

Proof

Let us put $\tau(A) = \tau_A$ just to simplify our notation. Then, if r_1 and r_2 are, respectively, the first and second rows of A, then $\tau_A(e_1) = \varphi(r_1)$ and $\tau_A(e_2) = \varphi(r_2)$.

- τ is injective:

  Assume that $A, B \in M_{2\times 2}(\mathbb{R})$ are two matrices such that $\tau_A = \tau_B$. This means, in particular, that $\tau_A(e_1) = \tau_B(e_1)$ and $\tau_A(e_2) = \tau_B(e_2)$. But this implies that the first and second rows of A and B are equal. Since they only have two rows, this implies $A = B$, and so τ is injective.

- τ is surjective:

  Take any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Then by computing $f(e_1) = (a, b)$ and $f(e_2) = (c, d)$ we can then define $A^f := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Clearly, then, $\tau_{A^f} = f$, so $f \in \mathrm{Im}(\tau)$ and so τ is surjective.

This shows that τ is bijective, and so we have proven the result. □

With this, we can finally define how to act on 1×2 matrices:

Definition 2.3.2.5. *The evaluation map* $ev : \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^2, \mathbb{R}^2) \times \mathbb{R}^2 \to \mathbb{R}^2$ *induces a unique evaluation map* $M_{1\times 2}(\mathbb{R}) \times M_{2\times 2}(\mathbb{R}) \to M_{1\times 2}(\mathbb{R})$ *given by*

$$YA := \varphi^{-1}(\tau_A(\varphi_Y)).$$

Lemma 2.3.2.6. *For any* $A \in M_{2\times 2}(\mathbb{R})$ *and any* $Y \in M_{1\times 2}(\mathbb{R})$ *we have that*

$$YA = \begin{pmatrix} a_1y_1 + a_3y_2 & a_2y_1 + a_4y_2 \end{pmatrix}$$

where $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$ *and* $Y = \begin{pmatrix} y_1 & y_2 \end{pmatrix}$.

Proof

Let, once more, $E = \{e_1, e_2\}$ be the canonical base of $\mathbb{R}^2$. Then $\tau_A(e_1) = (a_1, a_2)$ and $\tau_A(e_2) = (a_3, a_4)$.
Therefore,

$$\begin{aligned} \tau_A(\varphi_X) &= \tau_A(y_1, y_2) \\ &= \tau_A(y_1e_1 + y_2e_2) \\ &= y_1\tau_A(e_1) + y_2\tau_A(e_2) \\ &= y_1(a_1, a_2) + y_2(a_3, a_4) \\ &= (a_1y_1, a_2y_1) + (a_3y_2, a_4y_2) = (a_1y_1 + a_3y_2, a_2y_1 + a_4y_2) \end{aligned}$$

and we see that

$$YA = \begin{pmatrix} a_1y_1 + a_3y_2 & a_2y_1 + a_4y_2 \end{pmatrix}$$

which proves the result. □

Corollary 2.3.2.7. *For any $Y \in M_{1\times 2}(\mathbb{R})$ we have that $YI = Y$.*

Definition 2.3.2.8. *We define the **transpose map** $(-)^t : M_{2\times 1}(\mathbb{R}) \to M_{1\times 2}(\mathbb{R})$ to be the map given by $X^t := \varphi^{-1}(\psi_X)$.*

In other words, if we write $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ then its transpose is $X^t = \begin{pmatrix} x_1 & x_2 \end{pmatrix}$.

Proposition 2.3.2.9. *The transpose map is a bijection between $M_{2\times 1}(\mathbb{R})$ and $M_{1\times 2}(\mathbb{R})$*

Proof

This is immediate from the definition of the transpose map: Since $(-)^t = \varphi^{-1} \circ \psi$ and both φ^{-1} and ψ are bijections, the result follows. □

Definition 2.3.2.10. *By an abuse of notation, we'll denote $((-)^t)^{-1} = (-)^t$. That is, the symbol $(-)^t$ means both the process of turning a 2×1 matrix into a 1×2 matrix as well as its inverse process of turning a 1×2 matrix into a 2×1 matrix.*

Now, let's do the same for 2×2 matrices:

Definition 2.3.2.11. *We define the **transpose map** (once again, by abuse of notation) $(-)^t : M_{2\times 2}(\mathbb{R}) \to M_{2\times 2}(\mathbb{R})$ to be the map given by $A^t = \tau^{-1}(\phi_A)$.*

In other words, if we write $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ then its transpose is $A^t = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Proposition 2.3.2.12. *The transpose map is a bijection between $M_{2\times 2}(\mathbb{R})$ and itself.*

Proof

This is immediate from the definition of the transpose map: Since $(-)^t = \tau^{-1} \circ \phi$ and both τ^{-1} and ϕ are bijections, the result follows. □

Corollary 2.3.2.13. *The inverse of the transpose map is itself.*

Proof

We need to prove that $\tau^{-1} \circ \phi \circ \tau^{-1} \circ \phi = \mathrm{id}_{M_{2\times 2}(\mathbb{R})}$.

Take any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $\phi_A(e_1) = (a, c)$ and $\phi_A(e_2) = (b, d)$. Therefore,

$$A^t = \tau^{-1}(\phi_A) = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

But now $\phi_{A^t}(e_1) = (a, b)$ and $\phi_{A^t}(e_2) = (c, d)$. Therefore,

$$(A^t)^t = \tau^{-1}(\phi_{A^t}) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A.$$

Combining all of this together, we see that

$$(\tau^{-1} \circ \phi \circ \tau^{-1} \circ \phi)(A) = (A^t)^t = A = \mathrm{id}_{M_{2\times 2}(\mathbb{R})}(A)$$

and so the result follows. □

Corollary 2.3.2.14. *By rephrasing the previous corollary we see that* $\tau^{-1} \circ \phi = \phi^{-1} \circ \tau$.

Now, we can take all of these together and get a very important result:

Theorem 2.3.2.15. *Let* $X \in M_{2\times 1}(\mathbb{R})$, $A \in M_{2\times 2}(\mathbb{R})$ *and* $Y \in M_{1\times 2}(\mathbb{R})$. *Then* $(YA)^t = A^t Y^t$ *and* $(AX)^t = X^t A^t$.

Proof

Let us remember our definitions:

- $X^t = \varphi^{-1}(\psi_X)$
- $AX = \psi^{-1}(\phi_A(\psi_X))$
- $A^t = \tau^{-1}(\phi_A)$
- $Y^t = \psi^{-1}(\varphi_Y)$
- $YA = \varphi^{-1}(\tau_A(\varphi_Y))$
- $A^t = \phi^{-1}(\tau_A)$

Therefore,

$$\begin{aligned}(AX)^t &= (\psi^{-1}(\phi_A(\psi_X)))^t \\ &= \varphi^{-1}(\psi(\psi^{-1}(\phi_A(\psi_X)))) \\ &= \varphi^{-1}(\phi_A(\psi_X)) \\ &= \varphi^{-1}(\tau(\tau^{-1}(\phi_A))(\varphi(\varphi^{-1}(\psi_X)))) \\ &= \varphi^{-1}(\tau_{A^t}(\varphi_{X^t})) = X^t A^t\end{aligned}$$

and

$$\begin{aligned}(YA)^t &= (\varphi^{-1}(\tau_A(\varphi_Y)))^t \\ &= \psi^{-1}(\varphi(\varphi^{-1}(\tau_A(\varphi_Y)))) \\ &= \psi^{-1}(\tau_A(\varphi_Y)) \\ &= \psi^{-1}(\phi(\phi^{-1}(\tau_A))(\psi(\psi^{-1}(\varphi_Y)))) \\ &= \psi^{-1}(\phi_{A^t}(\psi_{Y^t})) = A^tY^t\end{aligned}$$

so the result follows. □

Finally, to end this section, we'll introduce a new perspective which is better seen from the perspective of matrices:

Definition 2.3.2.16. *Let $Y \in M_{1\times 2}(\mathbb{R})$ and $X \in M_{2\times 1}(\mathbb{R})$. Inspired by the evaluation maps, we define the **multiplication of Y and X** to be the 1×1 matrix YX given by*

$$\begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1y_1 + x_2y_2 \end{pmatrix}.$$

Proposition 2.3.2.17. *There is a bijection $M_{1\times 1}(\mathbb{R}) \cong \mathbb{R}$.*

Proof

Let $\sigma : M_{1\times 1}(\mathbb{R}) \to \mathbb{R}$ be given by $\sigma\begin{pmatrix} a \end{pmatrix} := a$. Then it clearly is a bijection, which ends the proof. □

Definition 2.3.2.18. *The bijection $\sigma : M_{1\times 1}(\mathbb{R}) \to \mathbb{R}$ induces operations $X + Y$ and λX for all $X, Y \in M_{1\times 1}(\mathbb{R})$ and $\lambda \in \mathbb{R}$ given, respectively, by*

$$\begin{aligned} X + Y &:= \sigma^{-1}(\sigma_X + \sigma_Y) \\ \lambda X &:= \sigma^{-1}(\lambda\sigma_X) \end{aligned}$$

where $\sigma_X := \sigma(X)$ and $\sigma_Y := \sigma(Y)$.

Proposition 2.3.2.19. *If $X = \begin{pmatrix} x \end{pmatrix}$, $Y = \begin{pmatrix} y \end{pmatrix}$ and $\lambda \in \mathbb{R}$, then $X + Y = \begin{pmatrix} x + y \end{pmatrix}$ and $\lambda X = \begin{pmatrix} \lambda x \end{pmatrix}$.*

Proof

This follows trivially by definition of σ. □

To end this section, then, we'll provide a list of good properties of all matrices additions and multiplications we've done so far.

Lemma 2.3.2.20. *Let $Y, Y' \in M_{1\times 2}(\mathbb{R})$, $X, X' \in M_{2\times 1}(\mathbb{R})$, $A, A' \in M_{2\times 2}(\mathbb{R})$ and $\lambda \in \mathbb{R}$. Then the following hold:*

a) $Y(AX) = (YA)X$;

b) $(Y + Y')X = YX + Y'X$;

c) $Y(X + X') = YX + YX'$;

d) $(Y + Y')A = YA + Y'A$;

e) $A(X + X') = AX + AX'$;

f) $Y(A + A') = YA + YA'$;

g) $(A + A')X = AX + A'X$;

h) $Y(AA') = (YA)A'$;

i) $(AA')X = A(A'X)$;

j) $(\lambda Y)X = \lambda(YX) = Y(\lambda X)$;

k) $(\lambda Y)A = \lambda(YA) = Y(\lambda A)$;

l) $(\lambda A)X = \lambda(AX) = A(\lambda X)$.

Proof

Using

$$Y = \begin{pmatrix} y_1 & y_2 \end{pmatrix}, \quad Y' = \begin{pmatrix} y_1' & y_2' \end{pmatrix}$$

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad X' = \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$$

$$A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}, \quad A' = \begin{pmatrix} a_1' & a_2' \\ a_3' & a_4' \end{pmatrix}.$$

all the proofs follow by simple computation and will be left as an exercise to the reader. □

2.3.3 Distance between points

Definition 2.3.3.1. *Let $v, u \in \mathbb{R}^2$. We define the **inner product** of v and u to be the real number $\langle v, u \rangle$ given by*

$$\langle v, u \rangle := \sigma(\varphi^{-1}(v)\psi^{-1}(u))$$

.

Proposition 2.3.3.2. *For any two vectors $v = (v_1, v_2)$ and $u = (u_1, u_2)$ in $\mathbb{R}^2$, we have that $\langle v, u \rangle = v_1u_1 + v_2u_2$.*

Proof

If $v = (v_1, v_2)$ and $u = (u_1, u_2)$, then $\varphi^{-1}(v) = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$ and $\psi^{-1}(u) = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$.

From this, we can see that $\varphi^{-1}(v)\psi^{-1}(u) = \begin{pmatrix} v_1u_1 + v_2u_2 \end{pmatrix}$ and, therefore,

$$\langle v, u \rangle = \sigma(\varphi^{-1}(v)\psi^{-1}(u)) = \sigma\left(v_1u_1 + v_2u_2\right) = v_1u_2 + v_2u_2$$

which ends the proof. □

In other words, the inner product of two vectors $v, u \in \mathbb{R}^2$ is just the multiplication of v, written as a 1×2 matrix, and u, written as a 2×1 matrix.

The reason why we didn't simply define it using vectors, but instead went all the way around, passing through matrices, to define the inner product is twofold.

First, because it makes way more sense to multiply matrices than it does to multiply vectors.

Second, because further ahead we're gonna need to define what's called the **algebraic dual** of a vector space, and then we'll see that inner products arise naturally from algebraic dualization. However, in the case of $\mathbb{R}^2$, we'll see that the algebraic dual of $M_{2\times1}(\mathbb{R})$ is simply $M_{1\times2}(\mathbb{R})$ - so this mindset is a particular case of a more general approach we're gonna use later on and, as such, provides great insight on how to approach the general case.

Example(s)

Let $v = (1, 2)$, $u = (-1, -5)$ and $w = (\pi, 0)$. Then we can compute:

$\langle v, u \rangle = 1 \cdot (-1) + 2 \cdot (-5) = -1 - 10 = -11$

$\langle u, v \rangle = (-1) \cdot 1 + (-5) \cdot 2 = -1 - 10 = -11$

$\langle v, w \rangle = 1 \cdot \pi + 2 \cdot 0 = \pi + 0 = \pi$

$\langle v, w \rangle = \pi \cdot 1 + 0 \cdot 2 = \pi + 0 = \pi$

$\langle u, w \rangle = (-1) \cdot \pi + (-5) \cdot 0 = -\pi + 0 = -\pi$

$\langle u, w \rangle = \pi \cdot (-1) + 0 \cdot (-5) = -\pi + 0 = -\pi$

This gives us intuition about our next result:

Proposition 2.3.3.3. *The inner product is commutative - that is, for any two vectors $v, u \in \mathbb{R}^2$ we have $\langle v, u\rangle = \langle u, v\rangle$.*

Proof

Let $v = (v_1, v_2)$ and $u = (u_1, u_2)$. Then:

$$\langle v, u\rangle = v_1u_1 + v_2u_2 = u_1v_1 + u_2v_2 = \langle u, v\rangle$$

so the inner product is commutative and the result follows. □

Example(s)

Let $v = (2, 7)$, $u = (2, -3)$ and $w = (-6, 9)$. Then we can compute:

$$\langle v, u\rangle = 2 \cdot 2 + 7 \cdot (-3) = 4 - 21 = -17$$

$$\langle v, w\rangle = 2 \cdot (-6) + 7 \cdot 9 = -12 + 63 = 51$$

$$\langle u, w\rangle = 2 \cdot (-6) + (-3) \cdot 9 = -12 - 27 = -39$$

and once again we have an intuition about a new result:

Proposition 2.3.3.4. *The inner product preserves addition and scalar multiplication - that is, for any three vectors $v, u, w \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$ we have $\langle v, u + w\rangle = \langle v, u\rangle + \langle v, w\rangle$ and $\langle v, \lambda u\rangle = \lambda\langle v, u\rangle$.*

Proof

Let, once again, $v = (v_1, v_2)$, $u = (u_1, u_2)$ and $w = (w_1, w_2)$. Then $u + w = (u_1 + w_1, u_2 + w_2)$ and $\lambda u = (\lambda u_1, \lambda u_2)$, and so:

$$\begin{aligned}\langle v, u + w\rangle &= v_1(u_1 + w_1) + v_2(u_2 + w_2)\\ &= v_1u_1 + v_1w_1 + v_2u_2 + v_2w_2\\ &= (v_1u_1 + v_2u_2) + (v_1w_1 + v_2w_2) = \langle v, u\rangle + \langle v, w\rangle\end{aligned}$$

and

$$\langle v, \lambda u\rangle = v_1(\lambda u_1) + v_2(\lambda u_2) = \lambda(v_1u_1) + \lambda(v_2u_2) = \lambda(v_1u_1 + v_2u_2) = \lambda\langle v, u\rangle$$

so the inner product preserves addition scalar multiplication, which proves the result. □

Now let's quickly go back to thinking about vectors geometrically. If we take any vector $v = (v_1, v_2) \in \mathbb{R}^2$, we know that v can be written as $(v_1, 0) + (0, v_2)$ and that the vectors $(v_1, 0)$ and $(0, v_2)$ are perpendicular (since they lie on different axis - $(v_1, 0)$ lies on the X-axis and $(0, v_2)$ lies on the Y-axis).

Therefore, we know that $0v_1v$ is a right triangle - as is $0v_2v$. In particular, we know that the angle $0\hat{v}_1v$ is a right angle. So, by applying the Pythagorean Theorem, we see that $\overline{0v_1}^2 + \overline{v_1v}^2 = \overline{0v}^2$.

But $\overline{0v_1}$ is just $|v_1|$, and $\overline{v_1v}$ is just $|v_2|$ (since they are perpendicular). Not only that, but $\overline{0v}$ is just $\|v\|$, by definition. So the above equation becomes

$$\|v\|^2 = |v_1|^2 + |v_2|^2 .$$

Finally, since $|x| \geq 0$ and $x^2 \geq 0$ for all $x \in \mathbb{R}$, we can finally get our final equation

$$\|v\|^2 = v_1^2 + v_2^2.$$

You might be asking why we left all this for now, instead of doing it back when we were doing geometric computations with vectors. Well, there's one good reason...

Lemma 2.3.3.5. *For any vector $v \in \mathbb{R}^2$ we have that $\|v\| = \langle v, v\rangle^{\frac{1}{2}}$.*

Proof

This is trivial by the preceding discussion:
Let $v = (v_1, v_2)$. Then $\langle v, v\rangle = v_1v_1 + v_2v_2 = v_1^2 + v_2^2 = \|v\|^2$. □

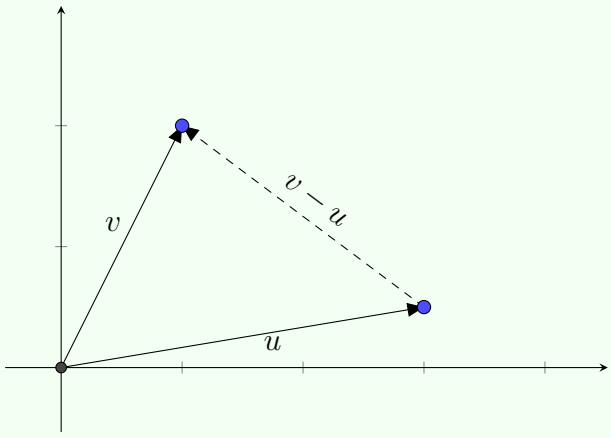
Corollary 2.3.3.6. *Let $v, u \in \mathbb{R}^2$ be any two vectors and $\theta \in [0, 2\pi)$ be the angle between them. Then $\cos\theta = \dfrac{\langle v, u\rangle}{\|v\|\|u\|}$.*

Proof

Remember, from geometry, that if ABC is a triangle with sides $a, b, c \in \mathbb{R}$ and angle $\theta \in [0, 2\pi)$ between a and b, then the Law of Cosines tells us that

$$c^2 = a^2 + b^2 - 2\,|a|\,|b|\cos\theta.$$

Take now any two vectors $v, u \in \mathbb{R}^2$. They determine a unique triangle $0vu$, whose sides measure $\overline{0v} = \|v\|$, $\overline{0u} = \|u\|$ and $\overline{vu}$. Let us first compute $\overline{vu}$.
We claim that $\overline{vu} = \|v - u\|$.



This is easy to see, because $(v-u)+u=v$ so if we overlay the vector $v-u$ at the end of the vector u we get the vector v (by definition of vector sum).
Now that that's taken care of, we can use the previous lemma to see that

$$\|v-u\|^2 = \langle v-u, v-u\rangle$$

and since we already know that the inner product preserves addition and scalar multiplication, that is simply

$$\|v-u\|^2 = \langle v-u, v-u\rangle = \langle v, v\rangle - \langle u, v\rangle - \langle v, u\rangle + \langle u, u\rangle$$

which, since we already know the inner product is commutative, is simply

$$\|v-u\|^2 = \langle v-u, v-u\rangle = \langle v, v\rangle - 2\langle v, u\rangle + \langle u, u\rangle.$$

Now we use that $\langle v, v\rangle = \|v\|^2$ and $\langle u, u\rangle = \|u\|^2$ to see that

$$\|v-u\|^2 = \|v\|^2 + \|u\|^2 - 2\langle v, u\rangle.$$

But, since $0vu$ is a triangle with sides $\|v\|$, $\|u\|$ and $\|v-u\|$, we have that the Law of Cosines tells us that

$$\|v-u\|^2 = \|v\|^2 + \|u\|^2 - 2\|v\|\|u\|\cos\theta,$$

where θ is the angle between v and u.
Finally, by comparing the two equalities we see that

$$\langle v, u\rangle = \|v\|\|u\|\cos\theta$$

so

$$\cos\theta = \frac{\langle v, u\rangle}{\|v\|\|u\|}$$

and the result follows. □

Corollary 2.3.3.7. *Two non-null vectors $v, u \in \mathbb{R}^2$ are perpendicular if, and only if, $\langle v, u\rangle = 0$.*

Proof

Assume $v \perp u$. Then if $\theta \in [0, 2\pi)$ is the angle between v and u, we know that $\theta = \frac{\pi}{2}$. But then $\cos\theta = 0$, so

$$\langle v, u\rangle = \|v\|\|u\|\cos\theta = \|v\|\|u\| \cdot 0 = 0$$

and we see that $\langle v, u\rangle = 0$.
Conversely, if $v, u \in \mathbb{R}^2$ are non-null vectors such that $\langle v, u\rangle = 0$, then

$$\cos\theta = \frac{\langle v, u\rangle}{\|v\|\|u\|} = \frac{0}{\|v\|\|u\|} = 0$$

so $\cos\theta = 0$.
This implies $\theta \in \{\frac{\pi}{2}, \frac{3\pi}{2}\}$, but since we're only measuring internal angles, $0 \leq \theta \leq \frac{\pi}{2}$. And so, the only possible case is that $\theta = \frac{\pi}{2}$, so $v \perp u$.
This finishes the proof. □

Corollary 2.3.3.8 (Cauchy-Schwarz). *For any two vectors $v, u \in \mathbb{R}^2$, $|\langle v, u\rangle| \leq \|v\|\|u\|$.*

Proof

Since $\langle v, u\rangle = \|v\|\|u\|\cos\theta$ and $-1 \leq \cos\theta \leq 1$, we see that

$$-\|v\|\|u\| \leq \langle v, u\rangle \leq \|v\|\|u\|$$

so $|\langle v, u\rangle| \leq \|v\|\|u\|$, as we had claimed. □

Proposition 2.3.3.9. *Let $v, u \in \mathbb{R}^2$ be any two vectors, and $\lambda \in \mathbb{R}$. Then the following hold:*

a) $\|\lambda v\| = |\lambda|\,\|v\|$;

b) $\|v\| \geq 0$, *with the equality happening if, and only if* $v = 0$;

c) $\|v + u\| \leq \|v\| + \|u\|$.

Proof

a)

$$\begin{aligned}\|\lambda v\|^2 &= \langle \lambda v, \lambda v\rangle \\ &= \lambda\langle v, \lambda v\rangle \\ &= \lambda^2\langle v, v\rangle = |\lambda|^2\,\|v\|^2\end{aligned}$$

and since $\|\lambda v\|$, $|\lambda|$ and $\|v\|$ are all positive,

$$\|\lambda v\| = \sqrt{\|\lambda v\|^2} = \sqrt{|\lambda|^2\,\|v\|^2} = |\lambda|\,\|v\|.$$

b) $\|v\| = \langle v, v\rangle^{\frac{1}{2}} = \sqrt{v_1^2 + v_2^2} \geq 0$.
If $\|v\| = 0$, then $v_1^2 + v_2^2 = 0$, so $v_1^2 = -v_2^2$, but they're both non-negative, so $v_1^2 = v_2^2 = 0$. This implies $v = 0$.
Clearly, if $v = 0$ then $\|v\| = 0$.

c)

$$\begin{aligned}\|v+u\|^2 &= \langle v+u, v+u\rangle \\ &= \langle v,v\rangle + \langle v,u\rangle + \langle u,v\rangle + \langle u,u\rangle \\ &= \langle v,v\rangle + 2\langle v,u\rangle + \langle u,u\rangle \\ &\leq \|v\|^2 + 2\,|\langle v,u\rangle| + \|u\|^2 \\ &\leq \|v\|^2 + 2\|v\|\|u\| + \|u\|^2 \\ &= (\|v\| + \|u\|)^2\end{aligned}$$

so $\|v+u\| \leq \|v\| + \|u\|$, as we had claimed.

This finishes the proof. □

Corollary 2.3.3.10. *For any vector $v \in \mathbb{R}^2$ we have that $\langle v, v\rangle \geq 0$. In particular, $\langle v, v\rangle = 0$ if, and only if, $v = 0$.*

Proof

This follows trivially from the previous proposition and the fact that $\langle v, v\rangle = \|v\|^2$. □

Corollary 2.3.3.11. *For any vector $v \in \mathbb{R}^2$, $\left\| \dfrac{v}{\|v\|} \right\| = 1$.*

Proof

This is trivial:

$$\left\| \frac{v}{\|v\|} \right\| = \left| \frac{1}{\|v\|} \right| \|v\| = \frac{\|v\|}{\|v\|} = 1$$

where the first equality follows from (1.) in the proposition above. □

It may not seem much, but this allows us to define:

Definition 2.3.3.12. *Let $v, u \in R^2$ be any two vectors. We define the **distance between** v **and** u to be $d(v,u) \in \mathbb{R}$, which is given by*

$$d(v,u) := \|v - u\|.$$

Proposition 2.3.3.13. *The distance we defined above satisfies the following properties for all $v, u, w \in \mathbb{R}^2$:*

a) $d(v,u) = d(u,v)$;

b) $d(v,u) \geq 0$, *with the equality happening precisely when* $v = u$;

c) $d(v,u) \leq d(v,w) + d(w,u)$.

Proof

a)

$$d(v,u) = \|v-u\| = \|-(u-v)\| = \|u-v\| = d(u,v).$$

b) This is obvious from the definition.

c)

$$\begin{aligned} d(v,u) &= \|v-u\| \\ &= \|v-w+w-u\| \\ &= \|(v-w)+(w-u)\| \\ &\leq \|v-w\| + \|w-u\| = d(v,w) + d(w,u). \end{aligned}$$

□

Corollary 2.3.3.14. *The letter (c) above is an equality if, and only if, v, u, w lie on the same line and w is between v and u.*

Proof

We'll prove this for lines through zero. The proof for an arbitrary line follows directly and is left as an exercise to the reader.
Let $r = \mathbb{R}v$ be a line through zero, such that $u \in r$. Then $u = \mu v$ for some $\mu \in \mathbb{R}$. Notice that we can assume that all three of $0, v, u$ are different, cause, otherwise, the result is trivially true. This implies $\mu \neq 1$ and both different from 0.
Then

$$\begin{aligned} d(v,u) &= \|v-u\| \\ &= \|v-\mu v\| \\ &= \|(1-\mu)v\| = |1-\mu|\,\|v\| \end{aligned}$$

and similarly $d(v,0) = \|v\|$ and $d(u,0) = \|u\| = |\mu|\,\|v\|$.
Therefore, if v is between 0 and u, it means that $0 < 1 < \mu$, so

$$\begin{aligned} d(0,u) &= |\mu|\,\|v\| \\ &= (\mu)\|v\| \\ &= (\mu - 1 + 1)\|v\| \\ &= (\mu-1)\|v\| + \|v\| \\ &= |\mu - 1|\,\|v\| + \|v\| = d(v,u) + d(0,v) \end{aligned}$$

and we see the equality holds.

Conversely, assume $v, u, 0 \in \mathbb{R}^2$ are such that the equality holds, that is,

$$d(0,u) = d(v,0) + d(u,v).$$

Then if either of these is zero, the result follows - for instance, $d(v,0) = \|v\|$ implies $v = 0$ (so they lie in the same line), and since between any two points there's always a line, this implies that u is also in a line with v and 0.
So we can assume all of these distances are non-zero, which means that u and v are non-zero vectors.
But we've already shown that

$$\|v-u\|^2 = \|v\|^2 - 2\langle v,u\rangle + \|u\|^2,$$

but the equality

$$d(0,u) = d(v,0) + d(u,v)$$

can be rewritten as

$$d(u,v) = d(0,u) - d(0,v)$$

which gives us

$$\begin{aligned} d(u,v) &= d(0,u) - d(0,v) \\ \|v-u\| &= \|u\| - \|v\| \\ \|v-u\|^2 &= (\|u\| - \|v\|)^2 \\ \|v-u\|^2 &= \|u\|^2 - 2\|u\|\|v\| + \|v\|^2 \end{aligned}$$

and by comparing the two expressions for $\|v-u\|^2$ we see that $\langle v,u\rangle = \|v\|\|u\|$.
But we've also shown that, for any two vectors (in particular for v,u), $\langle v,u\rangle = \|v\|\|u\|\cos\theta$, where θ is the angle between v,u.
Therefore, once again by comparing the two expressions for $\langle v,u\rangle$, we see that $\cos\theta = 1$, so $\theta = 0$.
This tells us that $v \parallel u$ and finishes the proof. □

Example(s)

Let $v = (2,3)$, $u = (4,-1)$. Then $d(v,u) = \|v-u\| = \|(-2,4)\| = \sqrt{(-2)^2+4^2} = \sqrt{4+16} = \sqrt{20} = 2\sqrt{5}$.
Notice that $d(e_1,e_2) = \|(1,0)-(0,1)\| = \sqrt{1^2+1^2} = \sqrt{2}$ which is precisely the diagonal of a square whose side length is 1.
These examples show that this notion of distance is precisely the notion of distance that's intrinsic to the geometry of the Euclidean plane. We just have a proper way of computing this distances now.

Remark 2.3.3.15

The third property above is called the **Triangle Inequality**. It basically says that in any triangle you can't have a side that's bigger than the sum of the other two.

Now that we have an algebraic notion of distance, you know what we're gonna do next, right?

2.3.4 Geometry results via linear algebra

In this section we're gonna prove some results about the geometry of $\mathbb{R}^2$.

For instance, we've already proven that any two lines $r, s \subseteq \mathbb{R}^2$ can be described, uniquely, by two pairs a vectors v_1, v_2, u_1, u_2 such that $r = \mathbb{R}v_1 + v_2$ and $s = \mathbb{R}u_1 + u_2$. Let us now prove that this is indeed a good description of lines:

Proposition 2.3.4.1. *Let $r, s \subseteq \mathbb{R}^2$ be two different lines through zero in $\mathbb{R}^2$. Then $r \cap s = \{0\}$.*

Proof

Let $r = \mathbb{R}v$ and $s = \mathbb{R}u$ for some $v, u \in \mathbb{R}^2$. We already know that $0 \in r \cap s$.
Assume there is some $x \in r \cap s$ such that $x \neq 0$. This means that $r = \lambda v = \mu u$, for some $\lambda, \mu \in \mathbb{R}$, both different from 0. But since λ is real, $\lambda v = \mu u$ implies $v = \frac{\mu}{\lambda} u$ and therefore $r = \mathbb{R}v = \mathbb{R}u = s$ (since $v \parallel u$).
But, by hypothesis, we're assuming that $r \neq s$, so this cannot happen. This means that we cannot assume that there's some $x \neq 0$ in $r \cap s$. It follows that $r \cap s = \{0\}$, which proves the result. □

Corollary 2.3.4.2. *Let $r, s \subseteq \mathbb{R}^2$ be two different lines in $\mathbb{R}^2$. Then $r \cap s$ is either $\varnothing$ or a single point.*

Proof

Let $r = \mathbb{R}v_1 + v_2$ and $s = \mathbb{R}u_1 + u_2$ for some $v_1, v_2, u_1, u_2 \in \mathbb{R}^2$.
If $r \cap s = \varnothing$, there's nothing to do.
Assume, then, that $r \cap s \neq \varnothing$. That means there is at least one $x \in r \cap s$. Assume, then, that there is also $y \in r \cap s$. We will show that $x = y$.
If that's the case (that is, $x, y \in r \cap s$), then there are real numbers $\lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{R}$ such that

$$x = \lambda_1 v_1 + v_2 = \mu_1 u_1 + u_2$$

$$y = \lambda_2 v_1 + v_2 = \mu_2 u_1 + u_2.$$

But then, $x - y = (\lambda_1 - \lambda_2)v_1 = (\mu_1 - \mu_2)u_1$ and, therefore, $x - y \in \mathbb{R}v_1 \cap \mathbb{R}u_1$.
If $\mathbb{R}v_1 = \mathbb{R}u_1$, then, by definition, $r \parallel s$ and therefore $r \cap s = \varnothing$. But this contradicts our assumption that there are $x, y \in r \cap s$. Therefore, $\mathbb{R}v_1 \neq \mathbb{R}u_1$.
This means that we can use the proposition to conclude that $\mathbb{R}v_1 \cap \mathbb{R}u_1 = \{0\}$. But we've just shown that $x - y \in \mathbb{R}v_1 \cap \mathbb{R}u_1$. It follows that $x - y = 0$, and so $x = y$.
This shows that if we have two points in $r \cap s$, then they are in fact the same point.
So $r \cap s$ has, at most, one point, which ends the proof. □

This allows us to prove some very useful results:

Definition 2.3.4.3. *A **rhombus** is a set of four points A, B, C, D in the Euclidean plane such that $\overline{AB} = \overline{BC} = \overline{CD} = \overline{DA}$.*

In other words, a rhombus is a quadrilateral which has four congruent sides.

Lemma 2.3.4.4. *Let $\mathbb{R}v, \mathbb{R}u$ be two lines through zero. Then $\mathbb{R}v \perp \mathbb{R}u$ if, and only if, $v \perp u$.*

Proof

Assume $\mathbb{R}v \perp \mathbb{R}u$. This means that any two vectors $v' \in \mathbb{R}v$ and $u' \in \mathbb{R}u$ are perpendicular. In particular, $v \in \mathbb{R}v$ and $u \in \mathbb{R}u$ implies that $v \perp u$.

Conversely, assume that $v \perp u$. Take now $v' \in \mathbb{R}v$ and $u' \in \mathbb{R}u$. We want to show that $v' \perp u'$.
Since $v' \in \mathbb{R}v$ and $u' \in \mathbb{R}u$, there are real numbers $\lambda, \mu \in \mathbb{R}$ such that $v' = \lambda v$ and $u' = \mu u$, by definition.
But then:

$$\langle v', u' \rangle = \langle \lambda v, \mu u \rangle = \lambda\mu \langle v, u \rangle = 0$$

since $v \perp u$. This shows that $v' \perp u'$.
The result follows. □

Corollary 2.3.4.5. *The diagonals of a rhombus are perpendicular to each other.*

Proof

First, notice that any rhombus can be considered as having vertices $0, v, v+u, u$ where $\|v\| = \|u\|$.
Take now the lines $\mathbb{R}(v+u)$ and $\mathbb{R}(v-u)+v$, which contain the diagonals of the rhombus. To show that the diagonals are perpendicular, it suffices to show that these two lines are perpendicular.
But notice that since $\mathbb{R}(v-u)+v \parallel \mathbb{R}(v-u)$, it suffices to show that $\mathbb{R}(v+u) \perp \mathbb{R}(v-u)$.
Finally, notice that, by the preceding lemma, showing $\mathbb{R}(v+u) \perp \mathbb{R}(v-u)$ is the same as showing that $(v+u) \perp (v-u)$.
This can be done by a simple computation:

$$\begin{aligned} \langle v+u, v-u \rangle &= \langle v, v \rangle - \langle v, u \rangle + \langle u, v \rangle - \langle u, u \rangle \\ &= \|v\|^2 - \langle v, u \rangle + \langle v, u \rangle - \|u\|^2 = 0 \end{aligned}$$

since $\|v\| = \|u\|$.
This shows that $(v+u) \perp (v-u)$, which, by the preceding observations, ends the proof. □

Definition 2.3.4.6. *A **rectangle** is a set of four points A, B, C, D in the Euclidean plane such that $A\hat{B}C$, $B\hat{C}D$, $C\hat{D}A$ and $D\hat{A}B$ are right angles.*

In other words, a rectangle is a quadrilateral which has four congruent angles (which are, therefore, right angles).

Lemma 2.3.4.7. *Let $v, u \in \mathbb{R}^2$ be two non-null vectors. Then $(v+u) \parallel (v-u)$ if, and only if, $v \parallel u$.*

Proof

Assume $(v+u) \parallel (v-u)$ and consider $\mathbb{R}(v+u)$ and $\mathbb{R}(v-u)$. Since they are lines through zero, this means they are equal, $\mathbb{R}(v+u) = \mathbb{R}(v-u)$. Let's call $r = \mathbb{R}(v+u) = \mathbb{R}(v-u)$.
We've already shown that lines through zero are closed under addition, so $(v+u)+(v-u) \in r$.
But $(v+u)+(v-u) = 2v \in \mathbb{R}v$, and $v \neq 0$ implies, since two lines through zero meet only at zero, that $\mathbb{R}v = r$
Similarly, $(v+u)-(v-u)$ must lie in r, but $(v+u)-(v-u) = 2u \in \mathbb{R}u$, and, once again, since $u \neq 0$ and two lines through zero meet only at zero, this implies that $\mathbb{R}u = r$.
Combining these two, we see that $\mathbb{R}v = r = \mathbb{R}u$, and so $v \parallel u$.

Conversely, if $v \parallel u$, this means that $\mathbb{R}v = \mathbb{R}u$, and so, since lines through zero are closed under addition, this means that both $v+u$ and $v-u$ lie in $\mathbb{R}v = \mathbb{R}u$. Therefore $v+u \parallel v-u$, which ends the proof. □

Corollary 2.3.4.8. *The diagonals of a rectangle meet at their midpoint.*

Proof

First, notice that any rectangle can be considered as having vertices $0, v, v+u, u$ where $v \in \mathbb{X}$ and $u \in \mathbb{Y}$.
Since $v \nparallel u$, by hypothesis, the preceding lemma tells us that $(v+u) \nparallel (v-u)$.
In particular, since $\mathbb{R}(v-u)+v \parallel \mathbb{R}(v-u)$ this implies that the line $\mathbb{R}(v-u)+v$ crosses the line $\mathbb{R}(v+u)$, since they're not parallel.
Now take the midpoint of the line segment $0v+u$ - that is, $\frac{1}{2}(v+u)$. We claim that $\frac{1}{2}(v+u) \in \mathbb{R}(v-u)+v$.
Indeed,

$$\begin{aligned}\frac{1}{2}(v+u) &= \frac{1}{2}v + \frac{1}{2}u \\ &= \frac{1}{2}v + \frac{1}{2}u - v + v \\ &= -\frac{1}{2}v + \frac{1}{2}u + v \\ &= -\left(\frac{1}{2}v - \frac{1}{2}u\right) + v = -\frac{1}{2}(v-u) + v \in \mathbb{R}(v-u)+v\end{aligned}$$

so $\frac{1}{2}(v+u)$ lies in both $\mathbb{R}(v+u)$ and $\mathbb{R}(v-u)+v$. Since two lines meet at, at most, one point, it follows that $\mathbb{R}(v+u) \cap \mathbb{R}(v-u)+v = \{\frac{1}{2}(v+u)\}$.

Finally, it remains to show that $\frac{1}{2}(v+u)$ is the midpoint of the line segment vu.
To do this, let's simply compute:

$$\begin{aligned} d\left(\frac{1}{2}(v+u), v\right) &= \left\| \left(\frac{1}{2}(v+u)\right) - v \right\| \\ &= \left\| \frac{1}{2}v + \frac{1}{2}u - v \right\| \\ &= \left\| -\frac{1}{2}v + \frac{1}{2}u \right\| \\ &= \left\| -\frac{1}{2}(v-u) \right\| \\ &= \left| -\frac{1}{2} \right| \|v-u\| \\ &= \frac{1}{2}d(v,u) = \frac{1}{2}\overline{vu}. \end{aligned}$$

It follows that $\frac{1}{2}(v+u)$ is indeed the midpoint of vu, which ends the proof. □

Definition 2.3.4.9. *Given $x \in \mathbb{R}^2$ any vector and $r \in \mathbb{R}$ a non-negative real number, a **circle of radius** r **and center** x is the set $\mathcal{C}(x,r)$ given by*

$$\mathcal{C}(x,r) := \{v \in \mathbb{R}^2 \mid d(v,x) = r\}.$$

In other words, a circle is the set of all points that are a fixed distance from a fixed point.

Definition 2.3.4.10. *Given a line $r \subseteq \mathbb{R}^2$ and a point $v \in \mathbb{R}^2$ we define the **distance between** v **and** r to be the real number $d(v,r) \in \mathbb{R}$ given by*

$$d(v,r) := \min_{u \in r}\{d(v,u)\}.$$

That is, the distance between a point and a line is the length of the smallest line segment connecting the point to the line.

Lemma 2.3.4.11. *Let $r \subseteq \mathbb{R}^2$ be a line and $v \in \mathbb{R}^2$ be any point. Then the distance between r and v is well-defined.*

Proof

We need to prove that no matter which r and which v we choose, $d(v,r)$ always exists.
To do this, consider the following construction:

(i.) Take t the line perpendicular to r through v;

(ii.) Mark $u \in r \cap t$.

We claim that $d(v,u) = \min_{w \in r}\{d(v,w)\}$.

Notice that $\mathbb{R}$ is totally ordered, so it suffices to show that there's no one smaller than $d(v,u)$.
Take any $w \in r$ and consider the triangle vuw.
Since, by construction $vu \perp r$, it follows that vuw is a right-triangle with hypotenuse vw.
But, by Pythagoras' Theorem, we know that $d(v,w)^2 = d(v,u)^2 + d(u,w)^2$.
In particular, $d(v,w)^2 \geq d(v,u)^2$, which implies (since distances are always non-negative) that $d(v,w) \geq d(v,u)$.
This shows that $d(v,u) \leq d(v,w)$ for all $w \in r$, and so $d(v,u) = \min_{w \in r}\{d(v,w)\}$, which ends the proof. □

Corollary 2.3.4.12. *Let $v \in \mathbb{R}^2$ and $r \subseteq \mathbb{R}^2$ be any line. Let also $u \in r$ be such that $d(v,u) = d(v,r)$. Then $v - u \perp r$.*

Proof

By the same argument as before: Assume $w \in r$ is the point obtained by the construction of the preceding proof. So $v - w \perp r$, by construction.
It follows that vuw is a right-triangle at w. Once again, by using Pythagoras' Theorem, we see that $d(v,w)^2 = d(v,u)^2 + d(u,w)^2$.
But $d(v,u) = d(v,w)$ implies $d(u,w)^2 = 0$, which is simply $d(u,w) = 0$. Now, by the properties of distance, this happens if, and only if, $u = w$.
Since $v - w \perp r$ by construction, this implies $v - u \perp r$, which ends the proof. □

Lemma 2.3.4.13. *Let $r = \mathbb{R}v + u \subseteq \mathbb{R}^2$ be any line, and $R \in \mathbb{R}$ any non-negative number. Then the set*

$$\mathcal{D}(R,r) := \{w \in \mathbb{R}^2 \mid d(w,r) = R\}$$

is just the set $s \cup t$ where $s = r + \overrightarrow{R}$ and $t = r - \overrightarrow{R}$, and $\overrightarrow{R} \in \mathbb{R}^2$ is any vector such that $\overrightarrow{R} \perp r$ and $\|\overrightarrow{R}\| = R$.

Proof

Take $x \in \mathcal{D}(R,r)$. This means that there's some $y \in r$ such that $d(x,y) = R$.
Define then $\overrightarrow{R} := x - y$. By definition, $\|\overrightarrow{R}\| = R$ and, by the preceding corollary, $\overrightarrow{R} \perp r$.
Since $y \in r = \mathbb{R}v + u$ and $\overrightarrow{R} = x - y$, we see that

$$x = y + \overrightarrow{R} \in r + \overrightarrow{R} = s.$$

This shows that $\mathcal{D}(R,r) \subseteq s \cup t$.

Conversely, take $x \in s \cup t$ and any $\overrightarrow{R} \in \mathbb{R}^2$ such that $\|\overrightarrow{R}\| = R$ and $\overrightarrow{R} \perp r$. This means that $x = \lambda v + u \pm \overrightarrow{R}$ for some $\lambda \in \mathbb{R}$.

Take now $y = \lambda v + u$. Clearly, $y \in r$. Now,

$$\begin{aligned} d(x,y) &= \|x - y\| \\ &= \|(\lambda v + u \pm \overrightarrow{R}) - (\lambda v + u)\| \\ &= \|\pm\overrightarrow{R}\| \\ &= \|\overrightarrow{R}\| = R. \end{aligned}$$

Notice, however, that $x - y = \pm\overrightarrow{R} \perp r$, so, by the preceding lemma, we see that $x \in \mathcal{D}(R, r)$. This shows that $s \cup t \subseteq \mathcal{D}(R, r)$, and, therefore, $\mathcal{D}(R, r) = s \cup t$, which ends the proof. □

Lemma 2.3.4.14. *If a line passes through the center of a circle, then it crosses the circle.*

Proof

Let $\mathcal{C}(x, r)$ be a circle for some $x \in \mathbb{R}^2$ and some non-negative $r \in \mathbb{R}$, as well as $s := \mathbb{R}v + u$ be a line for some $v, u \in \mathbb{R}^2$.
If $x \in s$ this means that $x = \lambda v + u$ for some $\lambda \in \mathbb{R}$.
We claim that $x + v' \in \mathcal{C}(x, r)$, for $v' := \frac{r}{\|v\|}v$.
Indeed,

$$\begin{aligned} d(x + v', x) &= \|(x + v') - x\| \\ &= \|v'\| \\ &= \left\| \frac{r}{\|v\|} v \right\| \\ &= \left| \frac{r}{\|v\|} \right| \|v\| \\ &= \frac{r}{\|v\|}\|v\| = r \end{aligned}$$

Now,

$$x + v' = (\lambda v + u) + \left(\frac{r}{\|v\|}v\right) = \left(\lambda + \frac{r}{\|v\|}\right) v + u \in \mathbb{R}v + u$$

and so $x+v'$ is a point in $\mathbb{R}v+u$ that is at distance r from x - this means that $x+v' \in r \cap \mathcal{C}(x, r)$ and, so, the proof is done. □

Lemma 2.3.4.15. *If a line contains a point in the inside of a circle, then it crosses the circle.*

Proof

Consider the line $s' := \mathbb{R}v$, the point $x' := x - u$ and the circle $\mathcal{C}' := \mathcal{C}(x', r)$. Then for all $y' \in s'$ we have

$$\begin{aligned} d(y', x') &= \|y' - x'\| \\ &= \|\lambda v - (x - u)\| \\ &= \|\lambda v + u - x\| \\ &= \|y - x\| = d(y, x) \end{aligned}$$

where $y = \lambda v + u \in s$. This shows that $d(\lambda v + u, x) = d(\lambda v, x - u)$ for all $\lambda \in \mathbb{R}$.
This means that we can rephrase the problem to work only with lines through zero.
So, assume $u = 0$, that is, s is a line through zero.
Take now any $y \in s$ such that $d(y, x) < r$ (that is, y is inside the circle $\mathcal{C}(x, r)$). Since s is a line through zero, every scalar multiple of y also belongs to s.
Assume that $d(y', x) < r$ for all $y' \in s$. Then:

$$d(y, y') \leq d(y, x) + d(y', x) \leq r + r = 2r$$

so we're saying that no matter which point $y' \in s$ we pick, its distance from y can never exceed $2r$.
But that's absurd - take, for instance, $y' = y + \frac{3r}{\|y\|} y = \left(1 + \frac{3r}{\|y\|}\right) y \in s$.
Then

$$\begin{aligned} d(y, y') &= \|y - y'\| \\ &= \left\| y - \left(1 + \frac{3r}{\|y\|}\right) y \right\| \\ &= \left\| \left(-\frac{3r}{\|y\|}\right) y \right\| \\ &\left| -\frac{3r}{\|y\|} \right| \|y\| \\ &= \frac{3r}{\|y\|} \|y\| = 3r \end{aligned}$$

and so $d(y, y') = 3r > 2r$.
This shows that there are some (infinitely many) points in s which are outside of $\mathcal{C}(x, r)$.
Since lines are connected and there are both points of s which are inside $\mathcal{C}(x, r)$ and points of s which are outside $\mathcal{C}(x, r)$, then s must cross $\mathcal{C}(x, r)$ at some point.
This ends the proof. □

Now, let's introduce a new tool to allow us to prove even more results:

Lemma 2.3.4.16. *Let $c \in \mathbb{R}^2$ and $r \in \mathbb{R}$ any non-negative real number. Then $\mathcal{C}(c, r) = \mathcal{C}(0, r) + c$, that is, any point in $\mathcal{C}(c, r)$ is just a point in $\mathcal{C}(0, r)$ added to c.*

Proof

Call $\mathcal{C} = \mathcal{C}(0, r)$ and $\mathcal{C}' = \mathcal{C}(c, r)$. We want to show that $\mathcal{C} + c = \mathcal{C}'$. Fix, then, $c = (c_1, c_2)$. Take then $v = (v_1, v_2) \in \mathcal{C}$, so $d(v, 0) = r$. Therefore, $v + c = (v_1 + c_1, v_2 + c_2) \in \mathcal{C} + c$. Then:

$$d(v + c, c) = \|(v + c) - c\| = \|v\| = \|v - 0\| = d(v, 0) = r$$

and so $v + c \in \mathcal{C}'$, by definition.
This shows that $\mathcal{C} + c \subseteq \mathcal{C}'$.

Conversely, take $v \in \mathcal{C}'$, that is, $d(v, c) = r$. We claim that $v - c \in \mathcal{C}$. Indeed:

$$d(v - c, 0) = \|(v - c) - 0\| = \|v - c\| = d(v, c) = r$$

so $v - c \in \mathcal{C}$.
Clearly, then, $v = (v - c) + c \in \mathcal{C} + c$ so $\mathcal{C}' \subseteq \mathcal{C} + c$.
These two together show that $\mathcal{C}' = \mathcal{C} + c$ and the result follows. □

Lemma 2.3.4.17. *Let $\mathcal{C}_1 = \mathcal{C}(c_1, r_1), \mathcal{C}_2 = \mathcal{C}(c_2, r_2)$ be any two circles, and let $I = \mathcal{C}_1 \cap \mathcal{C}_2$. Then for all $v \in \mathbb{R}^2$ we have that*

$$I \cong (\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v).$$

In other words, the number of meeting points of two circles doesn't depend on where they are.

Proof

First, we'll prove that $(\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v) = I + v$. Then we'll argue that $I + v \cong I$.
Take $x \in (\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v)$. This means that both $d(x, c_1 + v) = r_1$ and $d(x, c_2 + v) = r_2$ hold, by definition.
But then, since

$$\begin{aligned} r_1 &= d(x, c_1 + v) \\ &= \|x - (c_1 + v)\| \\ &= \|x - c_1 - v\| \\ &= \|(x - v) - c_1\| = d(x - v, c_1) \end{aligned}$$

and a similar equation holds for c_2 (that is, $r_2 = d(x, c_2 + v) = d(x - v, c_2)$) we see that $x - v \in I$, so $x \in I + v$.
This shows $(\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v) \subseteq I + v$.

Conversely, given any $x \in I + v$, this means that $x - v \in I$, and so

$$\begin{aligned} r_1 &= d(x - v, c_1) \\ &= \|(x - v) - c_1\| \\ &= \|x - (c_1 + v)\| = d(x, c_1 + v) \end{aligned}$$

and a similar equation holds for c_2 (that is, $r_2 = d(x - v, c_2) = d(x, c_2 + v)$). But this implies that $x \in (\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v)$, and so $I + v \subseteq (\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v)$.
This proves that $I + v = (\mathcal{C}_1 + v) \cap (\mathcal{C}_2 + v)$.

Finally, consider the function $f : I \to I + v$ given by $x \mapsto x + v$, as well as the function $g : I + v \to I$ given by $y \mapsto y - v$.
Clearly, $f \circ g = \mathrm{id}_{I+v}$ and $g \circ f = \mathrm{id}_I$, so $g = f^{-1}$ and f is an isomorphism.
This finishes the proof. □

Corollary 2.3.4.18. *Any two circumferences meet at either no points, a single point or two points.*

Proof

Let $\mathcal{C}_1 = \mathcal{C}(0, r_1)$ and $\mathcal{C}_2 = \mathcal{C}(c, r_2)$ for some $c \in \mathbb{R}^2$, $r_1, r_2 \in \mathbb{R}$ non-negative, $D = d(0, c)$, $R = r_1 + r_2$ and $R' = r_1 - r_2$. Let also $r_1 \geq r_2$.
Using the preceding lemma, it suffices to prove the result for this case, because every other case is just a translation of this case by some $T \in \mathbb{R}^2$.
Finally, we'll use $I = \mathcal{C}_1 \cap \mathcal{C}_2$ just to save on notation.
Then, we have three cases:

- $\underline{D > R \text{ or } D < R'}$:

  If there is a point $v \in I$, then $0, v, c$ are either colinear, or not.

  Assume they aren't. Then

  $$d(0, c) \leq d(0, v) + d(c, v),$$

  but the LHS is D and the RHS is R. This contradicts $D > R$.

  On the other hand, if they are in a line, then either v is between $0, c$, 0 is between v, c or c is between $0, v$.

  In each case we get

  $$\begin{aligned} d(0, c) &= d(0, v) + d(v, c) \\ d(v, c) &= d(0, v) + d(0, c) \\ d(0, v) &= d(0, c) + d(v, c), \end{aligned}$$

  respectively.

But the first is simply $D = R$, which contradicts $D > R$; the second one is simply $r_2 = r_1 + D$ which is a contradiction, since we're assuming $r_1 \geq r_2$; and the third one is simply $D = R'$, which contradicts $D < R'$.

So we have proven that if there were a $v \in I$ then it doesn't matter if it lies in the line between $0, c$ or not - this v always leads to a contradiction. So there can be no elements in I - that is, $I = \varnothing$.

- $\underline{D = R \text{ or } D = R'}$:

  Assume, like before, that $v \in I$. Then we can do the same proof so see that v cannot not be colinear with $0, c$.

  However, contrary to the previous case, v *can* be colinear with $0, c$:

  Indeed, either
  $$d(0, c) = d(0, v) + d(v, c)$$
  or
  $$d(0, v) = d(0, c) + d(v, c)$$
  hold. In the first case, we see that $D = R$ and v is between 0 and c, and in the second case we see that $D = R'$ and c is between 0 and v.

  All that's left is to show that there is at least one $v \in I$. Notice, however, that it suffices to show that there's a point v in either $\mathcal{C}_1$ or $\mathcal{C}_2$ such that v is between $0, c$ or c is between $0, v$.

  To do that, take the line $r = \mathbb{R}c$. This is a line through the centers of $\mathcal{C}_1$ and $\mathcal{C}_2$ and, as such, crosses both circles.

  Take $v \in \mathcal{C}_1 \cap r$. Then either v is between $0, c$ (yay!), c is between $0, v$ (yay!!) or 0 is between v, c (nay).

  If 0 is between v, c, we claim that 0 is not between $-v, c$.

  Indeed, if 0 is between v, c then
  $$d(v, c) = d(v, 0) + d(c, 0) = \|v\| + \|c\|.$$
  Assume that 0 is also between $-v, c$. Then
  $$d(-v, c) = d(-v, 0) + d(c, 0) = \|-v\|\|c\| = r_1\|c\| = d(v, c),$$
  but $v \in \mathbb{R}c$ implies $v = \lambda c$, so $-v = -\lambda c$.

  Therefore
  $$d(v, c) = \|v - c\| = |\lambda - 1| \, \|c\|$$
  and
  $$d(-v, c) = \|-v - c\| = |-1| \, \|v + c\| = |\lambda + 1| \, \|c\|,$$
  so they're equal if, and only if, $|\lambda - 1| = |\lambda + 1|$.

But then:

$$|\lambda - 1| = |\lambda + 1|$$
$$(\lambda - 1)^2 = (\lambda + 1)^2$$
$$\lambda^2 - 2\lambda + 1 = \lambda^2 + 2\lambda^1$$
$$-2\lambda = 2\lambda$$
$$-\lambda = \lambda$$
$$\lambda = 0$$

which would imply (since $d(v, 0) = r_1$, by definition) that $r_1 = 0$. But this contradicts our taking $r_1 > 0$. Therefore, 0 cannot be both between v, c and $-v, c$. This means that either one of these hold:

- v is between $0, c$;
- c is between $0, v$;
- $-v$ is between $0, c$;
- c is between $0, -v$.

But no matter which one of these holds, this shows that either v or $-v$ is in I - so there's always at least one point in I.

Finally, this reasoning also proves that v (or $-v$) is the unique point in I.

Indeed, assume $v, u \in I$ and that they're both between $0, c$. Then the equation

$$d(0, c) = d(0, v) + d(v, c) = d(0, u) + d(u, c)$$

holds. But since $v, u \in I$, this implies $v, u \in \mathcal{C}_1$, so $d(0, v) = d(0, u)$. Since $d(0, c)$ is fixed, it follows that $d(v, c) = d(u, c)$, and since we've already shown that they must lie in the line $\mathbb{R}c$, this implies $v = u$.

If, on the other hand, one of them (say, v) is between $0, c$ and c is between the other (u) and 0, then both

$$d(0, c) = d(0, v) + d(v, c)$$

and

$$d(0, u) = d(0, c) + d(c, u)$$

hold. So we can replace the expression for $d(0, c)$ given by the first equation in the second equation to derive

$$d(0, u) = (d(0, v) + d(v, c)) + d(c, u).$$

But $v, u \in I$ implies $v, u \in \mathcal{C}_1$, so $d(v, 0) = d(u, 0) = r_1$. Therefore, the above equality becomes $d(v, c) + d(c, u) = 0$, which is a contradiction, since both $d(v, c)$ and $d(c, u)$ are positive.

Therefore, the only possible case for $v, u \in I$ is $v = u$.

This shows that, in the case of $D = R$ or $D = R'$, we can conclude that I is a single point.

- $\underline{R' < D < R}$:

  In this case, by definition, we can construct a triangle whose sides measure r_1, r_2 and D. Indeed:

$$r_1 + r_2 = R > D$$

$$r_1 + D > r_2$$

$$r_2 + D > r_2 + R' = r_2 + (r_1 - r_2) = r_1$$

  where the first inequality follows by our assumption that $R > D$, the second inequality follows from $r_1 > r_2$ and the third inequality follows from $D > R'$.

  In particular, we can construct this triangles as having vertices $0, c$ since $d(0, c) = D$. This gives us a third point v such that $d(0, v) = r_1$ (and so $v \in \mathcal{C}_1$) and $d(v, c) = r_2$ (and so $v \in \mathcal{C}_2$). This shows us that $v \in I$.

  Take now $v' \in \mathbb{R}^2$ such that $d(v, \mathbb{R}c) = d(v', \mathbb{R}c)$, $d(v', 0) = d(v, 0)$ and $d(v, c) = d(v', c)$. This is done by taking the perpendicular to $\mathbb{R}c$ through v, marking its intersection point with $\mathbb{R}c$ (say, w), and then defining $v' = w - (v - w)$.

  Now it's easy to see that both $v, v' \in \mathbb{R}(v - w) + w$, that

$$d(v', \mathbb{R}c) = d(v', w) = \|v' - w\| = \|v - w\| = d(v, w) = d(v, \mathbb{R}c),$$

  that

$$d(v', 0)^2 = d(v', w)^2 + d(w, 0)^2 = d(v, w)^2 + d(w, 0)^2 = d(v, 0)^2$$

  so $d(v, 0) = d(v', 0)$, and that

$$d(v', c)^2 = d(v', w)^2 + d(w, c)^2 = d(v, w)^2 + d(w, c)^2 = d(v, c)^2$$

  so $d(v, c) = d(v', c)$.

  It follows that $v' \in I$ as well, and $v \neq v'$ (since $v' - v = 2(w - v) \neq 0$ because $v \notin \mathbb{R}c$, and hence $v \neq w$).

  It remains to show that there's no other point in I. But this is easy:

  Take $u \in I$. Then, as we've shown, $d(u, \mathbb{R}c) = d(v, \mathbb{R}c) = d(v', \mathbb{R}c)$. But we've already proven that the locus of points at a fixed distance from a given line is two lines. This means that either $u \in \mathbb{R}c + v$ or $u \in \mathbb{R}c + v'$.

In either case there's a point $w' \in \mathbb{R}c$ such that $d(u, \mathbb{R}c) = d(u, w')$ So if $u \in \mathbb{R}c + v$ we get

$$\begin{gathered} d(u,c)^2 = d(u,w')^2 + d(w',c)^2 \\ d(v,c)^2 = d(v,w)^2 + d(w',c)^2 \\ d(w',c)^2 = d(v,c)^2 - d(v,w)^2 \\ d(w',c)^2 = d(w,c)^2 \end{gathered}$$

so $d(w,c) = d(w',c)$, and since $w, w' \in \mathbb{R}c$ this implies $w' = w$, which, in turn, implies $u = v$ (since v is the point where the line perpendicular to $\mathbb{R}c$ through w meets $\mathbb{R}c + v$, and u is the point what the line perpendicular to $\mathbb{R}c$ through w' meets $\mathbb{R}c + v$).

Doing a similar computation in the case $u \in \mathbb{R}c + v'$ shows us that $u = v'$.

This shows that any point $u \in I$ is either v or v' - and so $I = \{v, v'\}$.

This finishes the proof. □

Finally, to end this section we're gonna give some formulaic presentations to some of our geometric entities:

Definition 2.3.4.19. *For any vector $v = (v_1, v_2) \in \mathbb{R}^2$ we define $v^\perp \in \mathbb{R}^2$ to be the vector given by*

$$v^\perp := (-v_2, v_1).$$

Proposition 2.3.4.20. *For any $v, u \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$ the following hold:*

a) $\|v\| = \|v^\perp\|$;

b) $(v + u)^\perp = v^\perp + u^\perp$;

c) $(\lambda v)^\perp = \lambda v^\perp$.

Proof

Let, once and for all, $v = (v_1, v_2)$ and $u = (u_1, u_2)$. Now all of them follow by simple computations:

a)

$$\|v^\perp\| = \sqrt{(-v_2)^2 + v_1^2} = \sqrt{v_2^2 + v_1^2} = \|v\|.$$

b)

$$\begin{aligned}(v+u)^\perp &= (v_1+u_1, v_2+u_2)^\perp \\ &= (-(v_2+u_2), v_1+u_1) \\ &= (-v_2-u_2, v_1+u_1) \\ &= (-v_2, v_1) + (-u_2, u_1) = v^\perp + u^\perp.\end{aligned}$$

c)

$$\begin{aligned}(\lambda v)^\perp &= (\lambda v_1, \lambda v_2)^\perp \\ &= (-\lambda v_2, \lambda v_1) \\ &= \lambda(-v_2, v_1) = \lambda v^\perp.\end{aligned}$$

□

Lemma 2.3.4.21. *For any vector $v \in \mathbb{R}^2$, $\langle v, v^\perp \rangle = 0$, that is, $v \perp v^\perp$.*

Proof

This follows from a simple computation:

$$\langle v, v^\perp \rangle = v_1(-v_2) + v_2 v_1 = -v_1 v_2 + v_1 v_2 = 0.$$

□

Corollary 2.3.4.22. *For any non-null vectors $v, u \in \mathbb{R}^2$ we have that $v \perp u$ if, and only if, $u \in \mathbb{R}v^\perp$.*

Proof

If $u \in \mathbb{R}v^\perp$ then $u = \lambda v^\perp$, for some $\lambda \in \mathbb{R}$. Then

$$\langle v, u \rangle = \langle v, \lambda v^\perp \rangle = \lambda \langle v, v^\perp \rangle = \lambda \cdot 0 = 0$$

so $v \perp u$.

Conversely, if $v \perp u$, then, writing $v = (v_1, v_2)$ and $u = (u_1, u_2)$ we see that $0 = \langle v, u \rangle = v_1 u_1 + v_2 u_2$. Since v is non-null, either or both of v_1 and v_2 is non-null.
If v_1 is non-null, then $v_1 u_1 + v_2 u_2 = 0$ implies

$$u_1 = -\frac{v_2}{v_1} u_2,$$

so if we define $p = -\frac{v_2}{v_1}$ we see that

$$u = (u_1, u_2) = (p u_2, u_2) = u_2(p, 1)$$

so $u \in \mathbb{R}(p,1)$.
But, since v_1 is non-null,

$$v^\perp = (-v_2, v_1) = \left(-v_1\frac{v_2}{v_1}, v_1\right) = (v_1 p, v_1) = v_1(p,1)$$

so $v^\perp \in \mathbb{R}(p,1)$. This shows that $u \parallel v^\perp$ so $u \in \mathbb{R}v^\perp$.
Analogously, if v_2 is non-null, then we proceed the same way to conclude that

$$u_2 = -\frac{v_1}{v_2}u_1$$

so we can define $p' = -\frac{v_1}{v_2}$ and see that both u and $v^\perp$ lie in $\mathbb{R}(1,p')$, and, therefore, $u \in \mathbb{R}v^\perp$.
This ends the proof. □

Proposition 2.3.4.23. *Let $r := \mathbb{R}v$ be a line through zero. Then there are uniquely determined $a, b \in \mathbb{R}$ such that*

$$r = \{(x,y) \in \mathbb{R}^2 \mid ax + by = 0\}.$$

Proof

Let $v = (v_1, v_2)$. Then $v^\perp = (-v_2, v_1)$, by definition.
By what we've shown, it is clear that any $u \in \mathbb{R}v$ satisfies $u \perp v^\perp$. This means that for any $(x,y) \in \mathbb{R}v$ we have that $(-v_2)x + v_1 y = 0$, and, therefore,

$$\mathbb{R}v \subseteq \{(x,y) \in \mathbb{R}^2 \mid (-v_2)x + v_1 y = 0\}.$$

Conversely, if $(x,y) \in \mathbb{R}^2$ is such that $(-v_2)x + v_1 y = 0$ then, by definition, $\langle (x,y), v^\perp \rangle = 0$ so $(x,y) \perp v^\perp$ and so, by what we've shown, $(x,y) \in \mathbb{R}v$, so

$$\{(x,y) \in \mathbb{R}^2 \mid (-v_2)x + v_1 y = 0\} \subseteq \mathbb{R}v.$$

This ends the proof. □

Corollary 2.3.4.24. *Let $r = \mathbb{R}v + u$ be a line. Then there are some uniquely determined $a, b, c \in \mathbb{R}$ such that*

$$r = \{(x,y) \in \mathbb{R}^2 \mid ax + by = c\}.$$

Proof

Let $v = (v_1, v_2)$ as before and $u = (u_1, u_2)$, and consider the line $r' = r - u$. Then r' is a line parallel to r and through zero, so we know that

$$r' = \{(x,y) \in \mathbb{R}^2 \mid (-v_2)x + v_1 y = 0\}$$

by the preceding result.

But now, since $r' = r - u$, given any $(x, y) \in r$ we know that $(x, y) - (u_1, u_2) = (x - u_1, y - u_2) \in r'$ so $(-v_2)(x - u_1) + v_1(y - u_2) = 0$.
Now we can expand this:

$$\begin{aligned}(-v_2)(x - u_1) + v_1(y - u_2) &= 0\\ (-v_2)x + v_2u_1 + v_1y - v_1u_2 &= 0\\ (-v_2)x + v_1y &= v_1u_2 - v_2u_1\end{aligned}$$

so we see that $(x, y) \in r$ if, and only if, $(-v_2)x + v_1y = v_1u_2 - v_2u_1$. This proves the result. □

Proposition 2.3.4.25. *Let $r \in \mathbb{R}$ be any non-negative number. Then*

$$\mathcal{C}(0, r) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}.$$

Proof

Take any $v \in \mathcal{C}(0, r)$. Then

$$r = d(v, 0) = \|v - 0\| = \|v\| = \sqrt{v_1^2 + v_2^2}$$

therefore $v_1^2 + v_2^2 = r^2$ and $v \in \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$.
This shows $\mathcal{C}(0, r) \subseteq \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$.

Conversely, take $v \in \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$. So

$$d(v, 0) = \|v - 0\| = \|v\| = \sqrt{v_1^2 + v_2^2} = \sqrt{r^2} = r$$

which implies $v \in \mathcal{C}(0, r)$ and, therefore, $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\} \subseteq \mathcal{C}(0, r)$.
This shows

$$\mathcal{C}(0, r) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\},$$

which ends the proof. □

Corollary 2.3.4.26. *Let $r \in \mathbb{R}$ be any non-negative number and $c = (c_1, c_2) \in \mathbb{R}^2$ any point. Then*

$$\mathcal{C}(c, r) = \{(x, y) \in \mathbb{R}^2 \mid (x - c_1)^2 + (y - c_2)^2 = r^2\}$$

Proof

Let $\mathcal{C}' = \mathcal{C}(c, r)$ and $\mathcal{C} = \mathcal{C}(0, r)$.
By the preceding proposition, we know that $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$ and we also know that $\mathcal{C}' = \mathcal{C} + c$.
Combining these two we see that for every $v = (v_1, v_2) \in \mathcal{C}'$ we have that $v - c \in \mathcal{C}$ and so,

if we put $c = (c_1, c_2)$, this tells us that $(v_1 - c_1)^2 + (v_2 - c_2)^2 = r^2$.
This proves the result. □

2.4 Final results

2.4.1 Determinants

Let's retrace a few steps.

In one of the last corollaries of the previous section we showed that the line $\mathbb{R}v + u$ can be described as

$$\{(x, y) \in \mathbb{R}^2 \mid (-v_2)x + v_1 y = v_1 u_2 - v_2 u_1\},$$

where $v = (v_1, v_2)$ and $u = (u_1, u_2)$. But we've also seen that this line passes through zero if, and only if,

$$\mathbb{R}v + u = \{(x, y) \in \mathbb{R}^2 \mid (-v_2)x + v_1 y = 0\}.$$

But this is equivalent to saying that $v_1 u_2 - v_2 u_1 = 0$.

However, we know, from the general theory, that the line $\mathbb{R}v + u$ passes through zero if, and only if, $u \in \mathbb{R}v$. This allows us to prove the following result:

Lemma 2.4.1.1. *Let $v = (v_1, v_2)$ and $u = (u_1, u_2)$. Then $u \in \mathbb{R}v$ if, and only if, $v_1 u_2 - v_2 u_1 = 0$.*

So instead of trying to find a scalar λ such that $u = \lambda v$ - which can be a cumbersome task - we can just do a one-line computation to determine whether or not two vectors are parallel. This leads to the following definition:

Definition 2.4.1.2. *Let $(a, b), (c, d) \in \mathbb{R}^2$. We define the **determinant of** $((a, b), (c, d))$ to be the number $\det((a, b), (c, d))$ given by*

$$\det((a, b), (c, d)) := ad - bc.$$

This gives us the following equivalent characterization of the determinant as a corollary of the preceding lemma:

Corollary 2.4.1.3. *Let $v, u \in \mathbb{R}^2$. Then $\det(v, u) = -\langle v, u^\perp \rangle$.*

Proof

If $v = (v_1, v_2)$ and $u = (u_1, u_2)$, then $u^\perp = (-u_2, u_1)$, so

$$\langle v, u^\perp \rangle = v_1(-u_2) + v_2 u_1 = -v_1 u_2 + v_2 u_1 = -(v_1 u_2 - v_2 u_1) = -\det(v, u).$$

□

Furthermore, this gives us a further characterization of parallel vectors:

Proposition 2.4.1.4. *Let $v, u \in \mathbb{R}^2$. Then the following are equivalent:*

a) $v \parallel u$;

b) $\mathbb{R}v = \mathbb{R}u$;

c) $v \perp u^{\perp}$;

d) $\det(v,u) = 0$.

All of these proofs have already been done at some point, so we won't bother redoing them. Before advancing, let us a prove a quick and useful result:

Proposition 2.4.1.5. *Let* $v, u \in \mathbb{R}^2$. *Then* $\det(v,u) = -\det(u,v)$.

Proof

Let $v = (v_1, v_2)$ and $u = (u_1, u_2)$. Then:

$$\det(v,u) = v_1u_2 - v_2u_1 = -(u_1v_2 - u_2v_1) = -\det(u,v).$$

□

Corollary 2.4.1.6. *For any two vectors* $v, u \in \mathbb{R}^2$ *we have that* $\det(v,u) = \langle u, v^{\perp}\rangle$.

Our next task is to find a geometric interpretation for the determinant of two vectors.

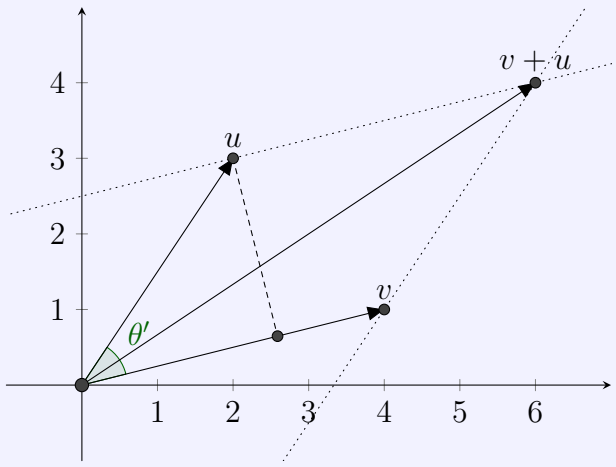
Example(s)

Consider the following case: $v = (4,1)$ and $u = (2,3)$. Then $\det(v,u) = 12 - 2 = 10$. But we also know that $\det(v,u) = -\langle v, u^{\perp}\rangle$ and that $\langle v, u^{\perp}\rangle = \|v\|\|u^{\perp}\|\cos\theta$ where $\theta \in [0, 2\pi)$ is the angle between v and $u^{\perp}$.
But $\|u^{\perp}\| = \|u\|$ and $\theta = \theta' + \frac{\pi}{2}$ where θ' is the angle between v and u.
Aside from that, we know that $\cos(t + \frac{\pi}{2}) = -\sin(t)$ for any $t \in \mathbb{R}$.
So $\langle v, u^{\perp}\rangle = \|v\|\|u^{\perp}\|\cos\theta$ can be rewritten as

$$\langle v, u^{\perp}\rangle = \|v\|\|u\|(-\sin\theta')$$

and since $\det(v,u) = -\langle v, u^{\perp}\rangle$ we can finally conclude that

$$\det(v,u) = \|v\|\|u\|\sin\theta'.$$



Looking at this picture, however, it's not hard to see that the dashed line has lenght precisely $\|u\| \sin \theta'$.
Therefore, $\|v\|\|u\| \sin \theta'$ is precisely the area of the parallelogram with vertices $0, v, v+u, u$.
But $\|v\|\|u\| \sin \theta' = \det(v, u)$. So **the determinant of v and u tells us the area of the parallelogram defined by v and u.**

However, what about $\det(u, v)$? Well, we know that $\det(v, u) = -\det(u, v)$ so if $\det(v, u)$ is positive, we get that $\det(u, v)$ is negative. But area, just like distance, is always a non-negative number.
What's going on here?
Well, it turns out that just like we can "extend" the notion of distances to negative numbers based on an "orientation" (that is, the sign means only the direction), we can also do the same for areas.
To make that more precise: Think about the real number line.
For instance, the numbers -5 and 5. They both have the "same value", in some sense, but in "opposite directions": 5 is "to the right" while -5 is "to the left".
We will do the same for areas: We will say that the area between two vectors v and u is **positive** if the angle between v and u is positive (just like in the picture).
In other words, the sign in the area will just serve as a notation to show us which vector "comes first in a counter-clockwise rotation" - if v comes before u, then $\det(v, u) \geq 0$; if u comes before v, then $\det(v, u) \leq 0$.
Notice also that $\det(v, u) = -\det(u, v)$ implies that $|\det(v, u)| = |\det(u, v)|$ so the preceding discussion does indeed make sense: The value of $\det(v, u)$ tells us the area of the parallelogram between v and u, and the sign tells us who comes first: v or u.

To finish off this example, try computing, using usual means, the area of the parallelogram determined by v and u. I'll even give you the value of θ' - it's $42.2°$.
See how hard that is? What even is the sine of $42.2°$? And then we have to multiply that by $\|v\| = \sqrt{4^2 + 1^1} = \sqrt{17}$ and then by $\|u\| = \sqrt{2^2 + 3^2} = \sqrt{13}$, all three of which are irrational numbers.
So in order to compute this area normally you'd have to use at least three approximations - one for $\sin(42.2°)$, one for $\sqrt{17}$ and one for $\sqrt{13}$. So it seems like the result would be very imprecise, and the more you decrease your uncertainty (by increasing decimal places in the approximation) the closer you get to the actual result.
But instead of doing a bunch of hard computations, if you just do $4 \cdot 3 - 1 \cdot 2$ you'll get the **exact** result, without any approximations or hard computations - just simple real addition and multiplication.

Now that we know that the determinant is just a different inner product, some results follow immediately:

Proposition 2.4.1.7. *Let $v, u, w \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$. Then the following hold:*

a) $\det(v + u, w) = \det(v, w) + \det(u, w)$*;*

b) $\det(v, u+w) = \det(v,u) + \det(v,w)$;

c) $\det(\lambda v, u) = \lambda \det(v,u) = \det(v, \lambda u)$;

d) $\det(v,u) = 0$ *if, and only if,* $v \parallel u$.

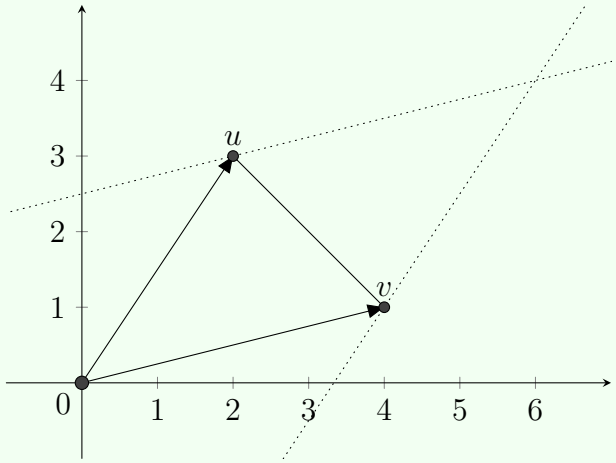
These all follow directly from the fact that determinants are inner products and have already been proven with that generality.

Now, to end this section, we want to use determinants to compute other kinds of areas.

Proposition 2.4.1.8. *Let* $v, u \in \mathbb{R}^2$. *Then the (unsigned) area of the triangle* $0vu$ *is just* $\dfrac{|\det(v,u)|}{2}$.

Proof

This follows trivially from the fact that $|\det(v,u)|$ is the (unsigned) area of the parallelogram determined by v, u, and that the diagonals of any parallelogram divide it into two congruent triangles trivially.



□

Corollary 2.4.1.9. *Let* $v, u, w \in \mathbb{R}^2$. *Then the (unsigned) area of the triangle with vertices* v, u, w *is just*

$$\frac{|\det(v-w, u-w)|}{2}.$$

Proof

This follows trivially from the proposition above, since translation preserves area, so the area of wvu is the same as the area of $0(v-w)(u-w)$. □

Corollary 2.4.1.10. *For any three points in the plane* $v, u, w \in \mathbb{R}^2$ *the (unsigned) area of the triangle determined by them is given by*

$$\frac{|\det(v,u) + \det(u,w) + \det(w,v)|}{2}.$$

Proof

From the preceding corollary we know that the unsigned area between v, u, w is

$$\frac{|\det(v-w, u-w)|}{2}.$$

But using the properties of the determinant we can expand that numerator:

$$\begin{aligned}\det(v-w, u-w) &= \det(v,u) + \det(-w,u) + \det(v,-w) + \det(-w,-w)\\ &= \det(v,u) - \det(w,u) - \det(v,w) + 0\\ &= \det(v,u) + \det(u,w) + \det(w,v)\end{aligned}$$

so the result follows. □

Finally, we're not gonna do the computations, but this allows us to compute the area of **any** polygon in the plane.

To start, given an n-gon $v_1v_2\cdots v_n$, notice that $v_1v_iv_{i+1}$ is always a triangle, where $2 \leq i \leq n-1$, and that those triangles completely cover up your n-gon's area.

Then we can compute the area of the n-gon as the sum of the areas of each of the triangles.

Example(s)

Let $v_1 = (1,1)$, $v_2 = (0,2)$, $v_3 = (1.5,3)$, $v_4 = (3,2)$ and $v_5 = (2,1)$ be a pentagon. Then, if $A(v,u,w)$ denotes the area of the triangle vuw, we can compute:

$$\begin{aligned}A(v_1,v_2,v_3) &= \frac{|\det(v_1-v_3, v_2-v_3)|}{2}\\ &= \frac{|\det((-0.5,-2),(-1.5,-1))|}{2}\\ &= \frac{|(-0.5)(-1)-(-2)(-1.5)|}{2}\\ &= \frac{|0.5-3|}{2}\\ &= \frac{|-2.5|}{2}\\ &= \frac{2.5}{2} = 1.25\end{aligned}$$

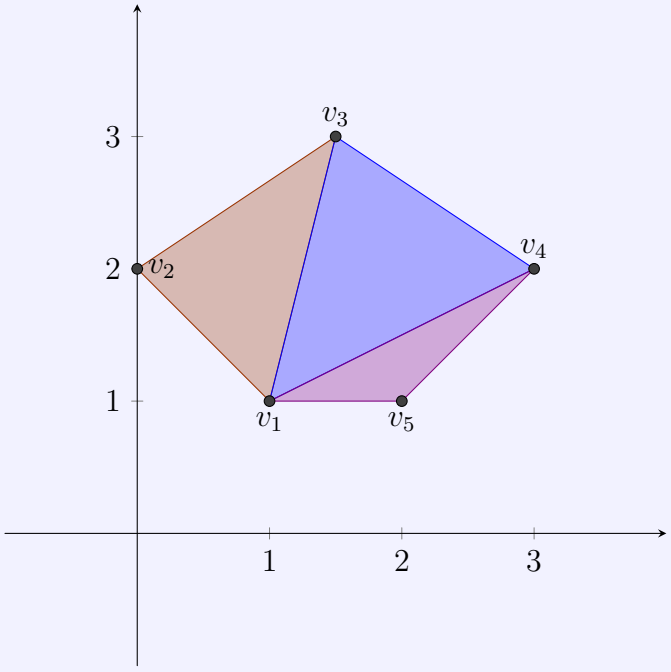
$$\begin{aligned} A(v_1, v_3, v_4) &= \frac{|\det(v_1 - v_4, v_3 - v_4)|}{2} \\ &= \frac{|\det((-2, -1), (-1.5, 1))|}{2} \\ &= \frac{|-2 - 1.5|}{2} \\ &= \frac{3.5}{2} = 1.75 \end{aligned}$$

$$\begin{aligned} A(v_1, v_4, v_5) &= \frac{|\det(v_1 - v_5, v_4 - v_5)|}{2} \\ &= \frac{|\det((-1, 0), (1, 1))|}{2} \\ &= \frac{|-1 - 0|}{2} \\ &= \frac{|-1|}{2} \\ &= \frac{1}{2} = 0.5 \end{aligned}$$

So the area $A(v_1, v_2, v_3, v_4, v_5)$ is just $A(v_1, v_2, v_3) + A(v_1, v_3, v_4) + A(v_1, v_4, v_5)$ which we have computed to be each respectively equal to $1.25, 1.75$ and 0.5, so

$$A(v_1, v_2, v_3, v_4, v_5) = 1.25 + 1.75 + 0.5 = 3.5$$

is the area of our pentagon.



One last lemma that could've saved us some time on the previous example and we're done:

Lemma 2.4.1.11. *Let $v_1v_2\cdots v_n$ be an n-gon in $\mathbb{R}^2$. Then its **signed** area is just*

$$\frac{\det(v_1,v_2)+\det(v_2,v_3)+\cdots+\det(v_{n-1},v_n)+\det(v_n,v_1)}{2}.$$

Proof

Follows directly from the reasoning above: We already know that

$$A(v_1,v_2,\cdots,v_n)=A(v_1,v_2,v_3)+A(v_1,v_3,v_4)+\cdots+A(v_1,v_{n-1},v_n)$$

and that

$$A(v,u,w)=\frac{\det(v,u)+\det(u,w)+\det(w,v)}{2}$$

for any $v,u,w\in\mathbb{R}^2$. So $A(v_1,v_i,v_{i+1})$ becomes

$$\frac{\det(v_1,v_i)+\det(v_i,v_{i+1})+\det(v_{i+1},v_1)}{2}$$

for all $2\leq i\leq n-1$.
But notice that $A(v_1,v_{i-1},v_i)$ is

$$\frac{\det(v_1,v_{i-1})+\det(v_{i-1},v_i)+\det(v_i,v_1)}{2}$$

so when we add up $A(v_1,v_{i-1},v_i)+A(v_1,v_i,v_{i+1})$ we get

$$\frac{\det(v_1,v_{i-1})+\det(v_{i-1},v_i)+\det(v_i,v_1)}{2}+\frac{\det(v_1,v_i)+\det(v_i,v_{i+1})+\det(v_{i+1},v_1)}{2}$$

and since $\det(v,u)=-\det(u,v)$ the terms $\det(v_1v_i)$ and $\det(v_i,v_1)$ in the middle cancel each other out, so this whole sum becomes just

$$\frac{\det(v_1,v_{i-1})+\det(v_{i-1},v_i)+\det(v_i,v_{i+1})+\det(v_{i+1},v_1)}{2}.$$

So we proceed by proving that this holds for all $n\in\mathbb{N}$:

- It clearly holds for $n=0$;
- Assume it holds for n. Let's show it holds for $n+1$:

  Take $v_1,v_2,\cdots,v_n,v_{n+1}$ an $(n+1)$-gon. We know that its area can be broken up into

  $$A(v_1,v_2,\cdots,v_n,v_{n+1})=A(v_1,v_2,v_3)+A(v_1,v_3,v_4)+\cdots+A(v_1,v_{n-1},v_n)+A(v_1,v_n,v_{n+1})$$

  but $A(v_1,v_2,v_3)+A(v_1,v_3,v_4)+\cdots+A(v_1,v_{n-1},v_n)$ is just $A(v_1,v_2,\cdots,v_n)$, and we're assuming the proposition holds for any n-gon. Therefore,

  $$A(v_1,v_2,\cdots,v_n)=\frac{\det(v_1,v_2)+\det(v_2,v_3)+\cdots+\det(v_{n-1},v_n)+\det(v_n,v_1)}{2}.$$

Finally, since $A(v_1, v_2, \cdots, v_n, v_{n+1}) = A(v_1, v_2, \cdots, v_n) + A(v_1, v_n, v_{n+1})$ and

$$A(v_1, v_n, v_{n+1}) = \frac{\det(v_1, v_n) + \det(v_n, v_{n+1}) + \det(v_{n+1}, v_1)}{2}$$

by the reasoning above, we see that when adding up $A(v_1, v_2, \cdots, v_n) + A(v_1, v_n, v_{n+1})$, the terms $\det(v_1, v_n)$ and $\det(v_n, v_1)$ cancel each other out, so the resulting sum is just

$$\frac{\det(v_1, v_2) + \det(v_2, v_3) + \cdots + \det(v_{n-1}, v_n) + \det(v_n, v_{n+1}) + \det(v_{n+1}, 1)}{2},$$

so the proposition holds for $n + 1$.

Finally, as usual, we have shown that the subset of natural numbers for which the proposition holds contains 0 and is closed under taking successors - which implies that this subset is actually the whole set of natural numbers.
This finishes the proof. □

Corollary 2.4.1.12. *For any n-gon $v_1, v_2, \cdots, v_n$ its **unsigned** area is just*

$$\frac{|\det(v_1, v_2) + \det(v_2, v_3) + \cdots + \det(v_{n-1}, v_n) + \det(v_n, v_1)|}{2}.$$

Example(s)

Revisiting the example above, then, we now know that the unsigned area of the pentagon is just

$$\frac{|\det(v_1, v_2) + \det(v_2, v_3) + \det(v_3, v_4) + \det(v_4, v_5) + \det(v_5, v_1)|}{2}$$

each one of which we can easily compute:

$$\begin{aligned}
\det(v_1, v_2) &= 2 \\
\det(v_2, v_3) &= -3 \\
\det(v_3, v_4) &= -6 \\
\det(v_4, v_5) &= -1 \\
\det(v_5, v_1) &= 1
\end{aligned}$$

so the unsigned area of the pentagon is just

$$\frac{|2 - 3 - 6 - 1 + 1|}{2} = \frac{|-7|}{2} = 3.5$$

which is precisely the same as we had gotten previously.

2.4.2 Determinants and linear functions

Now that we have the determinant on our side, let us use it to prove some results about linear functions.

First things first, we'll need to extend the definition of determinant for matrices instead of pairs of vectors:

Definition 2.4.2.1. *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *be a* 2×2 *real matrix. We define its* ***determinant*** *to be the real number* $\det A$ *given by*

$$\det A := ad - bc.$$

Lemma 2.4.2.2. *For any two matrices* $A, B \in M_{2\times 2}(\mathbb{R})$ *and* $\lambda \in \mathbb{R}$ *the following hold:*

a) $\det(AB) = \det A \cdot \det B$

b) $\det(\lambda A) = \lambda^2 \det A$

c) $\det A^t = \det A$.

Proof

Let, once and for all, $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$.

Then:

a)

$$\begin{aligned}
\det(AB) &= \det\left(\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}\right) \\
&= \det\begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} \\
&= (a_{11}b_{11} + a_{12}b_{21})(a_{21}b_{12} + a_{22}b_{22}) - (a_{11}b_{12} + a_{12}b_{22})(a_{21}b_{11} + a_{22}b_{21}) \\
&= a_{11}b_{11}a_{21}b_{12} + a_{11}b_{11}a_{22}b_{22} + a_{12}b_{21}a_{21}b_{12} + a_{12}b_{21}a_{22}b_{22} \\
&\quad -a_{11}b_{12}a_{21}b_{11} - a_{11}b_{12}a_{22}b_{21} - a_{12}b_{22}a_{21}b_{11} - a_{12}b_{22}a_{22}b_{21} \\
&= a_{11}a_{22}b_{11}b_{22} - a_{11}a_{22}b_{12}b_{21} - a_{12}a_{21}b_{11}b_{22} + a_{12}a_{21}b_{12}b_{21} \\
&= (a_{11}a_{22} - a_{12}a_{21})(b_{11}b_{22} - b_{12}b_{21}) = \det A \cdot \det B
\end{aligned}$$

b)

$$\begin{aligned}
\det(\lambda A) &= \det\begin{pmatrix} \lambda a_{11} & \lambda a_{12} \\ \lambda a_{21} & \lambda a_{22} \end{pmatrix} \\
&= (\lambda a_{11})(\lambda a_{22}) - (\lambda a_{12})(\lambda a_{21}) \\
&= \lambda^2(a_{11}a_{22}) - \lambda^2(a_{12}a_{21}) \\
&= \lambda^2(a_{11}a_{22} - a_{12}a_{21}) = \lambda^2 \det A
\end{aligned}$$

c)

$$\begin{aligned}\det A^t &= \det \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} \\ &= a_{11}a_{22} - a_{21}a_{12} \\ &= a_{11}a_{22} - a_{12}a_{21} = \det A\end{aligned}$$

And we see that the lemma follows. □

Proposition 2.4.2.3. *Let A be a 2×2 real matrix. Then $\det A = 0$ if, and only if, the columns of A are scalar multiples of each other, which happens if, and only if, the rows of A are scalar multiples of each other.*

Proof

By the preceding lemma, $\det A = \det A^t$ so the second equivalence of this statement follows trivially from the first.

Assume $\det A = 0$ for some real matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. This means that $ad - bc = 0$ or, in other words, $ad = bc$.

- If $a = c = 0$:

  Then the result is trivially true, since $(0, 0)$ is always a scalar multiple of any vector: Indeed, $(a, c) = (0, 0) = 0(b, c)$ so the first column is a scalar multiple of the second column and the result follows.

- If $a \neq 0$:

  Then the equality $ad = bc$ implies $d = \frac{b}{a}c$, and we know that $b = \frac{b}{a}a$.

  So

  $$(b, d) = \left(\frac{b}{a}a, \frac{b}{a}c\right) = \frac{b}{a}(a, c)$$

  and we see that (b, d) is a scalar multiple of (a, c).

- If $c \neq 0$:

  Then the equality $ad = bc$ implies $b = \frac{d}{c}a$, and we know that $d = \frac{d}{c}c$.

  So

  $$(b, d) = \left(\frac{d}{c}a, \frac{d}{c}c\right) = \frac{d}{c}(a, c)$$

  and we see that (b, d) is a scalar multiple of (a, c).

This shows that no matter the value of (a, c), the column (b, d) will always be a scalar multiple of it.

Conversely, if $(b, d) = \lambda(a, c)$ for some $\lambda \in \mathbb{R}$, then

$$\det A = ad - bc = a(\lambda c) - (\lambda a)c = \lambda(ac) - \lambda(ac) = 0.$$

This shows that if two columns on a matrix are scalar multiples of each other, than the determinant of that matrix is zero.

This finishes the proof. □

Definition 2.4.2.4. *Let $\mathcal{B} = \{v_1, v_2\} \subseteq \mathbb{R}^2$ be a base of $\mathbb{R}^2$ and let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be any linear function. We define* ***the matrix of f in the base $\mathcal{B}$*** *to be the matrix $A_f^{\mathcal{B}}$ constructed as follows:*

By definition of base, there are uniquely determined $a, b, c, d \in \mathbb{R}$ such that $f(v_1) = av_1 + cv_2$ and $f(v_2) = bv_1 + dv_2$. So we put

$$A_f^{\mathcal{B}} := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where a, c are the coefficients of $f(v_1)$ in the base $\mathcal{B}$, and b, d are the coefficients of $f(v_2)$ in the base $\mathcal{B}$.

Lemma 2.4.2.5. *For every base $\mathcal{B} \subseteq \mathbb{R}^2$, the matrix of $\mathrm{id}_{\mathbb{R}^2}$ in that base is the identity matrix $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.*

Proof

Let $\mathcal{B} = \{b_1, b_2\}$ be a base of $\mathbb{R}^2$. Then, by definition, the matrix of any linear function f in this base is the matrix whose first column is the coefficients of $f(b_1)$ in $\mathcal{B}$ and whose second column is the coefficients of $f(b_2)$ in $\mathcal{B}$.
But if we take $f = \mathrm{id}_{\mathbb{R}^2}$, then $f(b_1) = b_1$ and $f(b_2) = b_2$. Not only that, but the unique way to write $f(b_1)$ and $f(b_2)$ in $\mathcal{B}$ is now trivial: $f(b_1) = b_1 = 1 \cdot b_1 + 0 \cdot b_2$ and $f(b_2) = b_2 = 0 \cdot b_1 + 1 \cdot b_2$. Therefore,

$$A_{\mathrm{id}_{\mathbb{R}^2}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

is the identity matrix. □

Theorem 2.4.2.6. *Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Then f is a linear isomorphism if, and only if, for every base $\mathcal{B} \subseteq \mathbb{R}^2$ we have that $\det A_f^{\mathcal{B}} \neq 0$, where $A_f^{\mathcal{B}}$ is the matrix of f in the base $\mathcal{B}$.*

Proof

Assume that f is a linear isomorphism. This means that there's a unique linear function $f^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ such that $f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}_{\mathbb{R}^2}$.
Take now any base $\mathcal{B} \subseteq \mathbb{R}^2$, and let $A, A' \in M_{2\times 2}(\mathbb{R})$ be the matrices of f and f^{-1}, respectively, in the base $\mathcal{B}$.
By the preceding lemma, we know that the matrix of $\mathrm{id}_{\mathbb{R}^2}$ in any base (in particular, in the base $\mathcal{B}$) is I. So we get the equation

$$AA' = I,$$

by definition of matrix multiplication.
But now, since determinants preserve matrix multiplication, we see that

$$\det(AA') = \det A \cdot \det A'.$$

But on the other hand, the above equality tells us that

$$\det(AA') = \det I,$$

and we know that $\det I = 1$.
This tells us that $\det A \cdot \det A' = 1$. Therefore, neither $\det A$ nor $\det A'$ can be 0 - otherwise, we'd have $0 = 1$.
So by assuming that f has an inverse, we were able to show that $\det A_f^{\mathcal{B}} \neq 0$ for any base $\mathcal{B}$.

Conversely, assume that we are given a linear function f such that for any base $\mathcal{B} \subseteq \mathbb{R}^2$ we have that $\det A_f^{\mathcal{B}} \neq 0$.
In particular, for the canonical base $E = \{e_1, e_2\}$, we see that $\det A_f^E \neq 0$. But this means that the columns of A_f^E are non-parallel, and we know, by definition, that the columns of A_f^E are just $f(e_1)$ and $f(e_2)$, respectively. Therefore, $f(E) = \{f(e_1), f(e_2)\}$ is also a base of $\mathbb{R}^2$.
Since E is a base, there are some $a, b, c, d \in \mathbb{R}$ such that

$$f(e_1) = ae_1 + ce_2 \tag{1}$$
$$f(e_2) = be_1 + de_2. \tag{2}$$

Similarly, since $f(E)$ is a base, there are some $a', b', c', d' \in \mathbb{R}$ such that

$$e_1 = a'f(e_1) + c'f(e_2) \tag{3}$$
$$e_2 = b'f(e_1) + d'f(e_2). \tag{4}$$

Now we can work with these two pairs of equations: For instance, using (3) and (4) we can substitute them in (1) and (2), respectively, to obtain

$$f(e_1) = a(a'f(e_1) + c'f(e_2)) + c(b'f(e_1) + d'f(e_2)) = (aa' + cb')f(e_1) + (ac' + cd')f(e_2)$$

$$f(e_2) = b(a'f(e_1) + c'f(e_2)) + d(b'f(e_1) + d'f(e_2)) = (ba' + db')f(e_1) + (bc' + dd')f(e_2)$$

but $f(e_1) = 1 \cdot f(e_1) + 0 \cdot f(e_2)$ and $f(e_2) = 0 \cdot f(e_1) + 1 \cdot f(e_2)$ tell us that

$$aa' + cb' = 1 \tag{5}$$
$$ac' + cd' = 0 \tag{6}$$
$$ba' + db' = 0 \tag{7}$$
$$bc' + dd' = 1. \tag{8}$$

On the other hand, using (1) and (2) and substituting them on (3) and (4), respectively, gives us that

$$e_1 = a'(ae_1 + ce_2) + c'(be_1 + de_2) = (a'a + c'b)e_1 + (a'c + c'd)e_2$$

$$e_2 = b'(ae_1 + ce_2) + d'(be_1 + de_2) = (b'a + d'b)e_1 + (b'c + d'd)e_2$$

and since $e_1 = 1 \cdot e_1 + 0 \cdot e_2$ and $e_2 = 0 \cdot e_1 + 1 \cdot e_2$, we can see that

$$a'a + c'b = 1 \tag{9}$$
$$a'c + c'd = 0 \tag{10}$$
$$b'a + d'b = 0 \tag{11}$$
$$b'c + d'd = 1. \tag{12}$$

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $A' = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$.
Now we can compute AA':

$$AA' = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{pmatrix}$$

and see that equations (9)-(12) tell us that

$$AA' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I.$$

Analogously, if we compute $A'A$ we'll see that

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

and we can use equations (5)-(8) to see that

$$A'A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I.$$

Finally, we notice that $A = A_f^E$. So A' is the matrix of some linear function g in the base E such that $f \circ g = g \circ f = \mathrm{id}_{\mathbb{R}^2}$.
This shows that f is a linear isomorphism, and ends the proof. □

Definition 2.4.2.7. *Let $A \in M_{2\times 2}(\mathbb{R})$. We say that A is **invertible** if there is some matrix $B \in M_{2\times 2}(\mathbb{R})$ such that $AB = I = BA$.*

Corollary 2.4.2.8. *The following are equivalent for any matrix $A \in M_{2\times 2}(\mathbb{R})$:*

a) A is invertible;

b) $A = A_f^{\mathcal{B}}$ for some base $\mathcal{B} \subseteq \mathbb{R}^2$ and some linear isomorphism f;

c) The columns of A are non-parallel;

d) The rows of A are non-parallel;

e) A^t is invertible;

f) $\det A \neq 0$.

Corollary 2.4.2.9. *If a matrix is invertible, then its inverse is uniquely determined.*

2.4.3 Our First Named Theorem

This section ends the chapter with a big and very important result: The **Isomorphism Theorems**.

In order to even state those, however, we need to remember some things that might have been forgotten already:

- A **relation** on a set X is any subset of $\mathcal{P}(X)$;
- An **equivalence relation** is a relation that is reflexive ($x \sim x$), symmetric ($x \sim y$ implies $y \sim x$) and transitive ($x \sim y$ and $y \sim z$ implies $x \sim z$);
- The **quotient over an equivalence relation** is the set of equivalence classes that is, if $\sim$ is an equivalence relation on X, then $[x] \in X/\sim$ is the collection of all $y \in X$ such that $x \sim y$;
- The **kernel** of a linear function is the set of all vectors that go to zero under that function;
- The **A subspace** of $\mathbb{R}^2$ is a subset $X \subseteq \mathbb{R}^2$ such that for all $x, y \in X$ and $\lambda \in \mathbb{R}$, $x + \lambda y \in X$;
- Given two subspaces $X, Y \leq \mathbb{R}^2$, their **sum** is the subspace $X + Y$ whose elements are just sums of elements on X and Y;
- Given two subspaces $X, Y \leq \mathbb{R}^2$, their **intersection** $X \cap Y$ is also a subspace.

With this we can make an example:

Example(s)

Let $v \in \mathbb{R}^2$ and consider the line $r = \mathbb{R}v$. We already know that lines through zero in $\mathbb{R}^2$ are subspaces, so r is a subspace.
In lemma 2.3.4.13 we showed that the set of all points at a fixed distance from any line is two lines.
Not only that, but that for each choice of non-negative real number $D \in \mathbb{R}$, these two lines are completely determined as being $s_D = r + \overrightarrow{D}$ and $t_D = r - \overrightarrow{D}$, where $\overrightarrow{D}$ is defined as

$$\overrightarrow{D} := \frac{D}{\|v\|} v^{\perp},$$

that is, it's a vector which is perpendicular to v and has size D.
From this, it's easy to see that for any non-negative $D \in \mathbb{R}$ we have that

$$r = s_D - s_D.$$

This means that no matter which elements of s_D you pick, their difference will always lie in r.
Indeed: Let $x = \lambda v + \overrightarrow{D} \in s_D$ and $y = \mu v + \overrightarrow{D} \in s_D$. Then

$$x - y = (\lambda v + \overrightarrow{D}) - (\mu v + \overrightarrow{D}) = \lambda v - \mu v = (\lambda - \mu)v \in r.$$

We can then extend this construction to include negative D: We just put $s_D = t_{-D}$ (so, for instance, s_{-2} is defined to be t_2 and t_{-2} is defined to be s_2), so it doesn't matter the sign of D, s_D and t_D are always well-defined, and they always add up to r.

This allows us to define the following relation: $x \sim y$ if, and only if, $x - y \in r$.

- $\sim$ is reflexive:

  Clearly, for all $x \in \mathbb{R}^2$ we have that $x - x = 0 \in r$ so $x \sim x$;

- $\sim$ is symmetric:

  If $x \sim y$ it means that $x - y = \lambda v$ for some $\lambda \in \mathbb{R}$. But then

  $$y - x = -(x - y) = -(\lambda v) = \lambda v$$

  so $y - x \in r$ and hence $y \sim x$;

- $\sim$ is transitive:

  Let $x \sim y$ and $y \sim z$. This means that $x - y = \lambda v$ and $y - z = \mu v$ for some $\lambda, \mu \in \mathbb{R}$. Then:

  $$x - z = x - y + y - z = (\lambda v) + (\mu v) = (\lambda + \mu)v$$

  so $x - z \in r$ so $x \sim z$.

Let $\mathcal{S} := \{s_D \mid D \in \mathbb{R}\}$. We claim that $\mathcal{S} = \mathbb{R}^2/\sim$.

We've already shown that $s_D - s_D \in r$, so clearly $\mathcal{S} \subseteq \mathbb{R}^2/\sim$.
Conversely, take $[x] \in \mathbb{R}^2/\sim$ to be the class of all $y \in \mathbb{R}^2$ such that $x - y \in r$ and let $D = d(x, r)$.
Then, clearly, either $x \in s_D$ or $x \in s_{-D}$. It really doesn't matter which, and makes no real difference, so let's assume $x \in s_D$. We will show that every $y \in [x]$ is also in s_D, and hence $\mathbb{R}^2/\sim \subseteq \mathcal{S}$.
Since $x \in s_D$, we can write it as $x = \lambda v + \overrightarrow{D}$ for some $\lambda \in \mathbb{R}$. But now, taking $y \in [x]$ means that $x - y \in r$, so $x - y = \lambda' v$ for some $\lambda' \in \mathbb{R}$, by definition. But then:

$$\begin{aligned} x - y &= \lambda' v \\ \lambda v + \overrightarrow{D} - y &= \lambda' v \\ \lambda v + \overrightarrow{D} - \lambda' v &= y \\ y = (\lambda - \lambda')v + \overrightarrow{D} &\in s_D \end{aligned}$$

which tells us that every $y \in [x]$ is also in s_D.
This shows that $\mathcal{S} = \mathbb{R}^2/\sim$.
Notice that the only thing that we had any control over to define this quotient was the line r. If we had chosen any other line through zero we'd have gotten a different quotient.
This gives us a nice intuition about our next result:

Proposition 2.4.3.1. *Any subspace $X \leq \mathbb{R}^2$ determines a unique equivalence relation by $x \sim y$ if, and only if, $x - y \in X$.*

Proof

- $\sim$ is reflexive:

  Since X is a subspace it must contain zero, so $x - x = 0 \in X$ for all $x \in \mathbb{R}^2$ implies $x \sim x$;

- $\sim$ is symmetric:

  Since X is a subspace, it's closed under scalar multiplication. In particular, for any $v \in X$, $-v$ is also in X.

  Therefore, $x \sim y$ implies $x-y \in X$, so $-(x-y)$ must also be in X. But $-(x-y) = y-x$ which shows us that $y \sim x$;

- $\sim$ is transitive:

  Since X is a subspace, it's closed under addition. So if $x \sim y$ and $y \sim z$ this means that $x - y$ and $y - z$ are in X. But then their sum is also in X. This means that $(x - y) + (y - z) = x - z$ is in X, and so $x \sim z$.

Since $\sim$ is reflexive, symmetric and transitive, it is, by definition, an equivalence. This ends the proof. □

Definition 2.4.3.2. *Let $X \leq R^2$ be a subspace and $\sim$ the equivalence relation uniquely determined by X. We'll denote the set $\mathbb{R}^2/\sim$ simply by $\mathbb{R}^2/X$ and call it the* ***quotient space of*** $\mathbb{R}^2$ ***over*** X.

Example(s)

Following up on the previous example, we have shown that $\mathbb{R}^2/\mathbb{R}v = \mathcal{S}$, that is, the quotient space of $\mathbb{R}^2$ over a line through zero is the set of all lines which are parallel to it - each uniquely determined by its signed distance to it.

Now we're ready to state the theorem we're trying to prove.

Theorem 2.4.3.3 (Isomorphism Theorems)**.** *Let $X, Y \leq \mathbb{R}^2$ be subspaces and $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Then the following hold:*

(1.) $\mathbb{R}^2/\operatorname{Ker} f \cong \operatorname{Im} f$;

(2.) $(X + Y)/X \cong X/(X \cap Y)$;

(3.) $Y \subseteq X$ *if, and only if,* $X/Y \subseteq \mathbb{R}^2/Y$.

And we've already got everything we need to prove the first one:

Theorem 2.4.3.4 (First Isomorphism Theorem)**.** *For any linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$ we have that $\mathbb{R}^2/\operatorname{Ker} f \cong \operatorname{Im} f$.*

Proof

Let $\phi : \mathbb{R}^2/\operatorname{Ker} f \to \operatorname{Im} f$ be defined as $\phi([x]) := f(x)$.
First, we need to show that this is indeed a function - that is, that if we take the same point in the domain, its image under ϕ doesn't change.
In our case, this means that we need to show that $[x] = [y]$ implies $\phi([x]) = \phi([y])$.
Indeed: If $[x] = [y]$ this means that $x - y \in \operatorname{Ker} f$, by definition. Therefore, $f(x - y) = 0$, by definition of kernel. Since f is linear, this implies $f(x) = f(y)$, and so

$$\phi([x]) = f(x) = f(y) = \phi([y])$$

and we see that ϕ is indeed a function.

Next, consider $\psi : \operatorname{Im} f \to \mathbb{R}^2/\operatorname{Ker} f$ defined by $\psi(x) := [f^{-1}(x)]$. Once again, we need to show that this makes sense to define: $f^{-1}(x)$ is a **set**, not a vector, so it could happen that $[f^{-1}(x)]$ is not an element of $\mathbb{R}^2/\operatorname{Ker} f$, but, instead, a subset.
To show that this doesn't happen, then, we need to show that if there are more than one point - say, $v, u \in f^{-1}(x)$ - then $v - u \in \operatorname{Ker} f$. This suffices, because then $[f^{-1}(x)] = [v] = [u]$ which is indeed a single element.
Take then $v, u \in f^{-1}(x)$. This means that $f(v) = f(u) = x$. But then, $f(v - u) = f(v) - f(u) = x - x = 0$ so $v - u \in \operatorname{Ker} f$. This shows that $[f^{-1}(x)] = [v]$ and so ψ is indeed a function.

Finally, we claim that ϕ and ψ are mutually inverse.
Indeed given any $[x] \in \mathbb{R}^2/\operatorname{Ker} f$ and any $y \in \operatorname{Im} f$ we have:

$$\begin{aligned}(\psi \circ \phi)([x]) &= \psi(\phi([x])) \\ &= \psi(f(x)) \\ &= [f^{-1}(f(x))] = [x]\end{aligned}$$

and

$$\begin{aligned}(\phi \circ \psi)(y) &= \phi(\psi(y)) \\ &= \phi([f^{-1}(y)]) \\ &= f(f^{-1}(y)) = y\end{aligned}$$

so $\psi \circ \phi = \mathrm{id}_{\mathbb{R}^2/\operatorname{Ker} f}$ and $\phi \circ \psi = \mathrm{id}_{\operatorname{Im} f}$.
This shows that $\psi = \phi^{-1}$ and hence ϕ is an isomorphism, as we wanted to show. □

This is, in my opinion, the single most important result in all of linear algebra. Let us give some examples:

Example(s)

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be the function given by $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ that is, $f(x, y) = (x + y, x + y)$. We can compute the kernel of f to be $\operatorname{Ker} f = \mathbb{R}(1, -1)$.
So, by the First Isomorphism Theorem, we know that $\mathbb{R}^2/\operatorname{Ker} f \cong \operatorname{Im} f$. Let us compute, then, $\mathbb{R}^2/\operatorname{Ker} f$: $[x] = [y]$ in $\mathbb{R}^2/\operatorname{Ker} f$ if, and only if, $x - y \in \operatorname{Ker} f$.
This means that writing $x = (x_1, x_2)$ and $y = (y_1, y_2)$ we see that $[x] = [y]$ if, and only if $(x_1 + x_2, x_1 + x_2) = (y_1 + y_2, y_1 + y_2)$. But this happens if, and only if, $x_1 + x_2 = y_1 + y_2$.
It follows that each class $[x]$ is uniquely determined by the sum $x_1 + x_2$ - since, as we've shown, two elements are in the same class if, and only if, they have the same sum.
Consider then all classes of the form $[(\lambda, 0)]$ where λ ranges over all real numbers. We can easily see that two such classes $[(\lambda, 0)]$ and $[(\mu, 0)]$ are equal if, and only if, $\lambda = \mu$.
Finally, notice that for any other class $[(x_1, x_2)]$, we have that

$$[(x_1, x_2)] = [(x_1 + x_2, 0)]$$

since they both have the same sum. It follows then that the classes $[(\lambda, 0)]$ are, in fact, **all** the possible classes.
But we can clearly make a bijection between $\mathbb{R}$ and the set $\mathcal{S}$ of all classes of the form $[(\lambda, 0)]$ simply by mapping each real number a to the class $[(a, 0)]$.
Therefore, we have just shown that $\mathbb{R}^2/\operatorname{Ker} f \cong \mathbb{R}$.
But the First Isomorphism Theorem tells us that $\mathbb{R}^2/\operatorname{Ker} f \cong \operatorname{Im} f$, and we know that $\operatorname{Im} f$ is a subspace.
Combining these two isomorphisms, we get that $\operatorname{Im} f$ is a subspace that is isomorphic to $\mathbb{R}$ - so it must be a **line through zero**.
This is indeed just an instance of a more general case:

Lemma 2.4.3.5. *Let $X \leq \mathbb{R}^2$ be any subspace. Then $\mathbb{R}^2/X$ is isomorphic to either 0, a line through zero or $\mathbb{R}^2$.*

Proof

We know that we have only three kinds of subspaces: 0, lines through zero and $\mathbb{R}^2$.
Let's consider each of those cases:

- $\underline{X = 0}$:

  In this case, two classes $[x], [y] \in \mathbb{R}^2/0$ are equal if, and only if, $x - y \in 0$. But $0 = \{(0, 0)\}$, so $x - y \in 0$ implies $x = y$.

  What this tells us is that each class $[x]$ in $\mathbb{R}^2/0$ has a single element - x. Therefore, there is a clear bijection between $\mathbb{R}^2/0$ and $\mathbb{R}^2$ taking each class $[x]$ to its unique element x.

  This shows $\mathbb{R}^2/0 \cong R^2$.

- X is a line through zero:

  In this case, let $X = \mathbb{R}v$ for some $v \in \mathbb{R}^2$. Then two classes $[x], [y] \in \mathbb{R}^2/\mathbb{R}v$ are equal if, and only if, $x - y \in \mathbb{R}v$.

  We already know that if we denote $D = d(x, \mathbb{R}v)$ then x is in either s_D or s_{-D}. It doesn't matter which one it is, though, because whatever it might be, y will also be there.

  For instance, if $x \in s_D$ it means that $x = \lambda_x v + \overrightarrow{D}$, and so $x - y = \lambda v$ implies $y = (\lambda_x - \lambda)v + \overrightarrow{D}$ which is also in s_D.

  Similarly, if $x \in s_{-D}$ it means that $x = \lambda_x - \overrightarrow{D}$, and so $x - y = \lambda v$ implies $y = (\lambda_x - \lambda)v - \overrightarrow{D}$.

  Either way, we see that $[x] = [y]$ implies that x and y lie at the same signed distance from $\mathbb{R}v$.

  In other words, each class $[x]$ is uniquely determined by its signed distance from $\mathbb{R}v$ - different classes have different distances, and different distances give us different classes.

  Since the set of all signed distances is just $\mathbb{R}$ we have a clear bijection between $\mathbb{R}^2/\mathbb{R}v$ and $\mathbb{R}$ which takes each class to its signed distance.

  This shows that $\mathbb{R}^2/\mathbb{R}v \cong \mathbb{R}$, and, finally, since $\mathbb{R}$ is isomorphic to any line (in particular, to lines through zero) it follows that $\mathbb{R}^2/\mathbb{R}v$ is isomorphic to a line through zero.

- $X = \mathbb{R}^2$:

  In this case, two classes $[x], [y] \in \mathbb{R}^2/\mathbb{R}^2$ are equal if, and only if, $x - y \in \mathbb{R}^2$. But this is true for any pair of vectors $x, y \in \mathbb{R}^2$ - difference of vectors is always a vector.

  This shows that for any pair of vectors $v, u \in \mathbb{R}^2$, we always have that $[v] = [u]$ - therefore, there's only one class in $\mathbb{R}^2/\mathbb{R}^2$ which contains every single vector.

  It follows that $\mathbb{R}^2/\mathbb{R}^2$ is a singleton and, hence, isomorphic to any other singleton - in particular, to the singleton $0 = \{(0, 0)\}$.

Since we have covered all possible cases of subspaces, this finishes the proof. □

Remark 2.4.3.6

Notice that, at this point, we cannot prove the converse to this result - that is, we cannot prove that if $X \leq \mathbb{R}^2$ is a subspace such that $\mathbb{R}^2/X$ is isomorphic to a line, then X is a line. This is because of the following result:

Theorem 2.4.3.7. *Let $n \in \mathbb{N}$ be any natural number. Then $\mathbb{R}^n \cong \mathbb{R}$.*

Proof

We will prove that $\mathbb{R}^2 \cong \mathbb{R}$. The result then follows by noticing that $\mathbb{R}^n \cong \mathbb{R}^{n-1} \times \mathbb{R}$.
First things first, let $I \subset \mathbb{R}$ be the following set

$$I := \{x \in \mathbb{R} \mid 0 < x < 1\},$$

that is, the set of all real numbers which are between 0 and 1, excluding 0 and 1. This is called the **open unit interval** or just the **open interval**.
We will prove that I is isomorphic the set of all non-negative real numbers. To do this, consider the map

$$f : I \to \mathbb{R}$$

defined by

$$f(x) = \frac{x}{1-x}.$$

Since $x \in I$ implies $0 < x < 1$, we have that $1 - x$ is always positive and x is always positive, so the ratio $\frac{x}{1-x}$ is also always positive. This shows that $\operatorname{Im} f \subseteq \mathbb{R}_+$ - where $\mathbb{R}_+$ is the full subset of all positive real numbers.
Consider now the function

$$g : \mathbb{R}_+ \to \mathbb{R}$$
$$g(y) = \frac{y}{y+1}.$$

Since $y > 0$ implies $y + 1 > 0$, we have that $\frac{y}{y+1} > 0$, and since $y + 1 > y$, we have that $\frac{y}{y+1} < 1$, so $0 < g(y) < 1$ for all $y \in \mathbb{R}_+$ and $\operatorname{Im} g \subseteq I$.
Now:

$$g(f(x)) = g\left(\frac{x}{1-x}\right)$$

$$= \frac{\dfrac{x}{1-x}}{\dfrac{x}{1-x} + 1}$$

$$= \frac{\dfrac{x}{1-x}}{\dfrac{x}{1-x} + \dfrac{1-x}{1-x}}$$

$$= \frac{\dfrac{x}{1-x}}{\dfrac{1}{1-x}}$$

$$= \frac{x}{1} \cdot \frac{1-x}{1-x} = x$$

and

$$f(g(y)) = f\left(\frac{y}{y+1}\right)$$

$$= \frac{\frac{y}{y+1}}{1 - \frac{y}{y+1}}$$

$$= \frac{\frac{y}{y+1}}{\frac{y+1}{y+1} - \frac{y}{y+1}}$$

$$= \frac{\frac{y}{y+1}}{\frac{1}{y+1}}$$

$$= \frac{y}{1} \cdot \frac{y+1}{y+1} = y$$

so $f \circ g = \mathrm{id}_{\mathbb{R}_+}$ and $g \circ f = \mathrm{id}_I$, so they are inverse and, therefore, an isomorphism.

The next step is to show that I is isomorphic not only to $\mathbb{R}_+$, but to $\mathbb{R}$.
To see this, first notice that $-I = \{x \in \mathbb{R} \mid -1 < x < 0\}$ is isomorphic, exactly as above, to the set of negative real numbers - $\mathbb{R}_-$.
This shows that the open interval $I' := \{x \in \mathbb{R} \mid -1 < x < 1\}$ is isomorphic to $\mathbb{R}$ - just send everyone smaller than 0 to the negatives, everyone greater than 0 to the positives and zero onto itself.
Now we claim that $I \cong I'$. This is easily seen: Take $f' : I \to \mathbb{R}$ given by

$$f'(x) = 2x - 1.$$

It's easy to see that since $0 < x < 1$, we have that $0 < 2x < 2$ and so $-1 < 2x - 1 < 1$, so $-1 < f'(x) < 1$, which tells us that $\operatorname{Im} f' \subseteq I'$.
Similarly as above, consider $g' : I' \to \mathbb{R}$ given by

$$g'(y) = \frac{y+1}{2}.$$

Once again, we can readily see that $-1 < y < 1$ implies $0 < y + 1 < 2$ and so $0 < \frac{y+1}{2} < 1$, so $0 < g'(y) < 1$ which tells us that $\operatorname{Im} g' \subseteq I$.

Once again, we can compute

$$\begin{aligned} g'(f'(x)) &= g'(2x-1) \\ &= \frac{(2x-1)+1}{2} \\ &= \frac{2x}{2} = x \end{aligned}$$

and

$$\begin{aligned} f'(g'(y)) &= f'\left(\frac{y+1}{2}\right) \\ &= 2 \cdot \frac{y+1}{2} - 1 \\ &= (y+1) - 1 = y \end{aligned}$$

so $f \circ g = \mathrm{id}_{I'}$ and $g \circ f = \mathrm{id}_I$, so they are inverse and, therefore, an isomorphism.
As a corollary, we have that $\mathbb{R} \cong I' \cong I$ implies $I \cong \mathbb{R}$.

We're gonna spare you the details, since they're basically the same, but essentially the same proof shows us that $I^2 \cong \mathbb{R}^2$.
We will finally prove this theorem, then, by showing that $I \cong I^2$.
To do this, take $x \in I$ and write its unique decimal expansion:

$$x = 0.x_0x_1x_2x_3...x_n...$$

with all $a_i \in \mathbb{N}$ such that $0 \leq a_i \leq 9$ for all $i \in \mathbb{N}$.
This tells us that we can write all $x \in I$ as a sequence $(a_0, a_1, a_2, \cdots, a_n, \cdots)$ with all terms being natural numbers between 0 and 9.
As an example, we would write $\frac{1}{3}$ as 0.333..., so, in this case, $a_i = 3$ for all $i \in \mathbb{N}$. So we can represent $\frac{1}{3}$ as the sequence $(3, 3, 3, \cdots)$.
Notice also that each such sequence determines a unique number in I - given the sequence $(a_0, a_1, a_2, \cdots)$, the number $x = 0.a_0a_1a_2...$ is uniquely determined and is in I.
Since we can think of I as the set of all sequences with natural entries between 0 and 9, we can think if I^2 as the set of all pairs of such sequences.
For instance, the element $((3, 3, 3, \cdots), (1, 2, 3, 4, \cdots)) \in I^2$ and corresponds to the point $(0.333..., 0.1234...) \in I^2$.
Now that that's out of the way, we can define the function $\phi : I \to I^2$ given by

$$\phi(a_0, a_1, a_2, \cdots) = ((a_0, a_2, a_4, \cdots), (a_1, a_3, a_5, \cdots))$$

which breaks each sequence $x \in I$ into a pair of sequences in I^2 simply by alternating between even- and odd-indexed entries in the sequence.
Conversely, we define $\psi : I^2 \to I$ by putting

$$\psi((a_0, a_1, a_2, \cdots), (b_0, b_1, b_2, \cdots)) = (a_0, b_0, a_1, b_1, \cdots)$$

which creates a new sequence in I from a pair of sequences in I^2 by alternating between them.
Finally, let's show that $\phi = \psi^{-1}$:

$$\psi(\phi(a_0, a_1, a_2, \cdots)) = \psi((a_0, a_2, a_4, \cdots), (a_1, a_3, a_5, \cdots)) = (a_0, a_1, a_2, \cdots)$$

$$\phi(\psi((a_0, a_1, a_2, \cdots), (b_0, b_1, b_2, \cdots))) = \phi(a_0, b_0, a_1, b_1, \cdots) = ((a_0, a_1, a_2, \cdots), (b_0, b_1, b_2, \cdots))$$

so $\psi \circ \phi = \mathrm{id}_I$ and $\phi \circ \psi = \mathrm{id}_{I^2}$, which tells us that they are mutually inverse, and therefore an isomorphism.

Combining all of this, we get that

$$\mathbb{R} \cong I \cong I^2 \cong \mathbb{R}^2$$

and so $\mathbb{R} \cong \mathbb{R}^2$.
This ends the proof. □

Corollary 2.4.3.8. *Any line in $\mathbb{R}^2$ is in bijection with $\mathbb{R}^2$.*

Remark 2.4.3.9

This is a problem, because we would like lines and planes to be distinguishable by isomorphisms, but this doesn't happen. This is another reason why we will, at some point, only focus our attention on **linear** isomorphisms - because, for them, lines can only be isomorphic to lines, and planes can only be isomorphic to planes.

Returning, then, to our goal, we need to define a new object before we proceed any further.

Definition 2.4.3.10. *Let $X, Y \leq \mathbb{R}^2$ be two subspaces such that $Y \subseteq X$. We define the **quotient set of** X **over** Y to be the set X/Y given by*

$$X/Y := \{[x] \in \mathbb{R}^2/Y \mid x \in X\}.$$

With this, we can now prove the two remaining theorems:

Theorem 2.4.3.11 (Second Isomorphism Theorem)**.** *Let $X, Y \leq \mathbb{R}^2$ be any two subspaces of $\mathbb{R}^2$. Then:*

$$\frac{X+Y}{Y} \cong \frac{X}{X \cap Y}.$$

Proof

First, notice that since $Y \subseteq X + Y$, $X \cap Y \subseteq X$ and that both $X + Y$ and $X \cap Y$ are subspaces, it makes sense to define the two quotients in the statement of this theorem.

Before trying to prove this result, then, let us explore these sets a bit more.
For instance, take two elements $[v] = [u] \in \dfrac{X+Y}{Y}$. By definition, it is the equivalence class of the vector v in the quotient set $\mathbb{R}^2/Y$, where $v \in X + Y$.
So $[u] = [v]$ means that $v - u \in Y$. But since $v \in X + Y$, we can write $v = x + y$ for some $x \in X$ and $y \in Y$. Similarly for u, we can write it as $u = x' + y'$, with $x' \in X$ and $y' \in Y$.
But now $v - u \in Y$ means that $(x + y) - (x' - y' \in Y$, which means that $x - x' \in Y$ (since $y - y'$ already lies in Y).
This tells us that two vectors $x + y, x' + y' \in X + Y$ have the same class in $\dfrac{X+Y}{Y}$ if, and only if, $x - x' \in Y$. But notice that, by definition, $x - x' \in X$. So we can further affirm that those two vectors are equal in the quotient if, and only if, $x - x' \in X \cap Y$.
But this tells us that $[x+0] = [x+y]$ for all $x \in X$, $y \in Y$ since, clearly, $x - x = 0 \in X \cap Y$.
But given two vectors $x, x' \in X$, we have that $x - x' \in X \cap Y$ if, and only if, $[x] = [x'] \in \dfrac{x}{X \cap Y}$.
Combining all of that, we see that $[x+0] = [x'+0] \in \dfrac{X+Y}{Y}$ if, and only if, $[x] = [x'] \in \frac{X}{X \cap Y}$.

This gives us an obvious correspondence:

$$f : \frac{X+Y}{Y} \to \frac{X}{X \cap Y}$$

sending each $[x + 0] \in \dfrac{X+Y}{Y}$ to $[x] \in \dfrac{X}{X \cap Y}$ and

$$g : \frac{X}{X \cap Y} \to \frac{X+Y}{Y}$$

sending each $[x] \in \dfrac{X}{X \cap Y}$ to $[x + 0] \in \dfrac{X+Y}{Y}$.
These functions are trivially inverse to each other, and so they are isomorphisms.
This ends the proof. □

Theorem 2.4.3.12 (Third Isomorphism Theorem)**.** *Let $X, Y \leq \mathbb{R}^2$ be any two subspaces. Then $Y \subseteq X$ if, and only if, $X/Y \subseteq \mathbb{R}^2/Y$.*

Proof

Assume $Y \subseteq X$. Then $[x] = [x'] \in X/Y$ if, and only if, $x - x' \in Y$. But $x, x' \in \mathbb{R}^2$ means that $x - x' \in Y$ if, and only if, $[x] = [x'] \in \mathbb{R}^2/Y$.
This shows that $[x] \in X/Y$ implies $[x] \in \mathbb{R}^2/Y$ and ends the proof. □

This is actually not the full statements of these theorems, but we don't have the tools to actually work with their full statements at this point in time.

Chapter 3

Linear algebra in higher dimensions

3.1 Introduction

Now that we're reasonably comfortable with $\mathbb{R}^2$ (don't worry, we'll come back to it) we're gonna take the next step: Consider the set $\mathbb{R}^3$ and study what are its vectors, how they behave etc.

But, as it's going to become apparent very soon, this is essentially **not different at all** from what we've already been doing for $\mathbb{R}^2$ - some proofs are actually literally the same, without changing anything.

As a consequence, this chapter shall be much shorter than the previous one. Here's the basic schema of this chapter:

- First, we're gonna define $\mathbb{R}^3$ and stablish its properties. At this point we'll notice how it's basically $\mathbb{R}^2$ all over again, with some few minor changes;
- Then we're gonna prove whatever is new for $\mathbb{R}^3$ and state which results from $\mathbb{R}^2$ still hold.

With that said, let us begin!

3.1.1 Generalizing to $\mathbb{R}^3$

By definition, $\mathbb{R}^3$ is the set of all ordered triples of real numbers - like $(1, 2, 3)$ or $(0, \pi, -4)$ etc. We won't go into as much detail as we did for $\mathbb{R}^2$, but we can prove the following result:

Proposition 3.1.1.1. *The Euclidean space E_* with distinguished point $*$ is in bijection with $\mathbb{R}^3$.*

Corollary 3.1.1.2. *Every point in $\mathbb{R}^3$ can be thought of as either a point in the space, or a vector from the origin to its endpoint.*

The idea is pretty simple - we can think of any ordered triple $(x, y, z) \in \mathbb{R}^3$ as a set of "coordinates" in a grid system which tells us how to move away from the distinguished point $*$: Move x steps away from $*$ in a certain direction, move y steps away from $*$ in a different direction and then z steps away from $*$ in a third direction.

And we can, as before, define:

Definition 3.1.1.3. *Given two vectors $v = (v_1, v_2, v_3), u = (u_1, u_2, u_3) \in \mathbb{R}^3$ we define their **sum** $v + u$ to be the unique vector given by*

$$v + u := (v_1 + u_1, v_2 + u_2, v_3 + u_3).$$

The intuition here is also very similar as it was for $\mathbb{R}^2$: $0, v, u$ define a unique triangle in the space and, by axiom, there's a unique plane containing this triangle. Therefore, $v+u$ is the parallelogram which is contained in that plane and has $0v$ and $0u$ as its sides.

Proposition 3.1.1.4. *The addition of vectors in $\mathbb{R}^3$ is associative, commutative, has identity element and inverses.*

Definition 3.1.1.5. *Let us define a few important subsets of $\mathbb{R}^3$:*

- *The set $\{0\} \times \mathbb{R} \times \mathbb{R}$ will be called the $\mathbb{YZ}$-plane;*
- *The set $\mathbb{R} \times \{0\} \times \mathbb{R}$ will be called the $\mathbb{XZ}$-plane;*
- *The set $\mathbb{R} \times \mathbb{R} \times \{0\}$ will be called the $\mathbb{XY}$-plane;*
- *The set $\mathbb{R} \times \{0\} \times \{0\}$ will be called the $\mathbb{X}$-axis;*
- *The set $\{0\} \times \mathbb{R} \times \{0\}$ will be called the $\mathbb{Y}$-axis;*
- *The set $\{0\} \times \{0\} \times \mathbb{R}$ will be called the $\mathbb{Z}$-axis.*

Definition 3.1.1.6. *Let $v = (v_1, v_2, v_3) \in \mathbb{R}^3$ and $\lambda \in \mathbb{R}$. We define the **scalar multiplication** of v and λ to be the vector $\lambda v \in \mathbb{R}^3$ given by*

$$\lambda v := (\lambda v_1, \lambda v_2, \lambda v_3).$$

Proposition 3.1.1.7. *Scalar multiplication of vectors in $\mathbb{R}^3$ is associative, commutative, has identity element and is distributive over both real and vector addition. Not only that, but $\lambda v = 0$ if, and only if, $\lambda = 0$.*

Definition 3.1.1.8. *Given any vectors $v, u \in \mathbb{R}^3$ we define:*

- $\mathbb{R}v := \{w \in \mathbb{R}^3 \mid w = \lambda v \text{ for some } \lambda \in \mathbb{R}\}$ *the* ***line through zero containing*** v*;*
- $\mathbb{R}v + \mathbb{R}u := \{w \in \mathbb{R}^3 \mid w = \lambda v + \mu u \text{ for some } \lambda, \mu \in \mathbb{R}\}$ *the* ***plane through zero containing*** v ***and*** u*.*

Definition 3.1.1.9. *A subset $X \subseteq \mathbb{R}^3$ is called a* ***subspace*** *if it is closed under addition and scalar multiplication.*

Proposition 3.1.1.10. *Given any two non-zero vectors $v, u \in \mathbb{R}^3$, then both $\mathbb{R}v$ and $\mathbb{R}v + \mathbb{R}u$ are subspaces of $\mathbb{R}^3$.*

Proof

Let $v', v'' \in \mathbb{R}v$ - that is, $v' = \lambda' v$ and $v'' = \lambda'' v$ for some $\lambda', \lambda'' \in \mathbb{R}$. Then, for all $\lambda \in \mathbb{R}$:

$$v' + v'' = (\lambda' v) + (\lambda'' v) = (\lambda' + \lambda'')v \in \mathbb{R}v$$

$$\lambda v' = \lambda(\lambda' v) = (\lambda\lambda')v \in \mathbb{R}v$$

so $\mathbb{R}v$ is closed under addition and scalar multiplication.
Let $w, w' \in \mathbb{R}v + \mathbb{R}u$ - that is, $w = \lambda v + \mu u$ and $w' = \lambda' v + \mu' u$. Then, for all $\omega \in \mathbb{R}$:

$$w + w' = (\lambda v + \mu u) + (\lambda' v + \mu' u) = (\lambda + \lambda')v + (\mu + \mu')u \in \mathbb{R}v + \mathbb{R}u$$

$$\omega w = \omega(\lambda v + \mu u) = (\omega\lambda)v + (\omega\mu)u \in \mathbb{R}v + \mathbb{R}u$$

so $\mathbb{R}v + \mathbb{R}u$ is closed under addition and scalar multiplication.
This ends the proof. □

Now let's give some geometric definitions and interpret them with linear algebra:

Definition 3.1.1.11. *Two lines are said to be*

- ***Parallel*** *if they lie in the same plane and don't meet;*
- ***Skew*** *if they don't lie in the same plane and don't meet;*
- ***Transversal*** *if they meet.*

Similarly, two planes are said to be ***parallel*** *if they don't meet, and transversal if they meet. Finally, a line and a plane are said to be* ***parallel*** *if they don't meet.*

Proposition 3.1.1.12. *Any line in $\mathbb{R}^3$ is of the form $\mathbb{R}v + u$ for some vectors $v, u \in \mathbb{R}^3$, and any plane in $\mathbb{R}^3$ is of the form $\mathbb{R}v + \mathbb{R}u + w$ for some vectors $v, u, w \in \mathbb{R}^3$.*

Proof

Let $r \subseteq \mathbb{R}^3$ be any line. Then $r \cap \mathbb{YZ}$ is either empty, a single point or $r \subseteq \mathbb{YZ}$.

- If $r \cap \mathbb{YZ}$ is a single point u, then the line $r' \subseteq \mathbb{R}^3$ which is parallel to r through zero is such that $r = r' + u$.
- If $r \cap \mathbb{YZ}$ is empty or $r \subseteq \mathbb{YZ}$, we then check $r \cap \mathbb{XY}$ which can be, again, either empty, a single point or $r \subseteq \mathbb{XY}$.
  - If $r \cap \mathbb{XY}$ is a single point, just do as we did above, taking a parallel through zero and adding this single point to it.
  - If $r \cap \mathbb{XY}$ is empty, then the fact that $r \cap \mathbb{YZ}$ is also empty implies that $r \cap \mathbb{XZ}$ is a single point and we can just iterate the construction above.
  - If $r \subseteq \mathbb{XY}$, then $r = \mathbb{XY} \cap \mathbb{YZ} = \mathbb{Y}$ so it's already a line through zero.

Either way, we can always show that r is parallel to a line through zero.

Let $\pi \subseteq \mathbb{R}^3$ be any plane. Then:

- If $\pi \cap \mathbb{YZ}$ is a line s, then we can take π' the plane parallel to π through zero, and $v = s \cap \mathbb{Y}$ Then clearly $\pi = \pi' + v$.
- If $\pi \cap \mathbb{YZ} = \varnothing$, then surely $\pi \cap \mathbb{X} \neq \varnothing$ (since $\mathbb{X} \perp \mathbb{YZ}$), and since $\pi \parallel \mathbb{YZ}$, we can take $v = \mathbb{X} \cap r$ and see that $\pi = \mathbb{YX} + v$.

Once again, we see that no matter what, π is always parallel to a plane through zero.
This ends the proof. □

3.1.2 Linear functions in $\mathbb{R}^3$

In this section we're going to define linear functions for $\mathbb{R}^3$ and see that there's not much new going on.

Definition 3.1.2.1. *Let $f : \mathbb{R}^3 \to \mathbb{R}^3$ be a function. We'll say that f is a **linear function** if $f(v+u) = f(v) + f(u)$ and $f(\lambda v) = \lambda f(v)$ for all $v, u \in \mathbb{R}^3$ and $\lambda \in \mathbb{R}$.*
We'll denote the set of all linear functions in $\mathbb{R}^3$ by $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^3, \mathbb{R}^3)$.

Proposition 3.1.2.2. *Let $f : \mathbb{R}^3 \to \mathbb{R}^3$. Then f is linear if, and only if,*

$$f(x, y, z) = (ax + by + cz, dx + ey + fz, gx + hy + iz)$$

for some $a, b, c, d, e, f, g, h, i \in \mathbb{R}$.

Lemma 3.1.2.3. *Let f be a linear function in $\mathbb{R}^3$. Then f is uniquely determined by how it acts on $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.*

Proof

Since f is linear, we have that for all $(x, y, z) \in \mathbb{R}^3$ the following equation holds:

$$\begin{aligned} f(x, y, z) &= f((x, 0, 0) + (0, y, 0) + (0, 0, z)) \\ &= f(x, 0, 0) + f(0, y, 0) + f(0, 0, z) = xf(1, 0, 0) + yf(0, 1, 0) + zf(0, 0, 1) \end{aligned}$$

So if we put $f(1, 0, 0) = v$, $f(0, 1, 0) = u$ and $f(0, 0, 1) = w$ we see that $f(x, y, z) = xv + yv + zw$.
This ends the proof. □

Corollary 3.1.2.4. *Let $f : \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\} \to \mathbb{R}^3$. Then there's a unique linear function f' in $\mathbb{R}^3$ such that $f'(x, y, z) := xf(1, 0, 0) + yf(0, 1, 0) + zf(0, 0, 1)$.*

Definition 3.1.2.5. *We'll denote the vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ by e_1, e_2 and e_3, respectively.*

Definition 3.1.2.6. *A finite set $X \subseteq \mathbb{R}^3$ is called a **base** of $\mathbb{R}^3$ if any linear function is uniquely determined by the image of X - that is, if we write $X = \{x_1, x_2, \cdots, x_n\}$, then for all $v \in \mathbb{R}^3$, there are uniquely determined $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$ such that $v = \lambda_1 x_1 + \cdots + \lambda_n x_n$ and thus the image of v under any linear function $f : \mathbb{R}^3 \to \mathbb{R}^3$ is given simply by $f(v) = \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n)$.*

Definition 3.1.2.7. *The set $E := \{e_1, e_2, e_3\}$ is called the **canonical base** of $\mathbb{R}^3$.*

3.1.3 Subspaces in $\mathbb{R}^3$

In this section we're going to extend the result that said that subspaces of $\mathbb{R}^2$ had to be zero, lines through zero or $\mathbb{R}^2$ to $\mathbb{R}^3$. With this, we'll be able to see that the sum of two different lines is a plane, the sum of two different lines or two different planes is $\mathbb{R}^3$ and that the intersection of two planes is a line.

Lemma 3.1.3.1. *Let $X, Y \leq \mathbb{R}^3$ be any two subspaces, and take $v \in X \cap Y$ and $t \in X + Y$. Then not only do we have $\mathbb{R}v \subseteq X \cap Y$ and $\mathbb{R}t \subseteq X + Y$, but also $X + \mathbb{R}v = X$ and $Y + \mathbb{R}v = Y$.*

Proof

Let $v \in X \cap Y$ and take $u \in \mathbb{R}v$ - that is, $u = \lambda v$ for some $\lambda \in R$.
Since $v \in X$ and X is a subspace, $\mu v \in X$ for all $\mu \in \mathbb{R}$ - in particular, $u = \lambda v \in X$. Similarly for Y, since $v \in Y$ and Y is a subspace, $\mu v \in Y$ for all $\mu \in \mathbb{R}$ and so $u = \lambda v \in Y$.
But this tells us that $u \in X \cap Y$, by definition of set intersection. This shows that any vector of $\mathbb{R}v$ is in $X \cap Y$ - so $\mathbb{R}v \subseteq X \cap Y$.

Similarly for t, since $t \in X + Y$, we can write $t = x + y$ for some $x \in X$ and $y \in Y$. Take then $t' \in \mathbb{R}t$ - that is, $t' = \tau t$ for some $\tau \in \mathbb{R}$.
Now, this implies that $t' = \tau t = \tau(x + y) = \tau x + \tau y$ and since X and Y are subspaces, $\tau x \in X$ and $\tau y \in Y$ - so $\tau x + \tau y \in X + Y$. This shows $\mathbb{R}t \subseteq X + Y$.

To show the second statement, it suffices to realize that, by the previous statement, every element of $\mathbb{R}v$ already lies in X and that X is a subspace. So taking $u \in \mathbb{R}v$ and $x \in X$, we see that $x + u \in X$, so $X + \mathbb{R}v \subseteq X$.
Conversely, any element of X is, by definition, in the sum $\mathbb{R}v + X$ - for instance, any $x \in X$ can be seen as $0 + x$, because, since $\mathbb{R}v$ is a subspace, $0 \in \mathbb{R}v$. This shows that $X \subseteq \mathbb{R}v + X$ and ends the proof (because the argument for Y is exactly the same). □

This result is mainly important because it tells us that subspaces are closed under taking subspaces - that is, if you take any subspace X and any element inside it, then the line through that element is still in X.

That, combined with the following corollaries, will show that, indeed, subset sum is, in some sense, the "extension" of the concept of set union for vector spaces.

Corollary 3.1.3.2. *Let $X, Y \leq \mathbb{R}^3$ be any two subspaces of $\mathbb{R}^3$ such that $Y \subseteq X$. Then $X + Y = X$.*

Proof

Clearly, we already have $X \subseteq X + Y$ by definition of addition of subspaces.
Take then any $x + y \in X + Y$. Since $Y \subseteq X$, we have that $y \in X$, and since X is a subspace, we have that $x + y \in X$. So $X + Y \subseteq X$.
This ends the proof. □

Remark 3.1.3.3

Compare this to the already known result for sets that $Y \subseteq X$ implies $X \cup Y = Y$.

Corollary 3.1.3.4. *Let $X, Y, Z \leq \mathbb{R}^3$ be any three subspaces such that both or either of $Z \subseteq X$ and $Z \subseteq Y$ hold. Then $X + Y + Z = X + Y$.*

Proof

Once again, by definition of subspace addition, we already have that $X + Y \subseteq X + Y + Z$. Take $x + y + z \in X + Y + Z$. Then, since both or either of $Z \subseteq X$ and $Z \subseteq Y$ holds, we see that $z \in X$ or $z \in Y$.
In the first case, we now use the fact that X is a subspace, so $x + z \in X$ and therefore $x + y + z = (x + z) + y \in X + Y$, which shows that $X + Y + Z \subseteq X + Y$.
In the second case, we use the fact that Y is a subspace, so $y + z \in Y$ and therefore $x + y + z = x + (y + z) \in X + Y$, which shows that $X + Y + Z \subseteq X + Y$.
Either way, the result holds and this finishes the proof. □

Remark 3.1.3.5

Once again, compare this to the already know result about sets that if $Z \subseteq X$ or $Z \subseteq Y$ then $X \cup Y \cup Z = X \cup Y$.

Lemma 3.1.3.6. *Let $X, Y, Z \leq \mathbb{R}^3$. Then*

$$X + (Y \cap Z) = (X + Y) \cap (X + Z)$$

$$X \cap (Y + Z) = (X \cap Y) + (X \cap Z).$$

Proof

Take $x \in X$ and $w \in Y \cap Z$. Then, by definition, $x + w \in X + (Y \cap Z)$. Since Y is a subspace and $w \in Y$, $x + w \in X + Y$. Similarly, since Z is a subspace and $w \in Z$, $x + w \in X + Z$. Therefore, $x + w \in (X + Y) \cap (X + Z)$, so

$$X + (Y \cap Z) \subseteq (X + Y) \cap (X + Z).$$

Conversely, take $v \in (X + Y) \cap (X + Z)$. However, $X \subseteq X + Y$ and $X \subseteq X + Z$ together imply, by definition of intersection, that $(X + Y) \cap (X + Z) \subseteq X$ - so $v \in X$. Finally, since $X \subseteq X + (Y \cap Z)$, this shows that $v \in X + (Y \cap Z)$, so

$$X + (Y \cap Z) = (X + Y) \cap (X + Z).$$

The second statement is analogous:

Clearly $X \cap (Y+Z)$ contains both $X \cap Y$ and $X \cap Z$: If we take $y \in X \cap Y$, it is, in particular, in Y and X. But since it is in Y, it is also in $Y+Z$. Now we see that this y is in both X and $Y+Z$, so it is in $X \cap (Y+Z)$. We can do the same reasoning to show that any $z \in X \cap Z$ is also in $X \cap (Y+Z)$.
Now, by definition of subspace addition, since $X \cap (Y+Z)$ contains both of $X \cap Y$ and $X \cap Z$, it must be contained in their sum - that is, $(X \cap Y) + (X \cap Z)$.
This shows $X \cap (Y+Z) \subseteq (X \cap Y) + (X \cap Z)$.

Conversely, any element of $(X \cap Y) + (X \cap Z)$ is of the form $v+u$ where $v \in X \cap Y$ and $u \in X \cap Z$. Notice, however, that both v and u lie, in particular, in X - and since X is a subspace, $v+u \in X$.
On the other hand, since $v \in X \cap Y$ and $u \in X \cap Z$, in particular we have $v \in Y$ and $u \in Z$, so $v+u \in Y+Z$.
Since $v+u \in X$ and $v+u \in Y+Z$ we can conclude that $v+u \in X \cap (Y+Z)$ and so

$$(X \cap Y) + (X \cap Z) \subseteq X \cap (Y+Z).$$

This shows that $X \cap (Y+Z) = (X \cap Y) + (X \cap Z)$ and ends the proof. □

Remark 3.1.3.7

This is once gain similar to what we had with sets - where unions and intersections distributed over each other.

Now that we're done with some more technical results, let's start characterizing stuffs:

Proposition 3.1.3.8. *Let $v, u \in \mathbb{R}^3$ be any two non-null vectors. Then for any non-null $w \in \mathbb{R}v$ we have that $\mathbb{R}v = \mathbb{R}w$ and for any non-null $t \in \mathbb{R}v + \mathbb{R}u$ such that $t \notin \mathbb{R}v$ and $t \notin \mathbb{R}u$ we have that $\mathbb{R}t + \mathbb{R}u = \mathbb{R}v + \mathbb{R}t = \mathbb{R}v + \mathbb{R}u$.*

Proof

If $w \in \mathbb{R}v$, then $w = \lambda v$ for some $\lambda \in \mathbb{R}$. Take then any other $w' \in \mathbb{R}v$. Once again, this means that $w' = \lambda' v$ for some $\lambda' \in \mathbb{R}$. But now, clearly we have

$$w' = \lambda' v = \lambda' \frac{\lambda}{\lambda} v = \frac{\lambda'}{\lambda} \lambda v = \frac{\lambda'}{\lambda} w$$

so $w' \in \mathbb{R}w$ and $\mathbb{R}v \subseteq \mathbb{R}w$.
Conversely, every $w' \in \mathbb{R}w$ is of the form $w' = \omega w$ for some $\omega \in \mathbb{R}$, but since $w = \lambda v$, we have that

$$w' = \omega w = (\omega\lambda) v$$

so $w' \in \mathbb{R}v$ and $\mathbb{R}w \subseteq \mathbb{R}v$.
Therefore $\mathbb{R}v = \mathbb{R}w$.

To prove the second statement, we proceed analogously: Take $t \in \mathbb{R}v + \mathbb{R}u$. This means that

$t = \tau_1 v + \tau_2 u$ for some $\tau_1, \tau_2 \in \mathbb{R}$. Notice that both of τ_1 and τ_2 must be non-zero, because otherwise t would be on either or both of $\mathbb{R}v$ and $\mathbb{R}u$.
Given any $t' \in \mathbb{R}v + \mathbb{R}u$, once again we can write it as $t' = \tau_1' v + \tau_2' u$.
Since $\tau_1 \neq 0$, we can then do

$$\begin{aligned} t' &= \tau_1' v + \tau_2' u \\ &= \tau_1' \frac{\tau_1}{\tau_1} v + \tau_2' u \\ &= \frac{\tau_1'}{\tau_1} \tau_1 v + \tau_2' u \\ &= \frac{\tau_1'}{\tau_1}(t - \tau_2 u) + \tau_2' u \\ &= \frac{\tau_1'}{\tau_1} t - \frac{\tau_1'}{\tau_1}\tau_2 u + \tau_2' u = \frac{\tau_1'}{\tau_1} t + \left(\tau_2' - \frac{\tau_1'}{\tau_1}\tau_2\right) u \end{aligned}$$

and we see that $t' \in \mathbb{R}t + \mathbb{R}u$.
Since $\tau_2 \neq 0$, we can then do

$$\begin{aligned} t' &= \tau_1' v + \tau_2' u \\ &= \tau_1' v + \tau_2' \frac{\tau_2}{\tau_2} u \\ &= \tau_1' v + \frac{\tau_2'}{\tau_2} \tau_2 u \\ &= \tau_1' v + \frac{\tau_2'}{\tau_2}(t - \tau_1 v) \\ &= \tau_1' v + \frac{\tau_2'}{\tau_2} t - \frac{\tau_2'}{\tau_2}\tau_1 v \\ &= \left(\tau_1' - \frac{\tau_2'}{\tau_2}\tau_1\right) v + \frac{\tau_2'}{\tau_2} t \end{aligned}$$

and we see that $t' \in \mathbb{R}v + \mathbb{R}t$.
So clearly, $t' \in \mathbb{R}v + \mathbb{R}u$ implies both of $t' \in \mathbb{R}t + \mathbb{R}u$ and $t' \in \mathbb{R}v + \mathbb{R}t$ - and therefore, $\mathbb{R}v + \mathbb{R}u$ is contained in both $\mathbb{R}v + \mathbb{R}t$ and $\mathbb{R}t + \mathbb{R}u$.

On the other hand, take $x \in \mathbb{R}t + \mathbb{R}u$. This means that $x = \tau t + \mu u$ for some $\tau, \mu \in \mathbb{R}$. But then, since $t = \tau_1 v + \tau_2 u$, this tells us that

$$x = \tau t + \mu u = \tau(\tau_1 v + \tau_2 u) + \mu u = (\tau\tau_1)v + (\tau\tau_2 + \mu)u$$

so $x \in \mathbb{R}v + \mathbb{R}u$ which shows that $\mathbb{R}t + \mathbb{R}u \subseteq \mathbb{R}v + \mathbb{R}u$.
Similarly, for all $y \in \mathbb{R}v + \mathbb{R}t$ we can write it as $y = \lambda v + \tau t$ for some $\lambda, \tau \in \mathbb{R}$, so

$$y = \lambda v + \tau t = \lambda v + \tau(\tau_1 v + \tau_2 u) = (\lambda + \tau\tau_1)v + (\tau\tau_2)u$$

and we see $y \in \mathbb{R}v + \mathbb{R}u$, which implies $\mathbb{R}v + \mathbb{R}t \subseteq \mathbb{R}v + \mathbb{R}u$.

Now, finally, we have

$$\mathbb{R}v + \mathbb{R}t \subseteq \mathbb{R}v + \mathbb{R}u \subseteq \mathbb{R}t + \mathbb{R}u$$

and

$$\mathbb{R}t + \mathbb{R}u \subseteq \mathbb{R}v + \mathbb{R}u \subseteq \mathbb{R}v + \mathbb{R}t$$

so this implies that $\mathbb{R}v + \mathbb{R}t = \mathbb{R}t + \mathbb{R}u$ and therefore both of them equal $\mathbb{R}v + \mathbb{R}u$.
This ends the proof. □

Remark 3.1.3.9

This first result is a generalization of a similar result in $\mathbb{R}^2$ which said that two vectors are parallel if, and only if, they lie in the same line through zero.
Indeed, this result tells us not only that, but also that a vector t is in the plane containing two other vectors v, u if, and only if, v is in the plane containing t, u and u is in the plane containing t, v.
This gives us a lot of insight for the following definitions:

Definition 3.1.3.10. *Let $v, u \in \mathbb{R}^3$ be two vectors such that $\mathbb{R}v = \mathbb{R}u$. In this case we say that v and u are **parallel**, which we denote by $v \parallel u$.*

Definition 3.1.3.11. *Let $v, u, w \in \mathbb{R}^3$ be three vectors. We say that they are **coplanar** if any one of them is in the plane through the origin containing the other two.*

In symbols: v, u, w are coplanar if any of $v \in \mathbb{R}u + \mathbb{R}w$, $u \in \mathbb{R}v + \mathbb{R}w$ and $w \in \mathbb{R}v + \mathbb{R}u$ hold. We denote this by ${}_v\overset{u}{\triangle}_w$.

Lemma 3.1.3.12. *Let $v, u, w \in \mathbb{R}^3$. Then $v \parallel u$ implies ${}_v\overset{u}{\triangle}_w$.*

Proof

Since $v \parallel u$, we have that $\mathbb{R}v = \mathbb{R}u$, so, clearly, $u \in \mathbb{R}v + \mathbb{R}w$ and thus ${}_v\overset{u}{\triangle}_w$. □

That is, if two vectors are in the same line, then they are already coplanar.
Let's now use these definitions to start breaking $\mathbb{R}^3$ into smaller pieces:

Lemma 3.1.3.13. *The planes $\mathbb{XY}$, $\mathbb{YZ}$ and $\mathbb{ZX}$ are just the sums $\mathbb{X} + \mathbb{Y}$, $\mathbb{Y} + \mathbb{Z}$ and $\mathbb{Z} + \mathbb{X}$, respectively.*

Proof

We'll show that $\mathbb{XY} = \mathbb{X} + \mathbb{Y}$. The other are analogous and will be left as an exercise to the reader.
By definition, $\mathbb{XY} = \mathbb{R} \times \mathbb{R} \times \{0\}$. This means that $v \in \mathbb{XY}$ if, and only if, $v = (x, y, 0)$ for

some $x, y \in \mathbb{R}$.
Clearly, then, for all $v \in \mathbb{XY}$ we can write it as $v = xe_1 + ye_2 + 0e_3 = xe_1 + y_2$ which shows that $v \in \mathbb{X} + \mathbb{Y}$ - and so $\mathbb{XY} \subseteq \mathbb{X} + \mathbb{Y}$.

Conversely, given any $u \in \mathbb{X} + \mathbb{Y}$ there exists some $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ such that $u = x + y$. But since $\mathbb{X} = \mathbb{R}e_1$, we see that $x = x'e_1$ and since $\mathbb{Y} = \Re_2$ we see that $y = y'e_2$, for some $x', y' \in \mathbb{R}$. This tells us that

$$u = x + y = x'e_1 + y'e_2 = x'e_1 + y'e_2 + 0e_3 = (x', y', 0)$$

and so $u \in \mathbb{XY}$.
This shows us that $\mathbb{XY} \subseteq \mathbb{X} + \mathbb{Y}$, and ends the proof. □

Corollary 3.1.3.14. *The following equation holds in* $\mathbb{R}^3$*:*

$$\mathbb{X} + \mathbb{Y} + \mathbb{Z} = \mathbb{R}^3 = \mathbb{XY} + \mathbb{YZ} + \mathbb{ZX}.$$

Proof

The first equation is trivial: $\mathbb{X} = \mathbb{R}e_1$, $\mathbb{Y} = \mathbb{R}e_2$ and $\mathbb{Z} = \mathbb{R}e_3$, by definition, and we know that $E = \{e_1, e_2, e_3\}$ is a base. This means that any vector $v \in \mathbb{R}^3$ can be written with uniquely determined scalars $v_1, v_2, v_3 \in \mathbb{R}$ as

$$v = v_1e_1 + v_2e_2 + v_3e_3$$

this tells us that $v \in \mathbb{X} + \mathbb{Y} + \mathbb{Z}$, and so $\mathbb{R}^3 \subseteq \mathbb{X} + \mathbb{Y} + \mathbb{Z}$.
Conversely, we have that $\mathbb{X} + \mathbb{Y} + \mathbb{Z} \subseteq \mathbb{R}^3$ simply by the fact that every element of $\mathbb{X} + \mathbb{Y} + \mathbb{Z}$ is, by definition, a sum of vectors, all of which lie in $\mathbb{R}^3$ - and hence so does their sum.

We would now like to show that $\mathbb{R}^3 = \mathbb{XY} + \mathbb{YZ} + \mathbb{ZX}$, but in light of the preceding lemma, $\mathbb{XY} + \mathbb{YZ} + \mathbb{XZ}$ is just $\mathbb{X} + \mathbb{Y} + \mathbb{Y} + \mathbb{Z} + \mathbb{Z} + \mathbb{X}$ which we already know is simply $\mathbb{X} + \mathbb{Y} + \mathbb{Z}$ - and this we've already proven to be equal to $\mathbb{R}^3$.
This ends the proof. □

Corollary 3.1.3.15. *We can weaken the preceding equation a bit:*

$$\mathbb{XY} + \mathbb{YZ} = \mathbb{YZ} + \mathbb{ZX} = \mathbb{ZX} + \mathbb{XY} = \mathbb{R}^3.$$

Proof

It follows by the simple observation that all the subspace additions above equal $\mathbb{X} + \mathbb{Y} + \mathbb{Z}$. □

Corollary 3.1.3.16. *Finally, we can weaken it even further:*

$$\mathbb{XY} + \mathbb{Z} = \mathbb{YZ} + \mathbb{X} = \mathbb{ZX} + \mathbb{Y} = \mathbb{R}^3.$$

Proof

The same proof as above, since all of these sums equal $\mathbb{X}+\mathbb{Y}+\mathbb{Z}$. □

Remark 3.1.3.17

All of these can be seen as a generalization of the fact that, in $\mathbb{R}^2$, $\mathbb{X}+\mathbb{Y}=\mathbb{R}^2$. Indeed, we're saying that $\mathbb{R}^3$ can be though of as either a set with three axes ($\mathbb{X}+\mathbb{Y}+\mathbb{Z}=\mathbb{R}^3$), two planes ($\mathbb{XY}+\mathbb{YZ}=\mathbb{YZ}+\mathbb{ZX}=\mathbb{ZX}+\mathbb{XY}=\mathbb{R}^3$) or an axis and a plane ($\mathbb{XY}+\mathbb{Z}=\mathbb{YZ}+\mathbb{X}=\mathbb{ZX}+\mathbb{Y}=\mathbb{R}^3$).

We can now use this to start classifying all subspaces of $\mathbb{R}^3$:

Lemma 3.1.3.18. *Let $v,u\in\mathbb{XY}$ be two non-null non-parallel vectors. Then $\mathbb{R}v+\mathbb{R}u=\mathbb{XY}$.*

Proof

Since both $v,u\in\mathbb{XY}$ we can write them as $v=(v_1,v_2,0)$ and $u=(u_1,u_2,0)$.
If v_2 or u_2 are zero, then we're done: For instance, if $v_2=0$, then $v=(v_1,0,0)\in\mathbb{X}$, and since v is non-null, $v_1\neq 0$. So we can define $v':=\frac{u_1}{v_1}v$ and see that

$$v'=\frac{u_1}{v_1}(v_1,0,0)=(u_1,0,0)$$

so

$$u-v'=(0,u_2,0)\in\mathbb{Y}.$$

But since $v\nparallel u$, we cannot have $u_2=0$. Therefore, we can write

$$e_1=\frac{1}{v_1}v\in\mathbb{R}v$$

$$e_2=\frac{1}{u_2}(u-v')\in\mathbb{R}v+\mathbb{R}u,$$

so $\mathbb{R}v+\mathbb{R}u$ contains both $\mathbb{X}$ and $\mathbb{Y}$ and hence it contains $\mathbb{XY}$.
If, instead of $v_2=0$ we had assumed that $u_2=0$ we would have arrived at a similar conclusion.

Similarly, we can do the same consideration for the cases where either v_1 or u_1 are zero. For instance, if $u_1=0$, then, since u is non-null, we have that $u_2\neq 0$, so $u':=\frac{v_2}{u_2}u$ is such that

$$u'=\frac{v_2}{u_2}(0,u_2,0)=(0,v_2,0)\in\mathbb{Y}$$

and

$$v-u'=(v_1,v_2,0)-(0,v_2,0)=(v_1,0,0)\in\mathbb{X}$$

so, once again, $\mathbb{R}v+\mathbb{R}u$ contains both e_1 and e_2 and hence it contains $\mathbb{XY}$.

Once more, we can do the same reasoning for the case where $v_1 = 0$ and arrive at the same conclusion.

Finally, let's assume that all of v_1, v_2, u_1, u_2 are non-zero. In this case, we define $v' := \frac{u_2}{v_2}v$ and see that

$$v' = \frac{u_2}{v_2}v = \left(\frac{u_2}{v_2}v_1, u_2, 0\right).$$

For the sake of simplicity, let's call $v_1' := \frac{u_2}{v_2}v_1$ so v' becomes just $(v_1', u_2, 0)$.
Now, once more, we see that

$$u - v' = (u_1 - v_1', 0, 0) \in \mathbb{X}.$$

We claim that $u_1 - v_1' \neq 0$ or, in other words, that $u_1 \neq v_1'$. Indeed, if they were equal we would have

$$u = (u_1, u_2, 0) = \left(v_1', u_2\frac{v_2}{v_2}, 0\right) = \left(\frac{u_2}{v_2}v_1, \frac{u_2}{v_2}v_2, 0\right) = \frac{u_2}{v_2}(v_1, v_2, 0) = \frac{u_2}{v_2}v$$

in other words, if $u_1 = v_1'$ then $v \parallel u$. But we're assuming $v \nparallel u$, so, therefore, $u_1 \neq v_1'$.
This shows that we can divide $u - v'$ by $u_1 - v_1'$ and get e_1, so $e_1 \in \mathbb{R}v + \mathbb{R}u$.

Similarly, since $v_1 \neq 0$, we can define $v'' := \frac{u_1}{v_1}v$, so

$$v'' = \left(u_1, \frac{u_1}{v_1}v_2, 0\right)$$

and, once again, we'll call $v_2'' := \frac{u_1}{v_1}v_2$, so $v'' = (u_1, v_2'', 0)$. Now, clearly,

$$u - v'' = (0, u_2 - v_2'', 0) \in \mathbb{Y}$$

and we can, using the same arguments as before, show that $u_2 = v_2''$ if, and only if, $v \parallel u$ - which tells us (since $v \nparallel u$) that if we divide $u - v''$ by $u_2 - v_2''$ we obtain e_2. This tells us that $e_2 \in \mathbb{R}v + \mathbb{R}u$.
Since both $e_1, e_2 \in \mathbb{R}v + \mathbb{R}u$ we can further see that $\mathbb{XY} \subseteq \mathbb{R}v + \mathbb{R}u$. This ends the proof. □

Corollary 3.1.3.19. *The same result holds for two non-null non-parallel vectors in $\mathbb{Y}Z$ and $\mathbb{Z}X$ with essentially the same proof.*

Remark 3.1.3.20

This shows that the smallest subspace of the coordinate planes containing two non-parallel vectors is the plane itself.
We will now generalize this for arbitrary planes through zero:

Proposition 3.1.3.21. *Let $v, u \in \mathbb{R}^3$ be two non-null non-parallel vectors and $\pi \subseteq \mathbb{R}^3$ be any plane through zero such that $v, u \in \pi$. Then $\mathbb{R}v + \mathbb{R}u = \pi$.*

Proof

This proof follows in the same spirit as before: Let $\pi = \mathbb{R}p + \mathbb{R}q$ be any plane through zero. We're gonna show that $\mathbb{R}v + \mathbb{R}u$ contains both p and q - and thus contains π.
Now, $v, u \in \pi$ implies the existence of $\lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{R}$ such that

$$v = \lambda_1 p + \lambda_2 q$$

$$u = \mu_1 p + \mu_2 q.$$

First, notice that we cannot have $\lambda_1 = \lambda_2 = \mu_1 = \mu_2 = 0$ because that would imply $v = 0 = u$ and we're assuming they're both non-null. This means that, at least one of λ_1, λ_2 and one of μ_1, μ_2 must be non-zero.
If either λ_2 or μ_2 are zero, then we're done: For instance, if $\lambda_2 = 0$ then $v = \lambda_1 p \in \mathbb{R}p$, and since v is non-null, λ_1 cannot be zero. So we can define $v' := \dfrac{\mu_1}{\lambda_1} v$ and see that

$$v' = \frac{\mu_1}{\lambda_1} v = \mu_1 p$$

so

$$u - v' = (\mu_1 p + \mu_2 q) - \mu_1 p = \mu_2 q.$$

Now, since $v \nparallel u$, this implies that μ_2 cannot be zero (otherwise, we'd have $v = \lambda_1 p$ and $u = \mu_1 p$, so they'd be parallel). So we can write

$$p = \frac{1}{\lambda_1} v \in \mathbb{R}v$$

$$q = \frac{1}{\mu_2}(u - v') \in \mathbb{R}v + \mathbb{R}u,$$

so $\mathbb{R}v + \mathbb{R}u$ contains both p and q - and thus it contains π.
If instead of λ_2 we had taken μ_2 to be non-zero, we would have arrived at a similar conclusion using essentially the same steps.

Similarly, we can do the same consideration for the cases where either λ_1 or μ_1 are zero. For instance, if $\mu_1 = 0$, then, since u is non-null, we have that $\mu_2 \neq 0$, so we can define $u' := \dfrac{\lambda_2}{\mu_2} u$ in such a way that

$$u' = \frac{\lambda_2}{\mu_2} u = \lambda_2 q,$$

so

$$v - u' = (\lambda_1 p + \lambda_2 q) - \lambda_2 q = \lambda_1 p.$$

Once again, since $v \nparallel u$, we cannot have $\lambda_1 = 0$. So we can write

$$q = \frac{1}{\mu_2} u$$

$$p = \frac{1}{\lambda_1}(v - u')$$

so $\mathbb{R}v + \mathbb{R}u$ contains both p and q - and thus contains π.

Finally, if none of $\lambda_1, \lambda_2, \mu_1, \mu_2$ are zero, we define, once more, $v' := \frac{\mu_2}{\lambda_2} v$ and see that

$$v' = \frac{\mu_2}{\lambda_2} v = \frac{\mu_2}{\lambda_2} \lambda_1 p + \mu_2 q.$$

For the sake of simplicity, let's call $\lambda_1' := \frac{\mu_2}{\lambda_2} \lambda_1$, so $v' = \lambda_1' p + \mu_2 q$.
Now, once again, we see that

$$u - v' = (\mu_1 p + \mu_2 q) - (\lambda_1' p + \mu_2 q) = (\mu_1 - \lambda_1') p.$$

We claim that $\mu_1 \neq \lambda_1'$. Indeed, if they were equal we would have

$$u = \mu_1 p + \mu_2 q = \lambda_1' p + \mu_2 \frac{\lambda_2}{\lambda_2} q = \frac{\mu_2}{\lambda_2} \lambda_1 p + \frac{\mu_2}{\lambda_2} \lambda_2 q = \frac{\mu_2}{\lambda_2} (\lambda_1 p + \lambda_2 q) = \frac{\mu_2}{\lambda_2} v,$$

which contradicts the fact that $v \nparallel u$. So $\mu_1 \neq \lambda_1'$.
This tells us that

$$p = \frac{1}{\mu_1 - \lambda_1'} (u - v')$$

so $p \in \mathbb{R}v + \mathbb{R}u$.

Analogously, if we define $v'' := \frac{\mu_1}{\lambda_1} v$ we can see that

$$v'' = \frac{\mu_1}{\lambda_1} v = \mu_1 p + \frac{\mu_1}{\lambda_1} \lambda_2 q.$$

For the sake of simplicity, let's call $\lambda_2'' := \frac{\mu_1}{\lambda_1} \lambda_2$, so $v'' = \mu_1 p + \lambda_2'' q$.
Once again, we can see that

$$u - v'' = (\mu_1 p + \mu_2 q) - (\mu_1 p + \lambda_2'' q) = (\mu_2 - \lambda_2'') q$$

and we can see, by the same reasoning, that $\mu_2 = \lambda_2''$ if, and only if, $v \parallel u$. Since $v \nparallel u$, then, we can conclude that $\mu_2 \neq \lambda_2''$ and so

$$q = \frac{1}{\mu_2 - \lambda_2''} (u - v'')$$

which tells us that $u \in \mathbb{R}v + \mathbb{R}u$.

Combining these two, we see that both p and q are in $\mathbb{R}v + \mathbb{R}u$ - and thus π is also in $\mathbb{R}v + \mathbb{R}u$.
This ends the proof. □

Remark 3.1.3.22

This tells us that just like lines are determined by any non-null vector in them, planes are determined by any two non-null non-parallel vectors in them.

From this it now follows that:

Corollary 3.1.3.23. *Let $v, u, w \in \mathbb{R}^3$ be three non-null non-coplanar vectors. Then $\mathbb{R}v + \mathbb{R}u + \mathbb{R}w = \mathbb{R}^3$.*

Proof

We'll prove that $\mathbb{R}v + \mathbb{R}u + \mathbb{R}w$ contains all of the planes $\mathbb{XY}$, $\mathbb{YZ}$ and $\mathbb{ZX}$. Since they're pretty much the same proof, we'll only prove it for $\mathbb{XY}$ and leave the other two as exercises to the reader.
To start it off, write $v = (v_1, v_2, v_3)$, $u = (u_1, u_2, u_3)$ and $w = (w_1, w_2, w_3)$.
Now, if any two of the last coordinates above are zero - say, $v_3 = u_3 = 0$ - then clearly $\mathbb{R}v+\mathbb{R}u$ already contains $\mathbb{XY}$ (since v, u, w are non-coplanar and, therefore, v, u are non-parallel).
Therefore, we can assume that, at most, one of the last coordinates is zero.
Assume, without loss of generality, that u_3 and w_3 are certainly non-zero (so v_3 can be zero).
Then we can define $u' := \frac{w_3}{u_3}u$ and see

$$u' = \frac{w_3}{u_3}u = (u'_1, u'_2, w_3)$$

where $u'_1 := \frac{w_3}{u_3}u_1$ and $u'_2 := \frac{w_3}{u_3}u_2$.
Now:

$$w - u' = (w_1, w_2, w_3) - (u'_1, u'_2, w_3) = (w_1 - u'_1, w_2 - u'_2, 0).$$

Now we have two possible cases:

- If $v_3 = 0$, then $v = (v_1, v_2, 0)$. We claim that $v \nparallel (w - u')$. Indeed, since $w - u' = w - \frac{w_3}{u_3}u \in \mathbb{R}u + \mathbb{R}w$, we have that if $v \parallel (w - u')$ then $v \in \mathbb{R}u + \mathbb{R}w$.

  But we're assuming, by hypothesis, that v, u, w are non-coplanar, so we get that v cannot be parallel to $w - u'$.

  This ends the reasoning for this case, because now v and $w - u'$ are two non-parallel vectors in $\mathbb{XY}$ and, thus, by the preceding proposition, $\mathbb{XY} = \mathbb{R}v + \mathbb{R}(w - u')$. But clearly, we have that $w - u' \in \mathbb{R}u + \mathbb{R}w$, which implies $\mathbb{R}(w - u') \subseteq \mathbb{R}u + \mathbb{R}w$ and thus

  $$\mathbb{XY} \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w.$$

- If, however, v_3 does not equal zero, we have to do a small adjustment: Let, as before, $v' := \frac{w_3}{v_3}v$, so

  $$v' = \frac{w_3}{v_3}v = (v'_1, v'_2, w_3)$$

where $v_1' = \dfrac{w_3}{v_3}v_1$ and $v_2' = \dfrac{w_3}{v_3}v_2$. Now:

$$w - v' = (w_1, w_2, w_3) - (v_1', v_2', w_3) = (w_1 - v_1', w_2 - v_2', 0).$$

We claim that $(w - v') \nparallel (w - u')$. Indeed, if they were parallel then there would be some $\lambda \in \mathbb{R}$ such that $(w - v') = \lambda(w - u')$. But then:

$$\begin{aligned}
w - v' &= \lambda(w - u') \\
w - \frac{w_3}{v_3}v &= \lambda\left(w - \frac{w_3}{u_3}u\right) \\
w - \frac{w_3}{v_3}v &= \lambda w - \frac{\lambda w_3}{u_3}u \\
-\frac{w_3}{v_3}v &= \lambda w - \frac{\lambda w_3}{u_3}u - w \\
-\frac{w_3}{v_3}v &= (\lambda - 1)w - \frac{\lambda w_3}{u_3}u \\
\frac{w_3}{v_3}v &= \frac{\lambda w_3}{u_3}u - (\lambda - 1)w
\end{aligned}$$

and since we're assuming that $w_3 \neq 0$, we see that

$$v = \frac{v_3}{w_3}\left(\frac{\lambda w_3}{u_3}u - (\lambda - 1)w\right) = \frac{\lambda v_3}{u_3}u - \frac{(\lambda - 1)v_3}{w_3}w$$

and so $v \in \mathbb{R}u + \mathbb{R}w$. But this contradicts our initial hypothesis that v, u, w are non-coplanar.

This shows that, indeed, $(w - v') \nparallel (w - u')$ and since they're both in $\mathbb{XY}$ and are non-parallel, we see, by the preceding proposition, that $\mathbb{XY} \subseteq \mathbb{R}(w - v') + \mathbb{R}(w - u')$.

Finally, $(w-v') \in \mathbb{R}v+\mathbb{R}w$ and $(w-u') \in \mathbb{R}u+\mathbb{R}w$ implies that $\mathbb{R}(w-v')+\mathbb{R}(w-u') \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w$.

Combining these two we get that $\mathbb{XY} \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w$.

Either way, we can conclude that if v, u, w are non-coplanar, then $\mathbb{XY} \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w$.

Like we said at the beginning, we can repeat this same argument by eliminating the second or first coordinates to show that both $\mathbb{YZ}$ and $\mathbb{ZX}$ are also in $\mathbb{R}v + \mathbb{R}u + \mathbb{R}w$.
This shows that $\mathbb{XY} + \mathbb{YZ} + \mathbb{ZX} \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w$.
But we've already proven that $\mathbb{XY} + \mathbb{YZ} + \mathbb{ZX} = \mathbb{R}^3$, so $\mathbb{R}^3 \subseteq \mathbb{R}v + \mathbb{R}u + \mathbb{R}w \subseteq \mathbb{R}^3$ shows us that

$$\mathbb{R}v + \mathbb{R}u + \mathbb{R}w = \mathbb{R}^3$$

which ends the proof. □

Now we're ready to fully classify all subspaces in $\mathbb{R}^3$:

Theorem 3.1.3.24. *The only possible subspaces in $\mathbb{R}^3$ are zero, lines through zero, planes through zero and $\mathbb{R}^3$.*

Proof

Let X be a subspace in $\mathbb{R}^3$. If X has only one point, then it's zero (since all subspaces have zero).
If it has more than one point, say v, then $\mathbb{R}v \subseteq X$. Now, if $X = \mathbb{R}v$, then X is a line through zero.
If not, then there's a point $u \in X$ such that $v \nparallel u$. So $\mathbb{R}v + \mathbb{R}u \subseteq X$. If $X = \mathbb{R}v + \mathbb{R}u$, then X is a plane through zero.
If not, then there's a point $w \in X$ which is non-coplanar with v, u. So $\mathbb{R}v + \mathbb{R}u + \mathbb{R}w \subseteq X$. But we now know that since v, u, w are non-coplanar, then $\mathbb{R}v + \mathbb{R}u + \mathbb{R}w = \mathbb{R}^3$, so this shows that $\mathbb{R}^3 \subseteq X \subseteq \mathbb{R}^3$ - and so $X = \mathbb{R}^3$.
This ends the proof. □

Corollary 3.1.3.25. *Let $X, Y \leq \mathbb{R}^3$ be two distinct subspaces of $\mathbb{R}^3$. Then:*

a) If $X = Y = 0$, then $X + Y = X \cap Y = 0$;

b) If X and Y are lines, then $X + Y$ is a plane and $X \cap Y = 0$;

c) If X and Y are planes, then $X + Y = \mathbb{R}^3$ and $X \cap Y$ is a line.

Proof

a) This item is trivial.

b) Let $X = \mathbb{R}v$ and $Y = \mathbb{R}u$ be two lines. We know that $X \neq Y$ if, and only if, $v \nparallel u$. But we also know that $v \nparallel u$ if, and only if, $\mathbb{R}v + \mathbb{R}u$ is a plane and $\mathbb{R}v \cap \mathbb{R}u = 0$. This proves this item.

c) Let $X = \mathbb{R}v + \mathbb{R}u$ and $Y = \mathbb{R}p + \mathbb{R}q$. We know that $X \cap Y$ is a subset of both X and Y which is also a subspace of $\mathbb{R}^3$. Therefore, $X \cap Y$ is either zero, a line through zero, a plane through zero or $\mathbb{R}^3$.

It certainly cannot be $\mathbb{R}^3$, since $X \subset \mathbb{R}^3$. It also cannot be a plane, otherwise there would be a plane inside both X and Y which would, at once, be different from both of them and contain two non-parallel vectors - which is impossible.

So it's either a line or zero.

But then, this tells us that at least one triple in $\{v, u, p, q\}$ is non-coplanar - and hence that the sum of the lines through them is $\mathbb{R}^3$ - so we already get that $X + Y = \mathbb{R}^3$.

For instance, if v, u, p are non-coplanar, then $X + Y = \mathbb{R}v + \mathbb{R}u + \mathbb{R}p = \mathbb{R}^3$. But this

means that there exist some $q_1, q_2, q_3 \in \mathbb{R}$ such that

$$q = q_1 v + q_2 u + q_3 p.$$

But we can rearrange this into

$$-q_3 p + q = q_1 v + q_2 u$$

which tells us that $-q_3 p + q \in \mathbb{R}v + \mathbb{R}u = X$. But $-q_3 p + q \in \mathbb{R}p + \mathbb{R}q = Y$, by definition, so we have that $-q_3 p + q \in X \cap Y$ and therefore, $\mathbb{R}(-q_3 p + q) \subseteq X \cap Y$.

So we know that $X \cap Y$ contains a line and isn't a plane - so it can only be that line.

This shows that $X \cap Y$ is a line and $X + Y = \mathbb{R}^3$.

This ends the proof. □

Remark 3.1.3.26

This tells us that these subspaces behave exactly like their namesake geometric counterparts: The whole space is bigger than planes, which are bigger than lines, which are bigger than points; planes meet in lines, and lines meet in points; two lines determine a unique plane, and two planes determine a unique space.

3.1.4 Spanning sets and linear dependency in $\mathbb{R}^3$

Like we said previously, this section wasn't really necessary for $\mathbb{R}^3$. In $\mathbb{R}^3$, although not entirely necessary, it starts to become a little less useless - even if only for introducing useful notation.

Definition 3.1.4.1. *Let $v \in \mathbb{R}^3$ be any vector. We define the **subspace spanned by** v to be the subspace* $\operatorname{span} v$ *defined by*

$$\operatorname{span} v := \mathbb{R}v.$$

*Analogously, given any finite set $X = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^3$, we define the **subspace spanned by** X to be the subspace* $\operatorname{span} X$ *defined by*

$$\operatorname{span} X := \mathbb{R}x_1 + \mathbb{R}x_2 + \cdots + \mathbb{R}x_n.$$

Now we're going to introduce a notation that will follow and haunt us forever: The sigma notation.

Definition 3.1.4.2. *Let $X = \{x_1, x_2, \cdots, x_n\}$ be a finite set. We denote the **sum of all elements of** X by the symbol $\sum_{i=1}^{n} x_i$. In other words,*

$$\sum_{i=1}^{n} x_i := x_1 + x_2 + \cdots + x_n$$

so it's just a shorthand notation for not writing long sums.

Remark 3.1.4.3

At this point there are two remarks that need to be made:
First, how to interpret the sigma notation. The i is called the summation index. The $i = 1$ below the $\sum$ symbol means "we'll start making $i = 1$ and the n above the *um* symbol means "we'll stop when $i = n$. They're called, respectively, the summation limits/extremes/starting and ending point.
So then the summation proceeds by starting at the starting index and then increasing one by one until it reaches the ending index.

Second, we've already proven in the set theory chapter that any finite set is in bijection with a natural number - that's why we can take any finite set X and give its elements numbered indices.
If X was infinite, however, notice that it wouldn't be possible: For instance, we cannot give natural numbers as indices to the elements of $\mathbb{R}$.

Example(s)

Let $X = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^3$. Then we can write

$$\operatorname{span} X = \sum_{i=1}^{n} \mathbb{R}x_i$$

which is way more compact.
Let E be the canonical base of $\mathbb{R}^3$. Then

$$\operatorname{span} E = \sum_{i=1}^{3} \mathbb{R}e_i = \mathbb{R}e_1 + \mathbb{R}e_2 + \mathbb{R}e_3 = \mathbb{X} + \mathbb{Y} + \mathbb{Z} = \mathbb{R}^3.$$

In some sense, the subspace spanned by a set is the smallest subspace containing that set. This can be made precise with the following statement:

Lemma 3.1.4.4. *Let $X \subseteq \mathbb{R}^3$ be a finite set and let*

$$\mathcal{X} := \{Y \leq \mathbb{R}^3 \mid X \subseteq Y\}$$

be the set of all subspaces of $\mathbb{R}^3$ containing X.

Then, if we denote by $\bigcap \mathcal{X}$ the "intersection of all elements of $\mathcal{X}$", we have

$$\operatorname{span} X = \bigcap \mathcal{X}.$$

Proof

First things first, do notice that X belongs to both sides.
Clearly, $\operatorname{span} X$ is a subspace of $\mathbb{R}^3$ containing X, so $\operatorname{span} X \in \mathcal{X}$, which tells us that

$$\bigcap \mathcal{X} \subseteq \operatorname{span} X.$$

On the other hand, let $Y \in \mathcal{X}$. We're going to show that $\operatorname{span} X \subseteq Y$.
But this is obvious: Since Y is a subspace, it is closed under addition and scaling. But since Y contains X, any sum and scaling of elements of X is also in Y. This means that the set of all sums and scalings of elements of X (that is, $\operatorname{span} X$) is contained in Y.
So

$$\operatorname{span} X \subseteq Y.$$

Finally, notice that $\bigcap \mathcal{X}$ is also a subspace of $\mathbb{R}^3$ containing X (because, as stated, $X \subseteq \bigcap \mathcal{X}$ and intersection of subspaces is a subspace) - that is, $\bigcap \mathcal{X} \in \mathcal{X}$. So, by what we just did, any element ov $\mathcal{X}$ contains $\operatorname{span} X$ - in particular, $\bigcap \mathcal{X}$ contains $\operatorname{span} X$ - that is,

$$\operatorname{span} X \subseteq \bigcap \mathcal{X}.$$

This ends the proof.

□

So it makes sense to think of the span of a set of vectors as the "smallest" vector space containing that set of vectors. Now we can start working with linear dependency:

Definition 3.1.4.5. *Let $X \subseteq \mathbb{R}^3$ be a finite set. We say that X is **linearly dependent** if there is a proper subset $Y \subset X$ such that* $\operatorname{span} X = \operatorname{span} Y$.

*Conversely, we say that X is **linearly independent** if for all proper subsets $Y \subset X$, we have that* $\operatorname{span} Y \subset \operatorname{span} X$.

Example(s)

Let $E = \{e_1, e_2, e_3\} \subseteq \mathbb{R}^3$. We claim that E is linearly independent:
Indeed, $\mathcal{P}(E) = \{\varnothing, \{e_1\}, \{e_2\}, \{e_3\}, \{e_1, e_2\}, \{e_1, e_3\}, \{e_2, e_3\}, E\}$, and we know that $\operatorname{span} E = \mathbb{R}^3$.
So:

$$\begin{aligned}
\operatorname{span} \varnothing &= 0 \subset \mathbb{R}^3 \\
\operatorname{span}\{e_1\} &= \mathbb{X} \subset \mathbb{R}^3 \\
\operatorname{span}\{e_2\} &= \mathbb{Y} \subset \mathbb{R}^3 \\
\operatorname{span}\{e_3\} &= \mathbb{Z} \subset \mathbb{R}^3 \\
\operatorname{span}\{e_1, e_2\} &= \mathbb{XY} \subset \mathbb{R}^3 \\
\operatorname{span}\{e_1, e_3\} &= \mathbb{XZ} \subset \mathbb{R}^3 \\
\operatorname{span}\{e_2, e_3\} &= \mathbb{YZ} \subset \mathbb{R}^3
\end{aligned}$$

and we see that any proper subset of E spans a proper subset of $\operatorname{span} E$ - so, by definition, E is linearly independent.

On the other hand, let $X = \{e_1, e_2, (1, -8, 0)\}$ We claim that X is linearly dependent.
To see that, first we need to compute $\operatorname{span} X$:

$$\begin{aligned}
\operatorname{span} X &= \mathbb{R}e_1 + \mathbb{R}e_2 + \mathbb{R}(1, -8, 0) \\
&= \mathbb{X} + \mathbb{Y} + \mathbb{R}(1, -8, 0) \\
&= \mathbb{XY} + \mathbb{R}(1, -8, 0) = \mathbb{XY}
\end{aligned}$$

since $(1, -8, 0 \in \mathbb{XY})$.
Now, clearly, $\{e_1, e_2\} \subset X$ is a proper subset such that

$$\operatorname{span}\{e_1, e_2\} = \mathbb{XY} = \operatorname{span} X$$

so X is linearly dependent.

This gives us great insight on the following proof:

Lemma 3.1.4.6. *Let $X \subseteq \mathbb{R}^3$ be a finite set. Then X is linearly dependent if, and only if, there is some $x \in X$ such that* $\operatorname{span}(X \setminus \{x\}) = \operatorname{span} X$.

Proof

Clearly, if there is some $x \in X$ such that $\operatorname{span}(X \setminus \{x\}) = \operatorname{span} X$, then $X \setminus \{x\}$ is a proper subset of X which spans the same subspace as X, so X is linearly dependent.

Conversely, if X is linearly dependent, then there's some proper subset $Y \subset X$ such that $\operatorname{span} Y = \operatorname{span} X$.
Now, let $Z := X \setminus Y$. Choose any $z \in Z$. We claim that $\operatorname{span}(X \setminus \{z\}) = \operatorname{span} X$.
This is obvious: Since $Y = X \setminus Z$, we have that $Y \subseteq X \setminus \{z\}$, so $\operatorname{span} Y \subseteq \operatorname{span}(X \setminus \{z\})$.
But $(X \setminus \{z\}) \subseteq X$, so $\operatorname{span}(X \setminus \{z\}) \subseteq \operatorname{span} X$.
Finally, since $\operatorname{span} Y = \operatorname{span} X$ we get the following expression:

$$\operatorname{span} X = \operatorname{span} Y \subseteq \operatorname{span}(X \setminus \{z\}) \subseteq \operatorname{span} X$$

and therefore $\operatorname{span}(X \setminus \{z\}) = \operatorname{span} X$.
This ends the proof. □

Corollary 3.1.4.7. *Let $X \subseteq \mathbb{R}^3$ be a finite set. Then X is linearly dependent if, and only if, there's some $x \in X$ such that $x \in \operatorname{span}(X \setminus \{x\})$.*

Proof

If there's some $x \in X$ such that $x \in \operatorname{span}(X \setminus \{x\})$, then $\mathbb{R}x \subseteq \operatorname{span}(X \setminus \{x\})$, so

$$\operatorname{span} X = \sum_{x_i \in X} \mathbb{R}x_i = \sum_{x_i \in X \setminus \{x\}} \mathbb{R}x_i + \mathbb{R}x = \operatorname{span}(X \setminus \{x\}) + \mathbb{R}x = \operatorname{span}(X \setminus \{x\})$$

and X is linearly dependent.

Conversely, if X is linearly dependent, then there's some $x \in X$ such that $\operatorname{span} X = \operatorname{span}(X \setminus \{x\})$. But this means that

$$\operatorname{span} X = \operatorname{span}(X \setminus \{x\}) + \mathbb{R}x = \operatorname{span}(X \setminus \{x\})$$

and hence that $\mathbb{R}x \in \operatorname{span}(X \setminus \{x\})$. But this happens if, and only if, $x \in \operatorname{span}(X \setminus \{x\})$.
This ends the proof. □

Corollary 3.1.4.8. *Let $X = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^3$ be a finite set. Then X is linearly dependent if, and only if, there are real numbers $a_1, a_2, \cdots, a_n \in \mathbb{R}$ not all zero such that*

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0.$$

Proof

Let X be linearly dependent. Then there's some $x \in X$ which is spanned by $X \setminus \{x\}$. Let $X \setminus \{x\} = \{x'_1, x'_2, \cdots, x'_{n-1}\}$, just to give those elements a name.
But then, $x \in \operatorname{span} X \setminus \{x\}$ implies the existence of some real numbers $a_1, a_2, \cdots, a_{n-1} \in \mathbb{R}$ such that

$$x = a_1 x'_1 + a_2 x'_2 \cdots + a_{n-1} x'_{n-1}$$

and hence

$$a_1 x'_1 + a_2 x'_2 + \cdots + a_{n-1} x'_{n-1} + (-1)x = 0.$$

So there are real numbers $a_1, a_2, \cdots, a_{n-1}, (-1) \in \mathbb{R}$ not all zero (since one of them is -1) such that

$$a_1 x'_1 + a_2 x'_2 + \cdots + a_{n-1} x'_{n-1} + (-1)x = 0.$$

Conversely, if there are real numbers $a_1, a_2, \cdots, a_n \in \mathbb{R}$ not all zero such that

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = 0,$$

let us assume (without loss of generality) that $a_1 \neq 0$. If it isn't you can just do the same reasoning by replacing the first non-null index with 1, and the proof will still work.
Since $a_1 \neq 0$, we can rewrite the above equation as

$$x_1 + \frac{a_2}{a_1} x_2 + \cdots + \frac{a_n}{a_1} x_n = 0.$$

Let $a'_i := \dfrac{a_i}{a_1}$ for all $i \leq n$. Then this is just

$$x_1 + a'_2 x_2 + \cdots + a'_n x_n = 0$$

which allows us to do

$$x_1 = -a'_2 x_2 - a'_3 x_3 - \cdots - a'_n x_n.$$

But this tells us that $x_1 \in \operatorname{span} X \setminus \{x_1\}$ - and so X is linearly dependent.
This ends the proof. □

So, all these results together tell us that a set being linearly dependent means that one of the vectors is "superfluous" in the sense that it is already spanned by the other vectors in the set. Conversely, a set being linearly independent means that none of the vectors are superfluous, and you need each one of them to span the whole set.

But this has an obvious consequence in $\mathbb{R}^3$:

Lemma 3.1.4.9. *Let $X \subseteq \mathbb{R}^3$ be a finite set. If $\#X > 3$ then X is linearly dependent.*

Proof

This is obvious. If $X = \{x_1, x_2, x_3, \cdots, x_n\}$ with $n > 3$, then we can consider cases:

- $\{x_1, x_2, x_3\}$ is linearly dependent.

  This would imply that one of them is already spanned by the other two (and, in particular, by all the other vectors in X) - and hence the whole X would be linearly dependent.

- $\{x_1, x_2, x_3\}$ is linearly independent.

  This would imply that none of them is spanned by the others. In particular, this implies that x_1 isn't spanned by $\{x_2, x_3\}$ - that is, x_1, x_2, x_3 are non-coplanar.

  But we know that any three non-null non-coplanar vectors satisfy

  $$\mathbb{R}x_1 + \mathbb{R}x_2 + \mathbb{R}x_3 = \mathbb{R}^3$$

  so $\text{span}\{x_1, x_2, x_3\} = \mathbb{R}^3$. This means that every other vector in $\mathbb{R}^3$ (in particular, any other vector in X) is already spanned by $\{x_1, x_2, x_3\}$ - so X is linearly dependent.

Either way, we can always see that X is linearly dependent.
This ends the proof. □

Corollary 3.1.4.10. *Equivalently, if X is linearly dependent, then $\#X \leq 3$.*

Proof

This is the same result, but phrased in another manner and thus doesn't require further proof. □

Similarly, we can talk about the size of spanning sets:

Lemma 3.1.4.11. *Let $X \subseteq \mathbb{R}^3$ be a finite set. If $\#X < 3$ then* $\text{span}\, X \subset \mathbb{R}^3$.

Proof

This has already been proven, in that you need at least three non-null non-coplanar vectors to span $\mathbb{R}^3$. □

Corollary 3.1.4.12. *Equivalently, if* $\text{span}\, X = \mathbb{R}^3$ *then $\#X \geq 3$.*

Proof

Once again, this is the same result, just stated in another manner, and so doesn't require further proof. □

Now we can combine these two lemmas into a very important lemma:

Lemma 3.1.4.13. *Let $X \subseteq \mathbb{R}^3$ be a finite set of linearly independent vectors such that* $\operatorname{span} X = \mathbb{R}^3$. *Then $\#X = 3$.*

Proof

Since X is linearly independent, $\#X \leq 3$. But since X spans $\mathbb{R}^3$, $\#X \geq 3$. Combining these, we see that $\#X = 3$, as we had claimed. □

With this we can now state a similar result to what we have already proved in $\mathbb{R}^2$.

Theorem 3.1.4.14. *Let $X \subseteq \mathbb{R}^3$ be a subset of $\mathbb{R}^3$. Then the following are equivalent:*

a) X is a base for $\mathbb{R}^3$;

b) X is linearly independent and has three elements;

c) X spans $\mathbb{R}^3$ and has three elements;

d) X is a linearly independent set which spans $\mathbb{R}^3$.

Proof

We already know that (d) implies both (b) and (c) by the preceding lemma. Let, once and for all, $X = \{x_1, x_2, \cdots, x_n\}$. Then:

- (a) ⇒ (c):

  Since X is a base, given any linear function f, the image of any point v under f is uniquely determined by the image of X under f.

  In particular, for the identity function, we see that for any $v \in \mathbb{R}^3$ there are uniquely determined $v_1, v_2, \cdots, v_n \in \mathbb{R}$ such that

  $$\mathrm{id}_{\mathbb{R}^3}(v) = \sum_{i=1}^{n} v_i \mathrm{id}_{\mathbb{R}^3}(x_i)$$

  which is just

  $$v = \sum_{i=1}^{n} v_i x_i$$

  and so $v \in \operatorname{span} X$.

  This tells us that any $v \in \mathbb{R}^3$ is spanned by X, so X spans $\mathbb{R}^3$.

  We claim that x_1, x_2, x_3 are non-coplanar (and hence $\operatorname{span} X = \operatorname{span}\{x_1, x_2, x_3\}$).

  Indeed, since X is a base, there's a unique way to write any vector in $\mathbb{R}^3$ using vectors in X - in particular, there's a unique way to write x_1, x_2 and x_3 in terms of X:

  $$x_1 = 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + \cdots + 0 \cdot x_n$$

$$x_2 = 0 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 + \cdots + 0 \cdot x_n$$
$$x_3 = 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 + \cdots + 0 \cdot x_n$$

this tells us that they're all non-parallel, and that no two of them span the other one - so they are non-coplanar. So $\operatorname{span} X = \operatorname{span}\{x_1, x_2, x_3\}$.

But notice that we can do the same for any $x_i \in X$:

$$x_i = 0 \cdot x_1 + \cdots + 1 \cdot x_i + \cdots + 0 \cdot x_n.$$

So if $n > 3$, we'd have two ways of writing any x_i in terms of X: Either as $x_i = 1 \cdot x_i$ or as being spanned by $\{x_1, x_2, x_3\}$. But X is a base, so there's only one way to write any vector in terms of X.

That tells us that X must have precisely 3 elements - otherwise there would be some elements which would be able to be written in multiple ways in terms of X, which contradicts the fact that X is a base.

This shows that a base is a spanning set with three elements.

- (b) $\iff$ (c):

  In this case, we're always assuming that X has three vectors. We need to show that X is, then, linearly independent if, and only if, it is a spanning set for $\mathbb{R}^3$.

  Take $v \in \mathbb{R}^3$, any vector. Then X is a spanning set if, and only if, $\operatorname{span} X = \mathbb{R}x_1 + \mathbb{R}x_2 + \mathbb{R}x_3$ is $\mathbb{R}^3$. But we have proven that this happens if, and only if, X is a set of non-null, non-coplanar vectors - which is the same as saying that X is linearly independent.

  So X is a spanning set with three elements if, and only if, it is a linearly independent set with three elements.

- (b) $\implies$ (d):

  If X is a spanning set with three elements, then we've already proven that it is also linearly independent. So X is a linearly independent set which spans $\mathbb{R}^3$.

- (d) $\implies$ (a):

  Let X be a linearly independent spanning set for $\mathbb{R}^3$. We've already proven that this set has precisely three elements - let them be $X = \{x_1, x_2, x_3\}$.

  Since X spans $\mathbb{R}^3$, given any $v \in \mathbb{R}^3$ there are (not necessarily unique) real numbers $v_1, v_2, v_3 \in \mathbb{R}$ such that

  $$v = \sum_{i=1}^{3} v_i x_i. \tag{1}$$

  Suppose that there were other numbers $v_1', v_2', v_3' \in \mathbb{R}$ such that

  $$v = \sum_{i=1}^{3} v_i' x_i. \tag{2}$$

We want to show that $v_i = v_i'$ for all $i \leq 3$ and so that there's a unique way to write each vector in $\mathbb{R}^3$ as being spanned by X.

But since (1) and (2) are just two expressions for v, we can write

$$\sum_{i=1}^{3} v_i x_i = v = \sum_{i=1}^{3} v_i' x_i.$$

But then we can rearrange this into

$$\sum_{i=1}^{3} (v_i - v_i') x_i = 0$$

and since X is linearly independent, this implies that $v_i - v_i' = 0$ for all $i \leq 3$. But this is just the same as saying $v_i = v_i'$ for all $i \leq 3$.

Therefore, there's a unique way to write each vector $v \in \mathbb{R}^3$ as being spanned by X.

Finally, let $f : \mathbb{R}^3 \to \mathbb{R}^3$ be any linear function, and let $v = \sum_{i=1}^{3} v_i x_i$ as before (but now we know that the v_is are uniquely determined). We can now write $f(v)$ as

$$f(v) = \sum_{i=1}^{3} v_i f(x_i),$$

by simply applying f on v and using the fact that f is linear. This shows that f is entirely determined by X, so X is a base.

This ends the proof: Indeed, we've proven

$$(a) \implies (b) \iff (c) \iff (d) \implies (a)$$

□