Pure Mathematics

Chapter 1

Foundations

1.0.1 The Axiom of Choice

1.0.1.1 Statement of the Problem - Intuitive

The problem that the Axiom of Choice addresses was famously described by Bertrand Russell using the example of choosing from an infinite number of pairs of shoes and an infinite number of pairs of socks. From pairs of shoes it is easy to see how you can have a strategy for making a choice but for pairs of identical socks it is not at all obvious how to achieve the same thing.

In fact, a set like a pair of socks cannot exist because the definition of a set of objects requires that the objects are distinguishable. As a result, just looking at Bertrand Russell's example, it is tempting to think that the problem wouldn't arise. However, we can instead think of an infinite collection of non-empty sets containing random (distinguishable) objects; in this case we still cannot develop a choice function because there is no single strategy that can be applied to make choices from all the sets. This is a more important point than it may at first seem because the problem of choosing a strategy for each of the infinite sets is a problem of making an arbitrary choice for an infinite number of sets - i.e. the very same problem that the Axiom of Choice is attempting to address!

1.0.1.2 Statement of the Problem - Formal

If a set B is non-empty then this can be expressed in first-order predicate logic as the truth of

$$\phi(B) := (\exists x)(x \in B).$$

Then, through the use of *existential instantiation* we can posit an element, say $c \in B$.

Definition. *In formal logic,* ***Existential Instantiation*** *is the inference,*

$$\exists x \; . \; P(x) \implies P(c)$$

where c is a new constant symbol. The symbol c must not have been previously used in the proof nor must it appear in the final conclusion.

So existential instantiation gives us a way, in formal logic, of "selecting" an arbitrary element from a single set without the need for a strategy for making a choice. Furthermore, we can concatenate atomic predicate clauses together to make an arbitrarily-long (but finite) compound predicate statement. For example,

$$\phi(B_1) \wedge \phi(B_2) \wedge \cdots \wedge \phi(B_n)$$

which enables us to say,

$$\exists(x_1, x_2, \ldots, x_n) \; . \; (x_1 \in B_1) \wedge (x_2 \in B_2) \wedge \cdots \wedge (x_n \in B_n)$$

and so instantiate an n-tuple $(c_1, c_2, \ldots, c_n)$ with an element c_i from each set B_i. In this way, formal mathematical logic allows us to make a "selection" from an arbitrary *finite* number of non-empty sets.

However, the first-order logic that is the standard for the logical formalization of mathematics is not able to model infinite domains as the Compactness Theorem implies that first-order logic cannot uniquely determine infinite sets. We cannot, therefore, use the same logical approach to formalize a choice from an infinite number of sets. For this reason the capability needs to be introduced as an axiom.

1.0.1.3 The Axiom of Choice

Definition. *Let A be a non-empty set of non-empty sets. A* ***choice function*** *on A is a function,*

$$f : A \longmapsto \bigcup_{a \in A} \; s.t. \; \forall a \in A, \; f(a) \in a.$$

Axiom 1. ***The Axiom of Choice:*** *If A is a non-empty set of non-empty sets then there exists a choice function for A.*

1.0.1.4 The Well-Ordering Theorem (a.k.a Zermelo's Theorem)

Definition. *A set is* ***well-ordered*** *by a strict total order if every non-empty subset has a minimal element under the ordering.*

Axiom 2. ***Well-Ordering Theorem:*** *Every set can be well-ordered.*

> *The* ***Well-Ordering Principle*** *is usually taken to be the proposition that the positive integers are well-ordered (which may be axiomatic or proven by induction depending on the method of constructing the natural numbers) but is sometimes used synonymously with the Well-Ordering Theorem.*

Although, for historic reasons, this is known as a *theorem*, it has been found to be unprovable from the axioms of mathematics and must be accepted as an axiom itself.

It has also been found to be equivalent to the Axiom of Choice. That's to say, every set can be well-ordered if every collection of sets has a choice function and every collection of sets has a choice function if their union is a well-ordered set.

A point of interest when proving the Axiom of Choice using the Well-Ordering Theorem: We assert the well-ordering on the union of the collection of sets rather than asserting a (potentially different) well-ordering on each of the sets - which would equally suffice as a choice function. The reason for this is that if we attempt to assert an individual well-ordering on each of the sets then we are again falling into the problem of existential instantiation over an infinite structure - which, itself, requires the Axiom of Choice. This is the same point as was mentioned in the problem statement.

For uncountable sets the well-ordering may be inexpressible. Specfically, in the case of the reals $\mathbb{R}$, it has been proven that any well-ordering of $\mathbb{R}$ must be inexpressible.

1.0.1.5 Zorn's Lemma

To understand Zorn's Lemma some concepts relating to partial orders are needed.

Partial Orders

Definition. *Let $(P, \leq)$ be a partial order and let $A \subseteq P$. An element $p \in P$ is an **upper bound** for A if $a \leq p$ for all $a \in A$.*

Definition. *Let $(P, \leq)$ be a partial order. An element $m \in P$ is a **maximal element** if there is no $p \neq m \in P$ such that $m \leq p$.*

Partial orders implicitly define totally ordered subsets, or *chains*, within which there may be a maximal element. So, there can be multiple maximal elements for each chain in the poset.

Example of a Partial Order

(1) Let $(\mathbb{N} \setminus \{1\}, \preceq)$ be the relation "divides by". More precisely, for $m, n \in \mathbb{N}$

$$n \preceq m \iff m|n.$$

Then, under this order, every prime number is a maximal element. If we were to modify the order slightly to include the element 1, the

partial order $(\mathbb{N}, \preceq)$ has a single, global maximal element - the number 1.

Axiom 3. ***Zorn's Lemma:*** *Let $(P, \leq)$ be a non-empty partial order such that every totally ordered subset has an upper bound. Then P has a maximal element.*

Proposition 1. *Zorn's Lemma implies the Axiom of Choice.*

Proof. Let A be a non-empty set of non-empty sets. Define a *partial choice function* over A to be a function that defines a choice from some of the sets in A but not others. Define an *extends* relation between functions such that if f and g are functions then g *extends* f if and only if $dom(f) \subseteq dom(g)$ and $g(x) = f(x)$ for all $x \in dom(f)$.
Now if we postulate the existence of a collection C of all the partial choice functions over A then the *extends* relation between the members of C is a partial order. Denote this partial order $(A, \leq)$. For any chain in C we can take the union of all the domains of the functions in the chain — which will actually be the domain of the final function in the chain, say f — and form a function g such that $g(x) = f(x)$ for all $x \in dom(f)$. The function f is an upper bound on the chain.
Therefore we can apply Zorn's lemma to assert the existence of a maximal element h. The function h, being maximal, cannot be extended by any function in C and must, therefore, have a domain equal to the entire set A. It follows then that h is a choice function over all the set A and satisfies the Axiom of Choice. □

Theorem 1. *Every vector space has a basis*

Proof. Let V be an arbitrary vector space and let S be the set of all linearly independent subsets of V. The inclusion relation $\subseteq$ is a partial order over the members of S. For every chain in this partial order $s_1 \subseteq s_2 \subseteq \cdots$ if we take the union of the sets in the chain $U \coloneqq s_1 \cup s_2 \cup \cdots$, then U is an upper bound of the chain and, since U is the union of sets of linearly independent vectors in S, $U \in S$ also. Therefore, Zorn's Lemma tells us that S has a

maximal element.
Let B be a maximal element of S. By membership of S, B is a linearly independent set of vectors in V. Since B is a maximal element in S it follows that $\forall s \in S,\ s \subseteq B$. Suppose that there is some vector v in V such that the set $B \cup \{\vec{\boldsymbol{v}}\}$ is linearly independent. Then the set $B \cup \{\vec{\boldsymbol{v}}\}$ is a linearly independent set of vectors in V and so is a member of S but is not a subset of B. This contradicts the maximality of B. We can therefore conclude that no such v exists.
Therefore, there exists a set B of linearly independent vectors in V that spans the space V and is the largest spanning set of linearly independent vectors that can be found in V. It is, therefore, a basis of V. □

> *Note that when looking for an upper bound to the chain in the set S we don't just take the last element of the chain because the chain can be infinite.*

references: http://www.math.toronto.edu/ivan/mat327/docs/notes/11-choice.pdf

For further study: https://www.mn.uio.no/math/tjenester/kunnskap/kompendier/acwozl.pdf.

1.1 Number Theory

1.1.1 Natural Numbers

1.1.1.1 Peano Axioms

Axiom 4. *Closure under addition:*
For all $a, b \in \mathbb{N}$ *we have* $a + b \in \mathbb{N}$.

Axiom 5. *Closure under multiplication:*
For all $a, b \in \mathbb{N}$ *we have* $a \times b \in \mathbb{N}$.

Axiom 6. *Commutative Law for addition:*
For all $a, b \in \mathbb{N}$ *we have* $a + b = b + a$.

Axiom 7. *Associative Law for addition:*
For all $a, b, c \in \mathbb{N}$ *we have* $(a + b) + c = a + (b + c)$.

Axiom 8. *Commutative Law for multiplication:*
For all $a, b \in \mathbb{N}$ *we have* $a \times b = b \times a$.

Axiom 9. *Associative Law for multiplication:*
For all $a, b, c \in \mathbb{N}$ *we have* $(a \times b) \times c = a \times (b \times c)$.

Axiom 10. *Multiplicative Identity:*
There is a special element of $\mathbb{N}$*, denoted by 1, which has the property that for all* $n \in \mathbb{N}$, $n \times 1 = n$.

Axiom 11. *Additive cancellation:*
For all $a, b, c \in \mathbb{N}$ *if* $a + c = b + c$ *then* $a = b$.

Axiom 12. *Multiplicative cancellation:*
For all $a, b, c \in \mathbb{N}$ *if* $a \times c = b \times c$ *then* $a = b$.

Axiom 13. *Distributive Law:*
For all $a, b, c \in \mathbb{N}$, $a \times (b + c) = (a \times b) + (b \times c)$.

Axiom 14. *Definition of "less than":*
For all $a, b \in \mathbb{N}$, $a < b$ *if and only if there is some* $c \in \mathbb{N}$ *s.t.* $a + c = b$.

Axiom 15. *Trichotomous property:*
For all $a, b \in \mathbb{N}$ exactly one of the following is true: $a = b$, $a < b$, $b < a$.

Notation. We also write ab for $a \times b$.

Proposition 2. *If $a, b \in \mathbb{N}$ satisfy $a \times b = a$, then $b = 1$.*

Proof.

$$\begin{aligned} & a \times b = a = a \times 1 && \text{by Multiplicative Identity axiom} \\ \iff \quad & b \times a = 1 \times a && \text{by Commutative Law for multiplication} \\ \iff \quad & b = 1 && \text{by Multiplicative cancellation} \end{aligned}$$

□

Proposition 3. *If $a, b, c \in \mathbb{N}$ and $a < b$ then $a \times c < b \times c$.*

Proof.

$$\begin{aligned} & a < b \implies a + d = b \text{ for some } d \in \mathbb{N} && \text{by Definition of "less than"} \\ \therefore \quad & b \times c = (a + d) \times c = (a \times c) + (d \times c) && \text{by Distributive Law} \\ \therefore \quad & a \times c < (a \times c) + (d \times c) = b \times c && \text{by defn. "less than" and closure} \end{aligned}$$

□

Proposition 4. *1 is the least element of $\mathbb{N}$.*

Proof. Assume m is the least element of $\mathbb{N}$. Then, also $m < 1$. So, by Proposition 3,

$$m < 1 \implies m \times m < 1 \times m = m$$

But, closure of multiplication and $m \times m < m$ together contradict the assumption that m is the least element of $\mathbb{N}$.
Therefore m cannot be less than 1. Since we know that $1 \in \mathbb{N}$ and that the minimum element of $\mathbb{N}$, m, cannot be less than 1, it follows that 1 must be the minimum element of $\mathbb{N}$ and $m = 1$. □

1.1.2 Integers

TODO: construction of the integers from peano natural numbers

1.1.2.1 Odd and Even Numbers

Definition. *An **even** number, $n \in \mathbb{Z}$, is one that satisfies,*

$$\exists\, m \in \mathbb{Z} \;\cdot\; n = 2m$$

Definition. *An **odd** number, $n \in \mathbb{Z}$, is one that satisfies,*

$$\exists\, m \in \mathbb{Z} \;\cdot\; n = 2m + 1$$

1.1.2.2 Consequences

Sum of even numbers, $m + n$:

$$\begin{aligned} m + n &= 2k + 2l \quad \text{where } k, l \in \mathbb{Z} && \text{by defn. of even no.s } m, n \\ &= 2(k + l) \\ &= 2q \quad \text{where } q \in \mathbb{Z} \end{aligned}$$

So $m + n$ is also even. However, if $m + n$ is even:

$$\begin{aligned} m + n &= 2k \quad \text{where } k \in \mathbb{Z} && \text{by defn. of even } m + n \\ k &= \frac{m}{2} + \frac{n}{2} \end{aligned}$$

So m and n are not necessarily even. A counterexample is

$$3 + 5 = 8 \iff \frac{3}{2} + \frac{5}{2} = 4$$

To summarize:

- m, n even $\implies$ $m + n$ even
- $m + n$ even $\implies$ m, n even Wrong!

1.1.2.3 The Fundamental Theorem of Arithmetic

Definition of prime number: An integer that is only divided cleanly by itself and one. More formally, an integer, p, is prime if it is greater than 1 and,

$$\exists! \, m, n \in \mathbb{Z} \; \cdot \; \frac{p}{m} = n \wedge (m \neq p \wedge m \neq 1)$$

Primality $\implies$ Unique Prime Factorization:

> "Any number either is prime or is measured by some prime number."
> *Euclid, Elements Book VII, Proposition 32*

So, if an integer n is not prime then,

$$\begin{aligned} \exists \, a, b \in \mathbb{Z} \; \cdot \; \frac{n}{a} &= b \\ \iff n &= ab \end{aligned}$$

Then, for a (the same applies to b),

$$\begin{aligned} \exists \, c, d \in \mathbb{Z} \, , \, c, d \notin \{1, a\} \; \cdot \; \frac{n}{a} &= b \\ \iff n &= cd \end{aligned}$$

We can continue to descend like this until we must eventually encounter one or more primes. Furthermore, if a number, n, has a prime factorization, $p_1 p_2$ then,

$$n = p_1 p_2 = p_3 p_4 \iff \frac{p_1}{p_3} = \frac{p_4}{p_2} = n$$

But $\frac{p_1}{p_3} = n$ contradicts the definition of primeness of p_1. Therefore prime factorizations are unique.

Proof of existence

Proof. It must be shown that every integer greater than 1 is either prime or a product of primes. First, 2 is prime. Then, by strong induction, assume this is true for all numbers greater than 1 and less than n. If n is prime, there is nothing more to prove. Otherwise, there are integers a, b where $n = ab$, and $1 < a \leq b < n$. By the induction hypothesis, $a = p_1 p_2 ... p_j$ and $b = q_1 q_2 ... q_k$ are products of primes. But then $n = ab = p_1 p_2 ... p_j q_1 q_2 ... q_k$ is a product of primes. □

Proof of uniqueness

Proof. Suppose, to the contrary, that there is an integer that has two distinct prime factorizations. Let n be the least such integer and write $n = p_1p_2...p_j = q_1q_2...q_k$, where each p_i and q_i is prime. (Note that j and k are both at least 2.) We see that p_1 divides $q_1q_2...q_k$, so p_1 divides some q_i by Euclid's lemma. Without loss of generality, say that p_1 divides q_1. Since p_1 and q_1 are both prime, it follows that $p_1 = q_1$. Returning to our factorizations of n, we may cancel these two terms to conclude that $p_2...p_j = q_2...q_k$. We now have two distinct prime factorizations of some integer strictly smaller than n, which contradicts the minimality of n. □

1.1.2.4 Modular Arithmetic

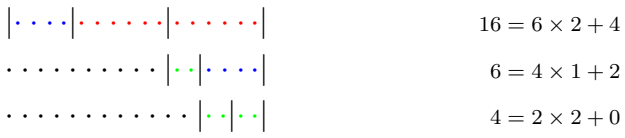
Greatest Common Divisor *(also called Highest Common Factor)*

Definition. *The* ***Greatest Common Divisor (gcd)*** *of two integers – say a and b – is an integer d that satisfies,*

$$d = z_1 a + z_2 b$$

for some $z_1, z_2 \in \mathbb{Z}$. This means that, whenever we are adding or subtracting multiples of the two numbers a and b, the result will always be a multiple of d and, therefore also, d is the smallest such result obtainable.

The greatest common divisor of 16 and 6 can be visualized as follows:



This implies the algorithm:

gcd(a, b) :
  if $b == 0$ then
    return a
  else
    return **gcd**$(b, a \bmod b)$
  end if

The greatest common divisor of a and b is the smallest difference of multiples of a and b. This is because – for any difference, d, of multiples of a and b – we have,

$$d = ma + nb \text{ for } m, n \in \mathbb{Z}$$

and, if $g = gcd(a, b)$ then, by definition, g divides both a and b and, therefore, also divides d. So any such sum (or difference) of integer multiples of a and b is a multiple of g.

Proposition 5. *For non-zero integers a and b, if $a = bq + r$ where $q, r \in \mathbb{Z}$, then $gcd(a, b) = gcd(b, r) = gcd(b, a \bmod b)$.*

Proof. $(a \bmod b) = r = a - bq$. For any m s.t. $m \mid a$ and $m \mid b$ we must also have $m \mid (a - bq)$ so the set of divisors of a and b is a subset of the set of divisors of b and $r = (a \bmod b)$. Conversely, for any m s.t. $m \mid b$ and $m \mid r = (a \bmod b)$ we have that $m \mid (bq + r) = a$ so the set of divisors of b and $r = (a \bmod b)$ is a subset of the set of divisors of a and b. So the sets are equal proving that they must have the same maximum element - the greatest common divisor. □

Proposition 6. *If $d = gcd(a, b)$ then there is no integer linear combination of a and b that equals any positive value less than d.*

Proof. Assume $d = gcd(a, b)$ and that $\exists\, e < d \in \mathbb{N}, m, n \in \mathbb{Z}$ s.t. $e = am + bn$. Then,

$$\begin{aligned} & e = am + bn = dz_1 m + dz_2 n = d(z_1 m + z_2 n) \quad \text{for } z_1, z_2 \in \mathbb{Z} \\ \iff \quad & z_1 m + z_2 n = \frac{e}{d} \notin \mathbb{Z} \end{aligned}$$

where we know that $\frac{e}{d} \notin \mathbb{Z}$ because $e < d$. But the field properties of the integers ensures that the integers are closed under integer linear combinations so that $z_1 m + z_2 n \in \mathbb{Z}$. Therefore such an e does not exist. □

Corollary 1. *If $d = gcd(a, b)$ then every integer linear combination of a and b is a multiple of d.*

Proof. Proposition 6 showed that there is no integer linear combination of a and b less than d. Suppose that we have,

$$e > d \in \mathbb{N}, m, n \in \mathbb{Z} \text{ s.t. } e = am + bn$$

then, because d divides both a and b,

$$e = adz_1 + bdz_2 = d(az_1 + bz_2), \quad z_1, z_2 \in \mathbb{Z}.$$

Now the closure of the integer field means that $z_3 = az_1 + bz_2 \in \mathbb{Z}$ so that,

$$e = z_3 d \implies d \mid e.$$

□

Proposition 7. *A number $x \in \mathbb{Z}_m$ has a multiplicative inverse if and only if $gcd(x, m) = 1$.*

Proof. Assume x^{-1} is a multiplicative inverse for $x \in \mathbb{Z}_m$. Then,

$$x^{-1}x = 1 \iff x^{-1}x \equiv 1 \pmod{m} \iff x^{-1}x = am + 1, \quad a \in \mathbb{Z}.$$

This means that we must have $1 = am + bx$ for some $a, b \in \mathbb{Z}$. Now if we have $d = gcd(x, m)$ then by Corollary 1 we must have $d \mid 1$. Therefore $d = 1$.

Clearly, also, if we have $gcd(x, m) = 1$ then we also have $1 = am + bx$ for some $a, b \in \mathbb{Z}$ and by following the previous logic in reverse we obtain that $b = x^{-1}$ is the multiplicative inverse of $x \in \mathbb{Z}_m$. □

Lowest Common Multiple

The lowest common multiple of two numbers is formed by the multiplication of all the prime factors that occur in the two numbers where repititions of prime factors are important. That's to say, the lowest common multiple of 4 and 8 is not 2 (which is the highest common factor/greatest common divisor) but 8 because in 8, the factor 2 occurs three times (as 2^3) and it occurs twice in 4,

$$lcm(4, 8) = lcm(2 \times 2, 2 \times 2 \times 2) = 2 \times 2 \times 2.$$

The general formula for the lowest common multiple may be expressed in terms of the gcd as follows

$$d = gcd(a, b) \implies lcm(a, b) = d \times (a/d) \times (b/d).$$

1.1.2.5 Some Proofs on the Integers

Proposition 8. *For any integer m, $\sqrt{m}$ is rational iff m is a square, i.e. $m = a^2$ for some integer a.*

To begin with we show the easier direction of implication: $(m = a^2) \implies (\sqrt{m}$ is rational).

Proof. Assume $m, a, b \in \mathbb{Z}$.

$$\begin{aligned} & m = a^2 \\ \iff & \sqrt{m} = |a| \\ & = a/b \text{ for } b = 1 \text{ or } -1. \end{aligned}$$

□

Now the other (harder) direction, $(\sqrt{m}$ is rational$) \implies (m = a^2)$.

Proof. Assume $m, a, b \in \mathbb{Z}$. $(\sqrt{m}$ is rational) can be formalized as:

$$\exists\, m, a, b \in \mathbb{Z} \cdot (\sqrt{m} = \frac{a}{b}) \ \wedge \ (a \text{ and } b \text{ are coprime})$$

$$\begin{aligned} \sqrt{m} &= \frac{a}{b} \\ \implies m &= \frac{a^2}{b^2} \\ \iff mb^2 &= a^2 \end{aligned}$$

But a and b are coprime so they don't share any prime factors. This means that a^2 and b^2 also don't share any prime factors. So, if $|b| > 1$, the prime factorization of mb^2 is necessarily different from that of a^2 meaning that $mb^2 \neq a^2$ contradicting the hypothesis of coprimality. On the other hand, if $|b| = 1$, then b has no prime factors (its prime factorization is empty) and so mb^2 has the same prime factorization as m which may be equal to that of a^2 in the case that $m = a^2$. □

Proposition 9. *For all nonnegative integers $a > b$ the difference of squares $a^2 - b^2$ does not give a remainder of 2 when divided by 4.*

Beginner's attempt - try proof by contradiction:

$$\begin{aligned} a^2 - b^2 &= 4n + 2 \\ 2k &= 4n + 2 && \text{by } a^2 - b^2 \text{ even} \\ k &= 2n + 1 \implies && k \text{ is some odd number.} \end{aligned}$$

So, proof by contradiction is our first instinct but doesn't seem to get us anywhere. Instead, proceed by cases:

Case a, b are even:

$$\begin{aligned} &\exists\, k, l \in \mathbb{Z} \;\cdot\; a = 2k, b = 2l \\ &\implies a^2 - b^2 = 4k^2 - 4l^2 \\ &\qquad\quad = 4\left(k^2 - l^2\right) \\ &\qquad\quad = 4m \;\text{ where }\; m \in \mathbb{Z} \end{aligned}$$

So 4 divides $a^2 - b^2$ with 0 remainder.

Case a, b are odd:

$$\begin{aligned} &\exists\, k, l \in \mathbb{Z} \;\cdot\; a = 2k + 1, b = 2l + 1 \\ &\implies a^2 - b^2 = \left(4k^2 + 4k + 1\right) - \left(4l^2 + 4l + 1\right) \\ &\qquad\quad = 4\left(k^2 + k - l^2 - l\right) \\ &\qquad\quad = 4m \;\text{ where }\; m \in \mathbb{Z} \end{aligned}$$

So, again, 4 divides $a^2 - b^2$ with 0 remainder.

Case a even, b odd:

$$\begin{aligned} &\exists\, k, l \in \mathbb{Z} \;\cdot\; a = 2k, b = 2l + 1 \\ &\implies a^2 - b^2 = 4k^2 - \left(4l^2 + 4l + 1\right) \end{aligned}$$

$$= 4\left(k^2 - l^2 - l\right) - 1$$
$$= 4m + 3 \quad \text{where} \quad m = k^2 - l^2 - l - 1 \in \mathbb{Z}$$

So, here, 4 divides $a^2 - b^2$ with 3 remainder. So the proposition is proven as we have proven all the possible cases.
TODO: There is also another approach given in the Cambridge University Discrete Mathematics lecture notes

1.1.3 Absolute Value

Definition. *The* ***absolute value*** *function is defined,*

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

Proposition 10. $|a|\,|b| = |ab|$.

Proof. By the definition of absolute value,

$$|ab| = \begin{cases} ab & ab \geq 0 \\ -ab & ab < 0. \end{cases}$$

Extending the definition to the product of absolute values,

$$|a|\,|b| = \begin{cases} ab & a, b \geq 0 \\ -ab & a < 0, b \geq 0 \\ -ab & a \geq 0, b < 0 \\ ab & a, b < 0. \end{cases}$$

We can see that these are equivalent because,

$$ab \text{ is } \begin{cases} \geq 0 & a, b \geq 0 \text{ or } a, b < 0 \\ < 0 & a < 0, b \geq 0 \text{ or } a \geq 0, b < 0. \end{cases}$$

□

1.1.3.1 The Triangle Inequality

$$|x| \geq x, |y| \geq y \implies |x| + |y| \geq x + y$$

$$|x+y| = \begin{cases} |x| + |y| & x, y \geq 0 \\ |-|x| + |y|| & x < 0, y \geq 0 \\ ||x| - |y|| & x \geq 0, y < 0 \\ |-(|x| + |y|)| & x, y < 0 \end{cases} \iff \begin{cases} ||x| + |y|| & x, y \geq 0 \text{ or } x, y < 0) \\ ||x| - |y|| & x < 0, y \geq 0 \text{ or } x \geq 0, y < 0 \end{cases}$$

Clearly, $||x| + |y|| \geq ||x| - |y||$ so that,

$$|x + y| \leq ||x| + |y|| = |x| + |y|$$

and this is known as the "triangle inequality".

Proposition 11. $|x - y| \leq |x - z| + |y - z|$

Proof.

$$|x - y| = |(x - z) + (z - y)| \leq |x - z| + |z - y| = |x - z| + |y - z|$$

□

Proposition 12. $|x - y| \geq ||x| - |y||$

Proof. Need to show $-|x - y| \leq |x| - |y| \leq |x - y|$. So, prove as two separate inequalities:

$$\begin{aligned} & |y| = |x + (y - x)| \leq |x| + |y - x| \\ \iff \quad & -|y - x| = -|x - y| \leq |x| - |y| \end{aligned}$$

$$\begin{aligned} & |x| = |(x - y) + y| \leq |x - y| + |y| \\ \iff \quad & |x| - |y| \leq |x - y| \end{aligned}$$

□

1.1.4 Complex Number

Definition. *The* ***modulus*** *of a complex number,* $z = a + bi$*, is the quantity defined as,*

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

TODO: modulus is also called "absolute value" and can be calculated as product with conjugate The modulus obeys the following properties:

- $|z_1 z_2| = |z_1|\,|z_2|$

Chapter 2

Algebra

2.1 Group Theory

2.1.1 Groups

Definition. *A binary operation is a function,*

$$f : G \times G \longmapsto G$$

which - by the definition of a function - maps a unique tuple from $G \times G$ to a unique value in the codomain G.

Definition. *Let G be a set and $*$ a binary operation on G and denote this $(G, *)$. Then $(G, *)$ is a* ***group*** *if:*

closure $\forall x, y \in G, x * y \in G$;

associativity $\forall x, y, z \in G, (x * y) * z = x * (y * z)$;

identity $\exists e \in G$ *s.t.* $\forall x \in G, e * x = x * e = x$;

inverse $\forall x \in G, \exists x^{-1} \in G$ *s.t.* $x * x^{-1} = x^{-1} * x = e$.

These are known as the ***group axioms****.*

Definition. *The group is an* ***Abelian group*** *if it has the additional property:*

commutativity $\forall x, y \in G, x * y = y * x \in G$.

Notation. from here on we will use juxtaposition notation for the group operation (so $xy = x * y$) and (usually) 1 for the identity element instead of e. This is known as *multiplicative notation.*

Theorem 2. *Suppose an associative law of composition is given on a set S. Then there is a unique way to define a product of n elements $a_1, \ldots, a_n$ for any $n \in \mathbb{N}$.*

Proof. Denote the product of n elements as $[a_1 \ldots a_n]$. We show that a product can be defined with the following properties:

(i) $[a_1] = a_1$;

(ii) $[a_1 a_2] = a_1 * a_2$ is defined by the law of composition;

(iii) for any integer i such that $1 \leq i \leq n$, $[a_1 \ldots a_n] = [a_1 \ldots a_i][a_{i+1} \ldots a_n]$.

Following a proof by induction, firstly note that the product is defined for $n \leq 2$ by (i) and (ii) and that (ii) also satisfies the requirement (iii). Then, assume that the product is defined for $n \leq 2$ and that this product is the unique product satisfying (iii).
Then the induction step is to show that,

$$[a_1 \dots a_n] = [a_1 \dots a_{n-1}][a_n]$$

TODO: complete this from Artin[56] □

2.1.1.1 Corollaries of the group axioms

The group operation is defined to map a unique tuple in $G \times G$ to a unique value in G so that if we have $x, y \in G$ then $f((x, y)) = f(x, y) = xy \in G$ and for $a, b, c \in G$,

$$a = b \iff (c, a) = (c, b) \implies f((c, a)) = f((c, b)) \iff ca = cb$$

$$\therefore a = b \implies ca = cb$$

Then, using all the group axioms - associativity, inverse and identity,

$$ca = cb \implies c^{-1}(ca) = c^{-1}(cb) \iff (c^{-1}c)a = (c^{-1}c)b \iff 1a = 1b \iff a = b$$

Therefore we have the principle of cancellation,

$$ca = cb \implies a = b$$

Note that, since we have used the axioms of inverse and identity and the definitions of these require these elements to exhibit these properties from both the left and the right, the principle of cancellation can also be shown from both the left and the right. So, also,

$$ac = bc \implies a = b$$

There are (at least) two approaches to finding the other consequences of the group axioms.

First approach. We begin by noticing that the law of cancellation implies that,

unique identity and inverses $\forall a, x, b \in G, ax = b$ has a unique solution because,

$$ax = ax' \iff x = x'$$

That unique solution is $a^{-1}b$. If $b = a$ we have $ax = a$ and x, by identity axiom, is an identity element. Since, the solution to this equation - x - is unique, it follows that there is a unique value that is the identity element. Then, if we let b be this unique identity element we have $ax = 1$ and the unique solution, x, is the inverse of a, i.e. a^{-1}. Therefore, the inverses of group elements are also unique.

Second approach. This approach begins by showing the uniqueness of the identity element solely using the defintion of the identity. Here, for clarity, we revert to using e to denote the identity element.

unique identity Assume there are two identity elements, e, e'. Then, by the definition of the identity $ee' = e'e = e = e'$ so that there is a single value that has the property of the identity element.

Then, using the definition of the inverse we have,

unique inverses Assume there are two distinct inverses of an element a: a^{-1} and a'. Then,

$$\begin{aligned} & aa^{-1} = 1 = aa' && \text{defn. of inverse, uniqueness of identity} \\ \iff \quad & a^{-1} = a' && \text{law of cancellation} \end{aligned}$$

Some Examples of Groups

- $(\mathbb{R} \setminus \{0\}, \times)$ is a group whereas $(\mathbb{R}, \times)$ is not a group because 0 has no multiplicative inverse.

- $(\mathbb{R}, +)$ is a group.
- The set of $n \times n$ invertible matrices is called the General Linear group and denoted GL_n
- Let G denote the set of matrices

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \middle| \, a, b \in \mathbb{Z}_7, \; a \neq 0 \right\}.$$

Then G is a group with respect to matrix multiplication (where all additions and multiplications are carried out in $\mathbb{Z}_7$). This is because closure can be shown by,

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a' & b' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} aa' & ab' + b \\ 0 & 1 \end{pmatrix} \in G$$

where the result is in G because $aa' \neq 0$. Next we need to show that we have inverses. So we need to show existence in G of matrices such that,

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a' & b' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \iff \begin{pmatrix} aa' & ab' + b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which implies that $aa' = 1$ and $ab' + b = 0$. Because we are in $\mathbb{Z}_7$ every non-zero element has a multiplicative inverse so we have,

$$a' = a^{-1} \quad \text{and} \quad b' = -a^{-1}b$$

so that,

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} a^{-1} & -a^{-1}b \\ 0 & 1 \end{pmatrix}.$$

2.1.1.2 Permutations and Symmetric Groups

Definition. *A **permutation** is a bijection from a set to itself. Since permutations are bijective, they are invertible and since they are functions, function composition defines an associative law of composition over them. As a result, they form a group.*

Definition. *The **symmetric group** defined over a set is the group whose elements are the permutations of the objects of the set and whose law of composition is the composition of functions. The name probably comes from the study of symmetries of geometric objects that were eventually realised to be equivalent to permutations of the vertices.*

Definition. *A **generating set of a group** is a subset such that every element of the group can be expressed as a combination (under the group operation) of finitely many elements of the subset and their inverses.*

Notation. The symmetric group over the integers from 1 to n is denoted S_n. The symmetric group over a set G may be denoted $Sym(G)$.

S_2 The symmetric group S_2 consists of the two elements i and τ which are, respectively, the identity map and the transposition which interchanges 1 and 2. The group composition law is described by the fact that the identity map is the identity of the composition and by the relation $\tau\tau = \tau^2 = i$. Which results in the multiplication table:

$$\begin{aligned} i \cdot i &= i \\ i \cdot \tau &= \tau \\ \tau \cdot i &= \tau \\ \tau \cdot \tau &= i \end{aligned}$$

Note that the law of composition is commutative.

S_3 The symmetric group S_3 contains 3! elements. It is the smallest group whose law of composition is not commutative. It can be described using any two permutations of $\{1, 2, 3\}$. For example, if we take,

$$x = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \; y = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then the permutations are,

$$\{1, x, x^2, y, xy, x^2y\} = \{\, x^i y^j \mid 0 \leq i \leq 2,\ 0 \leq j \leq 1 \,\}$$

These are the elements of the group. The composition law over these elements is the function composition of these permutation functions and its multiplication table is characterized by the rules:

$$x^3 = 1,\ y^2 = 1,\ yx = x^2y$$

These are derived directly from the permutations themselves. Note that this composition law is not associative as $yx \neq xy$.
Any product of the elements x, y and of their inverses can be brought into the form $x^i y^j$ with i, j taking the ranges given above by repeated application of the above rules. To do so, we move all occurrences of y to the right side using the last relation and bring the exponents into the indicated ranges using the first two relations:

$$\begin{aligned} x^{-1}y^3x^2y = x^2yx^2y = x^2(yx)xy = x^2(x^2y)xy = x^4(yx)y \\ = x^4(x^2y)y = x^6y^2 = (x^3)^2y^2 = 1 \cdot 1 = 1 \end{aligned}$$

Rules like these that determine a complete multiplication table are called *defining relations* for the group.

2.1.2 Subgroups

Definition. *A subset H of a group G is called a* ***subgroup*** *if it has the following properties:*

- ***Closure:*** *If $a \in H$ and $b \in H$ then $ab \in H$.*
- ***Identity:*** $1 \in H$.
- ***Inverses:*** *If $a \in H$ then $a^{-1} \in H$.*

These conditions show that the subset H is a group with respect to the induced law of composition created by applying the law of composition of G on the members of H. Note that the associative property is not mentioned because the associativity of the composition of members of G automatically carries over to H.

Notation. If H and G are groups then we may write $H \leq G$ to indicate that H is a subgroup of G.

Note that an alternative, more compact, formulation of the definition of a subgroup is as follows.

Let G be a group and $\emptyset \neq H \subseteq G$. Then H is a subgroup if

$$\boldsymbol{x, y \in H \implies x^{-1}y \in H}.$$

This is because,

$$[(x, y \in H \implies xy \in H) \wedge (x \in H \implies x^{-1} \in H)] \iff (x, y \in H \implies x^{-1}y \in H).$$

The implication,

$$[(x, y \in H \implies xy \in H) \wedge (x \in H \implies x^{-1} \in H)] \implies (x, y \in H \implies x^{-1}y \in H)$$

is obvious. In the other direction,

$$(x, y \in H \implies x^{-1}y \in H) \implies [(x, y \in H \implies xy \in H) \wedge (x \in H \implies x^{-1} \in H)]$$

is because, if we set $x = y$*,*

$$x^{-1}x = e \in H \implies x^{-1}e = x^{-1} \in H$$

and then, for $x \neq y$*,*

$$x^{-1}, y \in H \implies xy \in H.$$

Every group, at a minimum, has two trivial subgroups: the maximal subgroup - the group itself; and the minimal subgroup - the set containing just the identity. A subgroup that is neither of these is known as a *proper subgroup*.

Proposition 13. *Suppose* H *and* K *are subgroups of* G *such that neither* $H \subseteq K$ *nor* $K \subseteq H$*. Then* $H \cup K$ *is not a subgroup of* G*.*

Note that it might be possible, for this proof, to just show that we have no reason to think that $H \cup K$ *is a subgroup (i.e. the definitions don't require it) so, in general, it is not. But, actually, we can show something much stronger: that* $H \cup K$ ***cannot*** *be a subgroup. If we can easily show something stronger then in most cases it's going to add clarity.*

Proof. Since neither $H \subseteq K$ nor $K \subseteq H$ we can conclude that $H \setminus K$ and $K \setminus H$ are both non-empty. So, select an element from each,

$$h \in H \setminus K,\ k \in K \setminus H.$$

Then we have $h, k \in H \cup K$ and if $H \cup K$ were a group then the closure property of the group would require that

$$hk \in H \cup K.$$

Assume $hk \in H \cup K$. Then, $hk \in H$ or $hk \in K$. If $hk \in H$ then the group properties of H require that

$$h^{-1}hk = k \in H$$

which contradicts the selection of k. We have a similar situation if $hk \in K$. Therefore, $hk \notin H \cup K$. □

2.1.2.1 Additive Groups of Integers

Important examples are the subgroups of the additive group of integers $\mathbb{Z}^+$. Denote the subset of $\mathbb{Z}^+$ consisting of all multiples of a given integer b by $b\mathbb{Z}$ such that,

$$b\mathbb{Z} = \{\, n \in \mathbb{Z} \mid n = bk,\ k \in \mathbb{Z} \,\}$$

Proposition 14. *For any integer b, the subset $b\mathbb{Z}$ is a subgroup of $\mathbb{Z}^+$ and every subgroup of $\mathbb{Z}^+$ is of the form $b\mathbb{Z}$ for some integer b.*

Proof. $b\mathbb{Z}$ is a subgroup of $\mathbb{Z}^+$ because,

- $b(0) = 0 \in b\mathbb{Z}$;
- If $a_1, a_2 \in b\mathbb{Z}$ then $a_1 = bk_1, a_2 = bk_2$ for $k_1, k_2 \in \mathbb{Z}$ and so $a_1 + a_2 = bk_1 + bk_2 = b(k_1 + k_2) \in b\mathbb{Z}$
- For any $a = bk \in b\mathbb{Z}$, $-a = b(-k) \in b\mathbb{Z}$

Now we need to prove that any subgroup of $\mathbb{Z}^+$ is $b\mathbb{Z}$ for some b. Let H be an arbitrary subgroup of $\mathbb{Z}^+$. Then by subgroup properties,

- $0 \in H$;
- If $a_1, a_2 \in H$ then $a_1 + a_2 \in H$
- For any $a \in H$, $-a \in H$

We proceed to show that there is always some integer b such that $H = b\mathbb{Z}$. Firstly, if H is the minimal subgroup $\{0\}$ then H trivially conforms to $b\mathbb{Z}$ with $b = 0$.
Otherwise, $\exists a \in H$ s.t. $a \neq 0$ then also $\exists -a \in H$ s.t. $-a \neq 0$. One of these must be a positive non-zero integer so there is at least one such member of H. We take b to be the smallest positive non-zero integer in H. Then,

$b\mathbb{Z} \in H$

- $b \in H$ (by selection) so by subgroup properties $b+b \in H$ and $(b+b)+b \in H$ and $b + \cdots + b \in H$
- By subgroup properties $b \in H \implies -b \in H$

So, $\{\, bk \in \mathbb{Z} \mid k \in \mathbb{Z} \,\}$ is in H.

$H \in b\mathbb{Z}$ Take any $n \in H$. Using division with remainder and dividing by b we get,

$$n = bq + r \quad q \in \mathbb{Z},\, 0 \leq r < b$$

But, since $b\mathbb{Z} \in H$ this means that $bq \in H$ and so $-bq \in H$. Therefore $n - bq = r \in H$. But $0 \leq r < b$ and, by assumption, b is the smallest positive *non-zero* integer in H and so, $r = 0$. So, every $n \in H$ divides by b. □

2.1.2.2 Greatest Common Divisor

If we extend this to groups which are generated by two integers a, b, then we have a subgroup of $\mathbb{Z}^+$,

$$a\mathbb{Z} + b\mathbb{Z} = \{\, n \in \mathbb{Z} \mid n = ar + bs \quad r, s \in \mathbb{Z} \,\}$$

This is known as the subgroup *generated* by a, b because it is the smallest subgroup which contains a and b. Proposition 14 tells us that it has the form $d\mathbb{Z}$ for some integer d.

Corollary 2. *If d is the positive integer which generates the subgroup $a\mathbb{Z}+b\mathbb{Z}$ then d is the greatest common divisor of a and b and so,*

- *d can be written in the form $d = ar + bs$ for some integers r and s.*
- *d divides a and b.*
- *If an integer e divides a and b, it also divides d.*

Proof. The first property follows directly from the definition of the subgroup. The second property is a result of the fact that a, b are in the subgroup $a\mathbb{Z} + b\mathbb{Z}$ so that $d\mathbb{Z} = a$ and $d\mathbb{Z} = b$. The third property is evident because $d = ar + bs = ek_1r + ek_2s = e(k_1r + k_2s)$. □

2.1.2.3 Deductions about Subgroups

Proposition 15. *Suppose $G = \{g_1, g_2, \ldots, g_n\}$ is a finite group of order n and that $x \in G$. Then $\{xg_1, xg_2, \ldots, xg_n\} = G$.*

Proof. Let $X = \{xg_1, xg_2, \ldots, xg_n\}$. By closure in G, every element of X must be in G and by the inverses property in G every element of X is distinct. So, there are n distinct elements of X, each of which are members of G. Since the order of G is n we can conclude that $X = G$. □

Proposition 16. *Suppose that G is a finite group and that H is a non-empty subset of G such that $x, y \in H \implies xy \in H$. Then H is a subgroup.*

Proof. H is non-empty so it contains at least one element, say x. H is closed under the group operation so it must also contain $x^2, x^3, \ldots$. But G is finite so the order of x in G must be finite also and so $\exists n \in \mathbb{N}$ s.t. $x^n = e$. But also $x^n = e \iff x^{n-1} = x^{-1}$. Therefore, for every element in H, the inverse of the element is also in H. □

Proposition 17. *Suppose that p is a prime number and we have integers $1 \leq x, g, h < p$ such that $xg \equiv xh \pmod{p}$. Then $g = h$ and $x \in \mathbb{Z}_p^*$ has an inverse.*

Proof. If $xg \equiv xh \pmod{p}$ then the difference between xg and xh is a multiple of p. But Euclid's Lemma (https://en.wikipedia.org/wiki/Euclid's_lemma) tells us that, because p is prime,

$$p \mid (xg - xh) = x(g - h) \implies (p \mid x) \wedge (p \mid (g - h)).$$

But since we have $x, (g - h) < p$ it is impossible for p to divide either of them unless they are 0. Only $g - h$ can be 0. Therefore,

$$g - h = 0 \iff g = h.$$

So, if we define a function $f : \mathbb{Z}_p^* \longmapsto \mathbb{Z}_p^*$ such that $f(a) = xa$ for some fixed $x \in \mathbb{Z}_p^*$ then f is injective because

$$f(a) = f(b) \iff xa \equiv xb \pmod{p} \iff a \equiv b \pmod{p}.$$

This means that f maps the $p - 1$ different values of $\mathbb{Z}_p^*$ to $p - 1$ different values in $\mathbb{Z}_p^*$. Therefore f is a bijection and there exists an inverse function f^{-1} such that $f^{-1}(xa) = a$. □

2.1.2.4 Examples of Subgroups

(2) $(\mathbb{Z}_p^*, \otimes)$ for prime p is a subgroup of the integers. This can be seen as closure of modular multiplication is clear and the existence of inverses has been shown in Proposition 17.
This is not the case however, for non-prime p. For example $(\mathbb{Z}_6^*, \otimes)$ is not a subgroup as it does not have inverses. We can see this by looking at the values generated by selecting a non-identity element and multiplying it by all the elements in $\mathbb{Z}_6^*$:

$$2 \otimes 1 = \mathbf{2},\ 2 \otimes 2 = \mathbf{4},\ 2 \otimes 3 = \mathbf{0},\ 2 \otimes 4 = \mathbf{2},\ 2 \otimes 5 = \mathbf{4}.$$

As can be seen, it doesn't generate all the values of $\mathbb{Z}_6^*$ but repeats a subset of them. Compare with the same for $\mathbb{Z}_7^*$:

$$2 \otimes 1 = \mathbf{2},\ 2 \otimes 2 = \mathbf{4},\ 2 \otimes 3 = \mathbf{6},\ 2 \otimes 4 = \mathbf{1},\ 2 \otimes 5 = \mathbf{3},\ 2 \otimes 6 = \mathbf{5}.$$

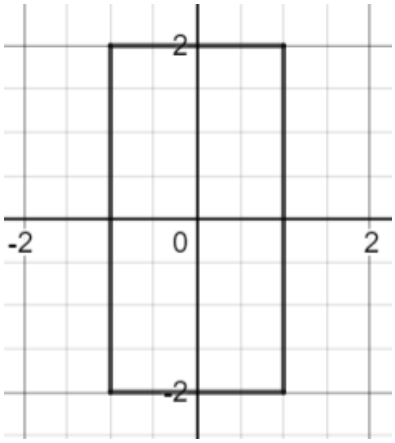
In the case of prime p all the values are generated so that multplication by other elements is a bijective function with a corresponding inverse.

(3) Let R be a non-square rectangle in $\mathbb{R}^2$ with corners having coordinates $(-1,-1), (-1,2), (1,2), (1,-1)$. Then there are four symmetries i, a, b, c of R, as follows:

- i is the identity
- a is reflection in the x-axis
- b is reflection in the y-axis
- c is a rotation of π radians around the origin.

These symmetry operations form a group whose group table is as follows.

	i	a	b	c
i	i	a	b	c
a	a	i	c	b
b	b	c	i	a
c	c	b	a	i



2.1.2.5 Cyclic Subgroups

Definition. *If we take a single member of a group (along with its inverse and the identity), the subgroup generated by that element takes the form (using multiplicative notation),*

$$H = \{x^{-(n-1)}, \cdots, x^{-2}, x^{-1}, 1, x, x^2, \cdots, x^{n-1}\}$$

where, either, $x^n = 1$ so that there are n distinct values in the group, or else n is infinite and the values never repeat. This is known as a ***cyclic group*** *and also as the* ***subgroup generated by x*** *and is denoted by $\langle x \rangle$.*

The cyclic subgroup, $\langle x \rangle$, generated by x is the smallest subgroup of G containing x in the sense that, if $H \leq G$ and $x \in H$ then $\langle x \rangle \subseteq H$.

Proposition 18. *Every cyclic group is Abelian.*

Proof. In a cyclic group every element has the form x^i for $i \in \mathbb{Z}$. So we have,

$$x^m x^n = x^{m+n} = x^n x^m$$

for all elements x^i in the group. □

Proposition 19. *The set S of integers n such that $x^n = 1$ is a subgroup of $\mathbb{Z}^+$.*

Proof. If $x^m = 1$ and $x^n = 1$, then $x^{m+n} = x^m x^n = 1$ also so we have closure of addition. Since $x^0 = 1$, 0 is in the subgroup so we have an identity. Finally, for some n in the subgroup, $x^n = 1 \iff x^{-n} = x^n x^{-n} = x^0 = 1$ so n being in the subgroup implies that $-n$ is also in the subgroup and we have inverses. □

Corollary 3. *It follows from S being a subgroup of $\mathbb{Z}^+$ and from Proposition 14 that S has the form $m\mathbb{Z}$ where m is the smallest positive integer such that $x^m = 1$. Therefore, in H, the m elements $1, x, x^2, \cdots, x^{m-1}$ are all different and any element in H will simplify to one of them: for $n \in S$, $n = mq + r$ such that $x^n = (x^m)^q x^r = 1^q x^r = x^r$.*

2.1.2.6 Order

Definition. *The **order** of a group G is the number of distinct elements it contains. It is typically denoted $|G|$.*
*An element of a group is said to have **order** m (possibly infinity) if the cyclic subgroup it generates has order m. This means that m is the smallest positive integer with the property $x^m = 1$ or, if the order is infinite, that, $x^m \neq 1$ for all $m \neq 0$.*

Theorem 3. *An element and its inverse have the same order.*

Proof. Firstly we need to consider the case that an element x has infinite order. In this case, $\nexists m \in \mathbb{N}$ s.t. $x^m = e$. Now suppose that $\exists n \in \mathbb{N}$ s.t. $(x^{-1})^n = e$. Then we have,

$$(x^{-1})^n = (x^n)^{-1} = e \iff e = x^n$$

which contradicts the hypothesis that x has infinite order. Therefore x^{-1} has infinite order also. Clearly also this argument can be used in reverse to show the reverse implication also holds.

Now consider the case that x has finite order. Let $x \in G$ be an arbitrary

member of an arbitrary group such that $x^m = e$. Then $x^{-1} = x^{m-1}$ and if we consider powers $i \in \mathbb{N}$ of the inverse $(x^{-1})^i = (x^{m-1})^i$ then the order is the lowest value of $i(m-1)$ such that $x^{i(m-1)} = e$. But we know that the lowest power of x equal to e is m so we're looking for the lowest multiple of m that has the form $i(m-1)$. So we require,

$$m \mid i(m-1) = im - i \iff (m \mid im) \wedge (m \mid i)$$

which clearly requires that $m \mid i$. Also, clearly, the lowest such i is $i = m$.

Another way to show this is to say that if $x^m = e$ then,

$$(x^{-1})^m = (x^m)^{-1} = e$$

so that the order of x^{-1} is less than or equal to m. Conversely, if x^{-1} has order n then,

$$x^n = ((x^{-1})^{-1})^n = ((x^{-1})^n)^{-1} = e^{-1} = e$$

so that the order of x is less than or equal to n. Thus we have,

$$m \leq n,\ n \leq m \implies m = n.$$

□

Theorem 4. *An element has order 2 iff it is equal to its inverse.*

Proof. Let $x \in G$ be an arbitrary member of an arbitrary group such that $x^2 = e$. Then by Theorem 3 we have,

$$e = x^2 = (x^{-1})^2 = x^{-2} \iff x = x^{-1}.$$

Also,

$$x = x^{-1} \iff x^2 = e.$$

□

Theorem 5. *A group of finite order cannot have any element of infinite order.*

Proof. If G is a group and $x \in G$ has infinite order then,

$$x^m = x^n$$

$$
\begin{aligned}
&\iff && x^{m-n} = 1 = x^{n-m} && \\
&\iff && x^{|m-n|} = 1 && \\
&\therefore && |m-n| = 0. && \text{because order of x is infinite}
\end{aligned}
$$

So, there are no two distinct powers of x that produce the same object so that $\langle x \rangle \leq G$, the cyclic group generated by x, is infinite. Since $\langle x \rangle \subseteq G$ this requires that G also be infinite. □

Theorem 6. *If a group element x has finite order m then:*

1. *Let $n \in \mathbb{Z}$. If $n = km + r$ where $k, r \in \mathbb{Z}$ and $0 \leq r \leq m-1$, then $x^n = x^r$.*
2. *For $n \in \mathbb{N}$, $x^n = 1 \iff m|n$.*
3. *$1, x, x^2, \ldots, x^{m-1}$ is a complete, repetition-free, list of elements of $\langle x \rangle$.*
4. *The subgroup $\langle x \rangle$ generated by x has cardinality m.*

Theorem 7. *In an Abelian group the set of all elements of finite order forms a subgroup.*

Proof. Let S be the set of all elements of finite order,

$$S = \{\, g \in G \mid \exists m \in \mathbb{N} \,.\, g^m = e \,\}.$$

- Firstly, S is non-empty because it contains the identity.
- Secondly, if we have $a, b \in S$ then $a^i = b^j = e$ for some $i, j \in \mathbb{N}$. Then, if we let $m = i \times j$,

  $$(ab)^m = a^m b^m = (a^i)^j (b^j)^i = e$$

  where $(ab)^m = a^m b^m$ is valid *only* because the group is Abelian.
- Lastly, S contains all inverses because,

  $$a^i = e \iff a^{i-1} = a^{-1} \iff (a^{-1})^i = (a^i)^{i-1} = e.$$

  Therefore $a^{-1} \in S$.

□

Theorem 8. *If every non-identity element of a group has order 2 then the group is Abelian.*

Proof. Let $x, y \in G$ be two arbitrary elements of order 2 of an arbitrary group. Then by Theorem 3 we have $x = x^{-1},\ y = y^{-1},\ xy = xy^{-1}$ and so,

$$xy = (xy)^{-1} = y^{-1}x^{-1} = yx.$$

Note that $(xy)^{-1} = y^{-1}x^{-1}$ relies only on the associativity of the group operation and is therefore valid for all groups.
We can also show it this way,

$$(xy)^2 = e \iff xyxy = e \iff yxy = xe = x \iff yx = xy.$$

□

Theorem 9. *If a finite group has even-numbered order then it must have at least one element of order 2.*

Proof. By the group properties we know that the group contains the identity element – which has order 1 – and, for every non-identity element, the group also contains its inverse. Also, since the group is finite, every element must have finite order. Now, if every non-identity element is distinct from its inverse then the order of the group will be odd (because of the identity and then every other element is paired with its inverse). For the group's order to be even we must have at least one non-identity element that is not distinct from its inverse which, by Theorem 4, is equivalent to having order 2. □

Corollary 4. *If a finite group has even-numbered order then it must have an odd number of elements of order 2.*

Proof. Let G be a finite group with even-numbered order and M be the number of elements that are distinct from their inverse and N be the number of elements that are not distinct from their inverse (these correspond to elements with order greater than 2 and elements with order 2 respectively). Then the order of G can be expressed as,

$$|G| = 1 + 2M + N.$$

Therefore, $|G|$ is even if N is an odd natural number. □

Theorem 10. *In an infinite cyclic group all elements have infinite order.*

Proof. Let $G = \langle x \rangle$ be an infinite cyclic group. Suppose there is some non-identity element of G, x^n with finite order m. Then,

$$(x^n)^m = e \iff x^{nm} = e$$

which contradicts the hypothesis that $\langle x \rangle$ is infinite. □

Note that the elements of an infinite cyclic group having infinite order does not mean that they generate the group. For example in an infinite cyclic group $\langle x \rangle$, the element x^2 generates the cyclic subgroup,

$$\ldots x^{-4}, x^{-2}, e, x^2, x^4, \ldots$$

which is infinite but clearly doesn't generate the whole group $\langle x \rangle$.

Theorem 11. *An infinite cyclic group has 2 generators.*

Proof. Let $G = \langle x \rangle$ be an infinite cyclic group and suppose there is some non-identity element of G, x^n that generates the group. To show this we only need to show that x^n can generate x because, since x is a member of the group, it is obviously necessary to generate it but, also, if we generate x then we can generate all the other members of the group since they are powers of x.
So let there be an integer a such that $(x^n)^a = x^{an} = x \iff x^n = x^{1/a}$. The cyclic group $\langle x \rangle$ only contains integer powers of x so it therefore follows that $|a| = 1$ which implies that $n = 1$ or -1 and $x^n = x$ or x^{-1}.

We could also say that,

$$x^{an} = x \iff x^{an-1} = e$$

but this implies that the order of x is finite and so contradicts the hypothesis that x generates an infinite cyclic group.

□

Theorem 12. *Let $G = \langle x \rangle$ be a finite cyclic group of order n. If r is a positive integer then $G = \langle x^r \rangle$ if and only if the greatest common divisor of n and r is 1.*

Proof. Members of $\langle x^r \rangle$ have the form $(x^r)^a$ for some $a \in \mathbb{Z}$. For integers b, i,

$$(x^r)^a = x^{ar} = x^{bn+i} = x^{bn}x^i = ex^i = x^i$$

so that the generated elements are x^i where $i = ar - bn$ is the remainder when dividing ar by n. If $d = gcd(n, r)$ then $d \mid i$ and the generated elements are powers of x that are multiples of d. Therefore, to generate every power of x it is necessary to have $d = 1$. Conversely, we can see – by the same argument in reverse – that it is sufficient if $d = 1$ to generate all the powers of x.

Alternatively, we can say if $d > 1$ then n/d is a positive integer less than n and r/d is a positive integer so,

$$(x^r)^{n/d} = (x^n)^{r/d} = e^{r/d} = e$$

which shows that the order of x^r is less than or equal to n/d which is less than n. Therefore the order of the cyclic group it generates is less than n and so it cannot be equal to G.
Conversely, if $d = 1$ then n and r are coprime and so we have,

$$(x^r)^m = e \iff x^{rm} = e \implies n \mid rm \implies n \mid m \implies m \geq n.$$

This says that the order of x^r in G is greater than or equal to n, the order of G. Well, clearly it cannot be greater than the order of G so it follows therefore, that the order of x^n is n. Since $|\langle x^r \rangle| = |G|$ we can conclude that $\langle x^r \rangle = G$. □

Theorem 13. *A group G is such that G contains at least 2 elements and the only subgroups of G are $\{e\}$ and G itself. Then G is a finite cyclic group of prime order.*

Proof. G contains at least 2 elements so there is at least one non-identity element x. The only subgroups of G are the whole group and $\{e\}$ but $\langle x \rangle$ cannot equal $\{e\}$ so it must equal G. Therefore G is the cyclic group generated by x.
But if G were the infinite cyclic group generated by x then only x and x^{-1} would generate the group and all other non-identity elements – say x^n for $n > 1 \in \mathbb{N}$ – would generate subgroups $\langle x^n \rangle \neq G$. Therefore, G cannot be infinite and is therefore finite.
Now, we have a finite cyclic group where every non-identity element generates

the group. Let $|G| = n$. Then, for every m s.t. $0 < m < n$, $\langle x^m \rangle = G$ and, by Theorem 12, m and n are coprime. Therefore, n is prime.
Here we could also use a proof by contradiction: Assume n is not prime and it has factors $r, s > 1 \in \mathbb{N}$. Then,

$$(x^r)^s = x^{rs} = x^n = e$$

so that the order of x^r in G is less than or equal to s which is less than n (because it is a factor of n). It follows then that $\langle x^r \rangle \neq G$, contradicting the definition of G. Therefore n is prime. □

Proposition 20. *Suppose the elements x, y in a group G have orders m, n respectively and that the $gcd(m, n) = 1$. Then $\langle x \rangle \cap \langle y \rangle = \{e\}$ and, if x and y commute, then the order of xy in G is mn.*

Proof. One way to approach this is to say that for $z \in \langle x \rangle \cap \langle y \rangle$ we have some $0 \leq i < m, 0 \leq j < n$ such that,

$$z = x^i = y^j \iff x^{im} = e = y^{jm},\ x^{in} = y^{jn} = e$$

so that $x^{in} = y^{jm} = e \iff (m \mid in \quad \text{and} \quad n \mid jm)$. Note that,

$$(m \mid in \quad \text{and} \quad n \mid jm) \iff m, n \mid in + jm.$$

Now, applying the fact that $gcd(m, n) = 1$ we see that both m and n must divide 1. But both m and n are orders of elements and so, by definition, greater than 1. The only other alternative is that both $i, j = 0$ which results in $z = x^0 = y^0 = e$.

We could also have said,

$$x^{im} = e = x^{in} \iff x^{im-in} = e \iff m \mid im - in.$$

In this case we can apply the fact that $gcd(m, n) = 1$ to the statement that $m \mid i(m - n)$ to deduce that: either $m \mid (m - n) \iff m \mid 1$ which is impossible because m must be greater than 1; or $m \mid i$ which is also impossible because $i < m$. So, again, we are only left with the alternative that $i = 0$ which results in $z = x^0 = y^0 = e$.

Next we prove the order of $xy \in G$ and we begin by noting that, if x and y commute, then $(xy)^r = x^r y^r$. So if we assume that xy has order r then we must have $m, n \mid r$ and the lowest such r is the order. Well, the lowest common multiple of m, n is defined according to the gcd as described in the Number Theory treatment of Modular Arithmetic (1.1.2.4) as,

$$d = gcd(m, n) \implies lcd(m, n) = d \cdot (m/d) \cdot (n/d).$$

Clearly then, if $d = gcd(m, n) = 1$, then the lowest common multiple is mn and so the order of $xy \in G$ is mn.

Or to describe it a different way: $(xy)^{mn} = x^{mn}y^{mn} = ee = e$ so that the order of xy,

$$|xy| \leq mn.$$

Conversely any r such that $(xy)^r = e$ must have $m, n \mid r$ and the lowest common multiple of m and n is mn so

$$r \geq mn.$$

Therefore, $|xy| = mn$. □

2.1.2.7 Examples of Cyclic Subgroups

(4) **Cyclic group with order 3**

$$G = \{1, x, x^2\}$$

where $x^3 = 1$ is a cyclic group of order 3 generated by the element x. Note that, since this is a group, it must also contain the inverses, x^{-1}, x^{-2} but $x^3 = 1$ so $x^{-1} = x^2$ and $x^{-2} = x$.

(5) **Symmetries of an equilateral triangle**

Consider an equilateral triangle with vertices labeled A, B, C:

$$\begin{array}{c} A \\ B \;\; C. \end{array}$$

Every permutation of the vertices is a transformation that produces an object that occupies the same space as the original, i.e. a *symmetry*. If we take one of them, say, the clockwise rotation one place that results in,

$$\begin{array}{cc} & B \\ C & A \end{array}$$

and we name this r, then clearly – since there are 3 vertices – performing this same rotation 3 times leaves us back where we started. So, using function composition as the law of composition and multiplicative notation, $r^3 = i$ where i is the identity transformation. Also the inverse of r is r^2. So, we have a group consisting of $\{i, r, r^2\}$ and function composition. Notice the resemblance of this group to the previous group $\{1, x, x^2\}$; this group is *isomorphic* to the cyclic group of order 3.

(6) **Group $(\mathbb{Z}_5^*, \otimes)$**

Consider the element 2 modulo 5. Using multiplicative notation we have,

$$2^2 = 4, 2^3 = 8 = 3, 2^4 = 16 = 1, 2^5 = 32 = 2.$$

So $2^1 = 2^5 \iff 1 = 2^4$ meaning that the element 2 has order 4 in the group and we see, as expected that the group it generates, $\langle 2 \rangle$ has 4 members. In this case, the members are all the members of the group – that's to say, *the element 2 generates the whole group*. If we consider the element 4 we have,

$$4^2 = 16 = 1, 4^3 = 64 = 4.$$

So this element oscillates between 1 and 4 and so, the cyclic subgroup that it generates $\langle 4 \rangle$ has order 2.
Since the group $(\mathbb{Z}_5^*, \otimes) = \langle 2 \rangle$ it can also be described as a cyclic group. This will be the case for any such group modulo a prime number – i.e. $(\mathbb{Z}_p^*, \otimes)$ where p is prime.

(7) **Cyclic group with infinite order**

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

under matrix multiplication (which is commutative in this case), generates a cyclic group of infinite order because

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

(8) **Cyclic groups in a non-Abelian group**

Consider the following two elements in $GL(2, \mathbb{R})$,

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

Both of these elements have finite order as,

$$(A^2)^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$(B^2)B = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

But their product AB does not have finite order.

$$AB = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \quad (AB)^n = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix}$$

for any $n \in \mathbb{N}$.

(9) **The Klein Four Group,** V is the simplest group that is not cyclic (it cannot be generated by a single element). It appears in many forms but, as an example, it can be realized as the group consisting of the four matrices,

$$\begin{bmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{bmatrix}$$

Any two non-identity elements generate V.

2.1.3 Isomorphisms

Definition. *An **isomorphism** is a bijection between two groups that preserves the structure of the groups by being compatible with the law of composition of both groups. More formally, two groups are **isomorphic** if there exists a bijection $\phi : G \longmapsto G'$ such that,*

$$\phi(ab) = \phi(a)\phi(b) \text{ for all } a, b \in G$$

where ab represents composition according to the law of composition of G and $\phi(a)\phi(b)$ represents composition according to the law of composition of G'.

*An **isomorphism** is a **bijection** between two **groups**. That's to say, it is already assumed in the definition of an isomorphism that the codomain G' is a group.*

Proposition 21. *As a consequence of this sole property that, across the bijection, the respective laws of composition are preserved, all other properties of the groups are also preserved.*

Proof. Let e be the identity in G and $e' = \phi(e) \in G'$, and $1'$ be the identity element in G' then,

- Since G' is a group, it has the inverses property that every element has an inverse so,

$$\begin{aligned} & e' = \phi(e) = \phi(ee) = \phi(e)\phi(e) = e'e' && \text{using preservation of law of composition} \\ \iff \quad & (e')^{-1}e' = ((e')^{-1}e')e' && \text{using the inverses property of } G' \\ \iff \quad & 1' = e' \end{aligned}$$

  which implies that e' is the identity in G' so that ϕ maps the identity in G to the identity in G'.

- We can use the fact just shown that $\phi(e) = e' = 1'$ to show,

$$1' = e' = \phi(e) = \phi(aa^{-1}) = \phi(a)\phi(a^{-1}) \quad \text{using preservation of law of composition}$$

$$\iff \qquad \phi(a)^{-1}1' = \phi(a)^{-1}\phi(a)\phi(a^{-1}) \qquad \text{using the inverses property of } G'$$
$$\iff \qquad \phi(a)^{-1} = \phi(a^{-1})$$

which shows that ϕ maps $a^{-1} \in G$ to $\phi(a)^{-1} \in G'$.

□

For example, if $e \in G$ is the identity of G mapped to an element $e' = \phi(e) \in G'$, then for any $a \in G$ mapped to $a' = \phi(a) \in G'$,

$$a' = \phi(a) = \phi(ea) = \phi(e)\phi(a) = e'a'$$

And $a' = e'a' = a'e'$ means that e' is the identity in G'. Furthermore, the order of elements in G and G' will also be the same as,

$$a^n = e \iff e' = \phi(e) = \phi(a^n) = \phi(a)^n = (a')^n$$

Since two isomorphic groups have the same properties, it is often convenient to identify them with each other when speaking informally. For example, the symmetric group S_n of permutations of $\{1, \cdots, n\}$ is isomorphic to the group of permutation matrices, a subgroup of $GL_n(\mathbb{R})$ and we often blur the distinction between these two groups.

Notation. Sometimes when two groups are isomorphic this is indicated using the notation,

$$G \approx G'$$

2.1.3.1 Examples

- Let $C = \{\cdots, a^{-2}, a^{-1}, 1, a, a^2, \cdots\}$ be an infinite cyclic group. Then the map,

  $$\phi : \mathbb{Z}^+ \longmapsto C \text{ s.t. } \phi(n) = a^n$$

  is an isomorphism where the preservation of the respective laws of composition can be seen as,

  $$\phi(m+n) = a^{m+n} = a^m a^n = \phi(m)\phi(n)$$

  and also $n + (-n) = 0$ and,

  $$\phi(-n) = a^{-n} = (a^n)^{-1}.$$

- Let G be the set of real matrices of the form,

$$\begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}$$

This is a subgroup of $GL_2(\mathbb{R})$ and so, its law of composition is the same as that of $GL_2(\mathbb{R})$, i.e. matrix multiplication.

$$\begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & y \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & x+y \\ 0 & 1 \end{bmatrix}$$

So, G is isomorphic to $\mathbb{R}^+$, the additive group of reals.

Definition. *The groups isomorphic to a given group G form what is called the* ***isomorphism class*** *of G. Groups are often classified into isomorphism classes, for example, there is one isomorphism class of groups of order 3 and there are two classes of groups of order 4 and five classes of 12.*

Proposition 22. *There is only one isomorphism class for each order of cyclic group.*

Proof. Any two cyclic groups of the same order are isomorphic because, if

$$G = \{1, x, x^2, \cdots, x^{n-1}\},\ G' = \{1, y, y^2, \cdots, y^{n-1}\}$$

are two cyclic groups of order n then the map $\phi(x^i) = y^i$ is an isomorphism. □

Proposition 23. ***Cayley's Theorem*** *states that every group is isomorphic to a group of permutations of the same underlying set, or in other words, to a subgroup of the symmetric group acting on the group.*

Proof. Let G be a group and $x \in G$ and define $f_x : G \longmapsto G$ as $f_x(g) = xg$. Then f_x is a bijection because it has an inverse ${f_x}^{-1}(g) = x^{-1}g = f_{x^{-1}}(g)$. Therefore f_x is a permutation.

Now we define a map, from G to the symmetric group of permutations of G, that maps each element x in G to the permutation defined by f_x. Let $\phi : G \longmapsto Sym(G)$ be defined as $\phi(x) = f_x$ then,

- ϕ is homomorphic because

$$(f_x \circ f_{x'})(g) = f_x(f_{x'}(g)) = x(x'g) = xx'g = f_{xx'}(g)$$

and so,

$$\phi(xx') = f_{xx'} = f_x \circ f_{x'} = \phi(x) \circ \phi(x').$$

- ϕ is injective because if $x, x' \in G$ such that $x \neq x'$ then $f_x(e) = x \neq x' = f_{x'}(e)$ is sufficient to show that $f_x \neq f_{x'}$. Or alternatively, the kernel of ϕ comprises the elements $k \in G$ such that $f_k(g) = kg = g \iff k = e$ so that the kernel is the trivial subgroup $\{e\}$.

Since ϕ is homomorphic, its image $im\,\phi$ is a subgroup of $Sym(G)$ and since ϕ is injective, it is in bijective correspondence with its image $im\,\phi \leq Sym(G)$. Therefore G is isomorphic to a subgroup of $Sym(G)$. □

2.1.3.2 Automorphisms

Definition. *The domain and codomain of an isomorphism can be the same set of objects so that $\phi : G \longmapsto G$. This is known as an* ***automorphism****.*

Example Let $G = \{1, x, x^2\}$ be a cyclic group of order 3 so that $x^3 = 1$. The transposition which interchanges x and x^2 is an automorphism of G,

$$\begin{array}{ccc} 1 & \longmapsto & 1 \\ x & \longmapsto & x^2 \\ x^2 & \longmapsto & x \end{array}$$

	1	x	x^2
1	1	x	x^2
x	x	x^2	1
x^2	x^2	1	x

$\longmapsto$

	1	x^2	x
1	1	x^2	x
x^2	x^2	x	1
x	x	1	x^2

This is because the group is cyclic and x and x^2 have the same order ($x^3 = 1$ and also $(x^2)^3 = x^6 = (x^3)^2 = 1^2 = 1$). So the law of composition is preserved.

2.1.3.3 Conjugation

The most important example of automorphism is conjugation.

Definition. ***Conjugation*** *by $b \in G$ is the map from G to itself defined by,*

$$\phi(a) = bab^{-1}$$

with the result that,

$$ba = \phi(a)b$$

so that we can think of conjugation of a by b as the way that we need to change a if we want to move the multiplication by b to the other side.

This is an automorphism (known as an *inner automorphism*) because it

- is compatible with law of composition,
$$\phi(xy) = bxyb^{-1} = bxb^{-1}byb^{-1} = \phi(x)\phi(y).$$
- has an inverse so it is bijective,
$$(\phi^{-1} \circ \phi)(a) = \phi^{-1}(\phi(a)) = b^{-1}(bab^{-1})b = (b^{-1}b)a(b^{-1}b) = a.$$
Note that this is different from the inverse element of a corresponding under the mapping ϕ,
$$\phi(a)\phi(a^{-1}) = bab^{-1}ba^{-1}b^{-1} = ba(1)a^{-1}b^{-1} = b(1)b^{-1} = 1.$$

A couple more important properties of the conjugate are as follows.

(i) In an abelian group where the composition law is commutative, conjugation becomes the identity map.
$$ba = ab \iff bab^{-1} = a \iff \phi(a) = a.$$

(ii) The inverse of the conjugate $bab^{-1} = b^{-1}ab$.

2.1.4 Homomorphisms

Definition. *A **homomorphism** is a mapping (not necessarily bijective) between two groups, $\phi : G \longmapsto G'$, such that,*

$$\phi(ab) = \phi(a)\phi(b) \text{ for all } a, b \in G$$

where ab represents composition according to the law of composition of G and $\phi(a)\phi(b)$ represents composition according to the law of composition of G'.

So, the difference between a *homomorphism* and a *isomorphism* is that the latter is bijective whereas the former is not. As a result, a *homomorphism* may be one-way only.

*A **homomorphism** is a **mapping** between two **groups**. That's to say, it is already assumed in the definition of a homomorphism that the codomain G' is a group.*

Examples of homomorphisms

(10) Let $C = \{a^{n-1}, \cdots, a^{-2}, a^{-1}, 1, a, a^2, \cdots, a^{n-1}\}$ be a finite cyclic group. Then the map,

$$\phi : \mathbb{Z}^+ \longmapsto C \text{ s.t. } \phi(n) = a^n$$

is a homomorphism. Note that if C were an infinite cyclic group then this would be an isomorphism.

(11) the sign of a permutation $sign : S_n \longmapsto \pm 1$

(12) the determinant function $det : GL_n(\mathbb{R}) \longmapsto \mathbb{R}^\times$

(13) an arguably trivial example is called the *inclusion* map $i : H \longmapsto G$ of a subgroup H into a group G, defined by $i(x) = x$. It functions as the identity for elements in the subgroup H but, since it is not surjective, there is no inverse mapping.

2.1.4.1 Image of a homomorphism

Since a homomorphism is not bijective it has an image different to the codomain group,

$$im\ \phi = \{\, x \in G' \mid \exists a \in G \text{ s.t. } \phi(a) = x \,\}$$

The image of a homomorphism is a subgroup of the codomain group G' because the homomorphism preserves the group structure as described in Proposition 21.

Notation. The image of the mapping ϕ with domain G is sometimes denoted $\phi(G)$.

2.1.4.2 Kernel of a homomorphism

Definition. *The* ***kernel*** *of a homomorphism is the set of elements in the domain that are mapped to the identity,*

$$ker\ \phi = \{\, a \in G \mid \phi(a) = 1' \,\}$$

Proposition 24. *The kernel of a homomorphism is a subgroup of the domain group G.*

Proof. If $a, b \in ker\ \phi$ then,

- closure: $\phi(ab) = \phi(a)\phi(b) = 1' \cdot 1' = 1'$ which shows that

$$a, b \in ker\ \phi \implies ab \in ker\ \phi.$$

- identity: By Proposition 21, $1' = e' = \phi(e)$ and so $e \in ker\ \phi$.
- inverses: Since $a \in ker\ \phi$, then

$$\begin{aligned} & 1' = e' = \phi(e) = \phi(aa^{-1}) = \phi(a)\phi(a^{-1}) = 1'\phi(a^{-1}) \\ \iff \quad & 1' = \phi(a^{-1}) \end{aligned}$$

  so that $a \in ker\ \phi \iff a^{-1} \in ker\ \phi$.

□

Proposition 25. *If $\phi : G \longmapsto G'$ is a group homomorphism with kernel N then, for $a, b \in G$,*

$$\phi(a) = \phi(b) \iff \exists n \in N, \ s.t. \ b = an$$

or, equivalently, $a^{-1}b \in N$.

Proof.

$$\begin{aligned}
& b = an \\
\implies \quad & \phi(b) = \phi(an) \\
\implies \quad & \phi(b) = \phi(a)\phi(n) && \text{by homomorphism property} \\
\implies \quad & \phi(b) = \phi(a)1' && \text{n is in the kernel} \\
\implies \quad & \phi(b) = \phi(a)
\end{aligned}$$

$$\begin{aligned}
& \phi(b) = \phi(a) \\
\implies \quad & \phi(a)^{-1}\phi(b) = 1' && \text{codomain is a group so has inverses} \\
\implies \quad & \phi(a^{-1})\phi(b) = 1' && \text{by Proposition 21} \\
\implies \quad & \phi(a^{-1}b) = 1' && \text{by homomorphism property} \\
\implies \quad & a^{-1}b = n \in N \\
\implies \quad & b = an
\end{aligned}$$

□

Theorem 14. *A homomorphism is injective iff its kernel is the trivial subgroup $\{e\}$.*

> *When asked to prove the proposition that a homomorphism is injective iff its kernel is the trivial subgroup $\{e\}$, it's tempting to begin proving each direction of the bidirectional implication with a proof by contradiction (e.g. "Assuming there is a non-identity element in the kernel...") but the direct positive proof can be made very quick and simple with the above corollary.*

Proof. Let $\phi : G \longmapsto G'$ be a homomorphism.

Assume that ϕ is injective and $k \in ker\,\phi$. Remembering that we always at least have $e_G \in ker\,\phi$,

$$\phi(k) = e_{G'} = \phi(e_G) \iff k = e_G$$

where the last implication is by the injectivity of ϕ.

Now assume that $ker\,\phi = \{e_G\}$, $a, b \in G$. Then,

$$\begin{aligned}
& \phi(a) = \phi(b) \\
\iff \quad & \phi(a)\phi(b)^{-1} = e_{G'} \\
\iff \quad & \phi(a)\phi(b^{-1}) = \phi(ab^{-1}) = e_{G'} && \text{using homomorphism properties} \\
\iff \quad & ab^{-1} \in ker\,\phi \\
\iff \quad & ab^{-1} = e_G && \text{by assumption } ker\,\phi = \{e_G\} \\
\iff \quad & a = b. && \square
\end{aligned}$$

Corollary 5. *A homomorphism is an isomorhpism if its kernel contains only the identity and its image is the whole of the codomain (i.e. it's surjective).*

2.1.4.3 Examples of Kernels of Homomorphisms

(14) The determinant function 12, $det : GL_n(\mathbb{R}) \longmapsto \mathbb{R}^\times$, has a kernel,

$$\{\, \text{real } n \times n \text{ matrices } A \mid det\, A = 1 \,\},$$

which is a subgroup of $GL_n(\mathbb{R})$ known as the *special linear group* $SL_n(\mathbb{R})$.

(15) The sign of a permutation 11 has a kernel that is the set of *even* permutations,

$$A_n = \{\text{even permutations}\},$$

which is a subgroup of the symmetric group S_n and is known as the *alternating group*, A_n.

(16) The map from the additive group of integers to a finite cyclic group 10,

$$\phi : \mathbb{Z}^+ \longmapsto C \;\text{ s.t. }\; \phi(n) = a^n$$

has the kernel,

$$ker\ \phi = \{\, n \in \mathbb{Z}^+ \mid a^n = 1 \,\}$$

which has been proven to be a subgroup in Proposition 19.

2.1.5 Equivalence Relations and Partitions

Notation. In the following treatment of equivalence relations we will use the notation $a \sim b$ to denote the equivalence of a and b; $\overline{a}$ to indicate the equivalence class of a; and $\overline{S}$ to indicate the partition of S comprised of equivalence classes such as the class $\overline{a} = \overline{b}$ which includes both a and b.

Any map of sets $\phi : S \longmapsto T$ defines an equivalence relation on the domain S such that $a \sim b$ iff $\phi(a) = \phi(b)$. We will refer to this as the *equivalence relation determined by the map.* The corresponding partition is made up of the sets of elements in the domain S that are mapped to the same element in the codomain T.

Definition. *Let $\phi : S \longmapsto T$ be a map, then the **inverse image** of an element $t \in T$ is defined as,*

$$\phi^{-1}(t) = \{\, s \in S \mid \phi(s) = t \,\}$$

and can also be applied to a set $U \in T$ as,

$$\phi^{-1}(U) = \{\, s \in S \mid \phi(s) \in U \,\}$$

*Note that in this notation, ϕ^{-1} **does not indicate an inverse function** as the inverse of the function may not exist but the inverse image is nevertheless defined.*
*The inverse images - the sets $\phi^{-1}(t)$ for all $t \in T$ - may also be called the **fibres** of the map ϕ.*

Clearly, the non-empty fibres of the map ϕ form a partition of S. We can express this partition of S as a bijection, that we shall call $\overline{\phi}$, between the

fibres of ϕ in S and the element of the image of S to which their members are mapped,

$$\overline{\phi} : \overline{S} \longmapsto im\ \phi$$

so that,

$$\overline{\phi}(\overline{s}) = \phi(s).$$

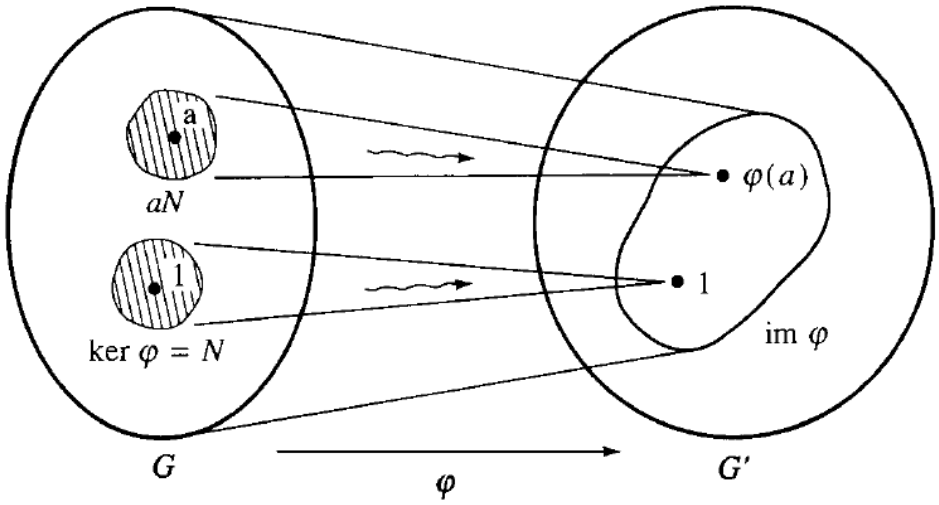


Figure 2.1: A schematic diagram of a group homomorphism

2.1.5.1 Congruence

Since a homomorphism maps the identity to the identity and inverses to inverses (Proposition 21), we can deduce that the inverse image of the identity in G' is going to contain at least the identity of G and that the inverse image of an element $(a')^{-1} \in G'$ will contain at least the element $a^{-1} \in G$. So, in terms of equivalence classes we can say that, for a homomorphism ϕ,

$$1 = e \in \phi^{-1}(1') \implies \overline{\phi}(\overline{1}) = 1'$$

$$a^{-1} \in \phi^{-1}((a')^{-1}) \implies \overline{\phi}(\overline{a^{-1}}) = (a')^{-1}$$

Definition. *The equivalence relation determined by a homomorphism is known as* ***congruence*** *and is commonly denoted using* $\equiv$ *instead of* $\sim$*. For a homomorphism* ϕ*,*

$$a \equiv b \iff \phi(a) = \phi(b).$$

Since ϕ *is a homomorphism we also have,*

$$a \equiv b \iff \phi(ac) = \phi(bc),\ \phi(a^{-1}) = \phi(b^{-1}).$$

More generally, a ***congruence relation*** *is an equivalence relation on an algebraic structure (such as a group, ring, or vector space) that is compatible with the structure in the sense that algebraic operations done with equivalent elements will yield equivalent elements.*

2.1.5.2 Congruence Examples

(17) The modulus function of complex numbers forms a homomorphism from the multiplicative group of complex numbers to the multiplicative group of reals,

$$\phi : \mathbb{C}^\times \longmapsto \mathbb{R}^\times \text{ s.t. } \phi(a) = |a|$$

and the induced equivalence relation is $a \equiv b \iff |a| = |b|$. The fibres of this map are the concentric circles about 0. They are in bijective correspondence with elements of $im\ \phi$, the set of positive reals.

2.1.6 Cosets

The set of elements of the form an - described in Proposition 25 - is denoted by aN and is called a *coset* of N in G.

Definition. *A coset can be defined for any subgroup H of a group G. A **left coset** is a subset of the form,*

$$aH = \{\, ah \mid h \in H \,\}.$$

Cosets are not, in general, subgroups. This can be easily seen as the left coset aH does not contain the identity as, although H contains the identity, aH contains $a1 = a$.

Note that the arbitrary subgroup H could also be thought of as a coset $1H = H$ and also that the left cosets aH are equivalence classes for the congruence relation,

$$a \equiv b \iff b = ah,\ h \in H.$$

This is a congruence because, for some arbitrary $c \in G$,

$$1 \equiv c \iff \exists h \in H \text{ s.t. } c = 1h = h.$$

That's to say, the elements that are congruent to the identity are precisely the members of the subgroup H so that it plays a similar role to the kernel N in Proposition 25. Furthermore, since the congruence relation is an equivalence relation it forms a partition of the domain G.

Proposition 26. *For a group G with a subgroup H and $x \in G$, the coset xH is equal to H iff $x \in H$.*

Proof. Assume $x \in H$. Then $\forall xh \in xH\ .\ xh \in H$. Therefore $xH \subseteq H$. Conversely, $x^{-1} \in H$ so,

$$\forall h \in H\ .\ x^{-1}h \in H \implies x(x^{-1}h) = h \in xH.$$

Therefore $H \subseteq xH$ and so $H = xH$.

Now assume that $H = xH$. Since $e \in H$ then $xe = x \in xH = H$ and so $x \in H$. □

Proposition 27. *The left cosets of a subgroup partition the group.*

Proof. The left cosets are equivalence classes and, as a result, they partition the group. □

2.1.6.1 Examples of cosets

(18) The coset of an element with the kernel N,

$$aN = \{\, g \in G \mid g = an,\ n \in N \,\}$$

is the set of all elements that are *congruent* to a. The *congruence classes* are precisely the cosets aN for each $a \in G$. They are also the nonempty *fibres* of the homomorphic map.

(19) 2.1.1.2 Continuing the example of the symmetric group S_3 represented as

$$G = \{1, x, x^2, y, xy, x^2y\}$$

with group multiplication rules,

$$x^3 = 1, y^2 = 1, yx = x^2y.$$

The element xy has order 2 so it generates a cyclic subgroup $H = \{1, xy\}$ of order 2. The left cosets of H in G are the three sets,

$$\{1, xy\} = 1H = xyH,\ \{x, x^2y\} = xH = x^2yH,\ \{x^2, y\} = x^2H = yH.$$

Note that they do partition the group G. Also, notice that the cosets aH for $a \in H$ produce the subgroup H itself as should be expected as the group properties of the subgroup dictate that all products of its elements are already present in the subgroup. For this reason, the cosets aH that are distinct from H are those such that $a \notin H$.

(20) Let $G = (\mathbb{R}^3, +)$ be the group of 3d vectors with vector addition and $\vec{\boldsymbol{w}} \in G$. Then if

$$H = \{\, \vec{\boldsymbol{x}} \in G \mid \vec{\boldsymbol{w}}^T \vec{\boldsymbol{x}} = \vec{\mathbf{0}} \,\}$$

then H is a subgroup, $H \leq G$. H is a vector space representing a plane through the origin in $\mathbb{R}^3$ and it's cosets are

$$\vec{\boldsymbol{v}} + H = \{\, \vec{\boldsymbol{v}} + \vec{\boldsymbol{h}} \mid \vec{\boldsymbol{v}} \in G, \vec{\boldsymbol{h}} \in H \,\}$$

which are the affine spaces representing the translated planes, parallel to H, but not passing through the origin. Once again we see that the cosets partition the space even if there may be an infinite number of them.

2.1.6.2 The index of a subgroup

Definition. *The **index** of a subgroup is the number of left cosets it forms in the parent group.*

Notation. The **index** of a subgroup H in G is denoted by $[G : H]$.

In the example (19) the index of H is 3. Note that if G were to contain infinitely many elements then the index of a subgroup may also be infinite.

Proposition 28. *Each coset aH has the same number of elements as H.*

Proof. As usual, equal cardinality is demonstrated by showing the existence of a bijection. It is clear that there is a bijective map between the subgroup H and any coset aH because the map $H \longmapsto aH$ is,

- injective because $ah = ah' \implies h = h'$ because by group properties a has an inverse in G;
- surjective because every $c \in aH$ has the form ah and is therefore mapped to by some $h \in H$.

□

2.1.6.3 Lagrange's Theorem

Since the left cosets of H in G form a partition of G and their order is the same as that of H we see that the order of G is the order of H multiplied by its index in G. This results in a formula known as the *Counting Formula* as follows,

$$|G| = |H| \cdot [G : H].$$

If G is of infinite order and H is finite, then the index of H in G will be infinite.

Theorem 15. ***Lagrange's Theorem:*** *Let G be a finite group, and let H be a subgroup of G. The order of H divides the order of G.*

Corollary 6. *Let G be a finite group, and let a be an element of G. Then the order of a divides the order of G. That's to say, the order of the cyclic group generated by a, $|\langle a \rangle|$, divides $|G|$.*

Corollary 7. *If G is a group of order n, then $g^n = e$ for every element g of G.*

Proof. This is clearly a consequence of the previous corollary. If we let the order of g be m, then by the previous corollary,

$$m \mid n \iff n = km \text{ for } k \in \mathbb{N} \iff g^n = g^{km} = (g^m)^k = e^k = e. \qquad \square$$

Corollary 8. *Suppose that a group G has p elements and that p is a prime integer. Let $a \in G$ be any element, not the identity. Then G is the cyclic group $\{1, a, \ldots, a^{p-1}\}$ generated by a.*

Proof. Since $a \neq 1$ by selection, it has order greater than 1. Since its order must divide the order of G, which is prime, its order is equal to the order of G, p. So, the order of the nonidentity element a is the same as the order of G and so it generates the whole group. $\square$

Corollary 9. *All groups with some prime order, p, are in the same isomorphism class.*

Proof. Any group with prime order p is the cyclic group of order p and by Proposition 22 there is only a single isomorphism class for each cyclic group of a given order. □

Proposition 29. *Suppose the elements x, y in a group G have orders m, n respectively and that the $gcd(m, n) = 1$. Then $\langle x \rangle \cap \langle y \rangle = \{e\}$.*

> *Here we will prove, using Lagrange's Theorem, something that we previously proved here (Proposition 20) using modular arithmetic. Notice how the proofs are similar but the proof with Lagrange's Theorem allows us to remain within Group Theory.*

Proof. Firstly, note that the intersection of the two cyclic groups,

$$H = \langle x \rangle \cap \langle y \rangle$$

is a subgroup both of the parent group G **and** of $\langle x \rangle$ and $\langle y \rangle$. So Lagrange's Theorem tells us that its order must divide into the order of the parent group **and** the orders of the cyclic groups of x and y. Therefore, we have,

$$|H| \mid m \quad \text{and} \quad |H| \mid n.$$

Now, applying the fact that the $gcd(m, n) = 1$ we see that $|H| \mid 1$ and therefore $|H| = 1$. Furthermore, any group of order 1 must be the minimal group $\{e\}$. □

Proposition 30. *Suppose that H is a subgroup of G and $x \in G$. Then there exists some $k \in \mathbb{N}$, $1 \leq k \leq [G : H]$ s.t. $x^k \in H$.*

Proof. Let $n = [G : H]$ be the index of H in G. Then there are precisely n cosets of H in G. But $x \in G \implies x^m \in G$ for any $m \in \mathbb{N}$ (we don't need to consider the negative powers of x because they are inverses of positive powers and are similar for these purposes) and so we have cosets of the form $x^m H$ for each $m \in \mathbb{N}$. Therefore, amongst the $n + 1$ cosets generated by, $x^i H$ for $i \in \{0, 1, \ldots, n\}$ we must have at least one repetition of the same coset.
So, for some fixed x^i, x^j with $0 \leq i, j \leq n$ and $i \neq j$, we have,

$$x^i H = x^j H$$

$$\begin{aligned}
&\iff && \forall h \in H \,.\, x^i h \in x^j H \\
&\iff && \forall h \in H \,.\, \exists h' \in H \,.\, x^i h = x^j h' \\
&\iff && \forall h \in H \,.\, \exists h' \in H \,.\, x^{i-j} = h^{-1}h' \in H
\end{aligned}$$

Since necessarily we have $1 \leq i - j \leq n$ we let $k = i - j$ and then $x^k \in H$ as required. □

2.1.6.4 Example applications of Lagrange Theorem

(21) **Fermat's Little Theorem**: *If p is a prime number then*

$$a^p \equiv a \bmod p \text{ for all } a \in \mathbb{Z}.$$

We need to be a little careful here. We might assume – given that we are multiplying the integer a in modulo p that the group we want to use is $(\mathbb{Z}_p, \otimes)$. However, this is not a group! The reason is that $\mathbb{Z}_p$ contains 0 which has no inverse under the proposed law of composition, multiplication.
If, however, we take $\mathbb{Z}_p^*$ where the $*$ means $\mathbb{Z}/\{0\}$ then we have a set of $p-1$ distinct elements. Over this set we can form the multiplicative group $G = (\mathbb{Z}_p^*, \otimes)$ because the primality of p means that every element has a multiplicative inverse.
*Note that this is **not** a group of prime order. The primality of p is essential to make sure that every element has a multiplicative inverse but, since we also have to eliminate 0 for the same reason, the order is $p-1$ which is not necessarily prime.*

Proof. Take the set $\mathbb{Z}_p^*$ under multiplication and some arbitrary $a \in \mathbb{Z}$.

(i) Primality of p means that it is possible to find $1 = na + mp$ for $m, n \in \mathbb{Z}$ (see Corollary 2). This implies that there exists a multiplicative inverse of every non-zero element in modulo p. Specifically, n is the inverse of a because $na = (-m)p + 1 \iff na \bmod p \equiv 1$.

(ii) Existence of the multiplicative inverses implies that we have a group $G = (\mathbb{Z}_p^*, \otimes)$.

(iii) G being a group implies that, for any element $a \in G$, by Corollary 7 we have $a^{p-1} = 1$.

(iv) In G, $a^{p-1} = 1 \iff a^p = a$ which translates to $a^p \equiv a \bmod p$.

□

2.1.6.5 Lagrange's Theorem and Homomorphisms

The Counting Formula can also be applied when a homomorphism is given. Let $\phi : G \longmapsto G'$ be a homomorphism. As we saw in coset example 18, the left cosets of $ker\ \phi$ are the fibres of the map ϕ. They are in bijective correspondence with the elements of the image. Therefore,

$$[G : ker\ \phi] = |im\ \phi| .$$

Which implies that,

Corollary 10. *If $\phi : G \longmapsto G'$ is a homomorphism of finite groups then,*

$$|G| = |ker\ \phi| \cdot |im\ \phi| .$$

As a result, $|ker\ \phi|$ divides $|G|$, and $|im\ \phi|$ divides both $|G|$ and $|G'|$.

2.1.6.6 Restriction of a Homomorphism to a Subgroup

A useful way of understanding the structure of a complicated group is to understand its subgroups and then derive an understanding of the parent group from knowledge about the subgroups it contains. This frequently involves the application of Lagrange's Theorem. *Restriction of a Homomorphism to a subgroup* refers to studying the behaviour of a homomorphism on subgroups of the parent group.

Suppose that $\phi : G \longmapsto G'$ is a homomorphism and that H is a subgroup of G. Then we may *restrict ϕ to H* to obtain a homomorphism whose domain is a subset of the original,

$$\phi|_H : H \longmapsto G'.$$

This *restriction* is a homomorphism because ϕ is a homomorphism and the restriction domain is a group. Clearly, the kernel of the restricted homomorphism is the intersection of the domain H with $ker\ \phi$.

2.1.6.7 Examples of using Lagrange's Theorem with a homomorphism restricted to a subgroup

(22) Referring again to the sign of a permutation (11) $S_n \longmapsto \{-1, 1\}$: the order of the codomain of this homomorphism is clearly 2. Suppose we form the restriction of this homomorphism to a subgroup H of S_n. Then, denoting the image by $\phi|_H(H)$, by Corollary 10 we have that $|\phi|_H(H)|$ divides both 2 and $|H|$.
So, if the subgroup H has odd order then $|\phi|_H(H)| = 1$ and – since $\phi|_H(H)$ must be a group because the group structure is preserved across the homomorphism – $\phi|_H(H) = \{1\}$. This means that H is in the kernel of the sign map and that the subgroup of permutations in S_n represented by H consists of only even permutations.
Therefore, every permutation whose order in S_n is odd is an even permutation (since the cyclic group that it generates has odd order). However, we can not make any conclusions about permutations of even order; they may be even or odd permutations.

2.1.6.8 Right Cosets

Right cosets also exist and are defined as,

$$Ha = \{\, g \in G \mid g = ha,\ h \in H \,\}$$

and these are equivalence classes for the *right congruence* relation,

$$a \equiv b \iff b = ha,\ h \in H.$$

Right cosets are not necessarily the same as left cosets. For instance, continuing the example in 19, the right cosets of the subgroup $\{1, xy\}$ of S_3 are,

$$\{1, xy\} = H1 = Hxy,\ \{x, y\} = Hx = H,\ \{x^2, x^2y\} = Hx^2 = Hx^2y.$$

Note that this generates a different partition of G then was generated by the left cosets.

2.1.7 Normal Subgroups and Centers

2.1.7.1 Normal Subgroups

Definition. *A subgroup N of a group G is called a* ***normal subgroup*** *if it has the property that,*

$$\forall\, a \in N, b \in G,\ bab^{-1} \in N$$

which is to say, that the conjugate by any element of G of any element in N is also in N.

Proposition 31. *A subset H of a group G is normal if and only if every left coset is also a right coset. If H is normal then,*

$$\forall a \in G,\ aH = Ha.$$

Proof. Suppose that H is normal. For any $h \in H$ and any $a \in G$,

$$ah = (aha^{-1})a.$$

Since H is normal, the conjugate by h of a is also in H, that's to say, $aha^{-1} \in H$ which implies that $(aha^{-1})a \in Ha$. Therefore, any arbitrary member of aH is also a member of Ha. Clearly, the same proof also works in the other direction so that any member of Ha is also a member of aH and the two cosets are equal. So, we have shown that (H is normal) $\implies$ (left and right cosets of H are equal).
Now we need to show that (left and right cosets of H are equal) $\implies$ (H is normal). Firstly, clearly the above logic doesn't apply if H is not normal; there will be at least one element whose conjugate is not in H so $aH \neq Ha$. However, it could still be the case that each left coset is also a right coset if, for every a in G, there is some b in G such that $aH = Hb$. However, this is not possible because aH and Ha both contain a which means that in a given partition of G they must be the same partition. So $aH \neq Ha$ implies that the partitions are different; Ha creates different equivalence classes. Therefore (left and right cosets of H are equal) $\implies$ (H is normal). □

This is really the point of normal subgroups: That their cosets contain multiplication from both sides. As a result, when the cosets themselves are used as members of groups (see: Quotient Groups 2.1.8.6), we can define a group composition operation between them such that $aHbH = abH$ and the operation is well defined (i.e. equal arguments give equal results) because,

$$aHbH = abHH = abH \quad and \quad aH = a'H \implies abH = aHb = a'Hb = a'bH$$

where $aH = Ha$ because H is normal. .

2.1.7.2 Examples of Normal Subgroups

(23) The kernel of a homomorphism is a normal subgroup because,

$$\begin{aligned}
& a \in ker\ \phi \iff \phi(a) = 1 \\
\implies \quad & \phi(bab^{-1}) = \phi(b) \cdot 1 \cdot \phi(b^{-1}) \\
\iff \quad & \phi(bab^{-1}) = \phi(b)\phi(b)^{-1} && \text{using Proposition 21} \\
\iff \quad & \phi(bab^{-1}) = 1.
\end{aligned}$$

For example,

a. $SL_n(\mathbb{R})$ 14 is a normal subgroup of $GL_n(\mathbb{R})$ *even though it is not Abelian*, which can be seen as for $M \in GL_n(\mathbb{R}), A, B \in SL_n(\mathbb{R})$,

$$AB \neq BA \quad \text{and} \quad det\, A = 1,\ det\, M^{-1} = 1/(det\, M)$$

so that,

$$det\, M^{-1}AM = (det\, M^{-1}) \cdot 1 \cdot (det\, M) = (det\, M)/(det\, M) = 1.$$

b. A_n 15 is a normal subgroup of the symmetric group S_n.

c. In fact, any subgroup of $GL_n(\mathbb{F})$ with some fixed determinant $d \in \mathbb{R}$ will be a normal subgroup.

$$N = \{\, A \in GL_n(\mathbb{F}) \mid det\, A = d \,\}$$

For $A \in N, M \in GL_n(\mathbb{F})$. $det(MAM^{-1}) = d$ by the same logic as in example a. These conjugate matrices are known as *similar* matrices.

(24) Any subgroup of an abelian group is normal because when the composition law is commutative, as was mentioned in the section on conjugation,

$$ba = ab \iff bab^{-1} = abb^{-1} = a$$

so that conjugation becomes the identity map and so, trivially, all conjugates of elements in a subgroup are also in the subgroup.

Subgroups of non-abelian groups, however, need not be normal. For example,

(25) Group T of invertible upper triangular matrices is not a normal subgroup of $GL_2(\mathbb{R})$. To show this note,

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, BAB^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

where $A \in T, B \in GL_2(\mathbb{R})$ but $BAB^{-1} \notin T$.

Proposition 32. *Let $\phi : G \longmapsto G'$ be a homomorphism and let H' be a subgroup of G'. Denote the inverse image $\phi^{-1}(H') = \{ x \in G \mid \phi(x) \in H' \}$ by $\tilde{H}$. Then,*

(i) $\tilde{H}$ is a subgroup of G.

(ii) If H' is a normal subgroup of G' then $\tilde{H}$ is a normal subgroup of G.

(iii) $\tilde{H}$ contains ker ϕ.

(iv) The restriction of phi to $\tilde{H}$ defines a homomorphism $\tilde{H} \longmapsto H'$ whose kernel is ker ϕ.

Proof. Proofs are as follows:

(i) $\tilde{H}$ is a subgroup of G because ϕ is a homomorphism and its image, H', is a group (which is required for a homomorphism).

(ii) If H' is a normal subgroup of G' then $\tilde{H}$ is a normal subgroup of G because for every element in $\tilde{H}$ the mapped element is in H'. Then, since H' is normal, the conjugates of the mapped element are also in H' which means that their inverse images are in $\tilde{H}$. Since the map is homomorphic, the inverse images of the conjugates in G' are the respective conjugates in G.

(iii) $\tilde{H}$ contains $ker\ \phi$ because it contains every element in G that maps to an element in H' and, since H' is a group, it includes the identity of G'. Therefore $\tilde{H}$ contains every element that maps to the identity of G' which is $ker\ \phi$.

(iv) The restriction of *phi* to $\tilde{H}$ is clearly a homomorphism and, since it contains $ker\ \phi$, its kernel is equal to the kernel of ϕ.

□

2.1.7.3 The Center of a Group

Definition. *The **center** of a group G is the set of elements that commute with every element of G,*

$$Z = \{\, z \in G \mid zx = xz,\ \forall x \in G \,\}.$$

We can also define,

$$C(x) = \{\, g \in G \mid gx = xg \,\}$$

as the set of elements in G that commute with a single fixed element x.

Notation. The **center** of a group G may be denoted by Z or by $Z(G)$.

The center of a group, Z, is a subgroup of G. This can be easily seen as, first of all, Z is non-empty because the identity is in the center of any group. Then, also, the center Z is closed under the group operation,

$$\forall a, b \in Z, x \in G\ .\ (ab)x = axb = x(ab)$$

and it contains the inverses,

$$\forall a \in Z, x \in G\ .\ ax = xa \iff a^{-1}ax = x = a^{-1}xa \iff xa^{-1} = a^{-1}x.$$

2.1.7.4 Examples of group centers

(26) Let $G = GL(2, \mathbb{R})$ be the group of invertible 2x2 matrices with real coefficients and take two elements in G,

$$M = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then we can identify the center of M by observing that an arbitrary matrix in $C(M)$ satisfies,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2a & b \\ 2c & d \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 2a & 2b \\ c & d \end{pmatrix}$$

which gives $b = 2b,\ c = 2c$ implying that b and c are 0. So,

$$C(M) = \left\{ \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \,\middle|\, a, d \in \mathbb{R} \setminus \{0\} \right\}.$$

While matrices in the center of N satisfy,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & a+b \\ c & c+d \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+c & b+d \\ c & d \end{pmatrix}$$

which gives $a = a + c,\ c + d = d,\ a + b = b + d$ implying that $c = 0$ and $a = d$. Therefore,

$$C(N) = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \,\middle|\, a, b \in \mathbb{R},\ a \neq 0 \right\}.$$

> *Note that in both cases some coefficients were required to be non-zero because to be members of the general linear group they must be invertible and so their determinant must be non-zero.*

(27) The center of the general linear group $GL_n(\mathbb{R})$ is the group of *scalar matrices* of the form cI for $c \in \mathbb{R}$, i.e. matrices of the form,

$$\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix}$$

in $GL_2(\mathbb{R})$. Note that, for *diagonal* matrices whose elements on the main diagonal are all non-zero but not-necessarily equal as in *scalar* matrices, multiplication is commutative with other diagonal matrices but not generally so with other matrices in the general linear group.

2.1.8 Products Groups and Quotient Groups

Definition. *If we take the cartesian product of two sets then:*

- *if the two sets are the underlying sets of two distinct groups then we have no way to combine them (as there is no common group operation) but we can take the pairing and define a component-wise multiplication over the pairs where each component is multiplied using that group's composition operation. In this way we create a new group over the pairs.*
- *if the two sets are subsets of a common group then there is a common group operation between them and so we can multiply them using this group operation. The result is another subset of the common group (not necessarily a subgroup).*

Both of these may at times be referred to as ***Product Groups*** *but the first one is more specifically referred to as a* ***Direct Product*** *and the second one may be referred to as a* ***Product Set****.*

2.1.8.1 Direct Products

Definition. *Let G, G' be two groups. The* ***direct product*** *is the set $G \times G'$ with component-wise multiplication using the group composition operation for the group corresponding to the component. Its order is the product of the orders of G and G'.*

Notation. The **direct product** of the two groups G, G' may be denoted by $G \times G'$ or GG'. In the case of Abelian groups the direct product may be referred to as the **direct sum** and denoted $G \oplus G'$.

So, if $a, b \in G$ and $a', b' \in G'$ then

- $(a, a'), (a, b'), (b, a'), (b, b') \in G \times G'$
- $(a, a')(b, b') = (ab, a'b')$

- the identity is $(1,1)$ and $(a,a')^{-1} = (a^{-1}, a'^{-1})$.

Definition. *The* ***projections*** *of a direct product $G \times G'$ are the maps p, p' such that,*

$$p(x,x') = x, \quad p'(x,x') = x'.$$

Proposition 33. *The* ***mapping property of direct products****: Let H be any group. The homomorphisms $\Phi : H \longmapsto G \times G'$ are in bijective correspondence to pairs (ϕ, ϕ') of homomorphisms*

$$\phi : H \longmapsto G, \qquad \phi' : H \longmapsto G'.$$

The kernel of Φ is the intersection $(ker\,\phi) \cap (ker\,\phi')$.

Proof. Given a pair of homomorphisms (ϕ, ϕ') we can define $\Phi(x) = (\phi(x), \phi'(x))$. Then this is homomorphic because,

$$\Phi(xy) = (\phi(xy), \phi'(xy)) = (\phi(x), \phi'(x))(\phi(y), \phi'(y)) = \Phi(x)\Phi(y).$$

Conversely, given such a Φ we can recover the pair of homomorphisms with the group projections as such (outer parentheses omitted for clarity),

$$\phi(x), \phi'(x) = p(\Phi(x)), p'(\Phi(x)).$$

Since the correspondence is invertible, it is a bijection.

Clearly, also,

$$\Phi(x) = (\phi(x), \phi'(x)) = (1,1) \iff (\phi(x) = 1) \wedge (\phi'(x) = 1)$$

so that $ker\,\Phi = (ker\,\phi) \cap (ker\,\phi')$. □

Proposition 34. *Let r, s be coprime integers. A cyclic group of order rs is isomorphic to the product of a cyclic group of order r and a cyclic group of order s.*

Proof. Let $C = \{1, x, x^2, \ldots, x^{rs-1}\}$, $C_1 = \{1, y, y^2, \ldots, y^{r-1}\}$, $C_2 = \{1, z, z^2, \ldots, z^{s-1}\}$ and define the map $\phi : C \longmapsto C_1 \times C_2$ as,

$$\phi(x^i) = (y^i, z^i).$$

Then ϕ is homomorphic because it is comprised of two homomorphisms (by the mapping proper Proposition 33),

$$\phi_1(x^i) = y^i \quad \text{and} \quad \phi_2(x^i) = z^i.$$

And ϕ is injective because,

$$\phi(x^i) = (1,1) \iff (y^i = 1) \wedge (z^i = 1) \iff (r \mid i) \wedge (s \mid i)$$

but r and s are coprime so this requires that $i = rs$ which is also the order of $x \in C$. So we have,

$$\phi(x^i) = (1,1) \iff x^i = x^{rs} = 1.$$

Therefore $ker\,\phi = \{1\}$ and, by Theorem 14, ϕ is injective.

Since ϕ is injective, its image has the same order as that of the domain C so we have,

$$|im\,\phi| = |C| = rs = |C \times C|$$

and ϕ is therefore surjective.

Therefore ϕ is a bijection and isomorphic. □

*Note that this is **only** the case for cyclic groups whose order is the product of two coprime numbers. For example, a cyclic group of order 4 is not isomorphic to a product of two cyclic groups of order 2 as every element in a product group $C_2 \times C_2$ has order 1 or 2. Whereas a cyclic group of order 4 has two elements of order 4 (the generating element and its inverse).*

Let $C_4 = \{1, x, x^2, x^3\}$, $C_2 = \{1, y\}$ and define the map $\phi : C_4 \longmapsto C_2 \times C_2$ as $\phi(x^i) = (y^i, y^i)$. Then,

$$\phi(x^i) = (1,1) \iff y^i = 1 \iff 2 \mid i$$

so that we have $ker\,\phi = \{1, x^2\}$ and so ϕ is not injective.

2.1.8.2 Product Sets

Definition. *Let A and B be subsets of the group G and denote the* ***product set*** *of A and B by*

$$AB = \{\, x \in G \mid x = ab \text{ for some } a \in A \quad \text{and} \quad b \in B \,\}.$$

Note that this notation is the same as one of the alternatives for the notation of the direct product so we need to be clear which is intended when we see this notation.

2.1.8.3 Relationship Between the Types of Product Groups

Proposition 35. *Let H and K be subgroups of G.*

(i) If $H \cap K = \{1\}$, the product map $p : H \times K \longmapsto G$ defined by $p(h,k) = hk$ is injective. Its image is the subset HK.

(ii) If either H or K is a normal subgroup of G, then the product sets HK and KH are equal and are subgroups of G.

(iii) If H and K are normal, $H \cap K = \{1\}$, and $HK = G$, then G is isomorphic to the direct product $H \times K$.

Proof. Proofs of each property are as follows.

(i) If we assume that $H \cap K = \{1\}$ then, for $h, h' \in H$, $k, k' \in K$,

$$p(h,k) = p(h',k') \iff hk = h'k' \iff (h')^{-1}h = k'k^{-1}$$

so that $(h')^{-1}h = k'k^{-1} \in H \cap K = \{1\}$ therefore,

$$(h')^{-1}h = 1 \iff h = h', \qquad k'k^{-1} = 1 \iff k' = k.$$

Therefore p is injective.

(ii) Assume w.l.o.g. that K is a normal subgroup. Then for all $k \in K, g \in G$, $g^{-1}kg \in K$ and, in particular, for $h \in H, h^{-1}kh \in K$. Therefore,

$$hk \in HK \implies h(h^{-1}kh) = kh \in HK$$

and conversely, using the fact that, $h^{-1} \in H \implies hkh^{-1} \in K$,

$$kh \in KH \implies (hkh^{-1})h = hk \in KH.$$

Therefore $HK = KH$ and this implies that HK is a subgroup because, for $h, h' \in H, k, k' \in K$,

- HK is closed because

  $$kh' \in KH = HK \implies kh' = h''k'' \in HK$$

  for some $h'' \in H$, $k'' \in K$, and so,

  $$hk, h'k' \in HK \implies (hk)(h'k') = h(kh')k' = h(h''k'')k' = (hh'')(k''k') \in HK.$$

- HK has inverses because for $hk \in HK$,

  $$h^{-1}k^{-1} \in HK \implies k^{-1}h^{-1} \in HK.$$

(iii) If $H \cap K = \{1\}$ then the product map p is injective and if $HK = G$ then $im\, p = G$ so p is surjective also and, therefore, is a bijection between $H \times K$ and G.

To show that p is a homomorphism between the direct product and the product set HK we need to show that,

$$p((h,k)(h',k')) = p((hh',kk')) = hh'kk' = hkh'k' = p((h,k))p((h',k'))$$

which will be true if $h'k = kh'$ which, in turn, will be the case if products in HK are commutative.
Now we have $H \cap K = \{1\}$ and so,

$$hk = kh \iff k^{-1}hk = h \iff k^{-1}hkh^{-1} = 1$$

implies that $H \cap K = \{k^{-1}hkh^{-1}\}$. So, if we can show that $k^{-1}hkh^{-1}$ is in both H and K then we have a homomorphism.
Since, in this case, both H and K are normal we have,

$$h, h^{-1} \in H,\ k \in K \implies k^{-1}hk \in H,\ hkh^{-1} \in K$$

which, by the group closure of H and K gives,

$$(k^{-1}hk)h \in H \quad \text{and} \quad k^{-1}(hkh^{-1}) \in K.$$

Therefore, if both H and K are normal subgroups and $H \cap K = \{1\}$, then $hk = kh$ for all $h \in H$, $k \in K$. This, in turn, means that the product map p is a homormorphism between the direct product $H \times K$ and the product set HK. Since p is also bijective, it is an isomorphism.

□

It is important to note that the product map of two subgroups $H \times K \longmapsto HK = G$ will not be a group homomorphism unless the two subgroups commute with each other.

2.1.8.4 Examples of Product Groups

(28) There is a group with subgroups of orders $1 \ldots 12$. It is a direct product of cyclic groups of orders

$$2 \times 2 \times 2 \times 3 \times 3 \times 5 \times 7 \times 11 = |G| = 27,720$$

so it looks like,

$$G = C_2 \times C_2 \times C_2 \times C_3 \times C_3 \times C_5 \times C_7 \times C_{11}.$$

2.1.8.5 Products of Cosets

It is possible to define a law of composition on the cosets of normal subgroups. This is because,

$$aH = Ha \implies aHbH = abHH = abH$$

so that we may define a law of composition such that $aH * bH = abH$ which closes over the set of cosets of H. The identity element of this composition is $eH = H$ and the inverse of the element aH is $a^{-1}H$.

*Note that this **only** applies to normal subgroups. The reason is that if H is not normal then there exists $h \in H$ and $a \in G$ such that $aha^{-1} \notin H$ which means that $S = aHa^{-1}H$ is not in any coset.*
This last claim can be proven if we observe – remembering that cosets partition the group – that S contains $a1a^{-1}1 = 1$ which means that it has to be in H (TODO: in the kernel?). However, S also contains $aha^{-1}1 = aha^{-1} \notin H$ so, since these are equivalence classes, S cannot be in H.

2.1.8.6 Quotient Groups

Definition. *Suppose N is a normal subgroup of a group G. Then the quotient group G/N is the set of cosets of N in G with the coset product. Its order is the index of N in G, $[G : N]$.*

Notation. Sometimes – when it is not necessary to specify the subgroup against which the cosets are being formed – the set of cosets in G is denoted $\overline{G}$ and a member coset aH is denoted $\overline{a} \in \overline{G}$.

Theorem 16. *Every normal subgroup of a group G is the kernel of a homomorphism.*

Proof. For any normal subgroup $N \leq G$, if we define the map,

$$\pi : G \longmapsto G/N$$

then π is homomorphic because $\pi(ab) = abN = aNbN = \pi(a)\pi(b)$.
Now, Proposition 26 tells us that

$$\pi(x) = 1N = N \iff x \in N$$

which implies that $ker\,\pi = N$. (We could also observe that the cosets are equivalence classes and the kernel is the equivalence class containing the identity. The coset that contains the identity is the original subgroup $N = 1N$.) □

Theorem 17. ***First Isomorphism Theorem****: Let $\phi : G \longmapsto G'$ be a surjective group homomorphism, and let $N = ker\,\phi$. Then G/N is isomorphic to G' by the map $\overline{\phi}$ which sends the coset $\overline{a} = aN$ to $\phi(a)$,*

$$\overline{\phi}(\overline{a}) = \phi(a).$$

Proof. The non-empty fibres of ϕ are the cosets aN as seen in the example 18. So, G/N can be thought of either as the cosets of the kernel of ϕ or as the non-empty fibres of ϕ. Then, $\overline{\phi}$ bijectively maps the cosets in G/N with the elements of the $im\,\phi$ and, because ϕ is surjective, we have $im\,\phi = G'$ so we have a bijection $\overline{\phi} : G/N \longmapsto G'$.
Also, the map $\overline{\phi}$ is homomorphic because coset multiplication is consistent with multiplication in the group,

$$\overline{\phi}(\overline{ab}) = \phi(ab) = \phi(a)\phi(b) = \overline{\phi}(\overline{a})\overline{\phi}(\overline{b}).$$

□

2.1.8.7 Examples of Quotient Groups

(29) Let $G = (\mathbb{Z}, +)$ and $H = \{\, 4n \mid n \in \mathbb{Z} \,\}$. Then the cosets of H are $\{\, z + 4n \mid z \in \mathbb{Z} \,\}$ and the product of two cosets,

$$(z_1 + H) + (z_2 + H) = (z_1 + z_2 + H).$$

(30) Let $G = (\mathbb{R}, +)$ and $H = \{\, 2n\pi \mid n \in \mathbb{Z} \,\}$. Then the cosets are the possible angles. This is an example of an infinite quotient group. The affine spaces in example 20 are another example of an infinite quotient group.

(31) In example 17 we saw that the modulus of complex numbers is a homomorphism from complex numbers to the reals. So, its kernel is the unit circle – the set of complex numbers of modulus 1. The cosets of the unit circle are the concentric circles,

$$C_r = \{\, z \mid |z| = r \,\}.$$

Applying the product of cosets gives us $C_r C_s = C_{rs}$ which works out because,

$$|(a + bi)(c + di)| = |(ac - bd) + (ad + bc)i|$$

$$
\begin{aligned}
&\iff && |(a+bi)(c+di)| = \sqrt{(ac-bd)^2+(ad+bc)^2} \\
&\iff && |(a+bi)(c+di)| = \sqrt{(a^2c^2+b^2d^2-2abcd)+(a^2d^2+b^2c^2+2abcd)} \\
&\iff && |(a+bi)(c+di)| = \sqrt{(a^2+b^2)(c^2+d^2)} \\
&\iff && |(a+bi)(c+di)| = \sqrt{(a^2+b^2)}\sqrt{(c^2+d^2)} \\
&\iff && |(a+bi)(c+di)| = |(a+bi)|\,|(c+di)|\,.
\end{aligned}
$$

TODO: Notes on Quotient Groups: Artin[81]

2.1.8.8 Modular Arithmetic

TODO: Describe modular arithmetic in terms of cosets: Artin[79] TODO: include in examples Abstract Maths ex. 14.9

2.2 Fields

2.2.1 Infinite Fields

Definition. *A **field** F is a set together with two laws of composition, addition and multiplication, satisfying the following axioms:*

(i) Addition makes F into an abelian group F^+ (or $(F, +)$). Its identity element is denoted 0.

(ii) Multiplication is associative and commutative and makes $(F/\{0\}, \times)$ into a group. Its identity element is denoted 1.

(iii) Distributive law: For all $a, b, c \in F$, $(a + b)c = ac + bc$.

The distributive law establishes a relationship between the two laws of composition such that, for $a \in \mathbb{F}$,

$$a + a = 1a + 1a = (1 + 1)a.$$

In so doing, it establishes a relationship between the two groups: the additive group and the multiplicative group.

2.2.1.1 Subfields of $\mathbb{C}$

Definition. *A field F is a **subfield of** $\mathbb{C}$ if the following properties hold:*

- *If $a, b \in F$, then $a + b \in F$.*
- *If $a \in F$, then $-a \in F$.*
- *If $a, b \in F$, then $ab \in F$.*
- *If $a \in F$ and $a \neq 0$, then $a^{-1} \in F$.*
- $1 \in F$.

Note that using the first, second and last of these axioms we can deduce that $1 - 1 = 0$ is an element of F.

*Also notice that addition on the field makes $(F, +)$ into an abelian group and multiplication makes $(F/\{0\}, \times)$ into an abelian group also. Conversely, any subset of F for which this is also true is a **subfield**.*

2.2.2 Finite Fields

TODO: Finite Fields (Artin[98])

2.2.3 Complex Numbers

Proposition 36. *For every $\alpha \in \mathbb{C}$, there exists a unique $\beta \in \mathbb{C}$ such that $\alpha + \beta = 0$.*

Proof. By contradiction: Say there are two such elements, β, γ such that,

$$\alpha + \beta = 0 = \alpha + \gamma$$

$$(\alpha + \beta) + \beta = (\alpha + \beta) + \gamma$$
$$0 + \beta = \beta = 0 + \gamma = \gamma \qquad \square$$

Proposition 37. *For every $\alpha \in \mathbb{C}$ with $\alpha \neq 0$, there exists a unique $\beta \in \mathbb{C}$ such that $\alpha\beta = 1$.*

Proof. By contradiction: Say there are two such elements, β, γ then,

$$\alpha\beta = 1 = \alpha\gamma$$
$$\beta = \frac{1}{\alpha} = \gamma \qquad \square$$

2.2.4 Complex Numbers Problems

Find all the roots of $x^3 = 1$ for $x \in \mathbb{C}$ complex numbers Since $x^3 - 1 = (x-1)(x^2 + x + 1)$, we have (via zero-factor theorem) possible roots from,

$$x - 1 = 0 \iff x = 1$$

$$x^2 + x + 1 = 0 \implies x = \frac{-1 \pm \sqrt{-3}}{2} = \frac{-1 \pm \sqrt{3}\,i}{2}$$

More generally,

$$(a + bi) + (a - bi) = 2a$$

and since also,

$$\left[\frac{-1 + \sqrt{3}\,i}{2}\right]^2 = \frac{-1 - \sqrt{3}\,i}{2}$$

as well as the reverse,

$$\left[\frac{-1 - \sqrt{3}\,i}{2}\right]^2 = \frac{-1 + \sqrt{3}\,i}{2}$$

this means that if $x = \frac{-1 \pm \sqrt{3}\,i}{2}$ then $x^2 + x$ is of the form $(a+bi)+(a-bi) = 2a$ and so we have that $x^2 + x = -1 \iff x^2 + x + 1 = 0$.

In addition,

$$(a + bi)(a - bi) = a^2 + b^2$$

which means that if $x = \frac{-1 \pm \sqrt{3}\,i}{2}$ then $x^3 = x^2 x$ is of the form $(a + bi)(a - bi) = a^2 + b^2$ so we have that $x^3 = \frac{-1}{2}^2 + \frac{\sqrt{3}}{2}^2 = \frac{1}{4} + \frac{3}{4} = 1$.

So we see that - allowing for complex x - the cubic polynomial $x^3 - 1$ has 3 roots as we should expect from the Fundamental Theorem of Algebra *(is this the correct interpretation of this?)*.

2.3 Matrices

2.3.1 Basic Algebra of Matrices

Definition. *Matrix* ***equality*** *is defined component-wise so that if $A = B$ then A and B must have the same dimension as well as equal values in each component.*

Definition. *An* ***identity*** *element e is defined as $ea = ae = a$.*

The definition of an identity element above is in any context (not just for matrices). For matrices this has certain consequences.

Proposition 38. *Identity matrices must be square*

Proof. For a matrix A and an identity matrix I, $AI = IA = A$ which means that AI, IA and A must all have the same dimensions. If A is of dimension $m \times n$ then I must have dimension $n \times m$ but then AI has dimension $m \times m$ while IA has dimension $n \times n$. We conclude that $m = n$ and both matrices are square. □

***If** A, B, C **are matrices s.t.** $AB = AC$**, can we, in general, conclude that** $B = C$**?**
The answer is no, as the following example shows:

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 \\ 3 & 5 \end{pmatrix}, \qquad C = \begin{pmatrix} 8 & 0 \\ -4 & 4 \end{pmatrix}$$

$$A = B = \begin{pmatrix} 0 & 0 \\ 4 & 4 \end{pmatrix}$$

This is because multiplication by A has no inverse (i.e. it's not a bijection and A^{-1} does not exist) as we can see by the fact that $|A| = 0$.
If** A, B, C **are matrices s.t.** $A+5B = A+5C$**, can we, in general, conclude that** $B = C$**?
The answer is yes because the matrix addition and scalar multiplication always have inverses. The inverse of $+A$ is $-A$ and the inverse of scalar multiplication by 5 is scalar multiplication by $\frac{1}{5}$. So we can say,

$$\begin{aligned} & A + 5B = A + 5C \\ \iff \quad & A + 5B - A = A + 5C - A \\ \iff \quad & 5B = 5C \\ \iff \quad & \left(\frac{1}{5}\right) 5B = \left(\frac{1}{5}\right) 5C \\ \iff \quad & B = C \end{aligned}$$

2.3.1.1 Matrix multiplication

Multiplication of matrices proceeds as a collection of dot-products of individual vectors. As a result, its properties are largely dependent on the properties of the dot-product (see: 81).

Matrix multiplication treats the two operand matrices as collections of vectors with the first matrix having the vectors as rows and the second having the vectors as columns.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}$$

This difference in orientation of the vectors in the two operands results in the multiplication not being commutative - the order matters. So, the first property of the dot-product is not preserved but the others are preserved (albeit with a slight modification for the last one).

$$\begin{aligned} \alpha \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} \alpha a & \alpha b \\ \alpha c & \alpha d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} &= \begin{bmatrix} \alpha(ae + bg) & \alpha(af + bh) \\ \alpha(ce + dg) & \alpha(cf + dh) \end{bmatrix} \\ &= \alpha \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix} \end{aligned}$$

$$\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix}\right)\begin{bmatrix} i & j \\ k & l \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}\begin{bmatrix} i & j \\ k & l \end{bmatrix}$$
$$= \begin{bmatrix} i(a+e)+k(b+f) & j(a+e)+l(b+f) \\ i(c+g)+k(d+h) & j(c+g)+l(d+h) \end{bmatrix}$$
$$= \begin{bmatrix} ia+kb & ja+lb \\ ic+kd & jc+ld \end{bmatrix} + \begin{bmatrix} ie+kf & je+lf \\ ig+kh & jg+lh \end{bmatrix}$$
$$= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} i & j \\ k & l \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix}\begin{bmatrix} i & j \\ k & l \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a^2+b^2 & ac+bd \\ ac+bd & c^2+d^2 \end{bmatrix}$$

So, to summarize:

If A, B, C are matrices and α is a scalar then,

- $\alpha AB = (\alpha A)B = A(\alpha B) = \alpha(AB)$
- $(A+B)C = C(A+B) = AC + BC$
- AA^T is a symmetric matrix with positive values along the diagonal

2.3.1.2 Block Matrix multiplication

In the description of matrix multiplication as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

$a, b, \ldots, h$ could also be blocks of matrices. In which case, the resulting calculations — $ae + bg$ etc. — refer to matrix multiplication ae where a and

e are treated as matrices and must have compatible dimensions. This can be seen as follows.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & \cdots \\ a_{21}b_{11} + a_{22}b_{21} & \cdots \end{bmatrix}$$

$$\begin{aligned} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} &= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & \cdots \\ \vdots & \end{bmatrix} \\ &= \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21}) + a_{13}b_{31} & \cdots \\ \vdots & \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} + \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} \begin{bmatrix} b_{31} & b_{32} \end{bmatrix}. \end{aligned}$$

So we have, for example, working in blocks,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{bmatrix},$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \implies AF = -BH,\, CE = -DG,\, AE{+}BG = CF{+}DH = I,$$

$$\begin{bmatrix} A \\ B \end{bmatrix} \begin{bmatrix} C & D \end{bmatrix} = \begin{bmatrix} I & E \\ F & G \end{bmatrix} \implies AC = I \iff C = A^{-1}.$$

2.3.1.3 Matrix transpose

Proposition 39. $(AB)^T = B^T A^T$

Proof. Denote the ith row of the matrix A as $A[i :]$ and the jth column of the matrix B as $B[: j]$ and a matrix whose components at (i, j) are the dot-products of the ith row of the matrix A with the jth column of the matrix B as $(\langle A[i :], B[: j]\rangle)$. Then,

$$(AB)^T = (\langle A[i :], B[: j]\rangle)^T = (\langle A[j :], B[: i]\rangle)$$
$$B^T A^T = (\langle B^T[i :], A^T[: j]\rangle) = (\langle B[: i], A[j :]\rangle)$$

So, by commutativity of dot-product, $(AB)^T = B^T A^T$. □

Proposition 40. $(A^T)^{-1} = (A^{-1})^T$

Proof.

$$\begin{aligned} & I = AA^{-1} = (AA^{-1})^T = (A^{-1})^T A^T \\ \iff \quad & I(A^T)^{-1} = (A^{-1})^T A^T (A^T)^{-1} \\ \iff \quad & (A^T)^{-1} = (A^{-1})^T. \end{aligned}$$ □

Proposition 41. $(A + B)^T = A^T + B^T$

Proof. In $A + B$ the (i, j) element is $A_{ij} + B_{ij}$ so the (i, j) element of $(A + B)^T$ is $A_{ji} + B_{ji}$ which clearly is also the (i, j) element of $A^T + B^T$. □

2.3.1.4 Matrix inverse

Definition. *Inverse property is: If there exists a matrix B such that $AB = BA = I$ then B is the **inverse** of A and A is the inverse of B.*

This definition is inherently bound up with the definition of the identity ($\exists$ a matrix I s.t. $AI = IA = A$) and both define the identity and inverse elements as commutatively producing their result under matrix multiplication. Since matrix multiplication is not, in general, commutative there is no guarantee that if $AB = I$ then $BA = I$. An example of this failing is,

$$A = \begin{bmatrix} 1 & 2 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$AB = \begin{bmatrix} 1 \end{bmatrix} = I_1, BA = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix} \neq I_2$$

But we could have guessed this because Proposition 38 showed that identity matrices must be square and its product with a matrix must be defined from both the left and the right, i.e. $IA = AI = A$ meaning that the matrix A must have the same dimensions as I. So, for non-square matrices, no identity can exist. If there is no identity, then the inverse is not defined either.

Proposition 42. *If the inverses of the matrices A and B both exist then so does the inverse of the product AB and it is equal to $B^{-1}A^{-1}$.*

Proof.

$$\begin{aligned} & (AB)(AB)^{-1} = I \\ \iff \quad & (A^{-1}A)B(AB)^{-1} = A^{-1}I \\ \iff \quad & (B^{-1}B)(AB)^{-1} = B^{-1}A^{-1} \\ \iff \quad & (AB)^{-1} = B^{-1}A^{-1} \end{aligned}$$

and since B^{-1} and A^{-1} both exist then their product exists. Furthermore, this holds for a product of any finite sequence of invertible matrices $A_1A_2\cdots A_n$ which can easily be shown by induction on the associative product. □

2.3.2 Basic properties of Matrices

Definition. *If a_{ij} is an entry of a matrix in the ith row and jth column then the **main diagonal** of the matrix is the collection of entries a_{ij} with $i = j$.*

The main diagonal is most often spoken of with respect to square matrices but the definition does not require that the matrix be square (wikipedia).

2.3.2.1 Trace

Definition. *The **trace** of a matrix is the sum of the diagonal entries.*

2.3.2.2 Symmetric Matrices

Definition. *A **symmetric** matrix is a matrix that is invariant under transposition.*

This obviously requires that the matrix be square and that the upper off-diagonal elements mirror the lower off-diagonal elements.

Proposition 43. *If A and B are symmetric matrices then $(BA)^T = AB$.*

Proof. Proposition 39 gives us the result that,

$$(BA)^T = A^T B^T$$

and the definition of symmetric matrices tells us that for symmetric A and B,

$$A^T = A \quad \text{and} \quad B^T = B.$$

Therefore $(BA)^T = AB$. □

2.3.2.3 Upper Triangular Matrices

An **upper triangular matrix** is also known as a **row echelon matrix**.

2.3.2.4 Reduced Row Echelon Form Matrices

Definition. *A **reduced row echelon form matrix** is an upper triangular matrix with the added conditions that the pivot values are all 1 and the other components in the same column as a pivot are all 0.*

Reading off the nullspace of a RREF matrix
Let

$$A = \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} I_2 & F \\ 0 & 0 \end{bmatrix}.$$

Then the nullspace of A is, for $t \in \mathbb{R}$

$$t \begin{bmatrix} -F \\ I_1 \end{bmatrix} = t \begin{bmatrix} -a \\ -b \\ 1 \end{bmatrix}.$$

So the general formula is

$$\begin{bmatrix} -F \\ I_k \end{bmatrix}$$

where k is the dimension of the nullspace (kernel) of A.
If the free variables are not next to each other then the blocks are broken up. For example for

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

the nullspace of A is

$$\begin{bmatrix} -a & -c \\ 1 & 0 \\ -b & -d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

2.3.2.5 Diagonal Matrices

Definition. *A **diagonal** matrix is a matrix whose entries outside the main diagonal are all zero.*

- *The definition of a diagonal matrix does **not** require that all the diagonal entries are nonzero so, for example, the zero matrix is a diagonal matrix.*
- *Diagonal matrices are **usually understood to be** a subset of symmetric matrices but this is not strictly necessary as the definition of the main diagonal permits an interpretation that includes rectangular matrices (see wikipedia).*

Proposition 44. *If $A = a_{ij}$, $B = b_{ij}$ are square diagonal matrices then multiplying them results in a square diagonal matrix $C = c_{ij}$ whose diagonal entries $c_{11}, \ldots, c_{nn}$ are the products of the corresponding diagonal entries in A and B. That's to say*

$$c_{11}, \ldots, c_{nn} = a_{11}b_{11}, \ldots, a_{nn}b_{nn}.$$

Proof. If $A = a_{ij}$, $B = b_{ij}$ are square diagonal matrices then,

$$AB = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & 0 & \cdots & 0 \\ 0 & b_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & \cdots & 0 \\ 0 & a_{22}b_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a_{nn}b_{nn} \end{bmatrix}.$$

□

Corollary 11. *Multiplication of square diagonal matrices is commutative.*

Proof. The multiplication described in Proposition 44 is commutative because the diagonal entries of the two matrices are the same whichever way around the multiplication is performed and the multiplication of the individual entries itself is commutative. □

Corollary 12. *Square diagonal matrices with nonzero diagonal entries are invertible.*

Proof. A result of the multiplication described in Proposition 44 is that, if A is a square diagonal matrix with nonzero diagonal entries, then we can obtain the identity matrix by multiplication with a matrix B whose diagonal elements are the reciprocal of the corresponding entries in A. So, if the jjth entry in A is a_{jj}, then the corresponding entry in B is $1/a_{jj}$. For example,

$$AB = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} 1/a_{11} & 0 & \cdots & 0 \\ 0 & 1/a_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1/a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Furthermore, by 11, this multiplication is commutative so we have,

$$AB = BA = I.$$

Therefore $B = A^{-1}$. Furthermore, this matrix always exists as long as the entries of the matrix are drawn from a field (so that they have multiplicative inverses) and the diagonal entries are nonzero. □

Proposition 45. *Square diagonal matrices with nonzero diagonal entries form an abelian group under multiplication.*

Proof. Let S be the set of $n \times n$ diagonal matrices with nonzero diagonal entries. Then,

- $I_n \in S$ so S is nonempty.
- By Proposition 44, if A, B are members of S, then AB is also a square diagonal matrix with nonzero diagonal entries and so is also a member of S.
- By 12 members of S are invertible and are members of S.

Therefore, $S \leq GL_n(\mathbb{F})$. □

2.3.3 Matrices as linear transformations

2.3.3.1 Multiplying a vector by a matrix on the left: $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{y}}$

Left multiplication of a matrix A of dimension $m \times n$ on a column vector $\vec{\boldsymbol{x}}$ of dimension $n \times 1$ transforms it to a column vector $\vec{\boldsymbol{y}}$ of dimension $m \times 1$.

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

This can be thought of a function from the space of n-dimensional vectors from which $\vec{\boldsymbol{x}}$ is drawn to the space of m-dimensional vectors in which $\vec{\boldsymbol{y}}$ resides. So, for real-valued vectors, the function would be a function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ such that,

$$f(x_1, \cdots, x_n) = \vec{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

Or else, this could be thought of as m n-ary functions of the form $f : \mathbb{R}^n \mapsto \mathbb{R}$,

$$\begin{aligned} f_1(x_1, \cdots, x_n) &= a_{11}x_1 + \cdots + a_{1n}x_n = y_1 \\ &\vdots \\ f_m(x_1, \cdots, x_n) &= a_{m1}x_1 + \cdots + a_{mn}x_n = y_m \end{aligned}$$

In this case, each row of the matrix is a real-valued function in n variables. Each of these functions is *homogenous linear* (a function of the form $a_1x_1 + \cdots + a_kx_k + c$ for scalars $a_1, \cdots, a_k, c$ and $c = 0$) and so the system of functions is called a *linear transformation*.

Example of $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{y}}$

(32) Consider the following system of equations (represented by an augmented matrix) for some constants a, b,

$$\left[\begin{array}{ccc|c} 1 & -1 & 2 & 4 \\ 3 & -1 & -1 & 0 \\ 1 & 1 & a & b \end{array}\right] \rightsquigarrow \left[\begin{array}{ccc|c} 1 & -1 & 2 & 4 \\ 0 & 1 & \frac{-7}{2} & -6 \\ 0 & 0 & a+5 & b+8 \end{array}\right]$$

where the second matrix represents a system of linear equations with the same solutions as the first. The second form tells us that:

- The system has **precisely one solution** if $a \neq -5$, so that we have an upper triangular matrix.
- The system has **infinitely many solutions** if $a = -5$ and $b = -8$, so that the bottom row is all zeroes and there is a free variable.
- The system has **no solutions** if $a = -5$ but $b \neq -8$, so that the bottom row of the matrix is zeroes but the corresponding component in the augmented column is non-zero; meaning that the system is inconsistent.

2.3.3.2 Multiplying a matrix of vectors by a matrix on the left: $AX = Y$

Looking at the matrix as a linear transformation from one co-ordinate space to another, consider $AX = Y$ where X is a matrix - which may be considered a collection of vectors - transformed by the matrix A into the matrix - or collection of vectors - Y.

2.3.3.3 Change of Co-ordinates

If a matrix A has columns comprised of the axes of a co-ordinate system, then $\vec{\boldsymbol{x}}$ defined against the standard basis is, in the other system, $A^{-1}\vec{\boldsymbol{x}}$. This is because the axes in A are defined relative to the standard basis axes. Therefore, if $\vec{\boldsymbol{x_A}}$ is a vector defined against the axes in A, then the same vector against the standard basis axes $\vec{\boldsymbol{x}}$ would be $A\vec{\boldsymbol{x_A}} = \vec{\boldsymbol{x}}$ and so $\vec{\boldsymbol{x_A}} = A^{-1}\vec{\boldsymbol{x}}$.

2.3.3.4 Types of Transformations

There are 3 basic types of transformation:

- **Rigid body** - preserves distances and angles.

  Examples: translation and rotation.

- **Conformal** - preserves angles.

  Examples: translation, rotation and uniform scaling.

- **Affine** - preserves parallelism.

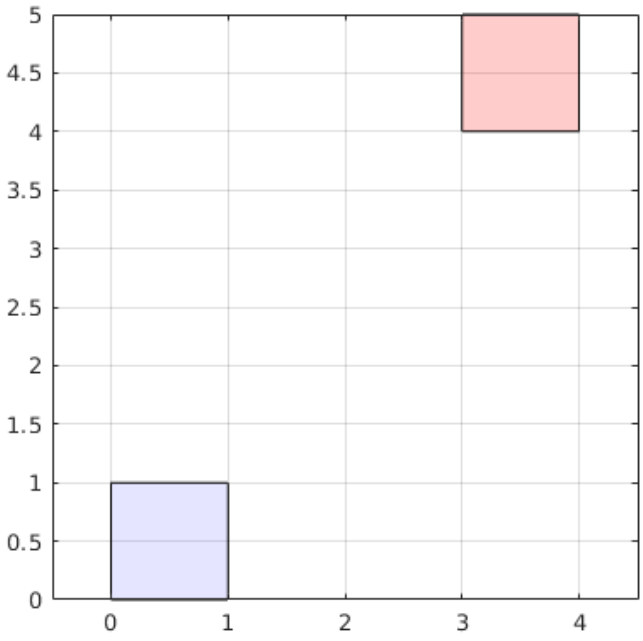
  Examples: translation, rotation, uniform and non-uniform scaling, shearing and reflection.

2.3.3.5 Rigid Body

Translation So as to perform the translation as multiplication by a transformation matrix we take the approach of homogeneous coordinates (see:https://en.wikipedia.org/wiki/Homogeneous_coordinates) so we form matrix with the identity in the first two columns and then a third column with the translation vector. Then, we add a row of ones to the vectors we will translate and the output vectors also have a 1 in the third row that is ignored.

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}, \; X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

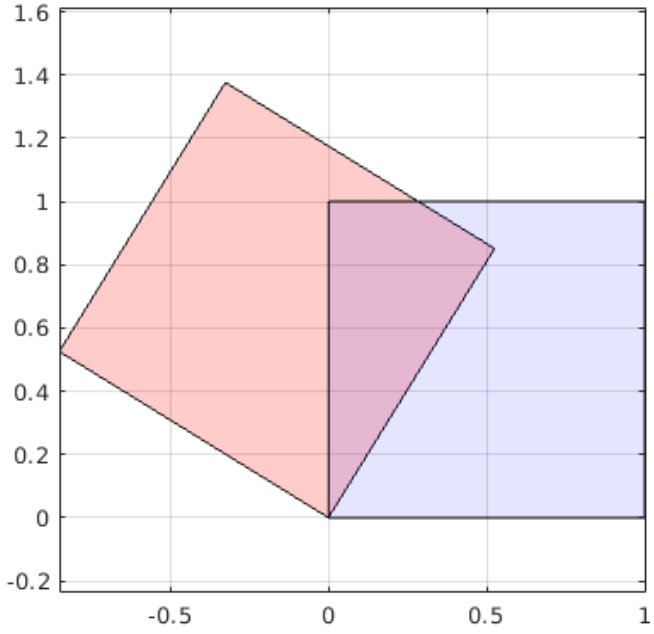
$$AX = Y = \begin{bmatrix} 3 & 4 & 4 & 3 \\ 4 & 4 & 5 & 5 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$



Rotation

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$AX = Y = \begin{bmatrix} 0 & 0.5253 & -0.3256 & -0.8509 \\ 0 & 0.8509 & 1.3762 & 0.5253 \end{bmatrix}$$

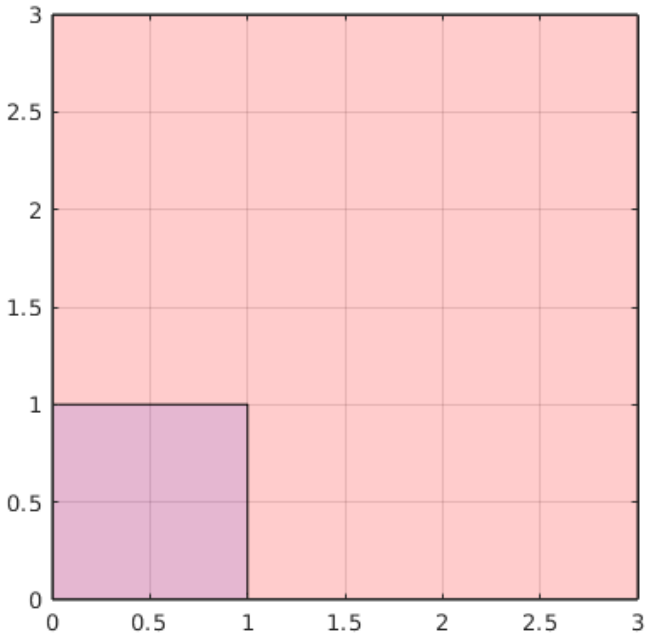


2.3.3.6 Conformal

Uniform Scaling is scaling by an equal amount in each dimension.

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \; X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$AX = Y = \begin{bmatrix} 0 & 3 & 3 & 0 \\ 0 & 0 & 3 & 3 \end{bmatrix}$$

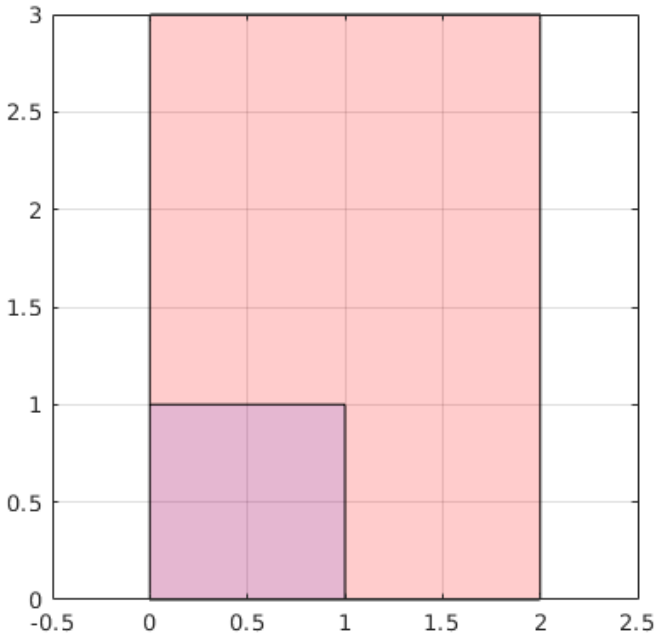


2.3.3.7 Affine

Non-uniform Scaling is scaling by different amounts in the different dimensions. (The example shown here preserves the angles but for other shapes, a triangle for example, the angles would not be preserved.)

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \; X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

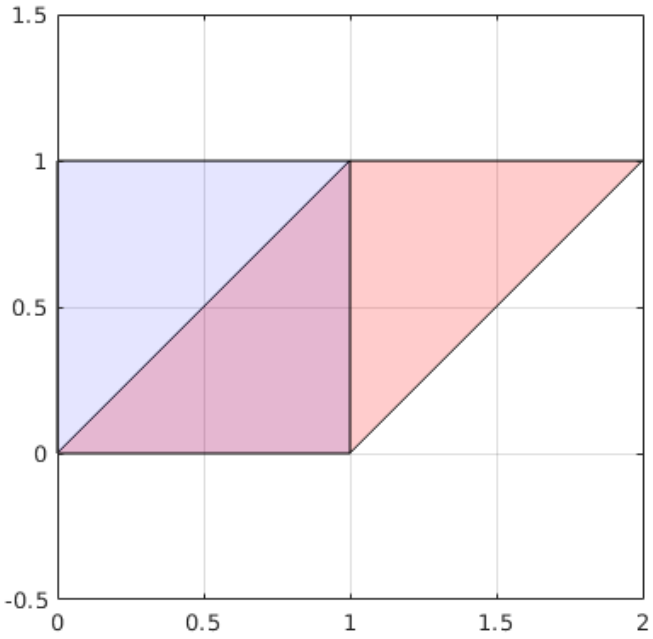
$$AX = Y = \begin{bmatrix} 0 & 2 & 2 & 0 \\ 0 & 0 & 3 & 3 \end{bmatrix}$$



Shearing

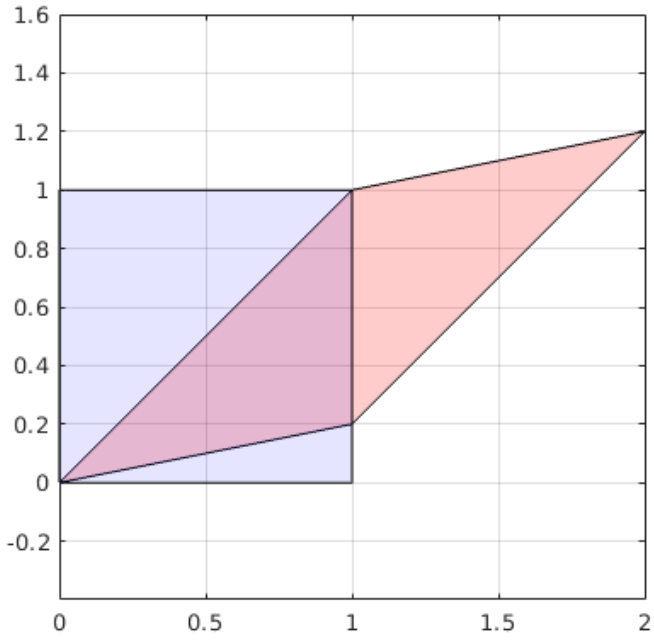
$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$AX = Y = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$



$$A = \begin{bmatrix} 1 & 1 \\ 0.2 & 1 \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

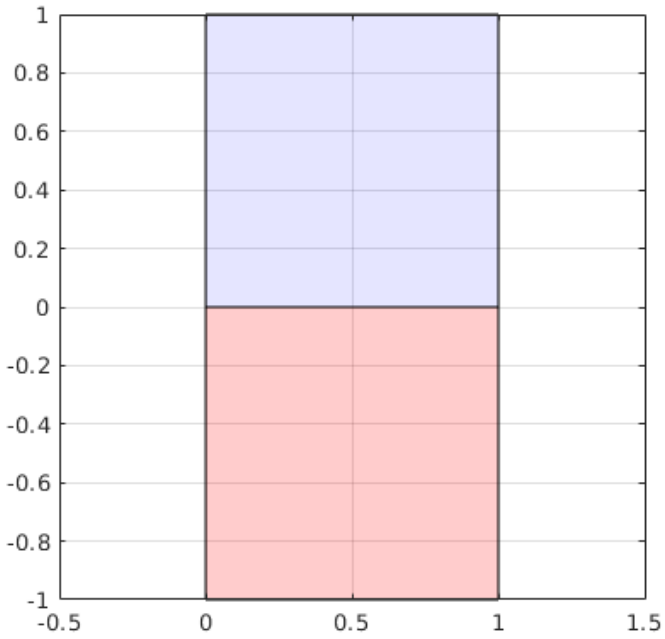
$$AX = Y = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 0 & 0.2 & 1.2 & 1 \end{bmatrix}$$



Reflection

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

$$AX = Y = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & -1 \end{bmatrix}$$


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.5
0
0.5
1
1.5

2.3.4 Elementary Matrices and Row Operations

Notation. The **matrix units** - matrices with a single non-zero component whose value is 1 are traditionally named $\boldsymbol{e_{ij}}$ where i, j is the matrix co-ordinate of the 1.

An arbitrary matrix $A = (a_{ij})$ may be expressed as a sum of such unit matrices as $A = a_{11}e_{11} + \cdots + a_{nn}e_{nn}$.

$$e_{ij} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & 1 & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

So matrix units can be used to analyse matrix addition but to analyse matrix multiplication some square matrices called **elementary matrices** are more useful.

Multiplying a matrix from the left (so doing row operations), there are 3 types of elementary matrix:
Adding rows: $\boldsymbol{I + ae_{ij}}$ ***for*** $\boldsymbol{i \neq j}$

$$\begin{bmatrix} 1 & & & \\ & \cdot & a & \\ & & \cdot & \\ & & & 1 \end{bmatrix}$$

This adds a times some row to another row.
Swapping rows: $\boldsymbol{I + e_{ij} + e_{ji} - e_{ii} - e_{jj}}$ ***for*** $\boldsymbol{i \neq j}$

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

This swaps the rows i and j.

Scalar-multiplying a row: $\boldsymbol{I + (c-1)e_{ii}}$ ***for*** $\boldsymbol{c \neq 0}$

$$\begin{bmatrix} 1 & & & \\ & \cdot & & \\ & & c & \\ & & & 1 \end{bmatrix}$$

This multiplies row i by c.

Proposition 46. *Elementary matrices are invertible and their inverses are also elementary matrices.*

Proof. Proceed by cases on the 3 elementary types of elementary matrices.

Case $\boldsymbol{I + ae_{ij}}$ If R_i is row i and R_j is row j, then this matrix performs $R_i + aR_j$. Clearly this can be "undone" by performing $R_i - aR_j$. So the matrix, $I - ae_{ij}$ is the inverse and clearly this is also an elementary matrix of the same type.

Case $\boldsymbol{I - e_{ii} - ejj + e_{ij} + e_{ji}}$ This matrix swaps 2 rows in a permutation that is its own inverse.

Case $\boldsymbol{I + (c-1)e_{ii}}$ This matrix performs cR_i and so it is "undone" by performing $c^{-1}R_i$ (which for a real-valued matrix would be $\left(\frac{1}{c}\right)R_i$) and this inverse matrix is also an elementary matrix of the same type. □

Proposition 47. *Suppose $AX = B$ and a series of elementary row operations on $[A \mid B]$ produces $[A' \mid B']$, then the solutions of $A'X = B'$ are the same as those of $AX = B$.*

Proof. First note that the series of elementary row operations is described as multiplication on the left by a series of elementary matrices say, $E_1, E_2, \cdots, E_n$ so that,

$$[A' \mid B'] = [(E_n \cdots E_2 E_1)A \mid (E_n \cdots E_2 E_1)B]$$

Now, let $(E_n \cdots E_2 E_1) = E$ and notice that, since each of the individual E_i is invertible the product of them is also invertible by Proposition 42 so,

$$A'X = B' \iff EAX = EB$$

and the existence of the inverse E^{-1} means that the law of cancellation is in effect so,

$$EAX = EB \iff AX = B$$
$$\therefore A'X = B' \iff AX = B. \qquad \square$$

Note that this is why Gaussian Elimination can employ row reduction to produce a linear system that is simpler to solve: elementary row operations on the matrix form linear combinations of the equations in a system of linear equations. This means that any vector that was a solution to the original system of equations (i.e. was a member of the intersection of the linear vector spaces represented by the equations) will also be a solution to the new system. Gaussian Elimination proceeds in this way until — if we obtain the row reduced echelon form — it has produced a system in which the linear vector spaces represented by the rows of the matrix are parallel to the co-ordinate system so that we can simply read off the co-ordinates of the solution (if it is a single point).

Proposition 48. *Let A be a square matrix. The following conditions are equivalent:*

- *A can be reduced to the identity by a sequence of elementary row operations.*
- *A is a product of elementary matrices.*
- *A is invertible.*
- *The system of homogeneous equations $AX = 0$ has only the trivial solution $X = 0$.*

Proof. If A can be reduced to the identity by a sequence of elementary row operations then,

$$(E_n \cdots E_2 E_1)A = I$$

and by Proposition 46 and Proposition 42 the matrix $(E_n \cdots E_2 E_1)$ is invertible so,

$$A = (E_n \cdots E_2 E_1)^{-1} I = (E_n \cdots E_2 E_1)^{-1} = {E_1}^{-1} {E_2}^{-1} \cdots {E_n}^{-1}$$

and, also by Proposition 46, A is a product of elementary matrices and is invertible.
Furthermore, if $AX = 0$ then $X = A^{-1}0 = 0$ - i.e. the only solution to $AX = 0$ is $X = 0$. □

2.3.4.1 The Reduced Row Echelon Form

The **reduced row echelon form** of a matrix is described in 2.3.2.4. It is obtained for a given matrix by a sequence of elementary row operations on the matrix according to the procedure known as Gaussian Elimination.

Example of Gaussian Elimination

(33) Let

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

Gaussian Elimination on A produces the reduced row echelon form (rref) matrix

$$A' = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}.$$

The rref matrix A' tells us that the third column of A is obtainable as a linear combination of the first two columns.

If we break up the matrix A into the blocks suggested by the rref

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

we see that

$$\begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}^{-1} = \frac{1}{6} \begin{bmatrix} 2 & -1 \\ 0 & 3 \end{bmatrix}$$

and

$$\frac{1}{6}\begin{bmatrix}2 & -1\\ 0 & 3\end{bmatrix}\begin{bmatrix}1\\ -4\end{bmatrix} = \begin{bmatrix}1\\ -2\end{bmatrix}$$

so that

$$\begin{bmatrix}\begin{bmatrix}3 & 1\\ 0 & 2\end{bmatrix}^{-1} & \begin{bmatrix}0\\ 0\end{bmatrix}\\ \begin{bmatrix}0 & 0\end{bmatrix} & \begin{bmatrix}0\end{bmatrix}\end{bmatrix}$$

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

Note that the matrix multiplication by blocks works out:

$$\begin{bmatrix}\begin{bmatrix}3 & 1\\ 0 & 2\end{bmatrix}^{-1} & \begin{bmatrix}0\\ 0\end{bmatrix}\\ \begin{bmatrix}0 & 0\end{bmatrix} & \begin{bmatrix}0\end{bmatrix}\end{bmatrix}\begin{bmatrix}\begin{bmatrix}3 & 1\\ 0 & 2\end{bmatrix} & \begin{bmatrix}1\\ -4\end{bmatrix}\\ \begin{bmatrix}3 & 2\end{bmatrix} & \begin{bmatrix}-1\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix} & \begin{bmatrix}1\\ -2\end{bmatrix}+\begin{bmatrix}0\\ 0\end{bmatrix}\\ \begin{bmatrix}0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0\end{bmatrix} & \begin{bmatrix}0\end{bmatrix}+\begin{bmatrix}0\end{bmatrix}\end{bmatrix}.$$

From this it's clear that the Gaussian Elimination algorithm amounts to left multiplication by the inverse of the invertible part of the matrix (in this case R).

Another way of looking at it arises from observing that

$$AA' = \begin{bmatrix}3 & 1 & 1\\ 0 & 2 & -4\\ 3 & 2 & -1\end{bmatrix}\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & -2\\ 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}3 & 1 & 1\\ 0 & 2 & -4\\ 3 & 2 & -1\end{bmatrix} = A.$$

So, the rref matrix of A is also a right identity of A. The final column of the rref matrix representing the free variable, is an alternative to the standard basis vector $\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}$ when taking linear combinations of the columns of A.

Using the rref to solve $A\vec{x} = \vec{b}$

If we are attempting to solve an equation of the form $A\vec{x} = \vec{b}$ then forming the augmented matrix and performing Gaussian Elimination left multiplies the vector $\vec{b}$ by the inverse of the invertible part of the matrix to obtain an element in the reverse image of $\vec{b}$.

For example, let $\vec{b} = \langle 3, 4, 5\rangle$ then, forming the augmented matrix $[A \,|\, \vec{b}]$ and performing row reduction to obtain the rref,

$$\left[\begin{array}{ccc|c} 3 & 1 & 1 & 3 \\ 0 & 2 & -4 & 4 \\ 3 & 2 & -1 & 5 \end{array}\right] \rightsquigarrow \left[\begin{array}{ccc|c} 1 & 0 & 1 & \frac{1}{3} \\ 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

from which we infer that

$$\vec{x} = \begin{bmatrix} \frac{1}{3} \\ 2 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

for $t \in \mathbb{F}$.

Note that:

- *If $\vec{b}$ were not in the span of the first two columns of A then elimination would show that the rows of A were inconsistent. In this case, for $\vec{b}$ to be in the span of A it had to be a point in the plane $x + \frac{y}{2} - z = 0$.*
- *If we were to swap the columns of A around we would find the same kernel but the particular solution would be different because it would be expressed w.r.t to different vectors.*

Nature of the Solution

The solution found by this method is, naturally, a coset of the kernel. In the language of linear algebra it is usually described as a particular solution + a general solution. Geometrically, in this case, the general solution (the kernel/nullspace) is a line and the particular solution is a point on the line. But, more interestingly, the particular solution found by this method is the

solution that effectively has the kernel zero-ed out. That's to say,

$$\begin{aligned}\vec{\boldsymbol{x}} &= \begin{bmatrix}\frac{1}{3}\\2\\0\end{bmatrix} = \begin{bmatrix}\frac{1}{3}\\2\\0\end{bmatrix} + \begin{bmatrix}0\\0\\0\end{bmatrix} \\ &= \begin{bmatrix}\frac{1}{3}\\2\\0\end{bmatrix} + 0\begin{bmatrix}-1\\2\\1\end{bmatrix}.\end{aligned}$$

For example, if we define $t' = t - 1$ then,

$$\vec{\boldsymbol{x}} = \begin{bmatrix}\frac{-2}{3}\\4\\1\end{bmatrix} + t'\begin{bmatrix}-1\\2\\1\end{bmatrix}$$

is also, equally, a solution. But the solution found by Gaussian Elimination will always consist of a vector (the particular solution) with zeroes in the components corresponding to any free-variable columns in the matrix.

2.3.5 Determinants

1×1
The determinant of a 1×1 matrix is just its unique component entry,

$$det \begin{bmatrix} a \end{bmatrix} = a$$

2×2
The determinant of a 2×2 matrix is given by the formula,

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

Returning to our example of a 2d operator:

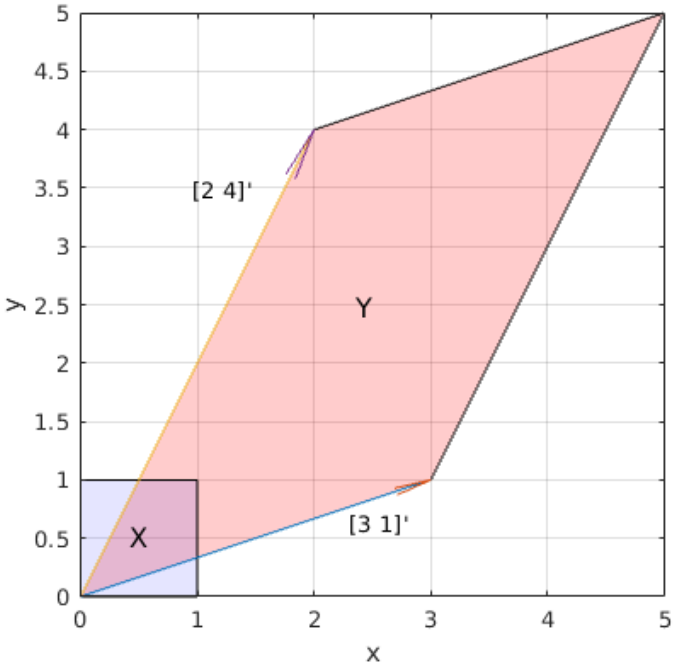
We see that $det\ A = 10$ and the parallelogram, Y, that is the image of the unit square, X, under the transformation represented by A has area,

$$\text{area } = b \cdot h = |\langle 3, 1 \rangle| \cdot |\langle 2 - 3, 4 - 1 \rangle| = \sqrt{10} \cdot \sqrt{10} = 10$$

And the determinant would be 0 in the case that the columns were proportional (representing co-linear vectors) and the determinant would be negative if the orientation of the output vectors were reversed w.r.t. the input vectors.
So, if we swap either the columns or the rows of the transformation matrix, A, the determinant comes out -10.
Swapping the columns:

$$A_c = \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

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

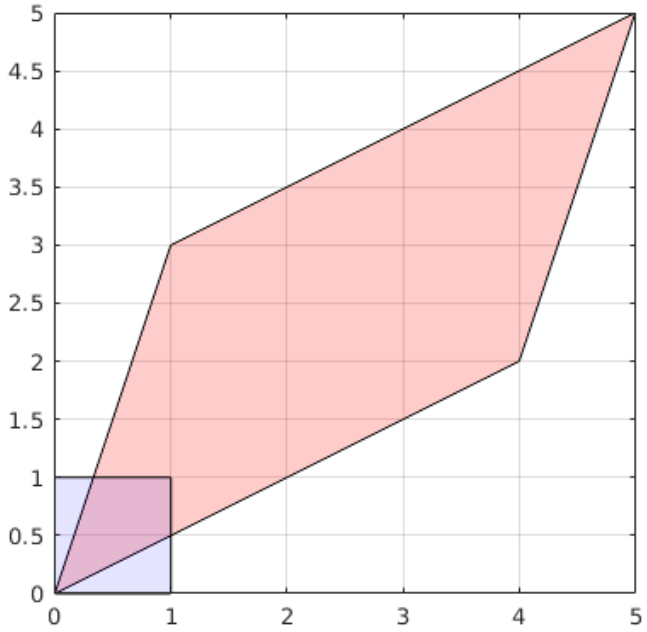


Note that the result looks exactly the same - it's just that now the x-vector $\langle 1, 0\rangle$, produces $\langle 4, 2\rangle$ and the y-vector, $\langle 0, 1\rangle$ produces $\langle 3, 1\rangle$.

Swapping the rows:

$$A_r = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \; X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

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



Note that if we swap **both** the columns and the rows then we get back to a transformation with determinant 10.

$$A_{rc} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

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

which produces the same parallelogram as the previous one but with columns reversed.

Summary So we find that,

$$\begin{bmatrix} 3 & 2 \\ 1 & 4 \end{bmatrix}, \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \text{ have determinant } > 0$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} \text{ have determinant } < 0$$

If the product of the components on the diagonal of the matrix is greater than the components on the bottom-left to top-right diagonal then the determinant is > 0, if the reverse is true then the determinant is < 0, and if they are equal then the determinant $= 0$.
Note that, for the determinant to be 0 in our example, we need something like $detA = (4 \times 3) - (4 \times 3)$ which, due to the commutativity of multiplication can be achieved by both,

$$\begin{bmatrix} 4 & 3 \\ 4 & 3 \end{bmatrix}, \begin{bmatrix} 4 & 4 \\ 3 & 3 \end{bmatrix}$$

but in both cases the columns are proportional and therefore co-linear.
$\boldsymbol{n \times n}$
The determinant of an $n \times n$ matrix is defined recursively as:

- if $n = 1$ then $det\, A = a_{11}$, i.e. the determinant is equal to the sole component.
- else if $n > 1$ then, defining A_{ij} as the matrix formed by leaving out the ith row and the jth column,

$$det\, A = a_{11} det\, A_{11} - a_{12} det\, A_{12} + \cdots \pm a_{1n} det\, A_{1n}$$

In the $n = 2$ case, each $det\, A_{ij}$ has $n = 1$ and so is simply equal to the sole component that is neither on the same row or column as the component a_{ij} that is multiplying it. This feature of the determinant calculation continues recursively for higher dimension matrices so that the calculation is always comprised of terms that are a product of components on each of the different columns and rows. In fact, it comprises the products of all such possible combinations of components.
For example, when $n = 2$ the only combinations are,

$$\{a_{11}, a_{22}\} \text{ and } \{a_{12}, a_{21}\}$$

so there are only 2 terms in the determinant calculation.
When $n = 3$ the possible combinations are,

$$\{a_{11}, a_{22}, a_{33}\},\ \{a_{11}, a_{32}, a_{23}\},$$

$$\{a_{12}, a_{21}, a_{33}\},\ \{a_{12}, a_{31}, a_{23}\},$$
$$\{a_{13}, a_{21}, a_{32}\},\ \{a_{13}, a_{31}, a_{22}\}$$

so there are 6 terms in the determinant calculation. Notice that each term is generated by a different permutation of the columns while holding the rows fixed in ascending order and that the sign of each term is governed by how many permutations the permutation of columns is away from ascending order, $1, 2, \cdots, n$.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22})$$
$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

2.3.5.1 Consequences

From this feature of the calculation we can see a number of the important properties of the determinant.

Proposition 49. *$det\, I = 1$*

Proof. Whatever the dimension of the identity matrix there will be only one combination of rows and columns that is the diagonal along which the 1s of the identity matrix reside. So, there will be a single term of the determinant calculation that is a product of 1s and all other terms will contain at 0s. In addition, the term that is along the diagonal has a positive sign in the determinant calculation. Therefore the result is 1. □

Proposition 50. *$det\, A$ is linear in the rows of the matrix*

Proof. If p and q are row vectors and we have matrices A_p, A_q, A_{pq} in which are present, respectively, the row vector p, the one q, and the row $p + q$, then linearity implies that $det\, A_{pq} = det\, A_p + det\, A_q$. This can be seen since every term of the determinant calculation of A_{pq} will contain one of the components in the row $p + q$. So each term of the calculation will take the form,

$$(p + q)a_{ij} \cdots a_{mn} = p(a_{ij} \cdots a_{mn}) + q(a_{ij} \cdots a_{mn})$$

The other implication of linearity is that - if a row is multiplied by a scalar, c, to produce A_c then $det\, A_c = c\, det\, A$. Using a similar reasoning to the

previous argument we have each term of the determinant taking the form,

$$c\, a_{ij} \cdots a_{mn}$$

which obviously results in the determinant being multiplied by c. □

Proposition 51. *If two columns are exchanged in the matrix then the determinant is multiplied by* -1

Proof. If columns p and q are exchanged then the components of p and q appear in terms with signs reversed. Since the components of p and q appear in every term of the determinant calculation, every term has the sign reversed. So the determinant is multiplied by -1. □

Proposition 52. *$det\, A = 0$ if there are two identical columns in the matrix*

Proof. If column p is identical to column q then we can swap columns p and q and we will have the same matrix so the determinant must also remain the same. But Proposition 51 proved that swapping two columns causes the determinant to be multiplied by -1. So, if A_{pq} is the matrix A after swapping the columns,

$$det\, A_{pq} = det\, A = -det\, A \iff det\, A = 0$$

□

Proposition 53. *Adding a multiple of one column to another leaves the determinant unchanged*

Proof. By combining Proposition 50 and Proposition 52 we find that if the columns of A are,

$$\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_p}}, \vec{\boldsymbol{x_q}}, \cdots, \vec{\boldsymbol{x_n}}$$

and the columns of A_c are,

$$\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_p}} + c\vec{\boldsymbol{x_q}}, \vec{\boldsymbol{x_q}}, \cdots, \vec{\boldsymbol{x_n}}$$

then,

$$\begin{aligned} det\, A_c &= det\,(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_p}}, \vec{\boldsymbol{x_q}}, \cdots, \vec{\boldsymbol{x_n}}) + c \cdot det\,(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_q}}, \vec{\boldsymbol{x_q}}, \cdots, \vec{\boldsymbol{x_n}}) \\ &= det\,(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_p}}, \vec{\boldsymbol{x_q}}, \cdots, \vec{\boldsymbol{x_n}}) + c \cdot 0 \\ &= det\, A \end{aligned}$$

□

2.3.5.2 Better formulation (from Rudin's Principles of Mathematical Analysis)

Let $a(i,j)$ be the component in the ith row and jth column of the matrix A and,

$$sign(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

$$s(j_1, \cdots, j_n) = \prod_{p<q} sign(j_q - j_p)$$

Then the determinant,

$$det\, A = \sum s(j_1, \cdots, j_n) a(1, j_1) \cdots a(n, j_n)$$

defined over all n-tuples of n distinct values, $j_1, \cdots, j_n$ with $1 \leq j_r \leq n$ (i.e, permutations of $[1, n] \subset \mathbb{N}$) with each term being produced by a different permutation.

From this we can see that,

- **The determinant of the identity matrix is 1**
  Every term of the determinant will contain at least one 0 apart from the term that traverses the main diagonal, which is all 1s. We can see that there is only one such term because the main diagonal has $i = j$ and so there is only one such $j_1, \cdots, j_n$ that satisfies this.

- **The determinant is linear in the rows or columns of the matrix, holding the others constant**
  If a column, j_r, is multiplied by a scalar α and another column, j_k, is added to it, then the resulting determinant takes the form,

$$\begin{aligned} & det\, A = \sum_i s(j_1, \cdots, j_n) a(i, j_1) \cdots (\alpha a(i, j_r) + a(i, j_k)) \cdots a(i, j_n) \\ \iff & det\, A = \alpha a(i, j_r) \sum_i s(j_1, \cdots, j_n) a(i, j_1) \cdots a(i, j_n) + \\ & \qquad a(i, j_k) \sum_i s(j_1, \cdots, j_n) a(i, j_1) \cdots a(i, j_n) \end{aligned}$$

- **If two columns are exchanged then the determinant is multiplied by -1**

If columns p and q are exchanged then this is equivalent to swapping j_p and j_q in the n-tuple so that $s(j_1, j_2, \cdots, j_n)$ changes sign and so the determinant is multiplied by -1.

- **If two columns are equal then the determinant will be 0**
  If two columns are the same then this is equivalent to a repetition of a value in the tuple $j_1, \cdots, j_n$ and so,

$$\exists\, p, q \text{ s.t. } sign(j_q - j_p) = 0 \implies s(j_1, \cdots, j_n) = 0$$

  which results in every term of the determinant being 0.
  This can also be proven by using the previous property that tells us that the determinant is multiplied by -1 when we exchange the identical columns but - since the columns are identical - the resultant matrix is the same - which means that the determinant remains unchanged. Therefore, the determinant must be 0.

Proposition 54. *If A and B are $n \times n$ matrices, then*

$$det\, BA = det\, A\, det\, B$$

Proof. Let the columns of A be the vectors, $\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}}$ so that for each column j,

$$\vec{\boldsymbol{x_j}} = \sum_i a(i,j)\vec{\boldsymbol{e_i}}$$

and define,

$$\Delta_B(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}}) = \Delta_B(A) = det\, BA$$

so that,

$$det\,(B\vec{\boldsymbol{x_1}}, B\vec{\boldsymbol{x_2}}, \cdots, B\vec{\boldsymbol{x_n}}) = \Delta_B(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}})$$

Since $B\vec{\boldsymbol{x_j}}$ is linear in $\vec{\boldsymbol{x_j}}$, $\Delta_B(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}})$ is linear in each $\vec{\boldsymbol{x_j}}$ and so,

$$\begin{aligned}
\Delta_B(\vec{\boldsymbol{x_1}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}}) &= \Delta_B(\sum_i a(i,1)\vec{\boldsymbol{e_i}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}}) \\
&= \sum_i a(i,1)\Delta_B(\vec{\boldsymbol{e_i}}, \vec{\boldsymbol{x_2}}, \cdots, \vec{\boldsymbol{x_n}}) \\
&= \sum_{i_1} a(i_1,1) \sum_{i_2} a(i_2,2) \cdots \sum_{i_n} a(i_n,n)\, \Delta_B(\vec{\boldsymbol{e_{i_1}}}, \vec{\boldsymbol{e_{i_2}}}, \cdots, \vec{\boldsymbol{e_{i_n}}}) \\
&= \sum a(i_1,1)a(i_2,2)\cdots a(i_n,n)\, \Delta_B(\vec{\boldsymbol{e_{i_1}}}, \vec{\boldsymbol{e_{i_2}}}, \cdots, \vec{\boldsymbol{e_{i_n}}})
\end{aligned}$$

the sum being extended over all n-tuples, $(i_1, \cdots, i_n)$ such that $1 \leq i_j \leq n$. Also, by referring again to the properties of the determinant we see that,

$$\Delta_B(\vec{\boldsymbol{e_{i_1}}}, \vec{\boldsymbol{e_{i_2}}}, \cdots, \vec{\boldsymbol{e_{i_n}}}) = t(i_1, i_2, \cdots, i_n)\Delta_B(\vec{\boldsymbol{e_1}}, \vec{\boldsymbol{e_2}}, \cdots, \vec{\boldsymbol{e_n}})$$

where $t(i_1, i_2, \cdots, i_n) = 1, 0, -1$ similar to the function s previously. So, we end up with,

$$det\, BA = \sum a(i_1, 1)a(i_2, 2) \cdots a(i_n, n)t(i_1, i_2, \cdots, i_n)\, det\, B = det\, A\, det\, B$$

□

Proposition 55. *A linear operator A on $\mathbb{R}^n$ is invertible if and only if $det\, A \neq 0$*

Proof. If A is invertible then, $AA^{-1} = I$ and, using Proposition 49 and Proposition 54, we have,

$$det\, AA^{-1} = det\, A \cdot det\, A^{-1} = 1$$

so $det\, A$ cannot be 0.

Furthermore, if the columns of A are not independent then there is some linear combination of the columns that produces $\vec{\mathbf{0}}$. Since, by Proposition 53 we know that adding multiples of columns to other columns leaves the determinant unchanged, this means that the determinant is equal to the determinant of a matrix with $\vec{\mathbf{0}}$ as a column. Such a matrix has determinant 0, so the determinant of A is also 0. □

Corollary 13. *For invertible matrices,*

$$det\, A^{-1} = \frac{1}{det\, A}$$

Corollary 14. *The determinant is the only function that has the described properties.*

Proof. Every matrix, A, can be transformed by multiplication by elementary matrices to a row-reduced form, R, which is either the identity matrix - in the case that A is invertible - or a matrix with the last row zeroes - in the case where A is not invertible. So, the determinant of the row-reduced matrix, R, is either 1 or 0. Meanwhile, the determinants of the elementary matrices are:

- Add multiple of row to another row - determinant is 1 because this operation maintains the determinant of the identity.
- Swap two rows - determinant is -1 - determinant is -1 because this operation multiplies the determinant of the identity by -1.
- Multiply a row by some scalar c - determinant is c because this operation multiplies the determinant of the identity by c.

So, we have,

$$R = E_1E_2 \cdots E_nA \implies det\, R = det\, E_1E_2 \cdots E_n \cdot det\, A$$

where $det\, E_1E_2 \cdots E_n$ is a known, non-zero quantity - say d_e. Since the determinant of R is either 0 or 1 this leaves the determinant of A being either 0 or $\frac{1}{d_e}$.

So, the value of the determinant of an arbitrary matrix, A, is wholly determined by the properties described. □

Proposition 56. *The determinant of any square matrix is equal to that of its transpose. That's to say, for an arbitrary square matrix A,*

$$det\, A^T = det\, A.$$

Proof. The determinant formula from 2.3.5.2 gives us:

$$det\, A = \sum s(j_1, \cdots, j_n)a(1, j_1) \cdots a(n, j_n)$$

where $a(1, j)$ is the (i, j)th element of the matrix A and the summation is over all n-tuples of n distinct values, $j_1, \cdots, j_n$ with $1 \leq j_r \leq n$. So we can also deduce that,

$$det\, A^T = \sum s(j_1, \cdots, j_n)a(j_1, 1) \cdots a(j_n, n).$$

Clearly for the identity permutation $j_1, \ldots, j_n = 1, \ldots, n$ the term of the summation generated is the same: in both cases it is the product of the elements along the main diagonal $a(1, 1) \cdots a(n, n)$. On the other hand, if we take a minimal permutation, for example, where $j_1 = 2$, $j_2 = 1$ so that the

first two elements are swapped, then we see that the term of the summation generated in the determinant of A is

$$-a(1,2)a(2,1)a(3,j_3)\cdots a(n,j_n)$$

while the term of the summation generated in the determinant of A^T is

$$-a(2,1)a(1,2)a(3,j_3)\cdots a(n,j_n)$$

so that the same permuatation in the n-tuple $j_1,\cdots,j_n$ produces the exact same terms of the summation in the determinant of A and of A^T.
It's easy to see in fact, that this will happen with any permutation as the permutation of the $j_1,\cdots,j_n$, in the determinant of A, permutes the column indices while selecting in order from the rows whereas, in the determinant of A^T, it permutes the row indices while selecting from the columns in order. So, the net result is the same in both cases. Since each permutation of the n-tuple produces equal terms of the summation, the complete summation produced by all the permutations will be the same.
Therefore $det\, A^T = det\, A$. □

2.3.6 Permutation Matrices

Definition. *A permutation p is a bijective map from a set S to itself. If a matrix P is the matrix associated with a permutation p then:*

- *the jth column of the matrix is the basis vector $e_{p(j)}$,*
- *P is a sum of the matrix units, $P = e_{p(1)1} + e_{p(2)2} + \cdots + e_{p(n)n} = \sum_j e_{p(j)j}$.*

Proposition 57. *If P, Q are permutation matrices associated with the permutations p, q then the matrix that corresponds to the permutation $p \circ q$ is PQ*

Proof. $pq(i) = p(q(i))$ and $PQX = P(QX)$ □

Proposition 58. *A permutation matrix P is invertible and its inverse is the transpose, $P^{-1} = P^T$*

Proof. A left-multiplying permutation matrix for a permutation, p, maps each row from the input matrix using a column j in the permutation matrix, to the output row, $p(j)$. Since the permutation, by definition, is bijective, we know that this mapping is one-to-one and invertible. If we transpose the matrix P to P^T, swapping rows and columns in the permutation matrix, then the new matrix, P^T maps input rows $p(j)$ into output rows j which is clearly the inverse permutation. □

Since a permutation matrix is a the result of permuting the rows of the identity matrix, clearly, its determinant is ± 1. A permutation is referred to as *odd* or *even* depending on whether its determinant is -1 or 1 respectively. Its determinant is called the *sign of the permutation*,

$$sign\, p = det\, p = \pm 1$$

The determinant of an arbitrary $n \times n$ matrix can be described as,

$$det\, A = \sum_p \left[det \sum_j a_{p(j)j} e_{p(j)j} \right]$$

$$
\begin{aligned}
&= \sum_p \left[(a_{p(1)1} \cdots a_{p(n)n}) \cdot \ det \sum_j e_{p(j)j} \right] \\
&= \sum_p \Big[(a_{p(1)1} \cdots a_{p(n)n}) \cdot \ det\, P \Big] \\
&= \sum_p \Big[(sign\, p)(a_{p(1)1} \cdots a_{p(n)n}) \Big]
\end{aligned}
$$

This is the same formula as earlier and is known as the *complete expansion* of the determinant.

2.3.7 Cramer's Rule

Expansion by minors on the jth column:

$$det\, A = (-1)^{j+1} a_{1j}\, det\, A_{1j} + (-1)^{j+2} a_{2j}\, det\, A_{2j} + \cdots + (-1)^{j+n} a_{nj}\, det\, A_{nj}$$

Expansion by minors on the ith row:

$$det\, A = (-1)^{i+1} a_{i1}\, det\, A_{i1} + (-1)^{i+2} a_{i2}\, det\, A_{i2} + \cdots + (-1)^{i+n} a_{in}\, det\, A_{in}$$

Definition. *If we form a matrix with elements* $\alpha_{ij} = (-1)^{i+j}\, det\, A_{ij}$ *and then transpose it we get the* ***adjoint matrix****.*

Notation. The adjoint of A is denoted $adj\, A$.

Following we use $[x]$ to denote a matrix as distinguished from a scalar.

Let $d = det\, A$. Then, if we multiply the adjoint matrix of $[A]$ by $[A]$ we get,

$$[adj\, A][A] = \begin{bmatrix} d & & & & & \\ & d & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & d \end{bmatrix}$$

The off-diagonal elements come out zero because they involve a determinant calculation that involves the same row (or column) repeated and so those determinants are zero.

Theorem 18.

$$[adj\, A][A] = (det\, A)[I] = [A][adj\, A]$$

Corollary 15.

$$\frac{1}{det\, A}[adj\, A][A] = [I] \iff [A^{-1}] = \frac{1}{det\, A}[adj\, A]$$

This formulation of the inverse of a matrix can be used to write the solution to a system of linear equations (reverting to the normal notation) $AX = B$ as, multiplying on the left by A^{-1},

$$X = A^{-1}B = \frac{1}{det\, A}(adj\, A)B$$

so that X is a vector whose components, x_j, are expressed as,

$$\begin{aligned} x_j &= \frac{1}{det\, A}(b_1\alpha_{1j} + \cdots + b_n\alpha_{nj}) \\ &= \frac{1}{det\, A}(b_1(-1)^{1+j}\, det\, A_{1j} + \cdots + b_n(-1)^{n+j}\, det\, A_{nj}) \end{aligned}$$

which is the expansion by minors of A on the jth column but with the components a_{ij} of A replaced with the components of the vector B, divided by the determinant of A.

2.3.8 Linear Algebra of Polynomials

2.3.8.1 Uniqueness of Polynomials

Proposition 59. *Three points uniquely identify a quadratic polynomial.*

Proof. Assume that there are two distinct quadratic polynomials $p(x)$, $q(x)$ that share 3 points. That is,

$$p(x_1) = q(x_1),\ p(x_2) = q(x_2),\ p(x_3) = q(x_3).$$

Then we can define another quadratic polynomial $r(x) = p(x) - q(x)$ with the property that,

$$r(x_1) = p(x_1) - q(x_1) = 0 = r(x_2) = r(x_3).$$

In other words, $r(x)$ has 3 roots: x_1, x_2, x_3. But $r(x)$ is a quadratic polynomial and has a maximum of 2 roots. Therefore there is non such (nonzero) polynomial. □

2.3.8.2 Uniqueness of Coefficients of Polynomials

TODO: is this the best place for this?

Theorem 19. *If a polynomial is identically zero (i.e. the zero function), then all coefficients are 0.*

Proof. An obvious way to prove this is to begin by assuming that we have a degree-m polynomial,

$$p(x) = a_0 + a_1 x + \cdots + a_m x^m$$

where any coefficients a_i with $i < m$ may be zero. Then we can reason that if $p(x)$ is zero for all $x \in \mathbb{R}$ then its derivative $p'(x)$ must also be identically zero. In this way we can extend the implication to lower and lower degree polynomials by differentiation until $p^{(m)}$ is a degree zero polynomial,

$$p^{(m)}(x) = m!\, a_m = 0 \implies a_m = 0.$$

But $a_m = 0$ contradicts the hypothesis that $p(x)$ is a degree-m polynomial.

Another (better?) proof is given in Linear Algebra Done Right as follows.

Let

$$x = \frac{|a_0| + |a_1| + \cdots + |a_{m-1}|}{|a_m|} + 1$$

so that $x \geq 1$ and so $|a_0| \leq |a_0|\, x$. Then, using the triangle inequality,

$$|a_0 + a_1 x| \leq |a_0| + |a_1 x| \leq |a_0 x| + |a_1 x| = (|a_0| + |a_1|)x.$$

Using induction with an inductive step,

$$\begin{aligned}
& \left|a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}\right| \leq (|a_0| + \cdots + |a_{n-1}|)x^{n-1} \\
\implies \qquad & |a_0 + a_1 x + \cdots + a_n x^n| \leq \left|a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}\right| + |a_n x^n| \\
& \qquad\qquad \leq (|a_0| + \cdots + |a_{n-1}|)x^{n-1} + |a_n|\, x^n \\
& \qquad\qquad \leq (|a_0| + \cdots + |a_{n-1}| + |a_n|)x^n
\end{aligned}$$

we can deduce that

$$\left|a_0 + a_1 x + \cdots + a_{m-1}x^{m-1}\right| \leq (|a_0| + \cdots + |a_{m-1}|)x^{m-1}.$$

Now,

$$\begin{aligned}
|a_m x^m| = |a_m|\, x^m &= (|a_m|\, x)x^{m-1} \\
&= \left[|a_m| \left(\frac{|a_0| + |a_1| + \cdots + |a_{m-1}|}{|a_m|} + 1\right)\right] x^{m-1} \\
&= (|a_0| + |a_1| + \cdots + |a_{m-1}| + |a_m|)\, x^{m-1}.
\end{aligned}$$

Therefore, combining with the previous result, we have,

$$\left|a_0 + a_1 x + \cdots + a_{m-1}x^{m-1}\right| \leq (|a_0| + \cdots + |a_{m-1}|)x^{m-1} < |a_m x^m|\,.$$

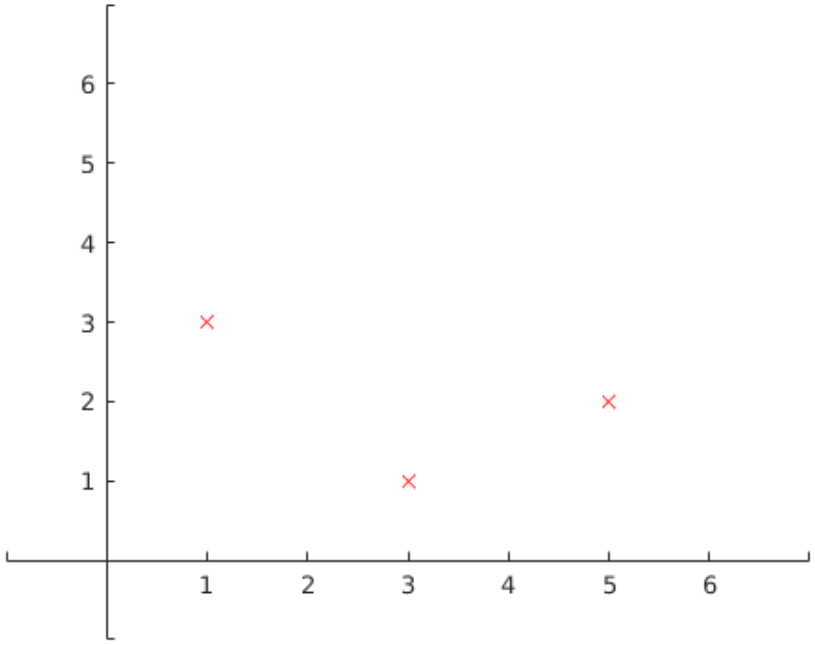
We can now reason that,

$$\begin{aligned}
& \left|a_0 + a_1 x + \cdots + a_{m-1}x^{m-1}\right| < |a_m x^m| \\
\implies \qquad & a_0 + a_1 x + \cdots + a_{m-1}x^{m-1} \neq -a_m x^m \\
\iff \qquad & a_0 + a_1 x + \cdots + a_{m-1}x^{m-1} + a_m x^m \neq 0.
\end{aligned}$$

□

2.3.8.3 Lagrange Polynomials

If we look for quadratic polynomials, $p(x)$, that pass throught the 3 points (1, 3), (3, 1) and (5, 2):



Then the first has roots at $x = 1, 3$ and passes through the point (5, 2). So, we have:

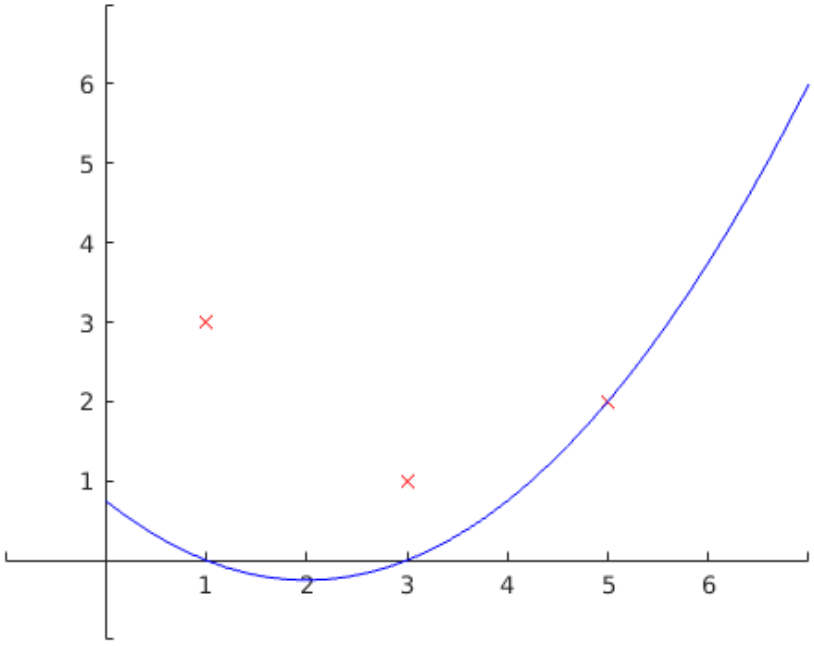
$$p(1) = p(3) = 0, p(5) = 2$$

meaning that $(x - 1)$ and $(x - 3)$ are factors. Therefore,

$$\begin{aligned} p(x) &= \alpha(x-1)(x-3) \\ &= \alpha(x^2 - 4x + 3) \end{aligned}$$

$$\begin{array}{lcl} & p(5) & = 2 \\ \implies & \alpha(5^2 - 4(5) + 3) & = 2 \\ \iff & 8\alpha & = 2 \\ \iff & \alpha & = \frac{1}{4} \end{array}$$

$$\therefore p(x) = \frac{1}{4}(x^2 - 4x + 3)$$

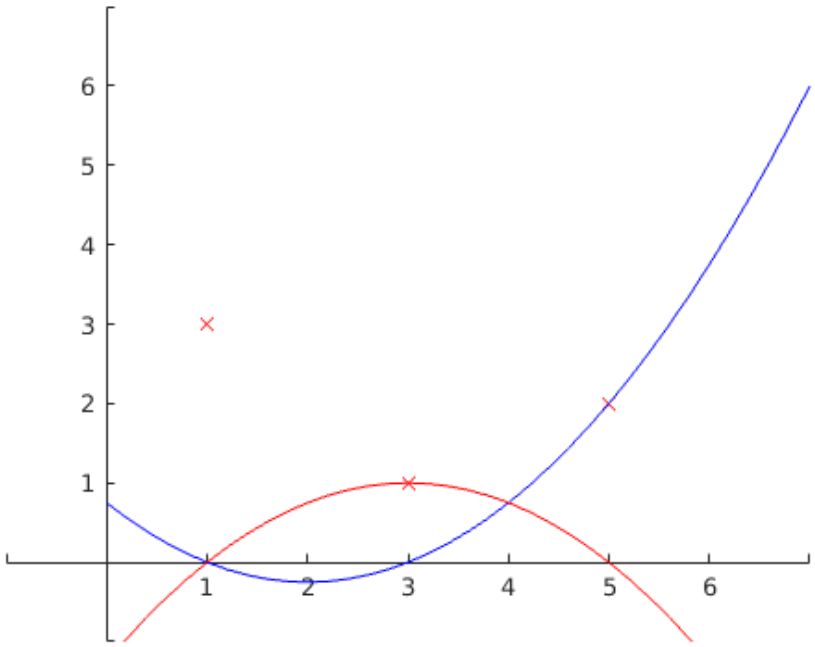


The second has roots at $x = 1, 5$ and passes throught the point (3, 1):

$$\begin{array}{rl} p(x) & = \alpha(x - 1)(x - 5) \\ & = \alpha(x^2 - 6x + 5) \end{array}$$

$$\begin{array}{rcl} & p(3) & = 1 \\ \implies & \alpha(3^2 - 6(3) + 5) & = 1 \\ \iff & -4\alpha & = 1 \\ \iff & \alpha & = -\frac{1}{4} \end{array}$$

$$\therefore p(x) = -\frac{1}{4}(x^2 - 6x + 5)$$

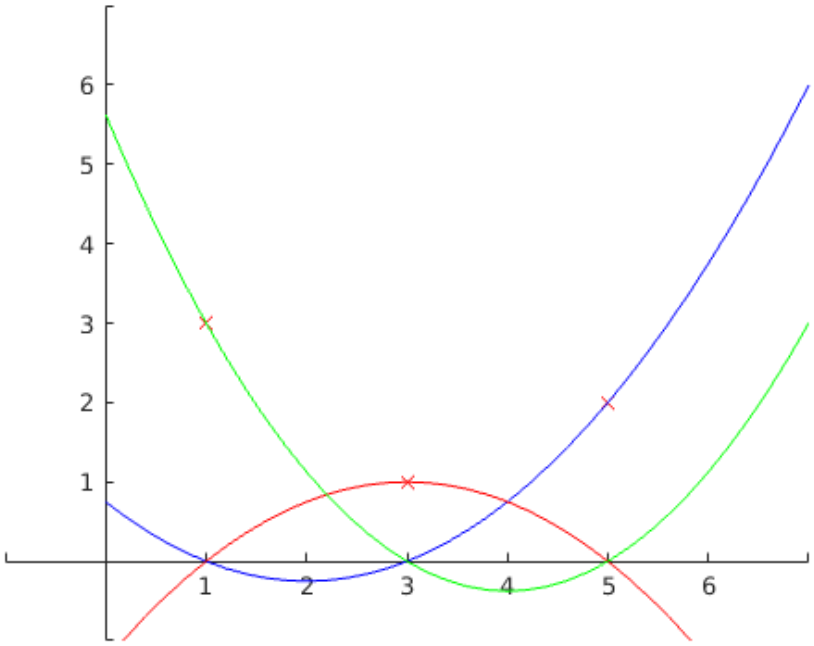


The third has roots at $x = 3, 5$ and passes through the point (1, 3):

$$\begin{array}{rl} p(x) & = \alpha(x - 3)(x - 5) \\ & = \alpha(x^2 - 8x + 15) \end{array}$$

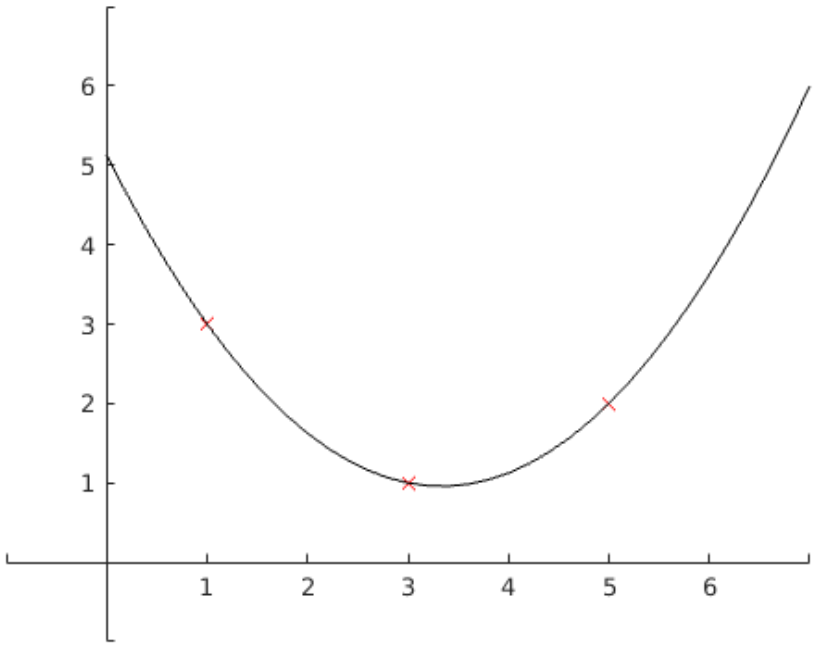
$$\begin{array}{lcl} & p(1) & = 3 \\ \implies & \alpha(1^2 - 8(1) + 15) & = 3 \\ \iff & 8\alpha & = 3 \\ \iff & \alpha & = \frac{3}{8} \end{array}$$

$$\therefore p(x) = \frac{3}{8}(x^2 - 8x + 15)$$



Adding them together we get,

$$\begin{aligned} & \tfrac{1}{4}(x^2 - 4x + 3) - \tfrac{1}{4}(x^2 - 6x + 5) + \tfrac{3}{8}(x^2 - 8x + 15) \\ = & (\tfrac{1}{4} - \tfrac{1}{4} + \tfrac{3}{8})x^2 + (-1 + \tfrac{3}{2} - 3)x + (\tfrac{3}{4} - \tfrac{5}{4} + \tfrac{45}{8}) \\ = & \ \tfrac{3}{8}x^2 - \tfrac{5}{2}x + \tfrac{41}{8} \end{aligned}$$



2.3.8.4 Matrix approach

We are looking for the unique quadratic polynomial $p(x) = \alpha_1 x^2 + \alpha_2 x + \alpha_3$ that satisfies

$$p(1) = 3,\ p(3) = 1,\ p(5) = 2$$

so this gives us the simultaneous equations:

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 &= 3 \\ 9\alpha_1 + 3\alpha_2 + \alpha_3 &= 1 \\ 25\alpha_1 + 5\alpha_2 + \alpha_3 &= 2 \end{aligned}$$

which can be expressed as a matrix equation,

$$\begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}.$$

Now,

$$\begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1/8 & -1/4 & 1/8 \\ -1 & 3/2 & -1/2 \\ 15/8 & -5/4 & 3/8 \end{bmatrix}$$

so we have,

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 1/8 & -1/4 & 1/8 \\ -1 & 3/2 & -1/2 \\ 15/8 & -5/4 & 3/8 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3/8 - 1/4 + 1/4 \\ -3 + 3/2 - 1 \\ 45/8 - 5/4 + 3/4 \end{bmatrix} = \begin{bmatrix} 3/8 \\ -5/2 \\ 41/8 \end{bmatrix}.$$

Therefore $p(x) = (3/8)x^2 + (-5/2)x + (41/8)$.

2.3.8.5 Unified view

Consider each of the component lagrange polynomials:

- $p_1(x) = \frac{1}{4}(x^2 - 4x + 3) = (1/4)x^2 - x + (3/4)$

$$\vec{\boldsymbol{p}}_1 = \begin{bmatrix} 1/4 \\ -1 \\ 3/4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix} \vec{\boldsymbol{p}}_1 = \begin{bmatrix} 1/4 - 1 + 3/4 \\ 9/4 - 3 + 3/4 \\ 25/4 - 5 + 3/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

- $p_2(x) = -\frac{1}{4}(x^2 - 6x + 5)$

$$\vec{\boldsymbol{p}}_2 = \begin{bmatrix} -1/4 \\ 3/2 \\ -5/4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix} \vec{\boldsymbol{p}}_2 = \begin{bmatrix} -1/4 + 3/2 - 5/4 \\ -9/4 + 9/2 - 5/4 \\ -25/4 + 15/2 - 5/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

- $p_3(x) = \frac{3}{8}(x^2 - 8x + 15)$

$$\vec{\boldsymbol{p}}_3 = \begin{bmatrix} 3/8 \\ -3 \\ 45/8 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix} \vec{\boldsymbol{p}}_3 = \begin{bmatrix} 3/8 - 3 + 45/8 \\ 27/8 - 9 + 45/8 \\ 75/8 - 15 + 45/8 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}.$$

So,

$$\begin{bmatrix} 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix} (\vec{\boldsymbol{p}}_1 + \vec{\boldsymbol{p}}_2 + \vec{\boldsymbol{p}}_3) = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}.$$

2.4 Vector Spaces

2.4.1 Definition of a Vector Space

Let F denote a field which is a subfield of $\mathbb{C}$ and V denote a vector space over F.

Definition. ***Addition, Scalar Multiplication***

- *An* ***addition*** *on a set V is a function that assigns an element $u+v \in V$ to each pair of elements $u, v \in V$.*
- *A* ***scalar multiplication*** *on a set V is a function that assigns an element $\lambda v \in V$ to each $\lambda \in F$ and each $v \in V$.*

Note that both functions are closed over V.

A vector space *is a set V along with an addition on V and a scalar multiplication on V such that the following properties hold:*

commutativity $\vec{u} + \vec{v} = \vec{v} + \vec{u}$ *for all* $\vec{u}, \vec{v} \in V$*;*

associativity $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$ *and* $(ab)\vec{v} = a(b\vec{v})$ *for all* $\vec{u}, \vec{v}, \vec{w} \in V$ *and all* $a, b \in F$*;*

additive identity *there exists an element* $\vec{0} \in V$ *such that* $\vec{v} + \vec{0} = \vec{v}$ *for all* $\vec{v} \in V$*;*

additive inverse *for every* $\vec{v} \in V$ *there exists* $\vec{w} \in V$ *such that* $\vec{v} + \vec{w} = \vec{0}$*;*

multiplicative identity $1\vec{v} = \vec{v}$ *for all* $\vec{v} \in V$*;*

distributive properties $a(\vec{u} + \vec{v}) = a\vec{u} + a\vec{v}$ *and* $(a + b)\vec{u} = a\vec{u} + b\vec{u}$ *for all* $a, b \in F$ *and* $\vec{u}, \vec{v} \in V$*;*

Proposition 60. *A vector space contains a unique additive identity element.*

Proof. If $\vec{\mathbf{0}}'$ is also an additive identity then by the additive identity property,

$$\vec{\mathbf{0}} + \vec{\mathbf{0}}' = \vec{\mathbf{0}}$$

but since $\vec{\mathbf{0}}$ is also an additive identity,

$$\vec{\mathbf{0}}' + \vec{\mathbf{0}} = \vec{\mathbf{0}}'$$

Then, by the commutativity of vector addition,

$$\vec{\mathbf{0}} = \vec{\mathbf{0}} + \vec{\mathbf{0}}' = \vec{\mathbf{0}}' + \vec{\mathbf{0}} = \vec{\mathbf{0}}' \qquad \square$$

Proposition 61. *A vector space contains a unique additive inverse for each element.*

Proof. If $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{w}}$ are both additive inverses of $\vec{\boldsymbol{u}}$ then, by the additive inverse property we have,

$$\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}} = \vec{\mathbf{0}} \text{ and also } \vec{\boldsymbol{u}} + \vec{\boldsymbol{w}} = \vec{\mathbf{0}}$$

using the uniqueness of the additive identity,

$$\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}} = \vec{\mathbf{0}} = \vec{\boldsymbol{u}} + \vec{\boldsymbol{w}}$$

Then, if we add one of the additive inverses of $\vec{\boldsymbol{u}}$ to both sides,

$$\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = \vec{\boldsymbol{u}} + \vec{\boldsymbol{w}} + \vec{\boldsymbol{v}}$$

and use the associativity of vector addition,

$$\begin{aligned} (\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}}) + \vec{\boldsymbol{v}} &= (\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}}) + \vec{\boldsymbol{w}} \\ \vec{\mathbf{0}} + \vec{\boldsymbol{v}} &= \vec{\mathbf{0}} + \vec{\boldsymbol{w}} \\ \vec{\boldsymbol{v}} &= \vec{\boldsymbol{w}} \end{aligned}$$

$\square$

Vector Subtraction

Definition. *Because additive inverses are unique we can use the notation $-\vec{\boldsymbol{v}}$ to denote the additive inverse of $\vec{\boldsymbol{v}}$. Then we define $\vec{\boldsymbol{w}} - \vec{\boldsymbol{v}}$ to mean $\vec{\boldsymbol{w}} + -\vec{\boldsymbol{v}}$.*

$$\vec{\boldsymbol{u}} - \vec{\boldsymbol{v}} \coloneqq \vec{\boldsymbol{u}} + -\vec{\boldsymbol{v}}$$

Proposition 62. $0\vec{v} = \vec{0}$ *for every* $\vec{v} \in V$.

Note that this proposition is asserting something about scalar multiplication and the additive identity of V. The only part of the definition of a vector space that connects scalar multiplication and vector addition is the distributive property. Therefore the distributive property must be used in this proof.

Proof. Firstly take,

$$\vec{v} + 0\vec{v} = 0\vec{v} + 1\vec{v}$$

and then use the properties of the underlying field to say

$$(0 + 1)\vec{v} = 1\vec{v} = \vec{v}$$

Now we have shown that,

$$\vec{v} + 0\vec{v} = \vec{v}$$

which, by the definition and uniqueness of the additive identity, shows that $0\vec{v} = \vec{0}$. But if we want to continue algebraically we can now add the additive inverse to both sides,

$$(\vec{v} + -\vec{v}) + 0\vec{v} = (\vec{v} + -\vec{v})$$

$$\vec{0} + 0\vec{v} = 0\vec{v} = \vec{0}$$ □

Another, simpler proof exists.

Proof. Using the underlying field properties and the distributivity of scalar vector multiplication,

$$0\vec{v} = (0 + 0)\vec{v} = 0\vec{v} + 0\vec{v}$$

and then adding the additive inverse to both sides,

$$(0\vec{v} + -(0\vec{v})) = (0\vec{v} + -(0\vec{v})) + 0\vec{v}$$
$$\vec{0} = \vec{0} + 0\vec{v} = 0\vec{v}$$ □

Proposition 63. $a\vec{0} = \vec{0}$ *for every* $a \in F$.

Proof. Using the distributivity of scalar multiplication of vectors and the additive identity,

$$a\vec{\mathbf{0}} = a(\vec{\mathbf{0}} + \vec{\mathbf{0}}) = a\vec{\mathbf{0}} + a\vec{\mathbf{0}}$$

Then, adding the additive inverse to both sides,

$$\begin{aligned}(a\vec{\mathbf{0}} + -(a\vec{\mathbf{0}})) &= a\vec{\mathbf{0}} + (a\vec{\mathbf{0}} + -(a\vec{\mathbf{0}})) \\ \vec{\mathbf{0}} &= a\vec{\mathbf{0}} + \vec{\mathbf{0}} = a\vec{\mathbf{0}}\end{aligned}$$

□

Proposition 64. $(-1)\vec{\boldsymbol{v}} = -\vec{\boldsymbol{v}}$ *for every* $\vec{\boldsymbol{v}} \in V$.

Proof. Using the distributivity of scalar multiplication of vectors and the underlying field properties we have,

$$(-1)\vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = (-1)\vec{\boldsymbol{v}} + 1\vec{\boldsymbol{v}} = (-1+1)\vec{\boldsymbol{v}} = 0\vec{\boldsymbol{v}} = \vec{\mathbf{0}}$$

Now we could add the additive inverse to both sides to show that,

$$\begin{aligned}(-1)\vec{\boldsymbol{v}} + (\vec{\boldsymbol{v}} + -\vec{\boldsymbol{v}}) &= \vec{\mathbf{0}} + -\vec{\boldsymbol{v}} \\ (-1)\vec{\boldsymbol{v}} + \vec{\mathbf{0}} &= \vec{\mathbf{0}} + -\vec{\boldsymbol{v}} \\ (-1)\vec{\boldsymbol{v}} &= \vec{\boldsymbol{v}}\end{aligned}$$

□

But we already have,

$$(-1)\vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = \vec{\mathbf{0}}$$

and this, by the definition of the additive inverse, proves that $(-1)\vec{\boldsymbol{v}}$ is an additive inverse of $\vec{\boldsymbol{v}}$. Since we have previously proven the uniqueness of the additive inverse in Proposition 61 we can conclude, in fact, that $(-1)\vec{\boldsymbol{v}} = -\vec{\boldsymbol{v}}$ the unique additive inverse of v.

2.4.1.1 Vectors as magnitude and direction

A vector is often described as an object with magnitude and direction. So how does this relate to the definition of vector spaces?

Magnitude

If we take the distributive property of scalar multiplication of vectors over addition and the multiplicative identity we get:

$$\vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = 1\vec{\boldsymbol{v}} + 1\vec{\boldsymbol{v}} = (1+1)\vec{\boldsymbol{v}} = 2\vec{\boldsymbol{v}}.$$

This is how magnitudes behave: if we sum a magnitude with itself we get 2 times the magnitude. If we compare this with sets (not in a measure space), for example, where addition of sets is set union and, if A is a set then,

$$A + A = A.$$

That's to say, sets (without a measure defined on them) do not behave as magnitudes; if we add a set to itself we get the original set. As a result, it would be difficult or inappropriate to model sets as vectors.

> *Sets have their own sort of algebra: see wikipedia.*

Direction

If we take Proposition 64 we see that

$$(-1)\vec{\boldsymbol{v}} = -v \quad \text{and} \quad (-1)\vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = \vec{\mathbf{0}}.$$

That's to say, a vector's magnitude is cancelled by being added to another vector with the same magnitude multiplied by -1. This is how the directionality comes into the definition. This abstract vector space definition only defines that a vector and its additive inverse have opposing direction; a relationship of direction between vectors that are not additive inverses is not defined. If an inner product is defined over the vector space then this provides a concept of angle between vectors.

> *A coordinate space is not necessarily required to define an inner product. For example, the expectation of the product of two random variables is an inner product: wikipedia.*

Minimal Representation

Since a magnitude and a direction are required to define vectors, the minimal numerical representation of a vector is a signed number. We can see this in basic physics where, if a system is defined in a single spatial dimension, forces are described as a signed real number with the sign indicating the direction.

2.4.1.2 Vector Spaces as Groups

Vector addition on a vector space V is an associative, commutative law of composition on the set of vectors so that $\boldsymbol{(V,+)}$ is an abelian group. Notice also, however, that scalar multiplication defines a law of composition between vectors in V and scalars in the field F. This is known as an **external law of composition** on the vector space. This is an important part of the definition of vectors as, it defines a relationship between the additive group of vectors V^+ and the field F in much the same way that the distributive law does for the additive and multiplicative groups inside fields (compare 2.2). Specifically the relationship takes the form,

$$(1+1)\vec{\boldsymbol{v}} = \vec{\boldsymbol{v}} + \vec{\boldsymbol{v}} = 2\vec{\boldsymbol{v}}.$$

It is typically the case that when modelling some system with vectors, the system contains more structure than is represented by a vector space. Furthermore, if we view a vector space as an abelian group then the group only describes a part of the vector space structure. These abstractions — vector space and group — allow us to see what is generalizable about a system and what is specific to the system in question.

Isomorphisms between vector spaces

Definition. *An* ***isomorphism*** $\phi : V \longmapsto V'$ *– where* V *and* V' *are defined over the same field* F *– is a bijective map compatible with the vector laws of composition. That's to say, for all* $\vec{\boldsymbol{v}}, \vec{\boldsymbol{v'}} \in V$, $c \in F$,

$$\phi(\vec{\boldsymbol{v}} + \vec{\boldsymbol{v'}}) = \phi(\vec{\boldsymbol{v}}) + \phi(\vec{\boldsymbol{v'}}) \quad \textit{and} \quad \phi(c\vec{\boldsymbol{v}}) = c\phi(\vec{\boldsymbol{v}}).$$

Examples

(34) The space $\mathbb{F}^n$ of n-dimensional row vectors is isomorphic to the space of n-dimensional column vectors.

(35) If we treat the set of complex numbers $\mathbb{C}$ as a vector space then the map $\phi : \mathbb{R}^2 \longmapsto \mathbb{C}$ defined as $\phi(a,b) = a + bi$ is an isomorphism.

2.4.2 The notation F^S

Notation. If S is a set then $\boldsymbol{F^S}$ denotes the set of functions $S \mapsto F$.

Addition is defined as, for $f, g, (f+g) \in F^S$,

$$(f+g)(x) = f(x) + g(x)$$

for all $x \in S$.

Scalar multiplication is defined as, for $\lambda \in F, \lambda f \in F^S$,

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in S$.

Example: If S is the interval $[0,1]$ and $F = \mathbb{R}$ then $\mathbb{R}^{[0,1]}$ is the set of real-valued functions on the interval $[0,1]$. $\mathbb{R}^{[0,1]}$ is a vector space with additive identity $0 : [0,1] \mapsto \mathbb{R}$ defined as $0(x) = 0$ and the additive inverse of some function $f \in \mathbb{R}^{[0,1]}$ is the function defined as $(-f)(x) = -f(x)$.
Any *non-empty* set S in conjunction with a subset of $\mathbb{C}$ would similarly produce a vector space. In fact, the vector space F^n can be thought of as the space of functions from the set $\{1, 2, 3, \ldots, n\}$ to F. For example, vectors in 3-dimensional space can be viewed as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \equiv f : \{1,2,3\} \mapsto \mathbb{R} \text{ with } f(t) = \begin{cases} x & t = 1 \\ y & t = 2 \\ z & t = 3 \end{cases}$$

2.4.3 Polynomials as a vector space

A very important example involves treating a polynomial as a vector. A function $p : F \mapsto F$ is called a polynomial with coefficients in F if there exist $a_0, \ldots, a_m \in F$ such that,

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$

for all $z \in F$.
Then we can define a vector space, $P(F)$, to be the set of all polynomials with coefficients in F.

Addition on $P(F)$ is defined as,

$$(p+q)(z) = p(z) + q(z) \qquad \text{for } p, q \in P(F), z \in F$$

whose associativity is clear from the definition and the commutativity can be shown by,

$$\begin{aligned}((p+q)+r)(z) &= (p+q)(z) + r(z) \\ &= p(z) + q(z) + r(z) \\ &= p(z) + (q+r)(z) \\ &= (p + (q+r))(z)\end{aligned}$$

Scalar multiplication on $P(F)$ is defined as,

$$(ap)(z) = ap(z) \qquad \text{for } p \in P(F), a, z \in F$$

whose associativity can be shown by substituting (ab) for a in the definition,

$$[(ab)p](z) = (ab)p(z)$$

Then, by the associativity of the multiplication of the elements of the field F we have,

$$(ab)p(z) = a[b(p(z)]$$

then we use the definition in reverse,

$$a[b(p(z)]) = a[(bp)(z)] = [a(bp)](z)$$

(compare with $(ab)\vec{\boldsymbol{v}} = a(b\vec{\boldsymbol{v}})$)

modeling Concretely, each $p(z) \in P(F)$ is a vector that could be modeled, say, as

$$\vec{\boldsymbol{p}} = \{\, (a_0, a_1, \ldots, a_m) \mid p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m \in P(F) \,\}$$

2.4.4 Subspaces of vector spaces

Definition. *A set U is a subspace of V if it is a subset of V and if the same addition and multiplication over U forms a vector space.*

Considering the required properties of a vector space, we can see that commutativity and associativity of the addition; associativity of the scalar multiplication; and distributivity of the scalar multiplication over the addition; will all be satisfied as we have the same addition and multiplication over a subset of the elements in V. That's to say, the vector space properties ensure that these properties hold $\forall \vec{\boldsymbol{v}} \in V$ and we have $\forall \vec{\boldsymbol{u}} \in U, \vec{\boldsymbol{u}} \in V$.
Furthermore, the multiplicative identity also holds $\forall \vec{\boldsymbol{v}} \in V$ so will also hold for every element of U.

So what remains to be proven to satisfy the requirements of a subspace?

- Existence of the additive identity
- Existence of an additive inverse for every element of U
- Closure of the addition and scalar multiplication over U

Note, however that - having proved in Proposition 64 that multiplication by -1 gives the additive inverse - closure of the scalar multiplication over U also implies the presence in U of the additive inverse of every element of U. So, actually, what we need to prove for U to be a subspace is only,

- $\vec{\mathbf{0}} \in U$
- Closure of the addition and scalar multiplication over U

2.4.4.1 Examples of Subspaces

(36) An example of a subspace of the polynomials, $P(F)$ is,

$$\{\, p \in P(F) \mid p(3) = 0 \,\}$$

Members of this subspace include:

- $p(z) = 3 - z$
- $p(z) = 9 - z^2$
- $p(z) = 3 - z + 3z^2 - z^3$
- $p(z) = 12z - 4z^2$
- ...etc.

To verify this we need to show that addition and multiplication are closed over this set and that $\vec{\mathbf{0}}$ is a member of the set. It's easy to see that $\vec{\mathbf{0}}$ is a member of the set as,

$$p(3) = 0 + 0(3) + 0(3)^2 + \cdots + 0(3)^m = 0$$

as required. Scalar multiplication is closed as,

$$ap(3) = a(0) = 0$$

whereas addition can be shown to be closed as,

$$(q + r)(3) = q(3) + r(3) = 0 + 0 = 0$$

Note that for values of $z \neq 3$, the closure of these functions is the same as for the general case of $P(F)$.

2.4.4.2 Sums and Direct Sums

Definition. *If $U_1, \ldots, U_m$ are subspaces of V then their sum is defined as*

$$U_1 + \cdots + U_m = \{\, \vec{\boldsymbol{u_1}} + \cdots + \vec{\boldsymbol{u_m}} \mid \vec{\boldsymbol{u_1}} \in U_1, \ldots, \vec{\boldsymbol{u_m}} \in U_m \,\}.$$

The sum of the subspaces of V is also a subspace of V because,

- Closure of addition

$$\begin{aligned}
&(\vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u_2}} + \cdots + \vec{\boldsymbol{u_m}}) + (\vec{\boldsymbol{u'_1}} + \vec{\boldsymbol{u'_2}} + \cdots + \vec{\boldsymbol{u'_m}}) \\
&= (\vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u'_1}}) + (\vec{\boldsymbol{u_2}} + \vec{\boldsymbol{u'_2}}) + \cdots + (\vec{\boldsymbol{u_m}} + \vec{\boldsymbol{u'_m}}) \\
&= \vec{\boldsymbol{v_1}} + \vec{\boldsymbol{v_2}} + \cdots + \vec{\boldsymbol{v_m}} \qquad \text{where } \vec{\boldsymbol{v_1}} \in U_1, \vec{\boldsymbol{v_2}} \in U_2, \ldots, \vec{\boldsymbol{v_m}} \in U_m
\end{aligned}$$

- Closure of scalar multiplication

$$\begin{aligned}
&a(\vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u_2}} + \cdots + \vec{\boldsymbol{u_m}}) \qquad \text{where } a \in F \\
&= a\vec{\boldsymbol{u_1}} + a\vec{\boldsymbol{u_2}} + \cdots + a\vec{\boldsymbol{u_m}} \\
&= \vec{\boldsymbol{v_1}} + \vec{\boldsymbol{v_2}} + \cdots + \vec{\boldsymbol{v_m}} \qquad \text{where } \vec{\boldsymbol{v_1}} \in U_1, \vec{\boldsymbol{v_2}} \in U_2, \ldots, \vec{\boldsymbol{v_m}} \in U_m
\end{aligned}$$

- Existence of $\vec{\mathbf{0}}$

$$\begin{aligned}
&U_1, U_2, \ldots, U_m \text{ are subspaces} \\
\implies &\vec{\mathbf{0}} \in U_1, \vec{\mathbf{0}} \in U_2, \ldots, \vec{\mathbf{0}} \in U_m \\
\implies &\vec{\mathbf{0}} + \vec{\mathbf{0}} + \cdots + \vec{\mathbf{0}} \in U_1 + U_2 + \cdots + U_m
\end{aligned}$$

> *Note though, that this may not be the only way of producing $\vec{\mathbf{0}}$ from the sum of vectors of these subspaces. That's to say, there could be some $(\vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u_2}} + \cdots + \vec{\boldsymbol{u_m}}) = \vec{\mathbf{0}}$ and this is a key difference from direct sums.*

Proposition 65. *$U_1 + U_2 + \cdots + U_m$ is the smallest subspace of V containing $U_1, U_2, \ldots, U_m$.*

Proof. $U_1 + U_2 + \cdots + U_m$ is a subspace of V that contains $U_1, U_2, \ldots, U_m$ because we can obtain U_i by setting all the u_j for $j \neq i$ to $\vec{\mathbf{0}}$.

If a subspace of V contains $U_1, U_2, \ldots, U_m$ then, by the closure of addition, it must also contain $U_1 + U_2 + \cdots + U_m$.

Therefore the smallest subspace of V that contains $U_1, U_2, \ldots, U_m$ is $U_1 + U_2 + \cdots + U_m$. $\square$

Definition. *If $U_1, \ldots, U_m$ are subspaces of V then their* ***direct sum*** *is defined as,*

$$U_1 \oplus \cdots \oplus U_m = \{ \vec{\boldsymbol{u_1}} + \cdots + \vec{\boldsymbol{u_m}} \mid \vec{\boldsymbol{u_1}} \in U_1, \ldots, \vec{\boldsymbol{u_m}} \in U_m \}$$

such that,

$$\vec{\boldsymbol{u_1}} + \cdots + \vec{\boldsymbol{u_m}} = \vec{\mathbf{0}} \implies \vec{\boldsymbol{u_1}} = \vec{\mathbf{0}}, \ldots, \vec{\boldsymbol{u_m}} = \vec{\mathbf{0}}.$$

In this case, the subspaces $U_1, \ldots, U_m$ are said to be ***independent****.*

That the unique way of obtaining $\vec{\mathbf{0}}$ is for all of the vectors from each of the subspaces to be $\vec{\mathbf{0}}$ is equivalent to there only being a single unique way of obtaining each resultant vector from an addition of the vectors from the individual subspaces. This can be seen as,

$$\begin{aligned} \vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u_2}} + \cdots + \vec{\boldsymbol{u_m}} &= \vec{\boldsymbol{u_1'}} + \vec{\boldsymbol{u_2'}} + \cdots + \vec{\boldsymbol{u_m'}} \\ (\vec{\boldsymbol{u_1}} + \vec{\boldsymbol{u_2}} + \cdots + \vec{\boldsymbol{u_m}}) - (\vec{\boldsymbol{u_1'}} + \vec{\boldsymbol{u_2'}} + \cdots + \vec{\boldsymbol{u_m'}}) &= \vec{\mathbf{0}} \\ (\vec{\boldsymbol{u_1}} - \vec{\boldsymbol{u_1'}}) + (\vec{\boldsymbol{u_2}} - \vec{\boldsymbol{u_2'}}) + \cdots + (\vec{\boldsymbol{u_m}} - \vec{\boldsymbol{u_m'}}) &= \vec{\mathbf{0}} \end{aligned}$$

Therefore, since vector spaces always contain $\vec{\mathbf{0}}$ and so we will always have the representation,

$$\vec{\mathbf{0}} + \vec{\mathbf{0}} + \cdots + \vec{\mathbf{0}} = \vec{\mathbf{0}}$$

if this is the unique representation of $\vec{\mathbf{0}}$ then it follows that,

$$\begin{aligned} (\vec{\boldsymbol{u_1}} - \vec{\boldsymbol{u_1'}}) = \vec{\mathbf{0}}, (\vec{\boldsymbol{u_2}} - \vec{\boldsymbol{u_2'}}) = \vec{\mathbf{0}}, \ldots, (\vec{\boldsymbol{u_m}} - \vec{\boldsymbol{u_m'}}) &= \vec{\mathbf{0}} \\ \implies \vec{\boldsymbol{u_1}} = \vec{\boldsymbol{u_1'}}, \vec{\boldsymbol{u_2}} = \vec{\boldsymbol{u_2'}}, \ldots, \vec{\boldsymbol{u_m}} &= \vec{\boldsymbol{u_m'}} \end{aligned}$$

which means that these are the same representation. And this clearly holds in reverse also as, if there is a single way of representing each resultant vector then there must be a single way of representing $\vec{\mathbf{0}}$ and due to the definition of a vector space we must always have the representation of all $\vec{\mathbf{0}}$. Therefore, this is the only representation of $\vec{\mathbf{0}}$.

Note that this is a condition on the contents of the subspaces and not on the way that the addition is performed. So, the difference between vector space sum $(U_1 + U_2)$ and vector space direct sum $(U_1 \oplus U_2)$ is not in the operator itself but in the operands they operate over.

For two subspaces, say, U_1, U_2 this condition on the subspaces reduces to the requirement that $U_1 \cap U_2 = \{\vec{\mathbf{0}}\}$ which can be seen as,

$$\begin{aligned} \vec{u_1} + \vec{u_2} &= \vec{0} \\ \vec{u_1} + -\vec{u}_1 + \vec{u_2} &= \vec{0} + -\vec{u}_1 \\ \vec{u_2} &= -\vec{u}_1 \\ \implies -\vec{u}_1 \in U_2 &\implies \vec{u_1} \in U_2 \end{aligned}$$

So, for two subspaces, obtaining $\vec{\mathbf{0}}$ as the sum of vectors from the subspaces implies a vector in common between them. So, for $\vec{\mathbf{0}} + \vec{\mathbf{0}}$ to be the only way of obtaining $\vec{\mathbf{0}}$ implies that $\vec{\mathbf{0}}$ is the only vector in common.
However, for more than two subspaces, say U_1, U_2, U_3, the situation is different as we could have,

$$\begin{aligned} \vec{u_1} + \vec{u_2} + \vec{u_3} &= \vec{0} \\ \iff \vec{u_1} + -\vec{u}_1 + \vec{u_2} + -\vec{u}_2 + \vec{u_3} &= \vec{0} + -\vec{u}_1 + -\vec{u}_2 \\ \iff \vec{u_3} &= -\vec{u}_1 + -\vec{u}_2 \end{aligned}$$

which does not imply any vectors held in common.

2.4.5 Vector Space Problems

Prove that $-(-\vec{v}) = \vec{v}$ for every $\vec{v} \in V$

$$\begin{aligned} -(-\vec{v}) &= -[(-1)\vec{v}] && \text{using Proposition 64} \\ &= (-1)[(-1)\vec{v}] && \text{using Proposition 64 again} \\ &= [(-1)(-1)]\vec{v} && \text{using associativity of scalar multiplication} \\ &= \vec{v} && \text{using field properties} \end{aligned}$$

Or, a quicker way is,

$$\begin{aligned} -\vec{v} + -(-\vec{v}) &= \vec{0} && \text{using additive identity of } -\vec{v} \\ (-\vec{v} + \vec{v}) + -(-\vec{v}) &= \vec{0} + \vec{v} && \text{adding } \vec{v} \text{ to both sides} \\ -(-\vec{v}) &= \vec{v} \end{aligned}$$

Prove that if $a \in F, \vec{v} \in V$, and $a\vec{v} = \vec{0}$, then $a = 0$ or $\vec{v} = \vec{0}$. We follow a proof by cases.

Case $a \neq 0$:

$$\begin{aligned} a\vec{v} = \vec{0}, a \neq 0 &\implies a^{-1}a\vec{v} = a^{-1}\vec{0} && \text{using field properties} \\ &\iff 1\vec{v} = b\vec{0} && \text{where } b = a^{-1} \in F \\ &\iff \vec{v} = \vec{0} && \text{using Proposition 63 and multiplicative identity} \end{aligned}$$

Case $\vec{v} \neq \vec{0}$:

$$\begin{aligned} a\vec{v} = \vec{0}, \vec{v} \neq \vec{0} &\implies a\vec{v} = a\vec{v} + -a\vec{v} \\ &\iff a\vec{v} = (a + -a)\vec{v} = 0\vec{v} && \text{using field properties} \\ \text{Wrong! } a\vec{v} = \vec{0} &\implies a\vec{v} = a\vec{v} + -a\vec{v} \\ &\text{without need for } \vec{v} \neq \vec{0} \end{aligned}$$

This indicates that you are proving something that doesn't need proving. In actual fact,

Case $a = 0$: Actually, in this case there is nothing to be proven as we know from Proposition 62 that $0\vec{v} = \vec{0}$. So we have collectively exhaustive cases by looking at $a = 0$ and $a \neq 0$ and we only need to show that $a \neq 0 \implies \vec{v} = \vec{0}$ which we have already done.

Case $\vec{v} \neq \vec{0}$:

$$\begin{aligned} & a\vec{v} = \vec{0}, \vec{v} \neq \vec{0} \implies a\vec{v} = \vec{0} = \vec{v} + -\vec{v} \\ \iff \quad & a\vec{v} + (-\vec{v}) = -\vec{v} && \text{applying additive inverse} \\ \iff \quad & a\vec{v} + (-1)\vec{v} = -\vec{v} \\ \iff \quad & (a + (-1))\vec{v} = (-1)\vec{v} && \text{using distributive law} \\ \iff \quad & a - 1 = -1 && \text{using injectivity of scalar multiplication} \\ \iff \quad & a = 0. && \text{using field properties} \end{aligned}$$

Give an example of a nonempty subset U of $\mathbb{R}^2$ such that U is closed under scalar multiplication but U is not a subspace of $\mathbb{R}^2$. For all $\lambda \in \mathbb{R}$ the set $\{ \lambda\vec{v} \mid \vec{v} \in \{(1,1)(-1,1)\} \}$ is closed under scalar multiplication but not addition.

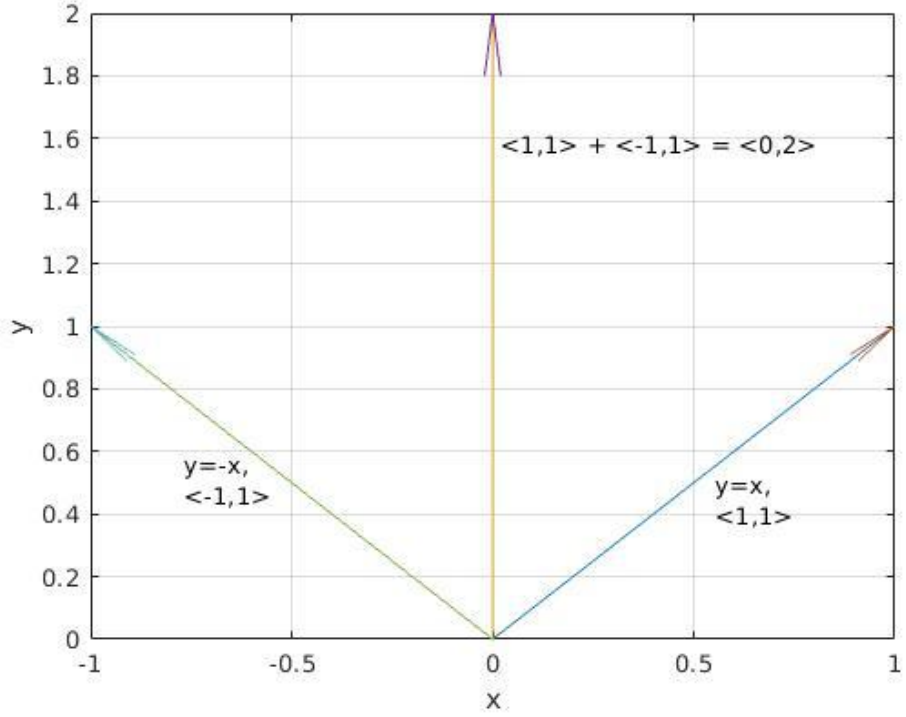


Figure 2.2: The blue arrows are vectors whose scalar multiples will all be in the same line as the blue arrows but the red arrow shows what happens if we add them; the result lies outside of both lines.

Is $\mathbb{R}^2$ a subspace of the complex vector space $\mathbb{C}^2$? The definition of a subspace of $\mathbb{C}^2$ is a set of vectors which is a subset of those in $\mathbb{C}^2$ and that forms a vector space under the same addition and scalar multiplication of $\mathbb{C}^2$. The scalar multiplication of the vector space $\mathbb{C}^2$ is multiplication by scalars $\lambda \in \mathbb{C}$.
For a vector, $\vec{v} \in \mathbb{R}^2$, scaling it by a complex number, $\lambda\vec{v}$ will result in a vector that is not necessarily in $\mathbb{R}^2$.

Is $\{ (a,b,c) \in \mathbb{C}^3 \mid a^3 = b^3 \}$ a subspace of $\mathbb{C}^3$? For $x \in \mathbb{C}$, x^3 has roots, $1, \frac{-1+\sqrt{3}i}{2}, \frac{-1-\sqrt{3}i}{2}$ so we don't have $a = b$ as we would if we were ranging over the reals.
Concretely, we can have, $(1, \frac{-1+\sqrt{3}i}{2}, 0)$ and $(1, \frac{-1-\sqrt{3}i}{2}, 0)$ but,

$$(1, \frac{-1+\sqrt{3}i}{2}, 0) + (1, \frac{-1-\sqrt{3}i}{2}, 0) = (2, -1, 0)$$

where $(2,-1,0) \notin \{ (a,b,c) \in \mathbb{C}^3 \mid a^3 = b^3 \}$ meaning that addition over this set is not closed. Therefore, this is not a subspace.

Give an example of a non-empty subset U of $\mathbb{R}^2$ such that U is closed under addition and under taking additive inverses (meaning $-\vec{u} \in U$ whenever $\vec{u} \in U$), but U is not a subspace of $\mathbb{R}^2$. First thought might be $\mathbb{R}^2 - \{\vec{0}\}$ but this is Wrong!. If the subset is closed under addition and under taking additive inverses then it means that $\vec{u} + -\vec{u} = \vec{0} \in U$ and so the set $\mathbb{R}^2 - \{\vec{0}\}$ is not closed under addition and taking additive inverses.
The set $\{ (x,y) \in \mathbb{R}^2 \mid x, y \in \mathbb{Z} \}$ however, is closed under addition because integer addition is closed and under taking additive inverses but scalar multiplication where the scalars range over the reals, will produce non-integer values for x and y.

Is the set of periodic functions over the reals a subspace of $\mathbb{R}^R$? vector problems, periodic functions The definition of two periodic functions over the reals is

$$\exists p > 0 \in \mathbb{R} \cdot f(x) = f(x+p)$$
$$\exists q > 0 \in \mathbb{R} \cdot g(x) = g(x+q)$$

Then for their sum to be periodic we need,

$$\exists \alpha, \beta \in \mathbb{Z}, m \in \mathbb{R} \cdot (m = \alpha p) \wedge (m = \beta q)$$
$$\iff \frac{q}{p} = \frac{\alpha}{\beta} \in \mathbb{Q}$$
$$\therefore (f+g)(x) = (f+g)(x+m) = f(x+m) + g(x+m)$$
$$\iff \frac{q}{p} \in \mathbb{Q}.$$

Prove that the union of two subspaces of V is a subspace of V if and only if one of the subspaces is contained within the other. Let A, B be subspaces of V and $\vec{\boldsymbol{a}} \in A,\ \vec{\boldsymbol{b}} \in B$ and,

$$C = A \cup B = \{\, \vec{\boldsymbol{c}} \mid \vec{\boldsymbol{c}} \in A \ \vee\ \vec{\boldsymbol{c}} \in B \,\}.$$

Since $\vec{\boldsymbol{a}}, \vec{\boldsymbol{b}} \in C$ we have (C subspace of V) $\iff \forall \alpha, \beta \in F \cdot (\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in C$. Then,

$$\begin{aligned} \vec{\boldsymbol{b}} \in A &\implies \forall \alpha, \beta \in F \cdot (\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in A \text{ (by subspace properties)} \\ &\implies (\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in C. \end{aligned}$$

A similar argument holds for $\vec{\boldsymbol{a}} \in B$. Conversely,

$$\begin{aligned} \forall \alpha, \beta \in F \cdot (\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in C &\implies ((\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in A) \vee ((\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in B) \\ &\implies ((\alpha\vec{\boldsymbol{a}} - \alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) = \beta\vec{\boldsymbol{b}} \in A) \vee ((\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}} - \beta\vec{\boldsymbol{b}}) = \alpha\vec{\boldsymbol{a}} \in B) \\ &\implies (\vec{\boldsymbol{b}} \in A) \vee (\vec{\boldsymbol{a}} \in B) \end{aligned}$$

$$\begin{aligned} \therefore \text{(C subspace of V)} &\iff \forall \alpha, \beta \in F \cdot (\alpha\vec{\boldsymbol{a}} + \beta\vec{\boldsymbol{b}}) \in C \\ &\iff (\vec{\boldsymbol{b}} \in A) \vee (\vec{\boldsymbol{a}} \in B) \\ &\equiv (B \subseteq A) \vee (A \subseteq B). \end{aligned}$$

Can a vector space over an infinite field be a finite union of proper subspaces? Assume that our vector space V is a finite union of proper subspaces, hence

$$V = \bigcup_{i=1}^{n} U_i.$$

Now, pick a non-zero vector $\vec{\boldsymbol{x}} \in U_1$, and pick another vector $\vec{\boldsymbol{y}} \in V \setminus U_1$.

There are infinitely many vectors $\vec{\boldsymbol{x}} + k\vec{\boldsymbol{y}}$, where $k \in K^*$ (K is our infinite field). Note that $\vec{\boldsymbol{x}} + k\vec{\boldsymbol{y}}$ is not in U_1, hence must be contained in some U_j where $j \neq 1$.

Then since $k \in K^*$, we can have $\vec{\boldsymbol{x}} + k_1\vec{\boldsymbol{y}}$, $\vec{\boldsymbol{x}} + k_2\vec{\boldsymbol{y}} \in U_j$, which implies that it also contains $\vec{\boldsymbol{y}}$ and hence also $\vec{\boldsymbol{x}}$, hence $U_1 \subset U_j$.

Explanation: There are infinitely many vectors $\vec{\boldsymbol{x}} + k\vec{\boldsymbol{y}}$ and only finitely many U_i so they cannot all be in different U_i so we have,

$$\begin{aligned} &\exists\, k_1,\, k_2 \in K^* \,\cdot\, \vec{\boldsymbol{x}} + k_1\vec{\boldsymbol{y}},\, \vec{\boldsymbol{x}} + k_2\vec{\boldsymbol{y}} \in U_j \\ \implies &(\vec{\boldsymbol{x}} + k_1\vec{\boldsymbol{y}}) - (\vec{\boldsymbol{x}} + k_2\vec{\boldsymbol{y}}) = (k_1 - k_2)\vec{\boldsymbol{y}} \in U_j \\ &\implies \vec{\boldsymbol{y}} \in U_j \implies \vec{\boldsymbol{x}} \in U_j \end{aligned}$$

Hence

$$V = \bigcup_{i=2}^{n} U_i.$$

Evidently, this can be continued, hence a contradiction arises.

Prove or give a counterexample: if U_1, U_2, W are subspaces of V such that $V = U_1 \oplus W$ and $V = U_2 \oplus W$ then $U_1 = U_2$. Counter example: $V = \mathbb{F}^2,\ U_1 = \{\,(x, 0) \in \mathbb{F}^2 \mid x \in F\,\},\ U_2 = \{\,(0, x) \in \mathbb{F}^2 \mid x \in F\,\},\ W = \{\,(x, x) \in \mathbb{F}^2 \mid x \in F\,\}$.

Let U_e denote the set of real-valued even functions on $\mathbb{R}$ and let U_o denote the set of real-valued odd functions on $\mathbb{R}$. Show that $\mathbb{R}^R = U_e \oplus U_o$. Every function $f \in \mathbb{R}^R$ can be expressed as the sum of an even function and an odd function as,

$$f(x) = \frac{f(x) + f(-x)}{2} + \frac{f(x) - f(-x)}{2} = g(x) + h(x)$$

where $g(x) \in U_e$ and $h(x) \in U_o$. So, $U_e + U_o$ spans $\mathbb{R}^R$.
Furthermore,

$$\begin{aligned} f(x) \in (U_e \cap U_o) &\implies (f(-x) = f(x)) \wedge (f(-x) = -f(x)) \\ &\implies f(x) = -f(x) \\ &\implies f(x) = 0 \end{aligned}$$

Since $f(x) = 0$ is the additive identity of this space, this shows that the intersection is $\vec{\mathbf{0}}$. So, $\mathbb{R}^R = U_e \oplus U_o$.

2.4.6 Finite Sets of Vectors

2.4.6.1 Span and Linear Independence

Definition. *The **span** of a nonempty finite set of vectors S – written span S – is defined as the set of **finite** linear combinations of elements of S,*

$$\{\, \alpha_1\vec{\boldsymbol{v_1}} + \alpha_2\vec{\boldsymbol{v_2}} + \cdots + \alpha_k\vec{\boldsymbol{v_n}} \mid \vec{\boldsymbol{v_i}} \in S,\, \alpha_j \in F \,\}.$$

The span of an empty set of vectors is defined to be $\{\vec{\mathbf{0}}\}$.

*The span of a set S is also sometimes known as the **subspace generated by** S.*

Definition. *A **linear relation** among a nonempty finite set of vectors S is any relation of a **finite** number of elements of S of the form,*

$$c_1\vec{\boldsymbol{v_1}} + \cdots + c_n\vec{\boldsymbol{v_n}} = \vec{\mathbf{0}}$$

where $c_i \in F$.
*A **linearly independent** set of vectors is a nonempty set among which there is no linear relation except the trivial relation where all $c_1, \ldots, c_n = 0$.*

The empty set is defined to be linearly independent.

Proposition 66. *If a set of vectors contains the zero vector $\vec{\mathbf{0}}$ then it cannot be linearly independent.*

Proof. Assume an arbitrary set of vectors $\vec{\boldsymbol{v_1}}, \ldots, \vec{\boldsymbol{v_n}}$ and assume it contains some $\vec{\boldsymbol{v_i}} = \vec{\mathbf{0}}$. Then we have the linear relation $c_i\vec{\boldsymbol{v_i}} = \vec{\mathbf{0}}$ with some $c_i \neq 0$. □

Corollary 16. *If a set of vectors contains any repitition then it cannot be linearly independent.*

Proof. If a set of vectors contains the same vector twice then subtracting one from the other is a non-trivial linear combination of the set of vectors equalling $\vec{\mathbf{0}}$. □

Corollary 17. *A set of vectors is linearly independent iff, for any vector that may be expressed as a linear combination of vectors in the set, the expression is unique.*

Proof. If the expression is not unique then the two different representations may be subtracted one from the other to produce a non-trivial linear combination resulting in $\vec{\mathbf{0}}$ which contradicts the hypothesis that they are linearly independent. Therefore linearly independent implies unique representation.

Conversely, if the set of vectors is not linearly independent then there exists some non-trivial linear combination that results in $\vec{\mathbf{0}}$. This, in turn, implies that there exists some linear combination of a subset of the vectors that is equal to a linear combination of the remaining vectors. Since these two linear combinations are equal, they represent two different expressions of the same resultant vector. Therefore unique representation implies linearly independent. □

Proposition 67. *The span of a list of vectors is the smallest subspace containing those vectors.*

Note that a vector space over $\mathbb{R}$ or $\mathbb{C}$ is an uncountable set as - while the dimensions of the vector space may be finite - closure under scalar multiplication means that the vectors in the space are continuously valued as the field providing the scalars is continuously valued.
This means that the notion of the *smallest* subspace cannot refer to the cardinality of the set and must refer to ordering based on subset. So, the smallest subspace containing a list of vectors is a subspace that contains the list of vectors and, of which, there is no proper subset which also contains the list of vectors.

Proof.

Let $S := span(\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}})$

$$:= \{\, \alpha_1\vec{\boldsymbol{v_1}} + \alpha_2\vec{\boldsymbol{v_2}} + \cdots + \alpha_k\vec{\boldsymbol{v_k}} \mid \alpha_1, \alpha_2, \ldots, \alpha_k \in F \,\}$$
$$\text{and let } V := \text{ the smallest vector space containing } \vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}.$$

then S contains every linear combination of $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}$ and nothing else and so is a vector space containing $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}$,

$$V \subseteq S$$

Additionally, any vector space containing the vectors $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}$ must contain all their linear combinations, $span(\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}})$,

$$S \subseteq V$$

Therefore there is no proper subset of $span(\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}})$ that is also a vector space containing $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}$, and so $span(\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}})$ is the smallest vector space containing $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \ldots, \vec{\boldsymbol{v_k}}$,

$$(V \subseteq S) \wedge (S \subseteq V) \iff V = S \qquad \square$$

Proposition 68. *Let L be a linearly independent set of vectors in V and $\vec{\boldsymbol{v}} \in V$. If we add $\vec{\boldsymbol{v}}$ to the set L then the resultant set L' is linearly independent iff $\vec{\boldsymbol{v}} \notin span\, L$.*

Proof. Clearly if $\vec{\boldsymbol{v}} \in span\, L$ then the resultant set is linearly dependent.
If $v \notin span\, L$ however, then if we attempt to form a linear relation of the vectors in L' then we find,

$$c_1\vec{\boldsymbol{v_1}} + \cdots + c_n\vec{\boldsymbol{v_n}} + b\vec{\boldsymbol{v}} = \vec{\boldsymbol{0}}$$

implies that $b \neq 0$ because that would leave a linear relation between the vectors of L which is not possible because L is linearly independent. Therefore,

$$\vec{\boldsymbol{v}} = -(c_1/b)\vec{\boldsymbol{v_1}} + \cdots + -(c_n/b)\vec{\boldsymbol{v_n}}$$

which contradicts the assumption that $v \notin span\, L$. $\square$

Proposition 69. *If we add a vector $\vec{\boldsymbol{v}} \in V$ to a set of vectors L in V to make a new set L', then $span\, L = span\, L'$ iff $\vec{\boldsymbol{v}} \in span\, L$.*

Proof. Clearly if $\vec{\boldsymbol{v}} \in span\, L$ then adding it to the set L doesn't change its span, so $span\, L = span\, L'$.
Conversely, by construction of L' we have $\vec{\boldsymbol{v}} \in L' \implies \vec{\boldsymbol{v}} \in span\, L'$ so if $span\, L = span\, L'$ then we also have $\vec{\boldsymbol{v}} \in span\, L$. □

Proposition 70. *Length of every linearly independent list in a space is less than or equal to the length of a spanning list in the same space.*

Proof. Let $U = \vec{\boldsymbol{u_1}}, \vec{\boldsymbol{u_2}}, \ldots, \vec{\boldsymbol{u_m}}$ be a linearly independent list of vectors in V and $W = \vec{\boldsymbol{w_1}}, \vec{\boldsymbol{w_2}}, \ldots, \vec{\boldsymbol{w_n}}$ be a spanning list of vectors in V.
If we take $\vec{\boldsymbol{u_1}}$ from U and add it to W then - since the other vectors in W are a spanning list - W must be linearly dependent. That's to say,

$$\begin{aligned}
\exists\, \alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R} \cdot \alpha_1 \vec{\boldsymbol{w_1}} + \cdots + \alpha_n \vec{\boldsymbol{w_n}} &= \vec{\boldsymbol{u_1}} \\
\iff \alpha_1 \vec{\boldsymbol{w_1}} + \cdots + \alpha_n \vec{\boldsymbol{w_n}} - \vec{\boldsymbol{u_1}} &= -\alpha_i \vec{\boldsymbol{w_i}} \\
\iff \frac{-\alpha_1}{\alpha_i} \vec{\boldsymbol{w_1}} + \cdots + \frac{-\alpha_n}{\alpha_i} \vec{\boldsymbol{w_n}} + \frac{1}{\alpha_i} \vec{\boldsymbol{u_1}} &= \vec{\boldsymbol{w_i}}
\end{aligned}$$

So, $\vec{\boldsymbol{w_i}}$ is in the span of $\vec{\boldsymbol{u_1}}, \vec{\boldsymbol{w_2}}, \ldots, \vec{\boldsymbol{w_n}}$ and we can drop $\vec{\boldsymbol{w_i}}$ from the list, W, and it will still span the vector space.
We can keep doing this with the remaining vectors in U - each time the vector to be removed will be some $\vec{\boldsymbol{w_i}}$ because all the $\vec{\boldsymbol{u_i}}$ are linearly independent - and all the while W remains a spanning list. We continue until we have replaced (potentially) all n vectors in W, which would happen if $m > n$. At this point we would have the spanning list $W = \vec{\boldsymbol{u_1}}, \vec{\boldsymbol{u_2}}, \ldots, \vec{\boldsymbol{u_n}}$ and $(m - n)$ remaining vectors in U.
Now, since W spans the space, the $(m - n)$ vectors that remain in U will be in the span of W. But, all the vectors that originally came from U were linearly independent, so it is impossible for any vectors in U to be in the span of W (which now comprises only vectors that originally came from U).
We therefore conclude that there can be no remaining vectors in U and, consequently that m cannot be greater than n, i.e. $m \leq n$. □

Summary of Span and Linear Independence

- *The span of a set of vectors changes when adding a vector not in the existing span.*
- *A linearly independent set of vectors continues to be linearly independent when adding a vector not in the existing span.*

The fundamental point of linear independence and span is that if the span of two sets of vectors is not completely disjoint (ignoring the zero vector) then the sets are not linearly independent. Any nonzero vector in common between the two spans implies a linear relation between the sets,

$$\begin{aligned} & \vec{u_1} + ... + \vec{u_n} = \vec{v} = \vec{w_1} + ... + \vec{w_n} \\ \iff \quad & \vec{0} = (\vec{w_1} + ... + \vec{w_n}) - (\vec{u_1} + ... + \vec{u_n}). \end{aligned}$$

2.4.6.2 Bases

Definition. *A **basis** for a vector space V is a set of vectors that is both linearly independent and spans the space V. The empty set is therefore a basis for the zero vector space $\{\vec{0}\}$.*
Since a basis of V spans the space, any vector in V may be expressed as as a linear combination of the vectors in the basis set and, since the basis set is linearly independent, this expression is unique.

*Compare with the **generating set of a group** (2.1.1.2).*

Proposition 71. *A set of vectors $B = \{\vec{v_1}, \ldots, \vec{v_n}\}$ in V is a basis iff every $\vec{w} \in V$ can be written in a single unique way as a linear combination of vectors in B.*

Proof. If B is a basis of V then, by definition, B spans V and so every $\vec{\boldsymbol{w}} \in V$ can be written as a linear combination of vectors in B. Furthermore, also by the definition of a basis, the vectors in B are linearly independent so, by corollary 17 the linear combination is unique.

Conversely if every $\vec{\boldsymbol{w}} \in V$ can be written in a single unique way as a linear combination of vectors in B then B both spans the space and is linearly independent by corollary 17. □

2.4.6.3 Examples of Bases

(37) If we take an arbitrary finite set of vectors $S = \vec{\boldsymbol{s_1}}, \ldots, \vec{\boldsymbol{s_n}}$ then the space $V = V(S)$ of linear combinations of elements of S is the set of all expressions of the form,

$$a_1\vec{\boldsymbol{s_1}}, \ldots, a_n\vec{\boldsymbol{s_n}}, \qquad a_i \in \mathbb{F}.$$

In this space addition and multiplication are carried out assuming no relations among the elements of S so that,

$$(a_1\vec{\boldsymbol{s_1}}+\cdots+a_n\vec{\boldsymbol{s_n}})+(b_1\vec{\boldsymbol{s_1}}+\cdots+b_n\vec{\boldsymbol{s_n}}) = (a_1+b_1)\vec{\boldsymbol{s_1}}+\cdots+(a_n+b_n)\vec{\boldsymbol{s_1}}$$

and

$$c(a_1\vec{\boldsymbol{s_1}} + \cdots + a_n\vec{\boldsymbol{s_n}}) = ca_1\vec{\boldsymbol{s_1}} + \cdots + ca_n\vec{\boldsymbol{s_n}}.$$

Then the mapping $\phi : \mathbb{F}^n \longmapsto V(S)$ defined as,

$$\phi(a_1, \ldots, a_n) = a_1\vec{\boldsymbol{s_1}}, \ldots, a_n\vec{\boldsymbol{s_n}}$$

is an isomorphism.

> *Note that if the assumption of no relation between the elements of S is not valid then ϕ may fail to be isomorphic.*

$V(S)$ is often referred to as *the space with basis S* or *the space of formal linear combinations of S*. If S is an infinite set then $V(S)$ is defined to be the set of all *finite* linear combinations of the elements of S.

This crops up frequently in applications, when taking weightings of different features for example; this isomorphism allows us to treat them as vectors in $\mathbb{F}^n$.

2.4.6.4 Finite-Dimensional Vector Spaces

Definition. *A vector space is called* ***finite-dimensional*** *if there is some finite set of vectors that spans the space.*

Proposition 72. *Any finite set which spans a finite-dimensional space contains a basis for the space.*

Proof. Let S be a spanning set of vectors in the space V.

If S is not linearly independent then there is some $\vec{\boldsymbol{v}} \in S$ such that $\vec{\boldsymbol{v}} \in span\,(S \setminus \{\vec{\boldsymbol{v}}\})$. So we can remove $\vec{\boldsymbol{v}}$ from the set and S still spans the space. We may continue doing this until S is linearly independent, at which point we have found the minimal subset of S that spans the space. This remaining subset is a basis of the space. □

Corollary 18. *Any finite-dimensional vector space has a basis.*

Proof. This follows from the previous proposition and the definition of a finite-dimensional vector space. □

Proposition 73. *Any set (including infinite sets) that spans a finite-dimensional vector space, contains a finite subset which spans the space.*

Proof. Let V be a finite-dimensional vector space and S be a spanning set of V. By the definition of V as finite-dimensional there exists some finite set that spans V. Let W be a finite set of vectors that spans V. Then, since S also spans V, every vector in W may be expressed as a finite linear combination of vectors in S. The set of all the members of S that participate in the linear combinations required to produce the set W is a finite subset of S that spans V. □

Proposition 74. *Let V be a finite-dimensional vector space. Any linearly independent set $L \subseteq V$ can be extended by adding vectors to obtain a basis of V.*

Proof. If L spans V then it is already a basis.
Assume L does not span V. Then there exists some $\vec{\boldsymbol{v}} \in V$ such that $\vec{\boldsymbol{v}} \notin \mathit{span}\, L$. If we add $\vec{\boldsymbol{v}}$ to L then the resulting set, say L', continues to be linearly independent and may or may not span V. If it does then L' is a basis. If not then we can continue to repeat the same process until it does span the space at which point we have a basis. □

Proposition 75. *For finite subsets of a vector space, any linearly independent set has cardinality less than or equal to that of any spanning set in the same space.*

> *Proposition 74 only tells us that, for any linearly independent set of vectors, there exists some basis whose cardinality is greater than or equal to that of the original set. What we want to prove here is the condition on the cardinality exists between any linearly independent set and spanning set in the same space.*

Proof. We will show two different ways of proving this: one using an algorithm on lists and the other using simultaneous equations.

(i) **proof using lists**

Let V be a finite-dimensional vector space, S a spanning list $\vec{\boldsymbol{s_1}}, \ldots, \vec{\boldsymbol{s_m}}$ and L a linearly independent list $\vec{\boldsymbol{l_1}}, \ldots, \vec{\boldsymbol{l_n}}$ in V, and assume that $m < n$.
Now, since S is a spanning list, every element of L is in its span and so if we remove the first element $\vec{\boldsymbol{l_1}}$ from L and add it to S then S will definitely contain a linear relation. If we then remove some element $\vec{\boldsymbol{s_i}}$ from the original spanning list that participates in a linear relation (i.e. $\vec{\boldsymbol{s_i}} \in \mathit{span}\, S$) then we have a modified list S, of the same length as the original, but with an element replaced by $\vec{\boldsymbol{l_1}}$ and this list continues to span the space.
We can repeat this task m times until all elements $\vec{\boldsymbol{s_i}}$ from the original spanning list have been removed and we have a spanning list $S = \vec{\boldsymbol{l_1}}, \ldots, \vec{\boldsymbol{l_m}}$ and the remaining $n - m$ elements are still in L. But the original $L = \vec{\boldsymbol{l_1}}, \ldots, \vec{\boldsymbol{l_n}}$ was linearly independent and no linear relation exists between them so it is impossible that the first m elements span

the space because this would mean that they would participate in a linear relation with the remaining $n - m$ elements.
So we have obtained a contradiction and we therefore conclude that $m \geq n \quad \square$.

(ii) **proof using simultaneous equations**

Let V be a vector space, S a finite spanning set $\vec{\boldsymbol{s_1}}, \ldots, \vec{\boldsymbol{s_m}}$ and L a finite linearly independent set $\vec{\boldsymbol{l_1}}, \ldots, \vec{\boldsymbol{l_n}}$ in V, and assume that $m < n$.
Since S spans the space, every $\vec{\boldsymbol{l_j}} \in L$ can be expressed as a linear combination of vectors $\vec{\boldsymbol{s_i}} \in S$ of the form,

$$\vec{\boldsymbol{l_j}} = a_{1j}\vec{\boldsymbol{s_1}} + \cdots + a_{mj}\vec{\boldsymbol{s_m}} = \sum_{i=1}^{m} a_{ij}\vec{\boldsymbol{s_i}}$$

for $1 \leq j \leq n$. A linear relation on the vectors of L would look like,

$$\begin{aligned} & c_1\vec{\boldsymbol{l_1}} + \cdots + c_n\vec{\boldsymbol{l_n}} = \vec{\boldsymbol{0}} \\ \iff \qquad & c_1 \sum_{i=1}^{m} a_{i1}\vec{\boldsymbol{s_i}} + \cdots + c_n \sum_{i=1}^{m} a_{in}\vec{\boldsymbol{s_i}} = \vec{\boldsymbol{0}} \end{aligned}$$

which expands into m simultaneous equations as follows.

$$\begin{gathered} c_1 a_{11}\vec{\boldsymbol{s_1}} + \cdots + c_n a_{1n}\vec{\boldsymbol{s_1}} = 0 \\ \vdots \\ c_1 a_{m1}\vec{\boldsymbol{s_m}} + \cdots + c_n a_{mn}\vec{\boldsymbol{s_m}} = 0 \end{gathered}$$

As we can see, each of the m equations has a has a factor $\vec{\boldsymbol{s_i}}$ in each term and so this may be factored out to give the following system.

$$\begin{gathered} \vec{\boldsymbol{s_1}}(c_1 a_{11} + \cdots + c_n a_{1n}) = 0 \\ \vdots \\ \vec{\boldsymbol{s_m}}(c_1 a_{m1} + \cdots + c_n a_{mn}) = 0 \end{gathered}$$

So now we have m equations such that the i-th equation will hold if either $\vec{\boldsymbol{s_i}} = \vec{\boldsymbol{0}}$ or $c_1 a_{i1} + \cdots + c_n a_{in} = 0$.

Assume that for all $\vec{s_i} \in S$, $\vec{s_i} \neq \vec{0}$. Then we end up with the following system.

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & \\ . & & \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ . \\ . \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ . \\ . \\ 0 \end{bmatrix}$$

It can be shown using matrix row reduction that a system such has this has non-trivial solutions if $m < n$. Therefore if $m < n$, there is a linear relation between the vectors of L contradicting its construction as linearly independent. Therefore $m \geq n$.

However, there may be some $\vec{s_i} \in S$ s.t. $\vec{s_i} = \vec{0}$ but only one because S is a set not a list. If that is all there is (i.e. $S = \{\vec{0}\}$) then the space $V = \{\vec{0}\}$ and the only linearly independent set of vectors in the zero space is the empty set. In this case $n = 0$ and so we cannot have $m < n$. Therefore $m \geq n$ also.
On the other hand, if there are non-zero elements in S then we can remove the equation i such that $\vec{s_i} = \vec{0}$ so that we will have $m - 1$ simultaneous equations in our system. But now the linear dependence of L follows if $(m-1) < n \iff m < (n+1)$ and clearly $m < n \implies m < n + 1$ so once again $m < n$ implies that L is not linearly independent.

Proposition 76. *Any two bases of the same finite-dimensional vector space have the same number of elements. In other words: For a given vector space, the cardinality of bases is fixed.*

Proof. Let L, L' be finite subsets of the finite-dimensional vector space V such that both are bases. Then both L and L' are linearly independent and span the space. Therefore we have,

$$|L| \leq \left|L'\right| \quad \text{and} \quad |L| \geq \left|L'\right|$$

which implies that $|L| = |L'|$. □

2.4.6.5 Dimension of Finite-Dimensional Vector Spaces

Definition. *The **dimension** of a finite-dimensional vector space v is the number of vectors in a basis. The dimension will be denoted by $dim\, V$.*

Theorem 20. *The dimension of a finite-dimensional vector space is an upper bound on the cardinality of a linearly independent set of vectors in the space and a lower bound on the cardinality of a spanning set of vectors in the same space.*

Proof. This theorem follows from Proposition 75. □

Theorem 21. *Any linearly independent set of vectors in a finite-dimensional vector space V of cardinality $dim\, V$ is a basis.*

Proof. Let $L \subset V$ be linearly independent set of vectors in V with $|L| = dim\, V$. By Proposition 74, any linearly independent set in V may be extended with zero or more vectors to obtain a basis ov V. But any basis of V has dimension $dim\, V = |L|$. Therefore we extend L with zero vectors to obtain a basis. □

Proposition 77. *If $W \subseteq V$ is a subspace of a finite-dimensional vector space then,*

$$dim\, W \leq dim\, V \quad and \quad (dim\, W = dim\, V) \iff (W = V).$$

Proof. Firstly note that W must also be finite-dimensional because by definition there is a finite set of vectors that spans V and since $W \subseteq V$, the same set must also span W.
Every basis of W is also a linearly independent set in V and so, by Proposition 75, it cannot have cardinality greater than any basis of V. Any basis in V has cardinality $dim\, V$ so the cardinality of any basis of W must be less than or equal to $dim\, V$. Therefore $dim\, W \leq dim\, V$.

If $dim\, W = dim\, V$ on the other hand, every basis of W has the same cardinality as every basis of V. Since every basis of W is also a linearly independent set in V with cardinality equal to $dim\, V$, by Theorem 21 it is a

basis of V. By similar logic in reverse, any basis of V is also a basis of W. Let B be such a basis. Then,

$$\vec{\boldsymbol{v}} \in W \iff \vec{\boldsymbol{v}} \in span\, B \iff \vec{\boldsymbol{v}} \in V.$$

Therefore $W = V$. □

Theorem 22. *Let W_1, W_2 be subspaces of a finite-dimensional vector space. Then,*

$$dim(W_1 + W_2) = dim\, W_1 + dim\, W_2 - dim(W_1 \cap W_2).$$

Proof. It is easy to show that the intersection of two subspaces is a subspace. So define a basis of $W_1 \cap W_2$ as $B = \{\vec{\boldsymbol{u_1}}, \ldots, \vec{\boldsymbol{u_r}}\}$ where $r = dim(W_1 \cap W_2)$. Then B is a linearly independent set in W_1 and can be extended to a basis of W_1,

$$B_{W_1} = \{\vec{\boldsymbol{u_1}}, \ldots, \vec{\boldsymbol{u_r}}, \vec{\boldsymbol{v_1}}, \ldots, \vec{\boldsymbol{v_{m-r}}}\}$$

where $m = dim\, W_1$.
By the same reasoning we can extend B to a basis of W_2,

$$B_{W_2} = \{\vec{\boldsymbol{u_1}}, \ldots, \vec{\boldsymbol{u_r}}, \vec{\boldsymbol{w_1}}, \ldots, \vec{\boldsymbol{w_{n-r}}}\}$$

where $n = dim\, W_2$.
Now if we can show that,

$$B' = B_{W_1} \cup B_{W_2} = \{\vec{\boldsymbol{u_1}}, \ldots, \vec{\boldsymbol{u_r}}, \vec{\boldsymbol{v_1}}, \ldots, \vec{\boldsymbol{v_{m-r}}}, \vec{\boldsymbol{w_1}}, \ldots, \vec{\boldsymbol{w_{n-r}}}\}$$

is a basis of $W_1 + W_2$ then the proof will follow easily.
Clearly, B' spans $W_1 + W_2$ as B' contains a basis of both W_1 and W_2 and so is able to express any sum of two vectors chosen from the two spaces.
Linear Independence is a little more complicated however. Consider that, since B_{W_1} is linearly independent, there can be no linear relation between the vectors $\vec{\boldsymbol{u_i}}$ and $\vec{\boldsymbol{v_i}}$ and similarly there can be no such relation between the vectors $\vec{\boldsymbol{u_i}}$ and $\vec{\boldsymbol{w_i}}$. Therefore, if there were to be a linear relation among the vectors it would have to involve the vectors $\vec{\boldsymbol{v_i}}$ with the vectors from B_{W_2} or the vectors $\vec{\boldsymbol{w_i}}$ with the vectors from B_{W_1}. If we model the linear relation as a relation between vectors that are linear combinations of the vectors $\vec{\boldsymbol{u_i}}, \vec{\boldsymbol{v_i}}$ and $\vec{\boldsymbol{w_i}}$ we get,

$$\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}} + \vec{\boldsymbol{w}} = \vec{\boldsymbol{0}} \iff \vec{\boldsymbol{v}} = -\vec{\boldsymbol{u}} - \vec{\boldsymbol{w}} \in W_2$$

therefore $\vec{\boldsymbol{v}} \in W_1 \cap W_2$ and so $\vec{\boldsymbol{v}}$ can be expressed as a linear combination of the vectors in B. But this implies a linear relation among the vectors of B_{W_1} which are linearly independent by construction. Therefore

$$\vec{\boldsymbol{v}} = \vec{\mathbf{0}} \implies -\vec{\boldsymbol{u}} - \vec{\boldsymbol{w}} = \vec{\mathbf{0}} \iff -\vec{\boldsymbol{u}} = \vec{\boldsymbol{w}}$$

which implies that there is a linear relation among the vectors of B_{W_2} which are also linearly independent by construction and so $\vec{\boldsymbol{u}} = \vec{\boldsymbol{w}} = \vec{\mathbf{0}}$. □

Corollary 19. *Let $W_1, \ldots, W_n$ be a set of subspaces of a finite-dimensional vector space. Then,*

$$dim(W_1 + \cdots + W_n) \leq dim\, W_1 + \cdots + dim\, W_n$$

with the equality case iff the spaces $W_1, \ldots, W_2$ are independent.

Proof. This follows by induction on Theorem 22.
If we take the case of $n = 2$ as our base case then,

$$dim(W_1 + W_2) = dim\, W_1 + dim\, W_2 - dim(W_1 \cap W_2) \leq dim\, W_1 + dim\, W_2.$$

Furthermore, if W_1, W_2 are independent then $W_1 \cap W_2 = \{\vec{\mathbf{0}}\}$ so that,

$$dim(W_1 + W_2) = dim\, W_1 + dim\, W_2 - 0 = dim\, W_1 + dim\, W_2.$$

For the induction step: Let $W' = W_1 + \cdots + W_{n-1}$ and take as the induction hypothesis that,

$$dim(W') \leq dim\, W_1 + \cdots + dim\, W_{n-1}$$

with equality being when the spaces are independent. Then observe that, by the associativity of addition of vectors and hence of vector spaces,

$$\begin{aligned} dim(W_1 + \cdots + W_n) &= dim((W_1 + \cdots + W_{n-1}) + W_n) \\ &= dim\, W' + dim\, W_n - dim(W' \cap W_n) && \text{by Theorem 22} \\ &\leq dim\, W' + dim\, W_n \\ &\leq dim\, W_1 + \cdots + dim\, W_{n-1} + dim\, W_n && \text{by induction hypothesis.} \end{aligned}$$

Furthermore, if $dim\, W' = dim\, W_1 + \cdots + dim\, W_{n-1}$ and if the spaces W' and W_n are independent then the intersection,

$$W' \cap W_n = \{\vec{\mathbf{0}}\}$$

which gives,

$$\begin{aligned} dim(W_1 + \cdots + W_n) &= dim((W_1 + \cdots + W_{n-1}) + W_n) \\ &= dim\, W' + dim\, W_n - 0 && \text{by Theorem 22} \\ &= dim\, W_1 + \cdots + dim\, W_{n-1} + dim\, W_n && \text{by induction hypothesis.} \end{aligned}$$

This shows that if the spaces are independent then

$$dim(W_1 + \cdots + W_n) = dim\, W_1 + \cdots + dim\, W_n$$

but it is also easy to reverse the logic and show that if the above equality holds then, if we add a space to the sum of other spaces, then the intersection of the added space with the sum of the other spaces must have dimension 0 and therefore must be $\{\vec{\mathbf{0}}\}$. In this way we can also prove the converse implication that the spaces must be independent. □

Proposition 78. *Two finite-dimensional vector spaces may be isomorphic only if they have the same dimension.*

Proof. We want to prove that if $\phi : W \longmapsto V$ is an isomorphism of vector spaces then it follows that $dim\, W = dim\, V$.
Assume for contradiction that ϕ is indeed an isomorphism between the vector spaces W and V but,

$$dim\, W = m > dim\, V = n.$$

Then any basis of $\vec{\boldsymbol{w_1}}, \ldots, \vec{\boldsymbol{w_m}} \in W$ is a linearly independent set of vectors in W with the property, for $c_1, \ldots, c_m \in \mathbb{F}$,

$$\begin{aligned} & c_1\vec{\boldsymbol{w_1}} + \cdots + c_m\vec{\boldsymbol{w_m}} = \vec{\mathbf{0}} \iff c_1, \ldots, c_m = 0 \\ \iff \quad & \phi(c_1\vec{\boldsymbol{w_1}} + \cdots + c_m\vec{\boldsymbol{w_m}}) = \phi(\vec{\mathbf{0}}) = \vec{\mathbf{0}} \iff c_1, \ldots, c_m = 0 \\ \iff \quad & c_1\phi(\vec{\boldsymbol{w_1}}) + \cdots + c_m\phi(\vec{\boldsymbol{w_m}}) = \vec{\mathbf{0}} \iff c_1, \ldots, c_m = 0. \end{aligned}$$

But we have $\phi(\vec{\boldsymbol{w_1}}), \ldots, \phi(\vec{\boldsymbol{w_m}}) \in V$ so that the result obtained implies that $\phi(\vec{\boldsymbol{w_1}}), \ldots, \phi(\vec{\boldsymbol{w_m}})$ is a linearly independent set of vectors in V of cardinality $m > n = dim\, V$. By Theorem 20 this cannot be and we have obtained a contradiction. □

TODO: example application of calculating the order of $GL_2(\mathbb{F})$ when $\mathbb{F} = \mathbb{F}_p$ is a prime field. Artin[114]

2.4.6.6 Direct Sums in Finite-Dimensional Vector Spaces

Proposition 79. *Let $W_1, \ldots, W_n$ be subspaces of a finite-dimensional vector space V, and let B_i be a basis for W_i. Then, the ordered set B obtained by listing the bases $B_1, \ldots, B_n$ in order is a basis of V iff V is the direct sum $W_1 \otimes \cdots \otimes W_n$.*

Proof. Let $B = \bigcup_i B_i$ be a basis of V. Then B is a linearly independent set and so,

$$\forall \vec{\boldsymbol{b}} \in B \,.\, \vec{\boldsymbol{b}} \notin span(B \setminus \{\vec{\boldsymbol{b}}\})$$

which also means that,

$$\forall B_i \subset B \,.\, span\, B_i \cap span(B \setminus B_i) = \vec{\mathbf{0}}.$$

This implies that the sets B_i are independent spaces. Since their union is a basis of V then together they span V. Since the span of each B_i is W_i then, $V = W_1 \oplus \cdots \oplus W_n$.
Conversely, if $V = W_1 \oplus \cdots \oplus W_n$ then each subspace W_i is independent so that, for every $\vec{\boldsymbol{b}} \in B_i$, $\vec{\boldsymbol{b}} \notin span(B \setminus B_i)$. This means that if we begin with B_1 and add to it the elements of B_2 then the resulting set remains linearly independent and we can continue this until we have $B = \bigcup_i B_i$ as a linearly independent set. Then B contains all the basis vectors of every W_i and therefore, by $V = W_1 \oplus \cdots \oplus W_n$, B spans the space V. It is therefore a basis of V. □

Proposition 80. *Let W be a subspace of a finite-dimensional vector space V. Then there is another subspace W' such that $W \oplus W' = V$.*

Proof. Any basis of W is a linearly independent set in V and can be extended to a basis of V by adding a set of linearly independent vectors S. Then S is a basis of a subspace W' such that $W \oplus W' = V$. □

2.4.7 Infinite Sets of Vectors

The definition of a vector space defines what it means to add two vectors and so, by extension, arbitrarily large *finite* sums of vectors but not what it means to add an infinite number of vectors. Therefore, we define the span and linear independence of infinite sets of vectors as conditions over finite subsets of the vectors.

2.4.7.1 Span and Linear Independence of Infinite Sets of Vectors

Definition. *The* ***span of an infinite set of vectors*** S *is defined to be the set of* ***finite*** *linear combinations of its elements,*

$$\{\, c_1\vec{\boldsymbol{v}_1} + \cdots + c_r\vec{\boldsymbol{v}_r} \mid \vec{\boldsymbol{v}_i} \in S,\, c_i \in \mathbb{F} \,\}$$

where r is finite but may be arbitrarily large.

Definition. *An* ***infinite set of vectors*** *is defined to be* ***linearly independent*** *if there is no linear relation among a* ***finite*** *subset of them,*

$$c_1\vec{\boldsymbol{v}_1} + \cdots + c_n\vec{\boldsymbol{v}_n} = \vec{\mathbf{0}}$$

where $c_i \in F$.

These definitions of span and linear independence are compatible with the corresponding definitions for finite sets of vectors.

2.4.7.2 Infinite-Dimensional Vector Spaces

Definition. *A vector space is called* ***infinite-dimensional*** *if there is no finite set of vectors that spans the space.*

2.4.7.3 Examples of Infinite-Dimensional Vector Spaces

(38) The space $\mathbb{R}^\infty$ of infinite real vectors $(a) = (a_1, a_2, a_3, \dots)$ can also be thought of as the space of sequences $\{a_n\}$ of real numbers. It has many important subspaces:

a. Convergent sequences: $C = \{\, (a) \in \mathbb{R}^\infty \mid \lim_{n\to\infty} a_n \text{ exists} \,\}$.

b. Bounded sequences: $l^\infty = \{\, (a) \in \mathbb{R}^\infty \mid \{a_n\} \text{ is bounded} \,\}$.

c. Absolutely convergent series: $l^1 = \{\, (a) \in \mathbb{R}^\infty \mid \sum_1^\infty |a_n| < \infty \,\}$.

d. Sequences with finitely many nonzero terms:

$$Z = \{\, (a) \in \mathbb{R}^\infty \mid a_n = 0 \text{ for all but finitely many } n \,\}.$$

2.4.7.4 Bases of Infinite-Dimensional Vector Spaces

Definition. *As with finite sets, a* ***basis of an infinite-dimensional vector space*** *is a linearly independent set which spans the space.*

(39) Let $S = (e_1, e_2, \dots)$ be the infinite set of standard basis vectors in $\mathbb{R}^\infty$. S does not span $\mathbb{R}^\infty$ because the vector $\vec{\boldsymbol{w}} = (1, 1, 1, \dots)$ is not a *finite* linear combination of the elements of S. It *is*, however, a basis of the vector space Z in 38d.

It can be shown, using the Axiom of Choice, that every vector space has a basis (Theorem 1) but a basis of $\mathbb{R}^\infty$ will be uncountably infinite and so will not be expressable as $(\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}}, \dots)$.

From the University of Michigan Maths Dept. Linear Algebra Supplement on Infinite-Dimensional Vector Spaces:

Let $\mathbb{R}$ be the set of real numbers considered as a vector space over the field $\mathbb{Q}$ of rational numbers. What could possibly be a basis? The elements $\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{6}, \dots$ can be shown to be linearly independent, but they certainly don't span $\mathbb{R}$, as we also need elements like $\pi, \pi^2, \pi^3, \dots$, which also form a linearly independent set. In fact, because $\mathbb{Q}$ is countable, one can show that the subspace of $\mathbb{R}$ generated by any countable subset of $\mathbb{R}$ must be countable. Because $\mathbb{R}$ itself is uncountable, no countable set can be a basis for $\mathbb{R}$ over $\mathbb{Q}$. This means that any basis for $\mathbb{R}$ over $\mathbb{Q}$, if one exists, is going to be difficult to describe.

If we were to look for a basis of the functions over the reals we could consider the indicator functions for all $r \in \mathbb{R}$,

$$i_r(x) \coloneqq \begin{cases} 1 & x = r \\ 0 & x \neq r \end{cases}.$$

This set of functions is clearly linearly independent and spans the set of functions $\mathbb{R} \longmapsto \mathbb{R}$. Since there is one such indicator function for each real number and the real numbers are uncountable, this set of indicator functions is also uncountable.

Theorem 23. *A linear operator over an infinite-dimensional vector space does not conform to the Dimension Formula for finite-dimensional vector spaces.*

Proof. The shift operator described in a note to Proposition 95) is an example of a linear transformation over an infinite-dimensional vector space that doesn't conform to the dimension formula. As noted there, the reason is that infinite sets may have proper subsets of equal cardinality. So we can have an image that is clearly "smaller" than the space but nevertheless has the same dimensionality and so the map has a trivial kernel anyway.

Conversely the differentiation operator has a non-trivial kernel (the set of constant polynomials) but when differentiating an infinite power series the result is still an infinite power series. □

There is much more to this topic. For more details see the University of Michigan Maths Dept. Linear Algebra Supplement on Infinite-Dimensional Vector Spaces.

2.4.8 Coordinate Vector Spaces

Definition. *A **coordinate vector** is a representation of a vector as an ordered list of numbers that describes the vector in terms of a particular ordered basis.*

Definition. *Let $B = \{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n\}$ be a basis of a finite-dimensional vector space V. Then for every $\vec{\boldsymbol{v}} \in V$ it is possible to express $\vec{\boldsymbol{v}}$ as a linear combination of the vectors in B in the form,*

$$\vec{\boldsymbol{v}} = c_1\vec{\boldsymbol{v}}_1 + \cdots + c_n\vec{\boldsymbol{v}}_n.$$

*Then, the **coordinate vector of $\vec{\boldsymbol{v}}$ with respect to the basis** B is,*

$$\vec{\boldsymbol{v}}_B = \langle c_1, \ldots, c_n \rangle.$$

*The definition of a basis 2.4.6.2 requires that it is a set of **linearly independent** vectors. It is worth noting what results if we attempt to use a linearly dependent set of vectors as a basis because in real-world situations this often occurs (for example in data analysis) when the full relationship between objects may not be known.*
Imagine a "basis" $B = \{x, 3x\}$. Co-ordinate vectors defined against this basis are not unique —

$$(3)x + (1)3x = (6)x + (0)3x = \cdots$$

In fact, from the basis and the co-ordinate vectors, we appear to working in a 2-dimensional space but in fact, because there is a linear relation between the two "basis" vectors we are actually working in a 1-dimensional space. As a result, algebraic analysis will be missing one relation. For example, take the vectors against "basis" B,

$$\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \end{bmatrix} \right\}.$$

It appears that there is a single linear relation between these vectors given by the nullspace of their matrix as follows:

$$\begin{bmatrix}1 & 1 & 1\\ 0 & 3 & 5\end{bmatrix} \rightsquigarrow \begin{bmatrix}1 & 0 & -2/3\\ 0 & 1 & 5/3\end{bmatrix} \implies \textit{nullspace} = t\begin{bmatrix}2/3\\ -5/3\\ 1\end{bmatrix} \qquad \textit{for } t \in \mathbb{R}.$$

Whereas in actual fact all vectors in this space are colinear, being of the form αx for some $\alpha \in \mathbb{R}$. If we apply the co-ordinate vectors to their "basis" we see that,

$$\begin{bmatrix}1\\0\end{bmatrix} = (1)x = \boldsymbol{x}, \begin{bmatrix}1\\3\end{bmatrix} = (1)x + (3)3x = \mathbf{10}\boldsymbol{x}, \begin{bmatrix}1\\5\end{bmatrix} = (1)x + (5)3x = \mathbf{16}\boldsymbol{x}.$$

Then we can see that the true extent of the relationship between the vectors is given by,

$$\begin{bmatrix}1 & 10 & 16\end{bmatrix} \rightsquigarrow s\begin{bmatrix}-10\\1\\0\end{bmatrix} + t\begin{bmatrix}-16\\0\\1\end{bmatrix}$$

for $s, t \in \mathbb{R}$.
As we can see, the true relation (with respect to x) between the vectors is 2-dimensional. The relation found when using the "basis" B is a 1-dimensional subspace within this space corresponding to the values $s = -5/3,\ t = 1$.

Definition. *The **dot product** of two coordinate vectors in an n-dimensional coordinate space is defined as,*

$$\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} = v_1 w_1 + \cdots + v_n w_n = \vec{\boldsymbol{v}}^T \vec{\boldsymbol{w}}.$$

Notation. The **dot product** may also be denoted $\langle \vec{\boldsymbol{v}}, \vec{\boldsymbol{w}} \rangle$ but this notation may also refer to a literal coordinate vector, i.e.

$$\begin{bmatrix}3\\7\end{bmatrix} = \langle 3, 7 \rangle = (3, 7)^T.$$

Proposition 81. *The properties of the dot product are:*

(i) $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{x}} \rangle$

(ii) $\alpha\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \alpha\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \vec{\boldsymbol{x}}, \alpha\vec{\boldsymbol{y}} \rangle$

(iii) $\langle \vec{\boldsymbol{x}} + \vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \rangle = \langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{z}} \rangle + \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \rangle$

(iv) $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{x}} \rangle \geq 0$ *and* $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{x}} \rangle = 0 \iff \vec{\boldsymbol{x}} = 0$

Proof. Proofs of these properties follow from the definition of the dot product.

(i) $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{x}} \rangle$ follows from the commutativity of multiplication in a field,

$$x_i y_i = y_i x_i.$$

(ii) $\alpha\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \alpha\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \rangle = \langle \vec{\boldsymbol{x}}, \alpha\vec{\boldsymbol{y}} \rangle$ follows from the associativity of multiplication in a field,

$$\alpha x_i y_i = x_i \alpha y_i.$$

(iii) $\langle \vec{\boldsymbol{x}} + \vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \rangle = \langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{z}} \rangle + \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \rangle$ follows from the distributivity of multiplication over addition in a field,

$$(x_i + y_i) z_i = x_i z_i + y_i z_i.$$

(iv) $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{x}} \rangle \geq 0$ and $\langle \vec{\boldsymbol{x}}, \vec{\boldsymbol{x}} \rangle = 0 \iff \vec{\boldsymbol{x}} = 0$ follows from its form as a sum of squares,

$$x_1^2 + x_2^2 + \cdots + x_n^2.$$

□

2.4.8.1 Examples of coordinate vectors

(40) Let $\vec{\boldsymbol{x}} = \langle 3, 3 \rangle \in \mathbb{R}^2$ and B be the set $\{\langle 2, 0 \rangle, \langle 1, 3 \rangle\}$ so that B is a non-standard basis of the space $\mathbb{R}^2$. Then we can also define $\vec{\boldsymbol{x}}$ with respect to the basis B as follows.

$$\vec{\boldsymbol{x}} = 1 \cdot \begin{bmatrix} 2 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Therefore, the **coordinate vector of $\vec{\boldsymbol{x}}$ with respect to the basis** B is defined as,

$$\vec{\boldsymbol{x}}_B = \langle 1, 1 \rangle.$$

2.4.8.2 Bases as Matrices

Notation. If $B = \{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n\}$ is a set of vectors then, when it seems necessary to clearly differentiate, the matrix whose columns are the elements of B will be denoted $[B]$. However, for convenience, when the mathematical context makes clear whether we are referring to a set of vectors or a matrix we will use B to indicate either the set or the matrix.

Proposition 82. *If $\vec{\boldsymbol{x}}_B$ is a coordinate vector of $\vec{\boldsymbol{x}}$ w.r.t. the basis B then left multiplication by the matrix $[B]$ whose columns are the elements of B converts $\vec{\boldsymbol{x}}_B$ to its standard coordinate form $\vec{\boldsymbol{x}}$,*

$$\vec{\boldsymbol{x}} = [B]\vec{\boldsymbol{x}}_B.$$

Proof. Let $B = \{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n\}$ be a basis and the basis matrix

$$[B] = \begin{bmatrix} \vec{\boldsymbol{v}}_1 & \ldots & \vec{\boldsymbol{v}}_n \end{bmatrix}$$

and $\vec{\boldsymbol{x}} = \langle x_1, \ldots, x_n \rangle$ is a vector in standard coordinates. Then if we calculate $\vec{\boldsymbol{x}}$ using the basis B,

$$\vec{\boldsymbol{x}} = x_{B1}\vec{\boldsymbol{v}}_1 + \cdots + x_{Bn}\vec{\boldsymbol{v}}_n$$

we get a coordinate vector w.r.t. B,

$$\vec{\boldsymbol{x}}_B = \langle x_{B1}, \ldots, x_{Bn} \rangle.$$

So, clearly, to recover the standard coordinate vector we need to apply $\vec{\boldsymbol{x}}_B$ to the basis against which it was defined,

$$[B]\vec{\boldsymbol{x}}_B = \begin{bmatrix} \vec{\boldsymbol{v}}_1 & \ldots & \vec{\boldsymbol{v}}_n \end{bmatrix} \begin{bmatrix} x_{B1} \\ \vdots \\ x_{Bn} \end{bmatrix} = x_{B1}\vec{\boldsymbol{v}}_1 + \cdots + x_{Bn}\vec{\boldsymbol{v}}_n = \vec{\boldsymbol{x}}. \qquad \square$$

Corollary 20. *If $\vec{\boldsymbol{x}}$ is a coordinate vector w.r.t. the standard basis and B is an alternative basis then left multiplication by the matrix $[B]^{-1}$ converts $\vec{\boldsymbol{x}}$ to $\vec{\boldsymbol{x}}_B$ its form w.r.t. the basis B.*

Proof.

$$[B]\vec{\boldsymbol{x}}_B = \vec{\boldsymbol{x}} \iff \vec{\boldsymbol{x}}_B = [B]^{-1}\vec{\boldsymbol{x}}.$$

$\square$

One way to think of the action of the basis matrix is that it is used to encode/decode coordinates into/from its basis coordinates.
For example, if $\vec{\boldsymbol{x}}_B$ is a coordinate vector w.r.t. the basis B, then left-multiplying it by the basis matrix $[B]$,

$$\vec{\boldsymbol{x}} = [B]\vec{\boldsymbol{x}}_B$$

decodes the coordinate vector into standard coordinates. Conversely we can use the inverse of the basis matrix,

$$\vec{\boldsymbol{x}}_B = [B]^{-1}\vec{\boldsymbol{x}}$$

to encode a standard coordinate vector into B-coordinates.

2.4.8.3 Relationship with Abstract Vector Spaces

Proposition 83. *Every vector space V of dimension n is isomorphic to the space $\mathbb{F}^n$ of column vectors.*

Proof. Let $\phi : \mathbb{F}^n \longmapsto V$ be defined as $\phi(\vec{\boldsymbol{x}}) = B\vec{\boldsymbol{x}}$ where B is a matrix whose columns are a basis of V. The map ϕ is surjective because the columns of B span the space V and it is injective because they are linearly independent. So ϕ is a bijection.
The structure of the vector space is preserved because,

$$\phi(\vec{\boldsymbol{x_1}} + \vec{\boldsymbol{x_2}}) = B(\vec{\boldsymbol{x_1}} + \vec{\boldsymbol{x_2}}) = B\vec{\boldsymbol{x_1}} + B\vec{\boldsymbol{x_2}} = \phi(\vec{\boldsymbol{x_1}}) + \phi(\vec{\boldsymbol{x_2}})$$

and

$$\phi(c\vec{\boldsymbol{x}}) = B(c\vec{\boldsymbol{x}}) = cB\vec{\boldsymbol{x}} = c\phi(\vec{\boldsymbol{x}}).$$

□

*Note that, by Proposition 78, $\mathbb{F}^n$ is **not** isomorphic to $\mathbb{F}^m$ for $m \neq n$. Every finite-dimensional vector space V is isomorphic to $\mathbb{F}^n$, for some uniquely determined integer n.*
So, the finite-dimensional vector spaces are completely classified by Proposition 83 and any problem on finite-dimensional vector spaces may

be reduced to a problem on column vectors and matrices.

It is a result of Proposition 83 that we can use coordinate spaces as vector spaces.

2.4.8.4 Using Coordinate Spaces to analyse Vectors

Span If we have a set of n vectors in $\mathbb{F}^m$ then we can determine if a vector $\vec{\boldsymbol{b}}$ is in the span of the set of vectors by solving the system,

$$A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$$

where $\vec{\boldsymbol{x}} \in \mathbb{F}^n$ and A is an $m \times n$ matrix whose columns are the set of vectors. If there is some $\vec{\boldsymbol{x}}$ that satisfies the equation then $\vec{\boldsymbol{b}}$ is in the span.

Linear Independence If we have a set of n vectors in $\mathbb{F}^m$ then we can determine linear independence by solving a system of homogeneous linear equations,

$$A\vec{\boldsymbol{x}} = \vec{\mathbf{0}}$$

where $\vec{\boldsymbol{x}} \in \mathbb{F}^n$ and A is an $m \times n$ matrix whose columns are the set of vectors. If there is a non-trivial solution — a non-zero $\vec{\boldsymbol{x}}$ for which $A\vec{\boldsymbol{x}} = \vec{\mathbf{0}}$ — then the set of vectors is not linearly independent.

2.4.8.5 Change of Basis

Definition. *If we represent an abstract vector $\vec{\boldsymbol{v}}$ as a coordinate vector with respect to two different bases $B = \{\vec{\boldsymbol{b}}_1, \ldots, \vec{\boldsymbol{b}}_n\}$ and $B' = \{\vec{\boldsymbol{b}}'_1, \ldots, \vec{\boldsymbol{b}}'_n\}$ then we have,*

$$\vec{\boldsymbol{v}} = a_1\vec{\boldsymbol{b}}_1 + \cdots + a_n\vec{\boldsymbol{b}}_n = a'_1\vec{\boldsymbol{b}}'_1 + \cdots + a'_n\vec{\boldsymbol{b}}'_n.$$

Using the notation $\vec{\boldsymbol{x}}_B = \langle a_1, \ldots, a_n\rangle$, $\vec{\boldsymbol{x}}_{B'} = \langle a'_1, \ldots, a'_n\rangle$ and letting B, B' from here on refer to the matrices whose columns are the elements of B, B', we can rewrite the previous equation with matrices as,

$$\vec{\boldsymbol{v}} = B\vec{\boldsymbol{x}}_B = B'\vec{\boldsymbol{x}}_{B'}.$$

*Then the **change of basis** from B to B' is the mapping,*

$$\vec{\boldsymbol{x}}_{B'} = (B')^{-1}B\vec{\boldsymbol{x}}_B = P\vec{\boldsymbol{x}}_B.$$

So we have,

$$(B')^{-1}B = P \iff B = B'P$$

*which shows that P is the mapping between the two bases. This is known as the **matrix of change of basis**.*

Note that another way to think about this is to say that $\vec{\boldsymbol{x}}$ is the coordinate vector of $\vec{\boldsymbol{v}}$ with respect to the standard basis and,

$$\vec{\boldsymbol{x}} = B\vec{\boldsymbol{x}}_B = B'\vec{\boldsymbol{x}}_{B'}$$

so we always use the standard basis as a reference.

Proposition 84. *If $B \in \mathbb{F}^{n\times n}$ is a basis of a finite vector space V then, for $P \in GL_n(\mathbb{F})$,*

$$BP^{-1} = B'$$

is another basis of V.

Proof. As a member of $GL_n(\mathbb{F})$, P is invertible and so is a bijective mapping. It follows then that each of the basis vectors forming the columns of B are in the span of the columns of B' and so the columns of B' must also span the space V. Furthermore, since we must also have $B' \in \mathbb{F}^{n\times n}$, there are n columns in B' and so they are a spanning set with cardinality equal to the set of columns of B which is a basis of V. Therefore, by Theorem 21, the columns of B' are a basis of V. □

2.4.8.6 Matrix of Change of Basis

Definition. *Let B, B' be two different bases of the same space. Then, the matrix P such that,*

$$B = B'P \quad \textit{and} \quad \vec{\boldsymbol{x}}_{B'} = P\vec{\boldsymbol{x}}_B$$

is known as the ***matrix of change of basis*** *between B and B'.*

The matrix of change of basis P contains the vectors of one basis defined w.r.t. to the other basis. So,

$$B = B'P$$

tells us that if we apply the coordinate vectors in P to the basis B' we get the basis vectors of B.
Conversely, if we have a coordinate vector $\vec{\boldsymbol{x}}_B$ defined w.r.t. to B and we treat P as though it were a basis and apply the coordinate vector $\vec{\boldsymbol{x}}_B$ to it,

$$\vec{\boldsymbol{x}}_{B'} = P\vec{\boldsymbol{x}}_B$$

we get the same vector defined w.r.t. to the basis B'.

2.4.8.7 Euclidean Coordinate Spaces

Definition. *The **Euclidean coordinate spaces** are the real coordinate spaces $\mathbb{R}^n$ equipped with the **standard norm**,*

$$|\vec{\boldsymbol{v}}| = \sqrt{v_1^2 + \cdots + v_n^2}$$

*which is considered to be the **length** of the vector in the space $\mathbb{R}^n$.*

- *The Euclidean norm (or standard norm) is the square root of the dot product of the vector with itself,*

$$\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} = v_1^2 + \cdots + v_n^2 = |\vec{\boldsymbol{v}}|^2 .$$

- *In a Euclidean coordinate space the metric (wikipedia:metric), or distance function, is the length of the vector between the two points considered to be position vectors in the corresponding Euclidean vector space. So, the distance d between the point whose position vector is $\vec{\boldsymbol{v}}$ and the point whose position vector is $\vec{\boldsymbol{w}}$ is,*

$$d = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}| = |\vec{\boldsymbol{w}} - \vec{\boldsymbol{v}}| .$$

Proposition 85. *The dot product of two distinct vectors in a Euclidean space is related to their lengths by the formula,*

$$\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} = |\vec{\boldsymbol{v}}| \, |\vec{\boldsymbol{w}}| \cos\theta$$

where θ is the angle between the vectors.

Proof. The properties of the dot product (Proposition 81) tell us that,

$$\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} = \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{v}} \quad \text{and} \quad (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot \vec{\boldsymbol{x}} = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{x}} - \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}}$$

so if we let $\vec{\boldsymbol{x}} = (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})$ also we obtain,

$$(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) = \vec{\boldsymbol{v}} \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) - \vec{\boldsymbol{w}} \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})$$

$$= \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} - \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} - \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{v}} + \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{w}}$$
$$= \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} + \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{w}} - 2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}}.$$

So this gives us a formula for the square of the length of the displacement vector between $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{w}}$ or, in other words, the square of the distance between $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{w}}$. Now, if we imagine $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{w}}$ and the vector $\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}$ forming a triangle we can also refer to the law of cosines which tells us that,

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

where θ is the angle subtended by a and b and c is the side opposite the angle. Letting,

$$a = |\vec{\boldsymbol{v}}|\,,\ b = |\vec{\boldsymbol{w}}|\,,\ c = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|$$

we obtain,

$$|\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|^2 = |\vec{\boldsymbol{v}}|^2 + |\vec{\boldsymbol{w}}|^2 - 2\,|\vec{\boldsymbol{v}}|\,|\vec{\boldsymbol{w}}|\cos\theta.$$

But the Euclidean norm is the square root of the dot product so also,

$$|\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|^2 = (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})$$

therefore

$$\begin{aligned}
& |\vec{\boldsymbol{v}}|^2 + |\vec{\boldsymbol{w}}|^2 - 2\,|\vec{\boldsymbol{v}}|\,|\vec{\boldsymbol{w}}|\cos\theta = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} + \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{w}} - 2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\
\iff \quad & |\vec{\boldsymbol{v}}|^2 + |\vec{\boldsymbol{w}}|^2 - 2\,|\vec{\boldsymbol{v}}|\,|\vec{\boldsymbol{w}}|\cos\theta = |\vec{\boldsymbol{v}}|^2 + |\vec{\boldsymbol{w}}|^2 - 2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\
\iff \quad & -2\,|\vec{\boldsymbol{v}}|\,|\vec{\boldsymbol{w}}|\cos\theta = -2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\
\iff \quad & |\vec{\boldsymbol{v}}|\,|\vec{\boldsymbol{w}}|\cos\theta = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}}.
\end{aligned}$$

□

This relationship is the foundation of analytic geometry.

2.4.8.8 Orthogonality

Definition. *Two nonzero vectors $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{w}}$ are said to be* ***orthogonal*** *if their dot product is zero, i.e.*

$$\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} = 0.$$

Theorem 24. *Two nonzero vectors $\vec{v}$ and $\vec{w}$ are perpendicular to each other if and only if their dot product is zero.*

Proof. If the dot product $\vec{v} \cdot \vec{w} = 0$ then, by Proposition 85,

$$\vec{v} \cdot \vec{w} = |\vec{v}|\,|\vec{w}| \cos\theta = 0$$

where θ is the angle between $\vec{v}$ and $\vec{w}$. Since $\vec{v}$ and $\vec{w}$ are both nonzero, the product of their lengths $|\vec{v}|\,|\vec{w}| > 0$. Then we must have $\cos\theta = 0 \iff \theta \in \{\pi/2, 3\pi/2\}$. So, $\vec{v} \cdot \vec{w} = 0$ implies that the vectors are perpendicular.
Conversely, we can use the same logic in reverse to show that, for perpendicular vectors, $\cos\theta = 0$ which means that the dot product will be 0. □

Corollary 21. *Geometric orthogonality — that's to say perpendicularity — is equivalent to the dot product definition of orthogonality.*

2.5 Linear Transformations

2.5.1 Basic Properties of Linear Transformations

The analogue for vector spaces of a homomorphism of groups is a map,

$$T : V \longmapsto W$$

from one vector space over a field $\mathbb{F}$ to another, which is compatible with addition and scalar multiplication:

$$T(\vec{\boldsymbol{v_1}} + \vec{\boldsymbol{v_2}}) = T(\vec{\boldsymbol{v_1}}) + T(\vec{\boldsymbol{v_2}}) \quad \text{and} \quad T(c\vec{\boldsymbol{v_1}}) = cT(\vec{\boldsymbol{v_1}}),$$

for all $\vec{\boldsymbol{v_1}}, \vec{\boldsymbol{v_2}} \in V$, $c \in \mathbb{F}$.

Note that another way of describing this is that ***linear combinations are preserved across linear transformations****. That's to say, if*

$$\vec{\boldsymbol{u}} = \alpha_1\vec{\boldsymbol{v}}_1 + \cdots + \alpha_n\vec{\boldsymbol{v}}_n \quad \textit{and} \quad \vec{\boldsymbol{w}} = \alpha_1 f(\vec{\boldsymbol{v}}_1) + \cdots + \alpha_n f(\vec{\boldsymbol{v}}_n)$$

then, if f is linear we also have,

$$f(\vec{\boldsymbol{u}}) = \vec{\boldsymbol{w}}.$$

Definition. *A homomorphism between two vector spaces that is also compatible with scalar multiplication is called a* ***linear transformation*** *or* ***linear map or mapping****.*

The compatibility with addition of vectors implies that a linear transformation is a homomorphism between additive groups of vectors.

Linear transformations ***preserve*** *linear combinations in their arguments but this must be distinguished carefully from* ***being equal*** *to linear combinations of their arguments. A typical linear transformation is* ***not*** *expressible as a linear combination of it's sole argument and — in fact — the image of a vector under a linear map only fails to be linearly independent to the original vector in the case that the original vector is an eigenvector $A\vec{\boldsymbol{v}} = c\vec{\boldsymbol{v}}$.*
It is, however, worth noting that a linear map between finite vector spaces may be thought of as the application of a coordinate vector to a different basis than that against which it was originally defined. Therefore, any linear combination of objects may be thought of as a linear map between the space of coefficients and the space of objects. For example, the linear combination,

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

may be thought of as a linear map from the space of coefficients α_i to the space of objects x_i,

$$T\left(\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{bmatrix}\right) = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n.$$

Corollary 22. *As with all homomorphisms, linear maps always map the identity to the identity. For linear maps this means mapping the zero vector to the zero vector.*

Proposition 86. *A linear map is homogeneous of degree 1.*

Proof. $L(\alpha\vec{\boldsymbol{v}}) = \alpha L(\vec{\boldsymbol{v}})$ for all $\alpha \in \mathbb{F},\ \vec{\boldsymbol{v}} \in V$. □

Proposition 87. *Linear Dependence is always preserved across any linear transformation.*

Proof. Let $\alpha_1\vec{\boldsymbol{v}}_1 + \alpha_2\vec{\boldsymbol{v}}_2 + \cdots + \alpha_n\vec{\boldsymbol{v}}_n = \vec{\mathbf{0}}$ be a linear relation between the vectors $\{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n\}$. If L is a linear map then,

$$\begin{aligned} & L(\alpha_1\vec{\boldsymbol{v}}_1 + \alpha_2\vec{\boldsymbol{v}}_2 + \cdots + \alpha_n\vec{\boldsymbol{v}}_n) = L(\vec{\mathbf{0}}) \\ \iff \qquad & \alpha_1 L(\vec{\boldsymbol{v}}_1) + \alpha_2 L(\vec{\boldsymbol{v}}_2) + \cdots + \alpha_n L(\vec{\boldsymbol{v}}_n) = \vec{\mathbf{0}}. \end{aligned}$$

□

2.5.1.1 The Kernel and the Image of a Linear Transformation

Definition. *Let $T : V \longmapsto W$ be any linear transformation. Then the* ***kernel (or nullspace)*** *of T is defined as,*

$$ker\, T = \{\, \vec{\boldsymbol{v}} \mid T(\vec{\boldsymbol{v}}) = \vec{\mathbf{0}} \,\}$$

and the ***image*** *of T as,*

$$im\, T = \{\, \vec{\boldsymbol{w}} \in W \mid \exists \vec{\boldsymbol{v}} \in V \,.\, \vec{\boldsymbol{w}} = T(\vec{\boldsymbol{v}}) \,\}.$$

Proposition 88. *The kernel of $T : V \longmapsto W$ is a subspace of V and the image is a subspace of W.*

Proof. T is a homomorphism between additive groups of vectors and so the proof that the kernel and image are subspaces is the same as for general homomorphisms (see 2.1.4.1). □

Proposition 89. *The fibres of a linear transformation are the additive cosets of the kernel.*

Proof. Let $T : V \longmapsto W$ be any linear transformation with kernel $K = ker\, T$. Then, for any fixed $\vec{\boldsymbol{v}} \in V$,

$$\forall \vec{\boldsymbol{k}} \in K \,.\, T(\vec{\boldsymbol{v}} + \vec{\boldsymbol{k}}) = T(\vec{\boldsymbol{v}}) + T(\vec{\boldsymbol{k}}) = T(\vec{\boldsymbol{v}}) + \vec{\mathbf{0}} = T(\vec{\boldsymbol{v}}).$$

So every element in the additive coset $\vec{\boldsymbol{v}} + K$ maps to the same value $T(\vec{\boldsymbol{v}})$ in W. Therefore we have,

$$im\, T = \{\, \vec{\boldsymbol{w}} \in W \mid \exists(\vec{\boldsymbol{v}} + K) \subseteq V \,.\, \{\vec{\boldsymbol{w}}\} = T(\vec{\boldsymbol{v}} + K)\,\}.$$

We could also express this using the inverse image as in 2.1.5 but the notation would be easily confused for the inverse transformation. □

Corollary 23. *Let $T : V \longmapsto W$ be any linear transformation with kernel $K = ker\, T$ and let there be $\vec{\boldsymbol{v}} \in V$, $\vec{\boldsymbol{w}} \in W$ such that*

$$T(\vec{\boldsymbol{v}}) = \vec{\boldsymbol{w}}.$$

Then,

$$T(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{w}} \iff \vec{\boldsymbol{x}} \in (\vec{\boldsymbol{v}} + K).$$

Proposition 90. *Linear Independence is preserved across a linear transformation iff the transformation is injective.*

Proof. Let $T : V \longmapsto W$ be any linear transformation with kernel $K = ker\, T$. If the kernel is nontrivial then there exists a nonempty basis of the kernel $B_K = \{\vec{\boldsymbol{k_1}}, \ldots, \vec{\boldsymbol{k_n}}\}$. Being a basis B_K is linearly independent so that,

$$\alpha_1\vec{\boldsymbol{k_1}} + \cdots + \alpha_n\vec{\boldsymbol{k_n}} = \vec{\boldsymbol{0}} \iff \alpha_1, \ldots, \alpha_n = 0.$$

However, $T(B_K)$ the image of B_K under T is

$$\{\vec{\boldsymbol{0}}, \ldots, \vec{\boldsymbol{0}}\}$$

which, by Proposition 66, is obviously not linearly independent.

We can also see it more directly from the definition of a linear relation.

$$\begin{aligned} & T(\alpha_1\vec{\boldsymbol{k_1}} + \cdots + \alpha_n\vec{\boldsymbol{k_n}}) = \vec{\boldsymbol{0}} \\ \iff \quad & \alpha_1 T(\vec{\boldsymbol{k_1}}) + \cdots + \alpha_n T(\vec{\boldsymbol{k_n}}) = \vec{\boldsymbol{0}} \\ \iff \quad & \alpha_1\vec{\boldsymbol{0}} + \cdots + \alpha_n\vec{\boldsymbol{0}} = \vec{\boldsymbol{0}} \end{aligned}$$

So clearly,

$$\alpha_1 T(\vec{\boldsymbol{k_1}}) + \cdots + \alpha_n T(\vec{\boldsymbol{k_n}}) = \vec{\boldsymbol{0}} \not\Longrightarrow \alpha_1, \ldots, \alpha_n = 0$$

which proves that if T is not injective then it does not preserve linear independence.

Conversely, if T does not preserve linear independence then, if $U = \{\vec{\boldsymbol{u}}_1, \ldots, \vec{\boldsymbol{u}}_n\} \subset V$ is a linearly independent set in V, there is a nontrivial linear relation between the vectors in $T(U)$ the image of U under T. That's to say,

$$\alpha_1 T(\vec{\boldsymbol{u}}_1) + \cdots + \alpha_n T(\vec{\boldsymbol{u}}_n) = \vec{\boldsymbol{0}} \quad \text{where } \prod_{i=1}^{n} \alpha_i \neq 0.$$

But this means that,

$$\begin{aligned} & \alpha_1 T(\vec{\boldsymbol{u}}_1) + \cdots + \alpha_n T(\vec{\boldsymbol{u}}_n) = \vec{\boldsymbol{0}} \\ \iff \quad & T(\alpha_1 \vec{\boldsymbol{u}}_1 + \cdots + \alpha_n \vec{\boldsymbol{u}}_n) = \vec{\boldsymbol{0}} \\ \iff \quad & \alpha_1 \vec{\boldsymbol{u}}_1 + \cdots + \alpha_n \vec{\boldsymbol{u}}_n \in ker\, T. \end{aligned}$$

Since U is linearly independent there is no nontrivial linear relation between its elements and since $\prod_{i=1}^{n} \alpha_i \neq 0$ we can conclude that,

$$\alpha_1 \vec{\boldsymbol{u}}_1 + \cdots + \alpha_n \vec{\boldsymbol{u}}_n \neq \vec{\boldsymbol{0}} \implies ker\, T \neq \{\vec{\boldsymbol{0}}\}$$

and therefore this proves that if T does not preserve linear independence then it is not injective. □

*Note that **linear dependence**, however, is preserved across any linear map (Proposition 87).*

2.5.1.2 Examples of Linear Transformations

(41) As previously seen in section 2.3.3, matrix multiplication on the left is a linear transformation. Let A be an $m \times n$ matrix with entries in $\mathbb{F}$ and consider A as an operator on column vectors $A : \mathbb{F}^n \longmapsto \mathbb{F}^m$. The kernel of A is the set of vectors that are solutions to $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{0}}$ while the image (or range) is the set of vectors $\vec{\boldsymbol{b}}$ such that $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$ has a solution.
The solutions $\{\, \vec{\boldsymbol{x}} \in \mathbb{F}^n \mid A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}} \,\}$ for some fixed $\vec{\boldsymbol{b}} \in \mathbb{F}^m$ are the additive coset $\vec{\boldsymbol{v}} + K$ where K is the kernel of A and $\vec{\boldsymbol{v}} \in \mathbb{F}^n$ is such that $A\vec{\boldsymbol{v}} = \vec{\boldsymbol{b}}$. Compare with 18.

(42) Also previously seen in 2.4.3 is that polynomials can be modeled as vectors. Let P_n be the vector space of real polynomials of degree $\leq n$. Then the derivative is a linear transformation $P_n \longmapsto P_{n-1}$. The kernel of the derivative is the set of degree 0 polynomials (i.e. constant functions) and the additive cosets of the kernel are $f(x) + c$, for $f(x) \in P_n$ and constant c.

2.5.1.3 The Dimension of a Linear Transformation

Definition. *The dimension of the image is called the* ***rank*** *while the dimension of the kernel is known as the* ***nullity****.*

Dimension and rank also exist for Groups (see wikipedia) where it refers to the minimal generating set for the group.

Theorem 25. ***The Dimension Formula:*** *Let $T : V \longmapsto W$ be a linear transformation, and assume that V is finite dimensional. Then,*

$$dim\, V = dim(ker\, T) + dim(im\, T) = rank + nullity.$$

Proof. Let $\{\vec{\boldsymbol{k_1}}, \ldots, \vec{\boldsymbol{k_m}}\}$ be a basis of $ker\, T$. Then, by Proposition 74, it may be extended to a basis of V,

$$B = \{\vec{\boldsymbol{k_1}}, \ldots, \vec{\boldsymbol{k_m}}, \vec{\boldsymbol{u_1}}, \ldots, \vec{\boldsymbol{u_n}}\}.$$

So, for any $\vec{\boldsymbol{v}} \in V$, $\vec{\boldsymbol{v}}$ may be expressed as a linear combination of the vectors in B. Therefore, for any $\vec{\boldsymbol{w}} \in im\, T$,

$$\begin{aligned}
& \vec{\boldsymbol{w}} = T(\alpha_1\vec{\boldsymbol{k_1}} + \cdots + \alpha_m\vec{\boldsymbol{k_m}} + \beta_1\vec{\boldsymbol{u_1}} + \cdots + \beta_n\vec{\boldsymbol{u_n}}) \\
\iff \quad & = T(\alpha_1\vec{\boldsymbol{k_1}}) + \cdots + T(\alpha_m\vec{\boldsymbol{k_m}}) + T(\beta_1\vec{\boldsymbol{u_1}}) + \cdots + T(\beta_n\vec{\boldsymbol{u_n}}) \\
\iff \quad & = \vec{\boldsymbol{0}} + T(\beta_1\vec{\boldsymbol{u_1}}) + \cdots + T(\beta_n\vec{\boldsymbol{u_n}}) \\
\iff \quad & = \beta_1 T(\vec{\boldsymbol{u_1}}) + \cdots + \beta_n T(\vec{\boldsymbol{u_n}})
\end{aligned}$$

This shows that $B' = \{T(\vec{\boldsymbol{u_1}}), \ldots, T(\vec{\boldsymbol{u_n}})\}$ spans $im\, T$. Furthermore, if there were a linear relation between the elements of B' then,

$$\begin{aligned}
& \beta_1 T(\vec{\boldsymbol{u_1}}) + \cdots + \beta_n T(\vec{\boldsymbol{u_n}}) = \vec{\mathbf{0}} \\
\iff \quad & T(\beta_1 \vec{\boldsymbol{u_1}} + \cdots + \beta_n \vec{\boldsymbol{u_n}}) = \vec{\mathbf{0}} \\
\iff \quad & \beta_1 \vec{\boldsymbol{u_1}} + \cdots + \beta_n \vec{\boldsymbol{u_n}} \in ker\, T \\
\iff \quad & \beta_1 \vec{\boldsymbol{u_1}} + \cdots + \beta_n \vec{\boldsymbol{u_n}} = \alpha_1 \vec{\boldsymbol{k_1}} + \cdots + \alpha_m \vec{\boldsymbol{k_m}}
\end{aligned}$$

where this last result implies a linear relation between the vectors of B. Since B is a basis this linear relation can only be the trivial relation and so $\beta_1, \ldots, \beta_n = 0$ and B' is linearly independent also. □

Notes about this:

- *This formula bears a resemblance to Lagrange's Theorem applied to homomorphisms of finite groups (10),*

$$|G| = |ker\, \phi| \cdot |im\, \phi|\,.$$

  The difference, however, is that the Dimension Formula of Linear Transformations is dealing with the generators of a group while Lagrange's Theorem is dealing with the orders of the groups. The orders of the groups are the number of elements in the group that are generated by the generators of the group. In the case of a real vector space, the vectors generated by the basis vectors are uncountably infinite due to scalar multiplication by real numbers and so cardinality doesn't apply in the same way.

- *This formula* ***only applies to finite-dimensional vector spaces****. This should be clear as we simply cannot do this kind of arithmetic with ∞. For example, if the rank is infinite then the dimension of the kernel would be $\infty - \infty =?$.*

2.5.2 Linear Transformations as Matrices

Proposition 91. *Left multiplication by a $m \times n$ matrix is a linear transformation $\mathbb{F}^n \longmapsto \mathbb{F}^m$.*

Proof. Let $T : \mathbb{F}^n \longmapsto \mathbb{F}^m$ be a linear transformation. Then T is a map from n-vectors to m-vectors that is compatible with the vector space operations. Let A be an $m \times n$ matrix then $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$ where $\vec{\boldsymbol{x}} \in \mathbb{F}^n$ and $\vec{\boldsymbol{b}} \in \mathbb{F}^m$ showing that $T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$ is a map of the form $\mathbb{F}^n \longmapsto \mathbb{F}^m$. Furthermore,

$$T(\vec{\boldsymbol{x_1}} + \vec{\boldsymbol{x_2}}) = A(\vec{\boldsymbol{x_1}} + \vec{\boldsymbol{x_2}}) = A\vec{\boldsymbol{x_1}} + A\vec{\boldsymbol{x_2}} = T(\vec{\boldsymbol{x_1}}) + T(\vec{\boldsymbol{x_2}})$$

and

$$T(c\vec{\boldsymbol{x}}) = A(c\vec{\boldsymbol{x}}) = cA\vec{\boldsymbol{x}} = cT(\vec{\boldsymbol{x}})$$

which shows that left multiplication preserves vector addition and scalar multiplication so that $T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$ is a linear map as required. □

Theorem 26. *Every linear transformation $\mathbb{F}^n \longmapsto \mathbb{F}^m$ is left multiplication by a particular $m \times n$ matrix.*

Proof. For any $\vec{\boldsymbol{x}} = \langle x_1, \ldots, x_n \rangle \in \mathbb{F}^n$ we can write it as,

$$x_1\vec{\boldsymbol{e_1}} + \cdots + x_n\vec{\boldsymbol{e_n}}.$$

Therefore if $T : \mathbb{F}^n \longmapsto \mathbb{F}^m$ then,

$$T(\vec{\boldsymbol{x}}) = T(x_1\vec{\boldsymbol{e_1}} + \cdots + x_n\vec{\boldsymbol{e_n}}) = T(\vec{\boldsymbol{e_1}})x_1 + \cdots + T(\vec{\boldsymbol{e_n}})x_n \in \mathbb{F}^m$$

and so letting $A \in \mathbb{F}^{m \times n}$ be,

$$A = \begin{bmatrix} T(\vec{\boldsymbol{e_1}}) & \cdots & T(\vec{\boldsymbol{e_n}}) \end{bmatrix}$$

we have,

$$T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}.$$ □

Corollary 24. *Any linear transformation between spaces isomorphic to $\mathbb{F}^n$ and $\mathbb{F}^m$ (refer to Proposition 83 and 37) is left multiplication by a particular $m \times n$ matrix.*

This is why linear transformations from a space to itself can be wholly characterized by what they do to the axes and also why every such linear transformation can be considered a change of basis and vice-versa.
Conceptually, a linear transformation changes the coordinates of a selection of transformed vectors. The confusion comes about because we implement the matrix of the linear transformation A by transforming the basis against which the coordinates are applied,

$$A = \begin{bmatrix} T(\vec{\boldsymbol{e_1}}) & \cdots & T(\vec{\boldsymbol{e_n}}) \end{bmatrix}.$$

Whereas a change of basis transforms the basis against which coordinates are applied ***and then updates the coordinates to balance out the change****. This can be seen if we deconstruct the change of basis formula:*

$$B\vec{\boldsymbol{x}}_B = B'\vec{\boldsymbol{x}}_{B'} \iff \vec{\boldsymbol{x}}_{B'} = (B')^{-1}B\vec{\boldsymbol{x}}_B$$

$$\vec{\boldsymbol{x}}_{B'} = P\vec{\boldsymbol{x}}_B = (B')^{-1}B\vec{\boldsymbol{x}}_B$$

This can be thought of as first obtaining the coordinates against the standard basis $B\vec{\boldsymbol{x}}_B$ and then applying the inverse of the target basis so as to obtain the equivalent coordinates against the target basis. But, note, we could also consider the whole thing as a linear transformation represented by P.

The biggest difference, however, is that a linear transformation can also be between different spaces — say from n-dimensional space to m-dimensional space — in which case it cannot be thought of as a change of basis as a vector $\vec{\boldsymbol{v}} \in \mathbb{F}^n$ cannot be equivalently expressed using a basis of $\mathbb{F}^m$ because the two spaces are not isomorphic.

2.5.2.1 Examples of Linear Transformations as Matrices

(43) Let $T : \mathbb{R}^2 \longmapsto \mathbb{R}^2$ be a linear transformation such that,

$$T(\vec{\boldsymbol{e_1}}) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad T(\vec{\boldsymbol{e_2}}) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

The transformation T has been completely described in this way because, for any $\vec{\boldsymbol{x}} = \langle x_1, x_2 \rangle \in \mathbb{R}^2$ we have,

$$T(\vec{\boldsymbol{x}}) = T(x_1\vec{\boldsymbol{e_1}} + x_2\vec{\boldsymbol{e_2}})$$

$$\iff \qquad T(\vec{\boldsymbol{x}}) = x_1 T(\vec{\boldsymbol{e_1}}) + x_2 T(\vec{\boldsymbol{e_2}})$$
$$\iff \qquad T(\vec{\boldsymbol{x}}) = x_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} x_1 - x_2 \\ 2x_1 \end{bmatrix}.$$

So, T is also left multiplication by the matrix,

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}.$$

(44) Consider a linear map $T : \mathbb{R}^n \longmapsto \mathbb{R}^m$ and the matrix representing it $A \in \mathbb{R}^{mn}$. Suppose the equation $A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}}$ has the known solution

$$\vec{\boldsymbol{x}} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \\ 0 \end{bmatrix} + s \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 1 \\ 0 \\ -1 \\ 1 \end{bmatrix} \qquad s, t \in \mathbb{R}.$$

What can be said about the linear transformation T from looking at the solution $\vec{\boldsymbol{x}}$?

- The dimension of the domain, $n = 5$, is clear since the solution $\vec{\boldsymbol{x}}$ is a vector in the domain space.
- The nullity - dimension of the kernel - is 2, since there are two free variables s, t. The basis of the 2-dimensional nullspace is the two vectors being multiplied by these variables.
- Using the dimension formula (Theorem 25) we can deduce that the rank of T is n - nullity. So the rank is $5 - 2 = 3$. This is also supported by the fact that the particular solution has 3 non-zero components.

What can **not** be said?

- The dimension of the codomain m cannot be derived from looking at the solution $\vec{\boldsymbol{x}}$. The dimension of the image of T is given by its range because the kernel maps to the origin in the codomain and so the image of the kernel has dimension zero. Since we are not told whether or not the linear map is surjective we cannot know the dimension of the codomain space - only the image of T in the codomain space.

(45) The minimal linear transformation is multiplication by a minimal matrix — a one-by-one matrix, i.e. a scalar. So

$$T(\vec{\boldsymbol{v}}) = A\vec{\boldsymbol{v}} = a\vec{\boldsymbol{v}}.$$

Since the real number line, for example, may be considered a one-dimensional vector space, multiplication of two real numbers, for example, may be considered as a linear transformation.

2.5.2.2 Linear Transformations and Change of Basis

It is often possible to achieve powerful simplifications of problems by selecting appropriate bases. In this section we will look at linear transformations represented by matrices between arbitrary bases of spaces. So here we are looking at the relationship between linear transformations and change of basis.

Definition. *If the matrix A of a linear transformation $T : V \longmapsto W$ is defined as, for $\vec{\boldsymbol{x}} \in V$,*

$$T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}} \in W$$

then ***the matrix of*** *T* ***with respect to the bases*** *$B \subset V$* ***and*** *$B' \subset W$ is defined as the matrix A that satisfies,*

$$A\vec{\boldsymbol{x}}_B = \vec{\boldsymbol{b}}_{B'} \in W$$

and also

$$T(\vec{\boldsymbol{x}}) = [B']A\vec{\boldsymbol{x}}_B = \vec{\boldsymbol{b}}$$

2.5.2.3 Intuition of the Matrix of T with respect to Bases

Let $T : V \longmapsto W$ be a linear transformation and let $B_V = \{\vec{\boldsymbol{v_1}}, \ldots, \vec{\boldsymbol{v_n}}\}$ be a basis of V and $B_W = \{\vec{\boldsymbol{w_1}}, \ldots, \vec{\boldsymbol{w_m}}\}$ be a basis of W. Then $T(\vec{\boldsymbol{v_i}}) \in W$ and so there is some $m \times n$ matrix $A = (a_{ij})$ such that,

$$\begin{bmatrix} \vec{\boldsymbol{w_1}} & \cdots & \vec{\boldsymbol{w_m}} \end{bmatrix} A = \begin{bmatrix} T(\vec{\boldsymbol{v_1}}) & \cdots & T(\vec{\boldsymbol{v_n}}) \end{bmatrix}.$$

where,

$$T(\vec{\boldsymbol{v_j}}) = \begin{bmatrix} \vec{\boldsymbol{w_1}} & \cdots & \vec{\boldsymbol{w_m}} \end{bmatrix} \begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} = a_{1j}\vec{\boldsymbol{w_1}} + \cdots + a_{mj}\vec{\boldsymbol{w_m}} = \sum_i a_{ij}\vec{\boldsymbol{w_i}}$$

so that the jth column of the matrix A is the coordinate vector of $T(\vec{\boldsymbol{v_j}})$ with respect to the basis B_W.

Substituting $B_W = \{\vec{\boldsymbol{w_1}}, \ldots, \vec{\boldsymbol{w_m}}\}$ we can obtain an expression for A,

$$\begin{aligned} & \begin{bmatrix} \vec{\boldsymbol{w_1}} & \cdots & \vec{\boldsymbol{w_m}} \end{bmatrix} A = \begin{bmatrix} T(\vec{\boldsymbol{v_1}}) & \cdots & T(\vec{\boldsymbol{v_n}}) \end{bmatrix} \\ \iff \quad & [B_W]A = \begin{bmatrix} T(\vec{\boldsymbol{v_1}}) & \cdots & T(\vec{\boldsymbol{v_n}}) \end{bmatrix} \\ \iff \quad & A = [B_W]^{-1} \begin{bmatrix} T(\vec{\boldsymbol{v_1}}) & \cdots & T(\vec{\boldsymbol{v_n}}) \end{bmatrix}. \end{aligned}$$

The matrix A is referred to as the **matrix of T with respect to the bases B_V and B_W** and conforms to,

$$A\vec{\boldsymbol{x}}_{B_V} = \vec{\boldsymbol{b}}_{B_W}.$$

This can be seen as,

$$\begin{aligned} & A\vec{\boldsymbol{x}}_{B_V} = [B_W]^{-1} \begin{bmatrix} T(\vec{\boldsymbol{v_1}}) & \cdots & T(\vec{\boldsymbol{v_n}}) \end{bmatrix} \vec{\boldsymbol{x}}_{B_V} \\ \iff \quad & A\vec{\boldsymbol{x}}_{B_V} = [B_W]^{-1}\vec{\boldsymbol{b}}_{B_V} \\ \iff \quad & A\vec{\boldsymbol{x}}_{B_V} = \vec{\boldsymbol{b}}_{B_W} \end{aligned}$$

so that $\vec{\boldsymbol{x}}_{B_V}$ — the coordinate vector with respect to B_V — is first transformed by applying the coordinates to the transformed version of the basis B_V and then these coordinates are converted to B_W coordinates by left multiplication by $[B_W]^{-1}$.

If we chose different bases for the spaces we would get a different matrix. If the bases are the standard bases then the matrix is the standard matrix for the transformation.

2.5.2.4 Examples of Linear Transform Matrices w.r.t. Bases

(46) Let $T : \mathbb{R}^2 \longmapsto \mathbb{R}^2$ be a linear transform defined (against the standard basis) by,

$$T\left(\begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}2\\3\end{bmatrix} \quad \text{and} \quad T\left(\begin{bmatrix}0\\1\end{bmatrix}\right) = \begin{bmatrix}3\\2\end{bmatrix}.$$

Then, if we define the matrix of T with respect to the standard basis only then we have,

$$A = \begin{bmatrix} T\left(\begin{bmatrix}1\\0\end{bmatrix}\right) & T\left(\begin{bmatrix}0\\1\end{bmatrix}\right) \end{bmatrix} = \begin{bmatrix}2 & 3\\3 & 2\end{bmatrix}$$

and, if we define a vector $\vec{\boldsymbol{x}} = \langle 1, 1\rangle$ against the standard basis we can see that,

$$T(\vec{\boldsymbol{x}}) = 1 \cdot T\left(\begin{bmatrix}1\\0\end{bmatrix}\right) + 1 \cdot T\left(\begin{bmatrix}0\\1\end{bmatrix}\right) = A\vec{\boldsymbol{x}} = \begin{bmatrix}2 & 3\\3 & 2\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}5\\5\end{bmatrix}.$$

If we now define B, an alternative basis of $\mathbb{R}^2$, as

$$B = \left\{\begin{bmatrix}3\\0\end{bmatrix}, \begin{bmatrix}0\\2\end{bmatrix}\right\}$$

then we have,

$$[B] = \begin{bmatrix}3 & 0\\0 & 2\end{bmatrix} \quad \text{and} \quad [B]^{-1} = \frac{1}{6}\begin{bmatrix}2 & 0\\0 & 3\end{bmatrix}$$

and the linear transform matrix **of coordinate vectors w.r.t. the basis** B is defined as,

$$A = \begin{bmatrix} T\left(\begin{bmatrix}3\\0\end{bmatrix}\right) & T\left(\begin{bmatrix}0\\2\end{bmatrix}\right) \end{bmatrix} = \begin{bmatrix}6 & 6\\9 & 4\end{bmatrix}.$$

Now *this* matrix A expects coordinate vectors w.r.t. B and so if we convert $\vec{\boldsymbol{x}}$ to basis B as follows,

$$\vec{\boldsymbol{x}}_B = [B]^{-1}\vec{\boldsymbol{x}} = \frac{1}{6}\begin{bmatrix}2 & 0\\ 0 & 3\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1/3\\1/2\end{bmatrix}$$

then we find that,

$$T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}_B = \begin{bmatrix}6 & 6\\ 9 & 4\end{bmatrix}\begin{bmatrix}1/3\\1/2\end{bmatrix} = \begin{bmatrix}6/3+6/2\\9/3+4/2\end{bmatrix} = \begin{bmatrix}5\\5\end{bmatrix}.$$

If we were to define another basis of $\mathbb{R}^2$ called B' and construct the matrix of T with respect to B and B' then the matrix A would become,

$$A = [B']^{-1}\begin{bmatrix}T\left(\begin{bmatrix}3\\0\end{bmatrix}\right) & T\left(\begin{bmatrix}0\\2\end{bmatrix}\right)\end{bmatrix}$$

and to get the result in standard coordinates we would need to apply the result to the basis vectors of B',

$$T(\vec{\boldsymbol{x}}) = [B']A\vec{\boldsymbol{x}}_B.$$

In the case where we want the result to be in the same basis as the argument vector $\vec{\boldsymbol{x}}_B$ we still need to modify the matrix A. In this case A becomes,

$$A = [B]^{-1}\begin{bmatrix}T\left(\begin{bmatrix}3\\0\end{bmatrix}\right) & T\left(\begin{bmatrix}0\\2\end{bmatrix}\right)\end{bmatrix}$$

This A still expects $\vec{\boldsymbol{x}}$ to be in B coordinates but outputs a result defined in B coordinates rather than standard coordinates.

$$A\vec{\boldsymbol{x}}_B = \frac{1}{6}\begin{bmatrix}2 & 0\\ 0 & 3\end{bmatrix}\begin{bmatrix}6 & 6\\ 9 & 4\end{bmatrix}\vec{\boldsymbol{x}}_B = \begin{bmatrix}2 & 2\\ 9/2 & 2\end{bmatrix}\begin{bmatrix}1/3\\1/2\end{bmatrix} = \begin{bmatrix}5/3\\5/2\end{bmatrix}.$$

So, to get the result in standard coordinates we need to apply the result to the basis vectors of B',

$$T(\vec{\boldsymbol{x}}) = [B](A\vec{\boldsymbol{x}}_B) = \begin{bmatrix}3 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}5/3\\5/2\end{bmatrix} = \begin{bmatrix}5\\5\end{bmatrix}.$$

Proposition 92. *Let A be the matrix of a linear transformation $T : V \longmapsto W$ with respect to the bases B_V, B_W of dimension n and m respectively. The matrices A' which represent T with respect to other bases are those of the form,*

$$A' = QAP^{-1}$$

where $Q \in GL_m(\mathbb{F}), P \in GL_n(\mathbb{F})$.

Proof. Let $B'_V = \{\vec{\boldsymbol{v}}'_1, \ldots, \vec{\boldsymbol{v}}'_n\}, B'_W = \{\vec{\boldsymbol{w}}'_1, \ldots, \vec{\boldsymbol{w}}'_n\}$ be alternative bases with respect to which we want to find A', the matrix of T. Also let,

$$B_V = B'_V P \quad \text{and} \quad B_W = B'_W Q.$$

Then for $\vec{\boldsymbol{v}}'_i \in B'_V$, $T(\vec{\boldsymbol{v}}') \in \mathit{span}\, B'_W$ so there exists a matrix A' such that, using the notation $T(B_V)$ to indicate the image of the set B_V under T and $[B_V]$ for the matrix whose columns are the elements of B_V,

$$\begin{aligned}
& \begin{bmatrix} \vec{\boldsymbol{w}}'_1 & \cdots & \vec{\boldsymbol{w}}'_m \end{bmatrix} A' = \begin{bmatrix} T(\vec{\boldsymbol{v}}'_1) & \cdots & T(\vec{\boldsymbol{v}}'_n) \end{bmatrix} \\
\iff \quad & [B'_W]A' = [T(B'_V)] \\
\iff \quad & A' = [B'_W]^{-1}[T(B'_V)] \\
\iff \quad & A' = [B'_W]^{-1}[T(B_V P^{-1})] \\
\iff \quad & A' = Q[B_W]^{-1}[T(B_V)]P^{-1} \qquad \text{P is coefficient matrix} \\
\iff \quad & A' = QAP^{-1}. \qquad \square
\end{aligned}$$

2.5.2.5 Simplification of the Matrix of a Transformation

Proposition 93. *Let $T : V \longmapsto W$ be a linear transformation of rank r. Bases B_V, B_W may be chosen so that the matrix of T takes the form,*

$$A = \begin{bmatrix} I_r & \vdots \\ \cdots & 0 \end{bmatrix}.$$

Proof. Let $U = \{\vec{\boldsymbol{u}}_1, \ldots, \vec{\boldsymbol{u}}_k\}$ be a basis of the kernel of T where $k = dim(ker\, T)$. Then U may be extended to a basis of V (Proposition 74),

$$B_V = \{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_r, \vec{\boldsymbol{u}}_1, \ldots, \vec{\boldsymbol{u}}_k\}.$$

Then, let $T(\vec{\boldsymbol{v}}_i) = \vec{\boldsymbol{w}}_i$ so that,

$$[T(B_V)] = \begin{bmatrix} \vec{\boldsymbol{w}}_1 & \cdots & \vec{\boldsymbol{w}}_r & \vec{\mathbf{0}} & \cdots & \vec{\mathbf{0}} \end{bmatrix}.$$

As shown in Theorem 25, $\{\vec{\boldsymbol{w}}_1, \ldots, \vec{\boldsymbol{w}}_r\}$ is a basis of the image of T and can also be extended to a basis of W,

$$B_W = \{\vec{\boldsymbol{w}}_1, \ldots, \vec{\boldsymbol{w}}_r, \vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_{m-r}\}.$$

So, the matrix A of T with respect to the bases B_V, B_W satisfies,

$$\begin{aligned} & & A &= [B_W]^{-1}[T(B_V)] \\ \iff & & [B_W]A &= [T(B_V)] \\ \iff & \begin{bmatrix} \vec{\boldsymbol{w}}_1 & \ldots & \vec{\boldsymbol{w}}_r & \vec{\boldsymbol{x}}_1 & \ldots & \vec{\boldsymbol{x}}_{m-r} \end{bmatrix} A &= \begin{bmatrix} \vec{\boldsymbol{w}}_1 & \ldots & \vec{\boldsymbol{w}}_r & \vec{\mathbf{0}} & \ldots & \vec{\mathbf{0}} \end{bmatrix}. \end{aligned}$$

If we look at the components of the matrices we see,

$$\begin{bmatrix} w_{11}\ldots & w_{1r} & x_{11}\ldots & x_{1(m-r)} \\ \vdots & & & \\ w_{r1}\ldots & w_{rr} & x_{r1}\ldots & x_{r(m-r)} \\ \vdots & & & \\ w_{m1}\ldots & w_{mr} & x_{m1}\ldots & x_{m(m-r)} \end{bmatrix} \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \\ a_{r1} & \ldots & a_{rn} \\ \vdots & & \\ a_{m1} & \ldots & a_{mn} \end{bmatrix} = \begin{bmatrix} w_{11}\ldots & w_{1r} & 0\ldots & 0 \\ \vdots & & & \\ w_{r1}\ldots & w_{rr} & 0\ldots & 0 \\ \vdots & & & \\ w_{m1}\ldots & w_{mr} & 0\ldots & 0 \end{bmatrix}$$

which makes it clear that A has the form,

$$A = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ \vdots & & & & & \\ 0 & \ldots & 0 & 0 & \ldots & 0 \end{bmatrix}$$

as required. □

Corollary 25. *By Proposition 91, left multiplication by any matrix is a linear transformation and so is equivalent to left multiplication by a matrix of the form*

$$\begin{bmatrix} I_r & \vdots \\ \cdots & 0 \end{bmatrix}$$

but with reference to different coordinate systems.

2.5.2.6 Example of Simplification by Selecting Bases

(47) Continuing the example 33 of Gaussian Elimination for row reducing a matrix we have a matrix

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

whose rref form is

$$A' = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}.$$

The rref tells us that the first two columns of A are a basis of the image and the kernel is

$$c \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \qquad c \in \mathbb{R}.$$

We can use this information to form a matrix A_{PQ} — which is the matrix A expressed with respect to the bases P and Q — such that A_{PQ} is maximally simplified. Following the procedure used in the proof of Proposition 93, we begin by finding a basis of the domain space by extending a basis of the kernel.

Since the kernel is one-dimensional and the domain is three-dimensional, we can extend the basis of the kernel given above to a basis of the domain by adding two of the three standard basis vectors. So, the chosen basis of the domain is

$$P = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \right\}.$$

Again following the procedure from the same proof, we next form a basis of the codomain space that extends a basis of the image which,

in this case, is the first two columns of the matrix A. So, we can choose the basis,

$$Q = \left\{ \begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

We could have chosen anything for the final column just so long as it is linearly independent of the first two columns.

Then,

$$\begin{aligned} A_{PQ} &= Q^{-1}AP \\ &= \begin{bmatrix} 3 & 1 & 0 \\ 0 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 & 1 \\ 0 & 2 & -4 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 2 & -1 & 0 \\ 0 & 3 & 0 \\ -6 & -3 & 6 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ 0 & 2 & -4 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 2 & -1 & 0 \\ 0 & 3 & 0 \\ -6 & -3 & 6 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 0 & 2 & 0 \\ 3 & 2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

2.5.3 Linear Operators and Eigenvectors

Definition. *A **linear operator** is a linear transformation from a space to itself. That's to say, the domain and codomain of the transformation are the same space and considered with respect to the same basis.*

Definition. *A linear operator is called **singular** if it does not have an inverse and **nonsingular** if it has an inverse.*

Proposition 94. *A linear operator $T : V \longmapsto V$ is bijective iff it has a trivial kernel.*

Proof. By Theorem 14 of homomorphisms, if T has a trivial kernel then it is injective which implies that $|im\, T| = |V|$ so that it is also surjective (because the domain is equal to the codomain). Therefore, it is bijective.
Conversely, if it is bijective then it is injective and so it has a trivial kernel. □

Corollary 26. *A linear operator $T : V \longmapsto V$ is isomorphic iff it has a trivial kernel.*

Corollary 27. *A linear operator is nonsingular iff it has a trivial kernel.*

Note that these are not true of linear transformations in general because, for a transformation between different vector spaces of different dimensions, injectivity does not imply bijectivity and invertibility.

Proposition 95. *The following conditions on a linear operator $T : V \longmapsto V$ on a finite-dimensional vector space are equivalent:*

(i) $ker\, T > 0$.

(ii) $im\, T < V$.

(iii) *If A is the matrix of the operator with respect to an arbitrary basis, then $\det A = 0$.*

(iv) *T is singular*

Proof. The first two of these properties follow from the dimension formula for finite-dimensional vector spaces. The third follows since an operator with a non-trivial kernel is noninvertible (Proposition 94) and so its matrix will also be noninvertible; noninvertible matrices have determinant 0. The last property follows from the definition of singular and the fact that T is noninvertible. □

Again it's worth noting the contrast with transformations in general: for a transformation whose domain vector space is lower dimension than its codomain, the transformation may be injective but not surjective (kernel is trivial but $\operatorname{im} T < V$), and conversely, if the domain is higher dimension than the codomain then the transformation may be surjective while not injective ($\ker T > 0$ but image covers codomain). Furthermore, the third condition relating to the determinant, only applies to operators because the determinant is a property of square matrices.

Note that the first two properties do not hold for infinite-dimensional vector spaces. For example, let $V = \mathbb{R}^\infty$ be the space of sequences $a_1, a_2, \ldots$. Then the "shift operator" is defined as,

$$T(a_1, a_2, \dots) = (0, a_1, a_2, \dots).$$

This is a linear operator because,

$$\begin{aligned} \alpha T(a_1, a_2, \dots) + \beta T(b_1, b_2, \dots) &= \alpha(0, a_1, a_2, \dots) + \beta(0, b_1, b_2, \dots) \\ &= (0, \alpha a_1, \alpha a_2, \dots) + (0, \beta b_1, \beta b_2, \dots) \\ &= (0, \alpha a_1 + \beta b_1, \alpha a_2 + \beta b_2, \dots) \\ &= T(\alpha a_1 + \beta b_1, \alpha a_2 + \beta b_2, \dots). \end{aligned}$$

However, while clearly for this operator we have $\operatorname{im} T < V$, nevertheless the kernel of this operator is the trivial kernel $\{\vec{\mathbf{0}}\}$. This just shows further that the dimension formula (Theorem 25) for finite-dimensional vector spaces does

not apply for infinite-dimensional spaces.
The explanation of this is that infinite sets can have proper subsets that have equal cardinality. So we can have an image whose dimensions are a proper subset of the dimensions of the space but the image, nevertheless, has the same dimensionality as the space. This can only happen with infinite-dimensional spaces.

Proposition 96. *Let A be the matrix of a linear operator T with respect to the basis B of dimension n. The matrices A' which represent T with respect to other bases are those of the form,*

$$A' = PAP^{-1}$$

where $P \in GL_n(\mathbb{F})$.

Proof. A linear operator is a specialization of a linear transformation where the domain and codomain are the same set so we can use the definition (2.5.2.2) of the matrix of a linear transformation w.r.t. to two different bases and simply set the two bases to the same set B. This produces,

$$T(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}_B = \vec{\boldsymbol{b}}_B$$

so that,

$$[B]A = [T(\vec{\boldsymbol{b}}_1) \cdots T(\vec{\boldsymbol{b}}_n)] \iff A = [B]^{-1}[T(\vec{\boldsymbol{b}}_1) \cdots T(\vec{\boldsymbol{b}}_n)].$$

Now let there be another basis B' related to B by,

$$[B] = [B']P \iff [B]P^{-1} = [B'],$$

$$P\vec{\boldsymbol{x}}_B = \vec{\boldsymbol{x}}_{B'}.$$

To define a matrix A' that performs the same linear transformation as A w.r.t. to the basis B' we need,

$$T(\vec{\boldsymbol{x}}) = A'\vec{\boldsymbol{x}}_{B'} = \vec{\boldsymbol{b}}_{B'}$$

and

$$[B']A' = [T(\vec{\boldsymbol{b}}'_1) \cdots T(\vec{\boldsymbol{b}}'_n)] \iff A' = [B']^{-1}[T(\vec{\boldsymbol{b}}'_1) \cdots T(\vec{\boldsymbol{b}}'_n)].$$

If we have the vectors of B' encoded in B-coordinates then we can use the transformed version of the basis B to produce the transformed version of the basis B',

$$[T(\vec{\boldsymbol{b}}'_1)\cdots T(\vec{\boldsymbol{b}}'_n)] = [T(\vec{\boldsymbol{b}}_1)\cdots T(\vec{\boldsymbol{b}}_n)][B]^{-1}[B'] = [T(\vec{\boldsymbol{b}}_1)\cdots T(\vec{\boldsymbol{b}}_n)]P^{-1}.$$

If we now note that,

$$A = [B]^{-1}[T(\vec{\boldsymbol{b}}_1)\cdots T(\vec{\boldsymbol{b}}_n)] \quad \text{and} \quad A' = [B']^{-1}[T(\vec{\boldsymbol{b}}'_1)\cdots T(\vec{\boldsymbol{b}}'_n)]$$

then,

$$\begin{aligned} & A' = [B']^{-1}[T(\vec{\boldsymbol{b}}_1)\cdots T(\vec{\boldsymbol{b}}_n)]P^{-1} \\ \iff \quad & A' = [B']^{-1}[B]AP^{-1} \\ \iff \quad & A' = PAP^{-1}. \end{aligned}$$ □

2.5.3.1 Similar Matrices

Definition. *If two matrices A, A' are related by,*

$$A' = PAP^{-1}$$

*for some $P \in GL_n(\mathbb{F})$ then they are known as **similar** matrices or **conjugates**.*

Note:

- *Similar matrices represent the same linear transformation expressed with respect to different bases.*
- *In general, linear maps are not invertible so the matrix A is not necessarily a member of a group. So, this use of the term "conjugation" only equates to the term in group theory within the general linear group $GL_n(\mathbb{F})$.*

- *Similar matrices have the same determinant which means that they expand or contract the space by the same amount. This is to be expected as they represent the same linear transformation defined against different bases.*

2.5.3.2 Intuition of Similar Matrices

$$A' = PAP^{-1} \iff A'P = PA$$

P is the change of basis matrix such that $\vec{\boldsymbol{x}}_{B'} = P\vec{\boldsymbol{x}}_B$. So P is $B_{B'}$ the basis vectors of B w.r.t. to B'. Another way of looking at it: $P = [B']^{-1}[B]$ so it decodes coordinates in B-coords to standard coordinates and then encodes them into B'-coords. Meanwhile A transforms the basis vectors of B and then encodes the result in B-coords so the schematic of A is $A = [B]^{-1}[T(B)]$.
So

$$PA = P[B]^{-1}[T(B)] = [B']^{-1}[B][B]^{-1}[T(B)] = [B']^{-1}[T(B)].$$

By similar reasoning the schematic of A' is $A' = [B']^{-1}[T(B')]$. But also we have,

$$A' = PAP^{-1} = ([B']^{-1}[B])([B]^{-1}[T(B)])([B]^{-1}[B']) = [B']^{-1}[T(B)][B]^{-1}[B']$$

so that,

$$\begin{aligned} & [B']^{-1}[T(B')] = [B']^{-1}[T(B)][B]^{-1}[B'] \\ \iff \quad & [T(B')] = [T(B)][B]^{-1}[B']. \end{aligned}$$

What this last result is saying is that if we take the basis vectors of B' and encode them in B-coordinates and then apply the result to the transformed basis of B then the result is the transformed basis of B' defined in standard coordinates.

So, we have,

$$A'P = PA = [B']^{-1}[T(B)]$$
$$P^{-1}A' = AP^{-1} = [B]^{-1}[T(B)][B]^{-1}[B'] = [B]^{-1}[T(B')].$$

Proposition 97. *Similarity of matrices is an equivalence relation.*

Proof. For any $M, N \in GL_n(\mathbb{F})$, similarity is an equivalence relation because of the following properties.

Reflexivity:

$$\begin{aligned} & N = I^{-1}NI \\ \therefore\ & N \sim N \end{aligned}$$

Symmetry:

$$\begin{aligned} & N = P^{-1}MP \\ \iff\ & NP^{-1} = P^{-1}M(PP^{-1}) \\ \iff\ & NP^{-1} = P^{-1}M \\ \iff\ & PNP^{-1} = (PP^{-1})M \\ \iff\ & PNP^{-1} = M \\ \iff\ & R^{-1}NR = M, \ \ R \in X \\ \therefore\ & N \sim M \iff M \sim N \end{aligned}$$

Transitivity:

$$\begin{aligned} & N = P^{-1}MP, \ \ M = Q^{-1}AQ \\ \implies\ & N = P^{-1}(Q^{-1}AQ)P \\ \iff\ & N = (P^{-1}Q^{-1})A(QP) \\ \iff\ & N = R^{-1}AR, \ \ R \in X \\ \therefore\ & (N \sim M) \wedge (M \sim Q) \iff (N \sim Q) \end{aligned}$$

□

Proposition 98. *Similar matrices have the same determinant.*

Proof. Let $A' = PAP^{-1}$ where P is a change of basis matrix. Then,

$$\begin{aligned} \det A' &= \det PAP^{-1} \\ &= (detP) \cdot (\det A) \cdot (\det P^{-1}) \\ &= (\det P) \cdot (\det A) \cdot (1/\det P) \\ &= \det A. \end{aligned}$$

□

2.5.3.3 Invariant Subspaces

Definition. *Let $T : V \longmapsto V$ be a linear operator on a vector space. A subspace W of V is called an* ***invariant subspace*** *or a* ***T-invariant subspace*** *if it is carried to itself by the operator,*

$$TW \subset W.$$

In other words, W is T-invariant if $T(\vec{\boldsymbol{w}}) \in W$ for all $\vec{\boldsymbol{w}} \in W$. In this case, T may be referred to as defining an operator on W that is the ***restriction of T to W.***

Let $W \subseteq V$ be a T-invariant subspace of V with basis $B_W = \vec{\boldsymbol{w}}_1, \ldots, \vec{\boldsymbol{w}}_k$. Then, by Proposition 74, B_W can be extended to a basis of V,

$$B_V = \{\vec{\boldsymbol{w}}_1, \ldots, \vec{\boldsymbol{w}}_k, \vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_{n-k}\}.$$

If we look at the matrix A of T with respect to this basis B_V we find a characteristic pattern.

$$A = [B_V]^{-1}[T(B_V)] = [B_V]^{-1}[T(\vec{\boldsymbol{w}}_1) \cdots T(\vec{\boldsymbol{w}}_k)\ T(\vec{\boldsymbol{v}}_1) \cdots T(\vec{\boldsymbol{v}}_{n-k})]$$

with $T(\vec{\boldsymbol{w}}_1), \ldots, T(\vec{\boldsymbol{w}}_k) \in W$ so that,

$$T(\vec{\boldsymbol{w}}_i) = \alpha_1 \vec{\boldsymbol{w}}_1 + \cdots + \alpha_k \vec{\boldsymbol{w}}_k.$$

This means that when we express $T(\vec{\boldsymbol{w}}_i)$ with respect to the basis B_V the coordinate vectors will take the form,

$$(\alpha_1, \ldots, \alpha_k, 0, \ldots, 0)^T.$$

As a result, the matrix A will take the form,

$$A = \begin{bmatrix} C & D \\ 0 & E \end{bmatrix}$$

where C is a $k \times k$ matrix that represents the restriction of T to W.

A description of a matrix of the form of the description of A here is known as a ***block decomposition****.*

On the other hand, if $V = W_1 \oplus W_2$ where *both* W_1 *and* W_2 are T-invariant subspaces then A takes the form,

$$A = \begin{bmatrix} C & 0 \\ 0 & E \end{bmatrix}$$

where, as before, C is a $k \times k$ matrix that represents the restriction of T to W_1 but, this time, also E is a $(n-k) \times (n-k)$ matrix that represents the restriction of T to W_2.

Matrices with the form of the matrix,

$$\begin{bmatrix} C & 0 \\ 0 & E \end{bmatrix}$$

are known as ***block diagonal matrices*** *or* ***diagonal block matrices****.*

2.5.3.4 Eigenvectors

Definition. *An* ***eigenvector*** *of a linear operator T is a nonzero vector $\vec{\boldsymbol{v}}$ with the property under T that,*

$$T(\vec{\boldsymbol{v}}) = c\vec{\boldsymbol{v}}$$

for some constant $c \in \mathbb{F}$. The constant c is called an ***eigenvalue****.*

Eigenvectors may also be referred to as ***characteristic vectors*** *and eigenvalues as* ***characteristic values****.*

The ***eigenspace*** *of an eigenvalue is the subspace formed by the eigenvectors associated with the eigenvalue and the zero vector.*

Note:

- *To have eigenvectors, T must be an operator as the image of the eigenvector $\vec{\boldsymbol{v}} \in V$ under T,*

$$T(\vec{\boldsymbol{v}}) = c\vec{\boldsymbol{v}} \in V$$

is, by definition, in the same space as the eigenvector itself (if there were a change of basis then we would not consider it to be the same vector). In fact, this is the key of the relationship between eigenvectors and invariant subspaces: The space spanned by eigenvectors of T is T-invariant. The bases of T-invariant subspaces, however, need not be eigenvectors. For example, a rotation of a plane in a 3d space has the plane as a T-invariant subspace but the only eigenvector is the axis of rotation, perpendicular to the plane.

- *An eigenvector may not be $\vec{\mathbf{0}}$ but an eigenvalue may be 0. As a consequence of this, every nonzero vector in the kernel is an eigenvector (with eigenvalue 0). As a consequence of this, if W is a 2-dimensional T-invariant subspace, for example, and its image under T is one-dimensional (a line inside the plane of W) then vectors in W that are not in the line of the image will be in the kernel of T and so also eigenvectors but with the eigenvalue 0. It is also worth noting that they are still considered to be parallel to their image under T because, by convention, the zero vector $\vec{\mathbf{0}}$ is considered to be parallel to all vectors.*

- *If we speak of an eigenvector of a square matrix then we are referring to an eigenvector of **left multiplication** by the matrix, i.e. a nonzero vector $\vec{\boldsymbol{x}}$ such that,*

$$A\vec{\boldsymbol{x}} = c\vec{\boldsymbol{x}}.$$

Clearly if $\vec{\boldsymbol{x}}_B$ is the coordinate vector of $\vec{\boldsymbol{v}} \in V$ with respect to B — a basis of V — and A is the matrix of T with respect to the basis B, then

$$A\vec{\boldsymbol{x}}_B = c\vec{\boldsymbol{x}}_B \iff T(\vec{\boldsymbol{v}}) = c\vec{\boldsymbol{v}}.$$

As a result, all similar matrices have the same eigenvalues.

Theorem 27. *The eigenspace of an eigenvalue is a vector subspace.*

Proof. Let S be the set of all eigenvectors of a linear operator T corresponding to a particular eigenvalue c. Then the eigenspace is defined as,

$$E = S \cup \{\vec{\mathbf{0}}\}.$$

Then E is a subspace because it contains the zero vector and

$$\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}} \in E \implies T(\alpha\vec{\boldsymbol{v}}+\beta\vec{\boldsymbol{w}}) = \alpha(c\vec{\boldsymbol{v}})+\beta(c\vec{\boldsymbol{w}}) = c(\alpha\vec{\boldsymbol{v}}+\beta\vec{\boldsymbol{w}}) \implies \alpha\vec{\boldsymbol{v}}+\beta\vec{\boldsymbol{w}} \in E.$$

□

Corollary 28. *Any linear combination of eigenvectors with eigenvalue c is also an eigenvector with eigenvalue c.*

2.5.3.5 Examples of Eigenvectors and Eigenvalues

(48) If we take the matrix,

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

we can determine its eigenvectors by finding the solutions to the equation,

$$\begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This gives us two simultaneous equations in three unknowns: the two vector dimensions x_1, x_2 and the eigenvalue c,

$$\begin{aligned} 3x_1 + x_2 &= cx_1 \\ 2x_2 &= cx_2 \end{aligned}$$

which imply that,

$$6x_1 = 2cx_1 - cx_2 \iff (6/c - 2)x_1 = -x_2 \iff x_2 = (2 - 6/c)x_1.$$

So, if $x_1 = 1$ then $x_2 = 2 - (6/c)$ and we have the follwing eigenvector/eigenvalue pairs.

$$c = 3 : \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; c = 1 : \begin{bmatrix} 1 \\ -4 \end{bmatrix} \text{ Wrong!}$$

Why does this not work? There is an additional constraint not expressed in the linear system here: eigenvectors are nonzero by definition. This means that we have the additional constraint,

$$x_1 x_2 \neq 0$$

which cannot be expressed in a linear system of equations. When $c \notin \{2, 3\}$ both x_1 and x_2 are zero and the vector is not an eigenvector. So, how can we systematically restrict the values of c to only those that produce valid eigenvectors? If we follow the alternative logic:

$$\begin{aligned}
& \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
\iff \quad & \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - c \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \vec{\mathbf{0}} \\
\iff \quad & \left(\begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} I - cI \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \vec{\mathbf{0}} \\
\iff \quad & \begin{bmatrix} 3-c & 1 \\ 0 & 2-c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \vec{\mathbf{0}}
\end{aligned}$$

Now if we let

$$A = \begin{bmatrix} 3-c & 1 \\ 0 & 2-c \end{bmatrix}$$

then the nullspace of A is the space of vectors $(x_1, x_2)^T$ that satisfy this equation. If A is invertible and nonsingular then this space is the trivial space $\{\vec{\mathbf{0}}\}$ but, if A is singular however, then there exist nonzero vectors that satisfy this equation. Therefore, it is precisely the nonzero vectors that are in the nullspace of this matrix A when it is singular that are the eigenvectors we are looking for. So, if we first determine the values of c for which A is singular, we can then determine the eigenvectors. Since, by Proposition 95, A is singular if and only if $det\, A = 0$, we are looking for *precisely the values of c which make $det\, A = 0$.*

(49) The minimal linear transformation seen in 45 — which is just scalar multiplication — has every vector as its eigenvectors because

$$T(\vec{\boldsymbol{v}}) = A\vec{\boldsymbol{v}} = a\vec{\boldsymbol{v}}$$

for all $\vec{v} \in V$. So every vector in V is an eigenvector with eigenvalue a.

2.5.3.6 Matrices of Eigenvectors

If $\vec{v}_1 \in V$ is a eigenvector of a linear transformation T and we extend the set $\{\vec{v}_1\}$ (by Proposition 74) to a basis of V, say $\{\vec{v}_1, \ldots, \vec{v}_n\}$, then the matrix of T will have the block form,

$$\begin{bmatrix} c & B \\ 0 & D \end{bmatrix} = \begin{bmatrix} c & \ldots & \ldots \\ 0 & \ldots & \ldots \\ \vdots & \ldots & \ldots \\ 0 & \ldots & \ldots \end{bmatrix}$$

where c is the eigenvalue of $\vec{v}_1$. This is the same block decomposition as that shown for T-invariant spaces in 2.5.3.3 with the case of a 1-dimensional invariant subspace.

Proposition 99. *If A is the matrix of a linear operator T with respect to a basis B then the matrix A is diagonal iff every basis vector in B is an eigenvector of T.*

Proof. The defining property of the matrix A is that the j-th column is the coordinates of the image of the j-th basis vector in B under T,

$$A(:j) = T(\vec{v}_j) = a_{1j}\vec{v}_1 + \cdots + a_{nj}\vec{v}_n.$$

For an eigenvector $\vec{v}_j$, $T(\vec{v}_j) = c\vec{v}_j = a_{jj}\vec{v}_j$ so that $a_{jj} = c$ the eigenvalue and for all a_{ij} such that $i \neq j$, $a_{ij} = 0$. □

Corollary 29. *The matrix of a linear operator T over a vector space V is similar to a diagonal matrix iff there exists some basis of V solely comprised of eigenvectors of T.*

Proposition 100. *Similar matrices have the same eigenvalues.*

Proof. Similar matrices represent the same transformation with respect to different bases and for a matrix A representing a transformation T with respect to an arbitrary basis B,

$$T(\vec{\boldsymbol{v}}) = c\vec{\boldsymbol{v}} \iff A\vec{\boldsymbol{v}}_B = c\vec{\boldsymbol{v}}_B.$$

That's to say, the eigenvalues are not dependent on the basis with respect to which a coordinate vector is defined. □

2.5.4 Diagonalisation

Definition. *The process of determining a diagonal matrix that is similar to a given matrix of a linear operator is known as* ***diagonalisation****.*

2.5.4.1 Existence of Eigenvectors

- Every linear operator on a complex vector space has at least one eigenvector and, in most cases, these form a basis.
- Linear operators over real vector spaces need not have eigenvectors (e.g. rotation of the plane $\mathbb{R}^2$ by an angle θ has no eigenvector unless $\theta = 0$ or π).
- Real matrices that are *positive* (having only positive components) are guaranteed to have at least one positive eigenvector.

2.5.4.2 The Effect of Multiplication by a Positive Matrix

TODO: Artin[134]

2.5.4.3 Determining the Eigenvectors

The process of finding eigenvectors is to first determine the eigenvalues and then calculate the eigenvectors that correspond to those eigenvalues. Let I be the identity operator. Then,

$$T(\vec{\boldsymbol{v}}) = c\vec{\boldsymbol{v}} \iff T(\vec{\boldsymbol{v}}) - c\vec{\boldsymbol{v}} = \vec{\mathbf{0}} \iff [T - cI](\vec{\boldsymbol{v}}) = \vec{\mathbf{0}}$$

where the expression $T - cI$ is a linear combination of linear transformations and so is also a linear transformation. Furthermore, it is an operator as both its terms are operators (in fact they need to have the same dimensions in order for the expression to make sense).
Two things are clear from this expression:

(i) The matrix of the linear operator $T - cI$ is $A - cI$ where I is the identity matrix.

(ii) The eigenvector $\vec{\boldsymbol{v}}$ is in the kernel of $T - cI$ and so is in the nullspace of $A - cI$.

Proposition 101. *A linear operator T has a nontrivial kernel iff 0 is an eigenvalue of T.*

Proof. This follows from the fact that, if we let $\vec{\boldsymbol{v}} \neq \vec{\mathbf{0}} \in ker\, T$, then

$$T(\vec{\boldsymbol{v}}) = \vec{\mathbf{0}} = 0\vec{\boldsymbol{v}} = c\vec{\boldsymbol{v}}$$

for $c = 0$ so that a nontrivial kernel implies that 0 is an eigenvalue. Conversely, if 0 is an eigenvalue we must have $0\vec{\boldsymbol{v}} = \vec{\mathbf{0}} = T(\vec{\boldsymbol{v}})$ and, since if a vector $\vec{\boldsymbol{v}}$ is an eigenvector, by definition, $\vec{\boldsymbol{v}} \neq \vec{\mathbf{0}}$, this therefore implies that the kernel contains a nonzero vector. □

Corollary 30. *A linear operator T has all the properties in Proposition 95 iff 0 is an eigenvalue of T.*

Proposition 102. *The eigenvalues of a linear operator T are the scalars $c \in \mathbb{F}$ such that the linear operator $[T - cI]$ is singular.*

Proof. The eigenvalues of a linear operator T are the scalars $c \in \mathbb{F}$ such that there exists a nonzero vector $\vec{\boldsymbol{v}}$ with $[T - cI](\vec{\boldsymbol{v}}) = \vec{\mathbf{0}}$. If such a vector exists then the kernel of $T - cI$ is nontrivial and so, by Proposition 95, $T - cI$ is singular. □

Corollary 31. *If $A - cI$ is the matrix of $T - cI$, **the eigenvalues of T are the scalars $c \in \mathbb{F}$ such that $det(A - cI) = 0$.***

Corollary 32. *The eigenvalues of $A - cI$ are the same as the eigenvalues of $cI - A$.*

Proof. If A is a $n \times n$ matrix representing the operator T then,

$$det(-A) = (-1)^n det(A).$$

So, if $det(A) = 0$ then $det(-A) = 0$ also. Therefore,

$$det(A - cI) = 0 \iff det(cI - A) = 0.$$ □

2.5.4.4 The Characteristic Polynomial

Notation. It is customary to use the variable t to denote the eigenvalue in the characteristic polynomial.

Definition. *The **characteristic polynomial** of a linear operator T is the polynomial,*

$$p(t) = det(tI - A) = \sum s(j_1, \cdots, j_n) a_{1j_1} \cdots a_{nj_n}$$

where the sum is defined over all permutations $j_1, \cdots, j_n$ of $\{1, \ldots, n\}$ and $s(j_1, \cdots, j_n)$ is the sign of the permutation.

The determinant is an expression in which every term is the product of a component in every column and row of the matrix with no column or row appearing more than once in each term,

$$tI - A = \begin{bmatrix} (t - a_{11}) & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & (t - a_{22}) & \cdots & -a_{2n} \\ \vdots & & & \vdots \\ -a_{n1} & \cdots & \cdots & (t - a_{nn}) \end{bmatrix}.$$

It can be seen in the matrix of $tI - A$ that the highest power of t will be obtained in the term of the determinant that forms the product of all the diagonal terms $a_{11} \cdots a_{nn}$ which occurs when $j_1, \cdots, j_n = 1, \ldots, n$. This term of the determinant will be the product of precisely n terms containing the eigenvalue t. Therefore the result is a polynomial of degree n in the eigenvalue t.

Theorem 28. *The eigenvalues of a linear operator are the roots of its characteristic polynomial.*

Proof. If $p(t)$ is the characteristic polynomial of a linear operator T then the values of t for which $p(t) = 0$ are the values of t such that $det(tI - A) = 0$ and these are precisely the eigenvalues. □

Proposition 103. *The eigenvalues of an upper or lower triangular matrix are its diagonal entries.*

Proof. The determinant of a triangular matrix is equal to the product of its diagonal entries and if A is a triangular matrix then $tI - A$ is also triangular. Therefore, the characteristic polynomial is simply,

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

and so the eigenvalues are the diagonal entries $a_{11}, \ldots, a_{nn}$. □

Proposition 104. *A positive matrix (a matrix whose entries are all positive) has at least one eigenvector with positive coordinates.*

> *An abstract vector does not have coordinates so when we refer to a "positive" vector with positive-valued coordinates, this is with respect to a particular basis. In this context, the basis in question is the basis with respect to which the matrix outputs the transformed vectors.*

Proof. TODO: review: this "proof" is not really a proof, more an example using a 2x2 matrix. □

Proposition 105. *The characteristic polynomial of a linear operator does not depend on the basis with respect to which the matrix of the operator is defined.*

Proof. For two similar matrices representing the same linear operator T we have,

$$A' = PAP^{-1}$$

where P is the matrix of change of basis between the bases of A and A'. If we form the characteristic polynomial of A',

$$\begin{aligned}
& tI - A' = tI - PAP^{-1} && \\
\iff \quad & = PtIP^{-1} - PAP^{-1} && PtIP^{-1} = tPP^{-1} = tI \\
\iff \quad & = P(tI - A)P^{-1} && \text{by distributivity of matrix multiplication.}
\end{aligned}$$

Then,

$$\begin{aligned} & det(tI - A') = det(P(tI - A)P^{-1}) \\ \iff & \qquad = det\, P \cdot det(tI - A) \cdot det\, P^{-1} \\ \iff & \qquad = det(tI - A). \end{aligned}$$

This result, $det(tI - A') = det(tI - A)$, must hold for all t and therefore implies that, for p, p' characteristic polynomials of A and A' respectively,

$$\forall t \in \mathbb{F} \; . \; p(t) = p'(t).$$

This implies that the characteristic polynomials are equal. □

Proposition 106. *The characteristic polynomial $p(t)$ of a matrix A has the form*

$$p(t) = t^n - (tr\, A)t^{n-1} + \cdots + (-1)^n(det\, A),$$

where $tr\, A$ is the trace of A (see: 2.3.2.1):

$$tr\, A = a_{11} + \cdots + a_{nn}.$$

Proof. Calculation of the characteristic polynomial of a matrix A is calculation of $p(t) = det(tI - A)$ the determinant of the matrix $tI - A$ which takes the form,

$$tI - A = \begin{bmatrix} (t - a_{11}) & \cdots & \cdots & \vdots \\ \vdots & (t - a_{22}) & \cdots & \vdots \\ \vdots & & & \vdots \\ \vdots & \cdots & (t - a_{(n-1)(n-1)}) & -a_{(n-1)n} \\ \vdots & \cdots & -a_{n(n-1)} & (t - a_{nn}) \end{bmatrix}.$$

This calculation proceeds with terms of products of elements from each row and a permutation of the column indices (see: 2.3.5.2) of which we examine

the first two terms.

The first term is for the identity permutation $j_1, \ldots, j_n = 1, \ldots, n$ along the diagonal:

$$\begin{aligned} a_{11} \cdots a_{nn} &= (t - a_{11}) \cdots (t - a_{nn}) \\ &= t^n - (a_{11} + \cdots + a_{nn}) t^{n-1} \\ &\quad + (a_{11}a_{22} + a_{11}a_{33} + \cdots + a_{(n-1)(n-1)}a_{nn}) t^{n-2} \\ &\quad \cdots + (-1)^n (a_{11} \cdots a_{nn}) \end{aligned}$$

The second term is the permutation one swap away from identity

$$j_1, \ldots, j_n = 1, \ldots, n, (n-1)$$

which is an odd permutation, so the sign is -1:

$$\begin{aligned} &(-1) a_{11} \cdots a_{(n-2)(n-2)} a_{(n-1)n} a_{n(n-1)} \\ &= (-1)(t - a_{11}) \cdots (t - a_{(n-2)(n-2)})(-a_{(n-1)n})(-a_{n(n-1)}) \\ &= -a_{(n-1)n} a_{n(n-1)} t^{n-2} + a_{(n-1)n} a_{n(n-1)} (a_{11} + \cdots + a_{(n-2)(n-2)}) t^{n-3} \\ &\quad \cdots - (-1)^n (a_{11} \cdots a_{(n-2)(n-2)} a_{(n-1)n} a_{n(n-1)}) \end{aligned}$$

From these first two terms we can discern enough about the general pattern of the characteristic polynomial to see that the first two terms in t^n and t^{n-1} are produced by the first permutation and take the form

$$t^n - (a_{11} + \cdots + a_{nn}) t^{n-1} = t^n - (tr\, A) t^{n-1}$$

as required. We can also see that the final terms of each permutation — those that involve no powers of t — are going to sum up to the value of $(-1)^n (det\, A)$. Therefore the characteristic polynomial takes the form,

$$p(t) = t^n - (tr\, A) t^{n-1} + \cdots + (-1)^n (det\, A)$$

as claimed. □

Corollary 33. *The trace of a matrix of a linear operator is independent of the basis with respect to which the matrix is defined.*

Proof. Let T be a linear operator and A be the matrix of T with respect to a basis B. Then Proposition 105 tells us that any matrix of the same linear operator defined with respect to some other basis (i.e. a similar matrix to A) has the same characteristic polynomial. Proposition 106 tells us that the coefficients of this characteristic polynomial include the trace of the matrix A. Therefore, if A' is a matrix of T defined against the basis B' and P is the change of basis matrix such that $B'P = B$ then,

$$A' = PAP^{-1} \iff p'(t) = p(t) \iff tr\, A' = tr\, A. \qquad \square$$

As a result of this we can refer to the characteristic polynomial, determinant and trace of a linear operator T without reference to a particular matrix or basis.

Proposition 107. *Let T be a linear operator on a finite-dimensional vector space V.*

(i) If V has dimension n, then T has at most n eigenvalues.

(ii) If $\mathbb{F}$ is the field of complex numbers and $V \neq 0$, then T has at least one eigenvalue, and hence it has an eigenvector.

Proof. (i) For any field $\mathbb{F}$, a polynomial of degree n can have at most n different roots (see Artin[373]). Since T is defined over a vector space of dimension n, the degree of the characteristic polynomial of T is n. Then, by Theorem 28 we can have a maximum of n eigenvalues.

(ii) Every polynomial of positive degree with complex coefficients has at least one complex root. This fact is called the Fundamental Theorem of Algebra (wikipedia).

$\square$

Proposition 108. *Let T be a linear operator on a finite-dimensional complex vector space V. There is a basis B of V such that the matrix A of T is upper triangular.*

Proof. By Proposition 107, T has at least one eigenvector. We can extend this eigenvector to a basis of V, say,

$$B' = \{\vec{\boldsymbol{v}}'_1, \ldots, \vec{\boldsymbol{v}}'_n\}.$$

Then the first column of the matrix A' of T with respect to B' will be

$$(c_1, 0, \ldots, 0)^T$$

where c_1 is the eigenvalue of $\vec{\boldsymbol{v}}'_1$. Therefore A' has the form

$$A' = \begin{bmatrix} c_1 & \cdots \\ 0 & D \end{bmatrix}$$

where D is a $(n-1) \times (n-1)$ matrix and, if $P = [B']^{-1} I = [B']^{-1}$ is the change of basis matrix, then

$$A' = PAP^{-1}.$$

Now we can use induction on the dimension of the matrix n and the induction hypothesis will be that there exists some upper triangular

$$Q' = QDQ^{-1}.$$

Define

$$Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}.$$

Then

$$(Q_1 P)A(Q_1 P)^{-1} = Q_1 PAP^{-1} {Q_1}^{-1} = Q_1 A' {Q_1}^{-1}$$

takes the form

$$\begin{bmatrix} c_1 & \cdots \\ 0 & QDQ^{-1} \end{bmatrix}$$

which is upper triangular. This proves the induction step. □

Note that this proof is over the complex number field but the same proof would work over any field that contains all the roots of the characteristic polynomial.

Proposition 109. *Let $\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_r \in V$ be eigenvectors for a linear operator T, with distinct eigenvalues $c_1, \ldots, c_r$. Then the set $\{\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_r\}$ is linearly independent.*

Proof. Assume for contradiction that there exists a linear relation between the set of eigenvectors,

$$\alpha_1 \vec{\boldsymbol{v}}_1 + \cdots + \alpha_r \vec{\boldsymbol{v}}_r = \vec{\boldsymbol{0}}.$$

Linearity of T gives us,

$$T(\alpha_1 \vec{\boldsymbol{v}}_1 + \cdots + \alpha_r \vec{\boldsymbol{v}}_r) = \alpha_1 T(\vec{\boldsymbol{v}}_1) + \cdots + \alpha_r T(\vec{\boldsymbol{v}}_r) = T(\vec{\boldsymbol{0}}) = \vec{\boldsymbol{0}}$$

while the eigenvector property gives us,

$$\alpha_1 T(\vec{\boldsymbol{v}}_1) + \cdots + \alpha_r T(\vec{\boldsymbol{v}}_r) = \alpha_1 c_1 \vec{\boldsymbol{v}}_1 + \cdots + \alpha_r c_r \vec{\boldsymbol{v}}_r$$

so we have the simultaneous equations,

$$\begin{aligned} \alpha_1 \vec{\boldsymbol{v}}_1 + \cdots + \alpha_r \vec{\boldsymbol{v}}_r &= \vec{\boldsymbol{0}} \\ \alpha_1 c_1 \vec{\boldsymbol{v}}_1 + \cdots + \alpha_r c_r \vec{\boldsymbol{v}}_r &= \vec{\boldsymbol{0}}. \end{aligned}$$

If we multiply the first equation by c_r and subtract the second equation from it we get,

$$\alpha_1 (c_r - c_1) \vec{\boldsymbol{v}}_1 + \cdots \alpha_{r-1} (c_r - c_{r-1}) \vec{\boldsymbol{v}}_{r-1} = \vec{\boldsymbol{0}}.$$

Since all the eigenvalues are distinct, for $i \neq j$, $c_i - c_j \neq 0$ and the eigenvectors $\vec{\boldsymbol{v}}_i$, by definition, are nonzero. So, this equation implies that either $\alpha_1, \ldots, \alpha_{r-1} = 0$ or there is a linear relation between the vectors $\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_{r-1}$. This dependence of the properties of the r-length list on the properties of the $(r-1)$-length list signals that we can set up a proof by induction using the hypothesis that a k-length list is linearly independent.
If we use $k = 2$ as the base case, set up the linear relation and use the eigenvector property as before then this results in,

$$\alpha_1 (c_2 - c_1) \vec{\boldsymbol{v}}_1 = \vec{\boldsymbol{0}}.$$

As before, both $c_2 - c_1$ and $\vec{\boldsymbol{v}}_1$ are nonzero so this implies that $\alpha_1 = 0$. This, in turn, implies that $\alpha_2 = 0$ also and so, the list of length $k = 2$ is linearly independent.
Then the induction step is to assume that the list of length $k = r - 1$ is

linearly independent and show that this implies that the list of length $k = r$ is linearly independent. We have already shown that if we set up a linear relation on a list of eigenvectors of length $k = r$ then the eigenvector property implies that,

$$\alpha_1(c_r - c_1)\vec{\boldsymbol{v}}_1 + \cdots \alpha_{r-1}(c_r - c_{r-1})\vec{\boldsymbol{v}}_{r-1} = \vec{\mathbf{0}}.$$

Now we can use the induction hypothesis to assert that $\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_{r-1}$ are linearly independent implying that $\alpha_1, \ldots, \alpha_{r-1} = 0$. This, in turn, implies that $\alpha_r = 0$ meaning that $\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_r$ is linearly independent. □

Theorem 29. *Let T be a linear operator on a vector space V of dimension n over a field F. Assume that its characteristic polynomial has n* ***distinct*** *roots in F. Then there is a basis for V with respect to which the matrix of T is diagonal.*

Proof. If the characteristic polynomial has n distinct roots then there are n distinct eigenvalues along with their associated eigenvectors. By Proposition 109, these eigenvectors form a linearly independent set. Since the dimension of the space V is n, by Theorem 21, these eigenvectors form a basis of V. Then, by Proposition 99, the matrix of T with respect to this basis is diagonal. □

Note that:

- *The diagonal entries of the matrix of a linear operator with respect to a basis of eigenvectors are the eigenvalues. For this reason, the set of values is wholly determined by the linear operator although the order is determined on the order of the vectors in the basis set (which is not significant);*
- *If a matrix A is found to be similar to a diagonal matrix B via a change of basis expressed in the matrix P then,*

  $$A^m = (P^{-1}BP)^m = P^{-1}B^mP.$$

  This is because — remembering that matrices don't, in general, commute — A^m takes the form,

  $$A^m = (P^{-1}BP)(P^{-1}BP)\ldots(P^{-1}BP)(P^{-1}BP)$$

$$= P^{-1}B(PP^{-1})BP \dots P^{-1}B(PP^{-1})BP$$
$$= P^{-1}B^m P.$$

Corollary 34. *If a linear operator on a vector space of dimension n over a field F has a characteristic polynomial with n distinct roots then its determinant is equal to the product of its eigenvalues and its trace is equal to the sum of its eigenvalues.*

Proof. By Proposition 98 and Corollary 33, the determinant and trace are independent of the basis with respect to which the matrix is defined and so the existence of a basis with respect to which the matrix is diagonal means that we can look at the determinant and trace of the diagonal matrix. □

2.5.4.5 Examples of Diagonalization using the Characteristic Polynomial

(50) Let the real-valued matrix A be,

$$A = \begin{bmatrix} 4 & 0 & 4 \\ 0 & 4 & 4 \\ 4 & 4 & 8 \end{bmatrix}.$$

Constructing the characteristic polynomial,

$$\begin{aligned} |A - \lambda I| &= \begin{vmatrix} 4-\lambda & 0 & 4 \\ 0 & 4-\lambda & 4 \\ 4 & 4 & 8-\lambda \end{vmatrix} \\ &= (4-\lambda)\begin{vmatrix} 4-\lambda & 4 \\ 4 & 8-\lambda \end{vmatrix} + 4\begin{vmatrix} 0 & 4-\lambda \\ 4 & 4 \end{vmatrix} \\ &= (4-\lambda)((4-\lambda)(8-\lambda) - 16 - 16) \\ &= (4-\lambda)(\lambda^2 - 12\lambda) \\ &= (4-\lambda)\lambda(\lambda - 12). \end{aligned}$$

So the eigenvalues are 4,0 and 12. To find an eigenvector for 4 we need to solve the equation $(A - 4I)\vec{\boldsymbol{x}} = \vec{\mathbf{0}}$ so we construct the matrix,

$$A - 4I = \begin{bmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 4 \end{bmatrix}$$

and then find the nullspace of this matrix using row reduction,

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

which gives $x_3 = 0$ and one free variable x_2. So

$$\vec{\boldsymbol{x}} = \begin{bmatrix} -t \\ t \\ 0 \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \qquad t \neq 0 \in \mathbb{R}.$$

Note that $t \neq 0$ because eigenvectors are nonzero by definition.

2.5.4.6 Jordan Canonical Form

TODO: from LSE

2.5.5 Orthonormal Bases and Orthogonal Operators

2.5.5.1 Orthonormal Bases

Definition. *A basis consisting of pairwise mutually orthogonal (see: 2.4.8.8) unit vectors is called an* ***orthonormal basis****.*

Proposition 110. *Let $U = \{\vec{\boldsymbol{u}}_1, \ldots, \vec{\boldsymbol{u}}_n\}$ be an orthonormal basis. Then, for any $\vec{\boldsymbol{u}}_i, \vec{\boldsymbol{u}}_j \in U$, $i \neq j$,*

$$\vec{\boldsymbol{u}}_i \cdot \vec{\boldsymbol{u}}_j = 0 \quad \textit{and} \quad \vec{\boldsymbol{u}}_i \cdot \vec{\boldsymbol{u}}_i = 1 = \vec{\boldsymbol{u}}_j \cdot \vec{\boldsymbol{u}}_j.$$

Proof. The elements of U are pairwise mutually orthogonal so — by the definition of orthogonal vectors (2.4.8.8) — $\vec{\boldsymbol{u}}_i \cdot \vec{\boldsymbol{u}}_j = 0$ for $i \neq j$. Furthermore,

$$\vec{\boldsymbol{u}}_i \cdot \vec{\boldsymbol{u}}_i = |\vec{\boldsymbol{u}}_i|^2 = 1$$

because the length of unit vectors is 1. □

2.5.5.2 Orthonormal Bases in Coordinate Vectors

If we consider what an orthonormal basis would look like in coordinate vectors. Obviously, the standard basis is an orthonormal basis as,

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0$$

and clearly for any i, j such that $i \neq j$, $\vec{\boldsymbol{e_i}} \cdot \vec{\boldsymbol{e_j}} = 0$ and $\vec{\boldsymbol{e_i}}, \vec{\boldsymbol{e_j}}$ have unit length.

We can form other orthonormal bases in $\mathbb{R}^3$ though. The vectors,

$$\begin{bmatrix}1\\1\\0\end{bmatrix} \cdot \begin{bmatrix}-1\\1\\0\end{bmatrix} = 0$$

are orthogonal but do not have unit length. We can make them unit length, though, by dividing them by $\sqrt{2}$ so that,

$$\begin{bmatrix}1/\sqrt{2}\\1/\sqrt{2}\\0\end{bmatrix}, \begin{bmatrix}-1/\sqrt{2}\\1/\sqrt{2}\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$$

is also an orthonormal basis of $\mathbb{R}^3$.

Another orthonormal basis of $\mathbb{R}^3$ is

$$\begin{bmatrix}\cos\theta\\\sin\theta\\0\end{bmatrix}, \begin{bmatrix}-\sin\theta\\\cos\theta\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$$

for any angle θ measured anticlockwise from the x-axis. Actually this is the general case of which the previous example is a special case when $\theta = \pi/2$.

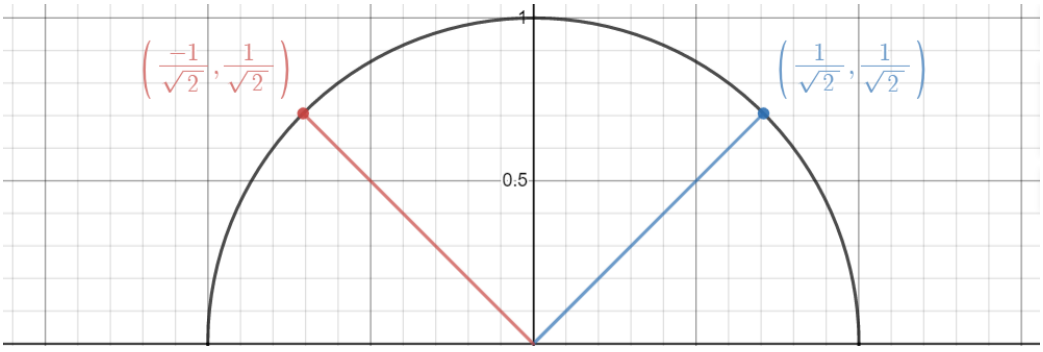


Figure 2.3: $(\cos\pi/4, \sin\pi/4)^T = (1/\sqrt{2}, 1/\sqrt{2})^T$, $(-\sin\pi/4, \cos\pi/4)^T = (-1/\sqrt{2}, 1/\sqrt{2})^T$

But for any value of the angle θ these remain orthogonal as can be seen if we generate the basis vectors for $\theta = \pi/3$.

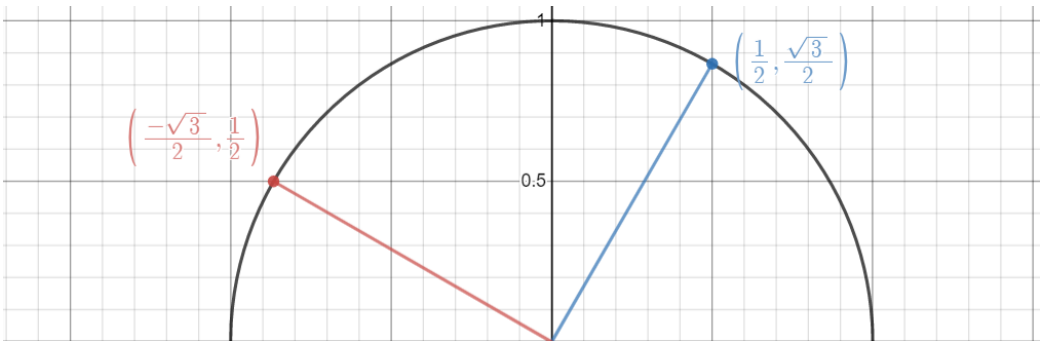


Figure 2.4: $(\cos \pi/3, \sin \pi/3)^T = (1/2, \sqrt{3}/2)^T$, $(-\sin \pi/3, \cos \pi/3)^T = (-\sqrt{3}/2, 1/2)^T$

2.5.5.3 Orthogonal Operators

Definition. *A matrix whose columns form an orthonormal basis is called an* ***orthogonal matrix****.*

The operation of left multiplication by such a matrix is called an ***orthogonal operator****.*

Proposition 111. *If a matrix A is orthogonal then $A^T = A^{-1}$.*

Proof. A is an orthogonal matrix so, by definition, its columns form an orthogonal basis. Then,

$$A^T A = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & & & \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}^2 + \cdots + a_{n1}^2 & a_{11}a_{12} + \cdots + a_{n1}a_{n2} & \cdots & \\ a_{12}a_{11} + \cdots + a_{n2}a_{n1} & a_{12}^2 + \cdots + a_{n2}^2 & \cdots & \\ \vdots & & & \\ \cdots & \cdots & \cdots & a_{1n}^2 + \cdots + a_{nn}^2 \end{bmatrix}$$

Along the main diagonal the components take the form

$$a_{1j}^2 + a_{2j}^2 + \cdots + a_{nj}^2 = a_j \cdot a_j$$

where a_j is the jth column of the matrix A. But the columns of A are vectors in an orthonormal basis and so $a_j \cdot a_j = 1$.
Furthermore, the off-diagonal values take the form

$$a_{1j}a_{1j'} + \cdots + a_{nj}a_{nj'} = a_j \cdot a_{j'}$$

for $j \neq j'$. Since the columns are from an orthonormal basis we know that $a_j \cdot a_{j'} = 0$.
Therefore the resultant matrix looks like,

$$A^T A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_n.$$

Clearly, the same effect would be seen for AA^T and so,

$$A^T A = AA^T = I \iff A^T = A^{-1}. \qquad \square$$

Proposition 112. *Left multiplication by an orthogonal matrix preserves the dot product. In other words, for all vectors $\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}}$, A is an orthogonal matrix if and only if,*

$$A\vec{\boldsymbol{v}} \cdot A\vec{\boldsymbol{w}} = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}}.$$

Proof. Using Proposition 111 and the matrix formula for the dot product we can deduce that, for all vectors $\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}}$,

$$\begin{aligned} A\vec{\boldsymbol{v}} \cdot A\vec{\boldsymbol{w}} &= (A\vec{\boldsymbol{v}})^T A\vec{\boldsymbol{w}} \\ &= \vec{\boldsymbol{v}}^T A^T A\vec{\boldsymbol{w}} \\ &= \vec{\boldsymbol{v}}^T \vec{\boldsymbol{w}} && A \text{ is orthogonal so } A^T A = I \\ &= \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \end{aligned}$$

which is to say that an orthogonal matrix preserves the dot product.

Conversely, if we assume that A preserves the dot product then, for all vectors $\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}}$,

$$\begin{aligned}
& A\vec{\boldsymbol{v}} \cdot A\vec{\boldsymbol{w}} = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\
\iff \quad & (A\vec{\boldsymbol{v}})^T A\vec{\boldsymbol{w}} = \vec{\boldsymbol{v}}^T \vec{\boldsymbol{w}} \\
\iff \quad & \vec{\boldsymbol{v}}^T A^T A\vec{\boldsymbol{w}} = \vec{\boldsymbol{v}}^T \vec{\boldsymbol{w}} \\
\iff \quad & \vec{\boldsymbol{v}}^T A^T A\vec{\boldsymbol{w}} - \vec{\boldsymbol{v}}^T \vec{\boldsymbol{w}} = 0 \qquad \text{both terms are scalars} \\
\iff \quad & \vec{\boldsymbol{v}}^T (A^T A - I)\vec{\boldsymbol{w}} = 0.
\end{aligned}$$

Now for any arbitrary matrix B,

$$e_i^T B e_j = b_{ij}$$

where b_{ij} is the (i, j)th element of B. Then, for

$$e_i^T B e_j = 0$$

to be true *for all possible* e_i, e_j would require that,

$$\forall i, j \; . \; b_{ij} = 0 \iff B = [0]$$

where $[0]$ is the zero matrix. Therefore,

$$\forall \vec{\boldsymbol{v}}, \vec{\boldsymbol{w}} \; . \; \vec{\boldsymbol{v}}^T (A^T A - I)\vec{\boldsymbol{w}} = 0 \iff A^T A - I = [0] \iff A^T A = I$$

and so if A preserves the dot product then it is orthogonal. □

Proposition 113. *The determinant of any orthogonal matrix is 1 or -1.*

Proof. If a matrix A is orthogonal then $A^T A = I$ which implies that,

$$det(A^T) det(A) = det(I) = 1.$$

By Proposition 56, $det(A^T) = det(A)$ so,

$$det(A^T) det(A) = det(A)^2 = 1 \iff \sqrt{det(A)} = 1 \iff det(A) = \pm 1. \quad \square$$

*If an orthogonal operator has determinant equal to 1 it is described as **orientation preserving** and if it is equal to -1 it is described as **orientation reversing**.*

Proposition 114. *The orthogonal matrices form a subgroup of $GL_n(\mathbb{F})$.*

Proof. Let $S = \{ A \in GL_n(\mathbb{F}) \mid A^T A = I \}$. Then,

- S is nonempty because $I \in S$.
- For $B, C \in S$,

  $$(BC)^T(BC) = C^T B^T (BC) = C^T (B^T B) C = I$$

  so $BC \in S$.
- For $B \in S$, by Proposition 111, $B^{-1} = B^T$ and

  $$(B^T)^T B^T = BB^T = BB^{-1} = I$$

  so S contains inverses.

Therefore $S \leq GL_n(\mathbb{F})$. □

*The subgroup of the general linear group formed by the orthogonal matrices is called the **orthogonal group** and is denoted O_n.*

2.5.6 Linear and Affine Transformations

2.5.6.1 Affine Spaces

Definition. *An **affine space** is a generalization of a Euclidean space in which there is no particular point designated as the origin. As a result vectors can be viewed as displacements rather than points.*

Let V be a vector space and P be a set of points. Then we can form an affine space over V and P by defining the vectors in V as displacements connecting members of P such that, for $Q_1, Q_2 \in P$, $\vec{\boldsymbol{v}} \in V$,

$$Q_1 + \vec{\boldsymbol{v}} = Q_2 \iff Q_2 - Q_1 = \vec{\boldsymbol{v}} \iff Q_1 - Q_2 = -\vec{\boldsymbol{v}}.$$

Definition. *A **frame** of an affine space is an extension of a basis of its underlying vector space to include a point designated as an origin. If $\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n$ is a basis of a vector space V and Q is a point in the set of points P, then $F = (\vec{\boldsymbol{v}}_1, \ldots, \vec{\boldsymbol{v}}_n, Q)$ is a frame of the affine space over V and P.*

Definition. *The **dimension** of an affine space is the dimension of the underlying vector space.*

Proposition 115. *Any linear combination of points in an affine space where the coefficients sum to 0 results in a vector.*

Proof. Let S be a sum of n points $Q_i \in P$ in an affine space associated with a vector space V such that,

$$S = \sum_{i=0}^{n} \alpha_i Q_i \quad \text{and} \quad \sum_{i=0}^{n} \alpha_i = 0.$$

Then, if we take the partial sum of the first two points,

$$S_2 = \alpha_1 Q_1 + \alpha_2 Q_2 = \alpha_1(Q_1 - Q_2) + (\alpha_1 + \alpha_2)Q_2$$

and then the next partial sum of the first three points,

$$\begin{aligned} S_3 &= \alpha_1(Q_1 - Q_2) + (\alpha_1 + \alpha_2)Q_2 + \alpha_3 Q_3 \\ &= \alpha_1(Q_1 - Q_2) + (\alpha_1 + \alpha_2)(Q_2 - Q_3) + (\alpha_1 + \alpha_2 + \alpha_3)Q_3 \end{aligned}$$

we can see that, by induction, the nth sum is,

$$\begin{aligned} S = \alpha_1(Q_1 - Q_2) \\ &+ (\alpha_1 + \alpha_2)(Q_2 - Q_3) \\ &+ (\alpha_1 + \alpha_2 + \alpha_3)(Q_3 - Q_4) \\ &\vdots \\ &+ (\alpha_1 + \cdots + \alpha_{n-1})(Q_{n-1} - Q_n) \\ &+ (\alpha_1 + \cdots + \alpha_n)Q_n. \end{aligned}$$

But we have $\alpha_1 + \cdots + \alpha_n = 0$ so the final term is 0. As a result S is a summation of terms of the form,

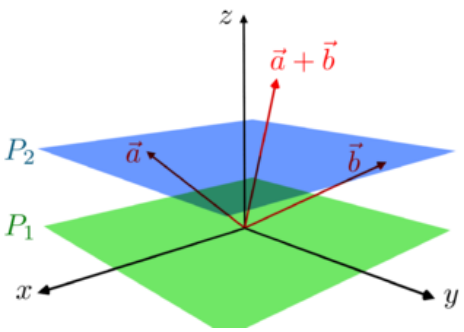
$$(\alpha_1 + \cdots + \alpha_{i-1})(Q_{i-1} - Q_i)$$

where $Q_{i-1} - Q_i$ is a vector. Therefore S is a linear combination of vectors in V and is therefore also a vector in V. □

Corollary 35. *Any linear combination of points in an affine space where the coefficients sum to 1 results in a point.*

Proof. In the preceding proof if we had, instead, $\alpha_1 + \cdots + \alpha_n = 1$ then the final term would equal Q_n and the resulting summation would be a vector plus the point Q_n. Therefore the sum is a point. □

2.5.6.2 Intuition of Affine Spaces



A translated linear subspace of a vector space like P_2 above that no longer passes through the origin is referred to as an **affine subspace**. It is not a vector space as $\vec{\mathbf{0}} \notin P_2$ and $\vec{\boldsymbol{a}}, \vec{\boldsymbol{b}} \in P_2$ but $\vec{\boldsymbol{a}} + \vec{\boldsymbol{b}} \notin P_2$. However, if we instead consider displacements between points — e.g. $\vec{\boldsymbol{b}} - \vec{\boldsymbol{a}}$ — then we see the relationship between affine spaces and vector spaces: $\vec{\boldsymbol{b}} - \vec{\boldsymbol{a}} \in P_1$. The displacements between points in P_2 form a linear subspace. So we can define an affine space A based on the set of points in P_2 and the vectors in P_1.

2.5.6.3 Affine Combinations

Definition. *An **affine combination** of vectors is a combination such that the coefficients sum to 1.*

The definition of an affine combination differs from that of a convex combination in that the coefficients of a convex combination are additionally required to be non-negative.

In an affine space there is no particular point designated as the origin but we can describe vector displacements between points as an ordered pair of points, for example, $(p, a) = \vec{\boldsymbol{pa}}$. Affine combinations of displacements agree on the resulting point with linear combinations in the corresponding Euclidean space. For example, imagine a point p in an affine space is at coordinates $(-1, 4)$ in the corresponding Euclidean space and similarly points a and b are at $(3, 4)$ and $(6, 1)$ respectively. Then, if we take an affine combination of the displacements to a and b the resulting point is independent of the chosen origin point.

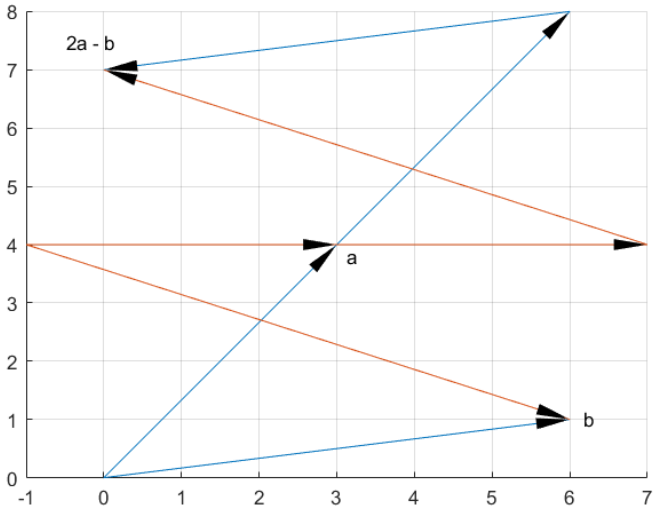


Figure 2.5: The diagram shows the affine combination $2\vec{\boldsymbol{a}} - \vec{\boldsymbol{b}} = 2\vec{\boldsymbol{pa}} - \vec{\boldsymbol{pb}}$ where $\vec{\boldsymbol{a}}$ denotes the usual position vector in the Euclidean space.

2.5.6.4 Affine Transformations

Definition. *Let $T : A_1 \longmapsto A_2$ be a mapping between the affine spaces A_1 and A_2. Then T is an **affine transformation** if:*

- *T maps vectors to vectors and points to points.*
- *T is a linear transformation over the vectors in the underlying vector space.*
- *$T(Q + \vec{\boldsymbol{v}}) = T(Q) + T(\vec{\boldsymbol{v}})$.*

*In an affine space a **translation** can be regarded as a change of frame in which we **change the origin point** while the vector basis may remain unchanged.*

Proposition 116. *Affine transformations preserve parallelism.*

Proof. Let A be an affine space defined over a set of points P and a vector space V and let $Q_1, Q_2 \in P$ and $\vec{\boldsymbol{v}} \in V$. Then, for $s, t \in \mathbb{F}$,

$$l_1 = Q_1 + t\vec{\boldsymbol{v}} \quad \text{and} \quad l_2 = Q_2 + s\vec{\boldsymbol{v}}$$

are parallel lines in A. Let T be an arbitrary affine transformation over A. Then,

$$l_1' = T(l_1) = T(Q_1) + tT(\vec{\boldsymbol{v}}) \quad \text{and} \quad l_2' = T(l_2) = T(Q_2) + sT(\vec{\boldsymbol{v}})$$

are also parallel. □

Proposition 117. *An affine transformation that preserves the dot product is left multiplication by an orthogonal matrix.*

Proof. Firstly, note that if a transformation m preserves the dot product and also fixes the standard basis vectors $\vec{\boldsymbol{e_i}}$ then,

$$m(\vec{\boldsymbol{e_i}}) = \vec{\boldsymbol{e_i}} \quad \text{and} \quad x_i = \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{e_i}} = m(\vec{\boldsymbol{x}}) \cdot m(\vec{\boldsymbol{e_i}}) = m(\vec{\boldsymbol{x}}) \cdot \vec{\boldsymbol{e_i}} = m(\vec{\boldsymbol{x}})_i.$$

Therefore, such a transformation m is the identity transformation.
Now, assume a transformation m' preserves the dot product (but does not necessarily fix the standard basis vectors) in $\mathbb{R}^n$ and let

$$B' = \{m'(\vec{\boldsymbol{e_1}}), \ldots, m'(\vec{\boldsymbol{e_n}})\}$$

be the transformed standard basis vectors. Then, if $[B']$ is the matrix whose columns are the elements of B' then, because m' preserves the dot product, $[B']$ is an orthogonal matrix as, by Proposition 114, is $[B']^{-1}$. So, composing this with m',

$$m'' = [B']^{-1} m'$$

is a transformation that both preserves the dot product and fixes the standard basis vectors. Therefore we have,

$$m'' = I_n = [B']^{-1} m' \iff [B'] = m'$$

so that m' is left multiplication by $[B']$, an orthogonal matrix. □

Properties of affine transformations
We can characterize affine transformations according to their form and behaviour in the Euclidean spaces $\mathbb{R}^2$ and $\mathbb{R}^3$. Specifically, whether the transformation:

1. Fixes the origin: $T(\vec{\mathbf{0}}) = \vec{\mathbf{0}}$
2. Preserves the dot product: $T(\vec{\boldsymbol{v}}) \cdot T(\vec{\boldsymbol{w}}) = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}}$, matrix is orthogonal
3. Preserves distances: $|T(\vec{\boldsymbol{v}}) - T(\vec{\boldsymbol{w}})| = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|$
4. Preserves angles: matrix is scalar multiple of orthogonal matrix
5. Preserves orientation: transformation does not include a reflection, determinant of matrix is positive
6. Preserves parallelism: transformation exhibits affine property (linearity under affine combinations)

As we can see here, the property of preserving parallelism depends only on the affine property so clearly, all affine transformations exhibit this behaviour. As a consequence of preserving parallelism, affine transformations preserve the dimension of affine subspaces (points, lines, planes, etc.). They do not preserve distances between points however, but they do preserve ratios of distances between points lying on a straight line.

Classes of affine transformations
The most common subclassifications of affine transformations are:

- Linear: preserves the origin.
- Conformal: preserves angles.
- Isometry (also known as a congruent transformation): preserves distances in metric spaces and so also implicity, angles. In a Euclidean space these transformations are known as a Euclidean isometries or rigid transformations.
- Rigid Motion: Euclidean isometry / rigid transformation that also preserves orientation.

2.5.6.5 Isometries

A Euclidean isometry or rigid motion, for example, carries a triangle to a congruent triangle. So, it preserves distances and angles but not necessarily orientation (a reflection flips the orientation).
The composition of two rigid motions is a rigid motion and the inverse of a rigid motion is also a rigid motion. Therefore, the rigid motions of $\mathbb{R}^n$ form a group under composition of operations. This group is called the *group of motions* and denoted M_n.

Proposition 118. *A rigid motion that fixes the origin preserves the dot product.*

Proof. Let m be a rigid motion that fixes the origin. Then m is an isometry and so preserves distances and also m maps the origin to the origin. So we have,

$$|m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})| = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}| \quad \text{and} \quad m(\vec{\mathbf{0}}) = \vec{\mathbf{0}}.$$

We can rewrite the isometry property as,

$$\begin{aligned} & \sqrt{(m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})) \cdot (m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}}))} = \sqrt{(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})} \\ \iff \quad & (m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})) \cdot (m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})) = (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}). \end{aligned}$$

Now, if we take the case where $\vec{\boldsymbol{w}} = \vec{\mathbf{0}} = m(\vec{\boldsymbol{w}})$,

$$\begin{aligned} & (m(\vec{\boldsymbol{v}}) - \vec{\mathbf{0}}) \cdot (m(\vec{\boldsymbol{v}}) - \vec{\mathbf{0}}) = (\vec{\boldsymbol{v}} - \vec{\mathbf{0}}) \cdot (\vec{\boldsymbol{v}} - \vec{\mathbf{0}}) \\ \iff \quad & m(\vec{\boldsymbol{v}}) \cdot m(\vec{\boldsymbol{v}}) = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}}. \end{aligned}$$

and we can deduce that $m(\vec{\boldsymbol{x}}) \cdot m(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{x}}$ for any vector $\vec{\boldsymbol{x}}$. Using this along with the properties of the dot product we obtain,

$$\begin{aligned} & (m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})) \cdot (m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})) = (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \\ \iff \quad & m(\vec{\boldsymbol{v}}) \cdot m(\vec{\boldsymbol{v}}) + m(\vec{\boldsymbol{w}}) \cdot m(\vec{\boldsymbol{w}}) - 2m(\vec{\boldsymbol{v}}) \cdot m(\vec{\boldsymbol{w}}) = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} + \vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{w}} - 2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\ \iff \quad & -2m(\vec{\boldsymbol{v}}) \cdot m(\vec{\boldsymbol{w}}) = -2\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}} \\ \iff \quad & m(\vec{\boldsymbol{v}}) \cdot m(\vec{\boldsymbol{w}}) = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{w}}. \end{aligned}$$

Therefore, m preserves the dot product and, by Proposition 117, is left multiplication by an orthogonal matrix. □

Corollary 36. *A rigid motion that fixes the origin is left multiplication by an orthogonal matrix and, therefore, also a linear operator.*

Proposition 119. *Left multiplication by any orthogonal matrix is a Euclidean isometry (a rigid motion) that fixes the origin.*

Proof. By, Proposition 112, left multiplication by an orthogonal matrix preserves the dot product. So, if m is an affine transformation such that $m(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}$ where A is an orthogonal matrix then,

$$\begin{aligned} & |m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})|^2 = m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) = (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) \cdot (\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|^2 \\ \iff & |m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})| = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}| \end{aligned}$$

But also, left multiplication by a matrix is a linear operator so $m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}) = m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})$ meaning that,

$$|m(\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}})| = |m(\vec{\boldsymbol{v}}) - m(\vec{\boldsymbol{w}})| = |\vec{\boldsymbol{v}} - \vec{\boldsymbol{w}}|$$

which is the isometry property for m. □

Linear
Isometries that fix the origin are linear operators and rigid motions in Euclidean space.

Translation
If T is a translation then, in general, $T(\vec{\mathbf{0}}) \neq \vec{\mathbf{0}}$ so translations do not fix the origin and, as a result, are **not** linear transformations. However, translations preserve distances and angles (and orientation because there is no reflection) so they are rigid motions. For example:

(51) Let $\vec{\boldsymbol{v}} = (v_1, \ldots, v_n)$ be any fixed vector in $\mathbb{R}^n$. Then translation by $\vec{\boldsymbol{v}}$ is the map,

$$t_v(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}} + \vec{\boldsymbol{v}} = \begin{bmatrix} x_1 + v_1 \\ \vdots \\ x_n + v_n \end{bmatrix}$$

which can be seen to be isometric (a rigid transformation) by,

$$\begin{aligned} & t_v(\vec{\boldsymbol{x}}) - t_v(\vec{\boldsymbol{y}}) = (\vec{\boldsymbol{x}} + \vec{\boldsymbol{v}}) - (\vec{\boldsymbol{y}} + \vec{\boldsymbol{v}}) = \vec{\boldsymbol{x}} - \vec{\boldsymbol{y}} \\ \implies \quad & |t_v(\vec{\boldsymbol{x}}) - t_v(\vec{\boldsymbol{y}})| = |\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}| \,. \end{aligned}$$

Proposition 120. *Every rigid motion m is the composition of an orthogonal linear operator and a translation. In other words, for some orthogonal matrix A and fixed vector $\vec{\boldsymbol{v}}$, it takes the form,*

$$m(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} + \vec{\boldsymbol{v}}.$$

Proof. Let $\vec{\boldsymbol{v}} = m(\vec{\mathbf{0}})$ and $t_v(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}} + \vec{\boldsymbol{v}}$ with inverse $t_{-v}(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}} - \vec{\boldsymbol{v}}$. Then composing this with m, the resulting transformation,

$$(t_{-v} \circ m)(\vec{\boldsymbol{x}}) = m(\vec{\boldsymbol{x}}) - \vec{\boldsymbol{v}}$$

continues to be isometric — because it is the composition of isometric transformations — and it fixes the origin because $(t_{-v} \circ m)(\vec{\mathbf{0}}) = m(\vec{\mathbf{0}}) - \vec{\boldsymbol{v}} = m(\vec{\mathbf{0}}) - m(\vec{\mathbf{0}}) = \vec{\mathbf{0}}$. It is therefore, by Corollary 36, left multiplication by an orthogonal matrix. So we can represent it as,

$$(t_{-v} \circ m)(\vec{\boldsymbol{x}}) = t_{-v}(m(\vec{\boldsymbol{x}})) = A\vec{\boldsymbol{x}}.$$

Since $t_{-v} = {t_v}^{-1}$ we can apply t_v to both sides of the equation,

$$m(\vec{\boldsymbol{x}}) = t_v(A\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} + \vec{\boldsymbol{v}}.$$

The obtained representation is uniquely determined by m as $\vec{\boldsymbol{v}} = m(\vec{\mathbf{0}})$ is clearly unique and then the translation t_{-v} is uniquely determined by $\vec{\boldsymbol{v}}$ and then $A = (t_{-v} \circ m)$ is unique for a given $\vec{\boldsymbol{v}}$ and m. □

For a rigid motion $m(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}} + \vec{\boldsymbol{v}}$, m is orientation-preserving if the matrix A is orientation-preserving and orientation-reversing if A is orientation-reversing.

Rotation

Rotations preserve distances, angles and orientation and so are rigid motions. Rotations also fix a vector which is known as the axis of rotation. If the axis of rotation contains the origin then they fix the origin and so are linear operators.

Theorem 30. *The rotations of $\mathbb{R}^2$ and $\mathbb{R}^3$ about the origin are the linear operators whose matrices with respect to the standard basis are orthogonal and have determinant 1.*

Proof. A rotation about the origin m involves rotating the standard basis vectors through an angle θ. It is in the definition of this rotation that the image of the standard basis vectors continue to subtend the same angle, $\pi/2$. Therefore, the rotation must preserve angles. It is also part of the definition that the image under rotation is not scaled so the rotation must preserve distances and must be a congruent transformation. Since the axis of rotation passes through the origin the origin is unchanged by this rotation and so these rotations are rigid motions that fix the origin and have the form,

$$m(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}$$

where A is an orthogonal matrix. Additionally, rotations do not change the orientation of an shape and so their matrices have determinant 1. □

The rotation matrices — orthogonal matrices with determinant 1 — form a subgroup of the group O_n of orthogonal matrices called the ***special orthogonal group*** *and denoted SO_n.*

Proposition 121. *Every member of the special orthogonal group $A \in SO_2$ is the matrix of a rotation.*

Proof. Let $A \in SO_2$. Then A is a 2×2 orthogonal matrix with determinant 1. Let $\vec{\boldsymbol{v}}_1$ be the first column of A which, since A is orthogonal, is a unit vector. Now assume that R is the matrix of a rotation whose first column is $\vec{\boldsymbol{v}}_1$ — which is possible because $\vec{\boldsymbol{v}}_1$ is a unit vector so R can be orthogonal. Then the matrix

$$B = R^{-1}A$$

fixes $\vec{\boldsymbol{e_1}}$ and also, as the composition of two orthogonal vectors, is orthogonal. Therefore the second column of B is a unit vector orthogonal to $\vec{\boldsymbol{e_1}}$ which could be $\vec{\boldsymbol{e_2}}$ or $-\vec{\boldsymbol{e_2}}$.
However, R is an orthogonal matrix with determinant 1 and so is a member of SO_2 which means that $R^{-1}A = B$ is also in SO_2.

This, in turn, means that B has determinant 1 which implies that the second column of B is not $-\vec{\boldsymbol{e_2}}$ and is, therefore, $\vec{\boldsymbol{e_2}}$. So, we have obtained the result that $B = I = R^{-1}A$ which implies that $R = A$. □

Rotating $\mathbb{R}^2$ about the origin

Rotating the 2-d plane about the origin means that the axis of rotation is just the origin (so the fixed vector is $\vec{\mathbf{0}}$).
For example:

(52) A rotation ρ_θ of the plane through an angle θ is a linear operator on $\mathbb{R}^2$ whose matrix with respect to the standard basis is

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

We can see that this is a rotation if we take $\vec{\boldsymbol{x}} = (x_1, x_2)^T \in \mathbb{R}^2$ and write it in polar coordinates,

$$\vec{\boldsymbol{x}} = (r, \alpha).$$

So, relating the polar and rectangular coordinates,

$$\vec{\boldsymbol{x}} = (r\cos\alpha, r\sin\alpha)^T.$$

When we left-multiply by R,

$$\begin{aligned} R\vec{\boldsymbol{x}} &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} r\cos\alpha \\ r\sin\alpha \end{bmatrix} \\ &= \begin{bmatrix} r\cos\alpha\cos\theta - r\sin\alpha\sin\theta \\ r\cos\alpha\sin\theta + r\sin\alpha\cos\theta \end{bmatrix} \\ &= \begin{bmatrix} r\cos(\alpha+\theta) \\ r\sin(\alpha+\theta) \end{bmatrix}. \end{aligned}$$

Note that R is orthogonal because

$$\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \cdot \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = -\cos\theta\sin\theta + \sin\theta\cos\theta = 0$$

and $det\, R = 1$,

$$\begin{vmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{vmatrix} = \cos^2\theta + \sin^2\theta = 1.$$

Rotating $\mathbb{R}^3$ about the origin

Definition. *Define ρ as a rotation in $\mathbb{R}^3$ around the origin if:*

(i) ρ is a rigid motion (orientation-preserving Euclidean isometry) that fixes the origin;

(ii) ρ also fixes a nonzero vector $\vec{\boldsymbol{v}}$;

(iii) ρ operates as a rotation on the plane P orthogonal to $\vec{\boldsymbol{v}}$.

Note that this definition could be described as selecting a 2-dimensional subspace of $\mathbb{R}^3$ and performing a 2-dimensional rotation on it as if it were $\mathbb{R}^2$.

Condition (i) implies, by Corollary 36, that ρ is left multiplication by an orthogonal matrix. Condition (ii) states that ρ has an eigenvector $\vec{\boldsymbol{v}}$ with eigenvalue 1. Then, because ρ preserves angles, the plane P referenced in condition (iii) that is orthogonal to the eigenvector $\vec{\boldsymbol{v}}$, must map to a plane that is orthogonal to the map of $\vec{\boldsymbol{v}}$ in the image of ρ. But $\vec{\boldsymbol{v}}$ is fixed by ρ and is unchanged in the image. Also $\vec{\boldsymbol{v}}$ uniquely identifies a plane orthogonal to it. Therefore the plane P is unchanged in the image also. In other words, P is an invariant subspace. So, condition (iii) says that the restriction of ρ to this invariant subspace is a rotation.

For example:

(53) A rotation of $\mathbb{R}^3$ about the origin can be described by a pair $(\vec{\boldsymbol{v}}, \theta)$ consisting of a unit vector $\vec{\boldsymbol{v}}$, a vector of length 1, which lies in the axis of rotation, and a nonzero angle θ, the angle of rotation. The two pairs $(\vec{\boldsymbol{v}}, \theta)$ and $(-\vec{\boldsymbol{v}}, -\theta)$ represent the same rotation. We also consider the identity map to be a rotation, though its axis is indeterminate.
The matrix representing a rotation through the angle θ about the vector $\vec{\boldsymbol{e_1}}$ is obtained easily from the 2×2 rotation matrix. It is

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}.$$

Multiplication by A fixes the first coordinate x_1 of a vector and operates by rotation on $(x_2, x_3)^T$. All rotations of $\mathbb{R}^3$ are linear operators, but their matrices can be fairly complicated.

Proposition 122. *Every element of SO_3 has eigenvalue 1.*

Proof. Let $A \in SO_3$. Then A is an orthogonal 3×3 matrix with determinant equal to 1. Reasoning from orthogonality of A we have,

$$\begin{aligned} & A^T A = I \\ \iff \quad & A^T A - A^T = I - A^T \\ \iff \quad & A^T(A - I) = I - A^T \\ \iff \quad & A^T(A - I) = (I - A)^T. \qquad \text{by Proposition 41} \end{aligned}$$

If we take the determinants of both sides of this equation we obtain,

$$\begin{aligned} & det(A^T) \cdot det(A - I) = det((I - A)^T) \qquad \text{by Proposition 54} \\ \iff \quad & det(A) \cdot det(A - I) = det(I - A) \qquad \text{by Proposition 56} \\ \iff \quad & det(A - I) = det(I - A). \qquad det(A) = 1 \end{aligned}$$

But the dimension of A being 3 implies that

$$det(-A) = (-1)^3 det(A) = -det(A)$$

so that,

$$det(A - I) = det(I - A) \iff det(A - I) = 0.$$

Therefore A has the eigenvalue 1. □

Proposition 123. *The elements of SO_3 are precisely the rotations about the origin of $\mathbb{R}^3$.*

Proof. Let $\rho : \mathbb{R}^3 \longmapsto \mathbb{R}^3$ be defined as $\rho(\vec{\boldsymbol{x}}) = A\vec{\boldsymbol{x}}$ where $A \in SO_3$. Then,

- by Proposition 119 and orthogonality of $A \in SO_3$, left multiplication by A is a rigid motion that fixes the origin. So ρ is isometric and fixes the origin;

- Proposition 122 shows that every $A \in SO_3$ has eigenvalue 1 which implies that ρ fixes a nonzero vector;
- if we let $\vec{\boldsymbol{v}}$ be the nonzero vector fixed by ρ (i.e. its eigenvector with eigenvalue 1), then we can normalize it to find the unit vector parallel to $\vec{\boldsymbol{v}}$, say $\vec{\boldsymbol{u}}_1$. Next we can find two unit vectors orthogonal to $\vec{\boldsymbol{u}}_1$ — say $\vec{\boldsymbol{u}}_2$ and $\vec{\boldsymbol{u}}_3$ — and these must be a basis for the plane orthogonal to $\vec{\boldsymbol{v}}$. Furthermore, if we select $\vec{\boldsymbol{u}}_2$ and $\vec{\boldsymbol{u}}_3$ to be orthogonal to each other than $B = \{\vec{\boldsymbol{u}}_1, \vec{\boldsymbol{u}}_2, \vec{\boldsymbol{u}}_3\}$ is an orthonormal basis of $\mathbb{R}^3$.
  Now if we define $P = [B]^{-1}$ then,

$$A' = P^{-1}AP$$

  is similar to the matrix A and so has the same determinant, 1. Furthermore, because B is an orthonormal basis, the matrices $[B]$, $[B]^{-1} = P$ are orthogonal. Since both P and A are orthogonal, $P^{-1}AP = A'$ is orthogonal also. Since A' is orthogonal and has determinant equal to 1, it is a member of SO_3.
  If we examine the structure of A', we see that the first column of A' is $\vec{\boldsymbol{v}}_1$ — the unit vector in the direction of $\vec{\boldsymbol{v}}$. Since $\vec{\boldsymbol{v}}$ is an eigenvector of ρ with eigenvalue 1, the first column of A' is $\vec{\boldsymbol{e_1}}$ and since A' is orthogonal, the other columns are orthogonal to the first. So the block structure of A' looks like,

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

  where R is a 2×2 matrix.
  We know that the determinant of A' is 1 and this implies that the determinant of R is also 1. Furthermore, R must also be orthogonal and so $R \in SO_2$. So, by Proposition 121, R is a rotation. Therefore R represents a rotation of the plane orthogonal to $\vec{\boldsymbol{v}}$ and this implies that ρ rotates the plane orthogonal to $\vec{\boldsymbol{v}}$ as required.

□

Uniform Scaling

Uniform scaling is a scalar multiple of an orthogonal matrix and is therefore a linear operator which means that it fixes the origin. It also preserves angles but not distances between points.

2.5.6.6 Conformal Transformations

Non-uniform Scaling
Non-uniform scaling, however, its matrix is not a scalar multiple of an orthogonal matrix and, as such, it does not preserve angles. An important example is:

(54) **Mercator projection**:

This is a map projection that was designed so that rhumb lines (lines of constant bearing over the surface of the earth) are straight lines on the map. To achieve this the projection ensures that a square on the surface of the earth presents as a square on the map. Then, modelling the earth as a sphere of radius R, if lines of latitude are horizontal grid lines across the map, then each actual line of latitude with circumference $2\pi R\cos\phi$ where ϕ is the angle of latitude will present on the map as the same length as the equator, which in reality is $2\pi R$. So they appear to be a line of length $\cos\phi$ times longer than they actually are, i.e. they are stretched by $\sec\phi$.
So, the map projection will stretch the width of a square on the surface of the earth by $\sec\phi$ and to maintain it as a square, it is necessary to stretch the height of the square by the same amount, $\sec\phi$. The actual square on the surface of the earth has height (approximately for a small square) $R\Delta\phi$ so, on the map we need a height $\Delta y \propto \Delta\phi\sec\phi$. Therefore, we have,

$$\frac{dy}{d\phi} = \sec\phi \implies y = \ln\left(\tan\phi + \sec\phi\right) + c$$

where c is a constant that we can set to 0. See https://www.math.ubc.ca/~israel/m103/mercator/mercator.html for a description of this derivation. For more information on conformal map projections generally see: Map Projection, York University, Toronto and Conformal Cartographic Representations - University of Barcelona.

Reflection
Reflection usually refers to a Euclidean Isometry (rigid transformation over a Euclidean space) that fixes a hyperplane (so a line in $\mathbb{R}^2$ and a plane in $\mathbb{R}^3$) but does not preserve orientation so is not a rigid motion. However, it may

also refer to a transformation that fixes an affine space of lower dimension than a hyperplane — for example, reflection in a point — in which case it does preserve orientation and is, therefore, a rigid motion (in fact, reflection in the origin in $\mathbb{R}^2$ is equal to rotation by π). Reflection may fix the origin or may not, depending on whether or not the origin is contained in the affine space fixed by the reflection.

2.5.6.7 Non-Rigid non-Conformal Transformations

Shear
Shear neither preserves distances nor angles. It does preserve parallelism though (as do all affine transformations) and it also fixes the origin, so it is a linear operator.

For more in-depth treatment of affine spaces and transformations see:

- *First two lectures of University of Texas - Multivariable Analysis.*
- *https://www.maa.org/sites/default/files/pdf/pubs/books/meg/meg_ch12.pdf*

Chapter 3

Analysis

3.1 Supremum and Infimum

3.1.1 Definitions

Definition. *An upper bound on a set A is a value x such that,*

$$\forall a \in A,\, a \leq x$$

and a lower bound is similarly defined as a value y such that,

$$\forall a \in A,\, a \geq y.$$

A set is said to be ***upper-bounded*** *if there exists some upper-bound on the set and is said to be* ***lower-bounded*** *if there exists some lower bound on the set. If there exists both upper and lower bounds then the set is said to be* ***bounded****.*

Definition. *The* ***supremum*** *of a upper-bounded set A is a value σ_A such that σ_A is an upper bound on A and,*

$$\sigma'_A < \sigma_A \iff \exists a \in A \ s.t. \ a > \sigma'_A$$

which is to say that if $\sigma'_A < \sigma_A$ then σ'_A is not an upper bound on A and, if σ'_A is not an upper bound on A then it must be less than σ_A since σ_A is an upper bound on A.
An alternative, equivalent definition is,

$$\forall \epsilon > 0, \exists a \in A \ s.t. \ a > \sigma_A - \epsilon.$$

Issue Note that there is an apparent paradox here: This second definition implies that

$$\begin{aligned} & \forall \epsilon > 0 \ . \ \exists a \in A \text{ s.t. } a + \epsilon > \sigma_A \\ \iff \quad & \exists a \in A \text{ s.t. } a \geq \sigma_A \end{aligned}$$

which result, when combined with the upper-bound property, gives

$$\begin{aligned} & \exists a \in A \text{ s.t. } (a \geq \sigma_A) \wedge (a \leq \sigma_A) \\ \iff \quad & \exists a \in A \text{ s.t. } a = \sigma_A \end{aligned}$$

which says that there is always an element in the bounded set that is equal to the supremum. This is not correct - the supremum may be in the set or external to it.
The initial implication is not true, however. We cannot infer that $\exists a \in A$ s.t. $a \geq \sigma_A$. This can be seen with another rearrangement,

$$\begin{aligned} & \forall \epsilon > 0 \ . \ \exists a \in A \text{ s.t. } a + \epsilon > \sigma_A \\ \iff \quad & \forall \epsilon > 0 \ . \ \exists a \in A \text{ s.t. } \epsilon > \sigma_A - a \end{aligned}$$

which shows us that for any positive epsilon there needs to be an a close enough to the value of σ_A that the difference in their values is less than epsilon. Since a can approach arbitrarily close to σ_A this is achievable for any positive epsilon. This property seems to be equivalent to the fact that σ_A is a *limit point* of A but that will be covered properly in Topology.

Definition. *The **infimum** of a lower-bounded set A is defined similarly to the supremum: as a value τ_A such that τ_A is a lower bound on A and,*

$$\tau'_A > \tau_A \iff \exists a \in A \ s.t. \ a < \tau'_A$$

or alternatively,

$$\forall \epsilon > 0, \exists a \in A \ s.t. \ a < \tau_A + \epsilon.$$

Notation. The supremum of A is denoted $sup\, A$ and the infimum is denoted $inf\, A$.

3.1.2 Deductions using the supremum and infimum

Proposition 124. *If a bounded set $A \subset \mathbb{R}$ has the property that,*

$$\forall x, y \in A \ . \ |x - y| < 1$$

then it follows that,

$$(sup\, A - inf\, A) \leq 1.$$

Proof. Let $\sigma_A = sup\, A$ and $\tau_A = inf\, A$ and w.l.o.g. assume that $x > y$. By the definitions of the supremum and infimum we have,

$$\begin{aligned} & \forall \epsilon > 0 \ . \ \exists x, y \in A \ . \ (x > \sigma_A - \epsilon) \wedge (y < \tau_A + \epsilon) \\ \iff \quad & \forall \epsilon > 0 \ . \ \exists x, y \in A \ . \ (x > \sigma_A - \epsilon) \wedge (-y > -\tau_A - \epsilon) \\ \iff \quad & \forall \epsilon > 0 \ . \ \exists x, y \in A \ . \ (x - y) > (\sigma_A - \tau_A) - 2\epsilon \end{aligned}$$

Now suppose, for contradiction, that $(\sigma_A - \tau_A) > 1$ then we can say that,

$$\exists r > 0 \ . \ (\sigma_A - \tau_A) = 1 + r.$$

If we then constrict ϵ such that,

$$\epsilon < \frac{r}{2} \iff 2\epsilon < r \iff r - 2\epsilon > 0$$

then the previous result tells us that, for $0 < \epsilon < \frac{r}{2}$,

$$\begin{aligned} & \exists x, y \in A \,.\, (x - y) > (\sigma_A - \tau_A) - 2\epsilon \\ \iff \quad & \exists x, y \in A \,.\, (x - y) > 1 + r - 2\epsilon > 1 \end{aligned}$$

which contradicts the set property that $\forall x, y \in A$, $|x - y| < 1$. So this shows that $(\sigma_A - \tau_A) \leq 1$. □

Proposition 125. *Let $A \subset \mathbb{R}$ be a bounded set and let B be the set defined by*

$$B = \{\, b \mid b = f(a),\ a \in A \,\}$$

where the function f is some strictly monotonic function. Then it follows that,

$$sup\, B = f(sup\, A).$$

Proof. A is bounded and so $\sigma_A = sup\, A$ exists. So, using the supremum properties we have,

$$\begin{aligned} & \forall a \in A \,.\, a \leq \sigma_A \\ \iff \quad & \forall a \in A \,.\, f(a) \leq f(\sigma_A) && \text{by monotonicity of f} \\ \iff \quad & \forall b \in B \,.\, b \leq f(\sigma_A) \end{aligned}$$

which is to say that $\sigma_B = f(\sigma_A)$ is an upper bound on B.
Furthermore, using the other supremum property, we have that,

$$\begin{aligned} & \sigma'_A < \sigma_A \implies \exists a \in A \text{ s.t. } a > \sigma'_A \\ \iff \quad & f(\sigma'_A) < f(\sigma_A) \implies \exists a \in A \text{ s.t. } f(a) > f(\sigma'_A) && \text{by strict monotonicity of f} \\ \iff \quad & \sigma'_B < \sigma_B \implies \exists b \in B \text{ s.t. } b > \sigma'_B. \end{aligned}$$

Therefore σ_B satisfies both requirements of the supremum and we have shown that,

$$sup\, B = f(sup\, A).$$

□

3.2 Limits

3.2.1 Limits of sequences

3.2.1.1 Problems with the informal description of a limit

If we say that a sequence tends to some value L when the terms of the sequence *gets closer and closer to* L we have the following problems:

- that the sequence gets closer and closer to many numbers so that this does not specify a single specific limit.
- that the sequence can have a limit but it's not the case that every term is closer than the previous term to the limit. For example,

  $$a_{2k} = 1/k,\ a_{2k-1} = \frac{1}{k+1}$$

  tends to 0 but $a_{2k} > a_{2k-1}$.

Definition. *A sequence a_n is said to* ***tend*** *to L or have the* ***limit*** *L iff,*

$$\forall \epsilon > 0 \in \mathbb{R},\ \exists N \in \mathbb{N}\ s.t.\ \forall n > N,\ |a_n - L| < \epsilon.$$

Definition. *The interval $(L-\epsilon, L+\epsilon)$ is called the* ***ϵ-neighbourhood of*** *L.*

Definition. *A sequence a_n is said to* ***tend to infinity*** *iff,*

$$\forall M > 0 \in \mathbb{R},\ \exists N \in \mathbb{N}\ s.t.\ \forall n > N,\ a_n > M$$

and ***tend to minus-infinity*** *iff,*

$$\forall M < 0 \in \mathbb{R},\ \exists N \in \mathbb{N}\ s.t.\ \forall n > N,\ a_n < M.$$

Definition. *A sequence that has a limit is called* ***convergent*** *and otherwise is called* ***divergent****. Note that* ***divergent*** *sequences include both sequences that remain bounded but oscillate without converging and those that tend to infinity (or minus-infinity).*

3.2.1.2 Examples of Convergence and Divergence

(55) **Non-convergent Oscillation**

The sequence $a_n = (-1)^n$ is divergent despite always remaining bounded within the interval $[-1, 1]$ as it neither converges to 1 or to -1.

(56) **Limit of an Infinite Recurrence**

Take the sequence given by,

$$a_1 = 1,\ a_{n+1} = \frac{a_n}{2} + \frac{3}{2a_n}\ \ (n \geq 1).$$

Assume there is an equilibrium value, a^*, then

$$\begin{aligned} & a^* = \frac{a^*}{2} + \frac{3}{2a^*} \\ \iff \quad & 2(a^*)^2 = (a^*)^2 + 3 \\ \iff \quad & (a^*)^2 = 3 \\ \iff \quad & a^* = \sqrt{3} \qquad\qquad \forall n, a_n \geq 0 \end{aligned}$$

So $\sqrt{3}$ is the steady-state value that this recurrence converges to as $n \to \infty$. If the recurrence didn't converge then the assumption of an equilibrium value would result in a contradiction. Note, however, that the fact that there is an equilibrium value does *not*, by itself, prove that this sequence converges (although this sequence does).

Proposition 126. *A sequence has at most one limit. In other words, a sequence can only converge, if at all, to a single unique value.*

Proof. Let L and L' both be limits of the sequence a_n, and the constant $\alpha = L - L'$. Then,

$$\forall \epsilon > 0 \in \mathbb{R},\ \exists N \in \mathbb{N} \text{ s.t. } \forall n > N,\ |a_n - L| < \epsilon$$

and

$$\forall \epsilon' > 0 \in \mathbb{R},\ \exists N' \in \mathbb{N} \text{ s.t. } \forall n' > N',\ \left|a_{n'} - L'\right| < \epsilon'.$$

But also we have,

$$\left|a_n - L'\right| = \left|(a_n - L) + (L - L')\right| = \left|(L - L') + (a_n - L)\right|$$

and using the triangle inequality,

$$\begin{aligned}
\left|L - L'\right| = \left|(L - L') + (a_n - L) - (a_n - L)\right| &\leq \left|(L - L') + (a_n - L)\right| + |-(a_n - L)| \\
\iff \qquad \left|L - L'\right| &\leq \left|(L - L') + (a_n - L)\right| + |a_n - L| \\
\iff \qquad \left|L - L'\right| - |a_n - L| &\leq \left|(L - L') + (a_n - L)\right| \\
\iff \qquad |\alpha| - |a_n - L| &\leq |\alpha + (a_n - L)|
\end{aligned}$$

Since $\alpha = L - L'$ is constant we can consider the situation when $\epsilon = \frac{|\alpha|}{2}$ then we have that,

$$\begin{aligned}
\exists N \in \mathbb{N} \text{ s.t. } \forall n > N,\ |a_n - L| < \epsilon = \frac{|\alpha|}{2} \\
\iff \qquad -|a_n - L| > -\frac{|\alpha|}{2} \\
\iff \qquad |\alpha| - |a_n - L| > |\alpha| - \frac{|\alpha|}{2} \\
\iff \qquad |\alpha| - |a_n - L| > \frac{|\alpha|}{2}.
\end{aligned}$$

Combining this with the previous result gives, for $\forall n > N$,

$$\frac{|\alpha|}{2} < |\alpha| - |(a_n - L)| \leq |\alpha + (a_n - L)| = \left|a_n - L'\right|$$

which, by rearranging a little, is,

$$\forall n > N,\ \left|a_n - L'\right| > \frac{|\alpha|}{2}.$$

But this means that if we also choose $\epsilon' = \frac{|\alpha|}{2}$ then there is no N' such that $\forall n' > N'$, $|a_{n'} - L'| < \epsilon'$ which contradicts the hypothesis that L' is also a limit of a_n. □

Proof. Another quicker way of proving the proposition is by letting $\epsilon = \epsilon' = \frac{|\alpha|}{2}$ so that,

$$2\epsilon = |\alpha| = \left|L - L'\right| = \left|L - a_n + a_n - L'\right| \leq |L - a_n| + \left|a_n - L'\right| = |a_n - L| + \left|a_n - L'\right|$$

which gives us,

$$2\epsilon \leq |a_n - L| + \left|a_n - L'\right|.$$

But by the limit definition,

$$|a_n - L| + \left|a_n - L'\right| < \epsilon + \epsilon'$$

and since we have set $\epsilon = \epsilon'$ then,

$$|a_n - L| + \left|a_n - L'\right| < 2\epsilon$$

which contradicts $2\epsilon \leq |a_n - L| + |a_n - L'|$. □

Definition. *If a_n is a sequence and $S = \{\, a_n \mid n \in \mathbb{N} \,\}$ then a_n is said to be **bounded below** if S has a lower bound and **bounded above** if S has an upper bound, and **bounded** if it is bounded above and below.*

Lemma 1. *Any finite set of elements from an ordered field has a minimum and a maximum.*

Proof. This can be proven quite easily using induction. Taking the base case of a set of cardinality one, clearly there is a maximum and a minimum both of which are the sole element of the set. Then, the induction step is to say, given a set S that has a maximum, s_{max}, and a minimum, s_{min}, if we add a new element e, then if e is greater than s_{max} it is the maximum of the new set and if it is less than s_{min} it is the minimum of the new set. Otherwise, the previous maximum and minimum also pertain to the new set. Therefore, adding a new element to a set that has a maximum and a minimum creates a new set with a maximum and a minimum. □

Proposition 127. *Any convergent sequence is bounded.*

Proof. Firstly, we need to prove that any finite sequence is bounded. We can do this simply by observing that any finite set of elements from an ordered field,

$$S = \{a_1, a_2, \ldots, a_n\}$$

has a minimum and a maximum.
Now, let a_n be an arbitrary convergent sequence so that,

$$\forall \epsilon > 0 \,.\, \exists N \in \mathbb{N} \,.\, \forall n > N \,.\, |a_n - L| < \epsilon$$

for some $L \in \mathbb{R}$.

Then, let S_{max} and S_{min} be the maximum and minimum respectively of the first N terms of a_n, $S = \{a_1, a_2, \ldots, a_N\}$, and,

$$\exists \epsilon > 0 \,.\, \forall n > N \,.\, |a_n - L| < \epsilon$$

so that, for $n > N$, the sequence a_n is bounded in the ϵ-neighbourhood of L. So, if we define $m = min\{S_{min}, L - \epsilon\}$ and $M = max\{S_{max}, L + \epsilon\}$, then the whole sequence a_n for all $n \in \mathbb{N}$ is bounded below by m and bounded above by M.
Therefore a_n is bounded. □

Definition. *An **increasing** sequence is a sequence a_n such that,*

$$\forall n \in \mathbb{N} \,.\, a_{n+1} \geq a_n$$

*and **decreasing** if,*

$$\forall n \in \mathbb{N} \,.\, a_{n+1} \leq a_n$$

*and **monotonic** if either increasing or decreasing.*

Proposition 128. *Any increasing sequence that is bounded above has a limit.*

Proof. Let a_n be an increasing sequence that is bounded above. Then,

$$\forall n \in \mathbb{N} \,.\, a_{n+1} \geq a_n$$

and let $S = \{\, a_n \mid n \in \mathbb{N} \,\}$. Since a_n is bounded above it has a supremum. Let $\sigma = sup\ S$ so that,

$$\forall a_n \in S \,.\, a_n \leq \sigma \quad \text{and} \quad \forall \epsilon > 0 \,.\, \exists a_n \in S \,.\, a_n > \sigma - \epsilon.$$

Therefore, for some arbitrary fixed $\epsilon > 0$,

$$\exists a_n \in S \,.\, a_n > \sigma - \epsilon$$

and setting $N = n$ so that $a_N > \sigma - \epsilon$, we have,

$$\forall n > N \in \mathbb{N} \,.\, a_n \geq a_N > \sigma - \epsilon$$

and so, recalling that σ is an upper bound on S,

$$\begin{aligned}
& \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, (a_n > \sigma - \epsilon) \wedge (a_n \leq \sigma) \\
\iff \quad & \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, a_n \leq \sigma < a_n + \epsilon \\
\iff \quad & \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, 0 \leq \sigma - a_n < \epsilon \\
\implies \quad & \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |\sigma - a_n| < \epsilon.
\end{aligned}$$

But ϵ was an arbitrary positive value so,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |\sigma - a_n| < \epsilon$$

and σ is, therefore, the limit of a_n. □

Corollary 37. *Any increasing sequence that is bounded above converges to the supremum of its elements (terms, values, etc.).*

Corollary 38. *A decreasing sequence that is bounded below converges to the infimum of its elements.*

3.2.2 Algebra of limits of sequences

Proposition 129. *Let a_n and b_n be convergent sequences with limits a and b, respectively. Let C be a real number and let k be a positive integer. Then as $n \to \infty$,*

a) $Ca_n \to Ca$

b) $|a_n| \to |a|$

c) $a_n + b_n \to a + b$

d) $a_n b_n \to ab$

e) $a_n^k \to a^k$

f) if, for all n, $b_n \neq 0$ and $b \neq 0$, then $\frac{1}{b_n} \to \frac{1}{b}$.

Proof. We prove each property individually in the given order.

3.2.2.1 Proof of (a) $Ca_n \to Ca$

If $C = 0$ then $Ca_n = 0 = Ca$ for all n and the proposition holds trivially. If $C \neq 0$ then, since $a_n \to a$,

$$\forall \epsilon' > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |a_n - a| < \epsilon'.$$

Now let $\epsilon = |C|\,\epsilon'$. Then,

$$\begin{aligned} & \forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |C|\,|a_n - a| < |C|\,\epsilon' = \epsilon \\ \iff \quad & \forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |Ca_n - Ca| < \epsilon. \qquad |x|\,|y| = |xy| \end{aligned}$$

3.2.2.2 Proof of (b) $|a_n| \to |a|$

$$\begin{aligned} & |a_n| = |a_n - a + a| \leq |a_n - a| + |a| && \text{the "triangle inequality"} \\ \iff & |a_n| - |a| \leq |a_n - a| \,. && (1) \end{aligned}$$

$$\begin{aligned} & |a| = |a - a_n + a_n| \leq |a - a_n| + |a_n| && \text{the "triangle inequality"} \\ \iff & |a| - |a_n| \leq |a - a_n| = |a_n - a| \\ \iff & |a_n| - |a| \geq -|a_n - a| \,. && (2) \end{aligned}$$

Putting (1) and (2) together we have,

$$\begin{aligned} & -|a_n - a| \leq |a_n| - |a| \leq |a_n - a| && \text{the "triangle inequality"} \\ \iff & ||a_n| - |a|| \leq |a_n - a| \,. \end{aligned}$$

The fact that a_n converges to a implies that $|a_n - a|$ converges to zero. Since it is an upper bound on the value of $||a_n| - |a||$, the value $||a_n| - |a||$ must also converge to zero. Specifically any value of n, ϵ such that $|a_n - a| < \epsilon$ will also satisfy $||a_n| - |a|| \leq |a_n - a| < \epsilon$.

3.2.2.3 Proof of (c) $a_n + b_n \to a + b$

Using, again, the "triangle inequality",

$$|(a_n + b_n) - (a + b)| = |(a_n - a) + (b_n - b)| \leq |a_n - a| + |b_n - b| \,.$$

So, $|a_n - a| + |b_n - b|$ is an upper bound on the value of $|(a_n + b_n) - (a + b)|$. If we take any arbitrary $\epsilon > 0$ then,

$$\exists N_1 \,.\, \forall n > N_1 \in \mathbb{N} \,.\; |a_n - a| < \frac{\epsilon}{2}$$

and

$$\exists N_2 \,.\, \forall n > N_2 \in \mathbb{N} \,.\; |b_n - b| < \frac{\epsilon}{2}.$$

Then, if we take $N = max\{N_1, N_2\}$, we have,

$$\forall n > N \in \mathbb{N} \,.\; |(a_n + b_n) - (a + b)| \leq |a_n - a| + |b_n - b| < \epsilon.$$

3.2.2.4 Proof of (d) $a_n b_n \to ab$

$$|a_n b_n - ab| = |a_n b_n - ab_n + ab_n - ab| \leq |b_n(a_n - a)| + |a(b_n - b)| \quad \text{the "triangle inequality"}$$
$$\iff \quad |a_n b_n - ab| \leq |b_n|\,|a_n - a| + |a|\,|b_n - b|$$

Since b_n converges, by Proposition 127, it is bounded. Therefore, $|b_n|$ has some upper bound which we shall call B. Then,

$$\forall \epsilon > 0 \,.\, \exists N_1 \,.\, \forall n > N_1 \in \mathbb{N} \,.\, |a_n - a| < \frac{\epsilon}{2B}$$

and

$$\forall \epsilon > 0 \,.\, \exists N_2 \,.\, \forall n > N_2 \in \mathbb{N} \,.\, |b_n - b| < \frac{\epsilon}{2\,|a|}.$$

Now, let $N = max N_1, N_2$. Then,

$$\forall n > N \in \mathbb{N} \,.\, B\,|a_n - a| + |a|\,|b_n - b| < \epsilon.$$

3.2.2.5 Proof of (e) $a_n^k \to a^k$

Using (d) - and because k is a positive integer - we can do induction on the power k.
Base cases 0 and 1 are clearly true as $k = 0$ results in a_n being the constant 1 for all n and so trivially converges to $a = 1$; and $k = 1$ results in the same sequence as a_n.
So, we perform the induction step for $k \geq 2$. Then, by the induction hypothesis, $a_n^{k-1} \to a^{k-1}$. But $a_n^k = a_n^{k-1} a_n$ and, by (d) and the induction hypothesis, we have that $a_n^{k-1} a_n \to a^{k-1} a = a^k$. Therefore, $a_n^k \to a^k$.

3.2.2.6 Proof of (f) $\forall n \,.\, b_n, b \neq 0 \implies \frac{1}{b_n} \to \frac{1}{b}$

Again invoking Proposition 127 and letting the upper bound on the sequence b_n be B,

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{b_n b}\right| = \left|\frac{b_n - b}{b_n b}\right| = \frac{|b_n - b|}{|b_n|\,|b|} \leq \frac{1}{B\,|b|}\,|b_n - b|\,.$$

Now, since $\frac{1}{B|b|}$ is a constant we can define the constant $C = \frac{1}{B|b|}$ and then we see that in (a) we have already proven that $C\,|b_n - b|$ converges to 0. In (a) we used that to prove that $Cb_n \to Cb$ but here it proves that $\frac{1}{b_n} \to \frac{1}{b}$. $\square$

3.2.3 Some theorems on limits of sequences

Theorem 31. *If* $|a| < 1$ *then* $\lim_{n\to\infty} a^n = 0$.

Proof. First of all, note that if $|a| = 0$ then $a = 0 = a^n$ for all n and so the limit holds trivially. For this reason, from here on, we will consider only the case where $a \neq 0$.
There are 3 parts to this proof:

1. $|a| < 1 \implies \lim_{n\to\infty} |a|^n = 0$,
2. $|a|^n = |a^n|$,
3. $\lim_{n\to\infty} |a^n| = 0 \implies \lim_{n\to\infty} a^n = 0$.

1. $\boldsymbol{|a| < 1 \implies \lim_{n\to\infty} |a|^n = 0}$

   It would be natural to prove this by showing that,

   $$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, \left||a|^n - 0\right| = |a|^n < \epsilon.$$

   We can show this by "reverse engineering" the value of N from the requirement that $|a|^n$ be less than ϵ,

   $$|a|^n < \epsilon \iff n \ln |a| < \ln \epsilon \iff n > \frac{\ln \epsilon}{\ln |a|}$$

   with the last step changing the direction of the inequality because we divide by $\ln |a|$ which - remembering that $|a| < 1$ - is a negative value. So, in this way, we have shown that $N = \frac{\ln \epsilon}{\ln |a|}$ is a general formula that relates a value of N with the required property with any arbitrary ϵ. However, this proof is not valid because it uses the concept of the logarithm which requires a lot of analysis that has not been proven at this stage. Since we are trying to build the fundamental basis of analysis, at this point we can only use concepts that are pre-requisites (axiomatic) in analysis or have been proven at this stage.

   An alternative way to show this, using the properties of limits of sequences just proven, is as follows: Let x_n be the sequence $x_n = |a|^n$. Then, because $0 < |a| < 1$,

   $$x_{n+1} = x_n \cdot x = |a|^n |a| < |a|^n = x_n$$

so that x_n is a decreasing sequence. Additionally, $\forall n \in \mathbb{R} \,.\, |a|^n \geq 0$ so 0 is a lower bound on the sequence. Therefore, the sequence converges to a limit (note we haven't yet established that 0 is the limit - only that it is a candidate). Furthermore, $x_{n+1} = |a|^{n+1} = |a|^n |a|$ and, if $x_n \to L$ then $x_{n+1} \to L$ also. But, putting these two facts together, along with property (d) of limits of sequences, means that,

$$L = \lim_{n\to\infty} |a|^{n+1} = \lim_{n\to\infty} |a|^n |a| = (\lim_{n\to\infty} |a|^n)(\lim_{n\to\infty} |a|) = |a| (\lim_{n\to\infty} |a|^n) = |a| L.$$

So,

$$L = |a| L \iff L(1 - |a|) = 0$$

and, since we know that $|a| \neq 1$, therefore L must be 0.

2. $\boldsymbol{|a|^n = |a^n|}$

   For $a \in \mathbb{R}, n \in \mathbb{N}$ it's easy to see that $|a| |a| \ldots |a| = |aa \ldots a|$. In actual fact, this appears to hold even for $n \in \mathbb{Q}$, e.g.

$$\left|(-1)^{\frac{1}{2}}\right| = |i| = 1 = \left|1^{\frac{1}{2}}\right| = |1| = 1$$

   but this should be checked when studying complex numbers more thoroughly. Also, the base a, can it also be complex?

3. $\boldsymbol{\lim_{n\to\infty} |a^n| = 0 \implies \lim_{n\to\infty} a^n = 0}$

   This can be proved directly from the definition of the limit.

$$\begin{aligned}
\lim_{n\to\infty} |a^n| = 0 &\iff \forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, ||a^n| - 0| < \epsilon \\
&\iff \forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |a^n| < \epsilon \\
&\iff \forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |a^n - 0| < \epsilon \\
&\iff \lim_{n\to\infty} a^n = 0.
\end{aligned}$$

Bear in mind that, in general, $\lim_{n\to\infty} |x_n| = L \not\implies \lim_{n\to\infty} x_n = L$. For example, if x_n converges to $-L$ then $|x_n|$ will converge to L.

```
code example
```

Furthermore, the example 56 showed how the fact of a_n and a_{n+1} converging to the same limit produces - when the sequence is expressed as a recurrence - an equilibrium value. □

3.2.3.1 The Sandwich Theorem

Proposition 130. *Let a_n, b_n, c_n be sequences such that,*

$$\textit{for all } n,\ a_n \leq b_n \leq c_n \quad \textit{and} \quad \lim_{n\to\infty} a_n = L = \lim_{n\to\infty} c_n.$$

Then $\lim_{n\to\infty} b_n = L$.

Proof. $\lim_{n\to\infty} a_n = L$ means that,

$$\begin{aligned} & \forall \epsilon > 0 \,.\, \exists N_1 \,.\, \forall n > N_1 \in \mathbb{N} \,.\, |a_n - L| < \epsilon \\ \iff \quad & \forall \epsilon > 0 \,.\, \exists N_1 \,.\, \forall n > N_1 \in \mathbb{N} \,.\, -\epsilon < a_n - L < \epsilon \\ \iff \quad & \forall \epsilon > 0 \,.\, \exists N_1 \,.\, \forall n > N_1 \in \mathbb{N} \,.\, L - \epsilon < a_n < L + \epsilon. \end{aligned}$$

By the same reasoning we also have,

$$\forall \epsilon > 0 \,.\, \exists N_2 \,.\, \forall n > N_2 \in \mathbb{N} \,.\, L - \epsilon < c_n < L + \epsilon.$$

So, if we let $N = max\{N_1, N_2\}$ then we have,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, L - \epsilon < a_n, c_n < L + \epsilon$$

and since we also know that $a_n \leq b_n \leq c_n$ it follows that,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, L - \epsilon < a_n \leq b_n \leq c_n < L + \epsilon.$$

This shows that,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |b_n - L| < \epsilon.$$

□

3.2.3.2 An example application of the Sandwich Theorem

(57) **Prove that $|x| < 1 \implies \lim_{n\to\infty} x^n = 0$**

We have already proven this using the properties of limits in Theorem 31 but here we are going to prove it using the Sandwich Theorem.

Proof. Firstly, as we showed in 2, $|x^n - 0| = |x^n| = |x|^n$. So to show that x^n tends to zero we can show that $|x|^n$ tends to zero. So, w.l.o.g. we take $x > 0$ (since $x = 0$ makes the proposition trivially true). Then, notice that $0 < x < 1 \implies x = \frac{1}{1+h}$ for some $h > 0$. Then, we can show inductively that $(1 + h)^n \geq 1 + hn$ as follows.

Base cases 0, 1: $(1+h)^0 = 1 = 1 + h(0),\ (1+h)^1 = 1 + h = 1 + h(1).$

Induction step $k > 1$

$$\begin{aligned} & (1+h)^k \geq 1 + hk \\ \iff \quad & (1+h)(1+h)^k \geq (1+h)(1+hk) \\ \iff \quad & (1+h)^{k+1} \geq 1 + hk + h + h^2k = 1 + h(k+1) + h^2k > 1 + h(k+1) \end{aligned}$$

This result implies that,

$$x^n = \frac{1}{(1+h)^n} \leq \frac{1}{1+hn}$$

so that,

$$0 < x^n \leq \frac{1}{1+hn}.$$

Since h is some fixed value, clearly,

$$\lim_{n\to\infty} \frac{1}{1+hn} = 0$$

and, obviously, the limit of the constant 0 is always 0 so, by the Sandwich Theorem,

$$\lim_{n\to\infty} x^n = 0.$$

□

3.2.4 Subsequences

Definition. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence and consider some strictly increasing natural numbers $(k_1, k_2, k_3, \dots)$ that is, $(k_1 < k_2 < k_3 < k_4 < \dots)$. Then the sequence $(a_{k_n})_{n\in\mathbb{N}}$ is called a* ***subsequence*** *of the sequence $(a_n)_{n\in\mathbb{N}}$.*
Note that a ***subsequence*** *is always infinite (I think).*

Theorem 32. *If a_n is a sequence that tends to a limit, then any subsequence of it tends to the same limit.*

Proof. Firstly, notice that if the nth index of some subsequence is k_n then $k_n \geq n$ (because the subsequence can only skip terms of the original - it can't add in terms). So then, if we have a sequence a_n that tends to a limit a and an arbitrary subsequence a_{k_n} then,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |a_n - a| < \epsilon \implies |a_{k_n} - a| < \epsilon$$

because $k_n \geq n > N$. □

Theorem 33. *Every sequence has a monotonic subsequence.*

Proof. Either there is an infinite number of terms that are greater than all the following terms or there is not. If there is not, then after the last such term all terms have a term that follows them that is greater than or equal to them.
In the first case we have a strict monotonic decreasing sequence and in the second we have a non-strict monotonic increasing sequence. □

If we put this theorem together with what we learned about bounded sequences in Proposition 128 and its corollaries - that monotonic bounded sequences are convergent - then we get one of the most famous results in analysis, The Bolzano-Weierstrass Theorem.

Theorem 34. ***The Bolzano-Weierstrass Theorem***
Every bounded sequence has a convergent subsequence.

3.2.5 Examples of limits of sequences

(58) **Let $(a_n)_{n\in\mathbb{N}}$ be a sequence, and let $(b_n)_{n\in\mathbb{N}}$ be the sequence defined by $b_n = |a_n|$ for $n \in \mathbb{N}$. Which of the following two statements implies the other?**

Answer: a_n converges $\implies$ b_n converges also but b_n converges $\not\implies$ a_n converges.
The first implication is because,

$$\begin{aligned} & |a_n| = |a_n - a + a| \leq |a_n - a| + |a| \\ \iff \quad & |a_n| - |a| \leq |a_n - a| \end{aligned}$$

$$\begin{aligned} & |a| = |a - a_n + a_n| \leq |a_n - a| + |a_n| \\ \iff \quad & |a| - |a_n| \leq |a_n - a| \\ \iff \quad & |a_n| - |a| \geq -|a_n - a| \end{aligned}$$

which both together imply that $||a_n| - |a|| \leq |a_n - a|$.
The latter non-implication is easy to see if one thinks of a sequence that consists of two subsequences that converge to 2 and -2. Then, their absolute value would converge to 2 but their values do not converge. Remember Theorem 32, for a sequence to converge to a limit, every subsequence of it must converge to the same limit.

(59) **What is the behaviour as $n \to \infty$ of the following:**

(i) $\frac{2n^3+1}{n+1}\left(\frac{3}{4}\right)^n$

$$0 < \frac{2n^3+1}{n+1}\left(\frac{3}{4}\right)^n < \frac{3n^3}{n}\left(\frac{3}{4}\right)^n = 3n^2\left(\frac{3}{4}\right)^n \to 0$$

(ii) $\frac{2^{2n}+n}{n^3 3^n+1}$

$$\frac{2^{2n}+n}{n^3 3^n+1} = \frac{4^n+n}{n^3 3^n+1} > \frac{4^n}{2n^3 3^n} = \frac{(4/3)^n}{2n^3} \to \infty$$

(60) **Let (a_n) be a sequence of non-negative numbers. Prove that if $a_n \to L$ as $n \to \infty$ then $\sqrt{a_n} \to \sqrt{L}$ as $n \to \infty$.**

Proof. We are told that $a_n \to L$ as $n \to \infty$ so we have,

$$\forall \epsilon' > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |a_n - L| < \epsilon'.$$

We are also told that (a_n) is non-negative so $L \geq 0$.
First, consider the possibility that $L = 0$. Then $|a_n| < \epsilon'$ for $n > N$. But,

$$|a_n| = (\sqrt{a_n})^2 < \epsilon' \iff \sqrt{a_n} < (\epsilon')^2$$

so that taking $\epsilon = (\epsilon')^2$ we obtain,

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, \left|\sqrt{a_n}\right| < \epsilon.$$

The remaining possibility is that $L > 0$. Notice that the expression we're looking to bound can be rewritten as follows:

$$\left|\sqrt{a_n} - \sqrt{L}\right| = \left|(\sqrt{a_n} - \sqrt{L})\frac{\sqrt{a_n} + \sqrt{L}}{\sqrt{a_n} + \sqrt{L}}\right| = \left|\frac{a_n - L}{\sqrt{a_n} + \sqrt{L}}\right|.$$

Furthermore, since $a_n \to L$ and $L > 0$, clearly $0 < L/2 < L$ and there will be some $\epsilon < L/2$ such that,

$$L - L/2 < L - \epsilon < a_n < L + \epsilon < L + L/2 \iff |a_n - L| < \epsilon < L/2.$$

This means that, inside this ϵ-neighbourhood of L, we can find a constant lower bound on $\sqrt{a_n} + \sqrt{L}$ as,

$$\sqrt{a_n} + \sqrt{L} > \sqrt{L/2} + \sqrt{L} = C$$

for constant C. Now we have

$$\left|\sqrt{a_n} - \sqrt{L}\right| = \left|\frac{a_n - L}{\sqrt{a_n} + \sqrt{L}}\right| < \frac{|a_n - L|}{C} < \frac{\epsilon'}{C}$$

so we can take $\epsilon = \epsilon'/C$ to obtain $|a_n - L| < \epsilon$ as required. □

Proof. This is an alternative proof of the $L > 0$ case using the fact that

$$\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}.$$

Choose ϵ such that $0 < \epsilon \leq \sqrt{L}$ and take N so that, for $n > N$, $|a_n - L| < \epsilon^2$, or in other words $L - \epsilon^2 < a_n < L + \epsilon^2$. Then we have

$$\sqrt{L} - \epsilon \leq \sqrt{L - \epsilon^2} < \sqrt{a_n} < \sqrt{L + \epsilon^2} \leq \sqrt{L} + \epsilon$$

which places $\sqrt{a_n}$ in ϵ-neighbourhood of L, so we're done.
Note that we need $\epsilon \leq \sqrt{L}$ for $\sqrt{L} - \epsilon \leq \sqrt{L - \epsilon^2}$ and also – so long as we are doing *real* analysis – this proof is only for the $L > 0$ case because, when $L = 0$, $\sqrt{L - \epsilon^2}$ will be a complex number. □

(61) **Let a_n be a positive decreasing sequence. Show that, if there exist numbers N and α such that**

$$0 < \frac{a_{n+1}}{a_n} < \alpha < 1 \qquad \forall n > N$$

then $a_n \to 0$ as $n \to \infty$. But if, on the other hand, we have

$$0 < \frac{a_{n+1}}{a_n} < 1 \qquad \forall n > N$$

then we cannot conclude that $a_n \to 0$.

The basic principle here is that, if a convergent sequence converges to a non-zero limit, then the ratio of consecutive terms must tend to 1. If the ratio of consecutive terms remains below 1 then the sequence must go to zero.

If $\frac{a_{n+1}}{a_n} < \alpha$ then α is an upper bound on the ratio of consecutive terms so we can deduce that,

$$a_{n+1} < \alpha a_n < a_n \qquad \text{since } \alpha < 1$$

and that, if we let a_N be the value of the sequence at $n = N$ and consider the sequence for $n > N$,

$$a_n < \alpha^{n-N} a_N$$

which is of the form,

$$a_n < \alpha^n C$$

for constant C. So we have bound the values of the sequence below $\alpha^n C$ which clearly goes to 0 as $n \to \infty$. Since we are told that the sequence is positive so that a_n is also bounded below by 0, we can conclude, by Sandwich Theorem, that $a_n \to 0$ as $n \to \infty$. $\square$

An alternative way of showing the same thing is to use the algebra of limits and the fact that, in a convergent sequence, any subsequence must also converge to the same limit (Theorem 32) to deduce that if $a_n \to L$ then we must also have $a_{n+1} \to L$. Then, if $L \neq 0$ we can apply the algebra of limits to obtain,

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \frac{L}{L} = 1$$

which contradicts $\frac{a_{n+1}}{a_n} < \alpha < 1$. Therefore, $L = 0$. $\square$

On the other hand, if $\frac{a_{n+1}}{a_n} \to 1$ as $n \to \infty$ then the sequence may converge to 0 or to some other value. For example:

(i) $\boldsymbol{a_n = \frac{1}{n} + 1}$

$$\frac{a_{n+1}}{a_n} = \frac{n(n+2)}{(n+1)^2} = \frac{n^2+2n}{n^2+2n+1} \to 1 \text{ as } n \to \infty$$

and clearly,

$$\lim_{n\to\infty} \frac{1}{n} + 1 = 1.$$

But also,

(ii) $\boldsymbol{a_n = \frac{1}{n}}$

$$\frac{a_{n+1}}{a_n} = \frac{n}{n+1} \to 1 \text{ as } n \to \infty$$

even though

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

3.2.6 Limits of functions

3.2.6.1 Definition of the limit of a function

Definition. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a function. We say that L is the* ***limit of $f(x)$ as x approaches a*** *if, for each $\epsilon > 0$ there exists $\delta > 0$ such that,*

$$0 < |x - a| < \delta \implies |f(x) - L| < \epsilon.$$

Definition. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a function. We say that* ***$f(x)$ tends to infinity as x approaches a*** *if, for each K there exists $\delta > 0$ such that,*

$$0 < |x - a| < \delta \implies f(x) > K.$$

3.2.6.2 Examples of limits of functions

(62) **Prove that if $f : \mathbb{R} \mapsto \mathbb{R}$ s.t. $f(x) = x^2 + x$ then $\lim_{x \to 2} f(x) = 6$.**

Let $0 < |x - 2| < \delta$ and consider some arbitrary $\epsilon > 0$. Then we have,

$$\left|(x^2 + x) - 6\right| = |(x - 2)(x + 3)| \leq |x - 2|\,|x + 3| < \delta\,|x + 3|\,.$$

It's tempting at this point to say that, since we are examining the behaviour when x approaches 2 so we can assume $x \approx 2 \iff (x + 3) \approx 5$ and then we can set $\delta = \frac{\epsilon}{6}$ so that,

$$0 < |x - 2| < \delta = \frac{\epsilon}{6} \implies |f(x) - 6| < \frac{|x + 3|}{6}\epsilon < \epsilon.$$

However, there is a subtle logical problem here: We have considered any arbitrary $\epsilon > 0$, meaning that ϵ could be large. Then, when we set $\delta = \frac{\epsilon}{6}$ we linked the value of δ to that of ϵ so that δ may also be arbitrarily large. But $0 < |x - 2| < \delta$ so that $|x - 2|$ may be

arbitrarily large also. This contradicts the assumption that $x \approx 2$ and so the argument we have here is only valid for small ϵ.

So, we need an alternative approach. Consider that,

$$|x+3| = |x-2+2+3| \leq |x-2| + |2+3| = |x-2| + 5 < \delta + 5.$$

This means that,

$$\Big|(x^2+x) - 6\Big| < \delta\,|x+3| < \delta(\delta+5).$$

But note, do **not** start trying to solve a quadratic. There is a much better way.

Now - here comes the clever bit - if we set $\delta = min\{1, \frac{\epsilon}{6}\}$ then *both* 1 and $\frac{\epsilon}{6}$ are an upper bound on the value of δ so that we can say,

$$\Big|(x^2+x) - 6\Big| < \delta(\delta+5) \leq 6\delta \leq 6\frac{\epsilon}{6} = \epsilon.$$

3.2.6.3 Algebra of limits of functions

Theorem 35. *Let $f, g : \mathbb{R} \mapsto \mathbb{R}$ be two functions and c be any real number. Suppose that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. Then,*

(i) $\lim_{x\to a}(cf)(x) = cL$

(ii) $\lim_{x\to a}(|f|)(x) = |L|$

(iii) $\lim_{x\to a}(f+g)(x) = L + M$

(iv) $\lim_{x\to a}(f-g)(x) = L - M$

(v) $\lim_{x\to a} f(x)g(x) = LM$

(vi) $\lim_{x\to a}(f/g)(x) = L/M$ *provided* $g(x) \neq 0$ *for any* x *in the neighbourhood of* a

(vii) $\lim_{x\to a} f(x)^c = L^c$.

Proof. see: Lamar University - Paul's Online Notes - Proofs of limit properties
First we need to prove a couple of more basic statements about limits as preliminaries.

- $\boldsymbol{\lim_{x\to a} c = c}$:

  $$\forall \epsilon > 0 \,.\, \exists \delta \,.\, 0 < |x - a| < \delta \implies |c - c| < \epsilon$$

  is clearly satisfied for any interval of x-values (since the constant function doesn't depend on x). So, the proposition is trivially satisfied by any ϵ, δ.

- $\boldsymbol{\lim_{x\to a} x = a}$:

  $$\forall \epsilon > 0 \,.\, \exists \delta \,.\, 0 < |x - a| < \delta \implies |x - a| < \epsilon$$

  is also trivially satisfied for any $\epsilon = \delta$.

(i) $\boldsymbol{\lim_{x\to a}(cf)(x) = cL}$:

Let

$$\epsilon > 0,\ \epsilon_1 = \frac{\epsilon}{|c|}.$$

By hypothesis,

$$\forall \epsilon_1 > 0 \,.\, \exists \delta_1 \,.\, 0 < |x - a| < \delta_1 \implies |f(x) - L| < \epsilon_1.$$

Furthermore,

$$\begin{aligned} & |f(x) - L| < \epsilon_1 \\ \iff \quad & |c|\,|f(x) - L| < |c|\,\epsilon_1 \\ \iff \quad & |cf(x) - cL| < |c|\,\epsilon_1 = |c|\,\frac{\epsilon}{|c|} = \epsilon. \end{aligned}$$

by Proposition 10

(ii) $\boldsymbol{\lim_{x\to a}(|f|)(x) = |L|}$:

Using properties of absolute value (see: 1.1.3),

$$|f(x) - L| \geq \Big||f(x)| - |L|\Big|$$

so that

$$|f(x) - L| < \epsilon \implies \Big||f(x)| - |L|\Big| < \epsilon.$$

This means that the limit hypothesis on $f(x)$ implies that

$$\lim_{x \to a}(|f|)(x) = |L|$$

as required.

(iii) $\boldsymbol{\lim_{x \to a}(f + g)(x) = L + M}$:

Let

$$\epsilon > 0,\ \epsilon_1 = \frac{\epsilon}{2}.$$

By hypothesis we have,

$$\forall \epsilon_1 > 0\ .\ \exists \delta_1\ .\ 0 < |x - a| < \delta_1 \implies |f(x) - L| < \epsilon_1,$$

$$\forall \epsilon_1 > 0\ .\ \exists \delta_2\ .\ 0 < |x - a| < \delta_2 \implies |g(x) - M| < \epsilon_1.$$

Let

$$\delta = \min\{\delta_1, \delta_2\}$$

so that, for $0 < |x - a| < \delta$ we have,

$$|f(x) - L| < \epsilon_1 \quad \text{and} \quad |g(x) - M| < \epsilon_1.$$

If we sum the two expressions and employ the absolute value properties (see: 1.1.3) we obtain,

$$\begin{aligned} & |f(x) - L| + |g(x) - M| < 2\epsilon_1 = \epsilon \\ \implies \quad & |(f(x) - L) + (g(x) - M)| < \epsilon && |x + y| \leq |x| + |y| \\ \iff \quad & |(f(x) + g(x)) - (L + M)| < \epsilon. \end{aligned}$$

(iv) $\boldsymbol{\lim_{x \to a}(f - g)(x) = L - M}$:

Arguing similarly to (iii), again, for x in the δ-neighbourhood of a, we have both

$$|f(x) - L| < \epsilon_1 \quad \text{and} \quad |g(x) - M| < \epsilon_1.$$

This time, we are going to subtract the two expressions and use a fact derived from the absolute value properties (see: 1.1.3) that,

$$||x| - |y|| \leq |x - y| \leq |x| + |y| \,.$$

$$\begin{aligned} & |(f(x) - L) - (g(x) - M)| \leq |f(x) - L| + |g(x) - M| < 2\epsilon_1 \\ \iff \quad & |(f(x) - L) - (g(x) - M)| < 2\epsilon_1 = \epsilon \\ \iff \quad & |(f(x) - g(x)) - (L - M)| < \epsilon. \end{aligned}$$

Note that the maximum "error" is the sum of the individual errors $\epsilon_1 + \epsilon_2$ just as it is for (iii).

(v) $\mathbf{\lim_{x \to a} f(x)g(x) = LM}$:

First, we prove a lemma that will be useful for the second part of this proof.

Lemma 2.

$$\lim_{x \to a} (f(x) - L)(g(x) - M) = 0$$

Proof. Let

$$\epsilon > 0, \ \epsilon_1 = \sqrt{\epsilon}.$$

By hypothesis we have,

$$\forall \epsilon_1 > 0 \ . \ \exists \delta_1 \ . \ 0 < |x - a| < \delta_1 \implies |f(x) - L| < \epsilon_1,$$

$$\forall \epsilon_1 > 0 \ . \ \exists \delta_2 \ . \ 0 < |x - a| < \delta_2 \implies |g(x) - M| < \epsilon_1.$$

Let

$$\delta = \min\{\delta_1, \delta_2\}$$

so that, for $0 < |x - a| < \delta$ we have,

$$|f(x) - L| < \epsilon_1 \quad \text{and} \quad |g(x) - M| < \epsilon_1.$$

Then,

$$|f(x) - L| \, |g(x) - M| < {\epsilon_1}^2 = \epsilon.$$

By Proposition 10,

$$|f(x) - L|\,|g(x) - M| = |(f(x) - L)(g(x) - M)|$$

so we have,

$$\forall \epsilon > 0\ .\ \exists \delta\ .\ 0 < |x - a| < \delta \implies |(f(x) - L)(g(x) - M) - 0| < \epsilon.$$

Which is to say,

$$\lim_{x \to a} (f(x) - L)(g(x) - M) = 0. \qquad \square$$

The second part begins by observing that,

$$f(x)g(x) = (f(x) - L)(g(x) - M) + Mf(x) + Lg(x) - LM.$$

If we take the limit of both sides of this equation as $x \to a$ then,

$$\begin{aligned}
\lim_{x \to a} f(x)g(x) &= \lim_{x \to a} [(f(x) - L)(g(x) - M) \\
&\qquad + Mf(x) + Lg(x) - LM] \\
&= \lim_{x \to a} (f(x) - L)(g(x) - M) \\
&\qquad + M \lim_{x \to a} f(x) + L \lim_{x \to a} g(x) - LM \\
&= 0 + ML + LM - LM = LM.
\end{aligned}$$

(vi) $\boldsymbol{\lim_{x \to a}(f/g)(x) = L/M}$ provided $g(x) \neq 0$ for any x in the neighbourhood of a:

First we prove the following lemma:

Lemma 3.

$$\lim_{x \to a} \frac{1}{g(x)} = \frac{1}{M}$$

Proof. Firstly observe that,

$$\frac{1}{g(x)} - \frac{1}{M} = \frac{M - g(x)}{Mg(x)}$$

and that

$$\begin{aligned} & |M| = |M - g(x) + g(x)| \leq |M - g(x)| + |g(x)| \\ \iff \quad & |g(x)| \geq |M| - |M - g(x)| \\ \iff \quad & \frac{1}{|g(x)|} \leq \frac{1}{|M| - |g(x) - M|}. \end{aligned}$$

By hypothesis there exists some δ_1 such that,

$$0 < |x - a| < \delta_1 \implies |g(x) - M| < \frac{|M|}{2} \implies \frac{1}{|M| - |g(x) - M|} < \frac{2}{|M|}.$$

Also, for any arbitrary $\epsilon > 0$ there exists some δ_2 such that,

$$0 < |x - a| < \delta_1 \implies |g(x) - M| < \frac{|M|^2}{2}\epsilon$$

which means that,

$$\frac{|M - g(x)|}{|Mg(x)|} = |g(x) - M| \cdot \frac{1}{|M|} \cdot \frac{1}{g(x)} < \frac{|M|^2}{2}\epsilon \cdot \frac{1}{|M|} \cdot \frac{2}{|M|} = \epsilon. \quad \square$$

Now we can use (v) to deduce that,

$$\lim_{x \to a} \frac{f(x)}{g(x)} = (\lim_{x \to a} f(x)) \left(\lim_{x \to a} \frac{1}{g(x)} \right) = L \cdot \frac{1}{M} = \frac{L}{M}.$$

(vii) $\boldsymbol{\lim_{x \to a} f(x)^c = L^c}$:

We need to prove this in progressive stages. First for a natural number power, then for any rational number and then for irrational numbers.

Lemma 4.

$$\lim_{x\to a} f(x)^n = L^n \qquad n \in \mathbb{N}$$

Proof. Prove by induction on n. Base case $n = 2$:

$$\begin{aligned}\lim_{x\to a} f(x)^2 &= \lim_{x\to a} f(x)f(x) \\ &= (\lim_{x\to a} f(x))(\lim_{x\to a} f(x)) && \text{by (v)} \\ &= \left[\lim_{x\to a} f(x)\right]^2.\end{aligned}$$

Induction step:

$$\begin{aligned}\lim_{x\to a} f(x)^{n+1} &= \lim_{x\to a} f(x)^n f(x) \\ &= (\lim_{x\to a} f(x)^n)(\lim_{x\to a} f(x)) \\ &= \left[\lim_{x\to a} f(x)\right]^n (\lim_{x\to a} f(x)) && \text{by induction hypothesis} \\ &= \left[\lim_{x\to a} f(x)\right]^{n+1}. && \square\end{aligned}$$

Lemma 5.

$$\lim_{x\to a} f(x)^{1/n} = L^{1/n} \qquad n \in \mathbb{N}$$

Proof.

$$\begin{aligned} & \left[\lim_{x\to a} f(x)^{1/n}\right]^n = \lim_{x\to a}(f(x)^{1/n})^n && \text{by Lemma 4} \\ \iff \quad & \left[\lim_{x\to a} f(x)^{1/n}\right]^n = \lim_{x\to a} f(x) \\ \iff \quad & \lim_{x\to a} f(x)^{1/n} = \left[\lim_{x\to a} f(x)\right]^{1/n}. && \square\end{aligned}$$

Lemma 6.

$$\lim_{x\to a} f(x)^q = L^q \qquad q \in \mathbb{Q}$$

Proof. Firstly consider the case that $q \geq 0$. Since q is a rational number, it can be expressed as m/n for $m, n \in \mathbb{N}$. Then,

$$\lim_{x\to a} f(x)^q = \lim_{x\to a} f(x)^{m/n}$$

$$
\begin{aligned}
&= \lim_{x\to a}(f(x)^{1/n})^m \\
&= \left[\lim_{x\to a} f(x)^{1/n}\right]^m && \text{by Lemma 4} \\
&= \left[\left[\lim_{x\to a} f(x)\right]^{1/n}\right]^m && \text{by Lemma 5} \\
&= \left[\lim_{x\to a} f(x)\right]^{m/n} = \left[\lim_{x\to a} f(x)\right]^q.
\end{aligned}
$$

Now consider the case where $q < 0$. Then we can express q as $-m/n$ for $m, n \in \mathbb{N}$. By (vi),

$$
\begin{aligned}
\lim_{x\to a} f(x)^q &= \lim_{x\to a} f(x)^{-m/n} \\
&= \lim_{x\to a}(f(x)^{-1})^{m/n} \\
&= \lim_{x\to a}(f(x)^{m/n})^{-1} \\
&= \left[\lim_{x\to a} f(x)^{m/n}\right]^{-1} && \text{by (vi)} \\
&= \left[\lim_{x\to a} f(x)\right]^{-q}. && \square
\end{aligned}
$$

What follows is a tentative, speculative proof for irrational powers and will hopefully be confirmed or corrected at a later date.

If we view an arbitrary irrational number r as a Cauchy sequence then it is a convergent sum of an infinite series of rational terms. So, for $q_1, q_2, \cdots \in \mathbb{Q}$

$$x^r = x^{(q_1+q_2+\cdots)} = x^{q_1}x^{q_2}\cdots$$

where

$$\lim_{i\to\infty} q_i = 0 \implies \lim_{i\to\infty} x^{q_i} = 1.$$

Therefore,

$$
\begin{aligned}
\lim_{x\to a} f(x)^r &= \lim_{x\to a} f(x)^{(q_1+q_2+\cdots)} \\
&= \lim_{x\to a} f(x)^{q_1} f(x)^{q_2}\cdots \\
&= \left[\lim_{x\to a} f(x)\right]^{q_1}\left[\lim_{x\to a} f(x)\right]^{q_2}\cdots
\end{aligned}
$$

$$= \left[\lim_{x \to a} f(x)\right]^{(q_1+q_2+\cdots)}$$
$$= \left[\lim_{x \to a} f(x)\right]^{r}.$$

□

3.2.6.4 One-sided limits

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that L is the* ***limit of $f(x)$ as x approaches a from the left (or from below)****, denoted by $\lim_{x \to a^-} f(x) = L$, if for each $\epsilon > 0$, there exists $\delta > 0$ such that,*

$$a - \delta < x < a \implies |f(x) - L| < \epsilon.$$

3.2.6.5 Limits at infinity

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that L is the* ***limit of $f(x)$ as x approaches ∞****, denoted by $\lim_{x \to \infty} f(x) = L$, if for each $\epsilon > 0$, there exists $M > 0$ such that,*

$$x \geq M \implies |f(x) - L| < \epsilon.$$

3.3 Continuity

3.3.1 Continuity

Definition. *A function f is* ***continuous at a point a*** *if*

- $f(a)$ *is defined,*
- $\lim_{x\to a} f(x) = f(a)$.

More formally, the definition of $\lim_{x\to a} f(x) = L$ *is,*

$$\forall \epsilon > 0 \;.\; \exists \delta \text{ s.t. } 0 < |x - a| < \delta \implies |f(x) - L| < \epsilon.$$

But if $f(a)$ is defined, then when $|x - a| = 0$ (i.e. when $x = a$) we have

$$f(x) = f(a) \implies |f(x) - f(a)| = 0 < \epsilon.$$

Therefore, if $f(a)$ is defined and $\lim_{x\to a} f(x) = f(a)$*, then*

$$\forall \epsilon > 0 \;.\; \exists \delta > 0 \text{ s.t. } |x - a| < \delta \implies |f(x) - f(a)| < \epsilon.$$

Definition. *A function is* ***continuous*** *if it is* ***continuous at every point***.

Definition. *A function is* ***continuous on the closed interval $[a, b]$*** *if it is*

- *continuous at every point in the open interval* (a, b),
- $\lim_{x\to a^+} f(x) = f(a)$,
- $\lim_{x\to b^-} f(x) = f(b)$.

Definition. *A function is* ***left continuous*** *or* ***continuous on the left*** *at* a *if,*

$$\lim_{x \to a-} f(x) = f(a).$$

Obviously, from the other side, a function can be ***right continuous****.*

Theorem 36. *Let* $f, g : \mathbb{R} \to \mathbb{R}$ *be functions that are continuous at* $a \in R$ *and* c *be any real number. Then* $|f|, (cf), (f-g), (f+g), (f(x)g(x))$ *are all continuous at* a*, and* (f/g) *is continuous provided* $g(x) \neq 0$ *for any* x *in some neighbourhood of* a*.*

Proof. This follows from the algebra of limits of functions given in Theorem 35. □

Corollary 39. *It follows then that every polynomial is continuous. This can be seen as the most simple polynomial is a constant - which is clearly continuous; also* $f(x) = x$ *is clearly continuous. Then powers of* x *are continuous as they are products of continuous functions and when multiplied by coefficients this is a constant multiplying a continuous function so the resultant function is continuous. Then any polynomial is a summation of such terms so the result is continuous as the sum of continuous functions is continuous.*

Theorem 37. *If* g *is a function which is continuous at* a*, and* f *is a function which is continuous at* $g(a)$*. Then* $(f \circ g)$ *is continuous at* a*.*

Proof. Continuity of f at $g(a)$ guarantees that for any $\epsilon > 0$ there exists some $\delta' > 0$ such that $|x' - g(a)| < \delta' \implies |f(x') - f(g(a))| < \epsilon$ and continuity of g at a guarantees that for any $\delta' > 0$ there exists some $\delta > 0$ such that $|x - a| < \delta \implies |g(x) - g(a)| = |x' - g(a)| < \delta'$. □

Corollary 40. $\lim_{x \to a}(f \circ g)(x) = \lim_{x \to a} f(g(x)) = f(\lim_{x \to a} g(x))$

3.3.1.1 Continuity of functions over sequences

We now give an alternative definition of continuity which ties in the concept of limits for sequences.

Theorem 38. *A function* f *is continuous at* a *if and only if for each sequence* (x_n) *such that* $\lim_{n \to \infty} x_n = a$ *we have* $\lim_{n \to \infty} f(x_n) = f(a)$*.*

Before the proof, an important point to note is that this theorem applies to "each sequence" with the described limit. This is important as it is possible to find an individual sequence such that the inference is not valid. For example, the constant sequence $\forall n \in \mathbb{N} \,.\, x_n = a$ clearly tends to a as $n \to \infty$ but this would not imply continuity of f as the limit of $f(x_n)$ for such a sequence would amount to saying that $f(a) = f(a)$. The definition of continuity is assertion of the equality of the limit of f over values in the neighbourhood of a (but not at a itself) with the value of f at a. So, continuity is implied by the fact that the limit of $f(x_n)$ for the constant sequence $x_n = a$ is equal to the limit of $f(x_n)$ for all other sequences x_n whose limit is a.
Another way of looking at it is that f is continuous at a because the limit there equals $f(a)$ however the argument converges to a.

Proof. Breaking it down into two propositions we have,

$$(\forall x_n \text{ s.t. } \lim_{n\to\infty} x_n = a) \ \lim_{n\to\infty} f(x_n) = f(a) \tag{P_1}$$

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\, |x - a| < \delta \implies |f(x) - f(a)| < \epsilon \tag{P_2}$$

and we need to show that $P_1 \iff P_2$.

So, to begin with we'll assume show that $P_1 \implies P_2$.
Unpacking P_1 we have a function f such that

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\, |f(x_n) - f(a)| < \epsilon \tag{1}$$

where x_n is a sequence such that

$$\forall \delta > 0 \,.\, \exists N' \,.\, \forall n > N' \in \mathbb{N} \,.\, |x_n - a| < \delta. \tag{2}$$

Then, if we choose a particular $\epsilon > 0$, there exists some N such that for all $n > N$ we have $|f(x_n) - f(a)| < \epsilon$. Now there will also be some $N' \leq N$ so that $\forall n > N,\ n > N'$ and so there is some δ such that $|x_n - a| < \delta$. Since P_1 says that this is the case *whenever* we have an x_n such as this, P_1 may be rewritten as

$$\forall \epsilon > 0 \,.\, (\exists N \,.\, \forall n > N \in \mathbb{N}) \,.\, \exists \delta > 0 \,.\, |x_n - a| < \delta \implies |f(x_n) - f(a)| < \epsilon.$$

Now, if we remove reference to the number of the term of the sequences - N, n - we have

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\; |x - a| < \delta \implies |f(x) - f(a)| < \epsilon$$

which is the statement of continuity of f in P_2.

Next we prove $P_2 \implies P_1$.
So now we begin by assuming P_2 which was that,

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\; |x - a| < \delta \implies |f(x) - f(a)| < \epsilon.$$

Actually, it is quite easy to apply a similar logic as previously but in reverse to make the converse implication. We can choose any ϵ and there exists some δ such that P_2 holds. Then, as we have seen previously in (2), P_1 tells us that, for this value of δ,

$$\exists N' \,.\, \forall n > N' \in \mathbb{N} \,.\; |x_n - a| < \delta$$

and P_2 tells us that,

$$|x_n - a| < \delta \implies |f(x_n) - f(a)| < \epsilon.$$

Putting the two together we get,

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\, \exists N' \,.\, \forall n > N' \in \mathbb{N} \,.\; |x_n - a| < \delta \implies |f(x_n) - f(a)| < \epsilon$$

which implies that

$$\forall \epsilon > 0 \,.\, \exists N \,.\, \forall n > N \in \mathbb{N} \,.\; |f(x_n) - f(a)| < \epsilon.$$

So we have shown that P_2 implies that (2) implies (1) which is $P_2 \implies P_1$ as required.

For completeness, let's consider another way of proving that $P_1 \implies P_2$ using a proof by contradiction.
So, we are assuming P_1 but also assuming, for contradiction, that P_2 is false. Then we are negating the statement,

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\; |x - a| < \delta \implies |f(x) - f(a)| < \epsilon$$

so we are asserting that,

$$\forall \epsilon > 0 \,.\, \nexists \delta > 0 \,.\, |x - a| < \delta \implies |f(x) - f(a)| < \epsilon$$

which is equivalent to

$$\forall \epsilon > 0 \,.\, \forall \delta > 0 \,.\, |x - a| < \delta \not\implies |f(x) - f(a)| < \epsilon$$

or alternatively,

$$\forall \epsilon > 0 \,.\, \forall \delta > 0 \,.\, \exists x \,.\, (|x - a| < \delta) \wedge (|f(x) - f(a)| \geq \epsilon).$$

So, this says that we can choose any arbitrary $\epsilon > 0$ and for all $\delta > 0$ there will be some x in the δ-neighbourhood of a such that $|f(x) - f(a)| \geq \epsilon$.

Now, if we choose a value of δ that depends on a natural number n in such a way that $\delta \to 0$ as $n \to \infty$ – for example, if $\delta = 1/n$ – then the δ-neighbourhoods around a will get smaller as $n \to \infty$. Then we can select an x from the δ-neighbourhood that corresponds to a particular value of n and call it x_n and, in this way, we create a sequence x_n that converges to a. So we have $\lim_{n\to\infty} x_n = a$.
But now, for every δ there is an x_n in the δ-neighbourhood of a with $|f(x_n) - f(a)| \geq \epsilon$ and, if we choose this value for x_n, we have a constructed a sequence that converges to a but

$$\lim_{n\to\infty} f(x_n) \neq f(a)$$

which contradicts hypothesis P_1. □

Corollary 41. $\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n)$

3.3.1.2 Continuous Functions on Closed Intervals

Definition. *For a subset X of the domain of a function f, we say that f is bounded on X if there exists M such that $|f(x)| \leq M$ for each $x \in X$.*

Definition. *We define the supremum (or maximum) of f on X as sup $\{ f(x) \mid x \in X \}$ (or max $\{ f(x) \mid x \in X \}$ if it exists).*

3.3.1.3 Examples of bounded and unbounded functions

(63) The indicator function for rational numbers within the reals, known as the **Dirichlet Function**,

$$f(x) = \begin{cases} 0 & x \text{ is irrational} \\ 1 & x \text{ is rational} \end{cases}.$$

This function is nowhere continuous because between every two irrational numbers there is a rational number (and vice-versa) so $f(x)$ is flipping between 0 and 1 in every neighbourhood of every point however small a neighbourhood we consider. So this function never converges anywhere but *is* bounded because it only ever takes values of 1 or 0 so, clearly, its maximum is 1 and minimum is 0.

(64) The function $f(x) = 1/x$ over the interval $(0, 1]$ is continuous at every point but unbounded as it goes to infinity as $x \to 0$. If we consider the same function over the closed interval $[0, 1]$ then we no longer have continuity over this interval as there is a singularity at $x = 0$.

3.3.1.4 Extreme Value Theorem

Theorem 39. *Let f be continous on $[a, b]$. Then f is bounded on $[a, b]$ and it achieves its maximum; that's to say, the supremum is equal to the maximum.*

Note that, even if f is defined at every point in $[a, b]$, if it is not continuous then it may not be bounded. There do exist functions that are not continuous but bounded (for example the Dirichlet Function 63) but there also exist functions that are not continuous and unbounded such as the reciprocal function 64. So, functions that are not continuous on a closed interval may be bounded or not; and functions that are continuous on an open interval also may or may not be bounded (again, the reciprocal function is an example of a function that is continuous on an open interval but not bounded); but here we will show that functions that are continuous on a closed interval must be bounded on that interval.

Proof. It may be tempting to begin trying to prove this by reasoning as follows.

That f is continuous on $[a, b]$ means that, for any $c \in (a, b)$,

$$\forall \epsilon > 0 \,.\, \exists \delta > 0 \,.\, |x - c| < \delta \implies |f(x) - f(c)| < \epsilon.$$

This means that the value of $f(x)$ must be finite everywhere in (a, b) as, choosing any fixed point c in the open interval, $|x - c|$ is finite and, therefore, less than some δ thus implying that $|f(x) - f(c)| < \epsilon$ for some finite $\epsilon \ldots$

However this is **dead wrong!** If we take the example of the reciprocal function (64) over the interval $(0, 1]$: If we take x-values approaching 0, the value of $f(x)$ grows unbounded. For any given x-value it will be less than some finite ϵ but we can always find another x-value with a greater value of $f(x)$. So, there is no maximum ϵ and so, also no maximum value of $f(x)$ on the interval.

Another tempting way to prove this is as follows.

A closed interval is an interval such that every sequence of values in the interval converges to a point in the interval.
That f is continuous on $[a, b]$ implies that for every sequence, x_n, of values in $[a, b]$ that converges to some point in $c \in [a, b]$, $\lim_{n\to\infty} f(x_n) = f(c)$. Furthermore, that the interval is closed implies that every sequence of values in the interval converges to a point in the interval. Therefore, we can conclude that $f(x)$ at every point in $[a, b]$ exists and is finite so that f is bounded and obtains a maximum on the interval.

There are two issues with this:

1. The statement about the nature of a closed interval that the proof relies upon has not been proven.

2. That $f(x)$ is defined and finite for all $x \in [a, b]$ is taken as proof that the function is bounded and obtains a maximum in the interval.

To deal with issue (1) – if we're not going to prove the proposition about closed intervals as a pre-requisite of the proof – we need to develop a proof

that doesn't rely on this characteristic of closed intervals. To deal with (2) meanwhile, we need to explicitly prove boundedness and that f obtains a maximum.
The proof given in LSE Abstract Mathematics course material follows.

> Suppose first that f is unbounded above. For each $n \in N$, let x_n be a point in $[a, b]$ such that $f(x_n) > n$. The sequence (x_n) is bounded, so has a convergent subsequence (x_{k_n}), tending to some limit c (by Theorem 10.11). Necessarily $c \in [a, b]$. Since f is continuous at c, $f(x_{k_n}) \to f(c)$ as $n \to \infty$. But this contradicts the construction of the sequence (x_n), since $f(x_{k_n}) > n \to \infty$. So f is bounded above. Let $M = \sup\{\, f(x) \mid x \in [a, b] \,\}$. For each $n \in N$, let x_n be a point in $[a, b]$ such that $f(x_n) > M - \frac{1}{n}$. Again take a convergent subsequence (x_{k_n}) of (x_n), tending to some limit $c \in [a, b]$. Arguing as before, we see $f(c) = M$.

This proof says: Assume that f is unbounded on the interval. Then we can construct a sequence of x-values such that, for the nth value x_n, $f(x_n) > n$. This is possible because, if f is unbounded, for any value of n, there is some subinterval of x-values such that for each of them $f(x) > n$ and any interval of the real numbers contains an infinite number of real numbers and so, a sequence x_n with $n \to \infty$. Note that, at this point, x_n is an arbitrary sequence which is not necessarily convergent (it could bounce around the interval).
Then, we notice that this sequence x_n is necessarily bounded (as it is a subinterval of $[a, b]$) and so we invoke the Bolzano-Weierstrass Theorem, Theorem 34 (called Theorem 10.11 in the quoted proof), to deduce that it has a convergent subsequence which we call x_{k_n} and call its limit c.
At this point we can use the continuity of f on the interval to deduce that $f(x_{k_n}) \to f(c)$ as $n \to \infty$ by which we obtain a contradiction to the condition we set on the values of x_{k_n} when we constructed the sequence – namely that $f(x_{k_n}) > n$. Notice that this is constructing a sequence x_{k_n} such that $f(x_{k_n})$ grows without bound as $n \to \infty$ and then saying, "but the limit of x_{k_n} as $n \to \infty$ is c which is inside the interval and so (by continuity) the limit of $f(x_{k_n})$ as $n \to \infty$ is $f(c)$ – a fixed finite value". This is the point where the fact that the interval $[a, b]$ is closed comes into play – if the interval were not closed it would be possible that c was not inside the interval and then we would not be able to invoke continuity to assert that the limit of $f(x_{k_n})$ over this sequence was finite.

So, now we have shown that if a function is continuous over a closed interval then assuming that the function is unbounded produces a contradiction and, therefore, we can conclude that it is, in fact, bounded.
The last thing that needs to be proven is that f obtains its maximum in the interval. Having shown that the function is bounded on the interval we now know that there exists

$$M = \sup\{\, f(x) \mid x \in [a, b] \,\}$$

and we need to show that there is such an x-value in $[a, b]$ that $f(x) = M$. In an open interval this might not be the case as the supremum of the function on the interval might occur as the limit of f over a sequence of x-values converging to a point that lies outside the interval. So, in this case, we construct a convergent sequence such that $f(x_{k_n}) \to M$ as $n \to \infty$ by selecting x_n such that $f(x_n) = M - \frac{1}{n}$. This is possible because the definition of the supremum says that, because it is the *lowest* upper bound,

$$\forall \epsilon > 0 \,.\, \exists f(x_n) \,.\, f(x_n) > M - \epsilon$$

and so we are letting $\epsilon = \frac{1}{n}$. Then, as previously, we take a convergent subsequence of this sequence and name it x_{k_n}. So, as before, we have a sequence converging on some point, we'll call it $c \in [a, b]$, at which f is continuous so that $f(x_{k_n}) \to M$ as $n \to \infty$ implies that $M = f(c)$. This, in turn, means that f obtains a maximum in the interval. □

3.3.1.5 Intermediate Value Theorem

Theorem 40. *Let f be continuous on $[a, b]$ with $f(a) < f(b)$. Then, for all K s.t. $f(a) < K < f(b)$, there exists some $c \in (a, b)$ with $f(c) = K$.*

> *Note that this theorem is **not** written for an interval such that $f(a) \leq f(b)$ because if $f(a) = f(b)$ then the only K s.t. $f(a) \leq K \leq f(b)$ is $f(a) = K = f(b)$. But now it is **not** true to say that there exists some $c \in (a, b)$ with $f(c) = K$ as there is no reason why the value of $f(x)$ at the bounds of the interval should be repeated somewhere in the interior of the interval.*

We begin by proving a special case from which the general proof will follow.

Lemma 7. *Let f be continuous on $[a, b]$ with $f(a) < 0 < f(b)$. Then there exists some $c \in (a, b)$ with $f(c) = 0$.*

Proof. A first attempt at this proof is given below.

Continuity at the interval bounds a and b means that, for some $\epsilon_1, \epsilon_2 > 0$,

$$\exists \delta_1 \,.\, 0 \leq x - a < \delta_1 \implies \big|f(x) - f(a)\big| < \epsilon_1,$$

$$\exists \delta_2 \,.\, 0 \leq b - x < \delta_2 \implies \big|f(x) - f(b)\big| < \epsilon_2.$$

So, we have a lower neighbourhood around $f(a)$ and an upper neighbourhood around $f(b)$ as follows,

$$f(a) - \epsilon_1 < f(x) < f(a) + \epsilon_1,$$

$$f(b) - \epsilon_2 < f(x) < f(b) + \epsilon_2$$

If $\epsilon_1 > \big|f(a)\big|$ and $\epsilon_2 > \big|f(b)\big|$ then both neighbourhoods contain $f(x) = 0$. Therefore, there is some interval of x such that $f(x)$ lies inside both the lower and upper neighbourhood – in the overlap of the two. In the overlap we have,

$$f(b) - \epsilon_2 < f(x) < f(a) + \epsilon_1.$$

Now if we let ϵ_1 vary freely but make ϵ_2 a function of ϵ_1 like so,

$$\epsilon_2 = f(a) + f(b) + \epsilon_1$$

then we still have $\epsilon_1 > \big|f(a)\big|$ and $\epsilon_2 > \big|f(b)\big|$ as,

$$\epsilon_1 > \big|f(a)\big| \iff -\epsilon_1 < f(a) < \epsilon_1 \iff 0 < f(a) + \epsilon_1 < 2\epsilon_1$$

so $\epsilon_2 > f(b)$ and, conversely, if we assume that $\epsilon_2 > \big|f(b)\big|$ then,

$$\epsilon_2 > \big|f(b)\big| \iff -\epsilon_2 < f(b) < \epsilon_2 \iff -\epsilon_2 - f(b) < 0 < \epsilon_2 - f(b)$$

$$\epsilon_1 = \epsilon_2 - f(a) - f(b) > -f(a) = \big|f(a)\big| \qquad \text{since } f(a) < 0.$$

Therefore,

$$\epsilon_2 = f(a) + f(b) + \epsilon_1 \implies [\epsilon_1 > \big|f(a)\big| \iff \epsilon_2 > \big|f(b)\big|].$$

Now we have,

$$f(b) - \epsilon_2 = -f(a) - \epsilon_1 < f(x) < f(a) + \epsilon_1$$

which is equivalent to

$$\big|f(x)\big| < f(a) + \epsilon_1 = \epsilon_3.$$

Now, since $f(a) + \epsilon_1 > 0$ we also have $\epsilon_3 > 0$ and it can become arbitrarily small by making $\big|f(a)\big| - \epsilon_1$ arbitrarily small. So we have shown that continuity over the closed interval and $f(a) < f(b)$ imply that we can find subintervals of $[a, b]$ such that $f(x)$ is constricted to ever-decreasing neighbourhoods of 0. In other words, for some c s.t. $a < c < b$, $f(x) \to 0$ as $x \to c$ and since f is continuous on the interval this implies that $f(c) = 0$ also.

This is not bad but suffers from vagueness in a couple of areas: the overlap of the lower and upper neighbourhoods probably needs to be more precisely defined and, certainly, the final part of the proof stating that confining $f(x)$ to ever-decreasing neighbourhoods of 0 implies that there is some c such that $f(c) = 0$ needs to be drawn much more explicitly.

Here is the proof given in LSE Abstract Mathematics.

We construct a sequence of intervals $[a_n, b_n]$ such that

1. $f(a_n) < 0,\ f(b_n) > 0$ for each n
2. $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ for each n.

We start by letting $[a_1, b_1] = [a, b]$. Then for each $n \geq 1$, we define $[a_{n+1}, b_{n+1}]$ as follows.

Let $c_n = (a_n + b_n)/2$, be the midpoint of the previous interval. If $f(c_n) = 0$, then the theorem is proved and so we need not continue constructing intervals!

Otherwise, if $f(c_n) < 0$, we define $a_{n+1} = c_n$ and $b_{n+1} = b_n$. And if $f(c_n) > 0$, we define $b_{n+1} = c_n$ and $a_{n+1} = a_n$. Note that the condition 1. is satisfied by choosing our intervals in this manner. Moreover, note

that the $(n+1)$st interval is half the size of the nth interval and so $b_{n+1} - a_{n+1} \leq (b_1 - a_1)/2^n$. It follows that

$$\lim_{n\to\infty}(b_n - a_n) = 0. \tag{3}$$

Finally, note that (a_n) is increasing and bounded above (by b_1) and so it has a limit; similarly (b_n) is decreasing and bounded below and so has a limit. Thus by (3) (and algebra of limits) these limits are equal to, say, c. Thus by continuity (using Theorem 38),

$$f(c) = \lim_{n\to\infty} f(b_n) \geq 0,$$

where the last inequality follows from the fact that each $f(b_n) \geq 0$ (in fact > 0). Similarly,

$$f(c) = \lim_{n\to\infty} f(a_n) \leq 0.$$

Thus $f(c)$ must be equal to zero, and the proof is complete.

□

Clearly, the general proof of the Intermediate Value Theorem follows naturally from this because,

- If $f(a) > f(b)$ then we can consider the function $g(x) = -f(x)$ so that we have $g(a) < g(b)$,
- If we have K s.t. $f(a) < K < f(b)$ with $K \neq 0$ we can consider $g(x) = f(x) - K$ so that we have $g(a) < 0 < g(b)$.

So, the general problem posed in the Intermediate Value Theorem is reducible to the case we have proved.

Corollary 42. *Suppose that the real function f is continuous on the closed interval $[a, b]$ and that f maps $[a, b]$ into $[a, b]$. Then there is $c \in [a, b]$ with $f(c) = c$.*

Proof. Let $h(x) = f(x) - x$ so that $f(c) = c$ if and only if $h(c) = 0$. Then also we have,

$$a \leq f(x) \leq b \implies h(a) \geq 0,\ h(b) \leq 0.$$

But this means that, either one of $h(a)$ or $h(b)$ is equal to 0 or neither are. In the case that one of them is equal to 0 then we have found our c such that $f(c) = c$. Otherwise, if neither is equal to 0, then we must have $h(a) > 0$ and $h(b) < 0$. So, we may apply the Intermediate Value Theorem to conclude that there exists $c \in (a, b)$ such that $h(c) = 0$ which is to say $f(c) = c$. □

3.3.1.6 Examples of reasoning with the Intermediate Value Theorem

(65) *Suppose the real function f is continuous, positive and unbounded on $\mathbb{R}$ and that $inf\{\, f(x) \mid x \in \mathbb{R} \,\} = 0$. Use the Intermediate Value Theorem to prove that the range of f is $(0, \infty)$.*

Let $y \in (0, 1)$. We show that there is some $c \in \mathbb{R}$ such that $f(c) = y$. This shows that the range is the whole of $(0, 1)$. (The fact that it is no larger follows from the given fact that f is positive.)
Now, $\inf f(\mathbb{R}) = \inf\{\, f(x) \mid x \in \mathbb{R} \,\} = 0$, so, since $y > 0$, there must be some $y_1 \in f(\mathbb{R})$ with $y_1 < y$. This means there is some $x_1 \in \mathbb{R}$ such that $y_1 = f(x_1) < y$.
Similarly, because f is unbounded, which means $f(\mathbb{R})$ is unbounded, there must be some $y_2 \in f(\mathbb{R})$ with $y_2 > y$ and there will be some $x_2 \in \mathbb{R}$ such that $y_2 = f(x_2) > y$.
Then y lies between $f(x_1)$ and $f(x_2)$ and, since f is continuous, the Intermediate Value Theorem shows that there is some c between x_1 and x_2 with $f(c) = y$.

(66) *Suppose the real function g is continuous on $\mathbb{R}$ and that g maps $[a, b]$ into $[d, e]$ and maps $[d, e]$ into $[a, b]$ where $a < b,\ d < e$. By considering the function*

$$k(x) = g(g(x)),$$

prove that there are $p, q \in \mathbb{R}$ such that

$$g(p) = q,\ g(q) = p.$$

Hence show that there is $c \in \mathbb{R}$ such that $g(c) = c$.

The function k, being a composition of continuous functions,

is continuous and also maps $[a, b]$ into $[a, b]$ so that we can use Corollary 42 to deduce that there exists $c \in [a, b]$ such that $k(c) = c$. If we let $p = c$ and $q = g(p)$ then we have,

$$k(p) = p \iff g(g(p)) = p \iff g(q) = p.$$

Now we can employ the same trick again by defining

$$h(x) = g(x) - x$$

and then we have

$$h(p) = g(p) - p = q - p,\ h(q) = g(q) - q = p - q$$

so that $h(p) = -h(q)$ and, therefore, $h(x)$ changes sign between p and q. (Note that we don't know which of p and q is the lower end and upper end of the interval but we know that between the two values the function changes sign.) Then, applying the Intermediate Value Theorem we have some c between p and q such that,

$$h(c) = 0 \iff g(c) - c = 0 \iff g(c) = c.$$

3.3.2 Relationship Between Sequences and Functions

Let $f : \mathbb{R} \mapsto \mathbb{R},\ f(x) = y$. Now imagine that we take regular intervals on the domain, say, interval 1. We name the x-values at the upper bound of these intervals $x_1, x_2, \ldots$ for $x = 1, 2, \ldots$. Then, the corresponding y-values, $y = f(x_1), f(x_2), \ldots$ can be named $y_1, y_2, \ldots$. In this way, we have defined a sequence, $y_n = f(x_n)$ where $x_n \in \mathbb{N}$ (note that this is a different notation from that used before where a sequence $x_n = f(n)$ for $n \in \mathbb{N}$).
Looking at the derivative of the function,

$$\frac{dy}{dx} \approxeq \frac{\Delta y}{\Delta x} = \frac{f(x_{n+1}) - f(x_n)}{1} = f(x_{n+1}) - f(x_n).$$

While the ratio of consecutive terms is,

$$\frac{y_{n+1}}{y_n} = \frac{y_n + \Delta y}{y_n} = \frac{f(x_{n+1})}{f(x_n)} = \frac{f(x_n) + (f(x_{n+1}) - f(x_n))}{f(x_n)}.$$

Note, also, that

$$\frac{y_n + \Delta y}{y_n} = 1 + \frac{\Delta y}{y_n} \approxeq 1 + \frac{dy/dx}{y} = 1 + \frac{d}{dx} \ln y$$

which is to say that $\frac{\Delta y}{y_n}$ is the discrete form of the log derivative. Clearly, when $\Delta y > 0$, we can put this in the form

$$\frac{y_{n+1}}{y_n} = 1 + \frac{\Delta y}{y_n} = 1 + \frac{\Delta y / y_n}{1} = \frac{1 + h}{1}.$$

If $\Delta y < 0$ however,

$$\frac{y_n + \Delta y}{y_n} = \frac{1}{y_n/(y_n + \Delta y)} = \frac{1}{\frac{y_n + \Delta y - \Delta y}{y_n + \Delta y}} = \frac{1}{1 + \frac{-\Delta y}{y_n + \Delta y}} = \frac{1}{1 + h}.$$

If $\frac{\Delta y}{y_n}$ does not go to 0 then the ratio of consecutive terms stays below 1 and the sequence converges to 0. If it does go to 0 then the ratio of consecutive terms goes to 1 and the sequence may converge to 0 or to some non-zero value. These cases appear (proof?) to be distinguishable by looking at how fast $\frac{\Delta y}{y_n}$ goes to 0. For example,

(i) $\boldsymbol{a_n = \frac{1}{n} + 1}$

$$\lim_{n\to\infty} \frac{1}{n} + 1 = 1$$

$$\frac{\Delta a}{a_n} = \frac{-1}{(n+1)^2}$$

(ii) $\boldsymbol{a_n = \frac{1}{n}}$

$$\lim_{n\to\infty} \frac{1}{n} = 0$$

$$\frac{\Delta a}{a_n} = \frac{-1}{n+1}$$

Maybe the fact that $\frac{\Delta y}{y_n}$ goes to 0 faster as $n \to \infty$ in the first case indicates that it converges before it gets to 0?

Chapter 4

Calculus

4.1 Taylor Series

Derivation of Taylor's Theorem

Rolle's Theorem

Taken from https://en.wikipedia.org/wiki/Rolle%27s_theorem#Standard_version_of_the_theorem

If a real-valued function f is continuous on a closed interval $[a, b]$ and differentiable on the open interval (a, b) and $f(a) = f(b)$, then there exists at least one $c \in (a, b)$ such that $f'(c) = 0$.

Mean Value Theorem

Based on https://en.wikipedia.org/wiki/Mean_value_theorem#Proof

If a real-valued function f is continuous on a closed interval $[a, b]$ and differentiable on the open interval (a, b) and $f(a) \neq f(b)$, then we can define a number $M \in \mathbb{R}$ such that,

$$f(b) = f(a) + M(b - a)$$

then let $g(x) = f(x) - f(a) - M(x-a)$ so that $g'(x) = f'(x) - M$. Now, since by the definition of M, $g(a) = g(b) = 0$, we can apply Rolle's theorem so that,

$$\begin{aligned} g'(c) &= 0 \text{ for some } c \in (a,b) \\ \implies 0 &= f'(c) - M \\ \iff M &= f'(c) \\ \therefore f(b) &= f(a) + f'(c)(b-a) \text{ for some } c \in (a,b) \end{aligned}$$

Taylor's Theorem

Taken from *Walter Rudin, Principles of Mathematical Analysis.*

Suppose f is a real-valued function on $[a,b]$, n, is a positive integer, $f^(n-1)$ is continuous on $[a,b]$, $f^(n)$ exists for every $t \in (a,b)$. Let α, β be distinct points of [a, b], and define

$$P(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(t-\alpha)^k.$$

Then there exists a point between α and β such that

$$f(\beta) = P(\beta) + \frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n.$$

Note that if $n = 0$ this degenerates to the Mean Value Theorem:

$$\begin{aligned} f(\beta) &= \sum_{k=0}^{0} \frac{f^{(k)}(\alpha)}{k!}(t-\alpha)^k + \frac{f^{(1)}(x)}{1!}(\beta-\alpha)^1 \\ &= \frac{f^{(0)}(\alpha)}{0!}(t-\alpha)^0 + \frac{f^{(1)}(x)}{1!}(\beta-\alpha)^1 \\ &= f(\alpha) + f'(x)(\beta-\alpha) \end{aligned}$$

*

Proof Let M be the number defined by

$$f(\beta) = P(\beta) + M(\beta-\alpha)^n$$

and put

$$g(t) = f(t) - P(t) - M(t-\alpha)^n \qquad (a \le t \le b).$$

We have to show that $n!M = f^{(n)}(x)$ for some x between α and β. Taking the nth derivative of $g(t)$,

$$g^{(n)}(t) = f^{(n)}(t) - n!M \qquad (a < t < b).$$

The proof will be complete if we can show that $g^{(n)}(x) = 0$ for some x between α and β. Since $P^{(k)}(\alpha) = f^{(k)}(\alpha)$ for $k = 0, \ldots, n-1$ we have

$$g(\alpha) = g'(\alpha) = \ldots = g^{(n-1)}(\alpha) = 0.$$

Our choice of M shows that $g(\beta) = 0$, so that $g'(x_1) = 0$ for some $x_1 \in (\alpha, \beta)$ by the Mean Value Theorem. Since $g'(\alpha) = 0$ we conclude similarly that $g''(x_2) = 0$ for some $x_2 \in (\alpha, \beta)$. After n steps we arrive at the conclusion that $g^{(n)}(x_n) = 0$ for some $x_n \in (\alpha, x_{n-1})$, that is, between α and β.

Examples of Taylor Series

*

$\cos x$ for x close to 0 Note: Taylor series at zero are also called Maclaurin series.

$$\begin{aligned}\cos x &= \cos 0 + (-\sin 0)x + \frac{(-\cos 0)}{2!}x^2 + \ldots \\ &= 1 - \frac{x^2}{2} + \ldots\end{aligned}$$

*

$\cos 2h$ for h close to 0

$$\cos 2h \approx 1 - \frac{(2h)^2}{2} = 1 - 2h^2$$

Note that if we choose to differentiate wrt. h rather than (2h) then we get the same result,

$$\begin{aligned}\cos 2h &= \cos 0 + (-2\sin 0)h + \frac{(-4\cos 0)}{2!}h^2 + \ldots \\ &= 1 - 2h^2 + \ldots\end{aligned}$$

Finite Maclaurin Series and the Binomial Theorem

4.2 The Number e

Definition. *The* ***natural logarithm*** *is defined as,*

$$\ln x = \int_1^x \frac{1}{t}\,\mathrm{d}t.$$

Definition. *The number* $\boldsymbol{e}$ *can be defined in various ways. Some of the most common of these are:*

(i)

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

(ii)

$$e = \sum_{n=0}^{\infty}\frac{1}{n!} = \frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\cdots$$

(iii) e *is the unique number such that,*

$$\ln e = 1$$

Proposition 131. *The definition of the natural log implies:*

(i) $\ln 1 = 0$*;*

(ii) $\frac{\mathrm{d}}{\mathrm{d}x}\ln x = 1/x$*;*

(iii) $\ln x = \ln y \implies x = y$.

Proof. The proofs of each of the properties are the following.

(i) $\mathbf{\ln 1 = 0:}$

By the properties of integrals,

$$\ln 1 = \int_1^1 \frac{1}{t}\,\mathrm{d}t = 0.$$

(ii) $\mathbf{\frac{d}{dx}\ln x = 1/x:}$

This is a consequence of the Fundamental Theorem of Calculus and,

$$\ln x = \int_1^x \frac{1}{t}\,\mathrm{d}t.$$

(iii) $\mathbf{\ln x = \ln y \implies x = y:}$

This is a consequence of the previous property that,

$$\frac{\mathrm{d}}{\mathrm{d}x}\ln x = \frac{1}{x}.$$

For $x > 0$, the function $1/x$ is strictly positive. This, in turn, means that the function,

$$f(x) = \int_1^x \frac{1}{t}\,\mathrm{d}t$$

is strictly monotonically increasing for all $x > 0$. Therefore,

$$x > y \implies f(x) > f(y)$$

and so,

$$x \neq y \implies f(x) \neq f(y).$$ □

Proposition 132. *Let* $f : \mathbb{R} \longmapsto \mathbb{R}$ *be defined as*

$$f(x) = e^x.$$

Then the function f *is the inverse of the natural log function* ln.

Proof. Using the definition of the natural log and properties of integrals we have,

$$\begin{aligned}
\ln e^x &= \int_1^{e^x} \frac{1}{t}\,\mathrm{d}t \\
&= \int_1^{e} \frac{1}{t}\,\mathrm{d}t + \int_e^{e^2} \frac{1}{t}\,\mathrm{d}t + \cdots + \int_{e^{x-1}}^{e^x} \frac{1}{t}\,\mathrm{d}t \\
&= \sum_{i=1}^{x} \int_{e^{i-1}}^{e^i} \frac{1}{t}\,\mathrm{d}t \\
&= \sum_{i=1}^{x} \int_1^{e} \frac{e^{i-1}}{t}\,\mathrm{d}u = \sum_{i=1}^{x} \int_1^{e} \frac{1}{u}\,\mathrm{d}u && u = t/(e^{i-1}),\ e^{i-1}\,\mathrm{d}u = \mathrm{d}t \\
&= \sum_{i=1}^{x} \ln e = \sum_{i=1}^{x} 1 = x.
\end{aligned}$$

Conversely, using similar logic,

$$\begin{aligned}
\ln e^{\ln x} &= \int_1^{e^{\ln x}} \frac{1}{t}\,\mathrm{d}t \\
&= \sum_{i=1}^{\ln x} \ln e = \sum_{i=1}^{\ln x} 1 = \ln x.
\end{aligned}$$

Then, by the injectivity of the natural log (property (iii) of Proposition 131),

$$\ln e^{\ln x} = \ln x \implies e^{\ln x} = x.$$

So, we have shown that,

$$\ln e^x = x = e^{\ln x}$$

which implies that the functions are inverses. □

Proposition 133. *For any $x, r \in \mathbb{R}$,*

$$\ln x^r = r \ln x.$$

Proof. By Proposition 132, the functions e^x and $\ln x$ are inverses. Therefore,

$$\ln x^r = \ln (e^{\ln x})^r = \ln e^{r \ln x} = r \ln x. \qquad \square$$

Proposition 134. *Definitions (i) and (ii) of the number e are equivalent. That's to say,*

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n \iff e = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \cdots$$

Proof. Consider, for finite n, the binomial expansion,

$$\begin{aligned}
\left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{1}{n^k} \\
&= \frac{1}{0!} + \frac{n}{1!}\left(\frac{1}{n}\right) + \frac{(n)(n-1)}{2!}\left(\frac{1}{n^2}\right) + \frac{(n)(n-1)(n-2)}{3!}\left(\frac{1}{n^3}\right) + \\
&\qquad \cdots + \frac{(n)(n-1)\cdots(1)}{n!}\left(\frac{1}{n^n}\right) \\
&= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!}\left(\frac{n-1}{n}\right) + \frac{1}{3!}\left(\frac{(n-1)(n-2)}{n^2}\right) + \\
&\qquad \cdots + \frac{1}{n!}\left(\frac{(n-1)(n-2)\cdots(1)}{n^{n-1}}\right) \\
&= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \\
&\qquad \cdots + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right)
\end{aligned}$$

which tends, as $n \to \infty$, to

$$\frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots \qquad \square$$

The actual proof of this requires limit inferior and superior (see wikipedia) because we haven't shown that the expression derived from the binomial expression converges. The full proof can be found in Artin[73] and wikipedia.

Proposition 135. *Definitions (i) and (iii) of the number e are equivalent. That's to say,*

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \iff e \text{ is the unique number such that } \ln e = 1.$$

Proof. Assume that,

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n.$$

Then, taking logs of both sides,

$$\begin{aligned}
\ln e &= \ln \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \\
&= \lim_{n\to\infty} \ln\left(1+\frac{1}{n}\right)^n && \text{by Theorem 37} \\
&= \lim_{n\to\infty} \frac{\ln\left(1+\frac{1}{n}\right)}{1/n} && \text{using Proposition 133} \\
&= \lim_{n\to\infty} \frac{\ln\left(1+\frac{1}{n}\right) - \ln 1}{1/n} \\
&= \lim_{h\to 0} \frac{\ln(1+h) - \ln 1}{h} \\
&= \left.\frac{\mathrm{d}(\ln x)}{\mathrm{d}x}\right|_{x=1} = \left.\frac{1}{x}\right|_{x=1} = 1.
\end{aligned}$$

This shows that

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \implies \ln e = 1.$$

This number is unique by property (iii) of the natural log in Proposition 131.

Conversely, if we assume that e is the unique number such that $\ln e = 1$ then the fact already shown, that

$$\ln \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = 1$$

implies that

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e. \qquad \square$$

4.2.0.1 e^x

Taking the limit definition of e (definition (i)) and raising it to the power of x, by (vi) of Theorem 35, we have

$$e^x = \left[\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n\right]^x = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^{xn}.$$

If we expand this out using binomial theorem we get,

$$\begin{aligned}
\left(1 + \frac{1}{n}\right)^{xn} &= \sum_{k=0}^{xn} \binom{xn}{k} \frac{1}{n^k} \\
&= 1 + (xn)\left(\frac{1}{n}\right) + \frac{(xn)(xn-1)}{2!}\left(\frac{1}{n^2}\right) + \frac{(xn)(xn-1)(xn-2)}{3!}\left(\frac{1}{n^3}\right) + \cdots \\
&= 1 + x + \frac{1}{2!}\left(\frac{x(xn-1)}{n}\right) + \frac{1}{3!}\left(\frac{x(xn-1)(xn-2)}{n^2}\right) + \cdots \\
&= 1 + x + \frac{x}{2!}\left(x - \frac{1}{n}\right) + \frac{x}{3!}\left(x - \frac{1}{n}\right)\left(x - \frac{2}{n}\right) + \cdots
\end{aligned}$$

If we take the limit of this expression as $n \to \infty$ then we get,

$$e^x = \frac{1}{0!} + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

But also, if we take the expression (which arises if we infinitely compound interest at an interest rate of x),

$$\left(1 + \frac{x}{n}\right)^n$$

and again expand it using binomial theorem,

$$\begin{aligned}
\left(1 + \frac{x}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{x^k}{n^k} \\
&= 1 + \frac{nx}{n} + \frac{n(n-1)}{2!}\frac{x^2}{n^2} + \frac{n(n-1)(n-2)}{3!}\frac{x^3}{n^3} + \cdots \\
&= 1 + x + \frac{x^2}{2!}\frac{n-1}{n} + \frac{x^3}{3!}\frac{(n-1)(n-2)}{n^2} + \cdots \\
&= 1 + x + \frac{x^2}{2!}\left(1 - \frac{1}{n}\right) + \frac{x^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots
\end{aligned}$$

again if we take the limit of this expression as $n \to \infty$ then we get,

$$\frac{1}{0!} + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = e^x.$$

4.3 Differentiation

4.3.1 Differentiation as a Linear Transformation

Notation. The derivative of $y(x)$ with respect to x will be denoted $\mathrm{D}_x\, y$ and of $f(x)$ with $\mathrm{D}_x\, f$.

Definition. *The set $P_n \subset \mathbb{R}^{\mathbb{R}}$ of all univariate real-valued polynomials $p : \mathbb{R} \longmapsto \mathbb{R}$ of degree n is defined as,*

$$P_n = \{\, p \in \mathbb{R}^{\mathbb{R}} \mid p(x) = \sum_{i=0}^{n} \theta_i x^i, \quad \theta_i \in \mathbb{R} \,\}.$$

Or alternatively, in vector notation,

$$P_n = \{\, p \in \mathbb{R}^{\mathbb{R}} \mid p(x) = \vec{\boldsymbol{\theta}}^T \vec{\boldsymbol{x}}, \quad \vec{\boldsymbol{\theta}} \in \mathbb{R}^{n+1} \,\}$$

where $\vec{\boldsymbol{x}}$ is the standard basis of the space of degree-n polynomials,

$$(1, x, \ldots, x^n)^T.$$

4.3.1.1 Linear Algebra of First-order Differential Equations

First Order

$$\frac{dy}{dx} = ax + b$$

$$\begin{aligned}
&\iff && \int \frac{dy}{dx}\,\mathrm{d}x = \int ax + b\,\mathrm{d}x \\
&\iff && y = a'x^2 + bx + c. && a' = a/2,\ \text{c is any constant}
\end{aligned}$$

Integrating we see that a first order differential equation only determines the function upto a constant value. In order to determine a specific function we need a relation between a value of x and a value of y (i.e. a point, in graphical terms). Typically this is described as an initial condition.

Second Order

$$\begin{aligned}
&&& \frac{d^2y}{dx^2} = ax + b \\
&\iff && \int \frac{d^2y}{dx^2}\,\mathrm{d}x = \int ax + b\,\mathrm{d}x \\
&\iff && \frac{dy}{dx} = a'x^2 + bx + c && a' = a/2,\ \text{c is any constant} \\
&\iff && \int \frac{dy}{dx}\,\mathrm{d}x = \int a'x^2 + bx + c\,\mathrm{d}x \\
&\iff && y = a''x^3 + b'x^2 + cx + d. && a'' = a/6,\ b' = b/2,\ \text{d is any constant}
\end{aligned}$$

Integrating a second order equation twice we see that we introduced two constants of integration, c, d, and the last two terms $cx + d$ are undetermined. So a second order equation of this type has only determined a function upto a first-degree polynomial (a line). To determine a specific function, in this case, we require two relations between x and y (two points determine a line).

The derivative $\mathrm{D}_x\, p(x)$, of the univariate degree-n polynomial $p \in P_n$ can be described as a linear transformation as,

$$\mathrm{D}_x\, p(x) = (A\vec{\boldsymbol{\theta}})^T \vec{\boldsymbol{x}} = \vec{\boldsymbol{x}}^T A\vec{\boldsymbol{\theta}}$$

where A is the $n \times (n+1)$ matrix,

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & n \end{bmatrix}.$$

The matrix A clearly has rank n and a 1-dimensional kernel so D_x transforms from $n+1$-space to n-space. The nullspace of A being,

$$nullspace(A) = t \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad t \in \mathbb{R}$$

the interpretation of which is that the derivative of constant polynomials is zero.

In general the differential equation,

$$\mathrm{D}_x\, y(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1} x^{n-1} \tag{4.1}$$

has the vector form,

$$\vec{\boldsymbol{x}}^T A \vec{\boldsymbol{\theta}} = \vec{\boldsymbol{x}}^T \vec{\boldsymbol{\alpha}} \iff A\vec{\boldsymbol{\theta}} = \vec{\boldsymbol{\alpha}} \tag{4.2}$$

where the solution is

$$y(x) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n = \vec{\boldsymbol{x}}^T \vec{\boldsymbol{\theta}}.$$

Note that we can say

$$\vec{\boldsymbol{x}}^T A \vec{\boldsymbol{\theta}} = \vec{\boldsymbol{x}}^T \vec{\boldsymbol{\alpha}} \iff A\vec{\boldsymbol{\theta}} = \vec{\boldsymbol{\alpha}}$$

because $\vec{\boldsymbol{x}}^T$ is a basis and therefore a linearly independent set. By linear independence, the equation on the left implies that the coefficients are equal.

This is equivalent to analysing P_n, the $(n+1)$-dimensional vector space of polynomials, by working in the coordinate space formed by the coefficients. This space is $\mathbb{R}^{n+1}$ which is isomorphic to any $(n+1)$-dimensional vector space by Proposition 83.

The matrix form of Equation 4.2 is,

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 2 & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & n-1 & 0 \\ 0 & 0 & 0 & \cdots & \cdots & n \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \vdots \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \end{bmatrix}. \tag{4.3}$$

Constructing the augmented matrix and using Gaussian Elimination we obtain,

$$\left[\begin{array}{cccccc|c} 0 & 1 & 0 & \cdots & \cdots & 0 & \alpha_0 \\ 0 & 0 & 2 & \cdots & \cdots & 0 & \alpha_1 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & \cdots & n-1 & 0 & \alpha_{n-2} \\ 0 & 0 & 0 & \cdots & \cdots & n & \alpha_{n-1} \end{array}\right]$$

$$\rightsquigarrow \left[\begin{array}{cccccc|c} 0 & 1 & 0 & \cdots & \cdots & 0 & \alpha_0 \\ 0 & 0 & 1 & \cdots & \cdots & 0 & \alpha_1/2 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & \alpha_{n-2}/n-1 \\ 0 & 0 & 0 & \cdots & \cdots & 1 & \alpha_{n-1}/n \end{array}\right]$$

so that θ_0 is a free variable and we can read off the other values of the coefficients θ_i corresponding to the columns of the matrix as follows:

$$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{n-1} \\ \theta_n \end{bmatrix} = \begin{bmatrix} t \\ \alpha_0 \\ \alpha_1/2 \\ \vdots \\ \alpha_{n-2}/n-1 \\ \alpha_{n-1}/n \end{bmatrix} \qquad \text{for } t \in \mathbb{R} \tag{4.4}$$

and this implies that the solution to the differential equation is:

$$y(x) = t + \alpha_0 x + \frac{\alpha_1}{2}x^2 + \cdots + \frac{\alpha_{n-1}}{n}x^n, \qquad t \in \mathbb{R}. \tag{4.5}$$

We can rewrite the result in Equation 4.4 as,

$$\begin{bmatrix} 0 \\ \alpha_0 \\ \alpha_1/2 \\ \vdots \\ \alpha_{n-2}/n-1 \\ \alpha_{n-1}/n \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \qquad \text{for } t \in \mathbb{R} \tag{4.6}$$

which makes it clear that we have a particular solution (with $t = 0$) plus a 1-dimensional nullspace. However, in practical applications modeled by differential equations, there will typically be an initial condition — an initial value of y such as $y(0)$ for example — and this will be used to determine a particular solution. In this situation (known as IVP or Initial Value Problems) the particular solution involves finding a value of the parameter t that fits the initial condition. For this reason, the solution in Equation 4.10 is referred to as the general solution of the differential equation while the particular solution has a particular value of the parameter t.

4.3.1.2 Linear Algebra of Second-order Differential Equations

The most simple type of second-order linear differential equation looks like,

$$\mathrm{D}_x^2\, y(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-2} x^{n-2}. \tag{4.7}$$

We can add an extra row of zeros to the matrix of D_x to make it square so that we can take powers of it,

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & n \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

and output an extra coefficient with value zero.

Then the vector form of Equation 4.7 is

$$A^2\vec{\boldsymbol{\theta}} = \vec{\boldsymbol{\alpha}} \tag{4.8}$$

and the matrix form is

$$\begin{bmatrix} 0 & 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 6 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & n(n-1) \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \vdots \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-2} \\ 0 \\ 0 \end{bmatrix} \tag{4.9}$$

which gives the solution to the differential equation as:

$$y(x) = s + tx + \frac{\alpha_0}{2}x^2 + \frac{\alpha_1}{6}x^3 + \cdots + \frac{\alpha_{n-2}}{n(n-1)}x^n, \qquad s, t \in \mathbb{R}. \tag{4.10}$$

So, here the kernel is 2-dimensional and we need two initial values to determine a particular solution.

4.3.1.3 Eigenvectors of the Differentiation Operator

If we had the differential equation,

$$\mathrm{D}_x\, y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots \tag{4.11}$$

so that the derivative of $y(x)$ is an infinite power series (not an infinite polynomial as polynomials are by definition finite), then the general solution would take the form, for $t \in \mathbb{R}$,

$$y(x) = t + \alpha_0 x + \frac{\alpha_1}{2}x^2 + \frac{\alpha_2}{3}x^3 + \cdots \tag{4.12}$$

So, if we had

$$\alpha_0 = \alpha_1,\ \frac{\alpha_1}{2} = \alpha_2,\ \frac{\alpha_2}{3} = \alpha_3$$

then the general solution would be, for $t \in \mathbb{R}$,

$$y(x) = t + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots \tag{4.13}$$

which is a family of functions which includes the derivative $\mathrm{D}_x\, y(x)$ — the derivative being the member of this set of functions with $t = \alpha_0$.

Note that, for polynomials, the derivative is a transformation between finite-dimensional vector spaces and so the Dimension Formula (Theorem 25) of linear transformations applies. We can see this in that the derivative has a one-dimensional kernel (the set of constant polynomials) and the derivative maps from P_n to P_{n-1}, the image having one less dimension then the domain space because there is a one-dimensional kernel.
However, the coordinate space of the power series above is isomorphic to an infinite-dimensional vector space where the Dimension Formula no longer applies (see Theorem 23). So when we differentiate it we get a result of the same dimensionality despite there being a non-trivial one-dimensional kernel.

In order to achieve this we need the coefficients to form the series,

$$\alpha_0,\ \frac{\alpha_0}{1},\ \frac{\alpha_0}{1\times 2},\ \frac{\alpha_0}{1\times 2\times 3},\ \dots$$

so that

$$\begin{aligned} y(x) &= \alpha_0 + \frac{\alpha_0}{1}x + \frac{\alpha_0}{1\times 2}x^2 + \frac{\alpha_0}{1\times 2\times 3}x^3 + \dots \\ &= \alpha_0 + \frac{\alpha_0}{1!}x + \frac{\alpha_0}{2!}x^2 + \frac{\alpha_0}{3!}x^3 + \dots \\ &= \alpha_0\left(1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots\right) = \alpha_0 e^x. \end{aligned}$$

So, the general solution is,

$$y(x) = te^x \qquad t \in \mathbb{R}$$

and the particular solution, for $t = \alpha_0$, is

$$y(x) = \alpha_0 e^x.$$

Theorem 41. *The derivative of $e^{f(x)}$ is $f'(x)e^{f(x)}$ where f' is the derivative of the function f.*

Proof.

$$e^{f(x)} = 1 + \frac{f(x)}{1!} + \frac{(f(x))^2}{2!} + \frac{(f(x))^3}{3!} + \cdots$$

$$\implies \quad \frac{\mathrm{d}}{\mathrm{d}x} e^{f(x)} = f'(x) + f'(x)\frac{f(x)}{1!} + f'(x)\frac{(f(x))^2}{2!} + f'(x)\frac{(f(x))^3}{3!} + \cdots$$

$$= f'(x)e^{f(x)}.$$

□

Corollary 43. *Any function of the form $Ae^{f(x)}$ is an eigenfunction of the differentiation operator with eigenvalue equal to $f'(x)$.*

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}x} Ae^{f(x)} = A\frac{\mathrm{d}}{\mathrm{d}x} e^{f(x)} = Af'(x)e^{f(x)} = f'(x)(Ae^{f(x)}).$$ □

Corollary 44. *The set of functions $e^{\int f'(x)\,\mathrm{d}x}$ form an eigenspace of the differentiation operator with eigenvalue $f'(x)$.*

Proposition 136. *For any $a, x \in \mathbb{R}$,*

$$\frac{\mathrm{d}}{\mathrm{d}x} a^x = (\ln a)a^x.$$

Proof.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} a^x &= \frac{\mathrm{d}}{\mathrm{d}x} e^{(\ln a)x} && \text{by Proposition 132} \\
&= (\ln a)e^{(\ln a)x} && \text{by Theorem 41} \\
&= (\ln a)a^x && \text{by Proposition 132}
\end{aligned}$$

□

4.3.1.4 Problems with this approach to Differentiation

We can describe polynomial functions as finite vectors defined against the basis of monomials $(1, x, x^2, \dots)$ but to describe transcendental functions such as e^x in the same manner we would need an infinite linear combination of monomials which cannot be generally defined to create a vector space of such vectors.

4.4 Homogeneous Functions

Definition. *A* ***homogeneous*** *function is a multivariate function* $f(x_1, \ldots, x_n)$ *such that,*

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^d f(x_1, \ldots, x_n) \qquad d \in \mathbb{Z},\ \lambda \in \mathbb{R}.$$

The integer power d *is known as the* ***degree*** *so that* f *is described as* ***homogeneous of degree*** d.

Lemma 8. *If a function* f *is homogeneous of degree* d *then it can be expressed as a polynomial whose terms contain powers of variables such that the powers sum to* d.

Proof. Assume f can be expressed as a multivariate polynomial. Then $f(x_1, \ldots, x_n)$ can be expressed as the sum of terms of the form,

$$\alpha x_1^{i_1} \cdots x_n^{i_n}$$

for some constant $\alpha \in \mathbb{R}$. Therefore, for each term of $f(\lambda x_1, \ldots, \lambda x_n)$ we have,

$$\alpha(\lambda x_1)^{i_1} \cdots (\lambda x_n)^{i_n} = \alpha\lambda^{i_1} x_1^{i_1} \cdots \lambda^{i_n} x_n^{i_n} = \lambda^{(i_1+\cdots+i_n)}(\alpha x_1^{i_1} \cdots x_n^{i_n}).$$

Since f is homogeneous of degree d we also have,

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^d f(x_1, \ldots, x_n)$$

where each term of $\lambda^d f(x_1, \ldots, x_n)$ has the form,

$$\lambda^d(\alpha x_1^{i_1} \cdots x_n^{i_n})$$

which means that, for every term of the expression for $f(\lambda x_1, \ldots, \lambda x_n)$, we must have,

$$\lambda^{(i_1+\cdots+i_n)} = \lambda^d \iff i_1 + \cdots + i_n = d. \qquad \square$$

Proposition 137. ***Euler's Theorem of Homogeneous Functions****: If* $f(x_1, \ldots, x_n)$ *is a homogeneous function of degree* d *then,*

$$d \cdot f(x_1, \ldots, x_n) = x_1 \frac{\partial f}{\partial x_1} + \cdots + x_n \frac{\partial f}{\partial x_n}.$$

Proof. Assume f can be expressed as a multivariate polynomial. Then $f(x_1, \ldots, x_n)$ can be expressed as the sum of terms of the form,

$$\alpha x_1^{i_1} \cdots x_n^{i_n}$$

for some constant $\alpha \in \mathbb{R}$. If we take the partial derivative of f with respect to x_1, each term of the result will take the form,

$$i_1 \alpha x_1^{(i_1-1)} x_2^{i_2} \cdots x_n^{i_n}$$

and each term of the partial derivative with respect to x_2 will have the form,

$$i_2 \alpha x_1^{i_1} x_2^{(i_2-1)} \cdots x_n^{i_n}$$

and the n-th partial derivative will have terms,

$$i_n \alpha x_1^{i_1} x_2^{i_2} \cdots x_n^{(i_n-1)}.$$

Now if we look at the terms of the expression $x_1 f_{x_1}$,

$$x_1 \cdot i_1 \alpha x_1^{(i_1-1)} x_2^{i_2} \cdots x_n^{i_n} = i_1(\alpha x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n})$$

we see that the terms are the same as the terms of the original function $f(x_1, \ldots, x_n)$ except multiplied by the power of x_1 in that term. Therefore, each term of the expression $x_1 f_{x_1} + \cdots + x_n f_{x_n}$ has the form,

$$(i_1 + i_2 + \cdots + i_n)(\alpha x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}) = d(\alpha x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n})$$

where we have used Lemma 8 to determine that,

$$i_1 + \cdots + i_n = d.$$

Therefore,

$$x_1 f_{x_1} + \cdots + x_n f_{x_n} = d \cdot f(x_1, \ldots, x_n). \qquad \square$$

4.5 Integration

4.5.1 Univariate Integration

4.5.1.1 The Riemann Integral

Definition. *In the context of the Riemann Integral, a* ***partition*** *of an interval* $[a, b] \in \mathbb{R}$ *is a set,*

$$P = \{\, x_i \mid 0 \leq i \leq n \,\} \quad \textit{where} \quad a \leq x_i < x_{i+1} \leq b.$$

That's to say,

$$P = \{x_0, x_1, \ldots, x_n\} \quad \textit{where} \quad a = x_0 < x_1 < \cdots < x_n = b.$$

Definition. *The* ***lower estimate*** *of the area under the curve of a function* $f(x)$ *with respect to a particular partition is defined as,*

$$L(P) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) \min_{x_i \leq x \leq x_{i+1}} f(x)$$

where each $x_i \in P$*. The* ***upper estimate*** *is similarly defined as,*

$$U(P) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) \max_{x_i \leq x \leq x_{i+1}} f(x).$$

Definition. *Let $P(n)$ be a partition of cardinality $n+1$ over the interval $[a,b]$. If, for a function $f(x)$,*

$$\lim_{n\to\infty} L(P(n)) = \lim_{n\to\infty} U(P(n)) = I$$

then the ***Riemann Integral*** *of $f(x)$ over $P(n)$ is defined as,*

$$\int_a^b f(x)\,\mathrm{d}x = I.$$

Example

(67) Suppose $f(x) = e^x$ and we define a partition over the interval $[0,1]$,

$$P(n) = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}.$$

Then the lower estimate of the area under the curve of this function is given by,

$$\begin{aligned} L(P(n)) &= \sum_{i=0}^{n-1} (x_{i+1} - x_i) \min_{x_i \le x \le x_{i+1}} f(x) \\ &= \frac{1}{n} \sum_{i=0}^{n-1} \min_{\frac{i}{n} \le x \le \frac{i+1}{n}} e^x \\ &= \frac{1}{n} \sum_{i=0}^{n-1} e^{\frac{i}{n}} = \frac{1}{n}(1 + e^{\frac{1}{n}} + \cdots + e^{\frac{n-1}{n}}) \\ &= \frac{1}{n}\left(\frac{e-1}{e^{\frac{1}{n}} - 1}\right). \end{aligned}$$

Meanwhile, the upper estimate is given by,

$$U(P(n)) = \frac{1}{n} \sum_{i=0}^{n-1} \max_{\frac{i}{n} \le x \le \frac{i+1}{n}} e^x$$

$$= \frac{1}{n}\sum_{i=0}^{n-1} e^{\frac{i+1}{n}} = \frac{1}{n}(e^{\frac{1}{n}} + \cdots + e^{\frac{n-1}{n}} + e)$$

$$= \frac{e^{\frac{1}{n}}}{n}\left(\frac{e-1}{e^{\frac{1}{n}}-1}\right).$$

Taking the limits as $n \to \infty$,

$$\begin{aligned}
\lim_{n\to\infty} L(P(n)) &= \lim_{n\to\infty} \frac{1}{n}\left(\frac{e-1}{e^{\frac{1}{n}}-1}\right) \\
&= (e-1)\lim_{n\to\infty} \frac{1}{n}\left(\frac{1}{e^{\frac{1}{n}}-1}\right) \\
&= (e-1)\lim_{n\to\infty} \frac{1/n}{e^{\frac{1}{n}}-1} \\
&= (e-1)\lim_{n\to\infty} \frac{-1/n^2}{(-1/n^2)e^{\frac{1}{n}}} \qquad \text{by L'Hôpital} \\
&= (e-1)\lim_{n\to\infty} e^{-\frac{1}{n}} = (e-1)(1) = e-1,
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty} U(P(n)) &= \lim_{n\to\infty} \frac{e^{\frac{1}{n}}}{n}\left(\frac{e-1}{e^{\frac{1}{n}}-1}\right) \\
&= (e-1)\lim_{n\to\infty} \frac{e^{\frac{1}{n}}}{n}\left(\frac{1}{e^{\frac{1}{n}}-1}\right) \\
&= (e-1)\left(\lim_{n\to\infty} e^{\frac{1}{n}}\right)\left(\lim_{n\to\infty} \frac{1}{n}\left[\frac{1}{e^{\frac{1}{n}}-1}\right)\right] \\
&= (e-1)(1)(1) = e-1.
\end{aligned}$$

So, we see that

$$\lim_{n\to\infty} L(P(n)) = \lim_{n\to\infty} U(P(n)) = e-1$$

and this is the value of the riemann integral for e^x over the interval $[0, 1]$. Note that it is in agreement with the analytic solution,

$$\begin{aligned}\int_0^1 e^x &= [e^x]_0^1 \\ &= e^1 - e^0 = e - 1.\end{aligned}$$

4.5.1.2 FTC and the Chain Rule

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int_{p(t)}^{q(t)} f(x)\,\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^{q(t)} f(x)\,\mathrm{d}x - \frac{\mathrm{d}}{\mathrm{d}t}\int_0^{p(t)} f(x)\,\mathrm{d}x \\ &= \frac{\mathrm{d}q}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}q}\int_0^{q} f(x)\,\mathrm{d}x - \frac{\mathrm{d}p}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}p}\int_0^{p} f(x)\,\mathrm{d}x \\ &= \frac{\mathrm{d}q}{\mathrm{d}t}f(q) - \frac{\mathrm{d}p}{\mathrm{d}t}f(p).\end{aligned}$$

4.5.1.3 Definite and Indefinite Integration

Indefinite integration determines an antiderivative up to a constant so that, if $F(x)$ is some antiderivative of $f(x)$ and C is a constant, then

$$\int f(x)\,\mathrm{d}x = F(x) + C.$$

This expresses the fact that any value of C would produce a valid antiderivative of $f(x)$. In fact, these are the fibres – the cosets of the kernel – of the differentiation linear transformation. The kernel of differentiation is the set of constant-valued functions,

$$f(x) = C$$

for any $C \in \mathbb{R}$.

In the case of a definite integral, however, the constant cancels out:

$$\int_a^b f(x)\,\mathrm{d}x = [F(x)+C]_a^b = (F(b)+C)-(F(a)+C) = F(b)-F(a)$$

so that we are able to resolve the value of the definite integral to a specific constant value. In this case, the result is the sum of two elements of the kernel of the differentiation transform and so the result is also in the kernel.

However, if we consider a function of a variable, t, such that

$$h(t) = \int_a^t f(x)\,\mathrm{d}x$$

then $h(t)$ is also an antiderivative of $f(t)$ but also,

$$h(t) = \int_a^t f(x)\,\mathrm{d}x = [F(x)+C]_a^t = F(t)-F(a) = F(t)+C_2$$

where $C_2 = -F(a)$ is a constant. So, the definite integral – specifically, the initial value, a – has allowed us to resolve a particular antiderivative from the set of functions produced by the possible values of C. That's to say, the initial value a specified for us a particular element in the kernel of the differentiation operator – namely, $C_2 = -F(a)$.

Furthermore, since the definite integral is a particular antiderivative of $f(x)$ we can also relate the indefinite and definite integrals as follows,

$$\int f(x)\,\mathrm{d}x = \int_a^t f(x)\,\mathrm{d}x + C.$$

This amounts to – expressed in terms of group theory – the statement that, if G is a group with kernel K and $x, k \in G,\ k \in K$, then

$$xK = xkK$$

where

$$\begin{aligned} xK &= \int f(x)\,\mathrm{d}x = F(x)+C \\ xk &= \int_a^t f(x)\,\mathrm{d}x = F(t)-F(a) \\ xkK &= \int_a^t f(x)\,\mathrm{d}x + C = F(t)-F(a)+C. \end{aligned}$$

4.5.2 Multivariate Integration

4.5.2.1 Partial Differentiation of Integrals

$$\frac{\partial}{\partial x}\left(\int f(x,y)\,\mathrm{d}y\right) = \int \frac{\partial f(x,y)}{\partial x}\,\mathrm{d}y$$

4.5.2.2 Integration of Partial Derivatives

Suppose we have a function $f(x,y)$. Then,

$$\int \frac{\partial f(x,y)}{\partial x}\,\mathrm{d}x = g(x,y) + C(y) \quad \text{and} \quad \int \frac{\partial f(x,y)}{\partial y}\,\mathrm{d}x = g(x,y) + C(x).$$

So, the antiderivatives are only determined upto a function of the other variable. In this case, we can recover the original function $f(x,y)$ if we have both partial derivatives by setting the antiderivatives equal,

$$g_1(x,y) + C_1(y) = g_2(x,y) + C_2(x).$$

Any cross terms (terms containing both x and y) in f will appear in both $g_1(x,y)$ and $g_2(x,y)$ and the remaining term in g_1 will be equal to C_2 while the remaining term in g_2 will be equal to C_1.

Example

(68) Let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be such that for all $(x,y) \in \mathbb{R}^2$ the following equalities hold:

$$\frac{\partial \varphi}{\partial x}(x,y) = \underbrace{e^x \sin(y)}_{f(x,y)} \wedge \frac{\partial \varphi}{\partial y}(x,y) = \underbrace{e^x \cos(y)}_{g(x,y)}.$$

Integrating f with respect to x we get

$$\varphi(x,y) = \int e^x \sin(y) dx = e^x \sin(y) + \psi(y),$$

for some differentiable function $\psi : \mathbb{R} \to \mathbb{R}$. Differentiating with respect to y it follows that

$$e^x \cos(y) + \psi'(y) = g(x,y) = e^x \cos(y).$$

Therefore ψ' is always 0 and it follows that there exists $C \in \mathbb{R}$ such that for all $u \in \mathbb{R}$ we have $\psi(u) = C$ – that is, ψ is constant.
This gives you $\varphi(x, y) = e^x \sin(y) + C \in \mathbb{R}$, for some $C \in \mathbb{R}$. Taking $C \in \mathbb{R}$ arbitrarily will give you a possible φ.

In the case of definite integration though, whether the variables are independent is important (?). For example, suppose $y = y(x)$. Refer: https://math.stackexchange.com/questions/754742/integrating-a-partial-derivative and https://math.stackexchange.com/questions/2714061/integrating-partial-derivatives?noredirect=1&lq=1.

- If $y = y(x)$ then the result of the integral may not be meaningful.
- If we have both of the partial derivatives (wrt. x and y), then we can recover the original function regardless of whether the variables are independent or not.

4.6 Difference and Differential Equations

4.6.1 First-order Difference Equations

*A **difference equation** is also known as a **recurrence equation**.*

Definition. *Let y_t be the t-th value in a sequence (typically t represents time). Then,*

$$y_t = ay_{t-1} + b, \qquad t \geq 1$$

*is called a **first-order linear difference equation with constant coefficients**. The value y_0 is called an initial condition.*

A solution to such an equation is an explicit – or a closed-form (see: wikipedia) – expression for y_t in terms of t and y_0.

If $b = 0$ we have,

$$y_t = ay_{t-1} \iff y_t - ay_{t-1} = 0$$

*which is known as a **homogeneous** first-order linear difference equation with constant coefficients.*

Proposition 138. *A first-order linear difference equation with constant coefficients of the form $y_t = ay_{t-1} + b$ where $a = 1$ is a arithmetic progression.*

Proof. Let $y_t = y_{t-1} + b$ then,

$$y_1 = y_0 + b,\ y_2 = y_1 + b = (y_0 + b) + b = y_0 + 2b,\ y_3 = y_0 + 3b, \ldots$$

so we have $\boldsymbol{y_t = y_0 + tb}$. If we describe this as an arithmetic progression we have,

$$x_n = a + nd$$

where the zeroth term a corresponds to y_0, the common difference d corresponds to b and, clearly, t and n are both term indices. □

Proposition 139. *A first-order linear difference equation with constant coefficients of the form $y_t = ay_{t-1} + b$ where $b = 0$ is a geometric progression.*

Proof. Let $y_t = ay_{t-1} + 0$ then,

$$y_1 = ay_0,\ y_2 = ay_1 = a(ay_0) = a^2y_0,\ y_3 = a^3y_0, \ldots$$

so we have $\boldsymbol{y_t = a^t y_0}$. If we describe this as a geometric progression we have,

$$x_n = ar^n$$

where the zeroth term a corresponds to y_0, the common ratio r corresponds to a, and n is the term index corresponding to t. □

Proposition 140. *A first-order linear difference equation with constant coefficients of the form $y_t = ay_{t-1} + b$ where $a \neq 1$ and $b \neq 0$ has solution*

$$\begin{aligned} y_t &= a^t y_0 + b(a^{t-1} + a^{t-2} + \cdots + a + 1) \\ &= a^t y_0 + b\sum_{i=0}^{t-1} a^i. \end{aligned}$$

Proof.

$$\begin{aligned} y_t &= ay_{t-1} + b \\ &= a(ay_{t-2} + b) + b = a^2y_{t-2} + ab + b \\ &= a^2(a(y_{t-3} + b)) + ab + b = a^3y_{t-3} + a^2b + ab + b \\ &= a^3y_{t-3} + b(a^2 + a + 1) \\ &= a^t y_0 + b(a^{t-1} + \cdots + a + 1). \end{aligned}$$

□

Proposition 141. *A first-order linear difference equation with constant coefficients of the form $y_t = ay_{t-1} + b$ where $a \neq 1$ and $b \neq 0$ has solution*

$$y_t = a^t \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}$$

where the value

$$\frac{b}{1-a} = y^* = ay^* + b$$

is the ***equilibrium*** *or* ***steady-state*** *value of the recurrence.*

Proof. From Proposition 140 we know that the solution to the given recurrence is

$$y_t = a^t y_0 + b \sum_{i=0}^{t-1} a^i.$$

The summation in the second term is the sum of a geometric progression,

$$\begin{aligned} \sum_{i=0}^{t-1} a^i = 1 + a + \cdots + a^{t-1} &= \frac{a^t - 1}{a - 1} \\ &= \frac{a^t}{a-1} - \frac{1}{a-1} \\ &= \frac{1}{1-a} - \frac{a^t}{1-a}. \end{aligned}$$

So we see that the sum of a geometric progression can be separated into two terms: one term depends on the number of elements in the progression (here t), and the other term does not. This is the explanation of the sum of convergent geometric series: if $|a| < 1$ then, when $t \to \infty$, the t-dependent term disappears leaving only the steady-state term.

The solution to the recurrence therefore becomes

$$\begin{aligned} y_t &= a^t y_0 + b \left(\frac{1}{1-a} - \frac{a^t}{1-a} \right) \\ &= a^t \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}. \end{aligned}$$

□

4.6.1.1 First-order Linear Difference Equations as Affine Transformations

If we define a vectorized linear difference equation with constant coefficients,

$$\vec{\boldsymbol{y}}_t = A\vec{\boldsymbol{y}}_{t-1} + \vec{\boldsymbol{b}}$$

where A is a matrix, then clearly $\vec{\boldsymbol{y}}_t$ is an affine transformation of $\vec{\boldsymbol{y}}_{t-1}$.

We can linearize this by adding an extra dimension that takes the value 1 like so:

$$\begin{bmatrix} \vec{\boldsymbol{y}}_t \\ 1 \end{bmatrix} = \begin{bmatrix} A & \vec{\boldsymbol{b}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \vec{\boldsymbol{y}}_{t-1} \\ 1 \end{bmatrix}.$$

A minimal, one-by-one matrix is just a scalar (see: 45) and a minimal vector can be just a field element (see: 2.4.1.1) so we can define $A = a$ such that the linearized equation becomes

$$\begin{aligned} \begin{bmatrix} y_t \\ 1 \end{bmatrix} &= \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}^t \begin{bmatrix} y_0 \\ 1 \end{bmatrix}. \end{aligned}$$

Eigenvectors of the Transformation
In order to easily take powers of the transformation we find the eigenvectors.

Let M be the transformation matrix,

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

and, for an eigenvector $\vec{\boldsymbol{v}}$,

$$\begin{aligned} & M\vec{\boldsymbol{v}} = \lambda\vec{\boldsymbol{v}} \\ \iff \quad & \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \vec{\boldsymbol{v}} = \lambda\vec{\boldsymbol{v}} \end{aligned}$$

$$\iff \qquad \left(\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} - \lambda I \right) \vec{\boldsymbol{v}} = \vec{\mathbf{0}}$$

$$\iff \qquad \begin{bmatrix} a-\lambda & b \\ 0 & 1-\lambda \end{bmatrix} \vec{\boldsymbol{v}} = \vec{\mathbf{0}}$$

$$\begin{vmatrix} a-\lambda & b \\ 0 & 1-\lambda \end{vmatrix} = 0$$

$$\iff \qquad (a-\lambda)(1-\lambda) = 0$$

$$\iff \qquad \lambda \in \{1, a\}.$$

So, for eigenvalue 1:

$$\begin{bmatrix} a-1 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\iff \qquad (a-1)v_1 + bv_2 = 0$$

$$\iff \qquad (a-1)v_1 = -b \qquad \text{letting } v_2 = 1$$

$$\iff \qquad v_1 = \frac{-b}{a-1} = \frac{b}{1-a}$$

$$\therefore \qquad \vec{\boldsymbol{v}} = \begin{bmatrix} \frac{b}{1-a} \\ 1 \end{bmatrix}.$$

Note that $\frac{b}{1-a}$ is the steady-state solution of the difference equation.

For eigenvalue a:

$$\begin{bmatrix} 0 & b \\ 0 & 1-a \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\iff \qquad v_2 = 0$$

$$\therefore \qquad \vec{\boldsymbol{v}} = c\begin{bmatrix}1\\0\end{bmatrix} \text{ for } c \in \mathbb{R}.$$

Note that the second component of this vector adds the translation of the affine transformation so this is telling us that if $b = 0$ then any vector is an eigenvector with eigenvalue a. This corresponds to the homogeneous solution.

Diagonalization

So, if B is the set of two eigenvectors

$$\left\{\begin{bmatrix}\frac{b}{1-a}\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}\right\}$$

and we take this as a basis, then the change of basis matrix to this basis is

$$P = [B]^{-1} = \begin{bmatrix}\frac{b}{1-a} & 1\\1 & 0\end{bmatrix}^{-1} = \begin{bmatrix}0 & 1\\1 & \frac{-b}{1-a}\end{bmatrix}$$

and the diagonal matrix that represents the transformation with respect to this basis is

$$\begin{aligned} D &= PMP^{-1} \\ &= \begin{bmatrix}0 & 1\\1 & \frac{-b}{1-a}\end{bmatrix}\begin{bmatrix}a & b\\0 & 1\end{bmatrix}\begin{bmatrix}\frac{b}{1-a} & 1\\1 & 0\end{bmatrix} \\ &= \begin{bmatrix}0 & 1\\1 & \frac{-b}{1-a}\end{bmatrix}\begin{bmatrix}\frac{b}{1-a} & a\\1 & 0\end{bmatrix} \\ &= \begin{bmatrix}1 & 0\\0 & a\end{bmatrix}. \end{aligned}$$

Since D is diagonal we have

$$D^t = \begin{bmatrix}1 & 0\\0 & a^t\end{bmatrix}.$$

So, in order to calculate,

$$M^t = \begin{bmatrix}a & b\\0 & 1\end{bmatrix}^t$$

we can calculate,

$$\begin{aligned} & & D^t = (PMP^{-1})^t = PM^tP^{-1} \\ \iff & & P^{-1}D^tP = M^t. \end{aligned}$$

$$\begin{aligned} \therefore & & M^t &= \begin{bmatrix} \frac{b}{1-a} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & a^t \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & \frac{-b}{1-a} \end{bmatrix} \\ \iff & & M^t &= \begin{bmatrix} \frac{b}{1-a} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ a^t & \frac{-a^tb}{1-a} \end{bmatrix} \\ \iff & & M^t &= \begin{bmatrix} a^t & \frac{b(1-a^t)}{1-a} \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

So, the solution to the first-order linear recurrence is found to be

$$\begin{aligned} \begin{bmatrix} y_t \\ 1 \end{bmatrix} &= \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}^t \begin{bmatrix} y_0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} a^t & \frac{b(1-a^t)}{1-a} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ 1 \end{bmatrix} \end{aligned}$$

so that

$$y_t = a^t y_0 + \frac{b(1-a^t)}{1-a} = a^t \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}.$$

4.6.1.2 Difference as Rate of change

Geometric progression: Say we have a first-order linear difference equation with constant coefficients of the form $y_t = ay_{t-1} + b$ where $b = 0$ so that the terms follow a geometric progression:

$$y_1 = ay_0,\ y_2 = a^2y_0,\ y_3 = a^3y_0, \ldots$$

Then the rate of change is

$$\frac{\Delta y}{\Delta t} = y_t - y_{t-1} = ay_{t-1} - y_{t-1} = (a-1)y_{t-1}.$$

Note that the ratio of consecutive terms remains constant,

$$\frac{y_t}{y_{t-1}} = a$$

which is the "geometric" characteristic of a geometric progression, but the rate of change is proportional to the value.

The rate of change of the rate of change is, therefore,

$$\frac{\Delta^2 y}{\Delta t^2} = (a-1)\frac{\Delta y_{t-1}}{\Delta t} = (a-1)\frac{\Delta y}{\Delta t} = (a-1)^2 y_{t-1}.$$

4.6.1.3 Examples of first-order linear difference equations w/ const. coefficients

(69) Let y_t be an account balance after t years and r be the annual interest rate paid on the account. Suppose also, that the interest is compounded n times per year. Then the formula for y_t is,

$$y_t = \left(1 + \frac{r}{n}\right)^n y_{t-1} = y_0 \left(1 + \frac{r}{n}\right)^{nt}.$$

The rate of change per year is,

$$\frac{\Delta y}{\Delta t} = y_t - y_{t-1} = \left(1 + \frac{r}{n}\right)^n y_{t-1} - y_{t-1} = \left(\left(1 + \frac{r}{n}\right)^n - 1\right) y_{t-1}$$

as expected for a geometric progression.

If, instead, we let t be continuous we can find the instantaneous rate of change by considering the values for $t, t + \Delta t$,

$$\frac{\Delta y}{\Delta t} = \frac{y_{(t+\Delta t)} - y_t}{\Delta t} = \frac{y_0 \left(1 + \frac{r}{n}\right)^{n(t+\Delta t)} - y_0 \left(1 + \frac{r}{n}\right)^{nt}}{\Delta t}$$

$$= \frac{y_0 \left(1+\frac{r}{n}\right)^{nt} \left(\left(1+\frac{r}{n}\right)^{n\Delta t} - 1\right)}{\Delta t}$$

and then letting $\Delta t \to 0$ to obtain,

$$\begin{aligned}
\frac{\mathrm{d}y}{\mathrm{d}t} &= \lim_{\Delta t \to 0} \frac{y_0 \left(1+\frac{r}{n}\right)^{nt} \left(\left(1+\frac{r}{n}\right)^{n\Delta t} - 1\right)}{\Delta t} \\
&= y_0 \left(1+\frac{r}{n}\right)^{nt} \lim_{\Delta t \to 0} \frac{\left(1+\frac{r}{n}\right)^{n\Delta t} - 1}{\Delta t} \\
&= y_0 \left(1+\frac{r}{n}\right)^{nt} \lim_{\Delta t \to 0} \left[\frac{\mathrm{d}}{\mathrm{d}(\Delta t)} \left(1+\frac{r}{n}\right)^{n\Delta t}\right] && \text{by L'Hôpital's rule} \\
&= y_0 \left(1+\frac{r}{n}\right)^{nt} \lim_{\Delta t \to 0} \left[n \ln\left(1+\frac{r}{n}\right) \left(1+\frac{r}{n}\right)^{n\Delta t}\right] && \text{by Proposition 136} \\
&= n y_0 \ln\left(1+\frac{r}{n}\right) \left(1+\frac{r}{n}\right)^{nt}
\end{aligned}$$

But the question is: How meaningful is this really? If the interest is being compounded only n times per year and these are the only moments when the account balance changes, then change happens at certain discrete moments rather than continuously so is it really meaningful to talk about the instantaneous rate of change? We can recover the discrete-time rate of change per year from this instantaneous one by integrating over a year.

Now suppose that n, the number of times per year the interest is compounded, goes to infinity. Then (see 4.2.0.1) we have,

$$y_t = y_0 e^{rt}.$$

For discrete time we have,

$$\frac{\Delta y}{\Delta t} = y_0 e^{rt} - y_0 e^{r(t-1)} = y_0 e^{r(t-1)}(e^r - 1) = y_{t-1}(e^r - 1)$$

which should be no surprise as we still have a geometric progression as

$$\frac{y_t}{y_{t-1}} = e^r.$$

If we now let t be continuous as before then, by a similar logic using L'Hôpital's rule, the instantaneous rate of change obtained is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = ry_0e^{rt}.$$

Note that this is wholly consistent with the discrete time version as we can obtain the discrete time rate of change per year from this instantaneous rate of change by integrating over a year as follows,

$$\begin{aligned}
\int_{t-1}^{t} ry_0e^{rx}\,\mathrm{d}x &= ry_0\int_{t-1}^{t} e^{rx}\,\mathrm{d}x \\
&= ry_0\left[\frac{e^{rx}}{r}\right]_{t-1}^{t} \\
&= y_0\left[e^{rx}\right]_{t-1}^{t} \\
&= y_0(e^{rt} - e^{r(t-1)}) \\
&= y_0e^{r(t-1)}(e^r - 1).
\end{aligned}$$

Now suppose b is deposited at the end of each year so that,

$$\begin{aligned}
y_t &= \left(1+\frac{r}{n}\right)^n y_{t-1} + b \\
&= \left(1+\frac{r}{n}\right)^n \left[\left(1+\frac{r}{n}\right)^n y_{t-2} + b\right] + b \\
&= \left(1+\frac{r}{n}\right)^{nt} y_0 + b\sum_{i=0}^{t-1}\left(1+\frac{r}{n}\right)^{ni}.
\end{aligned}$$

But if we, again, let the number of compounds n, go to infinity, then,

$$\begin{aligned}
y_t &= e^r y_{t-1} + b \\
&= e^r(e^r y_{t-2} + b) + b \\
&= y_0e^{rt} + b\sum_{i=0}^{t-1} e^{ri} \\
&= y_0e^{rt} + b\left(\frac{e^{rt}-1}{e^r-1}\right)
\end{aligned}$$

$$= y_0 e^{rt} + b\frac{e^{rt}}{e^r - 1} - b\frac{1}{e^r - 1}$$
$$= \left(y_0 - \frac{b}{1 - e^r}\right) e^{rt} + \frac{b}{1 - e^r}$$

where $\frac{b}{1-e^r}$ is the steady-state value.

(70) Let M_t be an account balance at the end of year t. Let $Q(t)$ be an amount that is deposited into the account (or withdrawn if the value is negative) one time, at the end of year t and let the interest rate be a fixed rate of I. Then,

$$\begin{aligned} M_t &= IM_{t-1} + Q(t) \\ &= I(IM_{t-2} + Q(t-1)) + Q(t) \\ &= I(I(IM_{t-3} + Q(t-2)) + Q(t-1)) + Q(t) \\ &= M_0 I^t + \sum_{i=0}^{t-1} Q(t-i) I^i \end{aligned}$$

and the rate of change is,

$$M_t - M_{t-1} = M_0 I^{t-1}(I-1) + Q(1)I^{t-1} = I^{t-1}(M_0(I-1) + Q(1))$$

which is proportional to I^{t-1}.

If the interest rate I, is also a function of t, then – if we define $I(0) = 1$ – we end up with:

$$M_t = M_0 \prod_{i=0}^{t} I(i) + \sum_{i=0}^{t-1} Q(t-i) \prod_{j=0}^{i} I(i)$$

which is getting pretty messy. At this point it becomes easier to work with continuous time.

4.6.2 Second-order Difference Equations

Definition. *A recurrence equation of the form*

$$y_t = ay_{t-1} + by_{t-2} + c$$

where $a, b, c \in \mathbb{R}$, is known as a ***second-order linear recurrence with constant coefficients****.*

Following the same approach as with the first-order equations we have,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ 1 \end{bmatrix}.$$

Determining eigenvalues we get,

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

$$\iff \quad (1-\lambda)((-\lambda)(a-\lambda) - b) = 0$$

$$\therefore \quad \lambda \in \left\{ 1, \frac{a+\sqrt{a^2+4b}}{2}, \frac{a-\sqrt{a^2+4b}}{2} \right\}.$$

Due to the translation represented by the lone 1 in the final row of the matrix, there will always be the eigenvalue 1. This corresponds to the steady state value which may be thought of as the eigenvector of the translation. The other two eigenvalues are the eigenvectors of the linear transformation part of the affine transformation. The block structure is as follows.

$$\begin{bmatrix} \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} c \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ 1 \end{bmatrix} = \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix}$$

$$= A\vec{\boldsymbol{y}} + \vec{\boldsymbol{t}}.$$

Steady-state value

Considering the case of the eigenvalue $\lambda = 1$ we have,

$$\begin{bmatrix} a-1 & b & c \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

So, to find the nullspace of the matrix we can find the row-reduced echelon form,

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

$$\rightsquigarrow \quad \begin{bmatrix} 1 & \frac{b}{a-1} & \frac{c}{a-1} \\ 0 & -1-\frac{b}{a-1} & -\frac{c}{a-1} \\ 0 & 0 & 0 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & \frac{b}{a-1} & \frac{c}{a-1} \\ 0 & 1 & \frac{c}{a+b-1} \\ 0 & 0 & 0 \end{bmatrix}$$

$$\rightsquigarrow \quad \begin{bmatrix} 1 & 0 & \frac{c}{a+b-1} \\ 0 & 1 & \frac{c}{a+b-1} \\ 0 & 0 & 0 \end{bmatrix}$$

where the last step is because

$$\begin{aligned} \frac{c}{a-1} - \left(\frac{b}{a-1}\right)\frac{c}{a+b-1} &= \frac{c}{a-1}\left(1 - \frac{b}{a+b-1}\right) \\ &= \frac{c}{a-1}\left(\frac{a-1}{a+b-1}\right). \end{aligned}$$

So the nullspace is

$$t \begin{bmatrix} \frac{-c}{a+b-1} \\ \frac{-c}{a+b-1} \\ 1 \end{bmatrix}$$

for any $t \in \mathbb{R}$ and the steady-state value is

$$\frac{-c}{a+b-1}.$$

This is also sometimes referred to as a ***particular solution of the non-homogeneous equation****.*

Eigenvalues of the linear transformation

Let the discriminant $d = \sqrt{a^2 + 4b}$. Then the eigenvalues of the linear transformation are

$$\frac{a+d}{2} \quad \text{and} \quad \frac{a-d}{2}.$$

In the case of the eigenvalue $\frac{a+d}{2}$ we have,

$$\begin{bmatrix} \frac{a-d}{2} & b & c \\ 1 & -(\frac{a+d}{2}) & 0 \\ 0 & 0 & 1-(\frac{a+d}{2}) \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Again using row reduction to find the nullspace:

$$\begin{bmatrix} \frac{a-d}{2} & b & c \\ 1 & -(\frac{a+d}{2}) & 0 \\ 0 & 0 & 1-(\frac{a+d}{2}) \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & -(\frac{a+d}{2}) & \frac{2c}{a-d} \\ 1 & -(\frac{a+d}{2}) & 0 \\ 0 & 0 & 1-(\frac{a+d}{2}) \end{bmatrix} \quad \text{using } \frac{2b}{a-d}\frac{a+d}{a+d} = -(\frac{a+d}{2})$$

$$\rightsquigarrow \begin{bmatrix} 1 & -(\frac{a+d}{2}) & \frac{2c}{a-d} \\ 0 & 0 & \frac{-2c}{a-d} \\ 0 & 0 & 1-(\frac{a+d}{2}) \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & -(\frac{a+d}{2}) & \frac{2c}{a-d} \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\rightsquigarrow \begin{bmatrix} 1 & -(\frac{a+d}{2}) & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

So the nullspace is

$$t \begin{bmatrix} \frac{a+d}{2} \\ 1 \\ 0 \end{bmatrix}$$

for any $t \in \mathbb{R}$ and the value

$$\frac{a+d}{2} = \frac{a + \sqrt{a^2 + 4b}}{2}$$

is the solution to the homogeneous equation corresponding to the eigenvalue $(a+d)/2$. It's also not hard to see that the other eigenvalue $(a-d)/2$ results in the solution to the homogeneous equation

$$\frac{a-d}{2} = \frac{a-\sqrt{a^2+4b}}{2}.$$

Another common way of finding these solutions in this case is to form what is known as the ***auxiliary equation*** *as:*

$$\begin{aligned}
& y_t - ay_{t-1} - by_{t-2} = 0 && \\
\rightsquigarrow \quad & m^t - am^{t-1} - bm^{t-2} = 0 && \text{for some } m \neq 0 \in \mathbb{R} \\
\iff \quad & m^{t-2}(m^2 - am - b) = 0 && \\
\iff \quad & m^2 - am - b = 0. && \text{since } m \neq 0
\end{aligned}$$

Note that this is the characteristic polynomial of the linear transformation part of the matrix:

$$(-\lambda)(a-\lambda) - b = \lambda^2 - a\lambda - b.$$

In this case, determining the final equation for y_t using matrix powers results in a very complex matrix and formula which is only reasonably performed on a computer. But we can infer that the formula will involve the eigenvalues of the linear transformation, raised to the power t, and the steady-state value,

$$y_t = C_1\frac{(a+d)^t}{2^t} + C_2\frac{(a-d)^t}{2^t} + \frac{c}{a+b-1}.$$

Then we can use the initial values to solve for the constants C_1, C_2,

$$y_0 = C_1 + C_2 + \frac{c}{a+b-1} \quad \text{and} \quad y_1 = C_1\frac{a+d}{2} + C_2\frac{a-d}{2} + \frac{c}{a+b-1}.$$

The values

$$\frac{a+d}{2} = \frac{a+\sqrt{a^2+4b}}{2} \quad \text{and} \quad \frac{a-d}{2} = \frac{a-\sqrt{a^2+4b}}{2}$$

may be two distinct real values, one real value (if $a^2+4b=0$) or two distinct complex values (if $a^2+4b<0$).

In the case of one real value: the formula becomes,

$$y_t = (C_1 + C_2 t)\frac{(a+d)^t}{2^t} + \frac{c}{a+b-1}.$$

The explanation of this appears to be that the matrix $A - \lambda I$ has cyclic order 2. TODO: ? eh?

In the case of two complex values: expressing the eigenvalues in exponential form we have,

$$\begin{aligned}\frac{a}{2} \pm \frac{\sqrt{-a^2-4b}}{2}i &= \sqrt{\frac{a^2}{4} + \frac{-a^2-4b}{4}}\,\exp\left(i\arctan\pm\frac{\sqrt{-a^2-4b}}{a}\right)\\ &= \sqrt{-b}\,\exp\left(i\arctan\pm\sqrt{-1-\frac{4b}{a^2}}\right).\end{aligned}$$

Let

$$\theta = \arctan\sqrt{-1-\frac{4b}{a^2}} \quad \text{and} \quad -\theta = \arctan-\sqrt{-1-\frac{4b}{a^2}}.$$

Then the eigenvalues raised to the power of t are:

$$(\sqrt{-b})^t\,e^{i\theta t} \quad \text{and} \quad (\sqrt{-b})^t\,e^{-i\theta t}$$

which means that the formula becomes,

$$\begin{aligned}y_t &= C_1(\sqrt{-b})^t\,(\cos(\theta t) + i\sin(\theta t))\\ &\quad + C_2(\sqrt{-b})^t\,(\cos(-\theta t) + i\sin(-\theta t))\\ &\quad + \frac{c}{a+b-1}\\ &= C_1(\sqrt{-b})^t\,(\cos(\theta t) + i\sin(\theta t))\\ &\quad + C_2(\sqrt{-b})^t\,(\cos(\theta t) - i\sin(\theta t))\\ &\quad + \frac{c}{a+b-1}\\ &= (\sqrt{-b})^t\,(C_1 + C_2)\cos(\theta t)\end{aligned}$$

$$
\begin{aligned}
&+ i(\sqrt{-b})^t \left(C_1 - C_2\right) \sin(\theta t) \\
&+ \frac{c}{a+b-1} \\
&= (\sqrt{-b})^t \left(C_3 \cos(\theta t) + iC_4 \sin(\theta t)\right) + \frac{c}{a+b-1}.
\end{aligned}
$$

But, since we are looking for real-valued solutions and any linear combination of the homogeneous solutions is also a homogeneous solution, we can generate other, real-valued homogeneous solutions from linear combinations of these ones. In fact, we can just divide the imaginary solution by i so that our real-valued solution is:

$$
(\sqrt{-b})^t \left(C_3 \cos(\theta t) + C_4 \sin(\theta t)\right) + \frac{c}{a+b-1}.
$$

TODO: eh? if divide by i then the real part will be divided by i also

4.6.3 Markov Chains

4.6.3.1 Markov Matrices (a.k.a. Stochastic Matrices)

This section taken from Harvard.

Definition. *An $n \times n$ matrix is called a **Markov** or **Stochastic** matrix if all entries are nonnegative and the sum of each column vector is equal to 1.*

The matrix

$$A = \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix}$$

is a Markov matrix.

Many authors write the transpose of the matrix and apply the matrix to the right of a row vector.

Let's call a vector with nonnegative entries p_k for which all the p_k add up to 1 a stochastic vector. For a stochastic matrix, every column is a stochastic vector.

Theorem 42. *If p is a stochastic vector and A is a stochastic matrix, then Ap is a stochastic vector.*

Proof. Let $v_1, \ldots, v_n$ be the column vectors of A. Then

$$Ap = \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_n \end{bmatrix} = p_1 v_1 + \ldots + v_n v_n$$

If we sum this up we get $p_1 + p_2 + \ldots + p_n = 1$. □

Theorem 43. *A Markov matrix A always has an eigenvalue* 1. *All other eigenvalues are in absolute value smaller or equal to* 1.

Proof. For the transpose matrix A^T, the sum of the row vectors is equal to 1. The matrix A^T therefore has the eigenvector

$$\begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}.$$

Because A and A^T have the same determinant, also $A - \lambda I_n$ and $A^T - \lambda I_n$ have the same determinant so that the eigenvalues of A and A^T are the same (TODO: we can reference the proof that a matrix is similar to its transpose but does this explanation also make sense?). With A^T having an eigenvalue 1 also A has an eigenvalue 1 Assume now that v is an eigenvector with an eigenvalue $|\lambda| > 1$. Then $A^n v = |\lambda|^n v$ has exponentially growing length for $n \to \infty$. This implies that there is for large n one coefficient $[A^n]_{ij}$ which is larger than 1. But A^n is a stochastic matrix (see homework) and has all entries ≤ 1. The assumption of an eigenvalue larger than 1 can not be valid. □

For there to be a long-term distribution of the markov chain it is necessary that the eigenvalue 1 in the markov matrix have multiplicity 1. Otherwise, there may be more than one eigenvector corresponding with eigenvalue 1 or the matrix may not be diagonalizable. In the case that it is diagonalizable, there will only be one eigenvector with eigenvalue 1 that is also a distribution vector (stochastic vector).

4.6.4 Differential Equations

Differential equations are most often used to describe the evolving state of dynamical systems – that is, systems whose future state is a function of its current state. Therefore, that portion of the future state that is dependent on the previous state is compounded in a similar fashion to compound interest. For this reason, the solutions to differential equations typically involve the exponential function.

4.6.4.1 Types of Differential Equations

Definition. *A* ***differential equation*** *– that is, a single equation as opposed to a system of equations – is an equation that relates a single dependent variable's derivatives to each other and may or may not explicitly include the independent variable. A common convention is for the dependent variable to be y and the independent variable to be t – reflecting the fact that it is often modelling time – but x is often used also.*

Definition. *A differential equation that does not explicitly include the independent variable is known as an* ***autonomous*** *equation. It represents a time-invariant system if the independent variable is viewed as time. So, if $y(t) = g(t)$ is a solution then $y(t) = g(t + c)$, for constant c, is also a solution.*

Definition. *A* ***first-order*** *differential equation is an equation in which only derivatives upto and including the first derivative of the dependent variable appear. That's to say, an equation that relates the dependent variable to its first derivative and, potentially, to the independent variable.*

Definition. *Similarly, a* ***second-order*** *differential equation relates the dependent variable to both its first and second derivatives as well as, potentially, to the independent variable. Higher-order differential equations also exist – but are less common – with the order being given by the highest derivative present in the equation.*

Definition. *A* ***linear*** *differential equation is an equation containing only degree-one monomial terms in the derivatives of the dependent variable. So, the equation is linear in the derivatives of the dependent variable although it may also contain any function of the independent variable.*

Definition. *A* ***nonlinear*** *differential equation is an equation that contains nonlinear terms of the derivatives of the dependent variable.*

Definition. *A* ***separable*** *first-order equation is one where the first-derivative of the dependent variable may be expressed as a single term. That's to say, we can put the equation in the form,*

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)g(y).$$

Note that the definition of a separable equations means that – if A, B, C, D are constants – the following equation is separable,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ABt^2y^2 + ADt^2 + BCy^2 + CD$$

because it can be factorized into,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = (At^2 + C)(By^2 + D) = f(t)g(y).$$

Definition. *A **partial** differential equation is a differential equation that includes at least one partial derivative. Otherwise, a differential equation is known as an **ordinary** differential equation. The two terms are frequently abbreviated to ODE and PDE.*

4.6.4.2 Solutions

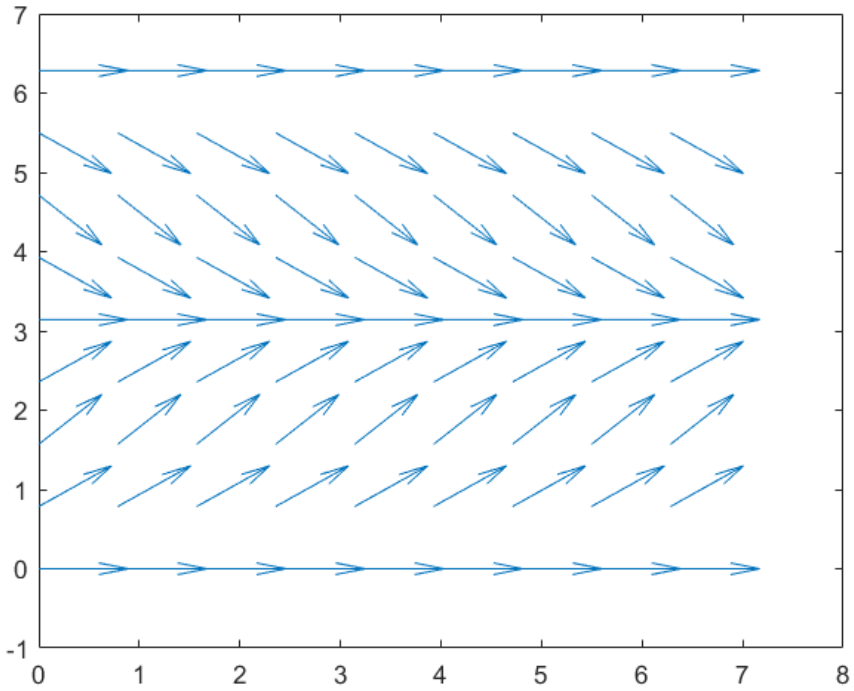
Definition. *A **solution** to a differential equation – also called a **general solution** – is a set of functions that share the relation expressed in the differential equation.*

This is similar to the solution to an indefinite integral being a set of antiderivatives. However, an indefinite integral determines an antiderivative upto an additive constant but the general solution of a differential equation may only determine a function upto a multiplicative constant.

Definition. *A **solution to the initial value problem** of a differential equation is a specific function where the initial value has determined a particular member of the set of functions in the general solution. That's to say, it is a function that **both** exhibits the relation in the differential equation and satisfies the initial condition(s) of the initial value problem.*

A general solution of a differential equation defines a direction field where at every co-ordinate – (x, y) or (t, y) – the gradient is defined by the equation $\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y)$. The solution to an IVP is a path through this direction field beginning at a point (t_0, y_0) representing the initial conditions.

Figure 4.1: The direction field of $y' = \sin(y)$. The field contains steady-states at $y = 0, \pi, \frac{pi}{2}$.



4.6.4.3 Equilibrium Points / Steady-states

Definition. *A **steady-state** solution is an IVP solution that is constant across all values of the independent variable. Therefore, the initial conditions of the IVP must be this constant value.*
For example, if $y(t) = C$ for some constant C, is a steady-state solution then, for the initial condition $y(t_0) = C$, the constant function $y(t) = C$ is a solution to the IVP. Since this function is constant for all values of t, the derivative $\frac{\mathrm{d}y}{\mathrm{d}t} = 0$ also at all values of t.

Although a function $y(t) = C$ is only a solution to the IVP $y(t_0) = C$ other solutions satisfying other initial conditions may converge to the value C. In this way, the steady-state may be the long-term behaviour of IVP solutions from a whole set of initial conditions.

If we take the general form of a first-order differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)g(y) + h(t),$$

then, for some value of the function y_s to be a steady-state solution we need, for all t,

$$\begin{aligned} & \frac{\mathrm{d}y}{\mathrm{d}t} = f(t)g(y_s) + h(t) = 0 \\ \iff \qquad & g(y_s) = -\frac{h(t)}{f(t)}. \end{aligned}$$

For example, the equation $y' = \frac{4t}{y} - 2t$ has a steady-state at $y = 2$.

Definition. *A steady-state is called **stable** if the set of initial conditions that converge to it is greater than its value alone. In other words, if $y(t) = C$ is a steady-state, then it is stable if the set of all initial values of the function $y_0 = y(t_0)$ whose IVP solutions converge to it,*

$$S = \{\, y_0 \mid y(t_0) = y_0 \implies \lim_{t\to\infty} y(t) = C \,\}$$

is a proper superset of $\{C\}$ – i.e. $\{C\} \subsetneq S$.

*Conversely, a steady-state is called **unstable** if the set of initial conditions that converge to it is only the value itself. That's to say,*

$$S = \{C\}.$$

*A steady-state can also be **semi-stable** if it is stable from one side and unstable from the other.*

*Stable steady-states happen when greater values of the function decrease to the steady-state value and lesser values of the function increase to the steady-state value. An **unstable** steady-state, conversely, is one where values of the function that are a little lesser and greater move away from the steady-state value. The states of real-world systems modelled by these unstable steady-states may sometimes not, in practice, be referred to as equilibria due to their instability.*

Examples of types of steady-state

(71) A ball rolling into a dip in the ground stays there whereas a ball at the top of a hump in the ground will likely roll off. If a dip in the ground is described by the curve $y = x^2$ so that the stable equilibrium is at $x = 0$ then, if the ball is pushed and rolls up the hill on the right side, the potential energy gained is proportional to x^2 and the tendency to return to the equilibrium is dependent on the rate at which the potential energy is released as the ball rolls back to the bottom, which is $\frac{\mathrm{d}x^2}{\mathrm{d}x} = 2x$. Since this is acting to reduce the value of the displacement x its sign is negative so $-2x$. Clearly the converse is the situation for a hump in the ground. So, in this model, the dip in the ground has a gradient field with a maximum at the equilibrium (similar to the curve of $y = -x^2$) while the hump in the ground has a gradient field with a minimum at the equilibrium (similar to the curve of $y = x^2$).

(72) The equation $y' = \sin(y)$ in figure 4.1 shows a stable equilibrium at $y = \pi$ and unstable equilibria at $y = 0, 2\pi$. This is because $\frac{\partial}{\partial y}\sin(y) = \cos(y)$ which is negative around $y = \pi$ – so this is a maximum and therefore, stable – and positive around the values $y = 0, 2\pi$ – so these are minimums and therefore, unstable.

(73) Another example: The equation $y' = \frac{4t}{y} - 2t$ has a **stable** steady-state at $y = 2$ for positive t but at the same y-value the steady-state is unstable for negative values of t. This is because $\frac{\partial}{\partial y}\left(\frac{4t}{y} - 2t\right) = -\frac{4t}{y^2} = -t$ at $y = 2$ which means that the steady-state is a maximum and stable for positive t and a minimum and unstable for negative t.

Note that, if t is modelling time, negative values of t may not be meaningful.

(74) The equation $y' = y^2$ has a semi-stable steady-state at $y = 0$: solution curves below it converge to it, but those above it diverge. Why? At $y = 0$ we have $\frac{\partial}{\partial y}\, y^2 = 2y = 0$ so this is an inflection point. We look at before and after and see that both have positive values of the gradient y'.

(75) A corner case: The equation $y' = 2\sqrt{y}$. This has a steady-state at $y = 0$ which is unstable from above because $\frac{\partial}{\partial y}\,(2\sqrt{y}) = \frac{1}{\sqrt{y}}$ is positive for positive y so this is a minimum when viewed from above. But the partial derivative has an infinite discontinuity as $y = 0$ and doesn't exist as a real number for negative y. Likewise the gradient y' is not a real number for negative y and so is undefined in this situation.

4.6.5 First-order Linear ODEs

The form $\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)y$

The most simple form has a single y-term whose coefficient may be a function of x. This form is separable as,

$$\begin{aligned}
& \frac{\mathrm{d}y}{\mathrm{d}x} = f(x)y \\
\iff \quad & \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} = f(x) \\
\iff \quad & \int \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x}\,\mathrm{d}x = \int f(x)\,\mathrm{d}x \\
\iff \quad & \ln|y| = F(x) + c && y \neq 0,\ F \text{ is an antiderivative of } f \\
\iff \quad & |y| = e^{F(x)} \cdot e^{c} \\
\iff \quad & y = ke^{F(x)} && k \in \mathbb{R}.
\end{aligned}$$

Check solution:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)ke^{F(x)} = f(x)y.$$

Note that the solution has the form,

$$y = ke^{F(x)}$$

where $F(x)$ is an antiderivative of $f(x)$, the coefficient of y in the original differential equation. Since the antiderivative is unique upto a constant factor, the other possible antiderivatives are achieved by the value of the coefficient k because,

$$e^{F(x)+c} = e^{F(x)} \cdot e^{c} = ke^{F(x)}.$$

I.V.P. Solution

If we have an initial value for the function – say $y(t_0) = y_0$ – then we need

$$y(t_0) = y_0 = y_0(1) = y_0e^0 = y_0e^{\int_{t_0}^{t} f(x)\,\mathrm{d}x}.$$

That's to say, with the addition of an initial value, the general solution,

$$y(t) = e^{\int f(t)\,\mathrm{d}t} = ke^{F(t)}$$

becomes a complete solution,

$$y(t) = y_0 e^{\int_{t_0}^{t} f(x)\,\mathrm{d}x}.$$

We can see the equivalence by setting the constant k in the general solution to $y_0/e^{F(t_0)}$,

$$y(t) = \frac{y_0}{e^{F(t_0)}}e^{F(t)} = y_0\frac{e^{F(t)}}{e^{F(t_0)}} = y_0 e^{(F(t)-F(t_0))} = y_0 e^{\int_{t_0}^{t} f(x)\,\mathrm{d}x}.$$

Separable Equations

Any equation of the form $\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)y$ is separable into the form,

$$f(x)\,\mathrm{d}x = g(y)\,\mathrm{d}y$$

and can then be solved by integrating both sides. As a result, the solution moves on level sets of $\int f(x)\,\mathrm{d}x - \int g(y)\,\mathrm{d}y$. This value, then, is invariant (i.e. constant) across all solutions and so, tends to represent a conserved quantity in physical systems. Therefore, these type of equations, when modelling physical systems, could be described as modelling closed systems with no external influence.

Examples

(76) Say we have an IVP (Initial Value Problem) such that our independent variable is t beginning at 0, with $y(0) = y_0$, and

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)y.$$

Then if we look at what happens at integral intervals of t we can see that:

$$y(0) = y_0$$

$$
\begin{aligned}
y(1) &= y_0 e^{\int_0^1 f(t)dt} \\
y(2) &= y_0 e^{\int_0^1 f(t)dt} e^{\int_1^2 f(t)dt} = y_0 e^{\int_0^2 f(t)dt} \\
&\vdots
\end{aligned}
$$

So, in general, at time t,

$$y(t) = y_0 e^{\int_0^t f(u)du}.$$

The form $\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)y + g(x)$
This form is not separable as it is. But if we multiply both sides by $e^{-F(x)}$, where $F(x)$ is an antiderivative of $f(x)$, then

$$
\begin{aligned}
& & \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x)y + g(x) \\
\iff & & e^{-F(x)}\frac{\mathrm{d}y}{\mathrm{d}x} &= e^{-F(x)}f(x)y + e^{-F(x)}g(x) \\
\iff & e^{-F(x)}\frac{\mathrm{d}y}{\mathrm{d}x} - e^{-F(x)}f(x)y &= e^{-F(x)}g(x) & \\
\iff & \frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-F(x)}y\right) &= e^{-F(x)}g(x) & \\
\iff & \int \frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-F(x)}y\right)\mathrm{d}x &= \int e^{-F(x)}g(x)\,\mathrm{d}x & \\
\iff & & e^{-F(x)}y &= \int e^{-F(x)}g(x)\,\mathrm{d}x \\
\iff & & y &= e^{F(x)}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x
\end{aligned}
$$

where the constant of integration of the integral on the left has been absorbed into the integral on the right.

Note, also, that

$$ke^{F(x)}\int \frac{g(x)}{ke^{F(x)}}\,\mathrm{d}x = ke^{F(x)}\frac{1}{k}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x = e^{F(x)}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x.$$

So, the solution has the form,

$$y = ke^{F(x)} + h(x)e^{F(x)}$$

where $h(x)$ is an antiderivative of $g(x)/e^{F(x)}$ and k is the constant of integration.

Check solution:

$$\begin{aligned} y &= e^{F(x)}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x \implies \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x)\left(e^{F(x)}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x\right) + e^{F(x)}\frac{g(x)}{e^{F(x)}} \\ &= f(x)\left(e^{F(x)}\int \frac{g(x)}{e^{F(x)}}\,\mathrm{d}x\right) + g(x) \\ &= f(x)y + g(x). \end{aligned}$$

and also,

$$\begin{aligned} y &= ke^{F(x)} + h(x)e^{F(x)} \implies \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x)ke^{F(x)} + f(x)h(x)e^{F(x)} + g(x)e^{-F(x)}e^{F(x)} \\ &= f(x)ke^{F(x)} + f(x)h(x)e^{F(x)} + g(x) \\ &= f(x)y + g(x) \end{aligned}$$

where we have used the fact that $h(x)$ is an antiderivative of $g(x)e^{-F(x)}$.

I.V.P. Solution

For the solution of the IVP we are going to want a definite integral. If the

exponent in the integrating factor has to be an antiderivative of $f(t)$, as a definite integral, we can use $e^{\int_a^t f(x)\,\mathrm{d}x}$ where a is any constant real number and t is our independent variable.

Note that the function,

$$F(t) = \int_a^t f(x)\,\mathrm{d}x$$

has derivative $F'(t) = f(t)$, by the FTC, and has $F(a) = 0$.

We can set the value of a to fit the initial conditions. For example, if $y(t_0) = y_0$ and $y(t) = y_0 e^{\int_a^t f(x)\,\mathrm{d}x}$ as in the separable case, then we can set $a = t_0$ so that the initial condition is met.

Now suppose we have an non-separable differential equation with an initial value $y(t_0) = y_0$. When we integrate using our integrating factor we want definite integration starting at t_0, so,

$$\begin{aligned}
& & \frac{\mathrm{d}y}{\mathrm{d}t} &= f(t)y + g(t) \\
\iff & & e^{-\int_a^t f(x)\,\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}t} - e^{-\int_a^t f(x)\,\mathrm{d}x} f(t)y &= e^{-\int_a^t f(x)\,\mathrm{d}x} g(t) \\
\iff & & \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-\int_a^t f(x)\,\mathrm{d}x} y\right) &= e^{-\int_a^t f(x)\,\mathrm{d}x} g(t) \\
\iff & & \int_{t_0}^t \frac{\mathrm{d}}{\mathrm{d}u}\left(e^{-\int_a^u f(x)\,\mathrm{d}x} y(u)\right)\mathrm{d}u &= \int_{t_0}^t e^{-\int_a^u f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u \\
\iff & & e^{-\int_a^t f(x)\,\mathrm{d}x} y(t) - e^{-\int_a^{t_0} f(x)\,\mathrm{d}x} y(t_0) &= \int_{t_0}^t e^{-\int_a^u f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u \\
\iff & & e^{-\int_a^t f(x)\,\mathrm{d}x} y(t) &= \int_{t_0}^t e^{-\int_a^u f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u + e^{-\int_a^{t_0} f(x)\,\mathrm{d}x} y(t_0) \\
\iff & & y(t) &= e^{\int_a^t f(x)\,\mathrm{d}x}\Big(\int_{t_0}^t e^{-\int_a^u f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u + e^{-\int_a^{t_0} f(x)\,\mathrm{d}x} y(t_0)\Big) \\
\iff & & y(t) &= \int_{t_0}^t e^{\int_u^t f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u + e^{\int_{t_0}^t f(x)\,\mathrm{d}x} y(t_0).
\end{aligned}$$

So the resultant solution form is:

$$y(t) = y_0 e^{\int_{t_0}^{t} f(x)\,\mathrm{d}x} + \int_{t_0}^{t} e^{\int_{u}^{t} f(x)\,\mathrm{d}x} g(u)\,\mathrm{d}u$$

for $y(t_0) = y_0$.

Examples

(77) $\boldsymbol{x\frac{\mathrm{d}y}{\mathrm{d}x} - 2y = 6}$:

The first step is to rearrange it to obtain an expression for the derivative.

$$\begin{aligned} & x\frac{\mathrm{d}y}{\mathrm{d}x} - 2y = 6 \\ \iff \qquad & \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{2}{x}y + \frac{6}{x}. \end{aligned}$$

Then we can apply the derived formula using the antiderivative $F(x) = 2\ln x$.

$$\begin{aligned} y &= e^{2\ln x} \int \frac{6/x}{e^{2\ln x}}\,\mathrm{d}x \\ &= 6x^2 \int \frac{1}{x^3}\,\mathrm{d}x \\ &= 6x^2 \left(-\frac{1}{2x^2} + c\right) \\ &= -3 + c'x^2. \qquad\qquad c' = 6c \end{aligned}$$

We can confirm this result by performing the calculation.

$$e^{-2\ln x}\frac{\mathrm{d}y}{\mathrm{d}x} = e^{-2\ln x}\frac{2}{x}y + e^{-2\ln x}\frac{6}{x}$$

$$
\begin{aligned}
\iff \qquad & x^{-2}\frac{\mathrm{d}y}{\mathrm{d}x} - x^{-2}\frac{2}{x}y = x^{-2}\frac{6}{x} \\
\iff \qquad & x^{-2}\frac{\mathrm{d}y}{\mathrm{d}x} - \frac{2}{x^3}y = \frac{6}{x^3} \\
\iff \qquad & \frac{\mathrm{d}}{\mathrm{d}x}(x^{-2}y) = \frac{6}{x^3} \\
\iff \qquad & \int \frac{\mathrm{d}}{\mathrm{d}x}(x^{-2}y)\,\mathrm{d}x = 6\int \frac{1}{x^3}\,\mathrm{d}x \\
\iff \qquad & x^{-2}y = 6(-\frac{1}{2x^2} + c) \\
\iff \qquad & x^{-2}y = -3\frac{1}{x^2} + c' \qquad\qquad c' = 6c \\
\iff \qquad & y = -3 + c'x^2.
\end{aligned}
$$

(78) Consider a bank account with variable interest. Let $M(t)$ be the amount of money in the account at time t, measured in years (though the specific unit isn't important conceptually), and let $I(t)$ be the rate of interest at time t: for example, 3% interest corresponds to $I(t) = 0.03$. Finally, let $Q(t)$ be the amount of money put in (or negative for removing money) in year t. Thus, $M(t)$ obeys the differential equation

$$\frac{\mathrm{d}M}{\mathrm{d}t} = I(t)M(t) + Q(t).$$

This illustrates a common trend: the first term indicates how it would grow independent of external forces, and the second term represents external influences.
This can be solved in different ways:

First, suppose $Q(t) = 0$ and $I(t) = I$ is constant. Then $M(t) = M(0)e^{It}$.

On the other hand, if $I(t)$ can vary, then $M(t) = M(0)e^{\int_0^t I(u)\,\mathrm{d}u}$ as explained in 71.

When $Q(t)$ isn't zero, we can begin to understand it by considering the case where money is only put in once at the end of each year — so we have $Q(0), Q(1), \ldots$ — giving the solution:

$$M(t) = M(0)e^{\int_0^t I(u)\,\mathrm{d}u} + Q(1)e^{\int_1^t I(u)\,\mathrm{d}u} + Q(2)e^{\int_2^t I(u)\,\mathrm{d}u} + \cdots$$

In the case of continuous deposit and withdrawal from the account we end up with:

$$M(t) = M(0)e^{\int_0^t I(u)\,\mathrm{d}u} + \int_0^t Q(x)e^{\int_x^t I(u)\,\mathrm{d}u}\,\mathrm{d}x.$$

We can relate this to the general formula for this form of differential equation using 4.5.1.3 as follows:

$$\begin{aligned} M(t) &= e^{\int I(t)\,\mathrm{d}t} \int Q(t)e^{-\int I(u)\,\mathrm{d}u}\,\mathrm{d}t \\ &= e^{\int_0^t I(u)\,\mathrm{d}u} \left(\int_0^t Q(x)e^{-\int_0^x I(u)\,\mathrm{d}u}\,\mathrm{d}x + C \right) \\ &= Ce^{\int_0^t I(u)\,\mathrm{d}u} + \int_0^t Q(x)e^{\int_x^t I(u)\,\mathrm{d}u}\,\mathrm{d}x \end{aligned}$$

which gives us $M(0) = C$ when we substitute in $t = 0$.

4.6.5.1 Number of Solutions

First-order linear differential equations always have a single solution. The linear algebra of first-order linear equations certainly supports a single unique solution: the matrix is non-singular. However, the proof of this lies in the explicit formula derived above,

$$y(t) = e^{\int_{t_0}^t f(x)\,\mathrm{d}x} \int_{t_0}^t g(x)e^{\int_x^t f(u)\,\mathrm{d}u}\,\mathrm{d}x.$$

4.6.6 First-order Nonlinear ODEs

For first-order nonlinear odes, separable equations can be solved in exactly the same manner as separable linear odes. Non-separable equations, however, cannot be solved as with linear equations.

4.6.6.1 Number of Solutions

For nonlinear differential equations, nothing general can be said about the number of *global* solutions. There could be 0, 1, or many. For *local* solutions over a restricted domain, two things can be said though:

Let the function $f(t, y)$ be differentiable on the interval $[t_0, t_1]$ and let $y' = f(t, y)$ with the initial condition $y(t_0) = y_0$. Then,

(i) There is at least one solution over some interval $[t_0, t_2]$ for $t_0 \leq t_2 \leq t_1$;

(ii) If $\frac{\partial f}{\partial y}$ is continuous over some interval $[t_0, t_2]$ for $t_0 \leq t_2 \leq t_1$, then there is one unique solution over the interval $[t_0, t_2]$.

$\frac{\partial f}{\partial y}$ represents the way that the derivative of y changes with the value of y. If it has an infinite discontinuity then this would seem to suggest an infinite sensitivity to initial conditions at the point of the singularity. Of course, any type of discontinuity may be a modelling problem.

Examples

(79) Continuing the example **??**, suppose we have the equation $y' = y^2$ and the initial value $y(0) = 1$. This is an autonomous equation and, as such, is separable. So,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = y^2$$

$$\iff \qquad \frac{1}{y^2}\,\mathrm{d}y = \mathrm{d}x$$

$$\iff \qquad \frac{-1}{y} = x + C$$

$$\iff \qquad y = \frac{-1}{x+C}.$$

Applying the initial value we have,

$$y(0) = \frac{-1}{0+C} = 1$$
$$\iff \qquad C = -1.$$

Therefore the solution is $y(x) = \frac{1}{1-x}$. But this is **not** a *global* solution! There is a singularity at $x = 1$ so this formula only provides *local* solutions over intervals of x that do not include 1.

(i) There is a solution $y(x) = \frac{1}{1-x}$ over any interval that does not include $x = 1$.

(ii) $\frac{\partial f}{\partial y} = 2y$ is continuous everywhere so solutions of this equation over an interval are unique. Note that there is a steady-state at $y = 0$ but it is unreachable from the initial condition of $y(0) = 1$.

(80) Continuing the example **??**, suppose we have the equation $y' = 2\sqrt{y}$ and the initial value $y(0) = 0$. This is also an autonomous equation and so is separable.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 2\sqrt{y}$$

$$\iff \qquad \frac{1}{\sqrt{y}}\,\mathrm{d}y = 2\,\mathrm{d}x$$

$$\iff \qquad 2\sqrt{y} = 2x + C$$

$$\iff \qquad \sqrt{y} = x + D$$
$$\iff \qquad y = (x+D)^2 = x^2 + 2Dx + D^2.$$

If we apply the initial value then we can determine that

$$y(0) = D^2 = 0 \implies D = 0.$$

This gives us the solution $y(x) = x^2$. But there is another solution: There is a steady-state at $y = 0$ and, since the initial condition is $y(0) = 0$, the initial condition is in the steady-state. Therefore $y(x) = 0$ is also a solution.

(i) There are 2 solutions $y(x) = x^2$ and $y(x) = 0$ over all intervals of x.

(ii) $\frac{\partial f}{\partial y} = \frac{1}{\sqrt{y}}$ is discontinuous at $y = 0$ – in fact, it is an infinite discontinuity – so, since $y = 0$ is the initial value, solutions over any interval with this initial value will not be unique. (And, in this case, there are 2 solutions).

(81) Consider a ball falling under the influence of gravity but resisted by air-resistance. (The example uses a spherical object because other shapes would be likely to have a much more complicated influence of air-resistance on them.) Let $v(t)$ be the velocity at a time t and g be the acceleration due to gravity.

We can model the air-resistance as $-av^2$ for some constant a. This model is quadratic in the velocity of the falling object because: the faster the object is falling through the air, the greater the resisting force produced by the air (by Newton's Second Law) but, also, the greater the amount of air that the object is coming into contact with in a given time interval. Furthermore, since the air-resistance is acting in the opposite direction to the movement of the object it has the opposite sign to the velocity.

Therefore, our final model of the falling ball is,

$$\frac{\mathrm{d}v}{\mathrm{d}t} = g - av^2.$$

This is a separable equation so we can proceed to solve it by the re-arrangement,

$$\frac{1}{g - av^2}\,\mathrm{d}v = \mathrm{d}t.$$

We have two (at least) possible approaches to the analytical solution of the integration,

$$\int \frac{1}{g - av^2}\,\mathrm{d}v.$$

We can use trigonometry functions or partial fractions. Using trig. we proceed as,

$$\begin{aligned}
\int \frac{1}{g - av^2}\,\mathrm{d}v &= \frac{1}{g}\int \frac{1}{1 - (a/g)v^2}\,\mathrm{d}v \\
&= \frac{1}{g}\int \frac{1}{1 - \sin^2\theta}\,\mathrm{d}(\sqrt{g/a}\sin\theta) \\
&= \frac{1}{g}\int \frac{\sqrt{g/a}\cos\theta}{1 - \sin^2\theta}\,\mathrm{d}\theta \\
&= \frac{\sqrt{g/a}}{g}\int \frac{\cos\theta}{\cos^2\theta}\,\mathrm{d}\theta \\
&= \frac{1}{\sqrt{ag}}\int \sec\theta\,\mathrm{d}\theta \\
&= \frac{1}{\sqrt{ag}}\ln(\sec\theta + \tan\theta) && \text{w/o the constant} \\
&= \frac{1}{\sqrt{ag}}\ln\left(\frac{1}{\sqrt{1 - (a/g)v^2}} + \sqrt{\frac{(a/g)v^2}{1 - (a/g)v^2}}\right) \\
&= \frac{1}{\sqrt{ag}}\ln\left(\frac{1 + \sqrt{\frac{a}{g}}\,v}{\sqrt{1 - (\frac{a}{g})v^2}}\right) \\
&= \frac{1}{\sqrt{ag}}\ln\left(\frac{\sqrt{g} + \sqrt{a}\,v}{\sqrt{g - av^2}}\right)
\end{aligned}$$

$$= \frac{1}{2\sqrt{ag}} \ln\left(\frac{g + 2\sqrt{ag}\, v + av^2}{g - av^2}\right)$$

$$= \frac{1}{2\sqrt{ag}} \ln\left(\frac{(\sqrt{g} + \sqrt{a}\, v)^2}{(\sqrt{g} + \sqrt{a}\, v)(\sqrt{g} - \sqrt{a}\, v)}\right)$$

$$= \frac{1}{2\sqrt{ag}} \ln\left(\frac{\sqrt{g} + \sqrt{a}\, v}{\sqrt{g} - \sqrt{a}\, v}\right).$$

Alternatively, using partial fractions we can proceed as,

$$\begin{aligned}
\int \frac{1}{g - av^2}\, \mathrm{d}v &= \int \frac{A}{\sqrt{g} + \sqrt{a}v} + \frac{B}{\sqrt{g} - \sqrt{a}v}\, \mathrm{d}v \\
&= \int \frac{1}{2\sqrt{g}(\sqrt{g} + \sqrt{a}v)} + \frac{1}{2\sqrt{g}(\sqrt{g} - \sqrt{a}v)}\, \mathrm{d}v \\
&= \frac{1}{2\sqrt{g}} \int \frac{1}{\sqrt{g} + \sqrt{a}v} + \frac{1}{\sqrt{g} - \sqrt{a}v}\, \mathrm{d}v \\
&= \frac{1}{2\sqrt{g}}\left(\frac{1}{\sqrt{a}} \ln(\sqrt{g} + \sqrt{a}v) + \frac{-1}{\sqrt{a}} \ln(\sqrt{g} - \sqrt{a}v)\right) \\
&= \frac{1}{2\sqrt{ag}}(\ln(\sqrt{g} + \sqrt{a}v) - \ln(\sqrt{g} - \sqrt{a}v)) \\
&= \frac{1}{2\sqrt{ag}} \ln\left(\frac{\sqrt{g} + \sqrt{a}v}{\sqrt{g} - \sqrt{a}v}\right).
\end{aligned}$$

So, either way, we arrive at,

$$\frac{1}{2\sqrt{ag}} \ln\left(\frac{\sqrt{g} + \sqrt{a}v}{\sqrt{g} - \sqrt{a}v}\right) = t + C$$

$$\iff \qquad \frac{\sqrt{g} + \sqrt{a}v}{\sqrt{g} - \sqrt{a}v} = Ae^{2\sqrt{ag}\, t} \qquad A = e^{2\sqrt{ag}\, C}$$

$$\iff \quad \sqrt{g} + \sqrt{a}v(1 + Ae^{2\sqrt{ag}\, t}) = \sqrt{g}Ae^{2\sqrt{ag}\, t}$$

$$\iff \qquad \sqrt{a}v(1 + Ae^{2\sqrt{ag}\,t}) = \sqrt{g}(Ae^{2\sqrt{ag}\,t} - 1)$$

$$\iff \qquad v = \frac{\sqrt{g}}{\sqrt{a}}\left(\frac{Ae^{2\sqrt{ag}\,t} - 1}{1 + Ae^{2\sqrt{ag}\,t}}\right)$$

$$\iff \qquad v = \sqrt{\frac{g}{a}}\left(\frac{Ae^{2\sqrt{ag}\,t} - 1}{Ae^{2\sqrt{ag}\,t} + 1}\right).$$

If the initial condition is $v(0) = 0$ – that's to say that the body starts at rest – then we can resolve the value of the constant $A = 1$. This then gives us the solution,

$$v(t) = \sqrt{\frac{g}{a}}\left(\frac{e^{2\sqrt{ag}\,t} - 1}{e^{2\sqrt{ag}\,t} + 1}\right) = \sqrt{\frac{g}{a}}\tanh(\sqrt{ag}\,t).$$

The terminal velocity is given by allowing $t \to \infty$ and the result is that $v \to \sqrt{\frac{g}{a}}$. As expected, the terminal velocity decreases with increasing values of the constant a.

Also, as expected from looking at the original equation and the partial derivative w.r.t. v, we have obtained a single unique solution – which is good because it could represent a problem for Newtonian physics if we didn't.

4.6.7 Autonomous, Separable and Exact ODEs

Autonomous Equations and Exact Equations

Definition. *An **autonomous** equation refers to a differential equation with no **explicit** dependence on the independent variable (typically in dynamical systems, t for time). These differential equations express change in a system based solely on the current value of the system. For example:*

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -ky \quad \textit{and} \quad \frac{\mathrm{d}y}{\mathrm{d}t} = ry(1-y).$$

Autonomous equations are always separable.

On the other hand, an equation of the form,

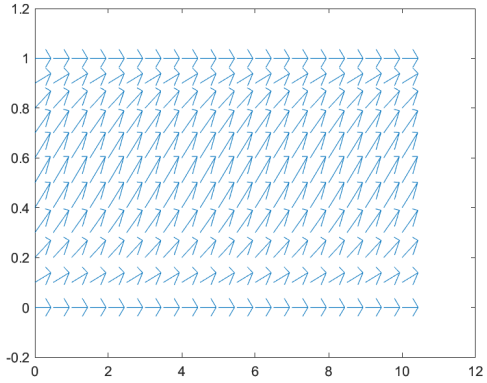
$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)y(1-y)$$

for example, would express that the growth rate intrinsic to the system is varying over time. Equations of this form are not autonomous but are separable.

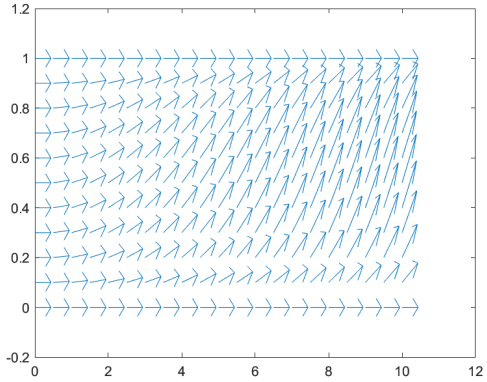
Meanwhile, an equation of the form,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -ky + f(x)$$

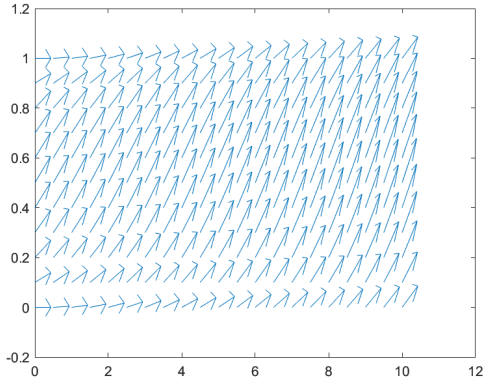
for example, would express that the rate of change of the system is being influenced by something that varies with the independent variable and is, in some way, extrinsic to the system (because it depends only on the independent variable and not on the state of the system). Equations of this form are neither autonomous nor separable.



(a) autonomous: $y' = y(1-y)$ - note that the field is invariant to a shift along the x-axis.



(b) separable: $y' = f(x)y(1-y)$ where $f(x) = 0.2x$



(c) non-separable: $y' = y(1-y) + f(x)$ where $f(x) = 0.02x$

Figure 4.1: autonomous, separable and non-separable direction fields

Definition. *An **exact** equation refers to a differential equation whose solutions all conserve the value of some property. All separable equations are exact equations because,*

$$f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x \implies \int f(y)\,\mathrm{d}y - \int g(x)\,\mathrm{d}x = 0.$$

However, there are also exact equations that are not separable such as,

$$y'\sin x + y\cos x + 2x = 0 \implies (y\sin x + x^2)' = 0 \implies y = \frac{C - x^2}{\sin x}.$$

Proposition 142. *Let y be any function satisfying the differential equation,*

$$A(x,y) + B(x,y)\,\frac{\mathrm{d}y}{\mathrm{d}x} = 0$$

with

$$\frac{\partial A}{\partial y} = \frac{\partial B}{\partial x}\,.$$

Then all such functions y preserve some invariant property $\Psi(x,y) = C$ for constant C.

Proof. The function Ψ may consist of terms of only x, terms of only y, terms containing both x and y and a constant term. So we can describe it as,

$$\Psi(x,y) = p(x)q(y) + r(x) + s(y) + k$$

where p, q, r, s are arbitrary functions and k is a constant. But, since Ψ will be a constant – and we are only interested in the fact that it is constant, not the value of the constant – we can ignore the constant term k because it will just change the value of the constant that Ψ is equal to. In other words, we can describe Ψ as,

$$\Psi(x,y) = p(x)q(y) + r(x) + s(y) = C$$

with k absorbed into the value of C.

Then, taking partial derivatives with respect to x and y we have,

$$\frac{\partial \Psi}{\partial x} = \frac{\mathrm{d}p}{\mathrm{d}x}\, q(y) + \frac{\mathrm{d}r}{\mathrm{d}x} \quad \text{and} \quad \frac{\partial \Psi}{\partial y} = p(x)\, \frac{\mathrm{d}q}{\mathrm{d}y} + \frac{\mathrm{d}s}{\mathrm{d}y}\,.$$

Furthermore, since Ψ is a constant function we know that,

$$\frac{\mathrm{d}\Psi}{\mathrm{d}x} = 0 = \frac{\partial \Psi}{\partial x} + \frac{\partial \Psi}{\partial y}\, \frac{\mathrm{d}y}{\mathrm{d}x}\,.$$

This has the required form $A(x,y) + B(x,y)\,\frac{\mathrm{d}y}{\mathrm{d}x} = 0$ with $\frac{\partial A}{\partial y} = \frac{\partial B}{\partial x}$. So, if we have a differential equation of the given form and we postulate the existence of a constant function Ψ such that $\frac{\partial \Psi}{\partial x} = A(x,y)$ and $\frac{\partial \Psi}{\partial y} = B(x,y)$, then,

$$\begin{aligned}
\int A(x,y)\,\mathrm{d}x = \int \frac{\partial \Psi}{\partial x}\,\mathrm{d}x &= \int \left(\frac{\mathrm{d}p}{\mathrm{d}x}\, q(y) + \frac{\mathrm{d}r}{\mathrm{d}x} \right) \mathrm{d}x \\
&= p(x)q(y) + r(x) + C_1(y), \\
\int B(x,y)\,\mathrm{d}y = \int \frac{\partial \Psi}{\partial y}\,\mathrm{d}y &= \int \left(p(x)\, \frac{\mathrm{d}q}{\mathrm{d}y} + \frac{\mathrm{d}s}{\mathrm{d}y} \right) \mathrm{d}y \\
&= p(x)q(y) + s(y) + C_2(x).
\end{aligned}$$

Comparing these two results,

$$\Psi(x,y) = p(x)q(y) + r(x) + C_1(y) = p(x)q(y) + s(y) + C_2(x),$$

we can see that the function $\Psi(x,y)$ that we are looking for is $\Psi(x,y) = p(x)q(y) + r(x) + s(y)$.

Another way that we could have resolved Ψ is to take the first integral,

$$\int A(x,y)\,\mathrm{d}x = \int \frac{\partial \Psi}{\partial x}\,\mathrm{d}x = p(x)q(y) + r(x) + C_1(y)$$

and say, if $\Psi(x,y) = p(x)q(y) + r(x) + C_1(y)$ then,

$$\frac{\partial \Psi}{\partial y} = p(x)\, \frac{\mathrm{d}q}{\mathrm{d}y} + \frac{\mathrm{d}C_1}{\mathrm{d}y}$$

and compare this with $B(x,y)$ to resolve the function C_1,

$$\begin{aligned} & p(x)\frac{\mathrm{d}q}{\mathrm{d}y} + \frac{\mathrm{d}C_1}{\mathrm{d}y} = B(x,y) = p(x)\frac{\mathrm{d}q}{\mathrm{d}y} + \frac{\mathrm{d}s}{\mathrm{d}y} \\ \iff \quad & \frac{\mathrm{d}C_1}{\mathrm{d}y} = \frac{\mathrm{d}s}{\mathrm{d}y} \\ \iff \quad & C_1(y) = s(y) + k \qquad \text{integrating both sides wrt. } y \end{aligned}$$

where k is a constant of integration which, in this case, can be ignored. □

Proposition 143. *All separable equations are exact equations.*

Proof. A separable differential equation is one that may be put in the form,

$$f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x.$$

We can re-arrange this,

$$\begin{aligned} & f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x \\ \iff \quad & f(y)\frac{\mathrm{d}y}{\mathrm{d}x} - g(x) = 0. \end{aligned}$$

Now, taking $A(x,y) = -g(x)$ and $B(x,y) = f(y)$ we have the form $A(x,y) + B(x,y)\frac{\mathrm{d}y}{\mathrm{d}x} = 0$ with $\frac{\partial A}{\partial y} = \frac{\partial B}{\partial x} = 0$ so this is an exact equation. □

All autonomous equations are separable (though the reverse is not generally true) and all separable equations are exact equations (although some exact equations are not separable). So we have,

$$\{\textit{Autonomous}\} \subset \{\textit{Separable}\} \subset \{\textit{Exact}\}.$$

Examples of Exact Equations

(82) Suppose we have the equation

$$(3x^2 + 2xy) + (2y + x^2)\,\frac{\mathrm{d}y}{\mathrm{d}x} = 0.$$

We have,

$$\frac{\partial(3x^2 + 2xy)}{\partial y} = 2x = \frac{\partial(2y + x^2)}{\partial x}$$

so this is an exact equation. We can solve for the constant function preserved by all solutions by setting

$$\begin{aligned}
&\int 3x^2 + 2xy\,\mathrm{d}x = \int 2y + x^2\,\mathrm{d}y \\
\iff \quad & x^3 + x^2y + C_1(y) = y^2 + x^2y + C_2(x) \\
\iff \quad & x^3 + C_1(y) = y^2 + C_2(x) \\
\therefore \quad & \Psi(x, y) = x^2y + y^2 + x^3 = C \qquad \text{for constant C.}
\end{aligned}$$

So, all solutions $y(x)$ of this differential equation preserve the value of Ψ and so, if we have an initial condition – say $y(0) = 1$ – then we can use this to find the value of $C = \Psi(0, 1)$ and this value will be preserved for all values of x and y.

$$\Psi(0, 1) = (0^2)(1) + (1)^2 + (0^3) = 1 = x^2y + y^2 + x^3 \qquad \forall x, y \in \mathbb{R}.$$

Now we can solve for y by treating x as a constant and arranging the equation as a quadratic in y:

$$y^2 + x^2y + (x^3 - 1) = 0$$

and then using the quadratic formula for y,

$$y = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-x^2 \pm \sqrt{x^4 - 4x^3 - 4}}{2}.$$

(83) Suppose we have the equation

$$(3x + 2y) + (\frac{2y}{x} + x)\,\frac{\mathrm{d}y}{\mathrm{d}x} = 0.$$

This is equation is *not* exact because,

$$\frac{\partial(3x + 2y)}{\partial y} = 2 \neq \frac{\partial(\frac{2y}{x} + x)}{\partial x} = \frac{-2y}{x^2} + 1.$$

However, if we multiply the equation by x then we get the equation in the previous example,

$$x(3x + 2y) + x(\frac{2y}{x} + x)\,\frac{\mathrm{d}y}{\mathrm{d}x} = (3x^2 + 2xy) + (2y + x^2)\,\frac{\mathrm{d}y}{\mathrm{d}x} = 0.$$

So there are times when multiplying by an integrating factor can result in an exact equation.

(84) Consider a body falling toward earth from a significant distance so that the change in acceleration due to gravity as the body gets closer to the earth is significant. Acceleration due to gravity is proportional to $1/r^2$ where r is the distance of the body from the centre of the earth – remember the mass of the body is a multiplier of the force of gravity but also of the inertia so in the expression for the resultant acceleration the mass cancels out – so we can model this situation as,

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{c}{r^2}$$

where c is a positive constant. Note that the expression on the right hand side is negative because r is always positive but, in the scenario, is decreasing so $\frac{\mathrm{d}r}{\mathrm{d}t}$ is negative, and it is getting more negative as r gets smaller, so the second derivative is also negative.

If we use classical Newtonian mechanics then we can model r as a displacement s from the center of the earth. Then the velocity $v = \frac{\mathrm{d}s}{\mathrm{d}t}$ is negative as the body is falling towards the earth. We also have,

$$a = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}s} \cdot \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}s} \cdot v.$$

Interestingly, we can see that,

$$\frac{\mathrm{d}v}{\mathrm{d}s} = \frac{a}{v} = \frac{\mathrm{d}v/\,\mathrm{d}t}{v}$$

which is the log-derivative of v representing the relative infinitesimal change (wikipedia) of the velocity and also,

$$\frac{\mathrm{d}v}{\mathrm{d}s} = \frac{\mathrm{d}(\mathrm{d}s/\,\mathrm{d}t)}{\mathrm{d}s}$$

the way that the rate of change of the displacement changes with the displacement.

If we consider the first-order equation

$$\frac{\mathrm{d}v}{\mathrm{d}s}\, v = -\frac{c}{s^2},$$

this has the form $y\,\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)$ and so, is separable.

$$\begin{aligned}
& \frac{\mathrm{d}v}{\mathrm{d}s}\, v = -\frac{c}{s^2} \\
\iff \qquad & \int v\,\mathrm{d}v = -c\int \frac{1}{s^2}\,\mathrm{d}s \\
\iff \qquad & \frac{v^2}{2} = -c\left(\frac{-1}{s}\right) + C \\
\iff \qquad & \frac{v^2}{2} = \frac{c}{s} + C \\
\iff \qquad & \frac{v^2}{2} - \frac{c}{s} = C.
\end{aligned}$$

So the quantity $\frac{v^2}{2} - \frac{c}{s}$ is preserved by all solutions of this differential equation. The gravitational potential energy of a body displaced from the earth is given (by Newton's 3rd law) as the work done if the body were to fall all the way to the earth's centre,

$$E_p = F \cdot s = m\, a(s) \cdot s$$

where m is the mass of the body, s is the displacement of the body from the centre of the earth and a is the acceleration due to gravity

as a function of the displacement. Substituting in the function for the acceleration due to gravity as a function of the distance from the earth's centre we have,

$$E_p = m\left(-\frac{c}{s^2}\right)s = -\frac{mc}{s}.$$

The kinetic energy of the falling body is given by,

$$E_k = \frac{mv^2}{2}.$$

Since the falling movement of the body is the potential energy turning to kinetic energy, the sum $E_p + E_k$ is conserved. So, for some constant K we have,

$$\begin{aligned} & \frac{mv^2}{2} - \frac{mc}{s} = K \\ \iff \qquad & m\left(\frac{v^2}{2} - \frac{c}{s}\right) = K \\ \iff \qquad & \frac{v^2}{2} - \frac{c}{s} = C \qquad \text{for } C = K/m. \end{aligned}$$

(85) Imagine a pendulum swinging on a rod (or a rope that remains taut) of length L. Let θ be the angle the rod makes with the vertical and g the acceleration due to gravity. The gravitational force acting on the pendulum has a component that is balanced by the tension in the rod maintaining the pendulum swinging along the circumference of a circle. This component acts in a direction perpendicular to the circumference of the circle in which the pendulum mass moves and along the radius of the circle of movement. The other component of g, orthogonal to this one, acts along the tangent of the circumference of movement of the pendulum mass. It is this component that creates the motion of the pendulum mass. (If this is not clear: Penn State page on pendulum oscillation.)

We have:

- Acceleration due to gravity with a component perpendicular to the pendulum rod given by: $a = -g \sin\theta$.
- Angular velocity of the pendulum mass: $\omega = \dfrac{\mathrm{d}\theta}{\mathrm{d}t}$.
- Tangential velocity of the pendulum mass: $v_t = L \dfrac{\mathrm{d}\theta}{\mathrm{d}t}$.

So, using $F = m\frac{\mathrm{d}v}{\mathrm{d}t}$, the equation describing the acceleration of the pendulum mass is,

$$\begin{aligned} & m \frac{\mathrm{d}v_t}{\mathrm{d}t} = -mg \sin\theta \\ \iff \qquad & \frac{\mathrm{d}v_t}{\mathrm{d}t} = -g \sin\theta \\ \iff \qquad & L \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -g \sin\theta. \end{aligned}$$

This is a second-order nonlinear equation. For small oscillations we can use the small-angle approximation of sine to linearize it (see example 83). If the oscillations are not small though, the small-angle approximation becomes too inaccurate to be useful and we must solve the nonlinear equation.

We can simplify it into a first-order nonlinear equation by describing the tangential velocity in relation to the angle θ and eliminating time from the model,

$$v = L\omega = L \frac{\mathrm{d}\theta}{\mathrm{d}t} \iff \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{v}{L},$$

$$L\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}\theta} \cdot \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}\theta} \cdot \frac{v}{L} = -g \sin\theta.$$

So we end up with the *separable* first-order nonlinear equation,

$$\begin{aligned} & v \frac{\mathrm{d}v}{\mathrm{d}\theta} = -gL \sin\theta \\ \iff \qquad & \int v \, \mathrm{d}v = -gL \int \sin\theta \, \mathrm{d}\theta \end{aligned}$$

$$\iff \qquad v^2 = 2gL\cos\theta + C$$

$$\iff \qquad v(\theta) = \pm\sqrt{2gL\cos\theta + C}.$$

Note that we have ended up with an expression for the tangential velocity as a function of the angle θ *only*. Also worth noting is that the constant quantity,

$$v^2 - 2gL\cos\theta = C$$

represents the conserved energy of the pendulum system – the v^2 term being proportional to the kinetic energy and the $2gL\cos\theta$ term being proportional to the potential energy.

If we consider an oscillation with maximum angle θ_{max} then the pendulum is momentarily at rest when $\theta = \theta_{max}$. It is convenient to use this as the initial condition of the angle of the pendulum $\theta_0 = \theta_{max}$ so that the pendulum begins at rest and so $v(\theta_0) = 0$ and then we can resolve the value of the constant,

$$v(\theta_0) = \pm\sqrt{2gL\cos\theta_0 + C} = 0$$

$$\therefore \qquad C = -2gL\cos\theta_0.$$

So, for the initial condition, we have resolved the tangential velocity w.r.t. to the angle of displacement of the pendulum as,

$$v(\theta) = \pm\sqrt{2gL(\cos\theta - \cos\theta_0)}.$$

We could also have obtained this result with definite integration from θ_0 to θ,

$$\int_{v(\theta_0)}^{v(\theta)} v\,\mathrm{d}v = -gL\int_{\theta_0}^{\theta}\sin t\,\mathrm{d}t$$

$$\iff \qquad \frac{1}{2}(v(\theta)^2 - v(\theta_0)^2) = gL(\cos\theta - \cos\theta_0)$$

$$\iff \qquad v(\theta)^2 = 2gL(\cos\theta - \cos\theta_0) \qquad \because v(\theta_0) = 0.$$

To recover the time information into the solution we can bring back the definition of the velocity as a function of time *as well as* the angle of displacement,

$$v = L\,\frac{\mathrm{d}\theta}{\mathrm{d}t}\,.$$

Substituting this back into the solution we get,

$$\begin{aligned} L\,\frac{\mathrm{d}\theta}{\mathrm{d}t} &= \pm\sqrt{2gL(\cos\theta - \cos\theta_0)} \\ \iff \qquad \mathrm{d}t &= \sqrt{\frac{L}{2g}}\,\frac{1}{\sqrt{\cos\theta - \cos\theta_0}}\,\mathrm{d}\theta. \end{aligned}$$

So, to find the time taken when the pendulum swings from its central position, with $\theta = 0$, to its amplitude, with $\theta_{max} = \theta_0$, we can integrate the expression on the right of this equation between θ_0 and 0. To get the whole time period of the oscillation we can multiply this time by 4,

$$T = 4\sqrt{\frac{L}{2g}}\int_{\theta_0}^{0}\frac{1}{\sqrt{\cos\theta - \cos\theta_0}}\,\mathrm{d}\theta.$$

This integral is a type of improper integral called an elliptic integral and it doesn't have an analytic solution. However, for small-angle oscillations we can use the small-angle approximation for cosine,

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

to obtain an approximate value of the integral,

$$\begin{aligned} &\int_{\theta_0}^{0}\frac{1}{\sqrt{(1-\frac{\theta^2}{2})-(1-\frac{\theta_0^2}{2})}}\,\mathrm{d}\theta \\ =&\int_{\theta_0}^{0}\frac{1}{\sqrt{\frac{1}{2}(\theta_0^2-\theta^2)}}\,\mathrm{d}\theta \\ =&\sqrt{2}\int_{\theta_0}^{0}\frac{1}{\sqrt{\theta_0^2-\theta^2}}\,\mathrm{d}\theta \\ =&\frac{\sqrt{2}}{\theta_0}\int_{\theta_0}^{0}\frac{1}{\sqrt{1-\left(\frac{\theta}{\theta_0}\right)^2}}\,\mathrm{d}\theta \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2}}{\theta_0} \int_{\theta_0}^{0} \frac{1}{\sqrt{1-\sin^2\alpha}} \, \mathrm{d}(\theta_0 \sin\alpha) \\
&= \frac{\sqrt{2}}{\theta_0} \int_{\sin^{-1}(1)}^{\sin^{-1}(0)} \frac{\theta_0 \cos\alpha}{\cos\alpha} \, \mathrm{d}\alpha \\
&= \sqrt{2}(\sin^{-1}(0) - \sin^{-1}(1)) \\
&= \sqrt{2}\left(-\frac{\pi}{2}\right) = -\frac{\pi}{\sqrt{2}}.
\end{aligned}
$$

In this case we can ignore the minus sign – it's merely a factor of the choice that $\sin^{-1}$ has to make in order to be a function; returning $\frac{\pi}{2}$ for $\sin^{-1}(1)$ instead of $-\frac{\pi}{2}$. So, the time period becomes,

$$T = 4\sqrt{\frac{L}{2g}}\left(\frac{\pi}{\sqrt{2}}\right) = 2\pi\sqrt{\frac{L}{g}}.$$

Note that the amplitude of the pendulum's oscillation θ_0 cancelled out in the calculation of the time period and that the resultant expression for the *time period of small oscillations* of a pendulum depends only on the length of the string and the acceleration due to gravity.

Homogeneous Differential Equations

Definition. *A* ***homogeneous differential equation*** *is an equation of the form,*

$$f(x,y)\frac{\mathrm{d}y}{\mathrm{d}x} = g(x,y)$$

where the functions f, g are both homogeneous of degree d.

The functions f, g need not be linear and so these differential equations are not necessarily linear.

The key insight here is that, due to the homogeneous function property (see: 4.4), if we define the function y to be $y(x) = x \cdot v(x)$ — which we may do because, due to the non-trivial kernel of differentiation, we are only able to determine a solution to a differential equation upto a constant term so we can set the constant term to 0 for convenience during the calculation — then,

$$f(\lambda x, \lambda y) = \lambda^d f(x, y) \implies f(x, xv) = x^d g(v).$$

When this is applied in a differential equation of the above form we obtain,

$$\begin{aligned}
& f(x,y)\frac{\mathrm{d}y}{\mathrm{d}x} = g(x,y) \\
\iff \quad & x^d f_1(v)\left(v + x\frac{\mathrm{d}v}{\mathrm{d}x}\right) = x^d f_2(v) && y = xv,\ y' = v + xv' \\
\iff \quad & f_1(v)\left(v + x\frac{\mathrm{d}v}{\mathrm{d}x}\right) = f_2(v) \\
\iff \quad & v + x\frac{\mathrm{d}v}{\mathrm{d}x} = \frac{f_2(v)}{f_1(v)} = f_3(v) \\
\iff \quad & x\frac{\mathrm{d}v}{\mathrm{d}x} = f_3(v) - v = f_4(v) \\
\iff \quad & \frac{1}{f_4(v)}\frac{\mathrm{d}v}{\mathrm{d}x} = \frac{1}{x} \\
\iff \quad & \int \frac{1}{f_4(v)}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = \int \frac{1}{x}\,\mathrm{d}x = \ln|x| + c \\
\iff \quad & \int \frac{1}{f_4(v)}\,\mathrm{d}v = \ln|x| + c \\
\iff \quad & \frac{1}{f_4'(v)}\ln|f_4(v)| = \ln|x| + c.
\end{aligned}$$

4.6.8 Comparison of Differential Equations with Difference Equations

TODO: Comparison of Differential Equations with Difference Equations

4.6.9 Second-order Linear ODEs

4.6.9.1 Examples of Second-order Linear

(86) **The Logistic:** The logistic model of population growth takes $P(t)$ to be the population at time t. The model takes some additional factor that illustrates that a space can only hold a fixed carrying capacity of the population, producing the equation

$$\frac{\mathrm{d}P}{\mathrm{d}t} = cP(t)\left(1 - \frac{P(t)}{A}\right).$$

From looking at the differential equation
Whatever the value of the derivative, we can see that the scalar c will increase its magnitude thereby amplifying changes. So, c is the growth rate and larger values of c cause the population to change more rapidly.

The steady-states of $P(t)$ are at the values such that the derivative is 0. Therefore,

$$cP(t)\left(1 - \frac{P(t)}{A}\right) = 0 \implies P(t) \in \{0, 1\}.$$

Looking at the steady-states one-by-one:

- $\boldsymbol{P(t) = 0}$: If $P(t)$ climbs above 0 then the derivative will be positive and so the function will be increasing away from the steady-state at 0. This is therefore an *unstable* steady-state.

  > *Note that, since $P(t)$ is a population, negative values don't make sense.*

- $\boldsymbol{P(t) = A}$: If $P(t)$ is less than A then the derivative will be positive and so the function will be increasing toward the steady-state at A. This is therefore a *stable* steady-state from below. Also, if $P(t)$ is greater than A, the derivative is negative and so the function is decreasing towards the steady-state value at A – so this a *stable* steady-state from above also. However, if the initial population value is less than A, then the population will never arrive at values above A and this is the way that this

function is normally used – taking values between 0 and A. From either side, the larger the value of the growth rate c, the faster the function will approach the steady-state.

Finding the solution

The equation is separable so we have,

$$\frac{\mathrm{d}P}{\mathrm{d}t} = cP\left(1 - \frac{P}{A}\right)$$

$$\iff \quad \frac{\mathrm{d}P}{\mathrm{d}t} = \left(\frac{c}{A}\right)P(A-P)$$

$$\iff \quad \frac{1}{P(A-P)}\,\mathrm{d}P = \left(\frac{c}{A}\right)\mathrm{d}t$$

$$\iff \quad \int \frac{1}{AP} + \frac{1}{A(A-P)}\,\mathrm{d}P = \int \left(\frac{c}{A}\right)\mathrm{d}t \qquad \text{by partial fractions}$$

$$\iff \quad \int \frac{1}{P} + \frac{1}{(A-P)}\,\mathrm{d}P = \int c\,\mathrm{d}t \qquad \text{by partial fractions}$$

$$\iff \quad \ln P - \ln(A-P) = ct + D \qquad D \text{ is const. of integration}$$

$$\iff \quad \ln\left(\frac{P}{A-P}\right) = ct + D$$

$$\iff \quad \frac{P}{A-P} = Be^{ct} \qquad B = e^{D}$$

$$\iff \quad P = Be^{ct}(A-P)$$

$$\iff \quad P(1+Be^{ct}) = ABe^{ct}$$

$$\iff \quad P = A\left(\frac{Be^{ct}}{1+Be^{ct}}\right).$$

Note that:

- *Another common way to write the logistic function is to divide by Be^{ct} to get:*

$$A\left(\frac{1}{Ee^{-ct}+1}\right)$$

where $E = \frac{1}{B}$.

- *As has been noted:* $0 \leq P(t) \leq A$. *So,*

$$\begin{aligned} & 0 \leq A\left(\frac{Be^{ct}}{1+Be^{ct}}\right) \leq A \\ \iff \qquad & 0 \leq \left(\frac{Be^{ct}}{1+Be^{ct}}\right) \leq 1. \end{aligned}$$

The value,

$$\frac{Be^{ct}}{1+Be^{ct}} = \frac{1}{Ee^{-ct}+1}$$

is a proportion, a value in $[0, 1]$.

- *Since we have,*

$$P(t) = A\theta$$

where θ *is a proportion and the derivative is given by,*

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = cA\theta(1-\theta) = cA\theta - cA\theta^2$$

we can see that if θ *is small then the derivative will be approximately,*

$$cA\theta = cP(t).$$

For this reason, when θ *is small – which happens when* t *is small – the growth of the function is approximately exponential.*

(87) **Simple Harmonic Motion:** $y'' = -ky$ TODO: S.H.M.

(88) This is a second-order non-linear equation but, for small angular displacements, we can linearize this equation using the small-angle approximation for sine $\sin\theta \approx \theta$ giving:

$$\begin{aligned} & L\tfrac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -g\theta \\ \iff \qquad & \tfrac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -\frac{g}{L}\theta \end{aligned}$$

$$\iff \qquad \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{g}{L}\theta = 0.$$

The auxiliary equation of this equation is $z^2 + \frac{g}{L} = 0$ – the roots of which are clearly complex and given by,

$$z = \pm\frac{\sqrt{-4(g/L)}}{2} = \pm\sqrt{-\frac{g}{L}} = \pm\sqrt{\frac{g}{L}}i.$$

TODO: 2nd order linearized pendulum equation

4.6.10 Second-order Nonlinear ODEs

4.6.10.1 Examples of Second-order Nonlinear