Abstract Algebra

Notes missing on the things I read…

Óscar Pereira[*]

14th June 2022

[*] {https://, oscar@}randomwalk.eu.

Contents

1 | Basics

1.1 Set theory

I seem to recall having read somewhere that set theory is the starting point for all mathematics. I don't know about the rest, but it does seem appropriate enough as a foundation for *algebra*...

1.1.1 The empty set

The very special set that contains nothing, usually denoted as $\varnothing$, is contained in all others sets. The usual proof of this is as follows: suppose that there existed a set S that did *not* contain $\varnothing$; this would mean that $\varnothing$ contains at least one element that is not in S—but this contradicts the definition of $\varnothing$ being the *empty* set.

This proof is not intuitionistically valid. An alternative is as follows: given any set A, we always have that for any set B, it holds that $A \setminus B \subseteq A$. Thus if we set $B = A$, we conclude that $A \setminus A = \varnothing \subseteq A$. This is intuitionistically valid, and setting $A = \varnothing$ yields $\varnothing \subseteq \varnothing$.

1.1.2 The powerset

Given a set S, its *powerset* is defined as the set of all of its subsets, and denoted as 2^S. The next result explains the reason for the terminology.

Theorem 1.1. *Given any finite set S, we have $|2^S| = 2^{|S|}$.*

Proof. We use induction on the size of S. It is true when $S = \varnothing$, because $|\varnothing| = 0$ and $2^0 = 1$ (note that $2^\varnothing = \{\varnothing\}$). But exponentiation to zero itself is a convention, so the argument might be more convincing if we start from one: indeed, for any singleton set, $\{a\}$, its powerset has two elements: $\{\varnothing, \{a\}\}$. And accordingly, $2^1 = 2$, so we have a base case.

Now suppose that $S = \{s_1, \ldots, s_n\}$ (we tacitly assume that all the s_i distinct), and that $|2^S| = 2^{|S|} = 2^n$, and let $S' = \{s_1, \ldots, s_n, s_{n+1}\}$. Let T be the set of all the elements of 2^S, plus all the elements of 2^S, with s_{n+1} added to them. That is, $T = 2^S \cup \left\{x \cup \{s_{n+1}\} \,|x \in 2^S\right\}$. Note that the two sets have no elements in common, i.e. they are disjoint. From the fact that $2^S \subseteq 2^{S'}$, it is immediate that $T \subseteq 2^{S'}$.[1] To show the reverse containment, consider an arbitrary element of $2^{S'}$. If it contains s_{n+1}, it is in T; if not, it is in 2^S—but by construction, this means that it is also in T. Hence, $2^{S'} \subseteq T$—and also $2^{S'} = T$.

Thus we conclude that the elements of $2^{S'}$ consist of all the elements of 2^S, plus those same elements, each joined ($\cup$) with $\{s_{n+1}\}$—hence, $2^{S'}$ has twice the size of 2^S. In particular, $|2^{S'}| = 2|2^S| = 2^{|S|+1} = 2^{|S'|}$. ■

Remark 1.2. This result can also be proved via Pascal's triangle, and the binomial theorem. Indeed given a set S with n elements, the number of distinct subsets is given by:

$$\sum_{k=0}^{n} \binom{n}{k} \tag{1.1}$$

which the binomial theorem tells us equals 2^n.[2] △

1.1.3 On set difference and complementation

Definition 1.3 (Set difference). *Given two sets A and B, the* **set difference** *of A minus B, denoted $A \setminus B$, is defined as the set $\{x \in A | x \notin B\}$.*

Theorem 1.4. *Given sets A, B and C, we have that $A \setminus B = A \setminus C$ if and only if $A \cap B = A \cap C$.*

Proof. We have that $A \setminus B \cup (A \cap B) = A$ and $A \setminus C \cup (A \cap C) = A$. As in both unions the sets are disjoint, it follows that if $A \setminus B = A \setminus C$ holds, then it must be case that $A \cap B = A \cap C$. And vice-versa. ■

Theorem 1.5. *Given sets A, B and C, the following hold:*

(i) $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

(ii) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$.

Proof. If $A = \varnothing$, then (i) and (ii) reduce to $\varnothing = \varnothing \cup \varnothing$ and $\varnothing = \varnothing \cap \varnothing$, which are trivially true—so let $A \neq \varnothing$. If $B = C = \varnothing$, then (i) and (ii) reduce respectively to $A = A \cup A$ and $A = A \cap A$, which are similarly true. So let at exactly one of B, C be different from $\varnothing$. Without loss of generality, let $B = \varnothing$, and $C \neq \varnothing$. Then $B \cap C = \varnothing$, and $B \cup C = C$, and thus:

- (i) reduces to $A = A \cup (A \setminus C)$, which is always true.
- (ii) reduces to $A \setminus C = A \cap (A \setminus C)$, which again is always true.

So let *both* $B \neq \varnothing$ and $C \neq \varnothing$. Then clearly $B \cup C \neq \varnothing$, but it could still happen that $B \cap C = \varnothing$. Then (i) reduces to $A = (A \setminus B) \cup (A \setminus C)$. Clearly, $x \in (A \setminus B) \cup (A \setminus C)$ implies that $x \in A$. Conversely, if $x \in A$, then the only way for it *not* to belong to $(A \setminus B) \cup (A \setminus C)$, would be if it belonged to $B \cap C$—which is against the hypothesis that $B \cap C = \varnothing$. Thus we conclude that $A = (A \setminus B) \cup (A \setminus C)$ holds.

So let A, B and C be all different from the empty set, and furthermore let both $B \cap C$ and $B \cup C$ be also not empty. Proceeding with the demonstration:

- (i) ($\rightarrow$) Let x belong to A but not to $B \cap C$. We have three cases:
  1. $x \in B \wedge x \notin C$, in which case $x \notin A \setminus B$ and $x \in A \setminus C$;
  2. $x \notin B \wedge x \in C$, in which case $x \in A \setminus B$ and $x \notin A \setminus C$;
  3. $x \notin B \wedge x \notin C$, and thus $x \in A \setminus B$ and $x \in A \setminus C$.

  In either way we end up with $x \in (A \setminus B) \cup (A \setminus C)$, thus concluding that $A \setminus (B \cap C) \subseteq (A \setminus B) \cup (A \setminus C)$.
  ($\leftarrow$) If $x \in A \setminus B$, then it belongs to $A \setminus (B \cap C)$, for if x does not belong to B, it cannot belong to $B \cap C$. Similarly, $x \in A \setminus C \rightarrow x \in A \setminus (B \cap C)$. Hence, it is obvious that if x belongs to either $A \setminus B$ or $A \setminus C$, it also belongs to $A \setminus (B \cap C)$—i.e. $A \setminus (B \cap C) \supseteq (A \setminus B) \cup (A \setminus C)$. Both results now imply that $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

- (ii) (→) Let $x \in A$. If $x \notin B \cup C$, then clearly $x \notin B$—meaning $x \in A \setminus B$—and $x \notin C$, which entails that $x \in A \setminus C$. I.e. $A \setminus (B \cup C) \subseteq (A \setminus B) \cap (A \setminus C)$.
  (←) Again let $x \in A$. $x \notin B$ and $x \notin C$ imply that x cannot belong to $B \cup C$, and hence $x \in A \setminus (B \cup C)$. I.e. $A \setminus (B \cup C) \supseteq (A \setminus B) \cap (A \setminus C)$.
  Both results now imply that $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$.

■

The same type of proof also gives us the classical results of distributivity:

Theorem 1.6. *The operations of* **conjunction** *(∩) and* **disjunction** *(∪) are* **distributive** *in relation to one another. That is to say, the following properties hold:*

(i) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(ii) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Proof.

- (i) (→) If $x \in A$, and $x \in B$, then $x \in A \cap B$, and hence it belongs to the right hand side. Similarly, if $x \in A$, and $x \in C$, then $x \in A \cap C$, and hence it belongs to the right hand side. So the left hand side is contained in the right hand side. (←) If $x \in A \cap B$, then it belongs to the left hand side. Similarly if $x \in A \cap C$. And if both $x \in A \cap B$ and $x \in A \cap C$, then again x belongs to left hand side. Hence, the right hand side is contained in the left hand side.
  The two statements together imply that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- (ii) (→) If $x \in A$, then it clearly belongs to the right hand side. If $x \notin A$, but $x \in B \cap C$, then x again belongs to the right hand side. Obviously, x also belongs to the right hand side if both conditions hold. This shows that the left hand side is contained in the right hand side. (←) Let x be an element of the right hand side. Then either $x \in A$, or, if $x \notin A$, then it must be that $x \in B$ and $x \in C$. But in both situations, x belongs to the left hand side, showing that the right hand side is contained on the left hand side.
  The two statements together imply that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

■

Definition 1.7. *Given two sets A and B, their* **symmetric difference***, denoted $A \bigtriangleup B$, defined as follows: $A \bigtriangleup B \overset{\text{def}}{=} (A \setminus B) \cup (B \setminus A)$. Equivalently, $A \bigtriangleup B \overset{\text{def}}{=} (A \cup B) \setminus (A \cap B)$.*

The equivalence can be shown by proving that any element on the set $A \bigtriangleup B$ according to one definition must also be in the set according to the other definition, and vice-versa. From this we can see that symmetric difference is commutative. Most of these equalities are proven via the same technique, i.e. showing that any element belonging to the LHS must also belong to the RHS, and vice-versa.

The symmetric difference of sets A and B could have also been defined as the set of all elements x that verify the condition:

$$(x \in A \wedge x \notin B) \vee (x \notin A \wedge x \in B) \tag{1.2}$$

Note this disjunction is exclusive. This way of looking at symmetric difference is useful for the next result.

Theorem 1.8. *Symmetric difference is associative.*

Proof. Given sets A, B and C, we want to show that $A \triangle (B \triangle C) = (A \triangle B) \triangle C$. From the LHS of the equation we know that exactly one of the following holds:

$$\begin{cases} x \in A \wedge x \notin (B \triangle C) \\ x \notin A \wedge x \in (B \triangle C) \end{cases}$$

The curly braces represent disjunction, although in this case, the nature of the propositions ensures that both cannot hold simultaneously—i.e. the disjunction is exclusive. We can expand this still further (both equations split into two):

$$\begin{cases} x \in A \wedge x \notin B \wedge x \notin C \\ x \in A \wedge x \in B \wedge x \in C \\ x \notin A \wedge x \in B \wedge x \notin C \\ x \notin A \wedge x \notin B \wedge x \in C \end{cases}$$

The exclusivity property still holds (any two of those four equations cannot hold at the same time). As $\wedge$ distributes over $\vee$, we can group the first and third equations, and the second and fourth, as follows:

$$\begin{cases} \big[(x \in A \wedge x \notin B) \vee (x \notin A \wedge x \in B)\big] \wedge x \notin C \\ \big[(x \in A \wedge x \in B) \vee (x \notin A \wedge x \notin B)\big] \wedge x \in C \end{cases}$$

But from the definition of symmetric difference, this is precisely:

$$\begin{cases} x \in (A \triangle B) \wedge x \notin C \\ x \notin (A \triangle B) \wedge x \in C \end{cases}$$

This is equivalent to $(A \triangle B) \triangle C$, which is what was required to show. ■

Remark 1.9 ($\triangle$ as a group operation). Fun fact: for a given set X, its powerset, together with the operation of symmetric difference (SD), *forms a group*. Indeed the SD of two subsets of X is bound to also be a subset of X, so we have closure. Associativity was just dealt with, and the SD of any set with $\varnothing$ is that set itself, so $\varnothing$ is the identity. Finally, the SD of a set with itself is precisely $\varnothing$, so each element is its own inverse. $\triangle$

Definition 1.10. *Given a set A, that belong to an ambience space Ω, we define A's* ***complement*** *(or it's* ***negation****) as:*

$$\overline{A} \overset{\text{def}}{=} \Omega \setminus A \tag{1.3}$$

Remark 1.11. It is immediate from the above definition that $\overline{\varnothing} = \Omega$, and $\overline{\Omega} = \varnothing$. $\triangle$

Theorem 1.12. $\overline{\overline{A}} = A$.

Proof. Let Ω be the universe. Then $\overline{\overline{A}} \overset{\text{def}}{=} \Omega \setminus (\Omega \setminus A)$. Furthermore, we can observe that *for any set A, $A \cup (\Omega \setminus A) = \Omega$ and $A \cap (\Omega \setminus A) = \varnothing$.* So given any element of Ω it belongs to one and only one of A or $\Omega \setminus A$. Hence all the elements not in $\Omega \setminus A$—i.e. $\Omega \setminus (\Omega \setminus A)$—must be in A. This shows that $\Omega \setminus (\Omega \setminus A) \subseteq A$. For the converse direction, observe that all the elements of A do *not* belong to $\Omega \setminus A$, and hence they belong to $\Omega \setminus (\Omega \setminus A)$, i.e. $A \subseteq \Omega \setminus (\Omega \setminus A)$. Thus $A = \Omega \setminus (\Omega \setminus A)$, and the theorem follows. ■

Theorem 1.13 (de Morgan's laws). *For sets B and C, the following holds:*

(i) $\overline{B \cap C} = \overline{B} \cup \overline{C}$.

(ii) $\overline{B \cup C} = \overline{B} \cap \overline{C}$.

Proof. Follows immediately from the definition of negation (def. 1.10), and from theorem 1.5, setting $A = \Omega$. ■

For the next theorem, an auxiliary result is needed.

Lemma 1.14. *Given sets A, B, we have $A \setminus B = A \cap \overline{B}$.*

Proof. Follows directly from the definitions of set difference and complementation. ■

Theorem 1.15. $\overline{A} \triangle \overline{B} = A \triangle B$.

Proof. $\overline{A} \triangle \overline{B} = (\overline{A} \cup \overline{B}) \setminus (\overline{A} \cap \overline{B}) = \overline{A \cap B} \cap (A \cup B) = (A \cup B) \setminus (A \cap B) = A \triangle B$. ■

1.1.4 Other properties

Theorem 1.16. *Given sets A, B and C, if either $A \subseteq B$ and $B \subset C$, or $A \subset B$ and $B \subseteq C$, then $A \subset C$.*

Proof. In the first case, as $B \subset C$, there is (at least) one element in C that is not in B; and as A is contained in B, then there is also at least one element in C that is not in A—and hence, $A \subset C$.

Similarly for the second case, there an element in B that is not in A; and as B is contained in C, there is also an element in C that is not in A—and so, $A \subset C$. ■

1.2 Induction *a la* Odifreddi

First of all, it more than customary to have induction start from 0—and Odifreddi is no exception. This is for convenience, of course, but we can also start the inductive process from a number superior to 0. Odifreddi, however, while approaching induction with a different formalism from what one usually encounters, also starts from 0: and this is also natural, for he wishes to study computable functions, which are partial computable functions which domain happens to be (the whole of) $\mathbb{N}$. As an exercise, here we rewrite his formalism, but with an arbitrary (positive) starting point. First a bit of notation.

$$(\exists x \le y)\varphi(x) \overset{\text{def}}{=} \exists x\,(x \le y \land \varphi(x)) \tag{1.4}$$

$$(\forall x \le y)\varphi(x) \overset{\text{def}}{=} \forall x\,(x \le y \to \varphi(x)) \tag{1.5}$$

Note that

$$(\forall x \le y)\varphi(x) \;\equiv\; \varphi(x) \land \varphi(x+1) \land \cdots \land \varphi(y) \tag{1.6}$$

Working the first two definitions above we get (as we would intuitively expect):

$$\neg(\exists x \le y)\varphi(x) \equiv (\forall x \le y)\neg\varphi(x) \tag{1.7}$$

$$\neg(\forall x \le y)\varphi(x) \equiv (\exists x \le y)\neg\varphi(x) \tag{1.8}$$

The case for $x \ge y$ is defined similarly (and similar remarks apply to negation).

Definition 1.17 (Axiom of simple induction). *If φ is a formula with one free variable and $r \in \mathbb{N}$ is an arbitrary natural number, then*

$$\Big\{\varphi(r) \land (\forall x \ge r)\big[\varphi(x) \to \varphi(\mathcal{S}(x))\big]\Big\} \to (\forall y \ge r)\varphi(y) \tag{1.9}$$

where $\mathcal{S}$ is the successor function.

The meaning of this axiom is perhaps best conveyed if instead of the "everyday" meaning of induction, we instead interpret $a \rightarrow b$ as meaning that we cannot have a true and b false. Simple induction is then merely the statement that if for a given proposition φ, it holds for a natural number r, and we can show that it cannot be true for a natural and false for its successor, then we axiomatically believe it holds for all naturals greater or equal than r.

But what if the truth of $\varphi(x+1)$—i.e. φ of the successor of x—depends not on the truth of $\varphi(x)$ alone, but on the truth of proposition φ for a subset of the previous values? Such a subset could even not include x, or on the other extreme, include *all* values smaller than $x+1$. Consider a nonempty set S for which elements the proposition φ holds (the base cases). You might be able show that this implies that it holds for the next value—and this might follow from φ holding just for some of the previous values (i.e. only some of the elements of S). And for the value after that, maybe its truth again follows from φ holding only for *some* of the previous values. And similarly for next value, and so on. In all of these scenarios, when we get to value x, we have already established that φ holds for *all* previous values—even though in each step, we might only use the fact that φ holds some subset of those previous values (and the same might again happen trying to prove that $\varphi(x+1)$ holds).

Hence as a next step we might try to see what happens on the assumption that φ holds for all values up to and including x. If from this it can be deduced that $\varphi(x+1)$ holds, then taking as a proposition the statement that φ holds for all values up to and including x, and applying weak induction on *that* "higher level" proposition, we get the principle of *strong induction*. The simplest way to state it, is to write it like this:

Theorem 1.18. *The* **naive** *form of strong induction:*

$$\varphi(r) \wedge (\forall z \geq r)\Big[\Big\{(\forall x\colon r \leq x \leq z)\varphi(x)\Big\} \rightarrow \varphi(z+1)\Big] \rightarrow (\forall y \geq r)\varphi(y) \tag{1.10}$$

where the quantifier in the inner implication is defined as would be expected:

$$(\forall x\colon r \leq x \leq z)\varphi(x) \overset{\text{def}}{=} \forall x\,(r \leq x \leq z \rightarrow \varphi(x)) \tag{1.11}$$

Proof. It follows from simple induction, as hinted above. Let

$$\Phi'(z) = (\forall x\colon r \leq x \leq z)\varphi(x) \tag{1.12}$$

i.e. equal to the antecedent of the inner implication.[3] To prove (1.10), suppose the antecedent of the outer implication (i.e. the inner implication) holds, for all $z \geq r$. This inner implication holds if and only if $\Phi'(z) \rightarrow \Phi'(z+1)$ also holds, also for all $z \geq r$ (cf. lemma 1.19).

And the fact that $\Phi'(z) \rightarrow \Phi'(z+1)$ holds, together with the fact that $\Phi'(r)$ (i.e. $\varphi(r)$) also holds, allow us to conclude, via simple induction, that $\Phi'(z)$ holds for all $z \geq r$—which can only happen if the same is true of $\varphi(z)$. ■

Lemma 1.19. *In the proof of theorem 1.18, we have that* $\Big\{(\forall x\colon r \leq x \leq z)\varphi(x)\Big\} \rightarrow \varphi(z+1)$ *holds if and only if* $\Phi'(z) \rightarrow \Phi'(z+1)$ *also holds.*

Proof. First, note that given two implications $a \rightarrow b$ and $c \rightarrow d$, to check that one holds if and only if the other also holds, it is easier to to check that their *negations* are equivalent. That is, to check that $a \wedge \neg b$ is equivalent to $c \wedge \neg d$. So we have the two implications

$$\textbf{a)}\ (\forall x\colon r \leq x \leq z)\varphi(x) \rightarrow \varphi(z+1) \quad \text{and} \quad \textbf{b)}\ \Phi'(z) \rightarrow \Phi'(z+1)$$

$\Phi'(z)$ is equal to the antecedent of a) by definition. And if $\varphi(z+1)$ is false, then clearly $\Phi'(z+1)$ is also false. Conversely, if $\Phi'(z+1)$ is false, *while* $\Phi'(z)$ *is true*, this can only be because $\varphi(z+1)$ is false. ■

Equality of conditionals. For ease of reference below, this is the name I have to the following statement: $a \to b$ is equivalent to $c \to d$, if and only if $a \wedge \neg b$ is equivalent to $c \wedge \neg d$. Both these equivalences can be verified by checking that a is true if and only if c is true, and that b is false if and only if d is false.

But this way of laying out strong induction requires an extraneous condition, which modern mathematical aesthetics denounces as inelegant. So if it can be suppressed… suppress it (of course, this new, more "elegant" form is harder to read, as it requires parsing through the implicit parts, but well, aesthetics seldom comes without a cost…). This is why this new more "aesthetic" form of strong induction prompts a more lengthy discussion afterwords. The superscript asterisks (*) mark the changes in relation to (1.10).

Theorem 1.20 (Strong induction). *For any proposition φ with one free variable, the following holds:*

$$(\forall z \geq r)\Big[\big\{(\forall x : r \leq x <^{*} z)\varphi(x)\big\} \to \varphi(z)^{*}\Big] \to (\forall z \geq r)\varphi(z) \tag{1.13}$$

Proof. Let $\Phi(z)$ denote the antecedent of the inner implication of (1.13), i.e.:

$$\Phi(z) \overset{\text{def}}{=} (\forall x : r \leq x < z)\varphi(x) \tag{1.14}$$

(Note the difference from Φ', viz. we have here $< z$ rather than $\leq z$.) Suppose the inner implication of (1.13) holds (for all $z \geq r$). This happens if and only if $\Phi(z) \to \Phi(\mathcal{S}(z))$ also holds (idem.). Indeed, both implications have exactly the same antecedent—$\Phi(z)$ is by definition the antecedent of the inner implication of (1.13). Thus, by the equality of conditionals principle, we need only to show that, assuming $\Phi(z)$ holds, $\varphi(z)$ is false if and only if $\Phi(\mathcal{S}(z))$ is false. From the definition of Φ, we obtain:

$$\Phi(\mathcal{S}(z)) = \varphi(r) \wedge \varphi(r+1) \wedge \cdots \wedge \varphi(z-1) \wedge \varphi(z) \tag{1.15}$$

Hence, if $\varphi(z)$ is false, then so is $\Phi(\mathcal{S}(z))$. Conversely, if $\Phi(\mathcal{S}(z))$ is false, and as we are assuming that $\Phi(z) = \varphi(r) \wedge \varphi(r+1) \wedge \cdots \wedge \varphi(z-1)$ is true, it must be the case that $\varphi(z)$ is false.

So assuming that the inner implication of (1.13) always holds, then so does $\Phi(z) \to \Phi(\mathcal{S}(z))$. Furthermore, $\Phi(r)$ is always (vacuously) true, and thus, applying simple induction to Φ, it follows that $(\forall z \geq r)\,\Phi(z)$—which is the same as $(\forall z \geq r)\,\varphi(z)$. (If there existed $z \geq r$ such that $\varphi(z)$ was false, then $\Phi(\mathcal{S}(z))$ would also be false. Conversely, if there existed $z \geq r$ such that $\Phi(\mathcal{S}(z))$ was false, then there would exist $y \in \{r, r+1, \ldots, z-1\}$ for which $\varphi(y)$ would be false.) ■

Where did the base case go? In the above proof, simple induction is applied to the proposition Φ, with the base $\Phi(r)$, which is vacuously true, and hence, always true. Does this mean we can skip checking the base case? Alas, no.

$\Phi(r)$ happens when $z = r$, and in the proof above it was shown that the implication $\Phi(z) \to \Phi(\mathcal{S}(z))$ must hold also in this case. I.e., $\Phi(r) \to \Phi(r+1)$ must hold, or equivalently, $\Phi(r) \to \varphi(r)$ must hold. Now as $\Phi(r)$ is always true, stating that this last implication holds is the same as stating that $\varphi(r)$ holds. But because $\Phi(r)$ is unconditionally true, it is very unlikely that the truth of $\varphi(r)$ will follow logically from the truth of $\Phi(r)$. In other words, we need to explicitly check that $\varphi(r)$ holds—which means we still very much need to check the base case.

Now consider $z = r + 1$. This gives rise to $\Phi(r+1) \to \Phi(r+2)$, which we can rewrite as $\varphi(r) \to \big(\varphi(r) \wedge \varphi(r+1)\big)$. Now if $\varphi(r)$ is true, then $\varphi(r) \wedge \varphi(r+1)$ is false if and only if $\varphi(r+1)$ is false. Hence, by the equality of conditionals principle:

$$\varphi(r) \to \big(\varphi(r) \wedge \varphi(r+1)\big) \Leftrightarrow \varphi(r) \to \varphi(r+1) \tag{1.16}$$

So the question now arises: do we need to check $\varphi(r+1)$? Well, it depends. It might be that $\varphi(r+1)$ follows logically from $\varphi(r)$—in which we do *not* need to check it explicitly. But it may also happen that it does *not* follow logically from $\varphi(r)$—in which case we do need to check $\varphi(r+1)$ explicitly, in addition to $\varphi(r)$. I.e., we have two base cases.[4]

Remark 1.21 (Defining the base case by convention). In [3, §1.1], the fact that *all nonzero* integers can be written (possibly non-uniquely) as a product of primes is proved as follows: if the integer is negative we can multiply by -1, and so we can work only with positive integers. Now by convention we set that a product of zero factors is 1. So 1 is clearly a product of (zero) primes. Now assume that all integers less than n can be written as a product of primes. If n is prime, is is the product of one prime. If it is not prime, we've covered the case $n=1$, so assume $n \neq 1$. Then it can be written as the product of at least two numbers, each smaller than n; by the induction hypothesis, these can be written as a product of primes—and thus so can n.

If the reader is left with the feeling that this involved a bit of cheating, I sympathise. One is indeed left with the feeling that the induction functions *because* of the convention establishing the base case. Actually however, it is the reverse: the base ($n=1$) is set—by convention—to such a value that, applying the inductive process to it, leads to the conclusion that the proposition is also true for the next value, $n=2$, which would be our "more natural" starting point anyway. Indeed, if we start instead with $n=1$, and move the next case $n=2$, well 2 is either prime (which is indeed the case), or, if 2 was not a prime, that would mean it could be written as product of 1's, which would in turn mean that 2 was also the result of the product of zero primes. Which is of course absurd, but goes on to show that the inductive reasoning is nonetheless valid: if 1 were a product of zero primes, then 2 would have also to either be prime or be the product of zero primes. And then the same reasoning can be done for $n=3$, because all numbers before it are products of (possibly zero) primes; and so on…

So in general it seems the pattern is always the same, in these sort of contrived examples: set by convention a starting point that actually leads (via the inductive process) to the more natural starting point—and from there onwards, it's "induction as usual". △

1.2.1 The Well-Ordering Principle

The above discussion shows that weak induction implies strong induction. The converse, which also holds, is usually shown by showing first that strong induction implies the *well-ordering principle* (see below), and then that the WOP implies weak induction (idem.). Note that this means that weak induction, strong induction, and the WOP are all equivalent, in some sense (again, more on this below).

Definition 1.22 (well-ordering principle (WOP)). *Let φ be a proposition, that is true for at least an element x. Then there exists an element y which is the* **smallest** *element for which φ is true.*

Remark 1.23. Given any set S, we can always define a function, sometimes called that set's *characteristic function*, that is true for any element that belongs to S, and false otherwise. Conversely, propositions that are either true or false—or more generally, functions with a binary output—implicitly define a set consisting of the elements for which they are true (and another one for those for which they are false). Hence, the WOP as stated above is equivalent to saying that every nonempty set has a smallest element. △

We can take an informal shortcut to argue that strong induction implies weak induction. Assuming we have a set of base case(s) already verified, and letting r be the smallest of said cases, strong induction means that if we never observe $\varphi(r), \varphi(r+1), \ldots, \varphi(z)$ being true, and

$\varphi(z+1)$ being false (for arbitrary r, z, with $z \geq r$), then we believe that φ is true for all $z \geq r$. Then, making $z = r$, this means we will also never observe $\varphi(z)$ being true and $\varphi(z+1)$ being false (again for an arbitrary $z \geq r$)—and hence we can (informally) say that this latter condition also warrants the belief that φ is true for $z \geq r$. But this is precisely weak induction!

On the meaning of saying that weak and strong induction, and WOP, are equivalent. Before showing that strong induction leads to the WOP, which in turn leads to weak induction, it is worth to pause for a moment and ask: what does it *mean* to say that weak induction implies strong induction? Or that strong induction implies the WOP, or that the WOP implies weak induction? Certainly any proposition that can be proved through weak induction can be also be proven from strong induction, but the converse is clearly false, because we assume more with strong induction. My view is that both weak and strong induction, as well as the WOP, are equivalent in the following sense: if we choose any one of them as an axiom, we can prove the other two forms, and thus "reach" (i.e. prove to be true) the same set of propositions. Thus, in the axiomatic model—where you have to start from somewhere—weak induction, strong induction and the WOP are all *equivalent starting points*.

From strong induction to the WOP. The intuition here is actually quite simple: suppose there is a set of natural numbers that has no smallest element. Then it clearly cannot contain 0, as it is the smallest element of $\mathbb{N}$. But then, it cannot contain 1 either, because then *it* would be the smallest element. And it also cannot contain 2, and so on... Thus, by the principle of (strong) induction, the original set must be empty—and this is the essence of the WOP: every **nonempty** set of naturals must have a smallest element.

Reasoning formally, we need only to take the contrapositive of (1.13):

$$(\exists y \geq r)\neg\varphi(y) \rightarrow (\exists z \geq r)[(\forall r \leq x < z)\varphi(x) \wedge \neg\varphi(z)] \tag{1.17}$$

Now write ψ for $\neg\varphi$; as φ is an arbitrary proposition, so is ψ. Thus we get:

$$(\exists y \geq r)\psi(y) \rightarrow (\exists z \geq r)[(\forall r \leq x < z)\neg\psi(x) \wedge \psi(z)] \tag{1.18}$$

This says that for any proposition ψ, if it holds for some value greater or equal to r, then there exists (in the same range) a *smallest* value z for which it also holds (which is not necessarily equal to r). Now in practice, when applying induction, it is customary to set r to the smallest element for which the proposition in question holds—the eponymous base case. But this needn't be so: formally, both (1.9) and (1.13)—hence also (1.17) and (1.18)—are *tautologies*, i.e., they are true for an arbitrary r (and indeed, an arbitrary φ). And as we are looking at smallest elements, we can set $r = 0$ and we get precisely the good old WOP: if ψ is true for some natural number y, then there exists a natural z which is the smallest element for which ψ is true.[5] Formally, we get the **well-ordering principle**:

$$(\exists y \geq 0)\psi(y) \rightarrow (\exists z \geq 0)[(\forall 0 \leq x < z)\neg\psi(x) \wedge \psi(z)] \tag{1.19}$$

Excursus: from weal induction to the WOP?! What would happen if we took the contrapositive to weak induction? The WOP involves a sort of "broad view"—the proposition holds for one element (the minimum) *and for no other below that minimum*—and so does *strong* induction, particularly when starting from 0: if a proposition holds for *all* elements below a given one, then it holds for that one as well. The result of taking the contrapositive of the weak induction principle (cf. (1.9), replacing $\neg\varphi$ with ψ) is coherent with this relation: you get that statement that if a proposition holds for any element greater than r, then either it does *not* hold for the smallest element (r), or there exists some other element x, greater than r, such that the proposition holds for x, but not for its predecessor (instead of it holding for x, and for *none* of its predecessors, as in the WOP). This is a weaker form of the WOP—and in both it and weak induction, the "broad view" from above is replaced with a more "localised" one.

From the WOP to weak induction. The last remaining bit to show to come full-circle—i.e. to show that weak induction, its strong(er) counterpart, and the WOP are equivalent, as discussed above—is to show that the WOP implies weak induction. By way of deriving a contradiction, suppose that the WOP is true, but that weak induction is *false*. That is, that there is a proposition ξ that is true for a natural r, and that holds for $n+1$ whenever it also holds for n, but that is *not* true for all naturals. Let S then be the set of elements for which ξ is false. If such a set is not empty, then it must have a minimal element, according to the WOP. Let m be that element. As the base case is r, we must have $m > r$; but as m is the smallest element for which ξ is false, it must be true for $m - 1$. But then, the assumed hypothesis implies that it must also be true for m—a contradiction! Thus the set S is indeed empty, and the principle of weak induction holds.

I end this section noting that induction does not apply only to the integers. In fact, it can be applied to other structures, as long as a special type of ordering relation—called a *well-order*—can be defined. It is the fact that the integers have such an ordering relation that allows the WOP to be meaningfully defined. This is the subject of the next section.

1.3 Binary Relations

Consider a set S; a *binary relation* is a subset of $S \times S$. For $a, b \in S$, if the ordered pair (a, b) belongs to that subset, it is denoted aRb. A relation can be seen as a generalisation of functions. Things get interesting when we impose some structure on those subsets.

To start, consider the following definition:

Definition 1.24. *A binary relation on a set S is called an* **order relation** *if, for all $a, b, c \in S$, it satisfies the following conditions:*

1. **Reflexivity**: *aRa;*
2. **Transitivity**: *$aRb \wedge bRc \Rightarrow aRc$;*
3. **Anti-symmetry**: *$aRb \wedge bRa \Rightarrow a = b$.*

Given any order relation R, we can define its **strict** version (denoted R^*), like so: $aR^*b \overset{\text{def}}{=} aRb \wedge a \neq b$. Conversely, we can define R from R^*: $R \overset{\text{def}}{=} aR^*b \vee a = b$ (see the appendix XXX).

The notion of strictness leads us to the notion of **totality**:

Definition 1.25. *An order relation R is said to be a* **total order** *if one, and only one, of the following holds: $a = b$, aR^*b, or bR^*a. Otherwise R is said to be a* **partial order**.

Informally, what the definition means is that any two elements are related. The criterion of totality requires an exclusive or, however, it can formalised without it, like so: $a = b \vee aR^*b \vee bR^*a$. This is because when one of those three conditions is true, the remaining two are false. To see this, first note that $aR^*b \to aRb$. Now:

- If $a = b$, it follows from the definition of R^* that aR^*b, and bR^*a are false.
- If aR^*b is true, then it again follows from the definition of R^* that $a = b$ is false. It just remains to be shown that bR^*a must also be false. This is done by contradiction, by supposing it was not. That is, suppose that in addition to aR^*b being true, bR^*a was also

true. As "$R^* \to R$" (see above), this means that both aRb and bRa are true—and from the anti-symmetry of R, it follows that $a = b$, which is contradictory. Hence, bR^*a must be false.

- Finally, if bR^*a is false, by a similar argument to that done for aR^*b above, it follows that both $a = b$ and aR^*b must be false.

So exclusive or is unneeded. However, note that all three conditions can be false—which means R is a partial, rather than a total, order.

Theorem 1.26 (Totality of non-strict order.). *A non-strict order relation R is total if and only if $aRb \vee bRa$ holds.*

Proof. We need to show only that $aRb \vee bRa$ is true if and only if $a = b \vee aR^*b \vee bR^*a$ is true. (→) If both aRb and bRa hold, then by the anti-symmetry of R, it follows that $a = b$. If aRb holds but bRa does not, then it must be the case that $a \neq b$—otherwise both would hold. But this means that aR^*b holds. If bRa holds but aRb does not, a similar argument shows that bR^*a holds.

(←) If $a = b$, then both aRb and bRa hold. If aR^*b holds, then so does aRb. Similarly for bR^*a and bRa. ■

Pre-orders. We can weaken the definition of an order relation, by allowing some elements to be "equivalent" to others. That is, we allow the possibility of having *different* elements a and b such that aRb and bRa hold. In other words, we ditch anti-symmetry:

Definition 1.27. *A binary relation that is reflexive, transitive, and* **not anti-symmetric** *is called a* **pre-order**.

Remark 1.28. Note that being non-anti-symmetric is not the same as being symmetric—on which, much more in a moment XXX. △

Take an element a, and consider the set of elements b such that aRb and bRa both hold. We shall take all the elements in this set to be "equivalent," in some sense.

Definition 1.29. *Given a pre-order R, and an element a, the set $[a] \overset{\text{def}}{=} \{x\colon aRx \wedge xRa\}$ is the* **equivalence class** *of the element a.*

Theorem 1.30. *Let R be a pre-order on a set S. The equivalence classes it induces on S are* **pairwise disjoint***, and moreover, they form a* **partition** *of S.*

Proof. Consider two distinct elements x, y of S, and suppose that their respective equivalence classes, $[x]$ and $[y]$, have a common element z. Then $[x] = [y]$: we have xRz and zRy, and so xRy; but we also have yRz and zRx, and so yRx also holds. Now let $x' \in [x]$. Then we have $x'Rx$ and xRy, and from the transitivity of R comes that $x'Ry$ holds—or equivalently, $x' \in [y]$, or still $[x] \subseteq [y]$. Now let $y' \in [y]$. Similarly we have $y'Ry$ and yRx, yielding $y'Rx$, which means that $y' \in [x]$, and so shows that $[y] \subseteq [x]$. This shows that $[x] = [y]$.

So if $[x]$ and $[y]$ are not disjoint (i.e., have common elements), then they are equal. This means that if they are not equal, then they must be disjoint. Thus the set of all equivalence classes must be pairwise disjoint—for if any two equivalence classes had a common element, they would be one and the same equivalence class.

Finally, to see that the set of equivalence classes forms a partition of S, it is enough to note that any element $a \in S$ is in an equivalence class—namely, its own, $[a]$. ■

Pre-orders retain the notion of *ordering*, except that now what is ordered are not the elements of S, but rather *their equivalence classes*. This is because in a pre-order R we can have elements x and y such that xRy holds, but yRx does *not*. Let $[R]$ be a binary relation *between equivalence classes* such that $[x][R][y]$ holds if and only if xRy holds (yRx may or may not hold). I will show that $[R]$ is an order relation. Before that, however, it must be shown that this binary relation is **well-defined**, in the sense that it does not depend on the chosen class representatives: given any $x' \in [x]$ and $y' \in [y]$, it must be shown that $[x][R][y]$ holds if and only if $[x'][R][y']$ holds. We use the transitivity or R: (→) we have $x'Rx$ and xRy, from which comes $x'Ry$. But also yRy', which yields $x'Ry'$—meaning that $[x'][R][y']$ holds; (←) The sequence is similar: xRx' and $x'Ry'$ both hold, yielding xRy', which in conjunction with $y'Ry$ yields xRy—and thus, $[x][R][y]$.

Now to show that $[R]$ is an order relation. Reflexivity is obvious. For transitivity, consider three equivalence classes, $[x]$, $[y]$ and $[z]$, such that $[x][R][y]$ and $[y][R][z]$ hold. Then xRy and yRz hold, and via the transitivity of R, comes xRz, which implies $[x][R][z]$. For anti-symmetry, suppose that $[x][R][y]$ and $[y][R][x]$ both hold. This means that both xRy and yRx hold—but R is *not* anti-symmetric, so we cannot conclude that $x = y$. But let $x' \in [x]$. Then we have $x'Rx$ and xRy, and as R *is* transitive, we conclude that $x'Ry$ holds—or equivalently, $x' \in [y]$, or still $[x] \subseteq [y]$. Now let $y' \in [y]$. Similarly we have $y'Ry$ and yRx, yielding $y'Rx$, which means that $y' \in [x]$, and so shows that $[y] \subseteq [x]$. Thus we conclude that $[x] = [y]$.

The totality of $[R]$ implies the totality of R. So we have shown that $[R]$ is an order relation and thus, by theorem 1.26, it is a total order if and only if $[x][R][y] \vee [y][R][x]$ holds, for any $x, y \in S$.[6] By the anti-symmetry of $[R]$, if both conditions hold, then $[x] = [y]$—so first suppose that only $[x][R][y]$ holds. Then xRy holds, by definition of $[R]$. Similarly, if $[y][R][x]$ that holds, then so does yRx. (Recall that $[R]$ is well-defined, and so it does not depend on the particular representatives of the equivalence classes; cf. above.) Either way, $xRy \vee yRx$ always holds—and this suggests generalising the above definition of totality (def. 1.25 and thm. 1.26, which were just for order relations), to pre-orders as well:

Definition 1.31. *A pre-order R is said to be* **total** *if $xRy \vee yRx$ always holds. Otherwise it is a* **partial pre-order**.

Remark 1.32. We could define a strict counterpart R^* of pre-order R, but we **cannot** define totality as in def. 1.25 (one and only one of $a = b$, aR^*b, bR^*a holding), because as R is not anti-symmetric, it is possible to have $a \neq b$ such that aR^*b and bR^*a both hold. And besides, such a definition could not be formalised as $a = b \vee aR^*b \vee bR^*a$—and exclusive-or would be required. I prefer to avoid such complications—and at any rate, def. 1.31 captures well the intuition of totality, viz., of having any two elements related to one another. △

Equivalence relations. Consider a nonempty set S, and a partition thereof: $S = \bigcup_i S_i$, with $S_i \cap S_j = \varnothing$ for $i \neq j$. Suppose further that each of the S_i have more than one element. Consider a binary relation over the set S defined as follows: xRy if and only if there exists i such that $x, y \in S_i$. This relation is a pre-order: it is clearly reflexive and transitive, and also not anti-symmetric, because all the S_i have at least two distinct elements, so there are different a, b verifying aRb and bRa. But more can be said.

In fact, it is immediate to check that $xRy \Rightarrow yRx$—and thus, the S_i are the equivalence classes induced by R on S. Moreover, this pre-order is not total: if $x \in S_i$ and $y \in S_j$ and $i \neq j$, then neither xRy nor yRx hold. So we have equivalence classes, but they are no longer ordered—they are all "equal," in a sense. This is an *equivalence relation*:

Definition 1.33. *An* **equivalence relation** *on a set S is a binary relation which is* **reflexive** *($\forall a \in S : aRa$),* **transitive** *($\forall a, b, c \in S : aRb \wedge bRc \Rightarrow aRc$) and* **symmetric***: $\forall a, b \in S : aRb \Rightarrow bRa$.*

A final thought. We have gone from order relations to equivalence relations. We can also think of the inverse direction: to go from equivalence relations to pre-orders, we order the equivalence classes. And to go from pre-orders to orders, we shrink the equivalence classes so that they are left with exactly one element each.

1.4 Binary Relations

Consider a set S; a *binary relation* is a subset of $S \times S$. For $a, b \in S$, if the ordered pair (a, b) belongs to that subset, it is denoted aRb. A relation can be seen as a generalisation of functions. Things get interesting when we impose some structure on those subsets.

Definition 1.34. *A binary relation on a set S is called an* **order relation** *if, for all $a, b, c \in S$, it satisfies the following conditions:*

1. **Reflexivity***: aRa;*
2. **Transitivity***: $aRb \wedge bRc \Rightarrow aRc$;*
3. **Anti-symmetry***: $aRb \wedge bRa \Rightarrow a = b$.*

Note that nothing is being said about *totality*; this will be the topic of §1.4.1. We can, in fact, in an order relation, have two elements a and b such that neither aRb nor bRa holds—in this case the order is said to be a *partial order*. For example, the integers ordered by the divisibility relation, $|$, is a partial order (as we can have two integers such that one is neither a multiple nor a divisor of the other, and vice-versa; e.g. 2 and 5). If any one element is related to every other element, then it is a *total order*.

Given a binary relation, if for two elements a and b, both aRb and bRa hold, this can be seen inducing a notion of "equality" between those elements, in some sense. With an order relation, that equality notion coincides with strict equality (because aRb and bRa can only happen if $a = b$). Thus, we indeed get a notion of *order*: as only one of aRb and bRa happen, then this is an absolute difference, i.e., if $a \neq b$ we can say that one of them is greater than the other.

We can go to the other extreme, and make R coincide with (what we shall mean by) equivalence between two elements. That is, we *loosen* the notion of equality, so that aRb means that (from the "point of view" of R) a is equivalent to b. Note that for this to be meaningful aRb must imply bRa; in indeed we have:

Definition 1.35. *An* **equivalence relation** *on a set S is a binary relation which is reflexive, transitive and* **symmetric***: $\forall a, b \in S : aRb \Rightarrow bRa$.*

Observe that with equivalence relations, a notion of order is impossible: indeed, as whenever aRb happens, so does bRa, there is never an absolute difference of the kind described above. Somewhere between these two extremes lie *pre-orders*, in which said absolute difference can exist—i.e. there exist a, b such that $a \neq b$, and only one of aRb or bRa happen. But there can also exist c, d with $c \neq d$, and where both cRd and dRc hold—meaning the binary relation is *not* anti-symmetric:

Definition 1.36. *An order relation which is reflexive and transitive is called a* **pre-order***.*

Very informally, pre-orders can be thought of as equivalence relations where the "equivalence classes" are smaller, and there is a loose ordering of such "classes." An order relation is then a degenerate case of a pre-order, where all the "equivalence classes" contain just one element; and an equivalence relation is a pre-order degenerated in the "other direction", i.e. the equivalence class of an element contains all other elements with which it is related.

To show that the hierarchy of "equivalence classes" in a pre-order is independent of the chosen representatives. Suppose that, for a given pre-order, only aRb holds (and bRa does not). Then given any element c such that cRa and aRc hold (i.e. it belongs to the "equivalence class" of a), and given any element d such that dRb and bRd hold ("equivalence class" of b), then we always have that only cRd holds (and not dRc). In words, if a is "smaller" than b, then any element in the "equivalence class" of a is "smaller" than any element in the "equivalence class" of b. This is consistent with the idea of an hierarchy of "equivalence classes," as outlined above.

Indeed, for element c, from transitivity we get that $cRa \wedge aRb \to cRb$. But bRc cannot hold, otherwise, again from transitivity, we would get $bRc \wedge cRa \to bRa$, which by hypothesis does not hold. And similarly for element d, from transitivity we get that $aRb \wedge bRd \to aRd$. But dRa cannot hold, for then we would get $bRd \wedge dRa \to bRa$, which again is against the hypothesis.

Finally, applying transitivity one last time, we get $cRb \wedge bRd \to cRd$; and the same conclusion also comes from $cRa \wedge aRd \to cRd$. But dRc cannot hold, otherwise we would get (for instance) $dRc \wedge cRa \to dRa$, which we shown above to be impossible. Thus c is "smaller" than d, and as both are arbitrary elements, our hierarchy of "equivalence classes" does not depend on the chosen representative.

Remark 1.37 (To be or not to be (strict)). Binary relations are said to be **strict**, or **irreflexive**, if $\forall x, \neg xRx$. However we will usually assume that R *is reflexive* (unless otherwise noticed), and denote its irreflexive counterpart by R^*. That is aR^*b is used to mean $aRb \wedge a \neq b$, or equivalently, aRb is used to mean $aR^*b \vee a = b$. The same distinction holds for orderings, and henceforth we will use the notation $(S, \preccurlyeq)$ to state that set S is ordered by $\preccurlyeq$, which is not strict (i.e. it is reflexive). Its strict counterpart is $a \prec b$, which means $a \preccurlyeq b \wedge a \neq b$ (or equivalently, $a \preccurlyeq b$ means $a \prec b \vee a = b$). I use this new notation henceforth when dealing explicitly with orders.[*] △

[*] For the formalism aficionado, we can redo the reasoning above in a more formal (and arguably more obscure) way. For example to show that, starting from the definition of aR^*b ($\overset{\text{def}}{=} aRb \wedge a \neq b$), we get $aRb \equiv aR^*b \vee a = b$, we can do:

$$\begin{aligned} & aR^*b \equiv aRb \wedge a \neq b \\ \Leftrightarrow\ & aR^*b \vee a = b \equiv (aRb \wedge a \neq b) \vee a = b \\ \Leftrightarrow\ & aR^*b \vee a = b \equiv (aRb \vee a = b) \wedge \underbrace{(a \neq b \vee a = b)}_{\text{always} = 1} \\ \Leftrightarrow\ & aR^*b \vee a = b \equiv aRb \vee a = b \end{aligned}$$

Whenever $a = b$ is true, aRb is also true. Hence $aRb \vee a = b$ has the same truth value as aRb—the only way for the truth values to be different, would be if aRb was false and $a = b$ was true, which is impossible (as R is non-strict). And so we conclude that

$$aR^*b \vee a = b \equiv aRb$$

A similar reasoning can be used to show that $aRb \wedge a \neq b \equiv aR^*b$ (starting from the definition of aRb, viz. that $aRb \overset{\text{def}}{=} aR^*b \vee a = b$).

Note that the negation of reflexivity is *not* irreflexivity. Indeed, we have

$$\neg(\forall a, aRa) = \exists a\colon \neg aRa \tag{1.20}$$

which is very different from the condition of irreflexivity, viz. $\forall a, \neg aRa$. Similarly, the negation of anti-symmetry is not symmetry, nor vice-versa. But if the condition for symmetry "always fails", in the sense that if we know that aRb holds, then we also that bRa does not, then we get the notion of *asymmetry*.

Definition 1.38. *A binary relation R is said to be* **asymmetrical** *if $aRb \to \neg bRa$, for all a, b.*

In line with what is said above, the negation of asymmetry—$\exists a, b\colon aRb \wedge bRa$—is neither symmetry nor anti-symmetry.

Lemma 1.39. *Asymmetry holds if and only if both irreflexivity and anti-symmetry hold.*

Proof. ($\to$) If R is asymmetrical it is irreflexive, because the only way the condition $\forall a, aRa \to \neg aRa$ is true, is if aRa is false for all a, which is exactly irreflexivity.

Furthermore, asymmetry also means the antecedent of the anti-symmetry condition is always false, hence anti-symmetry holds vacuously.

($\leftarrow$) Rewrite anti-symmetry as $a \neq b \to (\neg aRb \vee \neg bRa)$. From irreflexivity we know that if aRb, then $a \neq b$, and from the new form of anti-symmetry, we conclude that $\neg aRb \vee \neg bRa$ must hold. But as we have assumed aRb, then it must be the case that $\neg bRa$. Hence $aRb \to \neg bRa$, i.e. asymmetry. ■

An irreflexive relation can be symmetric; in this case the symmetry condition for when $a = b$, $aRa \to aRa$ will always be true, because both antecedent and consequent will be false. **An irreflexive relation can also be anti-symmetric**; in this case the anti-symmetry condition for when $a = b$, $aRa \wedge aRa \to a = a$, will be trivially true, because the antecedent will always be false. Also in this case, by lemma 1.39, such a (strict) relation will also be asymmetrical.

If anti-symmetry fails for all pairs, a restricted form of symmetry ensues. To see how, it is convenient to redefine that condition of anti-symmetry, in the following manner:

$$\begin{aligned}
&(aRb \wedge bRa) \Rightarrow a = b \\
\equiv\ &(\neg aRb \vee \neg bRa) \vee a = b \\
\equiv\ &(\neg aRb \vee a = b) \vee \neg bRa \\
\equiv\ &\neg(aRb \wedge a \neq b) \vee \neg bRa \\
\equiv\ &(aRb \wedge a \neq b) \Rightarrow \neg bRa
\end{aligned}$$

Via a similar reasoning to what was done above, if anti-symmetry fails for all pairs, in the sense that the implication $(aRb \wedge a \neq b) \to bRa$ now becomes true, this constitutes a restricted form of symmetry, viz. symmetry minus reflexivity (whereas regular symmetry can be or reflexive).

Transitivity. We have the following lemma.

Lemma 1.40. *Given a transitive relation, it is irreflexive if and only if it is asymmetrical.*

Proof. ($\to$) Set $a = c$ in the defining condition of transitivity (cf. definition 1.34); we get $aRb \wedge bRa \Rightarrow aRa$. If the relation is irreflexive, the consequent of this implication is false for all a; hence for it to be true (because the relation is transitive), the antecedent must also always be false. I.e. if, for instance, aRb is true, then bRa must be false. This is precisely the definition of asymmetry.

($\leftarrow$) If a relation is asymmetrical, then lemma 1.39 immediately shows that it must also be irreflexive. ■

We also have the two following results (where R denotes a reflexive relation, and R^* its strict counterpart).

If R^* is transitive, then so is R. As R^* is transitive, this means that for all a, b, c, $aR^*b \wedge bR^*c \to aR^*c$ holds. As whenever the antecedent is false, the implication is true, the only way to falsify that condition, when moving from R^* to R, is to set $a = c$, to see if we can falsify the consequent.

Remark 1.41 (Transitivity and (ir)reflexivity). If a non-strict relation R is transitive, then R^* needn't also be so: consider for example the relation R consisting of aRb, bRa, aRa and bRb. It is clearly transitive, however its strict counterpart, R^*, is not: $aRb \wedge bRa \to aRa$, and yet neither aRa or bRb are a part of R^*.[7]

However, if R^* is transitive, so is R. For reductio, suppose R is not transitive; then there must exist a, b and c such that $aRb \wedge bRc \wedge \neg aRc$ holds. Now as R is reflexive, $\neg aRc$ means $a \neq c$. But if $a = b$ holds, then the previous conjunction becomes $aRa \wedge aRc \wedge \neg aRc$, which is always false. Similarly, if we had $b = c$, said conjunction would instead become $aRc \wedge cRc \wedge \neg aRc$, which is also impossible. Hence a, b and c are all different—but this means that $aR^*b \wedge bR^*c \wedge \neg aR^*c$ also holds, meaning that R^* is not transitive, which is a contradiction. △

1.4.1 Totality and wellness

Definition 1.42. *A* **total relation** *is a relation where, for all $a, b \in S$, either aRb or bRa (or both).*

The property of totality means that any element is related to every other element. It implies reflexivity (the converse is of course false).

We can define the following two other properties for generic binary relations:

Definition 1.43. *A relation is said to be* **trichotomous** *if exactly one of the following holds: xR^*y, or yR^*x, or $x = y$.*

When we say that a non-strict relation R is trichotomous, it is understood that the trichotomy property applies to its strict version R^*.

Lemma 1.44. *Totality and anti-symmetry hold if and only if trichotomy holds.*

Proof. ($\to$) We reason by cases (regarding totality):

1. xRy and yRx true: by anti-symmetry $x = y$, and (by definition of R^*) both xR^*y and yR^*x are false.
2. xRy true and yRx false: $xRy \vee yRx \iff (xR^*y \vee x = y) \vee (yR^*x \vee x = y)$; if yRx is false that means $x \neq y$ and $\neg yR^*x$, which yields that only xR^*y is true.
3. xRy false and yRx true: identical to previous, concluding that only yR^*x is true.

($\leftarrow$) Conversely, trichotomy implies both totality and anti-symmetry:

1. If only xR^*y is true, then xRy, so totality holds, and anti-symmetry holds because, as xR^*y together with trichotomy implies $\neg yRb$, the antecedent (of the anti-symmetry condition) is false.
2. If only yR^*x is true, then the reasoning is similar.
3. If only $x = y$ is true, then totality holds (both conditions are true), and for anti-symmetry, both antecedent and consequent are true, so the implication is true as well.

■

Trichotomy implies irreflexivity $\forall x, \neg xRx$. Also, strong trichotomous relations obviously cannot be symmetric.

After this brief foray, we delve again into order relations.

Definition 1.45. *A* **total order** *is an order relation where, for all $a, b \in S$, either $a \preccurlyeq b$ or $b \preccurlyeq a$ (or both).*

An order relation which is not total is a *partial order*.

Definition 1.46 (Well-order). *A total order $\prec$ with the additional property that for any nonempty subset A of S there exists $a \in S$ such that for all $x \in S$ we have: $a \preccurlyeq x$ is called a* **well-order** *relation.*

A set together with a well-order relation is said to be a *well-ordered set*. Knuth XXX defines a well-order as a binary relation that, besides the "smallest element" property above, it is also transitive, and trichotomic. This latter condition implies totality (which implies reflexivity) and anti-symmetry—and so, we have all conditions of definition 1.46. For completeness, we restate Knuth's definition.

Definition 1.47 (Well-order a la Knuth). *A* **well-order** *is a relation $\prec$ that is transitive, trichotomic, and such that for any subset A of S there exists $a \in S$ such that for all $x \in S$ we have: $a \preccurlyeq x$.*

Example. We can construct a well-order for the set of *all* integers: let $x \prec y$ be a binary relation such that $|x| < |y| \vee (|x| = |y| \wedge x < 0 < y)$. Let us show that it is indeed a well-order; we use Knuth's definition.

Minimal element: for any integer x, we always have $0 \preccurlyeq x$.

Trichotomy: we want to show that exactly one of $x \prec y$, $y \prec x$ or $x = y$ holds. If $x = y$ it is clear that neither $x \prec y$ or $y \prec x$ can hold.

If $x \prec y$, then from the definition for $\prec$ it is clear that we cannot have $x = y$. Now if $x \prec y$, we have two cases: a) $|x| < |y|$, which entails that $|y| < |x|$ and $|x| = |y|$ are both *false*—which means $y \prec x$ cannot hold; b) $|x| = |y| \wedge x < 0 < y$, which implies that $|y| < |x|$ and $y < 0 < x$ are both *false*—again entailing that $y \prec x$ cannot hold.

If $y \prec x$, a similar reasoning shows that neither $x \prec y$ nor $x = y$ can hold. This establishes trichotomy.

Transitivity: we want to show that if $x \prec y$ and $y \prec z$ hold, then so must $x \prec z$. We must again break this down by cases.

- If $|x| < |y|$ and $|y| < |z|$ hold, then so does $|x| < |z|$ which means $x \prec z$.
- If $|x| < |y|$ and $|y| = |z| \wedge y < 0 < z$ hold, then so does $|x| < |z|$, which gives us $x \prec z$.
- If $|x| = |y| \wedge x < 0 < y$ and $|y| < |z|$ hold, then so does $|x| < |z|$, which gives us $x \prec z$.
- The missing case is when $|x| = |y| \wedge x < 0 < y$ and $|y| = |z| \wedge y < 0 < z$ hold, but this cannot happen, for y cannot be simultaneously greater and smaller than zero.

So the integers ordered by $\prec$ would be $0, -1, 1, -2, 2, -3, 3, \ldots$.

1.4.2 Induction, revisited

Writing condition 1.47 as an implication (the antecedent of which is another implication), and taking the contrapositive, we get the strong induction condition. Which leads to the question of why do we need transitivity and trichotomy, if it seems we get induction from the least element principle alone?

I think what this means is that well-ordering and induction are fundamentally equivalent, *irrespective of the ordering one chooses to use*. **But** the assumption (axiom) that allows us to extrapolate that from conditions for induction (grosso modo, $a_0 \wedge a_i \rightarrow a_{i+1}$), to the conclusion that the proposition in question holds for **all** elements of the domain, *only makes sense if the order relation is a well-order*. This is (somewhat...) easier to see if we picture induction “mechanically”, like dominoes falling in sequence: transitivity ensures there are no cycles, and trichotomy ensures that all elements can *eventually* be reached (if trichotomy was false, then there could exist two distinct elements related to each other, but not related to any other; if neither of these is the starting point of induction, then they would be unreachable).

1.4.3 Associativity

Given a binary operation $\cdot$ that is *associative*, the expression $a_1 a_2 \dots a_n$ is always well-defined, regardless of how one chooses to group the terms when evaluating it. This is shown by induction, where the definition of associativity—$a_1(a_2 a_3) = (a_1 a_2)a_3$—together with the previous, trivial cases of one and two terms, serves as the base case.

We begin with an auxiliary result. Consider a specific parenthesis disposition, defined inductively as $\prod_{i=1}^{n} a_i = \left(\prod_{i=1}^{n-1} a_i\right) a_n$. Applied to five terms for example, this would give $(((ab)c)d)e$; i.e. we associate left to right. Note that although associativity is only meaningfully defined for $n \geq 3$, $\prod$ is well-defined also for $n = 1, 2$. However, as stated above, the relevant base case is $n = 3$, which must always be explicitly verified.

Lemma 1.48. *Let m, n be positive integers. We have:*

$$\left(\prod_{i=1}^{m} a_i\right)\left(\prod_{i=m+1}^{m+n} a_i\right) = \left(\prod_{i=1}^{m+n} a_i\right) \tag{1.21}$$

For example, $((ab)c)((de)f) = ((((ab)c)d)e)f$.

Proof. The result is trivial to verify for $n = 1, 2$. For $n = 3$ we have:

$$\begin{aligned} &\left(\prod_{i=1}^{m} a_i\right)((a_{m+1}a_{m+2})a_{m+3}) \\ =&\left[\left(\prod_{i=1}^{m} a_i\right)(a_{m+1}a_{m+2})\right] a_{m+3} \\ =&\left(\prod_{i=1}^{m+2} a_i\right) a_{m+3} = \left(\prod_{i=1}^{m+3} a_i\right) \end{aligned}$$

This generalises easily; in fact, taking Eq. 1.21 as the inductive hypothesis, we obtain for $n+1$:

$$\begin{aligned}
&\left(\prod_{i=1}^{m} a_i\right)\left(\prod_{i=m+1}^{m+n+1} a_i\right)\\
&=\left[\left(\prod_{i=1}^{m} a_i\right)\left(\prod_{i=m+1}^{m+n} a_i\right)\right] a_{m+n+1}\\
&=\left(\prod_{i=1}^{m+n+1} a_i\right)
\end{aligned}$$

■

We can now state the general result.

Theorem 1.49 (General associativity rule). *Let n be a positive integer.* **Regardless of parenthesis layout***, we always have $a_1 \dots a_n = \prod_{i=1}^{n} a_i$, where $\prod$ is as defined above.*

Proof. The assertion is trivial for $n = 1, 2, 3$, and for $n = 4$ it is easily verifiable, but it depends on the assertion holding for *all* previous values of n. This strongly suggests using strong induction. Note that you **cannot** use **only** $n = 1$ or 2 as base cases here; the proof crucially depends on the property of associativity, which as remarked above, is only defined for $n \geq 3$—you have to check all the first 3 cases; I also checked $n = 4$ above just for good measure. So let us take the inductive step of saying that $a_1 \dots a_n = \prod_{i=1}^{n} a_i$ holds, *for all positive integers up to and equal to n.* For $n+1$, we note that for $a_1 \dots a_n a_{n+1}$, regardless of parenthesis disposition, there always exists $1 \leq k \leq n$ such that $a_1 \dots a_n a_{n+1} = (a_1 \dots a_k)(a_{k+1} \dots a_{n+1})$. Again, the parenthesis layout in both sub-groups of terms can be anything, but because both of them have n or less terms, by the induction hypothesis we get

$$(a_1 \dots a_k)(a_{k+1} \dots a_{n+1}) = \left(\prod_{i=1}^{k} a_i\right)\left(\prod_{i=k+1}^{n+1} a_i\right)$$

and by Lemma 1.48,

$$\left(\prod_{i=1}^{k} a_i\right)\left(\prod_{i=k+1}^{n+1} a_i\right) = \prod_{i=1}^{n+1} a_i$$

Thus, regardless of parenthesis, $a_1 \dots a_{n+1} = \prod_{i=1}^{n+1} a_i$ always holds, which proves the theorem—and also shows that parenthesis are unnecessary, because there is no ambiguity. ■

1.5 Functions

Let $f : A \to B$ be a function. If there exists $g : B \to A$ such that $g \circ f = id_A$, g is called the *left inverse* of f. And if $h : B \to A$ is such that $f \circ h = id_B$, then h is called the *right inverse* of f.

Theorem 1.50. ***(a)*** *f is* **injective** *if and only if it has a left inverse;* ***(b)*** *f is* **surjective** *if and only if it has a right inverse.*

Proof. **(a)**: $(\to)$ Assuming f is injective, to show that it has a left inverse we need only to construct it. Let $g : B \to A$ be defined as:

$$g(b) = \begin{cases} a, & \text{if there exists } a \text{ s.t. } f(a) = b \\ a', & \text{otherwise} \end{cases}$$

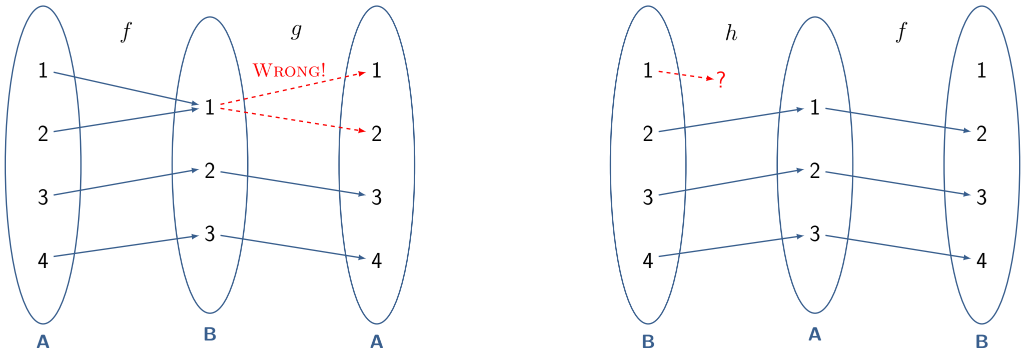
where a' is a random value of A. Because of the injectivity of f, if the first case happens, a is unique. This directly gives that $g(f(a)) = a, \forall a \in A$—i.e. id_A, which means g is the left inverse of f.

($\leftarrow$) Suppose f has a left inverse g, and that we have $f(x_1) = f(x_2)$. Applying g to both sides yields: $g(f(x_1)) = g(f(x_2)) \Longleftrightarrow x_1 = x_2$, which means f has is injective.

(b): ($\rightarrow$) If f is surjective, let $h : B \rightarrow A$ be such that $h(b) = a$, where a is a (possibly not unique) value such that $f(a) = b$. Because f is surjective, one such value always exists (it is unique if f is also injective). Then clearly $f(h(b)) = b$, i.e. id_B, that is, h is the right inverse of f.

($\leftarrow$) Suppose f has a right inverse $h : B \rightarrow A$; then for any element $b \in B$, we have $f(h(b)) = b$. That is, for every element of the codomain of f (i.e. B), there exists an element of the domain of f (i.e. A), namely $h(b)$, which is mapped by f to that codomain element. Hence f is surjective. ■

Remark 1.51. Note that the left inverses (in case **a)**) and right inverses (in case **b)**) defined above *are not unique*. △



(a) If f is not injective, it cannot have a left inverse—which sits on the right, when depicted.

(b) If f is not surjective, it cannot have a right inverse—which sits on the left, when depicted.

Figure 1.1: Intuition for theorem 1.50.

Theorem 1.52. *A function $f : A \rightarrow B$ is bijective if and only if there there exists $g : B \rightarrow A$ such that $g \circ f = id_A$ and $f \circ g = id_B$. Furthermore g is also a* **bijection**, *and it is* **unique**.

Proof. ($\rightarrow$) Construct g as follows. Because f is surjective, for any $b \in B$ there exists $a \in A$ such that $f(a) = b$; and because it is also injective, a is the only such value. Hence $g(b) = a$, where a is the unique value such that $f(a) = b$. The check that $g \circ f = id_A$ and $f \circ g = id_B$ is now routine.

($\leftarrow$) If g in the given conditions exists, then it is both the left and the right inverse of f. Theorem 1.50 then immediately implies that f is injective and surjective—and thus bijective.

For g being bijective, note that in both cases above, f is a left and right inverse of g, and hence g is (also per theorem 1.50) injective and surjective, and thus a bijection.

Now to show that g is unique, let g' be another inverse of f. For all x in the domain of f, both $g(f(x)) = x$ and $g'(f(x)) = x$ must hold. As f is bijective, when x ranges over all the domain of f, $f(x)$ ranges over all the domain of g and g'. Hence, g and g' are equal. ■

Theorem 1.53. *Given $f : A \rightarrow B$ and $g : B \rightarrow C$, we have:*

(i) *f and g injective $\Rightarrow$ $g \circ f$ injective;*

(ii) *f and g surjective $\Rightarrow$ $g \circ f$ surjective;*

(iii) *$g \circ f$ injective $\Rightarrow$ f injective;*

(iv) *$g \circ f$ surjective $\Rightarrow$ g surjective.*

Proof. All the cases are proved via the contrapositive:

(1): If $g \circ f$ is not injective, then there exist x_1 and x_2 such that $x_1 \neq x_2 \wedge g(f(x_1)) = g(f(x_2))$. But this directly implies that either f or g (or both) are not injective.

(2): If $g \circ f$ is not surjective, then either g is not surjective or, f is not surjective (or both).

(3): If f is not injective, then there exist x_1 and x_2 such that $x_1 \neq x_2 \wedge f(x_1) = f(x_2)$, which implies $g(f(x_1)) = g(f(x_2))$, i.e. $g \circ f$ is not injective.

(4): If g is not surjective, then $g \circ f$ cannot possibly be surjective, regardless of the domain of g, or the range of f. ■

Remark 1.54. From (1) and (2) above, we see that the composition of bijections is also bijective. △

2 | Groups

2.1 Groups

A set G with a binary operation that is closed, associative, and has an identity element, and inverses for all elements, forms a group.

An example are the so-called *diehedral groups*, which correspond to “rigid motions” of a regular polyhedron. To better convey what I mean by a rigid motion, imagine a jigsaw piece that happens to be shaped as a regular polyhedron. You remove it, rotate it, flip it, whatever—but in a way that allows you to *put it back in the same place*. How many motions of this kind are there? Well, take any side of the polyhedron, and label one of its vertices A, and the other one B. As the polyhedron has n vertices, for any rigid motion, you have n choices for where you want to place vertex A—but once this is done, you only have *two* available positions for vertex B. Hence, the diehedral group for an n side regular polyhedron, has $2n$ elements.

Weakened axioms. The group axioms for the existence of identity and inverses can be weakened, while still yielding the familiar properties of a group structure. Note that I denote the inverse of an element a as a^{-1}, for reasons that are explained in §2.2; especially circa definition 2.12.

Theorem 2.1. *Let X be a semigroup, i.e. a set with a binary associative operation, where there exists $e \in X$ such that, for all $a \in X$, $ea = a$ holds (that is, there exists a left identity). Furthermore, for all $a \in X$ there exists $a^{-1} \in X$ such that $a^{-1}a = e \in X$. That is, all elements have a left inverse. Then X is a group.*

Proof. We show this by showing first that any left inverse is also a right inverse. We want to come up with an expression where aa^{-1} appears, and which, through associativity, can either evaluate to aa^{-1} or e—thus proving that the left inverse is also a right inverse. $(a^{-1})^{-1}a^{-1}aa^{-1}$ is one such expression:

$$\begin{cases} \left((a^{-1})^{-1}a^{-1}\right)aa^{-1} = aa^{-1} \\ (a^{-1})^{-1}\left(a^{-1}a\right)a^{-1} = (a^{-1})^{-1}a^{-1} = e \end{cases}$$

Now it is only left to show that the left identity is also a right identity. We have $ae = a(a^{-1}a) = (aa^{-1})a = a$, QED. ■

We could have shown a similar result assuming only right-identity and right-inverses.

Theorem 2.2. *Let G be a group. Its identity element, e, is unique. The same holds for the inverse of any given element.*

Proof. Suppose there was another identity of G, say e'. Then, $ee' = e$ but also $ee' = e'$, so $e = e'$. Now let a' and a'' be two inverses of an element a. We have $(a''a)a' = a'$ but also $a''(aa') = a''$, and so $a' = a''$. ■

Corollary 2.3. *If two elements of a group have the same inverse, those two elements are equal.*

Proof. Let elements a' and a'' have the same inverse, namely a. We have $(a''a)a' = a'$ but also $a''(aa') = a''$, and so $a' = a''$. ■

Theorem 2.4. $(a^{-1})^{-1} = a$.

Proof. $(a^{-1})^{-1} = (a^{-1})^{-1}e = (a^{-1})^{-1}(a^{-1}a) = ((a^{-1})^{-1}a^{-1})a = a$. ■

We can now prove results like the following (Clark [1], article 26δ):

Theorem 2.5. *Let S be a semigroup with a finite number of elements. If the cancellation laws hold—that is if $ab = ac$ or $ba = ca$, then $a = b$—then S is a group.*

Proof. Let the elements of S be $s_1, \dots, s_n$, and right-multiply them all by an arbitrary $a \in S$. We obtain $s_1a, \dots, s_na$. These must all be distinct, otherwise we would have $s_ia = s_ja$ with $i \neq j$, but from the right cancellation law we have $s_i = s_j$, which is contradictory. So the s_ia are all distinct, and as they are in the same number as the elements in the original set, there exists an s_i such that $s_ia = a$, which means have a left identity element—denote it as e. The same reasoning shows that there must exist another element s_j such that $s_ja = e$, i.e. there also exist left inverses.[1] As a is an arbitrary element, this shows that all elements in S have a left inverse. It now follows from theorem 2.1 that S is a group. ■

Remark 2.6. The converse of theorem 2.5—that cancellation laws hold for any group—follows from the group axioms. △

This next result comes from the same place (Clark [1], article 29δ):

Theorem 2.7. *If G is a group such that each element is its own inverse, that is, $x^2 = e$ for all elements, then G is abelian.*

Proof. We have $(ab)(ba) = ab^2a = a^2 = e$, and so, ba is the inverse of ab. As the inverse is unique, and each element is its own inverse, ab and ba must be the same element—hence, the group is abelian. ■

2.2 Exponent Laws

A *monoid* M is a generalisation of a group, where we remove the condition that every element must have an inverse. I.e., it is a set with a binary operation that is closed, associative, and an identity element, e, for which it holds that $ea = ae = a$ for all $a \in M$.[2]

Of course, there can be invertible elements in a monoid—we just remove the requirement that *all* elements need to be invertible (if this happens the monoid is actually a group). If a is an invertible element of monoid M, that means there exists a' such that $a'a = aa' = e$. Note that this implies the inverse is unique, cf. theorem 2.2.

In a multiplicative monoid, exponentiation **to a non-negative power** is defined inductively as follows:

Definition 2.8. *Let a be an element of a monoid M, and $n \geq 0$. Then $a^{n+1} = a \cdot a^n$.*

This definition implies that $a^0 = e$;[3] also, due to associativity we have $a^{n+1} = a^n \cdot a$.

The usual exponent laws are valid in any monoid. For *non-negative exponents*, these are:

Theorem 2.9 (Exponent Laws). *Let a and b be elements on a monoid M. Then the following holds:*

(i) $a^n a^m = a^{n+m}$ *for all* $n \geq 0$ *and* $m \geq 0$.

(ii) $(a^n)^m = a^{nm}$ *for all* $n \geq 0$ *and* $m \geq 0$.

(iii) *If* $ab = ba$*, then* $(ab)^n = a^n b^n$ *for all* $n \geq 0$.

Proof. (1) Fix m, and verify that for $n = 0$ the property holds. Now assume it holds for an arbitrary n, and for $n+1$ we obtain $a^{n+1}a^m = aa^n a^m = aa^{n+m} = a^{(n+1)+m}$, where the last equality is due to definition 2.8 (as well as the commutativity and associativity of integer addition).

(2) Fix n, and verify that for $m = 0$ the property holds. Now assume it holds for an arbitrary m, and for $m+1$ we get $(a^n)^{m+1} = (a^n)(a^n)^m = a^n a^{nm} = a^{n(m+1)}$, where the last equality follows from (1).

(3) First prove that $ba^n = a^n b$: it holds for $n = 0$, and if we assume it holds for n, then for $n+1$: $ba^{n+1} = ba^n a = a^n ba = a^n ab = a^{n+1}b$. Now for $(ab)^n = a^n b^n$, it holds for $n = 0$; assuming it holds for n, for $n+1$ comes $(ab)^{n+1} = (ab)(ab)^n = (ab)a^n b^n = baa^n b^n = ba^{n+1}b^n = a^{n+1}bb^n = a^{n+1}b^{n+1}$. ■

Remark 2.10. Should you feel some unease due to the base being when either n or m is 0, feel free to start at 1—cf. remark 1.21. △

Remark 2.11. In property 3 above, as $ab = ba$, then it must also be that $(ab)^n = (ba)^n$. This means that $a^n b^n = b^n a^n$ also holds. △

Negative exponents. We now come to the main topic in this section, which is how we can generalise these laws to allow *any* integer exponent, including negative ones. Doing this however, first requires that one defines the exponentiation operation for negative powers. To give away the punchline:

negative exponents are only defined for invertible elements—i.e. units!

Let a be an invertible element of a monoid M, and let us represent said inverse as a^{-1}. Then, from the way have defined the inverse, we have that $aa^{-1} = a^{-1}a = a^0 = e$. This is in accordance with integer exponentiation, where given a common base, we add the exponents. This might lead us to consider an element like $(a^{-1})^n$, for some non-negative n. If $n = 0$ we obtain e, but if n is positive, what might the inverse of such an element be? Well, it is straightforward to verify that:

$$(a^{-1})^n a^n = e \tag{2.1}$$

and hence conclude that $(a^n)^{-1} = (a^{-1})^n$. A way of defining exponentiation of negative powers now suggests itself, especially when we take into account the desirability of maintaining that exponent addition property:

Definition 2.12. *Let the a be a unit of a monoid M, and n be a* **negative** *integer. Then* $a^n = (a^{-1})^{-n} = (a^{-n})^{-1}$.

On the other hand, if a is an invertible element with inverse a^{-1}, then a^{-1} is itself invertible:

Theorem 2.13. $(a^{-1})^{-1} = a$ *holds.*

Proof. $(a^{-1})^{-1} = (a^{-1})^{-1}e = (a^{-1})^{-1}a^{-1}a = ea = a$. ■

Both definition 2.12 and theorem 2.13 are in accordance with the usual rule to multiply the exponents. And the latter allows us to generalise the former definition:

Theorem 2.14. *$a^n = (a^{-1})^{-n} = (a^{-n})^{-1}$ holds for* **any** *integer n.*

Proof. The case for $n = 0$ is obvious, and for $n < 0$ is basically definition 2.12, so let $n > 0$. As $-n < 0$, we have from definition 2.12: $(a^{-1})^{-n} = ([a^{-1}]^{-1})^{-(-n)} = ([a^{-1}]^{-1})^n$. From theorem 2.13 now comes $([a^{-1}]^{-1})^n = a^n$.

Similarly, we have $(a^{-n})^{-1} = ([a^{-(-n)}]^{-1})^{-1} = ([a^n]^{-1})^{-1} = a^n$, from the same definition and theorem. ∎

Note that this means that the choice of representing the inverse of a as a^{-1} has paid off, because the property of multiplying exponents again holds, even when one of those exponents is -1 (and the other is any integer). This in turn allows us to generalise definition 2.8 ($aa^n = a^{n+1}$, for non-negative n):

Lemma 2.15. *For any $n \in \mathbb{Z}$, $a \cdot a^n = a^{n+1}$.*

Proof. If n is non-negative, this coincides with definition 2.8. If n is negative, then

$$\begin{aligned} & aa^n \\ &= a[a^{-1}]^{-n} && \text{(Definition 2.12)} \\ &= aa^{-1}[a^{-1}]^{-n-1} && \text{(Definition 2.8, as } -n \geq 1) \\ &= [a^{-1}]^{-n-1} = a^{n+1} && \text{(Theorem 2.14)} \end{aligned}$$

∎

One useful corollary is the following:

Corollary 2.16. *Given any integer l, another integer $k \geq 0$, and an invertible element a of a monoid M, we can always write $a^l = a^{l-k}a^k$.*

Proof. $a^l = aa^{l-1} = a^2a^{l-2} = \cdots = a^{l-k}a^k$. ∎

Generalisation of exponent laws for arbitrary integers. Now that we have constructed a way of making use of negative exponents, we prove the main result of this section, viz. that the usual exponent laws, proved above for non-negative exponents, also hold for negative ones. The trick turns out to be transforming the negative exponents into positive ones.

Theorem 2.17. *Let a and b be units in a monoid M. Then:*

(i) *$a^na^m = a^{n+m}$ for all $n, m \in \mathbb{Z}$.*

(ii) *$(a^n)^m = a^{nm}$ for all $n, m \in \mathbb{Z}$.*

(iii) *If $ab = ba$, then $(ab)^n = a^nb^n$ for all $n \in \mathbb{Z}$.*

Proof. The cases when the exponents are both non-negative have been dealt with in theorem 2.9, so we need deal only with the scenarios where either one, or both exponents are negative.

(i) We take first the case when both m and n are negative. We have $a^{m+n} = (a^{-1})^{-(m+n)}$ by theorem 2.14, and this latter expression is equal to $(a^{-1})^{-m+(-n)}$, where both $-m$ and $-n$ are positive. And as such, we apply theorem 2.9 to get $(a^{-1})^{-m}(a^{-1})^{-n}$, from where theorem 2.14 now allows us to write a^ma^n.

Now suppose that just one of the exponents is negative; without loss of generality, let it be $n < 0$. We have

$$\begin{aligned} a^m a^n &= a^m (a^{-1})^{-n} && \text{(thm. 2.14)} \\ &= \left(a^{m-(-n)} a^{-n}\right)(a^{-1})^{-n} && \text{(cor. 2.16, as } -n > 0) \\ &= a^{m-(-n)} e = a^{m+n} && (2.2) \end{aligned}$$

(ii) Again first consider the case where both m and n are negative. We have:

$$\begin{aligned} (a^m)^n &= \left\{[(a^{-1})^{-m}]^{-1}\right\}^{-n} && \text{(thm. 2.14)} \\ &= \left\{[(a^{-1})^{-1}]^{-m}\right\}^{-n} && \text{(idem.)} \\ &= (a^{-m})^{-n} && \text{(thm. 2.13)} \\ &= a^{mn} && \text{(thm. 2.9, as both } -m \text{ and } -n \text{ are } > 0) \end{aligned}$$

Now suppose only one of the exponents is negative, say m. Via a similar reasoning, we have $[(a^{-1})^{-m}]^n = (a^{-1})^{-mn} = [(a^{-1})^{-1}]^{mn} = a^{mn}$. If it were $n < 0$, we would have: $[(a^m)^{-n}]^{-1} = (a^{-mn})^{-1} = [(a^{mn})^{-1}]^{-1} = a^{mn}$.

(iii) First note that as $ab = ba$ by hypothesis, so is $a^{-1}b^{-1} = b^{-1}a^{-1}$. Now suppose $n < 0$, as the other case has been taken care of in thm. 2.9. We have:[4]

$$\begin{aligned} (ab)^n &= [(ab)^{-1}]^{-n} && \text{(thm. 2.14)} \\ &= [b^{-1}a^{-1}]^{-n} && (\text{as } abb^{-1}a^{-1} = b^{-1}a^{-1}ab = e) \\ &= [a^{-1}b^{-1}]^{-n} && \text{(above observation)} \\ &= [a^{-1}]^{-n}[b^{-1}]^{-n} && \text{(thm. 2.9, as } -n > 0) \\ &= a^n b^n && \text{(thm. 2.14)} \end{aligned}$$

■

On “additive” exponentiantion. The previous result can also be expressed in additive notation. In this case, a^n is represented as na or an, where a is a monoid element of whatever monoid we are considering, and n is an integer. The inverse of a—which, recall, has to exist for negative exponents to be allowed—is denoted $-a$. So we restate theorem 2.17 as follows:

Theorem 2.18 (Additive exponentiation). *Let a and b be units in an additively denoted monoid M. Then:*

(i) $an + am = a(n+m)$ *for all* $n, m \in \mathbb{Z}$.

(ii) $(an)m = a(nm)$ *for all* $n, m \in \mathbb{Z}$.

(iii) *If* $a + b = b + a$, *then* $(a+b)n = an + bn$ *for all* $n \in \mathbb{Z}$.

Remark 2.19. As remarked above, this new notational way of expressing exponentiation is “commutative”; e.g. in item 3 above, $an + bn = na + nb$. △

Also note that there is no problem if the elements of the monoid itself are integers; this is only troublesome for rings (cf. 3.1).

2.3 Subgroups

Definition 2.20 (Subgroup). *Let G be a group. A nonempty subset H of G is a* **subgroup** *if it also verifies the group axioms.*

Theorem 2.21 (Subgroup test). *Let G be a group and H a nonempty subset of G. H is a (sub)group if and only if $a, b \in H \Rightarrow ab^{-1} \in H$.*

Proof. (→) If H is a (sub)group, the theorem is obvious. (←) Conversely, if $a, b \in H \Rightarrow ab^{-1} \in H$, then associativity in H follows from the fact that $H \subseteq G$. Now as H is nonempty, it contains at least one element a. Thus $a \in H \Rightarrow aa^{-1} \in H$, i.e. $e \in H$, where e is the identity of G. Furthermore, for any $b \in H$, $eb^{-1} = b^{-1}$ must also be in H, i.e. H is closed for inverses. Finally, given $a, b \in H$, from the previous property we know that b^{-1} is also in H, and thus so is $a(b^{-1})^{-1} = ab$—i.e. H is closed for the group operation of G. ■

If the group is additive, the subgroup test says that it is sufficient to check for closure of "subtraction." For multiplicative groups, closure for "division" suffices. An equivalent form of this test is to check for closure for the group operation and inverses. This is stated in the next corollary.

Corollary 2.22 (Subgroup test v. 2). *Let G be a group and H a nonempty subset of G. H is a (sub)group if and only if $a, b \in H \Rightarrow ab \in H$ (closure of the group operation) and $a \in H \rightarrow a^{-1} \in H$ (closure of inverses).*

Proof. (→) If H is a subgroup, the result is clear. (←) If H is closed under inverses, then $b \in H \rightarrow b^{-1} \in H$, and from H's closure under the group operation, follows that $a, b \in H \rightarrow ab^{-1} \in H$. By theorem 2.21, H is a subgroup. ■

Subgroups containing an arbitrary set. Consider an arbitrary group G, and any nonempty subset S of that set G. We can always construct the "closure" of S, for the group operation of G, like so: $H = \left\{ s_1 s_2 \ldots s_m | s_i \in S \text{ or } s_i^{-1} \in S \text{ and } m \geq 0 \right\}$. It is straightforward that H is a subgroup of G, for by construction, H is closed under the group operation, and under inverses: if $s_{a_1} \ldots s_{a_r}$ and $s_{b_1} \ldots s_{b_t}$ are elements of H, then so is $s_{a_1} \ldots s_{a_r} s_{b_1} \ldots s_{b_t}$. For closure under inverses, if $s_1 s_2 \ldots s_m$ is an element of H, then its inverse is $s_m^{-1} \ldots s_2^{-1} s_1^{-1}$. As for each of the s_i's, it either belongs to S, or its inverse belongs to S, the same happens for each of the s_i^{-1}—and hence, $s_m^{-1} \ldots s_2^{-1} s_1^{-1}$ also belongs in H.

However, H is also the *smallest* subgroup of G that contains S. Reasoning by contradiction, suppose that there existed a subgroup H' of G that also contained S, but such that $|H'| < |H|$. Then there would exist in H (at least) one element of the form $a_1 a_2 \ldots a_t$, which would *not* belong to H'. But H' contains S, and as H' is a group, it must be closed under taking inverses, and thus, also contain the inverses of all the elements in S. Hence, all the a_i must be in H', and from closure of group operation, we conclude that $a_1 a_2 \ldots a_t$ must be in H', which contradicts the initial assumption—which proves that H is the smallest group containing S.

Theorem 2.23. *If G is an abelian group, the set $G\{m\} \overset{\text{def}}{=} \{a \in G | a^m = e\}$ is a subgroup.*

Proof. The set $G\{m\}$ is nonempty: $e^m = e$. $G\{m\}$ is clearly closed under the group operation: if $a^m = e$ and $b^m = e$, then as G is abelian, we have that $e = ee = a^m b^m = (ab)^m$, and hence, ab also belongs to $G\{m\}$. And obviously, $a^m = e \Longleftrightarrow (a^m)^{-1} = e^{-1} \Longleftrightarrow (a^{-1})^m = e$, so $G\{m\}$ is also closed under inverses. Hence, by the second form of the subgroup test (cf. thm. 2.22), $G\{m\}$ is a subgroup of G. ■

Iterating group elements. Consider an element a of some group G. The set $\langle a \rangle = \{a^m : m \in \mathbb{Z}\}$ is a subgroup of G: it is obviously closed under the group operation, and the inverse of any a^m is a^{-m}, so it is also closed under taking inverses.

Theorem 2.24. *Let a be an element of some group G. $\langle a \rangle$ is finite if and only if there exists a* **positive** *integer z such that $a^z = e$.*

Proof. (←) The set $\{m \in \mathbb{Z}^+ : a^m = e\}$ is nonempty, as it contains z (it also contains every multiple of z). Hence, by the WOP, it must contain a smallest element; call it n. I will show that $\langle a \rangle = \left\{a^0, a, \ldots, a^{n-1}\right\}$. Indeed, if s is any integer, by integer division we can write $s = nq + r$, with $0 \leq r < n$. Hence it follows that $a^s = a^{nq+r} = (a^n)^q a^r = ea^r = a^r$. As r belongs to the set $\{0, \ldots, n-1\}$, a^r (and thus a^s) belongs to $\left\{a^0, a, \ldots, a^{n-1}\right\}$. This shows that $\langle a \rangle \subseteq \left\{a^0, a, \ldots, a^{n-1}\right\}$; the reverse containment is immediate, and hence equality follows. Thus, $\langle a \rangle$ is finite.

(→) If $\langle a \rangle$ is finite, then there must exist some (distinct) integers i, j such that $a^i = a^j$. Without loss of generality, assume $i < j$. Then $a^i = a^j \iff a^{i-j} = e$. As $i - j$ is positive, we are done. ■

Remark 2.25. We could continue the second part of the proof, to discover the smallest positive value n for which $a^n = e$ (which may not be $i - j$). This would, as shown in the first part of the proof, allows us to write $\langle a \rangle$ as $\left\{a^0, a, \ldots, a^{n-1}\right\}$. In other words, if $\langle a \rangle$ is finite, it can be written as $\left\{a^0, a, \ldots, a^{n-1}\right\}$, for some n. △

Theorem 2.26. *Let G be a group, and H_1, H_2 be two subgroups. Then $H = H_1 \cap H_2$ is also a subgroup.*

Proof. Let $a, b \in H$. As H_1 is a subgroup, and a, b also belong to H_1, it follows from the subgroup test (cf. theorem 2.21) that $ab^{-1} \in H_1$. Similarly, we also have $ab^{-1} \in H_2$. But this means that $ab^{-1} \in H$ also holds. By the subgroup test, H is a subgroup. ■

2.4 Cosets and Normal/Factor (Sub)groups

Given a group G, a subgroup H, and an element $g \in G$, the set $gH \overset{\text{def}}{=} \{gh | h \in H\}$ is called a **(left) coset** of H, with representative g. The corresponding right coset would be Hg, defined as you would expect. Note that if $g \in H$, we have $gH = Hg = H$—i.e., both Hg and gH are (possibly different) permutations of H.[5] To see this, note that thanks to the closure of subgroups under the group operation, for any $h \in H$, $gh \in H$—and hence, $gH \subseteq H$; conversely, from the closure of subgroups under the group operation (and inverses), for any $h \in H$ there exists $h' \in H$ such that $gh' = h$, and hence $H \subseteq gH$. Thus $gH = H$, and a similar proof shows that for the right coset we also have $Hg = H$.

The implication shown above—that $g \in H$ implies that $gH = H$—also holds in reverse: if $gH = H$, then $g \in H$. Seeing this is an easy as $e \in H \rightarrow ge = g \in gH \rightarrow g \in H$. The same holds for Hg: if $Hg = H$, then $g \in H$. We can sum up this discussion in the following theorem.

Theorem 2.27. *Let H be a subgroup of a group G, and let $g \in G$. We have that $H = gH = Hg$ if and only if $g \in H$.*

Given $g_1, g_2 \in G$, saying that $g_1H = g_2H$ means that for any $h \in H$, we can always:

- find $h' \in H$ such that $g_1h = g_2h'$;

- find $h'' \in H$ such that $g_2h = g_1h''$.

Or in other words, saying that $g_1H = g_2H$ means that g_2H is a permutation of the set g_1H (and vice-versa, of course). Furthermore, if $g_1 = g_2$, then $g_1H = g_2H$—**but the converse is false!**

Now say $g_1H = g_2H$. Then, given any $h \in H$, there will exist some $h' \in H$ such that $g_1h = g_2h' \iff h = g_1^{-1}g_2h'$, which means $H \subseteq g_1^{-1}g_2H$. Conversely, from $g_1H = g_2H$ we also know that given any $h \in H$, there will exist some $h' \in H$ such that $g_2h = g_1h' \iff g_1^{-1}g_2h = h'$, from where it follows that $g_1^{-1}g_2H \subseteq H$. Thus we conclude that $g_1H = g_2H \iff g_1^{-1}g_2H = H$, and reasoning similarly (or better yet, by symmetry) we can also conclude that both sides of the previous equivalence are also equivalent to $g_2^{-1}g_1H = H$.

So to summarise, when given an expression like $g_1H = g_2H$, we can manipulate it as if it were an algebraic expression of sorts. This reasoning lays the foundation for the next result, which gives further conditions that follow from coset equality.

Theorem 2.28. *Let G be a group, H one of its subgroups, and g_1, g_2 two of G's elements. Then the following are equivalent:*

(i) $g_1H = g_2H$;

(ii) $Hg_2^{-1} = Hg_1^{-1}$;

(iii) $g_1H \subseteq g_2H$;

(iv) $g_2 \in g_1H$;

(v) $g_1^{-1}g_2 \in H$.

Proof. We shall prove 1 → 3 → 5 → 2 → 4 → 1.

(1 → 3) Equation 3 is a direct consequence of equation 1.

(3 → 5) There exists an $h \in H$ such that $g_1 = g_2h \iff h^{-1} = g_1^{-1}g_2$—and hence, $g_1^{-1}g_2 \in H$.

(5 → 2) $g_1^{-1}g_2 \in H \iff H = Hg_1^{-1}g_2 \iff Hg_2^{-1} = Hg_1^{-1}$. (Also cf. the discussion above, prior to the statement of this theorem.)

(2 → 4) From 2, it follows that there exists $h \in H$ such that $g_2^{-1} = hg_1^{-1} \iff g_2^{-1}g_1 = h \iff g_1^{-1}g_2 = h^{-1} \iff g_2 = g_1h^{-1}$, and hence, $g_2 \in g_1H$.

(4 → 1) We have that $g_2 = g_1h$, for some $h \in H$. Hence, $g_2H = g_1hH \iff g_2H = g_1H$. ■

Remark 2.29. The idea behind the equivalence of conditions $g_1H = g_2H$ and $Hg_2^{-1} = Hg_1^{-1}$ can be shortly stated like this: $g_1H = g_2H \iff g_1^{-1}g_2 \in H \iff Hg_1^{-1}g_2 = H \iff Hg_1^{-1} = Hg_2^{-1}$. △

Remark 2.30. By symmetry, all the results in theorem 2.28 holds if we swap g_1 and g_2. In particular, $g_1^{-1}g_2 \in H$ if and only if $g_1g_2^{-1} \in H$ (which also follows from H closure under taking inverses). △

Definition 2.31. *Given a group G, a subgroup N is said to be* **normal** *in G if for all $g \in G$, $gN = Ng$.*

Remark 2.32. Note that if G is abelian, all its subgroups are normal. △

Theorem 2.33. *If N is a normal subgroup of G, then we have $gNg^{-1} = N$, for any $g \in G$.*

Proof. We show this through double containment. From the definition of normal subgroup, for any $g \in G$ and any $n \in N$ there exists 1) $n' \in N$ such that $gn = n'g$ and 2) $n'' \in N$ such that $ng = gn''$. From condition 1) comes that $gng^{-1} = n'$, which means $gNg^{-1} \subseteq N$. From 2), comes $n = gn''g^{-1}$, from where $N \subseteq gNg^{-1}$. ■

Remark 2.34. The converse of theorem 2.33 is obviously true: $gNg^{-1} = N$ immediately implies that $gN = Ng$. △

Remark 2.35. By symmetry, we also have that if N is a normal subgroup of G, then we have $g^{-1}Ng = N$, for any $g \in G$. △

Actually, the conditions $N \subseteq gNg^{-1}$ and $gNg^{-1} \subseteq N$, for any $g \in G$, are equivalent (each of them implies the other). Take the first one, $N \subseteq gNg^{-1}$ for any $g \in G$: this means that given any $n \in N$, there exists $n' \in N$ such that $n = gn'g^{-1}$. But note that this holds, *for any* $g \in G$. As G is group, this also means this condition holds, *for the inverse of any* $g \in G$. Hence we can restate said condition as follows: given any $n \in N$, there exists $n' \in N$ such that $n = g^{-1}n'(g^{-1})^{-1}$. The last equality is equivalent to $gng^{-1} = n'$, for any $n \in N$ and $g \in G$—but is precisely to state that $gNg^{-1} \subseteq N$. In other words, we proved that $N \subseteq gNg^{-1} \to gNg^{-1} \subseteq N$, for any $g \in G$. The converse implication is proven via a similar reasoning.

Hence, given a group G and subgroup N, it suffices to know that either $N \subseteq gNg^{-1}$, or that $gNg^{-1} \subseteq N$, for any $g \in G$, to know that N is a normal subgroup of G. This allows us to prove the following property:

Theorem 2.36. *Let N_1 and N_2 be two normal subgroups of a group G. Then $N_1 \cap N_2$ is also a normal subgroup.*

Proof. $N_1 \cap N_2$ is a subgroup of G by property 2.26. To show that it is also normal, let $N = N_1 \cap N_2$. We know that for any $g \in G$, $gNg^{-1} \subseteq gN_1g^{-1} = N_1$, where the equality is because N_1 is normal. Similarly $gNg^{-1} \subseteq gN_2g^{-1} = N_2$. Hence, $gNg^{-1} \subseteq N_1 \cap N_2 = N$, entailing the normality of N. ■

Theorem 2.37 (Quotient group). *Given a group G, with a normal subgroup N, the set of left cosets of N, forms a group, under the operation $(aN)(bN) = (ab)N$. This group is called the* **factor group**, *or* **quotient group**.

Proof. N acts as the identity, and the inverse of aN is $a^{-1}N$. The operation is associative, because:

$$[(aN)(bN)](cN) = (abN)(cN) = ((ab)c)N = (a(cb))N \tag{2.3}$$

$$= (aN)[(cb)N] = (aN)[(bN)(cN)] \tag{2.4}$$

And it is well-defined, because if $aN = a'N$, and $bN = b'N$, then $a' = an_1N$ and $b' = bn_2N$, for some $n_1, n_2 \in N$. So:

$$(a'N)(b'N) = (an_1N)(bn_2N) = (aN)(bN) = (ab)N \tag{2.5}$$

because n_1N and n_2N are just permutations of N, due to the fact that $n_1, n_2 \in N$. ■

The previous result uses multiplicative notation, but we can illustrate the same principle using additive notation, with the group $\mathbb{Z}_n$, of integer addition modulo n.[6] The elements of this group are the equivalence classes, modulo n, of the integers $0, 1, \ldots, n-1$. But this group is also denoted as $\mathbb{Z}/n\mathbb{Z}$. Why? Because $\mathbb{Z}$ is an abelian additive group, and thus all subgroups are normal—and in particular, $n\mathbb{Z}$ is normal. Furthermore, the cosets $0+n\mathbb{Z}, 1+n\mathbb{Z}, \ldots, (n-1)+n\mathbb{Z}$ correspond exactly to the equivalence classes of $0, 1, \ldots, n-1$. And as we can always push extraneous n multiples into $n\mathbb{Z}$, coset addition in the quotient group corresponds to modular addition.

But note that this does NOT happen with $\mathbb{Z}_n^*$! Indeed, as $\mathbb{Z}$ is not a group under multiplication (only 1 and -1 have inverses), $n\mathbb{Z}$ cannot be a subgroup.

2.5 Permutation Groups

Theorem 2.38. *Let id be the identity permutation. If $id = \sigma_1\sigma_2 \ldots \sigma_r$, where each σ_i is a* **transposition**, *then r is* **even**.

Proof. Let $\sigma_r = (a\ b)$. Then, for $\sigma_{r-1}\sigma_r$, we have the following possibilities:

1. $\sigma_{r-1} = (a\ b)$. In this case, $\sigma_{r-1}\sigma_r = (a\ b)(a\ b) = id$.
2. $\sigma_{r-1} = (b\ c)$. That is, σ_{r-1} and σ_r have the element b in common. Hence, we have:

$$\sigma_{r-1}\sigma_r = (b\ c)(a\ b) = (a\ c\ b) = (a\ b)(a\ c) \tag{2.6}$$
$$(b\ a\ c) = (b\ c)(b\ a) \tag{2.7}$$
$$(c\ b\ a) = (c\ a)(c\ b) = \boxed{(a\ c)(c\ b)} \tag{2.8}$$

3. $\sigma_{r-1} = (a\ c)$. That is, σ_{r-1} and σ_r have the element a in common. Hence, we have:

$$\sigma_{r-1}\sigma_r = (a\ c)(a\ b) = (a\ b\ c) = (a\ c)(a\ b) \tag{2.9}$$
$$(c\ a\ b) = (c\ b)(c\ a) \tag{2.10}$$
$$(b\ c\ a) = (b\ a)(b\ c) = \boxed{(a\ b)(b\ c)} \tag{2.11}$$

4. $\sigma_{r-1} = (c\ d)$. That is, σ_{r-1} and σ_r are disjoint. And so, $\sigma_{r-1}\sigma_r = (c\ d)(a\ b) = \boxed{(a\ b)(c\ d)}$.

Now, a transposition being a cycle of length 2, that is to say, $(r\ s)$, with $r \neq s$, that having $id = (r\ s)$ is not possible. On the other hand, if $id = \sigma_1\sigma_2$, then clearly $\sigma_1 = \sigma_2^{-1}$—and as the inverse of a transposition is itself, we must have that $\sigma_1 = \sigma_2$. In other words, the proposition "if $id = \sigma_1\sigma_2 \ldots \sigma_r$, then r is even" is true for $r \leq 2$. Now suppose it is true for $r < n$; what happens for $r = n$? Well, $\sigma_{n-1}\sigma_n$ will fall into one of the four categories outlined above.

If it is case 1), then $\sigma_{n-1}\sigma_n = id$, and so we can just cancel this part, obtaining $id = \sigma_1\sigma_2 \ldots \sigma_{n-2}$. But $n-2 < n$, and so, by the induction hypothesis, it is even—and thus so must be n.

For the remaining cases 2), 3) and 4), the analysis above shows that we can rewrite $\sigma_{n-1}\sigma_n$ as a new pair of permutations $\sigma'_{n-1}\sigma'_n$, in which one of the elements of σ_n—in the enumeration above, this would be element a—only shows up in σ'_{n-1}. And we can continue to do this "leftward shift" of element a, with $\sigma_{n-2}\sigma'_{n-1}$, and so on. Eventually we would end up with the identity written as a sequence of transpositions in which the element a would only show up in the leftmost transposition—except this is not possible, because one such sequence of transpositions would not fix that element, and hence cannot be the identity (which fixes all elements). This forces the conclusion that, sometime during that "leftward shift" process done to the sequence of transpositions $\sigma_1\sigma_2 \ldots \sigma_n$ (which by assumption equals the identity), one must end up with scenario 1)—i.e., two equal transpositions juxtaposed, that thus cancel one another—leaving the identity written as a sequence of $n-2$ transpositions. Again by the induction hypothesis, $n-2$ is smaller than n, and thus even, and hence so must be n. ■

Theorem 2.39. *Let β be a permutation for which it holds that $\beta = \sigma_1 \ldots \sigma_r = \tau_1 \ldots \tau_s$, where both the σ_i and the τ_j are transpositions. Then r and s are either both even, or both odd.*

Proof. $id = \sigma_1 \ldots \sigma_r(\tau_1 \ldots \tau_s)^{-1} = \sigma_1 \ldots \sigma_r\tau_s^{-1} \ldots \tau_1^{-1} = \sigma_1 \ldots \sigma_r\tau_s \ldots \tau_1$. By the previous proposition (prop. 2.38), $r+s$ must be even—and so, r and s must either be both even, or both odd. ■

Remark 2.40. From the previous theorem it follows that no transposition can be written as the product of two transpositions, for it is already the product of one transposition, namely, itself. △

Corollary 2.41. *The inverse of an even (resp. odd) permutation is even (resp. odd).*

Proof. Suppose there was a permutation α that was even, but which inverse was odd (or vice-versa, it doesn't matter). Then, as $id = \alpha\alpha^{-1}$, we would be able to able to write the identity permutation as a sequence of odd transpositions, as even + odd = odd. This is impossible by theorem 2.38. ■

Theorem 2.42. *Let α and β be two permutations. If they are either both even or both odd, then $\alpha\beta$ is even. Otherwise—i.e., if α and β are "one of each," then $\alpha\beta$ is odd.*

Proof. Let $\gamma = \alpha\beta$. This is equivalent to $id = \alpha\beta\gamma^{-1}$. As the identity can only result from an even number of transpositions, if both α and β are both even or both odd, then their parity sums to even, and thus γ^{-1} must also be even, which in turn means that γ is even (cf. 2.41), which is the same as saying that $\alpha\beta$ is also even.

If α and β are "one of each," then their parity sums to odd, and via a similar reasoning, γ^{-1} must be odd, and so must γ, i.e., $\alpha\beta$. ■

Theorem 2.43. *Iterating n times a cycle of length, we obtain the identity. Furthermore, n is the smallest number for which this happens.*

Proof. A cycle is a bit like a collapsing sequence of dominoes: they all fall "in the same manner." Conversely, if one piece does not fall, then no piece fall. Thus it suffices to show that iterating the cycle n times leaves the first fixed, for it then follows that the same holds for all the other elements. Let $(a_1\ a_2\ \dots\ a_n)$ by a cycle of length n. Consider element a_1: after one iteration it goes to position 2, after two iterations to position 3, and so on, until after $n-1$ iterations it goes to position n, and after n iterations it goes to position $n+1$, which is actually position 1—i.e., it is fixed. Thus, so are all other elements, meaning we obtained the identity permutation. Moreover, this reasoning shows that it is not possible to fix element a_1 with less than iterations of the cycle $(a_1\ a_2\ \dots\ a_n)$. ■

Theorem 2.44. *Let β be a permutation written as a product of disjoint cycles. The minimum number of iterations required to obtain the identity, is the least common multiple of the lengths of all cycles.*

Proof. Begin by observing that the elements that are not in any cycle, are always fixed. Furthermore, as the cycles are disjoint, any element in any given cycle is fixed by all the other cycles. Hence, the "motion" of an element in a given cycle, depends only on that cycle. If said cycle has length n, then after n iterations all elements of that cycle are fixed (cf. the previous theorem, 2.43).

Now, those elements fixed after n iterations, are also fixed after $2n, 3n, \dots$ iterations. Applying the same reasoning to all the cycles in the permutation, we conclude that the minimal number of iterations needed to fix all elements of all cycles, is the least common multiple of the lengths of those cycles. ■

Example 2.45. Consider the permutation $(1\ 2\ 3)(4\ 5)$. According to the previous proposition (2.44), the cycle $(1\ 2\ 3)$ should be fixed at every 3 iterations, and the cycle $(4\ 5)$ at every 2—and the entire permutation at the end of 2 × 3 = 6 iterations. Which is exactly what happens (note that $(1\ 2\ 3)(4\ 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 5 & 4 \end{pmatrix}$):

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 5 & 4 \end{pmatrix} \Rightarrow \boxed{\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 4 & 5 \end{pmatrix}} \tag{2.12}$$

$$\Rightarrow \boxed{\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 5 & 4 \end{pmatrix}} \Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 4 & 5 \end{pmatrix} \tag{2.13}$$

$$\Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \end{pmatrix} \Rightarrow \boxed{\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}} \tag{2.14}$$

The boxed permutations are those one of the cycles is fixed (plus the last one which fixes both cycles). ◊

Theorem 2.46. *Any cycle in S_n can be written using at most $n-1$ transpositions.*

Proof. Given any cycle with k elements $(a_1\ a_2\ \dots\ a_k)$, it can always be represented using $k-1$ transpositions: $(a_1\ a_k)(a_1\ a_{k-1})\dots(a_1\ a_2)$. If the cycle belongs to S_n, then it has at most n elements—and hence, can be represented with $n-1$ transpositions. ■

Theorem 2.47. *Given a permutation σ in S_n that is* **not** *a cycle, it can be written using at most $n-2$ transpositions.*

Proof. If σ is not a cycle, then it can be written as the composition of at least two disjoint cycles. Suppose it takes exactly two disjoint cycles, and let a and b be their respective lengths. As $\sigma \in S_n$, $a + b \leq n$. Now by the previous result, theorem 2.46, each of those cycles can be written with at most $a-1$ and $b-1$ transpositions, respectively. Hence for σ we need a total of $(a-1)+(b-1) = (a+b)-2 \leq n-2$ transpositions.

Now suppose that σ requires *three* disjoint cycles, and let c be the length of that third cycle. We now have $a+b+c \leq n$, and so the total number of transpositions is $(a-1)+(b-1)+(c-1) = (a+b+c)-3 \leq n-3$. From this it is straightforward to see that as the number of disjoint cycles needed to represent σ increases, the number of required transpositions *decreases*. Given that σ is not a cycle, we need at *least* two disjoint cycles—and thus, we require at *most* $n-2$ transpositions. ■

Theorem 2.48. *Given a cycle of length r, iterating it s times yields a cycle if and only if $gcd(r,s) = 1$.*

Proof. We begin by proving a simpler result: if the length of the cycle, r, is odd, then iterating it two times yields another cycle. To see this, consider the cycle $(1\ 2\ \dots\ r)$, with r odd, and $s = 2$. We have:

$$(1\ 2\ 3\ \dots\ r)(1\ 2\ 3\ \dots\ r) = (1\ 3\ 5\ \dots\ r-2\ r\ 2\ 4\ 6\ \dots\ r-3\ r-1) \tag{2.15}$$

Now consider the more general case, where r might be any positive integer. We need to bring in the machinery from modular arithmetic, and for that reason, we will denote the cycle $(1\ 2\ \dots\ r)$ as $(0\ 1\ \dots\ r-1)$. To run this cycle once, means that 0 goes to 1, 1 to 2, and so on, until $r-2$ goes to $r-1$, and $r-1$ goes to 1. To run it twice, means that 0 goes to 2, 2 to 4, 4 to 6 and so on—but on this case, what happens at the end depends on the length of the cycle and the number of iterations. For example, consider the cycle (0 1 2 3), iterated twice. We have:

$$(0\ 1\ 2\ 3)(0\ 1\ 2\ 3) = (0\ 2)(1\ 3) \tag{2.16}$$

But if it is iterated *three* times, we again obtain a cycle:[7]

$$(0\ 1\ 2\ 3)^3 = (0\ 1\ 2\ 3)(0\ 2)(1\ 3) = (0\ 3\ 2\ 1) \tag{2.17}$$

This helps one to see that, more generally, for a cycle $(0\ 1\ \ldots\ r-1)$, iterated s times, 0 goes to s, s to $s+s=2s$, $2s$ to $3s$ and so on—where all the arithmetic is done modulo r. Note that if this results in a cycle, then it must be a cycle of length r; i.e., it must contain all the elements $0, 1, \ldots, r-1$. Otherwise, it would be a product of disjoint cycles—and thus, not a cycle.[8] Thus if this results in a cycle, it must be because all the r values $0, s, 2s, 3s, \ldots, (r-1)s$, all modulo r, yield the precisely the r values $0, 1, \ldots, r-1$, albeit in a possibly different order. This is because the value that comes after $(r-1)s$, is rs, which is equivalent to $0s$, or 0, modulo r—and thus, the cycle begins repeating itself, as expected. Conversely, if the result of having s iterations of $(0\ 1\ \ldots\ r-1)$ is a cycle of length r, denote it $(a_1\ a_2\ \ldots\ a_r)$, then that can only be because each of the a_i belongs to the sequence $0, s, 2s, 3s, \ldots, (r-1)s$, modulo r.

The above discussion leads us to conclude that the cycle $(0\ 1\ \ldots\ r-1)$, iterated s times, yields a cycle if and only if the equation $ks \equiv b \pmod{r}$, can always be solved (for unknown k) for any b (to avoid having to talk about equivalence classes here, let us consider that $k, b \in \{0, 1, \ldots, r-1\}$). An elementary result in number theory establishes that this happens if and only if $gcd(r,s) = 1$—which proves our result. ■

Remark 2.49. The above proof of theorem 2.48 glosses over one detail: the cycle $(0\ 1\ \ldots\ r-1)$, can also be written (for a sufficient large r) as, for example, $(3\ 4\ \ldots\ r-1\ 0\ 1\ 2)$. In this case the equation we required be solvable for any b, is $3 + ks \equiv b \pmod{r}$. This is equivalent to $ks \equiv b - 3 \pmod{r}$, and as b can be any of $\{0, 1, \ldots, r-1\}$, so can $b - 3 \pmod{r}$. We are thus led to the same conclusion: we must have $gcd(r,s) = 1$. And of course, 3 was an arbitrary starting point—we could have started the cycle at any $0 \le i < r$, obtaining $(i\ i+1\ i+2\ \ldots\ r-1\ 0\ 1\ 2\ \ldots\ i-1)$—and same reasoning would have led to the same conclusion: we must have $gcd(r,s) = 1$. △

* * *

Dihedral groups. In abstract terms, the dihedral group for a given $n \ge 3$, is a subset of S_n, with a given property. What property might that be? To get there, we need some background.

In more concrete terms, dihedral groups consist of all the "solid motions" of a regular polygon of n sides, with $n \ge 3$ (hereinafter, an n-gon). To illustrate what, exactly, one means by "solid motion," consider the square depicted in Figure 2.1. In how many different ways can we move it, so as to be able to "put it back in place?" To answer this, let us consider a motion in which vertex 1 is carried to position, say 3. If this is a solid motion, i.e. if it preserves the square "intact," so to speak, then vertex 2 either a) goes to position $3-1=2$ (i.e., it stays put), or b) it goes to position $3+1=4$. And once the new positions of vertices 1 and 2 have been fixed, *the position of all other remaining vertices are also determined*—otherwise, we would deform the shape of the square, making it a square no longer.

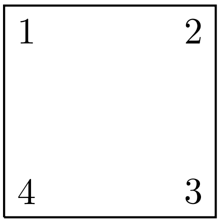


Figure 2.1: text

This reasoning applies, almost without change, to the more general case of a regular n-gon. If vertex 1 goes to position k, then vertex 2 goes to either $k+1$, or $k-1$—and once the position of vertex 2 has been set, the positions of all other remaining vertices are also set. So we have n possible positions for vertex 1, and having that set, two positions for vertex 2, and that sets the solid motion. Thus, there exists a total of $2n$ different solid motions.

Let us now return to the square of Figure 2.1. There exist $2 \times 4 = 8$ distinct solid motions. Four of these are *rotations*: Four of these are rotations: 90°, 180°, 270°, and 360°. It is easier for me to picture the motions of a dihedral group if I imagine the n-gon moving *clockwise*, so that will be the convention adopted hereinafter—but beware that this violates a longstanding convention in mathematics that positive angles stand for anticlockwise rotation. Furthermore, with the four rotation above, we do not need to consider *any* anticlockwise rotations, for we already have all possible cases covered: for example, an anticlockwise rotation of 90° is equivalent to a clockwise rotation of 270°.

To simplify the notation to be used further down, we shall think of a rotation of 360° as being the *trivial* rotation, i.e., a rotation of 0°. Let us now list the permutations that correspond to these four rotations.

$$0^\circ = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \qquad 90^\circ = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \tag{2.18}$$

$$180^\circ = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \qquad 270^\circ = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix} \tag{2.19}$$

Beyond these four rotations, there are also four *reflections*, induced by four *axis of symmetries*: vertical (|), horizontal (—), main diagonal $(1-3)$ and secondary diagonal $(2-4)$. The corresponding permutations are the following:

$$- = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \qquad | = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \tag{2.20}$$

$$1-3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \qquad 2-4 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} \tag{2.21}$$

Inspecting these 8 permutations, one can infer a more abstract notion of a "solid motion" of a regular n-gon actually means: it is a subset of S_n, with the property that the transformations of two *consecutive* vertices, uniquely identify any permutation in that subset. In fact, in the initial discussion above, where it is shown that there are $2n$ solid motions for a regular n-gon, we used precisely this property (with vertices 1 and 2), although without stating it explicitly. For example, if a given solid motion sends vertex 3 to position 1, and vertex 4 to position 2, then is speaking of the 180° rotation, for there no other permutation, in this subset of 8, which maps these two vertices in the same manner. Note, however, that this property **is not valid for non-consecutive vertices!** For example, taking vertex 2 to position 4 and vertex 4 position 2, could either be the $1-3$ reflection, or the 180° rotation.

Basing ourselves in the discussion hitherto, we will now define some notation. The non-trivial rotation with the smallest amplitute, $360^\circ/n$ clockwise, will be denoted r. The next one, $(360^\circ/n) \times 2$, will be r^2. Carrying on like this, we obtain $r, r^2, r^3, \ldots, r^{n-1}$. For r^n, we obtain the rotation by $(360^\circ/n) \times n = 360^\circ$, which is actually the trivial rotation, and hence will be denoted as r^0. That is, it corresponds to the rotation r, done zero times—i.e. we don't rotate anything.

Thus we obtain n distinct rotations, including the trivial one. What is missing are the *reflections*, the number of which is equal to the number of *axis of symmetry* that the n-gon actually has. Now, given any regular n-gon, for each vertex, there is always an axis that goes through that vertex, and splits the n-gon into two equal halves. Let us denote those axis—and, abusing the notation, the reflections they induce—by $s_1, s_2, \ldots, s_n$. If the number of vertices (and thus of edges), n, is odd, then the s_i are all of the axis of symmetry that the n-gon has; if n is even, things are slightly more complicated. As the case of the square in Figure 2.1 illustrates, the s_i coincide pairwise: $s_1 = s_{1+n/2}, s_2 = s_{2+n/2}, \ldots, s_{n/2} = s_{n/2+n/2} = s_n$. This means that the s_i only yield $n/2$ distinct axis of symmetry. There are, however, an additional $n/2$ axis of symmetry, that do not go through any vertex, but rather bisect parallel edges of the n-gon (for n even, the edges of a regular n-gon are pairwise parallel). There will be no need to refer explicitly to these edges, and hence, I introduce no notation for them.

So, for any regular n-gon, whether n is even or odd, we have n rotations, and n reflections, thus amounting to $2n$ solid motions, which, as shown above, is the total number of solid motions. The next step is to characterise these $2n$ motions a bit more formally.

Let us designate by s the reflection that fixes vertex 1, i.e., s_1. Let us now suppose that a given solid motion sends vertex 1 to position k. If vertex 2 is sent to position $k+1$, then the motion must be a rotation, namely r^{k-1}. If, on the other hand, vertex 2 is sent to position $k-1$, then this motion can never be a rotation. Then, it must be a reflection, for this is the only other kind of solid motion possible. However, the corresponding axis of symmetry is something that cannot be determined *a priori*—among other reasons, because it depends on whether n is even or odd. This turns out to not be much of a problem, though, because we can "deconstruct" this motion, as the combination of the reflection s, followed by the rotation r^{k-1}: s leaves vertex 1 in its place, and sends 2 to position n; r^{k-1} takes 1 to k, and 2 to $k-1$.

Thus we obtain, yet again, the set of $2n$ solid motions: n rotations—$r^0, r, \ldots, r^{n-1}$—and n reflections—$sr^0 (= s), sr, sr^2, \ldots, sr^{n-1}$. We can, one more time, go back to the example of the square of Figure 2.1, and check that s, sr, sr^2 and sr^3 are in fact the four possible reflections. We have:

$$s = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} = 90^\circ \tag{2.22}$$

And thus:

$$sr = \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}}_{r} \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}}_{s} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = | \tag{2.23}$$

$$sr^2 = (sr)r = \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}}_{r} \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}}_{sr} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} = 2-4 \tag{2.24}$$

$$sr^3 = (sr^2)r = \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}}_{r} \underbrace{\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}}_{sr^2} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} = - \tag{2.25}$$

Note that $s = sr^0 = 1-3$—and hence, we obtained *all* of the reflections of the square.

Important note on the order of parcels. Above, when writing explicitly the parcels of sr, I wrote first r, and then s—i.e., I reversed the order. Why? Because even though sr stands for applying the permutation s first, and the permutation r second, permutations are actually functions, and function composition usually happens right to left. Hence, when writing the actual functions, rather than the elements that denote them (e.g., s and r), I reversed the order.

Moving on, one can now ask: what is so special about s_1? Or in other words, could the set of n reflections $sr^0, sr, \ldots, sr^{n-1}$ have been described using some reflection *other than* s_1? The answer is yes because, given any two reflections, there is always some rotation that relates them. For example, let s_i and s_j be two distinct rotations. If s_i maps vertex 1 to position k, then it will have to map vertex 2 to position $k-1$. Similarly, if s_j sends vertex 1 to position l, then it will have to send vertex 2 to position $l-1$. It is now simple to observe that the rotation that takes vertex k to position l, will also take vertex $k-1$ to position $l-1$—relating two reflections. Let r' be that rotation; we have then that $s_j = s_i r'$.

This allows us to write $sr^0, sr, \ldots, sr^{n-1}$, using *any* reflection, rather than $s(= s_1)$. For example, let s' be some arbitrary reflection, and let r' be the rotation that relates them like this: $s = s'r'$. Via simple substitution, we can now rewrite $sr^0, sr, \ldots, sr^{n-1}$ as $s'r'r^0, s'r'r, \ldots, s'r'r^{n-1}$. Now, the application of two consecutive rotations, continues to be a rotation; one just has to add the exponents modulo n (because $r^n = r^0$): $r^i r^j = r^{i+j \pmod n}$. But this means that $r'r^0, r'r, \ldots, r'r^{n-1}$ are precisely the elements $r^0, r, \ldots, r^{n-1}$ (possibly rearranged). And hence, $s'r'r^0, s'r'r, \ldots, s'r'r^{n-1}$ must also be the same reflections as $sr^0, sr, \ldots, sr^{n-1}$. This shows that we can take s to be any of the n reflections of the n-gon. (Note that if n is even, this includes the reflections which axis corresponds to no vertex.)

Group axioms. We just described all the $2n$ elements of a dihedral group: n rotations—$r^0, r, \ldots, r^{n-1}$—and n reflections—sr^0, sr, sr^{n-1}. But we have *not* shown that these, together with the operation of function composition, actually yield a group. This is the next task.

First in the order business, is to show that the group operation is associative—but this is immediate because the permutations are functions, and function composition is associative. Next, we need an identity element—and with little surprise, this turns out to be r^0, which is, after all, the identity permutation. This leaves only the question of inverses.

For rotations it is easy: the inverse of r^k is r^{n-k}; and as $r^n = r^0$, the inverse of r^0 is itself—as one would expect to happen with the identity element. This is shown as follows: r^k takes vertex 1 to position $k+1$, and vertex 2 to position $k+2$. In turn, r^{n-k} takes vertex $k+1$ to position $n-k+k+1 = n+1 = 1$, and vertex $k+2$ to position $n-k+k+2 = n+2 = 2$. Thus, the application of r^k, followed by the application of r^{n-k}, yields fixes the consecutive vertices 1 and 2, and hence, also fixed the remaining vertices—which means we obtain the identity permutation, r^0.

As for the reflections sr^i, we intuitively expect that the inverse of a reflection, is that reflection itself—and this is indeed, what happens. To show as much, let $s = s_1$, and let r^i be an arbitrary rotation. I will show that sr^i fixes both vertices 1 and 2. We have seen that r^i sends a vertex k to position $i+k$. As for reflection $s = s_1$, it fixes vertex 1, and sends the remaining vertices k, with $2 \leq k \leq n$, to $n-k+2$. For example, vertex 2 goes to n, vertex 3 to $n-1$, and so on, until vertex n, that goes to position 2. If we apply the sequence of permutations sr^isr^i, we obtain:

	s	r^i	s	r^i
1	1	$1+i$	$n-(1+i)=n-i+1$	$n-i+1+i=n+1=1$
2	n	$n+i$	$n-(n+i)+2=2-i$	$2-i+i=2$

(2.26)

This shows that sr^i is its own inverse, or equivalently, that $sr^isr^i = r^0$. And thus, we have now shown that the set $\{r^0, r, \ldots, r^{n-1}, sr^0, sr, \ldots, sr^{n-1}\}$ (where s is an arbitrary reflection), together with the operation of function composition, indeed forms a group—the so-called **dihedral group**, denoted by D_n.

There are some important equalities to take notice of, but before going there, we make the following convention: $(r^i)^{-1} = r^{-i}$. This makes sense because, as we have seen, $(r^i)^{-1} = r^{n-i}$, and $r^i r^{n-i} = r^n = r^0$; but if we extrapolate this property of exponent summation to allow also negative exponents, we obtain $r^i r^{-i} = r^0$.

From the property that a reflection is its own inverse it also follows that:

$$sr^isr^i = r^0 \Longleftrightarrow sr^i = (sr^i)^{-1} = (r^i)^{-1}s^{-1} = r^{n-i}s \tag{2.27}$$

$$\Longleftrightarrow sr^i = r^{n-i}s \Longleftrightarrow sr^is^{-1} = r^{n-i}ss^{-1} \Longleftrightarrow \boxed{sr^is = r^{n-i}} \tag{2.28}$$

Now, we have seen above that $r^ar^b = r^{a+b \pmod n}$. There are other possibilities of solid motion combination. Here are some of them.

$$sr^asr^b = (sr^as)r^b = r^{n-a}r^b = r^{-a}r^b = r^{b-a} \tag{2.29}$$

Note that we could develop this in a alternative way: $r^{n-a}r^b \Longleftrightarrow r^{n-(a-b)}$ which is the inverse of r^{a-b}, and according to our convention, it equals r^{b-a}.

From $sr^is = r^{n-i}$ it also follows that $r^is = sr^{n-i}$. Moving on, we also have that:

$$r^asr^b = sr^{n-a}r^b = sr^{n-a+b} = sr^{b-a} \tag{2.30}$$

where the first equality is due to the previous property.

3 | Rings

3.1 Rings

A *ring* is an algebraic structure $(A, +, \cdot)$ such that $(A, +)$ is an abelian group, and $\cdot$ is associative and distributes over addition. If A contains an identity for the operation $\cdot$, it is called a *ring with unity*. $\lessdot$

Notations for $+$, $\cdot$, 0 **and** 1. Carrying over the similarity with the primordial ring, that of the integers $\mathbb{Z}$, we denote $+$ as addition and $\cdot$ as multiplication—although for a particular ring those operations could actually be something completely different. Moreover, for the same reason, the additive identity (which must always exist) is denoted by 0, and the (optional) multiplicative one, if it exists, is denoted by 1. (This is a significant departure from the case of groups, where the identity was usually denoted by e.) The additive inverse of an element a, is denoted $-a$. Obviously, these operations and elements might have nothing to do with their integer counterparts, but it significantly eases the notational burden.

The multiplicative identity, if it exists, is called an **unity**. Elements for which there exists a multiplicative inverse, if any, are called **units**.

Two types of multiplication. Speaking of notational burden, multiplication between elements of a ring, say between a and b, is often denoted by simply juxtaposing them, ab. This can cause confusion when want to represent the repeated addition of a, say, n times, which is sometimes done as na. To distinguish these two very different operations, the *latter one*, **repeated addition** of a, n times, will be represented as $n \cdot a$ (or $a \cdot n$). For example, we will write $a + a + a$ as $3 \cdot a$, or $a \cdot 3$. From this way of defining things, stems the next result.

Theorem 3.1. *Let a, b be elements of a ring R, and $s, t \in \mathbb{Z}$. We have: $(a \cdot s)(b \cdot t) = (st) \cdot (ab) = (ab) \cdot (st)$.*

Proof.

$$(a \cdot s)(b \cdot t) = \left(\sum_{i=1}^{s} a\right)\left(\sum_{j=1}^{t} b\right) \tag{3.1}$$

$$= \sum_{\substack{1 \le i \le s \\ 1 \le j \le t}} ab = (st) \cdot (ab) = (ab) \cdot (st) \tag{3.2}$$

■

Theorem 3.2. *For any ring, any element a multiplied by the* **additive** *identity (i.e. zero) equals that same identity.*

Proof. We have $0a = (0+0)a = 0a+0a$. Adding to both sides the additive inverse of $0a$, gives $0 = 0a$. For $a0$ the reasoning is analogous. ■

Theorem 3.3. *Given elements a, b of a ring R, we have:* $(-a)b = a(-b) = -ab$.

Proof. We have $(-a)b = (-a)b + ab - ab = (-a+a)b - ab = 0b - ab = -ab$. For $a(-b)$ the proof is similar. ■

Corollary 3.4. *Given elements a, b of a ring R, we have:* $(-a)(-b) = ab$.

Proof. $(-a)(-b) + a(-b) = (-a+a)(-b) = 0b = 0$, so $a(-b)$ is the additive inverse of $(-a)(-b)$. But by the previous theorem (3.3), $a(-b) = -ab$, and $-ab$ is the additive symmetric of ab. Hence $(-a)(-b)$ and ab have the same additive inverse, and so we must have $(-a)(-b) = ab$ (cf. corollary 2.3). ■

Corollary 3.5. *In a ring, the following holds:* $a(b-c) = ab - ac$ *and* $(b-c)a = ba - bc$.

Proof. $a(b-c) = a(b+(-c)) = ab + a(-c)$, which by theorem 3.3 equals $ab - ac$. For $(b-c)a$ the reasoning is similar. ■

Theorem 3.6. *For any ring with unity, the following hold:*

(i) $(-1)a = -a$.

(ii) $(-1)(-1) = 1$.

Proof. Note that -1 refers to the additive inverse of the multiplicative identity. For 1), $a + (-1)a = 1a + (-1)a = (1+(-1))a = 0a = 0$, and hence $(-1)a$ is the additive symmetric of a, which we denote as $-a$.

For 2), set $a = -1$ in 1), and recall that $-(-a) = a$.[1] Another way would be to set $a = b = -1$ in corollary 3.4. ■

3.2 Subrings

Unlike what happens with groups, subrings are not really that important—the more important notion turns out to be that of an *ideal*. But they are needed, and they are defined in the obvious way: given a ring R, a subset S of R is a subring if the axioms of rings also hold true for it. Of course, *whatever axioms we use for ring (e.g. with or without unity)* ***must also be used for the subring(s)!***

The following result applies to any ring, with or without unity.

Theorem 3.7. *Let $(R, +, \cdot)$ be a ring. A nonempty subset S of R is a* **subring** *if it is closed for $\cdot$ and $-$ (additive symmetric).*[2]

Proof. (→) If S is a subring, then it closed for $\cdot$, and as it is also an additive (sub)group, it is also closed for $-$ (cf. the subgroup test). (←) Conversely, the requirement that S be closed for $-$ implies, also via the subgroup test, that S will be an additive abelian (sub)group. Thus S is closed for $+$, and this, together with the assumption that it is closed for $\cdot$, shows that the distributivity laws hold for S (as it is a subset of R).[3] Associativity of $\cdot$ holds for the same reason. Hence S is a (sub)ring. ■

This is a very general definition, in particular because it doesn't require that R be a ring with unity (multiplicative identity). We could adapt the result for rings with unity, but that is a somewhat perilous course, because it is possible for a ring with unity, to have a subring also with unity—albeit a *different* unity:

Example 3.8. Consider the ring of integers modulo 6, $\mathbb{Z}_6$, and its subring $\{0, 2, 4\}$. The unity of $\mathbb{Z}_6$ is 1, but the unity of the subring is 4! ◊

3.3 Integral Domains

A specialisation of a ring, that sits between rings and fields, are *integral domains*.

Definition 3.9. *A commutative ring R with unity is an* **integral domain** *if it does* **not** *have any* **zero divisors***; i.e. if given any nonzero element $a \in R$, there is no nonzero element $b \in R$ such that $ab = 0$.*

For integral domains as defined above, the cancellation law holds: if $a \neq 0$, then $ab = ac \rightarrow b = c$. However we could have also defined integral domains as rings for which the cancellation law holds, as for such rings, there can be no zero divisors—so the two definitions are equivalent. This is shown in the next result.

Theorem 3.10. *Let a, b, c be elements of a ring, with $a \neq 0$. It has no zero divisors (i.e., it is an integral domain), if and only if the* **cancellation law** *holds, i.e., if $ab = ac$, then $b = c$.*

Proof. I will prove the equivalent statement that the ring has zero divisors if and only if the cancellation law *fails* to hold.

(→) Suppose the ring has zero divisors; in particular, suppose that $ab = 0$, and $b \neq 0$ (recall that a is already nonzero by hypothesis). Then the cancellation law fails: $ab = a0 = 0$, but $b \neq 0$.

(←) Suppose that cancellation fails, i.e., suppose that $ab = ac$, but that $b \neq c$. Then $a(b - c) = 0$, and $b - c \neq 0$. As by hypothesis, a is also nonzero, this establishes the existence of zero divisors. ■

A *field* is a ring which is also an abelian group for multiplication—i.e. where multiplication is commutative, and all elements except 0 have a multiplicative inverse. Obviously, all fields are rings—indeed, all fields are integral domains—they just have additional structure. We have the following result:

Theorem 3.11. *Any finite integral domain is a field.*

Proof. Take a nonzero element a. We must show that a has a multiplicative inverse. If $a = 1$ we are done, so assume $a \neq 1$. As the integral domain is finite, if we take the sequence of powers of $a, a^2, a^3, \ldots$, eventually we will hit repeated elements. Hence, there must exist positive integers i, j, with $i > j$ such that $a^i = a^j$. Cancelling, we obtain:

$$a^i = a^j \Leftrightarrow a^{i-j}a^j = 1a^j \Leftrightarrow a^{i-j} = 1 \tag{3.3}$$

If $i - j = 1$, that would mean that $a = 1$, which is not the case—so it must be that $i - j > 1 \Leftrightarrow i - j - 1 > 0$. And so $1 = a^{i-j} = a^{i-j-1}a$, meaning that a^{i-j-1} is the inverse of a. ■

3.4 Characteristic of a Ring

Definition 3.12. *Let R be a ring with unity. Its* **characteristic** *n is the least positive integer n such that $n \cdot 1 = 0$.*[4] *If no such integer exists, then the characteristic is 0.*

If a ring has characteristic n, then adding n times any of its nonzero elements, also yields 0: indeed let $x \neq 1$ be one such nonzero element. We have:

$$n \cdot x = n \cdot 1x = \underbrace{(1x + 1x + \cdots + 1x)}_{n \text{ times}} = \underbrace{(1 + 1 + \cdots + 1)}_{n \text{ times}}x = (n \cdot 1)x = 0x = 0 \tag{3.4}$$

For this reason, the **characteristic for a ring *without unity*** is defined as the smallest number of times we need to add *any* one of its elements to itself, in order to obtain 0.

Theorem 3.13. *The characteristic of an integral domain, is either* 0 *or a prime.*

Proof. We need to show that if the characteristic of an integral domain is positive, then it is prime. So let $n = st$ be the positive characteristic of an integral domain (with s, t, and of course n, positive integers). We have $0 = n \cdot 1 = st \cdot 1 = (s \cdot 1)(t \cdot 1)$, were the last equality is due to property 3.1. As we are dealing with an integral domain, one of $(s \cdot 1)$ or $(t \cdot 1)$ must be 0—but n is by hypothesis, the least positive integer such that $n \cdot 1 = 0$. Hence one of s or t must be equal to n, and the other is equal to 1 (the integer, not the unit of the integral domain). This entails the primality of n. ■

4 | Equations

The whole point of algebra, originally at least, was to solve equations...

4.1 Quadratic Equations

So we have an equation (in the reals) of the form $ax^2 + bx + c = 0$, and we want to solve it. Usually this requires factorising the expression. But how?

We could start by noting that $(x+d)^2 = x^2 + 2xd + d^2$. Since we want to obtain bx instead of $2dx$, we might set $b = 2d$, i.e. $d = b/2$:

$$(x + b/2)^2 = x^2 + bx + b^2/4 \tag{4.1}$$

This is closer to the goal, but what we want next is to have c instead of $b^2/4$:

$$x^2 + bx + c = 0 \Leftrightarrow x^2 + bx + b^2/4 + (c - b^2/4) = 0 \tag{4.2}$$

$$\Leftrightarrow (x + b/2)^2 = \frac{b^2 - 4c}{4} \Leftrightarrow x = -\frac{b}{2} \pm \frac{\sqrt{b^2 - 4c}}{2} \tag{4.3}$$

This is already pretty close. But suppose that in (4.1), we replace $b/2$ with $b/(2a)$. We obtain

$$\left(x + \frac{b}{2a}\right)^2 = x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} \tag{4.4}$$

which, if we multiply by a, yields:

$$a\left(x + \frac{b}{2a}\right)^2 = ax^2 + bx + \frac{b^2}{4a} \tag{4.5}$$

Our problem is now solved:

$$ax^2 + bx + c = 0 \Leftrightarrow ax^2 + bx + \frac{b^2}{4a} + \left(c - \frac{b^2}{4a}\right) = 0 \tag{4.6}$$

$$\Leftrightarrow a\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a} \Leftrightarrow \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \tag{4.7}$$

$$\Leftrightarrow x = -\frac{b}{2a} \pm \sqrt{\frac{b^2 - 4ac}{4a^2}} = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2|a|} \tag{4.8}$$

So if $a > 0$ (if $a = 0$ this is a degree 1 equation, and formula is inapplicable) we obtain:

$$\Leftrightarrow x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} = -\frac{b \pm \sqrt{b^2 - 4ac}}{2a} \tag{4.9}$$

If $a < 0$, then:

$$\Leftrightarrow x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2(-a)} = -\frac{b \mp \sqrt{b^2 - 4ac}}{2a} \tag{4.10}$$

But this means x takes exactly the same values as with formula (4.9)—which we take as the formula for the roots of second degree equations.

Notes

1. Basics

1. Observe that by construction, the set $\left\{x \cup \{s_{n+1}\} \,|\, x \in 2^S\right\}$ is in $2^{S'}$.
2. See theorem 1.1.3 in the Combinatorics report.
3. Equivalently, you can think of (1.12) as $\Phi'(z) = \varphi(r) \wedge \varphi(r+1) \wedge \cdots \wedge \varphi(z)$.
4. Of course, there might be more than two base cases; for example, for the associative property (§ 1.4.3), one of the base cases is $n = 3$, and it is a base case precisely because it does not follow from the previous cases, $n = 1$ and $n = 2$.
5. Also note that from the case $r = 0$, follow all others: if $\mathbb{N}$ has a smallest element (0), then so does every subset of $\mathbb{N}$.
6. We could also define a strict version of $[R]$, but it would be notationally cumbersome…
7. This could be prevented if R was asymmetrical—but then, by lemma 1.40, R would also be irreflexive, and hence it would coincide with R^*.

2. Groups

1. Note that s_j can be equal to s_i, when a is the identity.
2. There can be monoids in which one element has (say) two distinct left inverses (in which case said element cannot have a right inverse—why?).
3. If we set $n = 0$, then we get $a^{0+1} = aa^0 \iff a = aa^0 \iff a^0 = e$. Note that e is the group identity, sometimes represented by 1, yielding the perhaps more familiar form $a^0 = 1$.
4. The inversion property used in the second step—$(ab)^{-1} = b^{-1}a^{-1}$—is a particular case of a more general property, viz. that $(a_1 \cdots a_n)^{-1} = a_n^{-1} \cdots a_1^{-1}$, provided all the a_i are invertible. It is easily shown by induction.
5. Although the term *permutation* is usually reserved for *finite* sets, the ensuing discussion also applies if H has a (countably) infinite number of elements.
6. See the Number Theory wiki, section "Congruences", for more information on this group.
7. Keep in mind the cycles are permutations, which in turn are functions, and that function composition is associative.
8. See e.g., Judson [2], theorem 5.9, on p. 75, for the proof that any permutation can be written as a product of disjoint cycles.

3. Rings

1. Cf. theorem 2.13, which in additive notation is written precisely as $-(-a) = a$.

2. The reason why we cannot just require instead that S be closed for $+$ and $\cdot$ is that this is not sufficient to guarantee that S be an (additive) abelian group. Case in point: $\mathbb{Z}^+$ is closed for addition and multiplication, but is not an abelian group (it does not contain additive inverses).

3. In more detail, let a, b and c be elements of S. Then $a(b+c)$ is in S due to additive and multiplicative closure. And $ab+ac$ is also in S, due to multiplicative and additive closure. Similar reasoning applies to $(b+c)a$ and $ba+ca$. As the binary operation of S coincides with that of R, which must be well-defined—and as left and right distributivity hold in R—we conclude that the distributivity laws hold in S.

4. **Beware:** both the 0 and the 1 here refer to the additive and multiplicative identities, respectively! But in the next sentence, having characteristic 0, this 0 is an integer! Cf. the remarks at the beginning of the chapter about the two types of multiplications we have in rings (§3).

References

1. **Clark**, Allan (1984 [1971]). *Elements of Abstract Algebra*. New York: Dover Publications. ISBN: 978-0-486-64725-8. Cited on p. 24.
2. **Judson**, Thomas W. (2018). *Abstract Algebra*. Ann Arbor, Michigan: Orthogonal Publishing. ISBN: 978-1-944325-8. Cited on p. 46.
3. **Shoup**, Victor (2008). *A Computational Introduction to Number Theory and Algebra*, Second edition. Ebook version: Cambridge University Press. Cited on p. 9.

A | Miscellanea

A.1 Orders: Strict vs. non-strict

In the section on binary relations (§1.3) it is mentioned that given an order relation R, its strict (or *irreflexive*) counterpart, R^*, can be defined as [*] AppendixChapter

[*] foobar