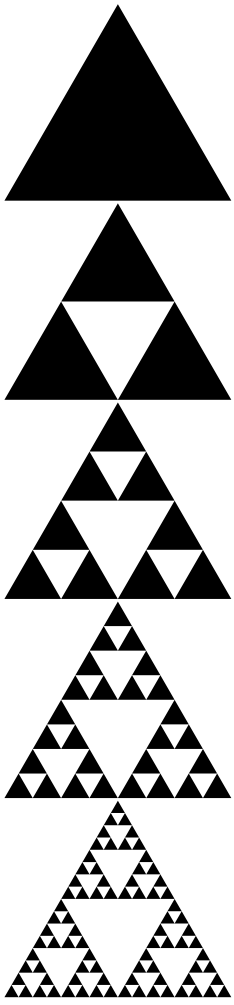
MATH 330: Lecture Notes

Binghamton University

David P Biddle

5/24/2022



1 Sets and Logic

A mathematical proof consists of a sequence of mathematical statements threaded together by rules of inference. We start by defining the simplest kind of mathematical statements, those that belong to propositional logic. A *statement* is a meaningful declarative sentence that is unamibiguously either *true* or *false*. For reference a declarative sentence should have a subject, a verb/predicate, and should end with a period. We can assign to a statement its truth value, which we will do here by assigning either 0 or 1 to the statement, and we write

this value by writing $tr(P) = 1$ if the statement P is true and $tr(P) = 0$ if the statement P is false. Propositional logic concerns itself with a way of putting simple statements together to form what we call *compound* statements, while at the same time providing a clear method to determine the truth value of these new statements. The way we put simple statements together is through the use of connectives; we introduce the fundamental logical connectives now.

Given statements P and Q we form the statement $P \wedge Q$, which is said as "P and Q", and we define its truth value by saying $tr(P \wedge Q) = 1$ exactly in the case $tr(P) = 1 = tr(Q)$ and in no other cases; the connective $\wedge$ is called logical *conjunction*. Likewise we can form the statement $P \vee Q$, which is said as "P or Q", and we define its truth value by saying $tr(P \vee Q) = 0$ exactly when both $tr(P) = 0 = tr(Q)$ and in no other cases; the connective $\vee$ is called logical *disjunction*. We emphasize here that our $\vee$ is inclusive in that $tr(P \vee Q) = 1$ if both $tr(P) = tr(Q) = 1$. There is another connective called the "exclusive disjunction" than we will not be using in these notes.

For any statement P we can form the statement $\neg P$, which is said as "not P", and we define its truth value by saying $tr(\neg P) = 1 - tr(P)$; the connective $\neg$ is called logical *negation*.

For the most part the three connectives $\wedge, \vee, \neg$ keep reasonably faithful to how most of us interpret the words "and", "or", and "not" in our regular use of language. If one says "David has a ferret and $1 + 1 = 3$", one is perfectly happy saying this statement is false since $tr(\text{David has a ferret}) = 1$ while $tr(1 + 1 = 3) = 0$, and we're more or less fine with the idea that the truth value of one of these simple statements does not affect the truth value of the other. David having ferrets is not the same subject as basic arithmetic unless we do something quite silly such as "David has a ferret Fergus and a ferret Freddie and one ferret plus one ferret equals three ferrets". We'll refrain from silliness and ferret arithmetic for the most part in these notes.

The next connective, implies, can cause a bit of confusion because of the way we tend to use 'implies' in our vernacular. Given statements P and Q we form the statement $P \Longrightarrow Q$, which is said "P implies Q", and we define its truth value by saying $tr(P \Longrightarrow Q) = 0$ exactly when $tr(P) = 1$ and $tr(Q) = 0$ and in no other cases. So $P \Longrightarrow Q$ is true unless P is true and Q is false. We note here that if $tr(P) = 0$ then $tr(P \Longrightarrow Q) = 1$ which might seem peculiar at first; we often say that $P \Longrightarrow Q$ is true vacuously in this case. We can think of the truth value of $P \Longrightarrow Q$ as being a promise rather than causality as in the following example: let P be the statement "your kid gets an A" and let Q be the statement "you buy them a car". Then we interpret $P \Longrightarrow Q$ as the promise "If your kid gets an A then you buy them a car" in the four scenarios as follows:

- If $tr(P) = 1$, your kid gets an A, and $tr(Q) = 1$, you buy them a car, you've honored the promise.
- If $tr(P) = 1$, your kid gets an A, and $tr(Q) = 0$, you don't but them a car, you broke your promise.
- If $tr(P) = 0$, your kid doesn't get an A, and $tr(Q) = 1$, you buy them a car, you've honored the promise because you haven't broken it. You only said you'd buy them a car if they did get an A, you never said anything about what would happen if they didn't get an A. You might have some parenting issues however.
- If $tr(P) = 0$, your kid doesn't get an A, and $tr(Q) = 0$, you don't buy them a car, there's no broken promise either.

In $P \implies Q$ there may be absolutely no relationship between P and Q just as in $P \wedge Q$ or $P \vee Q$. For example if P represents "$1 + 1 = 2$ and Q represents "π is irrational" then the truth value of $P \implies Q$ or "If $1 + 1 = 2$ then π is irrational" still just depends on the truth values of P and Q and not any perceived relationship between the statements. There are a lot of different ways we phrase implications in regular language: we might say "Q is necessary for P" or "Q follows from P" or "P only if Q" for $P \implies Q$.

Sometimes implications statements are a bit more subtle. For instance "all triangles are equilateral" and "the square of a real number is non-negative" might not seem like implications but can be rephrased as such: "if T is a triangle, then T is equilateral" and "if x is a real number, then $x^2 \geq 0$", have the same meaning as the aforementioned statements.

We add here one more logical connective that is more of a convenience than a necessity. Given statements P and Q we form the statement $P \iff Q$, which is said as "P if and only if Q, and we define its truth value by saying $tr(P \iff Q) = 1$ if $tr(P) = tr(Q)$ and in no other cases; the connective $\iff$ is called the logical biconditional. Other ways we may use this connective include 'P happens exactly when Q happens' or 'P and Q are logically equivalent'. The phrase "if and only if" is commonly abbreviated to "iff".

The use of the numbers 0 or 1 as truth values allows us to use a bit of algebra in our determination of truth values. In order for our arithmetic to make sense we define two operations on the symbols 0 and 1 which we call addition and multiplication:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Let $tr(P) = a, tr(Q) = b$. Then we have that $tr(\neg P) = 1 + tr(P) = 1 + a$ and $tr(P \wedge Q) = ab$ since if either P or Q is false then $P \wedge Q$ is false and $P \wedge Q$ is true exactly when both P and Q are true, the product $ab = a \cdot b$ completely captures these with multiplication. Finding $tr(P \vee Q)$ we see that $a + b + ab$ does the job since if both P and Q are false this is 0 and if P is true and Q is false this is $1 + 0 + 0 = 1$ and likewise if P is false and Q is true this is $0 + 1 + 0 = 1$. Lastly if P is true and Q is true this is $1 + 1 + 1 \cdot 1 = 0 + 1 = 1$. Note that the polynomials ab and $a + b + ab$ are symmetric in a and b in the sense that if we switch the roles of a and b we end up with the same expression. We expect this is not the same for implication. $tr(P \implies Q) = 1 + a + ab$ which works since if P is false this is 1 and if P is true and Q is true this is $1 + 1 + 1 = 0 + 1 = 1$ and lastly if P is true and Q is false this is $1 + 1 + 0 = 0 + 0 = 0$. Note that if we switch the roles of a and b in the polynomial $1 + a + ab$ we do not get the same polynomial, that order is very important to understanding this connective.

What about $tr(P \iff Q)$? This should certainly be symmetric and this is similar to $tr(P \wedge Q)$ in that $tr(P \iff Q) = a + b + 1$; we leave it to the reader to verify this.

Suppose we are given a compound statement Q that is constructed using various connectives applied to the simple statements $P_1, P_2, \ldots P_n$. We'd like the determine the truth value of Q based on the truth values of the P_i fairly systematically. One way of doing this that entertains

all possible truth values of Q, is by constructing a *truth table*. We give an example first before describing the process in general.

Suppose we have the compound statement:

$$Q := ((P_1 \implies P_2) \iff (\neg P_2 \implies \neg P_1)).$$

We build a table to look at all possible combinations of $tr(P_1), tr(P_2)$:

P_1	P_2	$\neg P_1$	$\neg P_2$	$P_1 \implies P_2$	$\neg P_2 \implies \neg P_1$	Q
1	1	0	0	1	1	1
1	0	0	1	0	0	1
0	1	1	0	1	1	1
0	0	1	1	1	1	1

We note the first two columns consist of all 4 truth combinations of $tr(P_1)$ and $tr(P_2)$ set up in a way that the rows are ordered *lexicographically* with respect to the order $1 > 0$ based on the truth values in these two columns: considering just the row entries in the first two columns, we arrange the rows so that if a row R has truth values a, b from left to right and R' has truth values a', b' we write $R > R'$ (and put R above R') if either

1. $a > a'$ or
2. $a = a'$ and $b > b'$.

In general if Q consists of n simple statements $P_1, P_2, \ldots, P_n$ we have 2^n possible combinations of truth values $tr(P_i)$, and we construct the first two columns of the truth table so that the rows are ordered so that if the row R has truth values $a_1, a_2, \ldots, a_n$ and row R' has truth values $a'_1, a'_2, \ldots, a'_n$ we say $R > R'$ (and put R above R' in the table) if either

$$a_1 > a'_1$$

or

$$a_1 = a'_1 \quad \text{and} \quad a_2 > a'_2$$

or

$$a_1 = a'_1, a_2 = a'_2 \quad \text{and} \quad a_3 > a'_3$$

and so on. We then construct the remainder of the columns of the table by analyzing each connective systematically.

Example 1.1. Suppose our compound statement is $Q := ((P_1 \implies P_2) \vee P_3) \wedge P_2$. Our truth table will consist of $2^3 = 8$ rows and because the statement Q is built up from three statements and three connectives, the table will have $3 + 3 = 6$ columns as follows:

P_1	P_2	P_3	$P_1 \implies P_2$	$(P_1 \implies P_2) \vee P_3$	Q
1	1	1	1	1	1
1	1	0	1	1	1
1	0	1	0	1	0
1	0	0	0	0	0
0	1	1	1	1	1
0	1	0	1	1	1
0	0	1	1	1	0
0	0	0	1	1	0

Definition 1.2. A statement without any explicit connectives is called an *atomic statement*. A *well-formed statement* is a statement Q that is defined recursively as follows: either Q atomic or Q is formed by applications of the connectives $\neg, \wedge, \vee, \implies$ to previously established well-formed statements P_1 and P_2 in the sense that if we know that P_1 and P_2 are well-formed we also know that the following are well-formed:

$$(\neg P_1) \quad (P_1 \wedge P_2) \quad (P_1 \vee P_2) \quad (P_1 \implies P_2).$$

We encourage the use of parentheses when building well-formed statements each time a connective is introduced but are content with leaving off an outer pair of parenthesis; note the example $Q := ((P_1 \implies P_2) \vee P_3) \wedge P_2$ which we consider well-formed if P_1, P_2, P_3 are atomic since it can be constructed in stages: first $(P_1 \implies P_2)$, second $((P_1 \implies P_2) \vee P_3)$, and lastly $(((P_1 \implies P_2) \vee P_3) \wedge P_2)$.

We will only be dealing with well-formed statements in these notes. This avoids a range of issues, for instance grammatically awkward or possibly ambiguous phrases like "$Q := S \vee T \wedge U$"; even if S, T, U are atomic it's unclear what the truth value of Q is given the truth values of S, T, U. Both $(S \vee T) \wedge U$ and $S \vee (T \wedge U)$ on the other hand are well-formed. Truth tables can always be constructed for well-formed statements, and the truth value of a well-formed statements can clearly and unambiguously be obtained from the truth value of its atomic components. From this point onward we will use the word 'statement' to mean 'well-formed statement'.

Definition 1.3. Suppose we have two statements Q_1 and Q_2 with $tr(Q_1) = tr(Q_2)$. We say that Q_1 and Q_2 are logically equivalent or propositionally equivalent and write $Q_1 \equiv Q_2$.

We note here that unless Q_1 and Q_2 are literally the same statement it is bad form to write $Q_1 = Q_2$. This is a very common mistake people make when learning propositional logic, we point it out to give the reader caution with the similarity of the symbols $=$ and $\equiv$.

The way we determine whether two statements Q_1, Q_2 are logically equivalent is by constructing a truth table which builds up both statements by considering all possible assignments of truth values in the atomic statements making up each statement Q_1, Q_2 and seeing if the columns under these statements are identical.

Example 1.4. We establish that $Q_1 := \neg(P_1 \wedge P_2)$ and $Q_2 := (\neg P_1) \vee (\neg P_2)$ are logically equivalent:

P_1	P_2	$P_1 \wedge P_2$	$Q_1 = \neg(P_1 \wedge P_2)$	$\neg P_1$	$\neg P_2$	$Q_2 = (\neg P_1) \vee (\neg P_2)$
1	1	1	0	0	0	0
1	0	0	1	0	1	1
0	1	0	1	1	0	1
0	0	0	1	1	1	1

We see that the columns corresponding to Q_1 and Q_2 are identical and so $Q_1 \equiv Q_2$. We encourage the reader to build truth tables for $\neg(P_1 \vee P_2)$ and $(\neg P_1) \wedge (\neg P_2)$ to show that these statements are also logically equivalent.

Definition 1.5. A statement that is always true is called a *tautology*. Thus if the column in a truth table for such a statement consists solely of 1's.

There are lots of simple tautologies such as $P \implies P$ and $P \vee \neg P$ for example as well as more complicated tautologies. Our previous example $((P_1 \implies P_2) \iff (\neg P_2 \implies \neg P_1)$ is also a tautology. We could also have verified in this example that the statements $P_1 \implies P_2$ and $\neg P_2 \implies \neg P_1$ are logically equivalent. We can phrase logical equivalence in terms of the connective $\iff$ in the sense that if $Q_1 \equiv Q_2$ then $Q_1 \iff Q_2$ is a tautology and vice a versa. For a list of useful and common tautologies click HERE.

A proof or more accurately a *deductive* proof of a statement Q is a process that starts with statements that are assumed or known to be true and concludes with Q through some valid method of reasoning. The methods of reasoning we will be most concerned with involve 'rules of inference'. A rule of inference is a rule that allows us to infer a conclusion from a premise in order to create an argument. Most rules of inference we will use stem from one that is called *modus ponens* in Latin. Modus ponens is the rule of inference which allows us to take the information P is true and $P \implies Q$ is true and allows us to infer that Q is true. One may associate to this rule of inference the tautology

$$(P \wedge (P \implies Q)) \implies Q,$$

which we ask the reader to verify with a truth table. Modus ponens is not the same as this tautology; the tautology is a statement that is always true which is distinct from the step by step process of deduction, but we will from time to time blur this distinction between tautology . We can use modus ponens to deduce other rules of inference: suppose we know that Q is false and that $P \implies Q$ is true; then we can infer that P must be false as well since if P were true then by modus ponens we would have Q is true contradicting our assumption that Q is false. This rule of inference is called *modus tollens* and is simply an application of modus ponens. We will see that this is the basis for 'proof by contrapositive'. Another natural rule of inference we will use frequently in deductive reasoning/deductive proofs is as follows: if both $(P \implies Q)$ and $(Q \implies R)$ are true then we can infer that $(P \implies R)$ is true; this is known as hypothetical syllogism and again stems from a straightforward tautology. We will not focus so heavily on the names of classical rules of inference much beyond what has already been stated and a few others discussed below, but the interested reader can click HERE or check out any text on mathematical logic for more.

Example 1.6 (Direct Proof)**.** Suppose we want to prove that "if x is a real number and $x > 1$ then $x^2 > x$". We proceed as follows: if x is real and $x > 1$ then x and $x - 1$ are both positive, and hence so is $x(x - 1)$. But then $x^2 - x > 0$ or $x^2 > x$.

In addition to modus ponens there are several other rules of inference which we will briefly outline. Each of these does have a background tautology in tow and so is 'reasonable' in the sense that modus ponens is.

Inference/Proof by Contrapositive: In order to show that $P \Longrightarrow Q$ is true it is enough to assume that Q is false and then show that P is false. Associated to this rule of inference is the tautology

$$(P \Longrightarrow Q) \Longleftrightarrow (\neg Q \Longrightarrow \neg P).$$

Inference/Proof by Contradiction: Assume that R is a false statement. Then in order to show that $P \Longrightarrow Q$ then it is enough to assume P is true and Q is false and derive the false statement R; this is based on the tautology $(P \Longrightarrow Q) \Longleftrightarrow ((P \wedge \neg Q) \Longrightarrow R)$ where R only takes on the truth value 0.

Disjunctive Conclusion Proof: In order to show that $A \Longrightarrow (B \vee C)$ is true it suffices to either:

- Assume A and $\neg B$ are both true and deduce C is true.
- Assume A and $\neg C$ are both true and deduce B is true.

This rule of inference is associated with the tautologies $(A \Longrightarrow (B \vee C)) \Longleftrightarrow ((A \wedge \neg B) \Longrightarrow C)$ and $(A \Longrightarrow (B \vee C)) \Longleftrightarrow ((A \wedge \neg C) \Longrightarrow B)$

Disjunctive Hypothesis Proof: In order to show that $A \vee B \Longrightarrow C$ is true it suffices to show that both $A \Longrightarrow C$ and $B \Longrightarrow C$ are true. This rule of inference is associated with the tautology $(A \vee B \Longrightarrow C) \Longleftrightarrow ((A \Longrightarrow C) \wedge (B \Longrightarrow C)$.

We encourage the reader to verify the above tautologies. There are other rules of inference besides the ones listed above, we've simply listed the most common rules of inference/proof methods we'll be using. Others will be discussed as necessary in these notes when appropriate. To construct a mathematical proof, namely to show that a 'proposition' or 'theorem' is true, one needs to begin with a collection of assumed statements (the hypotheses) along with other already understood to be true statements and definitions, and then find a sequence of statements connected by rules of inference leading to the conclusion statement.

In practice the above is often accomplished in a few slightly different but related ways. Let's say one is trying to establish $P \Longrightarrow Q$. The first way to do this is to lay out everything you can about statement P along with the statements of any definitions and any relevant previously established theorems and deduce from this entirety that the statement Q must be true. This is called the *forward way* of proving $P \Longrightarrow Q$. In this proof method one begins by looking at P and then produces a sequence of statements $P_1, P_2, \ldots P_n$ along with compatible rules of inference so that $P \Longrightarrow P_1$ and $P_1 \Longrightarrow P_2$, and so on until one infers $P_n \Longrightarrow Q$ and then connects these via hypothetical syllogism to get $P \Longrightarrow Q$. Here the progression is to start at P find P_1 then P_2 etc.... until arriving at Q

It might be more natural to instead of focusing extensively on P to focus on both P and Q and look for a statement P' so that one can focus on inferring $P \Longrightarrow P'$ and $P' \Longrightarrow Q$ (and hence $P \Longrightarrow Q$ as above) and possibly breaking these down further into maybe $P \Longrightarrow P''$ and $P'' \Longrightarrow P'$ and maybe $P' \Longrightarrow Q'$ and $Q' \Longrightarrow Q$ etc... The idea here is that one works a bit

forward from P and a bit backwards from Q and continues until a sequence of statements and appropriate inferences are made. This is called the *forward-backward way* of proving $P \Longrightarrow Q$.

Often times it is useful for us to have something similar to a statement but where one or more objects in the statement are indeterminate variables instead of fixed objects. A few examples of these more general types of statements are: "$x > 3$", "s is a number that equals $2t+1$". We can't in general determine if declarative sentences like these are true or false without being able to substitute in specific values for the variables. Let's let $Q(x) :=$ "$x > 3$". Then $Q(2)$ is the atomic statement "$2 > 3$", and we can infer that $tr(Q(2)) = 0$. Here it is quite reasonable to let $x = 2$ here but it would be a bit absurd to let $x =$ David. We want to make sure that there is a collection of allowable substitutions, and this gives rise to the concept of a predicate.

Definition 1.7. A *predicate* $P = P(x_1, x_2, \ldots, x_n)$ is a sentence that contains a finite number of variables and becomes a statement when specific values are substituted for the variables. The *domain* D_i of a predicate variable x_i is the set of all values that may be substituted in place of the variable. For each a_1 in D_1, a_2 in D_2, etc.... the domain we get a statement which we write $P(a_1, a_2, \ldots)$ so that either $tr(P(a_1, a_2, \ldots)) = 0$ or $tr(P(a_1, a_2, \ldots)) = 1$.

One of the most fundamental objects in mathematics is that of a *set*. A set is, very loosely, a collection of objects called its members or elements. The key relationship among sets is the relationship 'is a member of' or 'is an element of'. We denote most sets using capital letters like $X, Y, A, B, \ldots$. etc (though this is not necessary) and the relationship 'is an element of' using the symbol $\in$ and write $a \in X$ as shorthand for "a is an element of X". We'll be discussing a fair number of sets in this course that you are likely to have already seen in various informal capacities such as the integers or real numbers but with more of an emphasis on rigorous mathematical properties of these sets. In particular in these notes we refer to the natural numbers as the set $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, the positive natural numbers $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$, and the integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, \ldots\}$. The set of rational numbers is written $\mathbb{Q}$ and the set of real numbers $\mathbb{R}$; we will discuss these sets in much more detail in the course of the notes. Given a predicate statement $P(x)$ and a set X all of whose elements are in the domain of the predicate variable, we can form a new set $\{a \in X \,|\, P(a)\}$.

Example 1.8. Let $P(x) :=$ "$x < 3$" and let the domain D of this predicate be any numbers for which is makes sense to posit whether that number is less than 3. Let $X = \mathbb{N} = \{0, 1, 2, 3....\}$ the set of natural numbers. Then $\{a \in \mathbb{N} \,|\, P(a)\} = \{0, 1, 2\}$ while if we use a different set say $\mathbb{R}$ we might build a different set using this predicate: $\{a \in \mathbb{R} \,|\, P(a)\} = (-\infty, 3)$.

Let's focus for the moment on predicates $P(x), Q(x)$ which we assume to have the same domain (so the range of objects x for these predicates is the same). We can naturally extend our connectives $\neg, \wedge, \vee, \Longrightarrow$ to predicates $P(x), Q(x)$ by simply forming predicates:

$$\neg(P(x), P(x) \wedge Q(x), P(x) \vee Q(x), P(x) \Longrightarrow Q(x).$$

Given a predicate $P(x)$ we would like to be able to build certain statements exerting the existence of an object a in the domain of the predicate so that $P(a)$ is true; likewise we would also find it useful to establish that a predicate statement is always true over some collection of

objects in the domain of the predicate. Before we do this we introduce the concept of 'subset' and 'equality' of sets.

Definition 1.9. We say that $X \subseteq Y$ if $(a \in X \implies a \in Y)$ is true, and we say X is a *subset* of Y. Two sets X, Y are equal, written $X = Y$, exactly when the statements $X \subseteq Y$ and $Y \subseteq X$ are both true. Because $((P \implies Q) \wedge (Q \implies P)) \iff (P \iff Q)$ is a tautology we can say $X = Y \iff ((a \in X) \iff (a \in Y))$. In this very strong way sets are determined by their elements.

Example 1.10. Let $X_1 = \{a \in \mathbb{N} \,|\, a > 5\}$ and $X_2 = \{b \in \mathbb{N} \,|\, \exists c \in \mathbb{N}, b = c + 6\}$. We show that $X_1 = X_2$. Let $y \in X_1$ so that $y \in \mathbb{N}$ and $y > 5$. Then $y - 5 > 0$ and $y - 6 \in \mathbb{N}$. Set $c = y - 6$. Then $\exists c \in \mathbb{N}$ so that $c + 6 = (y - 6) + 6 = y$ and so $y \in X_2$ and hence $X_1 \subseteq X_2$. Let $x \in X_2$. Then $x \in \mathbb{N}$ and $\exists c \in \mathbb{N}$ so that $x = c + 6$. Since $x = c + 6 \geq 6$ we know that $x > 5$ and so $x \in X_1$ and we have $X_2 \subseteq X_1$ and so $X_1 = X_2$. Here we have $X_1 = X_2 = \{6, 7, 8, \dots\}$.

Suppose we have a predicate statement $P(x)$ with domain U. If $X \subseteq U$ we define the truth value of the statement $(\exists x \in X)P(x)$ to be 1 if $P(a)$ is true for some $a \in X$ and 0 otherwise. We call $\exists$ the *existential quantifier* and read $(\exists x \in X)P(x)$ as "there exists an x in X so that $P(x)$". Similarly we define $(\forall x \in X)P(x)$ to have truth value 1 if $P(a)$ is true as a ranges over every single member of the set X and 0 if $P(a)$ is false for any $a \in X$. We call $\forall$ the *universal quantifier* and read $(\forall x \in X)P(x)$ as "for all x in X, $P(x)$".

Example 1.11. Let $P(x) :=$ "$x < -3$" with domain the set of real numbers $\mathbb{R}$. Let $X_1 = \mathbb{N}, X_2 = \mathbb{Z}, X_3 = \{-\pi\}$. Then the statement $(\exists x \in X_1)P(x)$ is false since every natural number is non-negative; $(\exists x \in X_2)P(x)$ is true since $-10 \in X_2$ with $P(-10)$ being true; $(\exists x \in X_3)P(x)$ is also true since $-\pi \in X_3$ and $P(-\pi)$ is true. The statement $(\forall x \in X_1)P(x)$ is false since $0 \in X_1$ and $P(0)$ is false; the statement $(\forall x \in X_2)P(x)$ is false since $0 \in X_2$ and $P(0)$ is false; the statement $(\forall x \in X_3)P(x)$ is true however since there is only one element $-\pi$ of X_3 and $P(-\pi)$ is true.

If the domain of the predicate $P(x)$ say U is clear and we have no interest in restricting the predicate to a subset of the predicate's domain then we will often just write $(\exists x)P(x)$ and $(\forall x)P(x)$ instead of $(\exists x \in U)P(x)$ and $(\forall x \in U)P(x)$. We can also extend the idea of existential and universal quantification to include predicates $P(x_1, x_2, \dots, x_n)$ with more than one variable ($n \geq 2$) each with respective domains U_i as follows. Let $(\exists x_1)P(x_1, x_2, \dots, x_n)$ simply be the obvious predicate with one less variable (namely variables $x_2, x_3, \dots, x_n$).

Let $(\forall x_1)P(x_1, x_2, \dots, x_n)$ be the obvious predicate with one less variable. For example if our predicate is $P(a, b, c) :=$ "$a^2 + b^2 = c^2$" with domain $U = \mathbb{Z}$ then $(\exists a)P(a, b, c)$ is a predicate with two variables b, c and $(\exists a)(\exists b)P(a, b, c)$ is a predicate with one variable c and $(\exists a)(\exists b)(\exists c)P(a, b, c)$ is a statement, which happens to be true since $P(3, 4, 5)$ is true for instance.

We will often deal with statements that involve a mix of existential and universal quantifiers and we want to emphasize that in general the order in which the quantifiers are listed is paramount. For instance if $P(a, b) :=$ "$a < b$" with domain $\mathbb{R}$ we can form eight different statements using two quantifiers:

$$S_1 := (\forall a)(\forall b)(a < b)$$
$$S_2 := (\forall b)(\forall a)(a < b)$$
$$S_3 := (\forall a)(\exists b)(a < b)$$
$$S_4 := (\forall b)(\exists a)(a < b)$$
$$S_5 := (\exists a)(\forall b)(a < b)$$
$$S_6 := (\exists b)(\forall a)(a < b)$$
$$S_7 := (\exists a)(\exists b)(a < b)$$
$$S_8 := (\exists b)(\exists a)(a < b)$$

Note that S_1 and S_2 are saying the same thing; we will abbreviate these equivalent statements by writing $(\forall a, b)(a < b)$. Similarly S_7 and S_8 are saying the same thing;we will abbreviate these equivalent statements by writing $(\exists a, b)(a < b)$. S_1, S_2 are false whereas S_7, S_8 are true. The rest are a bit more delicate. S_3 is exerting that given any real number a we can find a bigger real number b, which is true (we could let $b = a + 1$ for instance). If we switch the order of these two quantifiers we get statement S_6 which claims that there is a real number b that is bigger than every other real number a, clearly false statement (b is not bigger than $a = b + 1$ for instance). S_4 exerts that given any real number b we can find smaller real number a, a true statement as we can always set $a = b - 1$. S_5 states that there is a real number a that is less than every other real number b, a false statement since we could look at $b = a - 1$. Order matters most of the time here unless all the quantifiers are of the exact same type.

Looking back at the connectives $\wedge$ and $\vee$ we see that they are somewhat complementary when we mix them with the negation connective in the sense that we have DeMorgan's tautologies $\neg(P \wedge Q) \iff (\neg P \vee \neg Q)$ and $\neg(P \vee Q) \iff (\neg P \wedge \neg Q)$; negation 'switches' these two connectives. Something similar happens when we look at how the negation connective works with universal and existential quantifiers. We note that given a predicate $P(x)$ we have the following logical equivalences:

$$[\neg(\forall x)P(x)] \equiv (\exists x)(\neg P(x))$$

and

$$[\neg(\exists x)P(x)] \equiv (\forall x)(\neg P(x))$$

Let $Q(x, y, z) :=$ "$x + y = z$" where our domain is the integers $\mathbb{Z}$. Let's examine the statement $(\forall x)(\forall z)(\exists y)Q(x, y, z)$ which states that given any two integers x and z we can find a third integer y so that $x + y = z$ (which happens to be true since we can let $y = z - x$ and the y defined this way is an integer). The negation of this statement, the statement $\neg(\forall x)(\forall z)(\exists y)Q(x, y, z)$ is logically equivalent to each of the following statements:

$$(\exists x)\neg(\forall z)(\exists y)Q(x, y, z)$$
$$(\exists x)(\exists z)\neg(\exists y)Q(x, y, z)$$

and

$$(\exists x)(\exists z)(\forall y)\neg Q(x, y, z)$$

where the last statement is $(\exists x)(\exists z)(\forall y)(x + y \neq z)$.

A key idea when working with sets is that we can build new sets from already understood sets by means of set operations of various types, several of which have a close relationship to the logical connectives and quantifiers of predicate logic. Three of the most fundamental set operations are that of union, intersection and set difference.

Definition 1.12. Let X, Y be sets. Define $X \cup Y := \{a \,|\, a \in X \vee a \in Y\}$ the union of X and Y and $X \cap Y := \{a \,|\, a \in X \wedge a \in Y\}$ the intersection of X and Y. Note that because $P \wedge Q \iff Q \wedge P$ and $P \vee Q \iff Q \vee P$ are tautologies, $X \cup Y = Y \cup X, X \cap Y = Y \cap X$. We define $X - Y := \{a \in X \,|\, a \notin Y\}$ and call this X minus Y. The language 'the difference of X and Y tends to cause a bit of ambiguity, it's not clear whether this refers to $X - Y$ or $Y - X$, and in general these two sets are not the same. $X - Y$ is a subset of X while $Y - X$ is a subset of Y.

Theorem 1.13. *Let A, B, C be sets. Then $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.*

Proof. We prove the first equality $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and leave the proof of the second inequality as a highly recommended exercise for the reader. Let $x \in A \cup (B \cap C)$. This happens iff $(x \in A) \vee (x \in B \cap C)$ by the definition of union; this in turn is equivalent to $(x \in A) \vee (x \in B \wedge x \in C)$ by the definition of intersection. Let $P :=$ "$x \in A$", $Q :=$ "$x \in B$", $R :=$ "$x \in C$". Then our statement is $P \vee (Q \wedge R)$. This is logically equivalent to $(P \vee Q) \wedge (P \vee R)$ or $(x \in A \vee x \in B) \wedge (x \in A \vee x \in C)$, which in turn is equivalent to $x \in (A \cup B) \cap (A \cup C)$ by the definition of union and intersection. So we have shown that $x \in A \cup (B \cap C) \iff x \in (A \cup B) \cap (A \cup C)$ and so $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. □

We say a set X is empty (or an empty set) if $\neg(a \in X)$ or $a \notin X$ is always true. In other words a set X is empty if $(\forall a)(a \notin X)$ is true. By our remarks on quantifiers $(\forall a)(a \notin X) \equiv \neg(\exists a)(a \in X)$.

Theorem 1.14 (Empty Set Theorem)**.** *If X and Y are both empty sets then $X = Y$.*

Proof. Given empty sets X and Y we need to prove the implications $(a \in X) \implies (a \in Y)$ and $(a \in Y) \implies (a \in X)$. The implication $(a \in X) \implies (a \in Y)$ is equivalent to its contrapositive $\neg(a \in Y) \implies \neg(a \in X)$, the latter being a true implication. Similarly we see that $(a \in Y) \implies (a \in X)$ is true. So we have that $X \subseteq Y$ and $Y \subseteq X$ by the definition of subset and lastly that $X = Y$ by the definition of set equality. □

The above theorem says that we can refer to *the* empty set, and we will write $\emptyset$ or $\{\}$ for this unambiguous set. We encourage the reader to provide a simple proof that given any set X we have $\emptyset \subseteq X$. Is there a 'largest' set? That's a murky question that gets into various logical issues at foundations of mathematics...the short answer is no. The discussion of sets in these notes avoids a lot of these important questions to favor a somewhat lighter introduction to the study of sets. For a more in depth look at set theory, in particular various forms of *axiomatic set theory* we encourage the reader to click HERE. In a manner quite similar to how we sometimes treated domains of various predicates, it's often convenient to restrict ourselves

to some 'universe' or 'universal set' if we are only focusing on objects of a particular type. We often denote such sets by U and make sure this set is very clearly mentioned or defined. This allows us to establish another useful set operation.

Definition 1.15. Given a universal set U and $X \subseteq U$ we define the complement of X (with respect to U) to be the set $X^c := U - X$, that is, the set of all elements in the universal set not in the set X. We see immediately that $U^c = \emptyset, \emptyset^c = U$ and that for any $X \subseteq U$ $(X^c)^c = X$.

Lemma 1.16. *Let $A, B \subseteq U$. Then $A \subseteq B \iff B^c \subseteq A^c$.*

Proof. We note that the contrapositive of the implication $x \in A \implies x \in B$ is the statement $x \notin B \implies x \notin A$, in other words:

$$(x \in A \implies x \in B) \iff (x \in B^c \implies x \in A^c)$$

which is exactly the statement of the lemma by definition of subset. □

Theorem 1.17 (DeMorgan's Theorem for sets)**.** *Let $A, B \subseteq U$. Then $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.*

We give two proofs of these two equalities.

Proof 1. Let $y \in (A \cup B)^c$. This happens iff $\neg(y \in A \cup B)$ by definition of complement. In turn this happens iff $\neg(y \in A \vee y \in B)$ by definition of union. This is logically equivalent to the statement $\neg(y \in A) \wedge \neg(y \in B)$ by DeMorgan's tautology. This happens iff $y \notin A \wedge y \notin B$ which in turn happens iff $y \in A^c \cap B^c$ by definition of intersection and complement. Hence $y \in (A \cup B)^c \iff y \in A^c \cap B^c$ and so $(A \cup B)^c = A^c \cap B^c$.

Let $y \in (A \cap B)^c$. This happens iff $\neg(y \in A \cap B)$ by definition of complement. In turn this happens iff $\neg(y \in A \wedge y \in B)$ by definition of intersection. This is logically equivalent to the statement $\neg(y \in A) \vee \neg(y \in B)$ by DeMorgan's tautology. This happens iff $y \notin A \vee y \notin B$ which in turn happens iff $y \in A^c \cup B^c$ by definition of union and complement. Hence $y \in (A \cap B)^c \iff y \in A^c \cup B^c$ and so $(A \cap B)^c = A^c \cup B^c$. □

Proof 2. We see that $A \cap B \subseteq A, B \subseteq A \cup B$ and so by an earlier lemma, $(A \cup B)^c \subseteq A^c, B^c \subseteq (A \cap B)^c$. From this we get that $(A \cup B)^c \subseteq A^c \cap B^c$ and $A^c \cup B^c \subseteq (A \cap B)^c$. Replacing A, B with A^c, B^c respectively in the previous two statements we get $(A^c \cup B^c)^c \subseteq A \cap B$ and $A \cup B \subseteq (A^c \cap B^c)^c$. We apply the lemma again to get that $(A \cap B)^c \subseteq A^c \cup B^c$ and $A^c \cap B^c \subseteq (A \cup B)^c$ and so we get $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$. □

We can generalize the ideas of union and intersection of two sets to an arbitrary collection of sets. Suppose we had

$$S = \{\{1, 2, 3, 4\}, \{4, 6\}, \{1, 2, 4, 6, 7, 8\}, \{4, 9\}\}.$$

We could form a new set which we will denote $\cap S$ to be the intersection of all elements of S: $\cap S = \{4\}$. We could also form another new set which we will denote $\cup S$ to the union of the elements of S: $\cup S = \{1, 2, 3, 4, 6, 7, 8, 9\}$. We make these definitions precise:

Definition 1.18. Let S be a collection of sets. We define the intersection of S, written $\cap S$, as follows: $x \in \cap S$ if $\forall A \in S$, $x \in A$. We define the union of S, written $\cup S$, as follows: $x \in \cup S$ if $\exists A \in S$, $x \in A$. We encourage the reader to state and prove more general forms of DeMorgan's Theorems for union and intersection of collections of sets. We can think of $A \cup B$ as $\cup\{A, B\}$ and $A \cap B$ as $\cap\{A, B\}$.

Example 1.19. Suppose we have

$$S = \{\{1, 2, 3, 4, 5, 6, 7\}, \{3, 5, 7, 9, 11\}, \{4, 5, 6, 7, 8, 9\}\}$$

. Then $\cap S = \{5, 7\}$ since the elements 5 and 7 are the only elements common to all three members of S. We see that $\cup S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 11\}$.

Example 1.20. Let $X = \mathbb{R}$ be the universal set of real numbers, and let $T = \{(-\infty, 1], (-1, 2], [1, \infty)\} = \{A_1, A_2, A_2\}$ be a collection of subsets of X. Then $\cap T = (-1, 1]$ and $\cup T = \mathbb{R}$. If we look at the collection of complements $T^c = \{A_1^c, A_2^c, A_3^c\} = \{(1, \infty), (-\infty, -1] \cup (2, \infty), (-\infty, -1)\}$, we have $\cap T^c = \emptyset$ and $\cup T^c = (-\infty, -1] \cup (1, \infty)$. Note that $\cap T^c = (\cup T)^c$ and $\cup T^c = (\cap T)^c$.

Another way to build new sets from old ones is to form ordered pairs or more generally for a given positive integer $n > 0$, ordered $n - tuples$.

Definition 1.21. Let A, B be sets. Then the (Cartesian) product of the sets $A \times B := \{(x, y) \,|\, x \in A, y \in B\}$ and more generally, if $A_1, A_2, \ldots, A_n$ are sets, the $n-$fold (Cartesian) product $A_1 \times A_2 \times \cdots \times A_n := \{(x_1, x_2, \ldots, x_n) \,|\, x_1 \in A_1, x_2 \in A_2, \ldots x_n \in A_n\}$.

Example 1.22.

Let $A = \{x, 1\}, B = \{s, \clubsuit\}$. Then $A \times B = \{(x, s), (x, \clubsuit), (1, s), (1, \clubsuit)\}$. Furthermore:

$$A \times B \times A = \{(x, s, x), (x, s, 1), (x, \clubsuit, x), (x, \clubsuit, 1),$$
$$(1, s, x), (1, s, 1), (1, \clubsuit, x), (1, \clubsuit, 1)\},$$

while

$$(A \times B) \times A = \{((x, s), x), ((x, s), 1), ((x, \clubsuit), x), ((x, \clubsuit), 1),$$
$$((1, s), x), ((1, s), 1), ((1, \clubsuit), x), ((1, \clubsuit), 1)\}.$$

Since the ordered pair $((1, s), x)$, which is an element of $(A \times B) \times A$, is not an element of $A \times B \times A$, we see that $(A \times B) \times A$ is not the same set as $A \times B \times A$. We encourage the reader to write out the sets $B \times A$ and $A \times (B \times A)$ and compare them to $A \times B$ and $(A \times B) \times A$ respectively to see that in general these are all different sets.

Example 1.23. If we let $A_i = A$ for every single $1 \leq i \leq n$ where $n \in \mathbb{N}^+$ we simple write A^n for the product $A_1 \times \cdots \times A_n$. For instance we write $\mathbb{R}^2$ for the set of ordered pairs of real numbers, $\mathbb{R}^3$ for the set of triples (x, y, z) of real number, etc...

2 Functions

There are many different ways one set may be related to another. One of the most fundamental relationships between two sets is that of a *function*. We are a bit more rigorous with our definition of function than one might have seen in a calculus course.

Definition 2.1. A function $f : A \to B$ consists of three pieces of information: a set A, the domain of the function, a set B, the codomain of the function, and an assignment of a unique element of B to each of the elements of the set A. We define $(\exists! x \in S)P(x)$ to mean $(\exists x \in S)P(x) \wedge ((y \in S \wedge P(y)) \implies x = y)$ where P is a predicate statement. We can write the definition of a function rule using quantifiers: $(\forall x \in A)\,(\exists! y \in B)\,(y = f(x))$. Sometimes we write $x \overset{f}{\mapsto} y$ and say f maps the element $x \in A$ to the element $y \in B$. We often write $dom(f) = A, codom(f) = B$. We define

$$Y^X := \{f : X \to Y\}$$

to be the set of all functions with domain X and codomain Y.

Example 2.2. Let $B = \{y\}$ be a set with a single element and suppose that A is any set. There is exactly one function from A to B since we would necessarily have $f(x) = y$ given $x \in A$.

Example 2.3. Writing "$f(x) = x^2$" isn't enough to specify a function. Depending on the domain and codomain we get can many different functions. Two functions are equal if they have the same domain, the same domain, and the exact same rule. So for instance $h : \mathbb{R} \to [0, \infty)$ given by $h(x) = x^2$ is a different function than $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = x^2$.

Note that given any function $f : A \to B$ we can define a subset of $A \times B$ given by $\{(x, f(x)) \,|\, x \in A\} \subseteq A \times B$ which we call the *graph* of f, and we write $graph(f)$ for this subset of $A \times B$. Often times we blur the distinction between a function and its graph, but in general if we only knew the graph of the function we would not know its codomain. For instance if we were told that a function had graph $\{(x, x^2) \,|, x \in \mathbb{R}\}$, we'd know the domain of the function was $\mathbb{R}$ and what the rule defining the function is, but it is not clear what the codomain of the function is with just the graph. Both functions h, g in the previous example have this set as their graphs.

Definition 2.4. Given any set X we can define a function that has domain and codomain X called the identity function on X and written $id_X :, X \to X$ with rule $id_X(a) = a$ for every $a \in X$. Note that if $X \neq Y$ then $id_X \neq id_Y$ even though they have the 'same' rule.

We can think of an infinite sequence $(a_k)_{k=0}^{\infty}$ where each $a_k \in X$ for some set X as simply a function with domain $\mathbb{N}$ and codomain X by letting $a : \mathbb{N} \to X$ be given by $a(k) = a_k$ for each $k \in \mathbb{N}$. A finite sequence of elements of X can be defined to be a function with domain $\{0, 1, 2 \ldots, n\}$ and codomain $\mathbb{N}$, namely $f : \{0, 1, 2 \ldots, n\} \to X$ gives us the sequence $(f(0), f(1), f(2), \ldots, f(n))$ of length $n + 1$.

Definition 2.5. A *binary operation* p on a set X is simply a function $p : X \times X \to X$. Likewise an n-ary operation q on a set X where $n \in \mathbb{N}^+$ is a function $q : X^n \to X$. We extend our previous definition of product to include the product of 0 copies of X and make $X^0 := \{()\}$ the set which contains the empty tuple. Then we can define a 0-ary operation q on a set X as a function $q : X^0 \to X$; since the domain of such an operation consists of a single element, we can identify q with the element $q(()) \in X$ and think of 0-ary operations as simply elements of X.

Example 2.6. $+ : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ given by $+(a, b) = a + b$ is a binary operation; so is subtraction and multiplication. Note that the function $+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by $+(a, b) = a + b$ is a binary operation on $\mathbb{N}$ but that subtraction is not; $1 - 3 = -2 \notin \mathbb{N}$. We'll address this later in this section when we discuss *partial functions*.

Example 2.7. There are lots of natural binary operations, but where do higher order operations show up? Let $T : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$T((x_1, y_1), (x_2, y_2), (x_3, y_3)) = T(A, B, C) = A - B + C = (x_1 - x_2 + x_3, y_1 - y_2 + y_3).$$

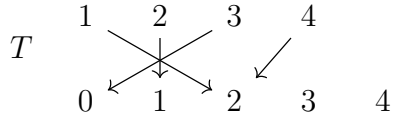
If $T(A, B, C) = D$ then $A - B + C = D$ or $A - B = D - C$ tells us that the lengths of the vectors $A - B$ and $C - D$ are the same, and we encourage the reader to draw a picture of the points A, B, C, D to verify that they form a parallelogram. If A, B, C are not collinear the parallelogram $A, B, C, T(A, B, C)$ has non-zero area.

We can build new functions out of previously defined functions in several ways, the most important being that of composition of functions.

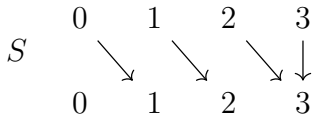
Definition 2.8. Let $f : A \to B, g : C \to D$ where $B \subseteq C$, in other words we have one function whose codomain is a subset of the domain of another function. We form $g \circ f : A \to D$ by the rule $g \circ f(a) = g(f(a))$ for each $a \in A$. We read this as f followed by g. Since $im(f) \subseteq B$, we extend the above definition of composition to include any functions $f : A \to B, g : C \to D$ where $im(f) \subseteq C$.

Example 2.9. Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = e^x$ and $g : [0, \infty) \to [0, \infty)$ given by $g(x) = \sqrt{x}$. Then we can do f first followed by g since $im(f) = (0, \infty) \subseteq [0, \infty) = dom(g)$ and get $g \circ f : \mathbb{R} \to [0, \infty)$ by $g \circ f(x) = g(f(x)) = g(e^x) = e^{\frac{x}{2}}$. Note we can also compose these functions in the opposite order since $im(g) = [0, \infty) \subseteq \mathbb{R} = dom(f)$ to get $f \circ g : [0, \infty) \to \mathbb{R}$ by $f \circ g(x) = f(g(x)) = f(\sqrt{x}) = e^{\sqrt{x}}$. Not only are these rules different, but the domains and codomains of these functions are not the same: $f \circ g \neq g \circ f$ in general.

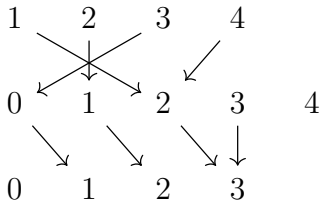
Let $T : \{1, 2, 3, 4\} \to \{0, 1, 2, 3, 4\}$ be given by the arrow diagram:



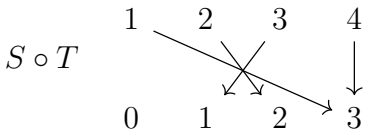
and let $S : \{0, 1, 2, 3\} \to \{0, 1, 2, 3\}$ be given by the arrow diagram:



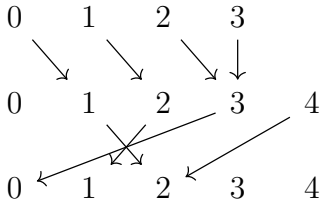
Since $im(T) = \{0, 1, 2\} \subseteq dom(S)$ we can form $S \circ T$, T followed by S, simply by stacking the arrow diagrams and then following the arrows from top to bottom:



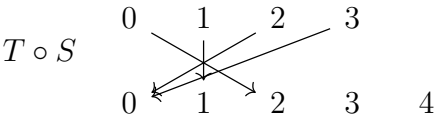
so $S \circ T$ has arrow diagram



Note also that $im(S) = \{1, 2, 3\} \subseteq dom(T)$ and so we can form the composition $T \circ S$, S follows by T by the same process of stacking the arrow diagrams and recording the result. First we stack:



and then recording the arrows from top to bottom we get



Theorem 2.10. *Function composition is associative, namely given $f : A \to B, g : C \to D, h : E \to F$ where $im(f) \subseteq C, im(g) \subseteq E$ then the functions $h \circ (g \circ f)$ and $(h \circ g) \circ f$ from A to F are the same function.*

Proof. The two functions have the same domain and codomain so we check the rule: let $x \in A$. Then $h \circ (g \circ f)(x) = h(g \circ f(x)) = h(g(f(x)) = (h \circ g)(f(x)) = (h \circ g) \circ f(x)$, same rule. □

Example 2.11. Let $\alpha : \mathbb{R} \to \mathbb{R}$ be $\alpha(x) = 2x - 1$. If we let $\beta : \mathbb{R} \to \mathbb{R}$ be given by $\beta(x) = \frac{x+1}{2}$ then notice $\beta \circ \alpha(x) = \beta(\alpha(x)) = \beta(2x - 1) = \frac{(2x-1)+1}{2} = x = id_{\mathbb{R}}(x)$ and $\alpha \circ \beta(x) = \alpha(\beta(x)) = \alpha(\frac{x+1}{2}) = 2(\frac{x+1}{2}) - 1 = x = id_{\mathbb{R}}(x)$ and so we see that $\beta \circ \alpha = id_{\mathbb{R}} = \alpha \circ \beta$. We say that α has a two-sided inverse, a function that when we compose on the left or the right we get an identity function. Note that this depends a lot on the sets involved and not just the

rule. If $\alpha' : \mathbb{Z} \to \mathbb{Z}$ is given by $\alpha'(x) = 2x - 1$ then just writing down an analog of β as the rule $\frac{x+1}{2}$ would not be well defined. But if we let

$$\beta(x) = \begin{cases} \frac{x+1}{2} & \exists y \in \mathbb{Z}, x = 2y - 1 \\ 0 & else \end{cases}$$

Then $\beta(\alpha(x)) = \beta(2x - 1) = \frac{(2x-1)+1}{2} = x = id_{\mathbb{Z}}(x)$. But one readily sees that if we compose these functions in the other order we do not get an identity function, for instance $\alpha(\beta(4)) = \alpha(0) = -1 \neq id_{\mathbb{Z}}(4)$. We say that β is a left inverse for α but not a right inverse. We will make these terms precise in a moment, but first we introduce some nice properties of functions which we will connect to the ideas of 'inverses'.

Definition 2.12. We say that $f : A \to B$ is injective or 1-1 (these are synonyms) if given $x_1 \neq x_2$ elements of A it follows that $f(x_1) \neq f(x_2)$. Equivalently if $f(x_1) = f(x_2)$ then $x_1 = x_2$ (why are these statements equivalent?).

Looking back at $\alpha : \mathbb{R} \to \mathbb{R}$ given by $\alpha(x) = 2x - 1$ we see that if $\alpha(x_1) = \alpha(x_2)$ then $2x_1 - 1 = 2x_2 - 1$ certainly implies $x_1 = x_2$ and so α is 1-1. Likewise α' above is 1-1 by a similar argument.

Definition 2.13. We say that $f : A \to B$ is surjective or onto (these are synonyms) if $im(f) = B$, in other words $\forall y \in B \ \exists x \in A$ so that $f(x) = y$.

Theorem 2.14. *Let $f : A \to B$ and $g : B \to C$. If f and g are $1-1$ then so is $g \circ f$. If f and g are onto then so is $g \circ f$.*

Proof. Suppose that f, g are $1 - 1$ and that we have $g \circ f(x_1) = g \circ f(x_2)$ for some $x_1, x_2 \in A$. Then $g(f(x_1)) = g(f(x_2))$ implies that $f(x_1) = f(x_2)$ since g is $1 - 1$. But then $x_1 = x_2$ since f is $1 - 1$. So we have shown $g \circ f(x_1) = g \circ f(x_2) \implies x_1 = x_2$ and so $g \circ f$ is $1 - 1$.

Now suppose that f, g are both onto and let $z \in C$. Since g is onto there exists $y \in B$ so that $g(y) = z$. Since f is onto there exists $x \in A$ so that $f(x) = y$. But then $g \circ f(x) = g(f(x)) = g(y) = z$ and so $g \circ f$ is onto. □

Looking back at $\alpha : \mathbb{R} \to \mathbb{R}$ given by $\alpha(x) = 2x - 1$ we see that if $y \in \mathbb{R} = codom(\alpha)$ then $\exists x \in \mathbb{R} = dom(\alpha)$ so that $\alpha(x) = y$, namely $x = \frac{y+1}{2}$ works $(\alpha(\frac{y+1}{2}) = y)$ and so α is onto. On the other hand α' is not onto: there is no $x \in \mathbb{Z}$ so that $\alpha'(x) = 0$ since this amounts to the equation $2x - 1 = 0$ having a solution that is an integer.

We saw that α had a two-sided inverse β while α' only had an inverse on one side. We formalize these concepts and then connect them to the properties of being injective/surjective.

Definition 2.15. Let $f : A \to B$. We say that $L : B \to A$ is a *left inverse* of f if $L \circ f = id_A$. We say that $R : B \to A$ is a *right inverse* of f if $f \circ R = id_B$.

Theorem 2.16. *Suppose a function $f :\to B$ has both a left inverse say $L : B \to A$ and a right inverse $R : B \to A$ then $R = L$.*

Proof. e have $R = (id_A) \circ R = (L \circ f) \circ R = L \circ (f \circ R) = L \circ id_B = L$ by the associativity of function composition. □

Theorem 2.17 (Left Inverse Theorem)**.** *Let $f : A \to B$ be a function where $A \neq \emptyset$. Then f is injective iff f has a left inverse.*

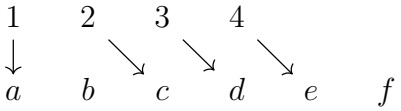
Proof. Assume f is 1-1 and take $a_0 \in A$. Define $g : B \to A$ by

$$g(b) = \begin{cases} a & \text{if } b \in im(f), \quad b = f(a) \\ a_0 & \text{if } x \notin im(f) \end{cases}$$

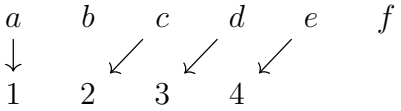
Then we first note that g is well-defined since for each $b \in im(f)$ there is a unique $a \in A$ so that $f(a) = b$ since f is 1-1. One easily checks that $g \circ f = id_A$ and so g is a left inverse.

Now let's assume f has a left inverse, call it $h : B \to A$ so that $h \circ f = id_A$ and assume that $f(x_1) = f(x_2)$ for $x_1, x_2 \in A$. Then applying h to the latter equation we get $x_1 = h(f(x_1)) = h(f(x_2)) = x_2$, showing that f is injective. □

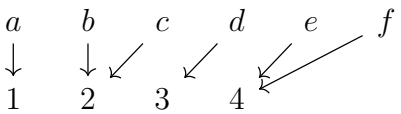
Example 2.18. Let's look at an example of finding a left inverse. Let $P : \{1, 2, 3, 4\} \to \{a, b, c, d, e, f\}$ be given by the arrow diagram



It's clear that P is $1 - 1$; no two arrows point at the same target. In building a left inverse we want to reverse any existing arrows to get:



The latter is not a function, there are no arrows leaving b and f. In order for us to have a function $Q : \{a, b, c, d, e, f\} \to \{1, 2, 3, 4\}$ we need to simply define $Q(b)$ and $Q(f)$; any choices will work, for instance one possible Q is



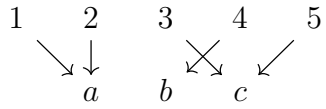
One easily checks that for all $x \in \{1, 2, 3, 4\}$ we have $Q(P(x)) = x$ so that $Q \circ P = id_{\{1,2,3,4\}}$. On the other hand, if we look at $P(Q(b)) = P(2) = c \neq b$ we see that it's false that $P \circ Q = id_{\{a,b,c,d,e,f\}}$. Note that the way we built Q was to take any existing arrows from P and reverse them, and then we added additional arrows in any way we please to ensure that Q is a function.

Theorem 2.19 (Right Inverse Theorem)**.** *Let $f : A \to B$ be a function. Then f is surjective iff f has a right inverse.*

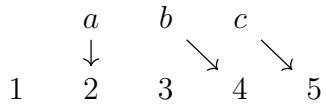
Proof. Assume that f is surjective. We build a right inverse p as follows: for each $b \in B$ choose an element $a \in A$ so that $f(a) = b$ (there might be many but there is at least one since f is onto). Define $p(b) = a$ and note that for each $b \in B, f \circ p(b) = f(p(b)) = f(a) = b = id_B(b)$, showing that p is a right inverse of f.

Let's assume that f has a right inverse q. Let $b \in B$ and look at $a := q(b)$. We see that $f(a) = f(q(b)) = b$ and so f is onto as desired. □

Example 2.20. Let's look at an example of finding a right inverse. Let $T : \{1, 2, 3, 4, 5\} \to \{a, b, c\}$ be given by the arrow diagram



T is clearly onto. We'll build a right inverse $S : \{a, b, c\} \to \{1, 2, 3, 4, 5\}$ by going through the set $\{a, b, c\}$ and for each element picking one of the T arrows and reversing it. One such S is given by



One checks that for each $x \in \{a, b, c\}$ we have $T(S(x)) = x$ so that $T \circ S = id_{\{a,b,c\}}$. Since $S(T(3)) = S(c) = 5 \neq 3$ we do not have $S \circ T = id_{\{1,2,3,4,5\}}$ in this case.

In general finding inverses of any type is tricky business. For instance looking at $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3 + 100\sin(x)$, it's not hard to see that since f is continuous and $\lim_{x\to\pm\infty} f(x) = \pm\infty$, the intermediate value theorem tells us $im(f) = \mathbb{R}$ and f is onto and hence has a right inverse. Such a right inverse $g : \mathbb{R} \to \mathbb{R}$ would have to satisfy $f(g(x)) = x$, and so finding an explicit formula for g would involve finding a solution $y = g(x)$ to the equation $y^3 + 100\sin(y) = x$ which is a nightmare. But nonetheless we know at least one such solution exists by the Right Inverse Theorem. f is not $1-1$ as $f(x) = 0$ has many solutions as one can verify by typing 'solve $x^3 + 100\sin(x) = 0$' into www.wolframalpha.com . By the Left Inverse Theorem no such left inverse of f exists.

One might have noticed that the left inverse q of the $1-1$ function p is onto, and the right inverse S of the onto function T is $1-1$. This is always the case and is fairly immediate from the Left and Right Inverse Theorems which we leave as an exercise to the reader:

Corollary 2.20.1. *Let $f : A \to B$ have left inverse $g : B \to A$. Then g is onto. Suppose $h : B \to A$ is a right inverse of f. Then h is $1-1$*

Definition 2.21. A function $f : A \to B$ that is 1-1 and onto is called a bijection or is said to be a 1-1 correspondence between A and B.

Example 2.22. Let X, Y, Z be sets. We show that $f : (X \times Y) \times Z \to X \times (Y \times Z)$ given by $f((a, b), c) = (a, (b, c))$ is a bijection. Suppose that $t \in X \times (Y \times Z)$. Then $\exists a \in A, b \in B, c \in C$ so that $t = (a, (b, c))$. But then $f((a, b), c) = t$ shows that f is onto. If $f(t_1) = f(t_2)$ for $t_1, t_2 \in (X \times Y) \times Z$ then $t_i = ((a_i, b_i), c_i)$ for $i = 1, 2$ and the condition $f(t_1) = f(t_2)$ is simply that $(a_1, (b_1, c_1)) = (a_2, (b_2, c_2))$. But then $a_1 = a_2$ and $(b_1, c_1) = (b_2, c_2)$ which implies $b_1 = b_2, c_2 = c_2$ and hence $t_1 = t_2$. So f is a bijection.

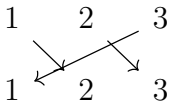
Based on the above theorems we know immediately that any function that is a bijection has a unique inverse, and vice a versa, any function that has both a left and right inverse not only has these one sided inverses equal, but it must be a bijection. If $f : A \to B$ is a bijection we write $f^{-1} : B \to A$ for the inverse of f.

Definition 2.23. A bijection $p : X \to X$ from a set to itself is called a *permutation of* X. We let Σ_X be the set of all permutations of X. $id_X \in \Sigma_X$ and if $f, g \in \Sigma_X$ then $f^{-1} \in \Sigma_X$ and $f \circ g \in \Sigma_X$.

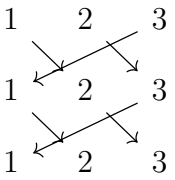
Example 2.24. Let $X = \{1, 2, 3\}$. We first figure out the size of the set Σ_X. For any permutation $p : X \to X$ there are 3 choices for $p(1)$, and since p is $1 - 1$ this leaves 2 choices for $p(2)$, and 1 choice for $p(3)$. Thus there are a total of $6 = 3 \cdot 2 \cdot 1$ permutations of X. It's not hard to see that if X has n elements the set Σ_X has size $n!$. We already know one of these 6 elements, id_X, so let us find the other 6.

Let $\sigma : X \to X$ be the permutation of X given by the arrow diagram:

σ

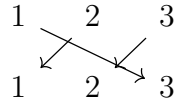


If we compose σ with itself we first get $\sigma \circ \sigma := \sigma^2$:



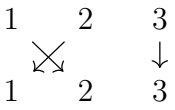
and so

σ^2

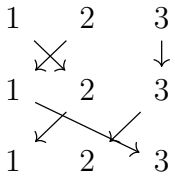


Note that $\sigma^2 = \sigma^{-1}$ by comparing their arrow diagrams. This tells us that $\sigma \circ \sigma^2 := \sigma^3$ is simply id_X. So we have three permutations so far: id_X, σ, σ^2. Let τ be the permutation of X given by the arrow diagram:

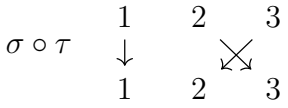
τ



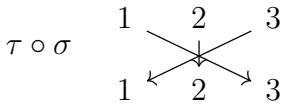
We see immediately that $\tau \circ \tau := \tau^2 = id_X$. Let's look at $\sigma \circ \tau$ by stacking and composing arrow diagrams:



gives us



Lastly if we compose τ and σ in the other order to get $\tau \circ \sigma$ we leave it to the reader to see that one gets



Note that each of $\tau, \sigma \circ \tau, \tau \circ \sigma$ are distinct from each other and distinct from id_X, σ, σ^2, and so altogether we have

$$\Sigma_X = \{id_X, \sigma, \sigma^2, \tau, \tau \circ \sigma, \sigma \circ \tau\}.$$

The set Σ_X with the operation of function composition is an example of a *group*; it is a set with an associative binary operation that has an identity element and an inverse for each element. We'll return to our discussion of permutations when we discuss Fermat's Little Theorem in the section on Number Theory.

We give a few examples of these concepts in the context of linear algebra and recall some basic facts about linear transformations (feel free to skip this if you have no background in linear algebra of course). Let $L = L_A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map given by left multiplication of an $m \times n$ real matrix A on a column vector $v \in \mathbb{R}^n$, namely $L(v) = Av$. L is onto exactly when $im(L) = \mathbb{R}^m$ or $rank(A) = dim(im(L)) = m$. We also see that L is 1-1 iff it has trivial kernel: suppose L is 1-1, and $v \in ker(L)$. Then $0 = L(0) = L(v)$ implies that $v = 0$ and if $ker(L) = \{0\}$ and $L(v_1) = L(v_2)$ then $L(v_1 - v_2) = 0$ and so $v_1 - v_2 \in ker(L)$ and so $v_1 = v_2$. By the Rank-Nullity Theorem we can then say L is 1-1 iff $rank(A) = n$. Now we are ready for some nice examples.

Example 2.25. Let $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$, which clearly has $rank(A) = 2$ and so the associated linear map $L_A : \mathbb{R}^3 \to \mathbb{R}^2$ is onto but not 1-1. Thus L_A has a right inverse but no left inverse. Let's try to find a right inverse, say $L_B : \mathbb{R}^2 \to \mathbb{R}^3$ given by some 3×2 matrix B. If $L_A \circ L_B = L_{I_2} = id_{\mathbb{R}^2}$ we'd have $v = L_A(L_B(v)) = L_A(Bv) = ABv$. We are looking for a matrix B so that

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

and there are lots of possibilities, one being to set $c = e = 0, a = 1, b = d = 0, f = 1$ (call this B). We can easily confirm that no left inverse L_C exists, since if there was such a linear map we would have $L_C \circ L_A(v) = (CA)v$ for each $v \in \mathbb{R}^3$ and $CA = I_3$. But $rank(I_3) = 3$ while $rank(C) \leq 2$ forces $rank(CA) \leq 2$. One sees what happens if we multiply BA:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

which is a rank two matrix not even close to the identity matrix.

Example 2.26. Let $M = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$ which clearly has $rank(M) = 2$ and so the associated linear map $L_M : \mathbb{R}^2 \to \mathbb{R}^4$ is 1-1 but not onto. So L_M has a left inverse but no right inverse, let's find try to find a left inverse say $L_N : \mathbb{R}^4 \to \mathbb{R}^2$ where N is 2×4. We would need $NM = I_2$ or

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} a+c+d & -a+b+d \\ c+g+h & -e+f+h \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Again there are infinitely many possibilities for N, one being $a = 1, b = -1, c = d = 0, h = 1, g = -1, e = f = 0$ (call this matrix N). Order really does matter here, MN is

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

which is a far cry from I_4.

It's fairly easy to find left and right inverses of linear transformations, but we remind the reader that finding inverses in general can be pretty horrible.

Example 2.27. Let $p : \mathbb{R} \to \mathbb{R}$ be given by $p(x) = x^{11} + 5x - \arctan(x)$. We know that $p'(x) = 11x^{10} + 5 - \frac{1}{1+x^2}$ and that $\frac{1}{1+x^2} \leq 1$ so that $p'(x) \geq 4 > 0$ and so p is strictly increasing. By a homework exercise this means that p is 1-1. We know $\lim_{x \to \pm\infty} p(x) = \pm\infty$ and p is continuous and so by the Intermediate Value Theorem we see that p is onto. So p is a bijection and so $p^{-1} : \mathbb{R} \to \mathbb{R}$ exists. Try finding a formula for this and see what trouble you run into.

If we are working with a function $f : A \to B$ we can form a new function $f^* : A \to im(f)$ given by the same rule $f^*(x) = f(x)$ for any $x \in A$. The function f^* here is always onto by its very construction. We can also take a function and restrict it to a subset of the domain as follows:

Definition 2.28. Let $f : A \to B$ and let $X \subseteq A$. Define $f|_X : X \to B$ to be the function with rule $f|_X(t) = f(t)$ for all $t \in X$. We call $f|_X$ the restriction of f to the subset X.

Example 2.29. Let $g : \mathbb{Z} \to \mathbb{Z}$ be given by $g(n) = n^2$. It's not hard to see that g is neither 1-1 nor onto. The function $h = g|_{\mathbb{N}}$ is 1-1 since if $h(n_1) = h(n_2)$ then $n_1^2 = n_2^2$ and so $0 = n_1^2 - n_2^2 = (n_1 - n_2)(n_1 + n_2)$ and so either $n_1 = n_2$ or $n_1 = -n_2$. But since $n_1, n_2 \in \mathbb{N}$ the only way $n_1 = -n_2$ could happen is if $n_1 = 0 = n_2$. Either way $n_1 = n_2$ and h is 1-1.

Example 2.30. In the previous example we saw that g was not 1-1 onto. We can put the elements $y \in \mathbb{Z}$ the codomain of g into one of three categories. There are elements like $y = 4$ that have more than one element x in the domain so that $g(x) = y$, for example $g(-2) = 4 = g(2)$. These are the perfect squares $y \in \{1, 4, 9, 16, \dots\}$. There are elements like $y = 0$ that have exactly one element x in the domain so that $g(x) = y$, in this case only $x = 0$ gives $g(x) = 0$. And there are elements y that aren't in the image of g at all, integers like -5 or 7.

Example 2.31. The function $s : \mathbb{R} \to [-1, 1]$ given by $s(x) = \sin(x)$ is certainly onto but far from being $1 - 1$ since $s(k\pi) = 0$ for any integer k. So s does not have an inverse function. However if we let $r = s|_{[-\frac{\pi}{2}, \frac{\pi}{2}]}$ be the function s restricted to the closed interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$, r *is* $1-1$ since r is strictly increasing on this interval. $r([-\frac{\pi}{2}, \frac{\pi}{2}]) = [-1, 1]$ and so $r : [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$ is a bijection and has an inverse $r^{-1} : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$. We call the function r^{-1} the 'inverse sine' or 'arcsine' function. We encourage the reader to revisit their calculus textbook for other examples of inverse trigonometric functions and see that they are constructed analogously.

We use the previous example to motivate our next definition. The function $s : \mathbb{R} \to [-1, 1]$ satisfies $s(k\pi) = 0$ for any $k \in \mathbb{Z}$. When solving $s(x) = 0$ for x we don't get a single value but in fact an entire set of $x's$ that work. Even though the inverse of s doesn't exists since s is not $1 - 1$, we'd like to still be able to talk about the set of all x so that $s(x) = 0$ and we will denote this by $s^{-1}(\{0\})$ or more simply $s^{-1}(0) = \{\cdots - 2\pi, -\pi, 0, \pi, 2\pi, \dots\}$. More precisely we make the following definition:

Definition 2.32. Let $f : A \to B$ and let $Y \subseteq B$. We define the preimage of Y via f, sometimes called the inverse image, to be the following subset of A:

$$f^{-1}(Y) := \{x \in A \,|\, f(x) \in Y\} \subseteq A.$$

Note that $x \in f^{-1}(Y)$ simply means $f(x) \in Y$, it does not mean that f is a bijection and f^{-1} exists. It is a subset of the domain. Two immediate observations are that $f^{-1}(B) = A$ and $f^{-1}(\emptyset) = \emptyset$.

Example 2.33. Revisiting the function g from example 2.12, $g^{-1}(\{4, 5, 6, 7, 8, 9\}) = \{-3, 2, 2, 3\}$ and $g^{-1}(\{\cdots - 2, -1\}) = \emptyset$

Lemma 2.34. *If $f : A \to B$ and $Y_1 \subseteq Y_2 \subseteq B$, then $f^{-1}(Y_1) \subseteq f^{-1}(Y_2)$.*

Proof. Let $x \in f^{-1}(Y_1)$. Then $f(x) \in Y_1$ and so $f(x) \in Y_2$ and hence $x \in f^{-1}(Y_2)$. □

Theorem 2.35. *If $f : A \to B$ and $Y_1, Y_2 \subseteq B$ then*

$$f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$$
$$f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2)$$

Proof. By the previous lemma we know that since $B_1 \cap B_2 \subseteq B_1, B_2 \subseteq B_1 \cup B_2$ we have $f^{-1}(B_1 \cap B_2) \subseteq f^{-1}(B_1), f^{-1}(B_2) \subseteq f^{-1}(B_1 \cup B_2)$ and so $f^{-1}(B_1 \cap B_2) \subseteq f^{-1}(B_1) \cap f^{-1}(B_2)$ and $f^{-1}(B_1) \cup f^{-1}(B_2) \subseteq f^{-1}(B_1 \cup B_2)$. If $x \in f^{-1}(B_1) \cap f^{-1}(B_2)$ then $f(x) \in B_1$ and $f(x) \in B_2$ and so $f(x) \in B_1 \cap B_2$ and so $x \in f^{-1}(B_1 \cap B_2)$ giving us $f^{-1}(B_1) \cap f^{-1}(B_2) = f^{-1}(B_1 \cap B_2)$. Similarly if $x \in f^{-1}(B_1 \cup B_2)$ then $f(x) \in B_1 \cup B_2$ and so either $f(x) \in B_1$ or $f(x) \in B_2$. By definition of preimage this implies $x \in f^{-1}(B_1)$ or $x \in f^{-1}(B_2)$ and so $x \in f^{-1}(B_1) \cup f^{-1}(B_2)$ and we have $f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$. □

Definition 2.36. Let $h : A \to B$ and $X \subseteq A$. We define a subset of B called the image of X via h as follows:

$$h(X) = \{y \in B \mid \exists x \in X, h(x) = y\} \subseteq A.$$

An immediate consequence of the definition is that $h(A)$ is the image or range of the function h.

Example 2.37. Let $g : [0, \infty) \to \mathbb{R}$ be $g(x) = \sqrt{x}$. Then

$$g([1, 4]) = [1, 2],$$
$$g(\{45\}) = \{\sqrt{45}\},$$
$$g(\emptyset) = \emptyset,$$
$$g([0, \infty)) = [0, \infty) = im(g).$$

Lemma 2.38. *Let $h : A \to B$ and $X_1 \subseteq X_2 \subseteq A$. Then $h(X_1) \subseteq h(X_2)$.*

Proof. Let $y \in h(X_1)$. Then $\exists x \in X_1$ so that $h(x) = y$. Since $X_1 \subseteq X_2$ we have that $x \in X_2$ and so $y \in h(X_2)$. □

Theorem 2.39. *Let $h : A \to B$ and let $A_1, A_2 \subseteq A$. Then we have:*

$$h(A_1 \cap A_2) \subseteq h(A_1) \cap h(A_2)$$

$$h(A_1 \cup A_2) = h(A_1) \cup h(A_2),$$

Proof. Let $y \in h(A_1 \cap A_2)$. Then $\exists x \in A_1 \cap A_2$ so that $h(x) = y$ by definition of image. Then by definition of intersection we have $x \in A_1 \wedge x \in A_2$ and so $y \in h(A_1) \wedge y \in h(A_2)$ by definition of image. This means $y \in h(A_1) \cap h(A_2)$ by definition of intersection and so $h(A_1 \cap A_2) \subseteq h(A_1) \cap h(A_2)$.

Let $y \in h(A_1 \cup A_2)$. Then $\exists x \in A_1 \cup A_2$ so that $h(x) = y$ by definition of image. Then by definition of union we have $x \in A_1 \vee x \in A_2$ and so $y \in h(A_1) \vee y \in h(A_2)$ by definition of image. This means $y \in h(A_1) \cup h(A_2)$ by definition of intersection and so $h(A_1 \cup A_2) \subseteq h(A_1) \cup h(A_2)$. If we take $z \in h(A_1) \cup h(A_2)$ then $z \in h(A_1)$ or $z \in h(A_2)$. If $z \in h(A_1)$ then there exists a $a \in A_1$ so that $h(a) = z$ by definition of image. So $a \in A_1 \cup A_2$ by definition of union and so $z = h(a) \in h(A_1 \cup A_2)$ by definition of image. The case where $z \in h(A_2)$ is completely analogous. So we have $h(A_1) \cup h(A_2) \subseteq h(A_1 \cup A_2)$ and so $h(A_1 \cup A_2) = h(A_1) \cup h(A_2)$. □

Example 2.40. Let $f : \{1, 2, 3, 4\} \to \{a, b, c, d, e\}$ be given by the following diagram:

$$\begin{array}{ccccccc} 1 & & 2 & & 3 & & 4 \\ \downarrow & & & \searrow & & \searrow & \downarrow \\ a & & b & & c & & d \qquad e \end{array}$$

Let $A_1 = \{1, 3\}$ and $A_2 = \{1, 4\}$. Then $f(A_1) = \{a, d\}, f(A_2) = \{a, d\}$, $f(A_1 \cap A_2) = f(\{1\}) = \{a\}$ and $f(A_1 \cup A_2) = f(\{1, 3, 4\}) = \{a, d\}$ and we check $f(A_1) \cap f(A_2) = \{a, d\} = f(A_1) \cup f(A_2)$. So here we see an example where $f(A_1 \cap A_2) \subsetneq f(A_1) \cap f(A_2)$.

In summary a function $f : A \to B$ consists of three pieces of information: a set A the domain, a set B the codomain, and rule that assigns to each element of A a unique element of the set B. As mentioned earlier we can specify the rule by listing the ordered pairs in the graph of the function. There are many useful relationships in mathematics that require a much looser framework than that of a function and we turn our attention to the following definition:

Definition 2.41. Let X and Y be sets and let $R \subseteq X \times Y$. Then we say that R is a relation with domain X and codomain Y or say R is a relation from X to Y. Technically the relation is the triple (R, X, Y) but we often just call the relation R when the domain and codomain are clear or implied. We will often write aRb to mean $(a, b) \in R$. If $X = Y$ then we say that R is a relation *on* X.

We always have $\emptyset \subseteq X \times Y$ for any X and Y; we call this the *empty relation* with domain X and codomain Y. Note that even though the empty set is unique there are many empty relations as we vary domain and codomain. In a similar vein given any sets X and Y we can let $R = X \times Y$ and declare every element of X to be related to every element of Y. Given any set X there is a natural relation $=$ with domain X and codomain X given by $(a, b) \in =$ to mean $a = b$; we call this the *equals relation* on the set X.

Any function $f : X \to Y$ is a relation from X to Y if we interpret the function rule as the set of ordered pairs $graph(f)$. A natural generalization of the function relation is that of a partial function from X to Y; we say that $P : X \to Y$ is a partial function from X to Y if P is any relation from X to Y (so $P \subseteq X \times Y$) so that $(x, y_1), (x, y_2) \in P \implies y_1 = y_2$. In other words P is a collection of ordered pairs so that no first coordinate is related to more than one second coordinate. An example of a partial function is $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by $((x, y), z) \in D$ exactly

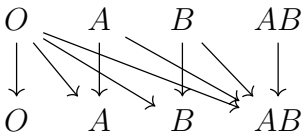
when $z = \frac{x}{y}$ (we can also write $D(x, y) = z$ here). This isn't a function since elements like $(3, 0)$ in the domain don't correspond to anything in the codomain, but it is a partial function since if $((x, y), z_1), ((x, y), z_2) \in D$ then $z_1 = \frac{x}{y} = z_2$. Another partial function is $S : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by $((a, b), c) \in S$ exactly when $c = a - b$ (we can also write $S(a, b) = a - b$. This isn't a function since elements like $(2, 4)$ in the domain don't correspond to anything in the codomain.

In these notes we will be focusing on relations R on a set X that satisfy some well-behaved properties. For a look at relations in general feel free to click HERE

3 Relations

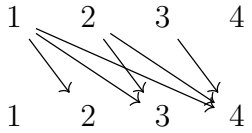
Let $R \subseteq X \times X$ be a relation on a set X. We write $(a, b) \in R$ or aRb and say "a is related to b via R" where $a, b \in X$.

Example 3.1. Let $T = \{O, A, B, AB\}$ and let BT be the relation on T given by the arrow diagram:



This is the 'blood type' relation on the set T of blood types, and the arrows reflect the 'giver-receiver' dynamic. Note that the arrows are absolutely essential, first and foremost because we make a mathematical distinction between the domain and the codomain, and secondly because giving the wrong blood type to someone could be extremely dangerous.

Example 3.2. Let $W = \{1, 2, 3, 4\}$ and let L be the relation on W given by the arrow diagram:



One can think of the relation L the following way: we have xLy exactly when $x < y$.

Definition 3.3. We say that R is reflexive if $\forall a \in X, (a, a) \in R$.

The relation BT above is clearly reflexive, since all four pairs $(O, O), (A, A), (B, B), (AB, AB)$ are part of the relation. The relation L on the other hand is not reflexive since $(1, 1) \notin L$

Definition 3.4. Let R_1, R_2 be relations on X. We define the relation:

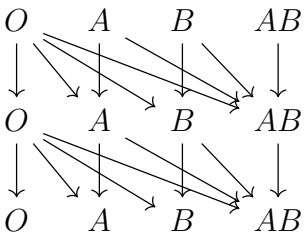
$$R_2 \circ R_1 = \{(a, c) \mid \exists b \in X, (a, b) \in R_1, (b, c) \in R_2\}$$

to be the composition of R_1 followed by R_2, and we also define

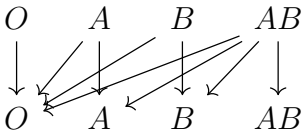
$$R_1^{-1} = \{(a, b) \mid (b, a) \in R_1\}$$

to be the inverse of the relation R_1.

We compute $BT \circ BT$ by stacking the relation BT on top of itself and reading off the composition in the form of an arrow diagram:



Evidently we have $BT \circ BT = BT$. We compute BT^{-1} by simply reversing all the arrows:



Definition 3.5. We say that R is symmetric if $R = R^{-1}$; this is equivalent to saying $\forall a, b \in X, (a, b) \in R \iff (b, a) \in R$.

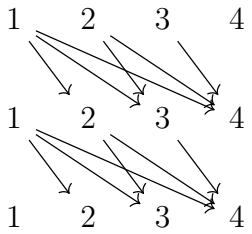
Since BT^{-1} and BT are different, BT is not symmetric.

Definition 3.6. We say that R is antisymmetric if $\forall a, b \in X$, if $(a, b), (b, a) \in R$ then $a = b$. In other words $R \cap R^{-1} \subseteq \{(a, a) \,|\, a \in X\}$.

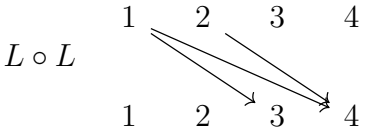
Since $BT \cap BT^{-1} = \{(O, O), (A, A), (B, B), (AB, AB)\}$ we have that the relation BT is antisymmetric.

Definition 3.7. We say R is transitive if $R \circ R \subseteq R$; this is equivalent to saying $\forall a, b, c \in X$, if $(a, b), (b, c) \in R$ then $(a, c) \in R$.

The relation BT is transitive since $BT \circ BT = BT$. Looking at the relation L we compute $L \circ L$ by stacking L on top of itself and composing arrows:



to get



We see that $L \circ L \subseteq L$ and so L is transitive. Note that we lost some arrows in the composition here, but since we didn't add any new ones this relation is transitive.

Example 3.8. Let $X = \{2, 3, 4 \dots\}$ and define a relation Q on x by mQn iff m and n have only the factors ± 1 in common (we say m, n are coprime or relatively prime). The relation Q is not reflexive since it's false that $2Q2$ for instance. The relation Q is certainly symmetric, but it is not antisymmetric since $3Q2$ and $2Q3$ yet $3 \neq 2$. And the relation Q is not transitive since $2Q3$ and $3Q2$ yet $2Q2$ is false. Let Q' be the relation on the same set X given by $mQ'n$ iff m and n share a common factor $p > 1$. Then Q' is reflexive since for any $a \in X$ we have a shares the factor a with itself, and so $aQ'a$ is always true. For the same reason that Q is symmetric, Q' is symmetric. Q' is not antisymmetric since $4Q'6$ and $6Q'4$ yet $4 \neq 6$. Lastly Q' is not transitive since $4Q'6$ and $6Q'3$ yet it's false that $4Q'3$.

We say that R is an *equivalence relation on* X if it is reflexive, symmetric, and transitive. We say R is a *partial order on* X, and call X a 'poset', if it is reflexive, antisymmetric, and transitive.

Example 3.9. If $X = \{1, 2\}$, then the relation

$R_1 = \{(1, 1), (1, 2)\}$ is not reflexive ($(2, 2) \notin R_1$), it is not symmetric since $R_1^{-1} \neq R_1$, but is transitive since $R_1 \circ R_1 = R_1$ and is antisymmetric since $R_1 \cap R_1^{-1} = \{(1, 1)\} \subseteq \{(1, 1), (2, 2)\}$.

$R_2 = \{(1, 2), (2, 1)\}$ is not reflexive since $(1, 1) \notin R_1$, it is symmetric, it is not transitive since $R_2 \circ R_2 = \{(1, 1), (2, 2)\}$, and is not antisymmetric since $R_2 \cap R_2^{-1} = R_2 \not\subseteq \{(1, 1), (2, 2)\}$.

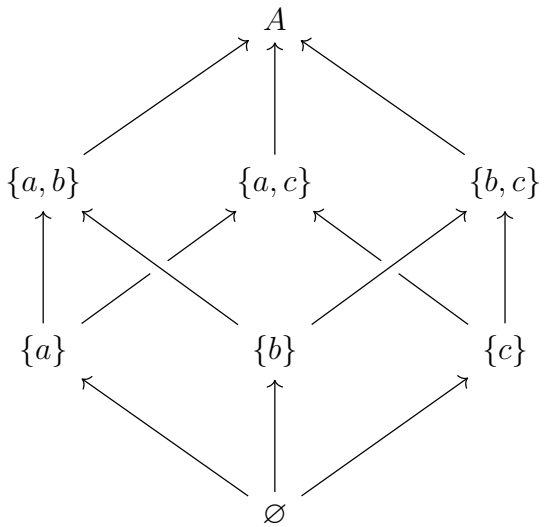
Example 3.10. Let $X = \mathbb{Z}$ and define $a|b$ to mean $\exists j \in \mathbb{Z}$ so that $b = aj$. We read $a|b$ as a *divides* b or b is a multiple of a. Note that $\forall a \in \mathbb{Z}$, $a|0$ and $\pm 1|a$. This defines a relation R on the integers where we say $(a, b) \in R \iff a|b$.

This relation is reflexive since $\forall a \in X, a|a$ and so $(a, a) \in R$. It is not symmetric since $(1, 2) \in R$ but $(2, 1) \notin R$. It is not antisymmetric since $1| - 1$ and $-1|1$ so $(1, -1), (-1, 1) \in R$ but $1 \neq -1$. It is transitive since if $a|b$ and $b|c$, $\exists j, k \in \mathbb{Z}$ so that $b = aj, c = kb$ and so $c = kb = k(aj) = (kj)a$, and so c is a multiple of a and hence $a|c$.

Example 3.11. Let X be any set and let $P(X) = \{A \mid A \subseteq X\}$ be the collection of all subsets of X (this is called the power set of X). Define a relation S on $P(X)$ by ASB iff $A \subseteq B$. This relation is reflexive since $A \subseteq A$ for any set A implies ASA, it is antisymmetric since if ASB and BSA then $A \subseteq B, B \subseteq A$ and hence $A = B$. It is clearly transitive: ASB, BSC implies $A \subseteq B, B \subseteq C$ and so $A \subseteq C$ and so ASC. It is symmetric exactly when $X = \emptyset$; if $X \neq \emptyset$ then $\exists a \in X$ and so $\{a\} \in P(X)$; in this case $\emptyset S\{a\}$ is true but $\{a\}S\emptyset$ is false.

For instance if $X = \{1, 2\}, P(X) = \{\emptyset, \{1\}, \{2\}, X\}$ and the above relation is $\{(\emptyset, \emptyset), (\emptyset, \{1\}), (\emptyset, \{2\}), (\emptyset, \{1, 2\}), (\{1\}, \{1\}),$
$(\{1\}, \{1, 2\}), (\{2\}, \{2\}), (\{2\}, \{1, 2\}), (\{1, 2\}, \{1, 2\})\}$.

If we let $A = \{a, b, c\}$ the relation S on $P(A)$ is best seen through the use of a *Hasse diagram* as follows:



Here we put the elements of the set $P(A)$ as the vertices, draw arrows to indicate inclusion, and make the assumptions that every vertex is related to itself (it's reflexive). For a proper definition and more examples of a Hasse diagram check out HERE

Example 3.12. Let $B = \{f \mid f : \mathbb{N} \to \{0,1\}\}$, in other words B is the set of infinite binary sequences. Define a relation R on B as follows:

$$fRg \iff f^{-1}(1) \subseteq g^{-1}(1).$$

Elements of B look like $f = 010011010101000010010000\ldots$. Then we have two sequences related if they have $1's$ in exactly the same positions. Letting $\overline{0} = 000000\ldots$ and $\overline{1} = 111111\ldots$ we see that $0100\overline{0}R011010\overline{0}$ and $011010\overline{0}R111011\overline{1}$ etc..

Clearly the relation is reflexive since checking $f^{-1}(1) \subseteq f^{-1}(1)$ amounts to seeing that any set is a subset of itself, and the relation is also transitive since $f^{-1}(1) \subseteq g^{-1}(1)$ and $g^{-1}(1) \subseteq h^{-1}(1)$ imply that $f^{-1}(1) \subseteq h^{-1}(1)$ and so $fRg \wedge gRh \implies fRh$. The relation is not symmetric however; if we define $z(n) = 0$ for every $n \in \mathbb{N}$ and $a(n) = 1$ for every $n \in \mathbb{N}$ we have $z^{-1}(1) = \emptyset, a^{-1}(1) = \mathbb{N}$ so zRa but it's false that aRz. This relation is antisymmetric since if $f^{-1}(1) \subseteq g^{-1}(1)$ and $g^{-1}(1) \subseteq f^{-1}(1)$ would imply $f^{-1}(1) = g^{-1}(1)$, and since the codomain consists of only two elements this forces $f^{-1}(0) = g^{-1}(0)$ as well, so $f = g$. So R is a partial order on B.

Note in the last example that if we changed the codomain to a different set, say $\{0,1,2\}$ or $\mathbb{R}$ we could still define a relation on these sets in a completely analogous manner and these relations would be reflexive and transitive, but with a larger codomain we would lose antisymmetry.

If $f : X \to X$ is a function with domain and codomain X then as we pointed out earlier $R = graph(f) = \{(a, f(a)) \mid a \in X\}$ is a relation on X. Here R^{-1} makes sense regardless of whether f has a left/right inverse, it just isn't usually the graph of a function. But $R^{-1} = graph(f^{-1})$ if f is indeed a bijection. If R is symmetric then f *must* be a bijection but the converse of this statement is in general false: if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = x^3$, f is a bijection and $R = graph(f) = \{(x, x^3) \mid, x \in \mathbb{R}\}$ while $R^{-1} = \{(x^3, x) \mid x \in \mathbb{R}\}$ is different than R since $(2,8) \in R$ but $(2,8) \notin R^{-1}$.

We recall that we say $\sim$ is an *equivalence relation* on a set X if it is reflexive, symmetric, and transitive. We often just say $(X, \sim)$ is an equivalence relation. For each $a \in X$ we define $[a] := \{b \mid a \sim b\}$ to be the *equivalence class* of a. Note that since $\sim$ is reflexive we always have that $a \in [a]$ and so equivalence classes are always non-empty.

Example 3.13. Let $X_1 = \{a, b, c, d\}$ and let $\sim_1$ be the equivalence relation:

$$\{(a,a), (b,b), (c,c), (d,d), (a,b), (b,a), (c,d), (d,c)\}.$$

Then $[a] = [b] = \{a, b\}$ and $[c] = [d] = \{c, d\}$ are the two distinct equivalence classes. Note that the same equivalence class can often have many names like in this example.

Example 3.14. Let $X_2 = \mathbb{R}$ and define $x \sim_2 y$ iff $x - y \in \mathbb{Q}$. One easily sees this is an equivalence relation. There are infinitely many equivalence classes, a few examples are $[0] = [\frac{1}{5}] = [-\frac{43}{642}] = [r]$ where r is any rational number (all the rational numbers are equivalent). Another class is $[\pi] = [\pi - 3.14] = [\pi - 3.1415926]$ etc...

One might also see here that every element of X is in one of these equivalence classes, and that these classes are either equal or disjoint. This is a general fact about equivalence relations:

Theorem 3.15 (Equivalence Class Theorem)**.** *Let $(X, \sim)$ be an equivalence relation. Then $\forall a, b \in X$ either:*

1. $[a] \cap [b] = \emptyset$ *or*
2. $[a] = [b]$

Proof. Assume $\exists x \in [a] \cap [b]$. Then $a \sim x$ and $b \sim x$ and so by symmetry $x \sim b$ and hence by transitivity $a \sim b$ and $b \sim a$. If $y \in [a]$ then $a \sim y$ and combining this with $b \sim a$ by transitivity we get $b \sim y$ or $y \in [b]$ and hence $[a] \subseteq [b]$. Similarly $[b] \subseteq [a]$ and so $[a] = [b]$. □

We denote the set of equivalence classes of $(X, \sim)$ by $X/\sim$, namely,

$$X/\sim := \{[a] \,|\, a \in X\}.$$

There is a natural surjective function $\pi : X \to (X/\sim)$ given by $\pi(a) = [a]$. Looking back at $(X_1, \sim_1)$ we see that $(X_1/\sim_1) = \{[a], [c]\}$ has two elements and $\pi : X_1 \longrightarrow (X_1/\sim_1)$ is given by $\pi(a) = [a] = \pi(b), \pi(c) = [c] = \pi(d)$. Note that $\pi^{-1}([a]) = \{a, b\} = [a] = [b]$ and $\pi^{-1}([c]) = \{c, d\} = [c] = [d]$ just returns the equivalence classes as pre-images.

The Equivalence Class Theorem says that equivalence relations give rise to a **partition** of the set involved via its equivalence classes. A partition $\mathbf{P}$ of a set X is a subset of the power set $P(X)$, excluding the empty set, so that $\cup\mathbf{P} = X$ and if $A \cap B \neq \emptyset$ are elements of $\mathbf{P}$ then $A = B$. Conversely we have the following:

Theorem 3.16. *Let $\mathbf{P}$ be a partition of a set X and define a relation $\sim$ on X by $a \sim b$ iff $\exists P \in \mathbf{P}$ so that $a, b \in P$. Then $\sim$ is an equivalence relation an X and $X/\sim = \mathbf{P}$.*

Proof. If $a \in X$ then since $\mathbf{P}$ is a partition there is some $P \in \mathbf{P}$ so that $a \in P$. But then $a \sim a$, and we see that $\sim$ is reflexive. Suppose $a, b \in X$ with $a \sim b$. This means there exists some $Q \in \mathbf{P}$ so that $a, b \in Q$. But then $b, a \in Q$ and hence $b \sim a$ as well; $\sim$ is symmetric. If $a, b, c \in X$ with $a \sim b, b \sim c$ then $\exists P_1, P_2 \in \mathbf{P}$ with $a, b \in P_1$ and $b, c \in P_2$. So $b \in P_1 \cap P_2$ implies that $P_1 = P_2$ since $\mathbf{P}$ is a partition, and so $a, b, c \in P_1 = P_2$ implies $a \sim c$. We leave it to the reader to readily check that $X/\sim$ is exactly $\mathbf{P}$. □

Example 3.17. Let $f : A \to B$ be any function. We can build another function based on f that is a bijection in the following way. First define a relation $\sim$ on A by setting $x \sim y$ if $f(x) = f(y)$; it's immediate that this is an equivalence relation on A. Define $\tilde{f} : (A/\sim) \to im(f)$ by $\tilde{f}([x]) = f(x)$. We first make sure this is well defined: if $[x] = [y]$ then $f(x) = f(y)$ and so $\tilde{f}([x]) = f(x) = f(y) = \tilde{f}([y])$ and the definition of $\tilde{f}$ does not depend on the choice of equivalence class representative. We also see that $\tilde{f}$ is a bijection since it's clearly onto, and if $\tilde{f}([x_1]) = \tilde{f}([x_2])$ then $f(x_1) = f(x_2)$ and so $x_1 \sim x_2$ and $[x_1] = [x_2]$ shows that it is 1-1.

Example 3.18. Let $s : \mathbb{R} \to \mathbb{R}$ be $s(x) = \sin(x)$, clearly s is neither $1-1$ nor onto. $\tilde{s} : (\mathbb{R}/\sim) \to [-1, 1]$ given by $\tilde{s}[x] = \sin(x)$ is a bijection.

Example 3.19. Let S be the relation on $X = \mathbb{Z} \times (\mathbb{Z} - \{0\})$ given by

$$(a, b)S(c, d) \iff ad = bc.$$

It's clear that given any $(a, b) \in X$ we have $(a, b)S(a, b)$, so S is reflexive. If $(a, b)S(c, d)$ then certainly $(c, d)S(a, b)$ since the conditions $ad = bc$ and $cb = ad$ are equivalent. So S is symmetric. If $(a, b)S(c, d)$ and $(c, d)S(x, y)$ then $ad = bc, cy = dx$ and so

$$ady = (ad)y = (bc)y = b(cy) = bdx,$$

and since $d \neq 0$ we can cancel it and get $ay = bx$. So $(a, b)S(x, y)$ and S is transitive and an equivalence relation.

Define $f : X/S \to \mathbb{Q}$ by $f([(a, b)]) = \frac{a}{b}$. We claim that f is a bijection, but first we need to establish that f makes sense, that it is well-defined. Suppose that $[(a, b)] = [(c, d)]$ so that $(a, b)S(c, d)$ or $ad = bc$. Then $\frac{a}{b} = \frac{c}{d}$ and so $f([(a, b)]) = f([(c, d)])$. In other words, our definition of f does not depend on the choice of equivalence class representative and hence is well-defined. Clearly f is onto. If $f([(a_1, b_1)]) = f([(a_2, b_2)])$ then $\frac{a_1}{b_1} = \frac{a_2}{b_2}$ and so $a_1b_2 = a_2b_1$. But then $(a_1, b_1)S(a_2, b_2)$ and $[(a_1, b_1)] = [(a_2, b_2)]$ and f is $1 - 1$.

4 Theory of Numbers

We will spend some time discussing/introducing a subject called arithmetic or number theory, which is the study of the integers $\mathbb{Z}$ and especially the non-negative integers $\mathbb{N}$. This will require a type of proof technique that involves proving statements of the following form:

$$\forall n \in \mathbb{N}, P_n \qquad (*)$$

where P_n is a predicate statement that depends on the natural number n. Some examples of such predicates are "$n < 2^n$" or $1 + 3 + 5 + \cdots + (2n + 1) = (n + 1)^2$. We introduce a proof technique to establish statements like $(*)$ above.

Theorem 4.1 (Principle of Mathematical Induction (version 1))**.** *Let P_n be a predicate statement that depends on $n \in \mathbb{N}$. Then in order to prove $\forall n \in \mathbb{N}\, P_n$ it suffices to establish:*

$$1.\, P_0$$
$$2. \forall k \in \mathbb{N}\, P_k \implies P_{k+1}$$

We often refer to establishing P_0 above as establishing the *base case* and assuming the hypothesis P_k of the conditional $P_k \implies P_{k+1}$ as the *inductive hypothesis* or *inductive assumption.*

Example 4.2. Suppose we want to prove $\forall n \in \mathbb{N}\, n < 2^n$. We could use the above principle, which we will often abbreviate 'PMI', as follows:

Let I_n be the predicate statement $n < 2^n$ where $n \in \mathbb{N}$. We proceed to prove $\forall n \in \mathbb{N}\, I_n$ by the PMI. We check the statement I_0 or $0 < 2^0$ which we know is true. Then let's assume I_k holds for some $k \in \mathbb{N}$, i.e, $k < 2^k$. Adding 1 to both sides of this inequality we get $k + 1 < 2^k + 1 \leq 2^k + 2^k = 2^{k+1}$ where we are using the fact that 2^k is an increasing sequence with smallest term 1. So I_{k+1} follows as a consequence of I_k and we've established $I_k \implies I_{k+1}$ and so $\forall n \in \mathbb{N}\, I_n$ by the PMI.

Example 4.3. Suppose we want to prove that $5^{n+1} > n^3 + 2n + 1$ for all $n \in \mathbb{N}$. Let T_n be the predicate statement $5^{n+1} > n^3 + 2n + 1$. We proceed to prove $\forall n \in \mathbb{N}\, T_n$. The base case is the statement T_0 or $5^1 > 0^3 + 2(0) + 1 = 1$ which is certainly true. Let's assume that T_k holds for some $k \in \mathbb{N}$, i.e. that $5^{k+1} > k^3 + 2k + 1$. Then

$$5^{k+1} = 5(5^k) > 5(k^3 + 2k + 1) = 5k^3 + 10k + 5$$

and since $k^3 \geq k^2 \geq k \geq 0$ when $k \geq 0$ we know that $5k^3 + 10k + 5 = k^3 + 4k^3 + 10k + 5 \geq k^3 + 4k^2 + 10k + 5 > k^3 + 3k^2 + 5k + 4 = (k + 1)^3 + 2(k + 1) + 1$. Altogether we have

$$5^{k+1} > (k + 1)^3 + 2(k + 1) + 1,$$

which is exactly the statement T_{k+1}. Since $I_k \implies I_{k+1}$ we have that $\forall n \in \mathbb{N}\, T_n$ by the PMI.

Example 4.4. Suppose we have an 8×8 checkerboard. We remove one square, any square. We would like to see if we can tile the remaining board B with one square removed by 3 square L shaped tiles so that no two tiles overlap, and that every square in B is covered by one of the L tiles. One can convince oneself that this can always be done by simply forming all possible B's and then brute force checking each case, but we turn to induction to find a simpler solution. We think of the 8×8 checkerboard as a $2^3 \times 2^3$ checkerboard and as the more general question: if we let B be the remaining board after a single square is removed from a $2^n \times 2^n$ checkerboard, can B always be tiled with 3 square L shapes?

In fact it can for every $n \in N^+$, and we prove this by induction on $n \geq 1$. Let $T_n :=$ "any $2^n \times 2^n$ checkerboard with one square removed can be tiled by 3 square L shapes". The statement T_1 is that every $2^1 \times 2^1$ checkerboard with one square removed can be tiled by a 3 square L shape; this statement is clearly true as the remaining board would be exactly a 3 square L tile. Assume that T_k holds for some $k \geq 1$ and let B be any $2^{k+1} \times 2^{k+1}$ board with a single square removed. Divide the board into quarters Q_1, Q_2, Q_3, Q_4 each of size $2^k \times 2^k$. We may assume that the missing square is in Q_1 or else we can renumber the quarters. Then Q_1 is a $2^k \times 2^k$ board with one square removed, and so by our inductive hypothesis this can be tiled by L shapes. Q_2, Q_3, Q_4 meet at the center of the board B and so we put an L tile so that it meets these three squares exactly once. This process covers up exactly one square from each of Q_2, Q_3, Q_4 and so by our inductive hypothesis we can tile what remains in these squares. Then the tiling of Q_1, along with the tilings of Q_2, Q_3, Q_4, as well as the L tile that meets all three of Q_2, Q_3, Q_4 provides a tiling of all of B. So T_{k+1} holds, and we are done by the PMI.

Example 4.5 (Backwards Induction)**.** There is a tendency to do induction somewhat backward when one is first getting used to writing proofs. Because this mistake is so common let's discuss it. Suppose we are trying to prove that $3|n^3 - n$ for each $n \geq 0$. Here is what 'backward induction' looks like:

Backward Proof. We proceed to prove $3|n^3 - n$ by induction on $n \geq 0$, the base case being $n = 0$, and we check that 3 divides $0^3 - 0 = 0$ which it does. Assume that $3|k^3 - k$ for some $k \geq 0$. This means $\exists a \in \mathbb{Z}$ so that $k^3 - k = 3a$. Then

$$3|(k+1)^3 - (k+1)$$

and so

$$3|k^3 + 3k^2 + 3k - k$$

hence

$$3|3k^2 + 3k + 3a$$

which implies

$$3|3(k^2 + k + a)$$

which is clearly true. Since we arrived at a true statement we are done by the PMI □

The pieces of a proper proof are all there, but the logic doesn't make any sense. One is working backwards starting with $3|(k+1)^3 - (k+1)$ instead of ending up there. To see what a proper proof by induction looks like we write out the steps that should follow after one states the inductive hypothesis:

Correct Induction. Assume that $3|k^3-k$ for some $k \geq 0$. This means $\exists a \in \mathbb{Z}$ so that $k^3-k = 3a$. Then

$$(k+1)^3 - (k+1) = k^3 + 3k^2 + 3k - k = 3(k^2 + k + a),$$

and since $k^2 + k + a \in \mathbb{Z}$ we know $3|(k+1)^3 - (k+1)$. So our $n = k$ case implies the $n = k+1$ case and we are done by the PMI.

□

Let is examine another example of 'backwards induction' and how to write things properly. Suppose we want to prove that $\forall n \in \mathbb{N}$ with $n > 2$ that $n^3 > 2n^2 + n + 1$ using the PMI. Here is what we want to avoid:

Backward Induction. We proceed by induction on $n > 2$, the base case being $n = 3$ or $3^3 > 2(3^2) + 3 + 1$ or $27 > 22$ which is true. Let's assume that the inequality holds for $n = k > 2$, namely that $K^3 > 2k^2 + k + 1$. We want to prove the case $n = k + 1$ or

$$(k+1)^3 > 2(k+1)^2 + 2(k+1) + 1$$

or

$$k^3 + 3k^2 + 3k + 1 > 2k^2 + 4k + 2 + 2k + 2 + 1$$

or

$$k^3 + 3k^2 + 3k > 2k^2 + 6k + 3$$

..... □

This gets nowhere really fast. One has to build the case $n = k + 1$ up from the case $n = k$. Here is the right way to do this proof by induction:

Proof. We proceed by induction on $n > 2$, the base case being $n = 3$ or $3^3 > 2(3^2) + 3 + 1$ or $27 > 22$ which is true. Let's assume that the inequality holds for $n = k > 2$, namely that $k^3 > 2k^2 + k + 1$. We want to prove the case $n = k + 1$ holds so we first note that since $k > 2$, $k^2 > 2k > 4$. Now

$$(k+1)^3 = k^3 + (3k^2 + 3k + 1) > (2k^2 + k + 1) + (3k^2 + 3k + 1),$$

where we use our I.H to make the inequality. The right hand side (RHS) of this is

$$2k^2 + 4k + 2 + 3k^2 > 2k^2 + 4k + 2 + 6k =$$

$$2k^2 + 5k + 2 + 5k > 2k^2 + 5k + 4,$$

using the inequalities $k^2 > 2k > 4$. Now $2k^2 + 5k + 4 = 2(k+1)^2 + (k+1) + 1$ and so altogether we have established that $(k+1)^3 > 2(k+1)^2 + 2(k+1) + 1$ which is precisely the statement $n = k + 1$. Thus by the PMI we are done. □

Definition 4.6. let $a, b \in \mathbb{Z}$. Then we write $a|b$ and say a *divides* b or b is a multiple of a if $\exists k \in \mathbb{Z}$ so that $b = ak$.

Theorem 4.7 (Even Lemma)**.** $\forall n \in \mathbb{N}\, 2|n(n+1)$.

Proof. Let $E_n :=$ "$2|n(n+1)$". We proceed to prove $\forall n \in \mathbb{N}\, E_n$ by the PMI, the base case or E_0 being the statement that $2|0(0+1)$ or $2|0$ which we know is true since every integer divides 0. Now assume that E_k holds for some $k \in \mathbb{N}$ so that $\exists t \in \mathbb{Z}$ so that $2t = k(k+1) = k^2 + k$. Then $(k+1)((k+1)+1) = k^2 + 3k + 2 = (k^2 + k) + (2k + 2) = 2t + 2k + 2 = 2(t + k + 1) = 2s$ where $s \in \mathbb{Z}$. This means that E_{k+1} follows and so we are done by the PMI. □

Example 4.8. Let's prove that $\forall n \in \mathbb{N}\ \ 6|n^3 - n$. Let $D_n := 6|n^3 - n$" and let's proceed to prove $\forall n \in \mathbb{N}\, D_n$ by the PMI. First we check that our base case D_0 holds, which means checking $6|0^3 - 0$ or $6|0$ which is true since every integer divides 0. Now assume D_k holds for some $k \in \mathbb{N}$ so that $\exists c \in \mathbb{Z}$ so that $6c = k^3 - k$ or $k^3 = 6c + k$. Then $(k+1)^3 - (k+1) = k^3 + 3k^2 + 2k = 6c + 3k^2 + 3k6 = 6c + 3k(k+1)$. But by the Even Lemma we know that $\exists t \in \mathbb{Z}$ so that $k(k+1) = 2t$ and so $6c + 3k(k+1) = 6c + 3(2t) = 6(c+t)$ and so $6|(k+1)^3 - (k+1)$ since $c + t \in \mathbb{Z}$. So D_{k+1} follows and we are done by the PMI. Note that since $(-n)^3 - (-n) = -(n^3 - n)$, we have that for any *integer* n, $6|n^3 - n$.

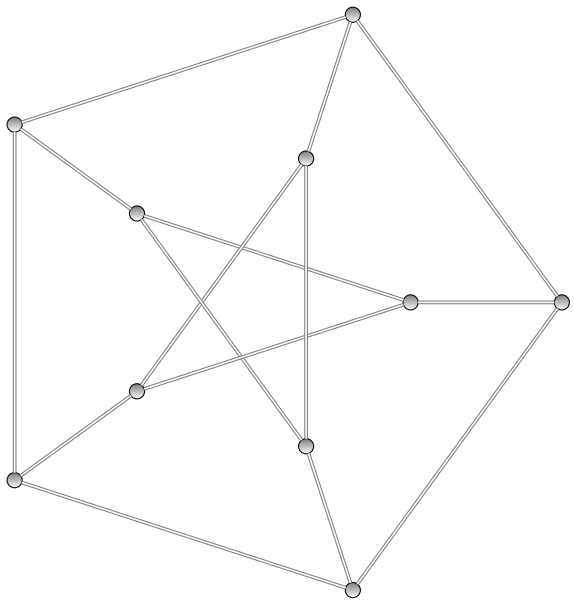
Example 4.9. Let's look at the sequence $5^n + 2 \cdot 11^n$ when $n = 0, 1, 2, \ldots$. The first few terms are $3, 27, 267, 2787, 29907$ which when factored are $3, 3^3, 3 \cdot 89, 3 \cdot 929, 3^2 \cdot 3323$, and we guess that each number of the form $5^n + 2 \cdot 11^n$ is divisible by 3. Let G_n be the predicate statement $3|5^n + 2 \cdot 11^n$; we want to prove that $\forall n \in \mathbb{N}\, G_n$ holds. The base case is when $n = 0$, and we've already checked that $3|3$. So let's assume that G_k is true for some $k \in \mathbb{N}$, namely that $3|5^k + 2 \cdot 11^k$. This means that $\exists t \in \mathbb{Z}$ so that $3t = 5^k + 2 \cdot 11^k$ or $5^k = 3t - 2 \cdot 11^k$. Then

$$5^{k+1} + 2 \cdot 11^{k+1} = 5(3t - 2 \cdot 11^k) + 2 \cdot 11^{k+1} = 15t - 10 \cdot 11^k + 2 \cdot 11^{k+1} =$$
$$15t - 11^k(-10 + 2 \cdot 11) = 15t - 11^k(12) = 3(5t - 4 \cdot 11^k),$$

and since $5t - 4 \cdot 11^k \in \mathbb{Z}$ we know that G_{k+1} holds. So $G_k \implies G_{k+1}$ and so by the PMI we have that $\forall n \in \mathbb{N}\, G_n$ holds.

Definition 4.10. A directed graph $\Gamma = (V, E)$ consists of a pair of sets, V are called the vertices of Γ and E is a collection of sets of the form $e = (v_1, v_2)$ where $v_i \in V$. A *path* in Γ consists of a collection of vertices $v_1, v_2, \ldots, v_k$ where for each $1 \leq i < k$ $(v_i, v_{i+1}) \in E$. A path is simple if its vertices are distinct.

Example 4.11. The Peterson graph P pictured below consists of 10 vertices (the black dots) and 15 edges:



Each relation R on a set X determines a directed graph Γ_R as follows: the vertices of Γ_R are the set X and the edges of Γ_R are just the ordered pairs $(a, b) \in R$. Let $R^0 = \{(a, a) | a \in X\}$ and $R^{n+1} = R^n \circ R$. Then $(a, b) \in R^k$ for some $k \geq 0$ exactly when there is a path of length k in Γ_R from a to b.

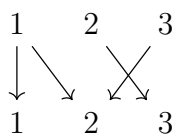
Definition 4.12. Let R be a relation on X. Let $R^* = R^0 \cup R^1 \cup R^2 \cup \dots$ be the collection of all ordered pairs (a, b) so that there is a path from a to b. We call R^* the *reflexive/transitive closure* of R.

Theorem 4.13. *For any R we have that R^* is transitive. Moreover if $R \subseteq S$ and S is reflexive and transitive, we must have $R^* \subseteq S$, i.e. that R^* is the smallest transitive relation containing R.*

Proof. The fact that $R^0 \subseteq R^*$ immediately tells us that R^* is reflexive. Let $(a, b), , (b, c) \in R^*$. Then $(a, b) \in R^k, (b, c) \in R^l$ for some $k, l \geq 0$. So there is a path of length k from a to b and a path of length l from b to c, so by traversing one path followed by the next, we have a path from a to c that is of length $k + l$ and so $(a, c) \in R^{k+1} \subseteq R^*$. So R^* is transitive. Suppose S is reflexive and transitive with $R \subseteq S$. We prove by induction that $R^n \subseteq S$ for all $n \geq 0$. The base case is $R^0 \subseteq S$, which holds since S s reflexive. If $R^k \subseteq S$ for some $k \geq 0$ we have that $R^{k+1} = R^k \circ R \subseteq S \circ S \subseteq S$ as well and we've proven our claim by the PMI. But then $R^* = R^0 \cup R^1 \cup \dots \subseteq S$ as well. □

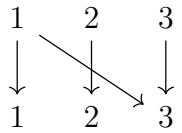
Example 4.14. Let $X = \{1, 2, 3\}$ and let R be given by the following arrow diagram:

R :



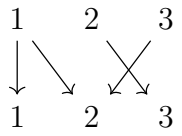
Then R^2 has diagram

R^2 :



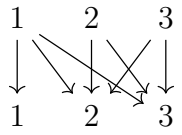
In particular since $R^2 \not\subseteq R$ then R is not transitive. R^3 has diagram

R^3 :



and so $R^3 = R$ and hence $R^4 = R^2, R^5 = R$, etc... So $R^* = R^0 \cup R \cup R^2$ has arrow diagram

R^* :



We encourage the reader to check that $R^* \circ R^* \subseteq R^*$ and to draw the graph Γ_R.

There are statements like $\forall n \in \mathbb{N}, n \geq 3, n^2 < 3^n$ where the truth of the predicate statement $n^2 < 3^n$ doesn't become consistent until we reach a certain integer, in this case $n = 3$, but where we would like to use the PMI. We can modify the PMI as follows:

Theorem 4.15 (Principle of Mathematical Induction (version 1*))**.** *Let P_n be a predicate statement that depends on an integer n and let $a \in \mathbb{Z}$ be a fixed integer. Then in order to prove $\forall n \in \mathbb{Z}, n \geq a, P_n$ it suffices to establish:*

$$1.P_a$$

$$2.\forall k \in \mathbb{Z}, k \geq a,\ P_k \Longrightarrow P_{k+1}$$

Example 4.16. Let us prove that $\forall n \in \mathbb{N}, n \geq 3$ that $n^2 < 3^n$. We proceed by induction on $n \geq 3$, the base case being when $n = 3$, and we check that $3^2 < 3^3$ holds, which it does. Assume that $k^2 < 3^k$ for some $k \geq 3$. Since $k \geq 3$ we have $k^2 \geq 3k \geq 9$. Then

$$3^{k+1} = 3(3^k) > 3k^2 = k^2 + 2k^2 \geq k^2 + 6k > k^2 + 2k + 1 = (k+1)^2,$$

namely that $(k^2 < 3^k) \Longrightarrow (3^{k+1} > (k+1)^2)$ when $k \geq 3$. So by the PMI version $1*$ we have that $\forall n \in \mathbb{N}, n \geq 3$ that $n^2 < 3^n$.

Example 4.17. We show that $(A_1 \cup A_2 \cup \cdots \cup A_n) \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$ for any sets $A_1, A_2, \ldots, A_n, B$. We proved this when $n = 2$ in the first chapter, now we wish to extend the result. We proceed to prove $(A_1 \cup A_2 \cup \cdots \cup A_n) \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$ by induction on $n \geq 2$, the base case being a theorem we've already proven. Assume that $(A_1 \cup A_2 \cup \cdots \cup A_k) \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_k \cap B)$ for any sets $A_1, A_2, \ldots, A_k, B$ where $k \geq 2$. Looking at $(A_1 \cup A_2 \cup \cdots \cup A_k \cup A_{k+1}) \cap B$ if we set $A' = (A_1 \cup A_2 \cup \cdots \cup A_k)$.

Then $(A_1 \cup A_2 \cup \cdots \cup A_k \cup A_{k+1}) \cap B = (A' \cup A_{k+1}) \cap B = (A' \cap B) \cup (A_{k+1} \cap B)$. By our inductive hypothesis $A' \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_k \cap B)$ and so altogether we have:

$$(A_1 \cup A_2 \cup \cdots \cup A_k \cup A_{k+1}) \cap B = ((A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_k \cap B)) \cup (A_{k+1} \cap B) =$$

$$(A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_k \cap B) \cup (A_{k+1} \cap B),$$

and so by the PMI version 1^* we are done.

Theorem 4.18 (Well Ordering Principle)**.** *Let $\emptyset \neq A \subseteq \mathbb{N}$. Then A has a least element, namely $\exists x \in A$ so that $\forall y \in A$ we have $x \leq y$.*

Proof. We prove the contrapositive, that if A has no least element that A must be empty. Let P_k be the statement that $k \in \mathbb{N} - A$ where $k \in \mathbb{N}$. We prove that $\forall k \in \mathbb{N}\, P_k$ by induction on $k \geq 0$. The base case or P_0 is the statement that $0 \in \mathbb{N} - A$. This is true because otherwise 0 would be the least element of A and we are assuming A has no least element. Assume that P_k holds for some $k \geq 0$. If any of the numbers $0, 1, 2, \ldots, k \in A$ then one of these would be the least element of A (why?). If $k+1 \in A$ then $k+1$ would be the least element of A which can't happen. So $k+1 \in \mathbb{N} - A$ and so P_{k+1} holds, and hence by the PMI we know that for all $n \in \mathbb{N}$ we have $n \in \mathbb{N} - A$. But this means that $\mathbb{N} - A = \mathbb{N}$ or that $A = \emptyset$. □

Definition 4.19. Let $n \in \mathbb{N}$ with $n \geq 2$. Define a relation $\equiv$ on $\mathbb{Z}$ by $a \equiv b$ if $n|(a-b)$. When we want to emphasize the role of the specific natural number n, called the modulus, we write $a \equiv b \pmod{n}$. This gives us a family of equivalence relations for each n (check this!), and we often say a is congruent to b mod (or modulo) n for $a \equiv b \pmod{n}$. We denote the set of equivalence classes of this relation by $\mathbb{Z}/n$.

We can show there are precisely n distinct equivalence classes in this relation as follows, first recalling a well known fact (which can be proven by induction on m):

Theorem 4.20 (Division Algorithm)**.** *Let $n \in \mathbb{N}^+$ and $m \in \mathbb{Z}$. Then $\exists q, r \in \mathbb{Z}$ with $0 \leq r < n$ so that $m = qn + r$. Moreover the q, r here are unique given m and n.*

Proof. Fix $n \in \mathbb{N}^+$. Case 1: $m \in \mathbb{N}$. We proceed by induction on $m \in \mathbb{N}$. If $m = 0$ we set $q = r = 0$ to see the base case holds here. Assume that we can write $k = qn+r$ where $0 \leq r < n$. Then $k+1 = qn + r + 1$ so if $r < n-1$ then we see that $r+1 < n$ and $k+1 = qn + (r+1)$ is the desired expression. If $r = n-1$ then $k+1 = (q+1)n + 0$ is the desired expression. Either way we get a non-negative remainder less than n. Case 2: $m < 0$. Then $-m > 0$ and we can apply case 1 to get $-m = qn + r$ where $0 \leq r < n$. If $r = 0$ then $m = (-q)n + 0$ gives us a remainder of zero. Otherwise $m = -qn - r = -qn - n + n - r = (-q-1)n + (n-r)$ gives us remainder $0 < n - r < 0$ since $0 < r < n$.

To prove uniqueness suppose we had $m = qn + r = q'n + r'$ where $0 \leq r' \leq r < n$. Then $0 \leq r' - r < n$ with $r' - r = n(q' - q)$. Since $n > 0$ this forces $q = q'$ and hence $r = r'$ as well. □

Returning to our example we establish some simple lemmas fixing a modulus n.

Lemma 4.21. *Given $m \in \mathbb{Z}$, we apply the Division Algorithm to m and n to get $m = qn + r$ with $0 \leq r < n$. Then $[m] = [r]$.*

Proof. We know $m - r = qn$ so $n|m - r$ and so $m \equiv r \pmod{n}$. □

Lemma 4.22. *For $0 \leq r < r' < n$ we have $[r] \neq [r']$.*

Proof. Assume this is false and that $[r] = [r']$ so that $n|(r' - r)$. Then $\exists k \in \mathbb{N}^+$ so that $r' - r = nk$ and since $k \geq 1$ we know $nk \geq n$ and so $r' - r \geq n$. But since $r < r' < n$ we know $0 < r' - r < n - r \leq n$, a contradiction. So $[r] \neq [r']$. □

Corollary 4.22.1. *$\mathbb{Z}/n$, the set of equivalence classes mod n, is*

$$\{[0], [1], \ldots, [n-1]\}$$

consisting of n distinct equivalence classes.

Proof: This is a consequence of the above two lemmas: every integer m is in the equivalence class determined by the Division Algorithm, and these classes are distinct. ♢

We now define two operations $+, *$ on the $\mathbb{Z}/n$ as follows. Let $a, b \in \mathbb{Z}$ and set

$$[a] + [b] := [a + b]$$
$$[a] * [b] := [ab].$$

We need to show that these operations are well-defined. Suppose that $[a] = [a'], [b] = [b']$. Then $a \equiv a' \pmod{n}, b \equiv b' \pmod{n}$. Then $a - a' = ns, b - b' = nt$ for some $s, t \in \mathbb{Z}$, and we check: $[a'] + [b'] = [a' + b'] = [(a - ns) + (b - nt)] = [a + b - (s + t)n] = [a + b] = [a] + [b]$ and $[a'] * [b'] = [a'b'] = [(a - ns)(b - nt)] = [ab - (s + t)n + stn^2] = [ab] = [a] * [b]$, showing that the two operations are well defined up to equivalence.

Example 4.23. In $\mathbb{Z}/6$ we have

$$[3] + [4] = [7] = [1]$$
$$[3][4] = [12] = [0]$$
$$[5][5] = [-1][-1] = [1]$$

etc... In $\mathbb{Z}/5$ we have $[2] + [3] = [0], [2][3] = [1], [4][4] = [1], [3][4] = [2]$ etc...

We make a note here that the statements $a \equiv b \pmod{n}$ and $[a] = [b]$ in $\mathbb{Z}/n$ are logically equivalent and are often used interchangeably. If there are multiple moduli involved and there is a risk of confusion, we can always use $[a]_n$ to denote the equivalence class of a modulo n.

Definition 4.24. Let $n \in \mathbb{N}$ with $n \geq 2$. We say n is prime if $k|n$ with $k \in \mathbb{N}$ implies $k = 1$ or $k = n$. If $n \geq 2$ is not prime we say it is composite.

Theorem 4.25 (Fundamental Theorem of Arithmetic)**.** *Every natural number greater than or equal to two can be factored uniquely as a product of primes.*

When we write 'factored uniquely' we mean up to rearrangement of the factors: (2)(3) and (3)(2) are to be considered the same factorization of the composite number 6. We prove part of the Fundamental Theorem of Arithmetic shortly that any factorization is necessarily unique, and save the existence of such a factorization for a later section when we introduce a revamped version of the PMI. First we need some tools.

Theorem 4.26. *There are infinitely many prime numbers.*

Proof. Suppose there were finitely many primes $p_1, p_2, \ldots, p_n$. Let $M = p_1 p_2 ... p_n + 1$ and note that for each $1 \leq k \leq n$ we have $p_k | p_1 ... p_n$. We also see that $p_k \nmid M$ or else $p_k | (M - p_1 p_2 ... p_n)$ or $p_k | 1$ which can't happen since $p_k > 1$. So M is a natural number greater than 2 that has no prime factors, contradicting the Fundamental Theorem of Arithmetic. □

A useful took for studying properties of prime and composite numbers is the concept of greatest common divisor. We give a slightly different definition of this from what most people see in middle school mathematics which is a bit more useful.

Definition 4.27. Let $a, b \in \mathbb{Z}$ with a, b not both zero. Define:

$$\gcd(a, b) = min\{k \in \mathbb{N}^+ \mid \exists s, t \in \mathbb{Z}, k = sa + tb\},$$

and define $\gcd(0, 0) = 0$. Note that we needed a, b not both zero so that the above set isn't empty.

For example if $a \neq 0$ then $\gcd(0, a) = min\{|a|, 2|a|, 3|a|, ...\} = |a|$. If $a = 4, b = -6$ we are looking at the smallest element of the set $\{4s + (-6)t \in \mathbb{N}^+ \mid s, t \in \mathbb{Z}\} = \{2, 4, 6, ...\}$ which is 2 since we can write $2 = 4(2) + (-6)(1) = 4(-1) + (-6)(-1)$ etc...Note that there the s, t are not unique in general as the last example demonstrates. What we do get right away is that the greatest common divisor of two integers is a linear combination of the integers with integer coefficients. In fact the definition is that it is the smallest positive linear combination.

Theorem 4.28 (GCD Theorem)**.** *For each $a, b \in \mathbb{Z}$ we have:*

1. $\gcd(a, b)|a$ *and* $\gcd(a, b)|b$
2. $k \in \mathbb{Z}, k|a, k|b$ *implies that* $\gcd(a, b)|k$

Proof. If a, b are both zero the theorem certainly holds, so let's assume $a \neq 0$. If $b = 0$ then $\gcd(a, b) = |a|$ and it's easy to see that the theorem holds in this case (why?). So let's further assume $b \neq 0$. We first examine the case $a, b > 0$. Let $g := \gcd(a, b)$. Then $\exists s, t \in \mathbb{Z}$ so that $g = sa + tb$. By the Division Algorithm we get that $a = gq + r$ where $0 \leq r < g$. If $r = 0$ then clearly $g|a$. Suppose that $r \neq 0$. Then

$$r = a - gq = a - (sa + tb)q = (1 - sq)a + (-tq)b$$

is a smaller positive linear combination of a and b than g which is not possible. So $r = 0$ and $g|a$. Similarly $g|b$ and we are done with part 1. Let $k \in \mathbb{Z}$ with $k|a$ and $k|b$. Then $a = kx, b = ky$ for some $x, y \in \mathbb{Z}$. Then $g = sa + tb = skx + tky = k(sx + ty)$ and so $g|k$ and we are done with part 2. Let $l \in \mathbb{Z}$. We leave it to the reader to supply the argument when one or both of a, b are negative. □

The two properties in the GCD Theorem are often taken as the definition of greatest common divisor, but we derive these properties as consequences of our definition.

Lemma 4.29. *If p is a prime, $a \in \mathbb{Z}$ then $\gcd(a, p) = 1$ or $\gcd(a, p) = p$. $\gcd(a, p) = 1$ iff $p \not| a$.*

Proof. $\gcd(a, p)|p$ and so $\gcd(a, p) = 1$ or p since p is prime. If $\gcd(a, p) = 1$ then by definition $\exists s, t \in \mathbb{Z}$ so that $sa + tp = 1$ and if $p|a$ we'd have $p|(sa + tp)$ or $p|1$ which can't happen. So $p \not| a$. If $\gcd(a, p) \neq 1$ then $\gcd(a, p) = p$ and so $p|a$. □

Lemma 4.30 (Euclid's Lemma)**.** *Let p be a prime number so that $p|ab$ where $a, b \in \mathbb{Z}$. Then $p|a$ or $p|b$.*

Proof. Suppose that $p \not| a$. Then $\gcd(a, p) = 1$ and so by definition of greatest common divisor $\exists s, t \in \mathbb{Z}$ so that $sa + tp = 1$. Then $bsa + btp = b$. Now $p|btp$ and $p|bsa$ since $p|ab$ and so $p|(bsa + btp)$ or $p|b$. □

Note that Euclid's lemma fails for composite numbers: we have $6|(3)(4)$ but $6 \not| 3$ and $6 \not| 4$ for instance. Note that the proof of Euclid's lemma depended on our funny way of defining greatest common divisor.

Proof of Uniqueness, FTA. Suppose $n \in \mathbb{N}$ with $n \geq 2$ has two factorizations $n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k} = q_1^{\beta_1} q_2^{\beta_2} \dots q_l^{\beta_l}$ where $p_1 \leq p_2 \leq \dots \leq p_k$ and $q_1 \leq q_2 \leq \dots \leq q_l$ are primes and assume without loss of generality that $k \leq l$. Since $p_1 | q_1^{\beta_1} q_2^{\beta_2} \dots q_l^{\beta_l}$ we must have by Euclid's lemma that p_1 must be one of the primes q_i and so $p_1 \geq q_1$. Since $q_1 | p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ the same argument gives us that $q_1 \geq p_1$ and so $p_1 = q_1$. So we can divide both sides of $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_2} = p_1^{\beta_1} q_2^{\beta_2} \dots q_l^{\beta_l}$ by p_1 to get $p_1^{\alpha_1 - 1} p_2^{\alpha_2} \dots p_k^{\alpha_2} = p_1^{\beta_1 - 1} q_2^{\beta_2} \dots q_l^{\beta_l}$. We keep dividing by p_1 until we run out of $p_1's$ on one side. There can't be any $p_1's$ left over on either side or we would contradict Euclid's Lemma. So $\alpha_1 = \beta_1$. We continue this dividing process with p_2 to get that $q_2 = p_2$ and $\alpha_2 = \beta_2$. We continue with dividing process with each successive $p_i = q_i$ until we finish with one side being 1. Then the other side must be 1 as well and this forces $k = l$, and these factorizations to be exactly the same. □

Definition 4.31. A class $[a] \in \mathbb{Z}/n$ is said to be a unit if $[a]x = [1]$ has a solution in $\mathbb{Z}/n$. We also say $a \in \mathbb{Z}$ is a unit modulo n if $ax \equiv 1 \pmod{n}$ has a solution. Note that these are essentially the same concept, one is expressed using the language of equivalence classes, the other using the equivalence relation itself. We write $(\mathbb{Z}/n)^*$ for the set of classes of units. Note that $[1]$ and $[-1] = [n-1]$ are always units and $[0]$ is never a unit.

Example 4.32. In $\mathbb{Z}/6 = \{[0], [1], [2], [3], [4], [5]\}$ there are exactly two units, $[1]$ and $[5] = [-1]$. We can rule out $[2], [3], [4]$, as follows: suppose $[2]x = [1]$ had a solution. Then $[3][2]x = [3][1]$ or $[0]x = [3]$. But the left hand side is $[0]$ for any x and this is a different class than $[3]$.

Example 4.33. In $\mathbb{Z}/10 = \{[0], [1], [2], \ldots, [8], [9]\}$ we find four units, $(\mathbb{Z}/10)^* = \{[1], [3], [7], [9]\}$.

Theorem 4.34. *a is a unit modulo n iff* $\gcd(a, n) = 1$.

Proof. Suppose that $ax \equiv 1 \pmod{n}$ has a solution. Then $\exists j \in \mathbb{Z}$ so that $ax - 1 = jn$. But then $ax + (-j)n = 1$ is a positive linear combination of a and n and since it is one it must be the greatest common divisor by definition. Likewise if $\gcd(a, n) = 1$ then $\exists s, t \in \mathbb{Z}$ so that $as + nt = 1$. But then $as = 1 - nt \equiv 1 \pmod{n}$ shows that $x = s$ is a solution to $ax \equiv 1 \pmod{n}$ and hence a is a unit modulo n. □

Theorem 4.35. *If a, b are both units modulo n then so is ab. If $a_1, a_2, \ldots, a_k$ are units modulo n where $k \in \mathbb{N}^+$ then $a_1 a_2 ... a_k$ is a unit modulo n.*

Proof. We know $ax \equiv 1 \pmod{n}$ and $by \equiv 1 \pmod{n}$ both have solutions $x, y \in \mathbb{Z}$. This means that $\exists j, k \in \mathbb{Z}$ so that $ax - 1 = jn, by - 1 = kn$ and so $(ab)(xy) = (ax)(by) = (1 + jn)(1 + kn) = 1 + (j + k)n + (jk)n^2 = 1 + n(\text{stuff})$ where stuff $\in \mathbb{Z}$ and so $(ab)(xy) \equiv 1 \pmod{n}$ and hence ab is a unit. The second claim is accomplished by induction on k (do this...). □

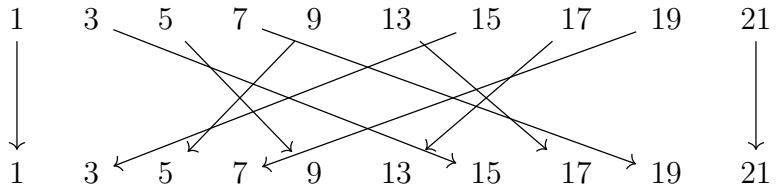
Definition 4.36. Let $a \in \mathbb{Z}$ and $n \geq 2$. If a is a unit modulo n (or equivalently if $[a]$ is a unit in $(\mathbb{Z}/n)^*$ and $ax \equiv 1 \pmod{n}$ or $[a]x = [1]$ in $\mathbb{Z}/n$, we write $a^{-1} \equiv x \pmod{n}$ or $[a]^{-1} = [x]$ and say x (or $[x]$) is the *inverse* of a modulo n.

Example 4.37. Let $n = 15$. We know that $(\mathbb{Z}/15)^* =$
$\{[1], [2], [4], [7], [8], [11], [13], [14]\}$ by checking greatest common divisors. We check that

$$[1]^{-1} = [1], [2]^{-1} = [13], [4]^{-1} = [4], [7]^{-1} = [13],$$
$$[11]^{-1} = [11], [13]^{-1} = [2], [14]^{-1} = [14].$$

Example 4.38. Let $n = 22$.

We know that $(\mathbb{Z}/22)^* = \{[1], [3], [5], [7], [9], [13], [15], [17], [19], [21]\}$. We draw an arrow diagram connecting each unit with its inverse:



We observe some patterns, the first being that there is left/right symmetry in this picture. This comes from the identity $xy \equiv 1 \pmod{22} \iff (-x)(-y) \equiv 1 \pmod{22}$, for instance, we see that $5 \cdot 9 \equiv 1 \pmod{22}$ and hence $(-5)(-9) \equiv 1 \pmod{22}$. But $-5 \equiv 17 \pmod{22}$ and $-9 \equiv 13 \pmod{22}$ and so we see that the arrow $5 \to 9$ produces the arrow $13 \to 19$ when we reflect the diagram about its center vertical axis. Note that the function that sends a unit to its inverse is also a permutation of the set of units.

If p is a prime number then $1, 2, 3, \ldots, p-1$ are all units modulo p in other words:

$$(\mathbb{Z}/p)^* = \{[1], [2], \ldots, [p-1]\}.$$

Lemma 4.39 (Cancellation mod n)**.** *Let $n \geq 2$ and $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Then $ax \equiv ay \pmod{n}$ implies that $x \equiv y \pmod{n}$. In other words, we can cancel units modulo n.*

Proof. Suppose that $ax \equiv ay \pmod{n}$ so that $ax - ay = a(x - y) = jn$ for some $j \in \mathbb{Z}$. Since $\gcd(a, n) = 1$ we know $a \neq 0$ and $a(x - y)$ is a product equal to a multiple of n, non of the prime factors of a can divide n and so they must divide j, in other words, $a|j$. But then $\frac{j}{a} \in \mathbb{Z}$ and so $x - y = \frac{j}{a}n$ gives that $x \equiv y \pmod{n}$. □

Theorem 4.40 (Fermat's Little Theorem)**.** *Let p be a prime and let $a \in \mathbb{Z}$ be a unit modulo p. Then $a^{p-1} \equiv 1 \pmod{p}$.*

Proof. Examine the set $S = \{[1], [2], [3], \ldots [(p-1)]\}$. Each element of this set is a unit modulo p and no two of these units are equal. Examine the set $S' = [a]S := \{[1a], [2a], [3a], \ldots, [(p-1)a]\}$ where we multiply each of the $p - 1$ units in S by $[a]$. This list consists of units as well and we claim that no two members of this list are equal. If we had $[ka] = [la]$ then by the cancellation property we would have $[k] = [l]$. So the list S' also consists of $p - 1$ distinct units modulo p. Since there are only $p - 1$ distinct units up to congruence modulo p, these lists contain the same members, perhaps in a different order. We multiply all the members of S together to get $[1] \cdot [2] \cdot \cdots \cdot [(p-1)] = [(p-1)!]$. We multiply all members of S' together to get $[1a] \cdot [2a] \cdot \cdots \cdot [(p-1)a] = [a]^{p-1}[(p-1)!]$. These products must be equal so we have

$$[(p-1)!] = [a]^{p-1}[(p-1)!].$$

But $[(p-1)!]$ is a product of units and hence a unit, and so we can cancel it from both sides of the above equation to get

$$[1] = [a]^{p-1}$$

which can be rephrased as $a^{p-1} \equiv 1 \pmod{p}$. □

Corollary 4.40.1 (Fermat's Little Theorem again)**.** *If p is a prime and $a \in \mathbb{Z}$ then $a^p \equiv a \pmod{p}$.*

Proof. If $\gcd(a, p) = p$ then $a^p \equiv a \equiv 0 \pmod{p}$ while
if $\gcd(a, p) = 1$ this follows from the above form of FLT by multiplying both sides by a. □

Example 4.41. Let's use Fermat's Little Theorem (or FLT) to find the inverse of the unit 13 modulo 17. FLT tells us that $13^{16} \equiv 1 \pmod{17}$ or $13(13^{15}) \equiv 1 \pmod{17}$ and hence $[13]^{-1} = [13]^{15}$. We can simplify this pretty easily: $13 \equiv -4 \pmod{17}$ and so $13^2 \equiv (-4)^2 \equiv 16 \equiv -1 \pmod{17}$. Then $13^4 \equiv (-1)^2 \equiv 1 \pmod{17}$ and $13^8 \equiv 1^2 \equiv 1 \pmod{17}$. Putting this altogether we find that $13^{15} = (13)^8(13)^4(13)^2(13) \equiv (1)(1)(-1)(-4) \equiv 4 \pmod{17}$. So $[13]^{15} = [4]$ and hence $[13]^{-1} = [4]$. We check that $(13)(4) = 52 \equiv 1 \pmod{17}$.

What we saw in the last example is that the inverse of $[a] \in (\mathbb{Z}/p)^*$ is $[a]^{p-2}$ based on Fermat's Little Theorem.

Example 4.42. Let's find the inverse of the unit 23 modulo 31. Using the idea mentioned above we see that $[23]^{29} = [23]^{-1}$. Now $23 \equiv -8 \pmod{31}$ and so $23^2 \equiv (-8)^2 \equiv 64 \equiv 2 \pmod{31}$ and so $23^4 \equiv 2^2 \equiv 4 \pmod{31}$, $23^8 \equiv 4^2 = 16 \pmod{31}$, $23^{16} \equiv 16^2 = 256 \equiv 8 \pmod{31}$. Altogether we get that

$$23^{29} = (23^{16})(23^8)(23^4)(23) \equiv (8)(16)(4)(-8) \equiv -2^{12} \pmod{31}.$$

Since $2^5 = 32 \equiv 1 \pmod{31}$ we have $-2^{12} = -2^2(2^5)^2 \equiv (-4)(1)^2 \equiv 27 \pmod{31}$ giving us that $[23]^{-1} = [27]$. We check that $(23)(27) \equiv (-8)(-4) \equiv 32 \equiv 1 \pmod{31}$.

This might seem like a cumbersome way of finding inverses of units, and it is restricted to finding inverses modulo a prime (we will see shortly that Fermat's Little Theorem has a generalization by Euler to include composite numbers as well; in fact, Euler gave the first proof of Fermat's Little Theorem when he generalized it). There are other much older ways we can find inverses, one way in particular begins with the following famous theorem/algorithm:

Theorem 4.43 (Euclidean Algorithm)**.** *Let $a, b \in \mathbb{N}^+$. Apply the Division Algorithm successively to get:*

$$\begin{aligned} a &= bq_1 + r_1 \qquad 0 \leq r_1 < b \\ b &= q_2 r_1 + r_2 \qquad 0 \leq r_2 < r_1 \\ r_1 &= q_3 r_2 + r_3 \qquad 0 \leq r_3 < r_2 \end{aligned}$$

etc..

$$r_{k-1} = q_{k+1}r_k + r_{k+1} \qquad 0 \leq r_{k+1} < r_k$$

Eventually there will be some first $r_n = 0$. Then $r_{n-1} = \gcd(a, b)$.

Proof. The fact that there will be some first $r_n = 0$ follows from the fact that the remainders in each step get strictly smaller while remaining non-negative. So we just need to establish that $r_{n-1} = \gcd(a, b)$. We see immediately that $r_{n-2} = q_n r_{n-1}$ and so $r_{n-1}|r_{n-2}$ by definition. In turn $r_{n-3} = q_{n-1}r_{n-2} + r_{n-1}$ and since $r_{n-1}|r_{n-1}$ and $r_{n-1}|r_{n-2}$ we know that $r_{n-1}|(q_{n-1}r_{n-2} + r_{n-1})$ or $r_{n-1}|r_{n-3}$. We keep going with this process to get that $r_{n-1}|r_k$ for each $1 \leq k \leq n-1$, and so $r_{n-1}|b$ and $r_{n-1}|a$ as well. So r_{n-1} is a common factor of a and b and hence by the GCD theorem $r_{n-1}|\gcd(a, b)$. Let $g = \gcd(a, b)$ and note that $g|(a - bq_1)$ or $g|r_1$. Likewise $g|(b - q_2r_1)$ or $g|r_2$. Repeating this and working backwards we see that $g|r_k$ for any k and so $g|r_{n-1}$. Since $g|r_{n-1}$ and $r_{n-1}|g$ we have $g = r_{n-1}$. □

Example 4.44. Let's put this to use: let's compute $\gcd(234, 42)$, letting $a = 234, b = 42$ in the EA we find that $234 = (5)(42) + 24, 42 = (1)(24) + 18, 24 = (1)(18) + 6, 18 = (3)(6) + 0$ and so $\gcd(234, 42) = 6$ which we know is correct.

Example 4.45. Let's show that 22 is a unit modulo 47 by using the Euclidean Algorithm to show $\gcd(22, 47) = 1$. $47 = (2)22 + 3, 22 = (7)3 + 1$ and so $\gcd(22, 47) = 1$.

Note that the Euclidean Algorithm doesn't require that we know how to factor a and b, it just requires that we can apply the division algorithm repeatedly, a much much simpler process. We can also use it to find inverses modulo n as the next example shows.

Example 4.46. In the example above where we compute $\gcd(22, 47) = 1$ we can solve for the remainders to get that $1 = 22 - 7(3), 3 = 47 - 2(22)$. Substituting the second equation into the first we get $1 = 22 - 7(47 - 2(22)) = (15)(22) + (-7)(47)$. Note that what we have just done is to express our greatest common divisor, in this case 1, as an explicit linear combination of 22 and 47. Then $1 \equiv (15)(22) + (-7)(47) \equiv (15)(22) \pmod{47}$ and so $[22]^{-1} = [15]$ in $\mathbb{Z}/47$.

Example 4.47. We find $[45]^{-1}$ in $\mathbb{Z}/101$. $101 = 2(45) + 11, 45 = 4(11) + 1$ and so we get $1 = 45 - 4(101 - 2(45)) = (9)(45) + (-4)(101)$ giving us $[45]^{-1} = [9]$. Let's compute $[2]^{-1}$ in $\mathbb{Z}/101$. $101 = 2(50) + 1$ and so $1 = (2)(-50) + (1)(101)$ and so $[2]^{-1} = [-50] = [51]$ in $\mathbb{Z}/101$.

We can use this method of finding inverses to solve linear congruences modulo n where by a linear congruence we mean a congruence of the form $ax \equiv b \pmod{n}$ where $a, b \in \mathbb{Z}$. We first make some general observations about such congruences.

Example 4.48. Suppose we are looking to solve $6x \equiv 4 \pmod{10}$. This is a fair bit different than solving the linear equation $6x = 4$ in $\mathbb{Z}$, which has no solution, or the linear equation $6x = 4$ in $\mathbb{R}$ which has a unique solution $x = 2/3$. We see that $x = 4$ and $x = 9$ are both solutions to $6x \equiv 4 \pmod{10}$ and moreover, any number congruent to either of these two solutions modulo 10 is also a solution.

Theorem 4.49. *Let $g = \gcd(a, n)$ where $a \in \mathbb{Z}$ and $n \geq 2$. Then the linear congruence $ax \equiv b \pmod{n}$ has no solutions if $g \not| b$ and has g solutions up to equivalence modulo n if $g|b$.*

Proof. Solving $ax \equiv b \pmod{n}$ is equivalent to finding a solution to $ax - b = ny$ or $ax - ny = b$ where $x, y \in \mathbb{Z}$. If $g \not| b$ then this is impossible since g divides a and n. If $g|b$ then we solve $ax - ny = b$ using the Euclidean Algorithm to find a solution $x = x_0, y = y_0$. Then $x = x_t := x_0 + (\frac{n}{g})t, y = y_t := y_0 + (\frac{a}{g})t$ are solutions where $0 \leq t < g$ since $ax_t - ny_t = a(x_0 + (\frac{n}{g})t) - n(y_0 + (\frac{a}{g})t) = (ax_0 - ny_0) + at\frac{n}{g} - nt\frac{a}{g} = b$. Moreover these g solutions are not equivalent modulo n in the following sense: if $x_k \equiv x_l \pmod{n}$ where $0 \leq k, l < g$ then $n|(x_k - x_l)$. But $|x_k - x_l| = |(x_0 + (\frac{n}{g})k) - (x_0 + (\frac{n}{g})l)| = \frac{n}{g}|k - l|$. Since $0 \leq k, l < g$ we know $|k - l| < g$ and so $0 \leq |x_k - x_l| < \frac{n}{g}g = n$ and so $x_l = x_k$ since this is the only multiple of n in this range. □

Looking back at $6x \equiv 4 \pmod{10}$ the above theorem tells us that since $\gcd(6, 10) = 2$ does indeed divide 4 there are two solutions to this congruence up to equivalence. Let's use the Euclidean Algorithm to find one of these, namely a solution to $6x - 10y = 4$. Well we can divide both sides by 2 and solve $3x - 5y = 2$. $5 = 3(1) + 2$ and so $2 = 3(-1) - 5(-1)$ gives us $x_0 = -1 = y_0$. Then $x_1 = -1 + \frac{10}{2}1, y_1 = -1 + \frac{6}{2}(1)$ or $x_1 = 4, y_1 = 2$. The two solutions up to equivalence to $6x \equiv 4 \pmod{10}$ are $x_0 = -1 \equiv 9 \pmod{10}, x_1 = 4$

If we start to examine a *system* of linear congruences things get a bit more complicated. For example:

$$\begin{aligned} x &\equiv 3 \pmod{4} \\ x &\equiv 10 \pmod{15} \\ x &\equiv 5 \pmod{7} \end{aligned}$$

does have solutions $x \in \mathbb{Z}$ that satisfy all three congruences at the same time, but the smallest positive solution is $x = 355$. Thankfully under the right circumstances these systems can often be solved very efficiently due to a theorem first used in 3rd-century China and brought to Europe a thousand years later by Leonardo De Pisa (Fibonacci). The following theorem gives a sufficient condition for a system of linear congruences to have a solution, and its proof gives an algorithm for finding such a solution.

Theorem 4.50 (Chinese Remainder Theorem)**.** *Let $n_1, n_2, \ldots, n_k$ be a collection of pairwise relatively prime natural numbers and $b_1, b_2, \ldots, b_k \in \mathbb{Z}$. Then the system*

$$\begin{aligned} x &\equiv b_1 \pmod{n_1} \\ x &\equiv b_2 \pmod{n_2} \\ &\ldots. \\ x &\equiv b_k \pmod{n_k} \end{aligned}$$

has a solution x and moreover this solution is unique up to congruence modulo $N = n_1 n_2 .. n_k$.

Proof. Let $N = n_1 n_2 ... n_k$ and set $m_i := N/n_i$ for $i = 1, 2, \ldots, k$. Then $\gcd(m_i, n_j) = 1$ if $i = j$ and $\gcd(m_i, n_j) = m_i$ if $i \neq j$, in other words m_i is a unit modulo n_i. Let $[m_i]^{-1} = [y_i]$ in $\mathbb{Z}/n_i$ for $i = 1, 2, \ldots, k$. Set $x = b_1 y_1 m_1 + b_2 y_2 m_2 + \cdots + b_k y_k m_k$. We check that

$$x = b_1 y_1 m_1 + b_2 y_2 m_2 + \cdots + b_k y_k m_k \equiv b_i \pmod{n_i}$$

for each $1 \leq i \leq k$ since if $j \neq i$ then $m_j \equiv 0 \pmod{n_i}$ and $b_i y_i m_i \equiv b_i \pmod{n_i}$ since $[y_i][m_i] = [1]$ in $\mathbb{Z}/n_i$.

To show uniqueness up to congruence modulo N let x and y both be solutions. For each $1 \leq i \leq k$, $x - y$ is divisible by n_i since $x - y \equiv b_i - b_i \equiv 0 \pmod{n_i}$ and since the n_i are relatively prime we get that $x - y$ is divisible by N. □

Example 4.51. Let's revisit the linear system

$$\begin{aligned} x &\equiv 3 \pmod{4} \\ x &\equiv 10 \pmod{15} \\ x &\equiv 5 \pmod{7} \end{aligned}$$

where we see the moduli are relatively prime. In this case $n_1 = 4, n_2 = 15, n_3 = 7$ and so $N = 420$ and $m_1 = 105, m_2 = 28, m_3 = 60$. First we find y_1 by computing $[105]^{-1} = [1]^{-1} = [1]$ in $\mathbb{Z}/4$ to get $y_1 = 1$. We find y_2 by computing $[28]^{-1} = [13]^{-1} = [7]$ in $\mathbb{Z}/15$ to get $y_2 = 7$. Lastly we find y_3 by computing $[60]^{-1} = [4]^{-1} = [2]$ in $\mathbb{Z}/7$ to get $y_3 = 2$. We assemble our solution $x = b_1 y_1 m_1 + b_2 y_2 m_2 + b_3 y_3 m_3 = (3)(1)(105) + (10)(7)(28) + (5)(2)(60) = 2875 \equiv 355 \pmod{420}$.

We mentioned earlier that Fermat's Little Theorem's first published proof was by Euler who in fact generalized the theorem to arbitrary natural numbers; in fact the proof we gave for FLT will easily generalize with a minimum of modifications. We introduce Euler's totient or phi function for this purpose.

Definition 4.52. Let $n \geq 2$ be a natural number. Define $\phi(n)$ to be the size of the following set:

$$\{k \in \mathbb{N} \,|\, 1 \leq k < n, \gcd(k, n) = 1\},$$

in other words $\phi(n)$ is the size of the set $(\mathbb{Z}/n)^*$, the set of units modulo n. For completion we set $\phi(1) = 1$ thus defining a function $\phi : \mathbb{N}^+ \to \mathbb{N}^+$. This is called Euler's totient or phi function.

Example 4.53. Let p be a prime. Then we have already seen that every number in the set $\{1, 2, 3, \ldots, p-1\}$ is a unit modulo p, there are $p - 1$ elements of this set, and so $\phi(p) = p - 1$. By looking at the set $\{1, 2, \ldots, p, p+1, \ldots, 2p, \ldots p^k - 1\}$ where $k \in \mathbb{N}$ we see that the only nonunits are $p, 2p, \ldots (p-1)p$ and so $\phi(p^k) = (p^k - 1) - (p - 1) = p^k - p = p^{k-1}(p - 1)$. Furthermore we state here without proof the following nice property of ϕ: if $a, b \in \mathbb{N}^+$ satisfy $\gcd(a, b) = 1$ then $\phi(ab) = \phi(a)\phi(b)$ (the proof of this uses the Chinese Remainder Theorem). We can check by hand simple facts like $\phi(pq) = (p - 1)(q - 1)$ where p, q are distinct primes.

Theorem 4.54. *Let $a, b \in \mathbb{N}^+$ with $\gcd(a,b) = 1$. Then $\phi(ab) = \phi(a)\phi(b)$.*

Proof. Let $S_1 = \{n \,|\, 1 \leq n < ab\}$ and note that S_1 has $\phi(ab)$ elements by definition of ϕ. Let $S_2 = \{(k,l) \,|\, 1 \leq k < a, \gcd(a,k) = 1, 1 \leq l < b, \gcd(b,l) = 1\}$. Since there are $\phi(a)$ possible choices for k and $\phi(b)$ possible choices for l we see that S_2 has $\phi(a)\phi(b)$ elements. Define $f : S_1 \to S_2$ by $f(n) = (n \mod a, n \mod b)$. We'll show that f is a bijection. First if $f(n_1) = f(n_2)$ where $1 \leq n_1, n_2 < ab$ then

$$n_1 \equiv n_2 \pmod{a}$$

$$n_1 \equiv n_2 \pmod{b}$$

and so $a|n_1 - n_2$ and $b|n_1 - n_2$. But since $\gcd(a,b) = 1$ we know $ab|n_1 - n_2$ so $n_1 \equiv n_2 \pmod{ab}$. But the only way this can happen is if $n_1 = n_2$ and so f is $1-1$. To show f is onto let $(k,l) \in S_2$. We need to find n so that $n \equiv k \pmod{a}, n \equiv l \pmod{b}$. But such a solution always exists by the Chinese Remainder Theorem. So f is a bijection. We'll see later in Chapter 7 that bijections preserve size and so the sizes of S_1, S_2 are the same and hence $\phi(ab) = \phi(a)\phi(b)$. □

Theorem 4.55. *Let $n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ be the prime factorization of $n \geq 2$. Then*

$$\phi(n) = n \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right)$$

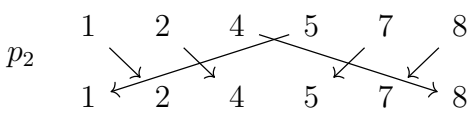
Proof. We have $\phi(n) = \phi(p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}) = \phi(p_1^{a_1})\phi(p_2^{a_2}) \dots \phi(p_k^{a_k})$. Dividing by n we get

$$\frac{\phi(n)}{n} = \frac{(p_1 - 1)(p_2 - 1)\dots(p_k - 1)p_1^{a_1 - 1} p_2^{a_2 - 1} \dots p_k^{a_k - 1}}{p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}} =$$

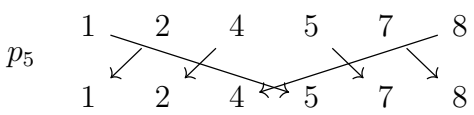
$$\left(\frac{p_1 - 1}{p_1}\right)\left(\frac{p_2 - 1}{p_2}\right) \dots \left(\frac{p_k - 1}{p_k}\right) = \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right).$$

□

Example 4.56. Let $n = 9$ so that the units modulo 9 are, up to equivalence, $(\mathbb{Z}/9)^* = \{1, 2, 4, 5, 7, 8\}$. Let $a = 2$ and look at the function $p_2 : (\mathbb{Z}/9)^* \to (\mathbb{Z}/9)^*$ defined by $p_2(x) = 2x \mod 9$; here we're being somewhat relaxed with notation, what we really mean is $p_2([x]) = [2x]$, but when there's no confusion we'll associate x with its equivalence class $[x]$. We draw an arrow diagram of this function:

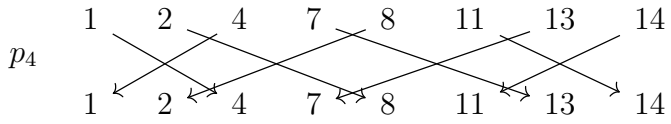


We see that this function is a bijection, i.e. a permutation of the set of units. Let $a = 5$ and let $p_5 : (\mathbb{Z}/9)^* \to (\mathbb{Z}/9)^*$ be defined by $p_2(x) = 5x \mod 9$. It has arrow diagram



which is clearly another permutation of $(\mathbb{Z}/9)^*$. We encourage the reader to draw diagrams for p_4, p_7, p_8 to see that they are indeed permutations.

Example 4.57. Let $n = 15$ so that $(\mathbb{Z}/15)^* = \{[1], [2], [4], [7], [8], [11], [13], [14]\}$. Let $a = 4$ so that $\gcd(4, 15) = 1$ and look at $p_4 : (\mathbb{Z}/15)^* \to (\mathbb{Z}/15)^*$ defined by $p_4(x) = 4x \pmod{15}$. It has arrow diagram



which is another permutation of $(\mathbb{Z}/15)^*$. We encourage the reader to draw p_7, p_8, p_{11} etc...

We'll see in the next theorem that if $\gcd(a, n) = 1$, so that a is a unit mod n, the function $p_a : (\mathbb{Z}/n)^* \to (\mathbb{Z}/n)^*$ defined by $p_a(x) = ax \pmod{n}$ is a permutation and gives us an extension of Fermat's Little Theorem:

Theorem 4.58 (Euler's Theorem)**.** *Let $n \in \mathbb{N}^+$ and $a \in \mathbb{Z}$ satisfying $\gcd(a, n) = 1$ (so that a is a unit modulo n). Then*

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

Proof. Examine the set $S = \{u_1, u_2, u_3, \ldots u_{\phi(n)}\}$ where each element of this set is a unit modulo n so that $1 \leq u_i < n$ and no two of these units are equal. Examine the set

$S' = aS := \{au_1, au_2, au_3, \ldots, au_{\phi(n)}\}$ where we multiply each of the $\phi(n)$ units in S by a. This list consists of units as well and we claim that no two members of this list are equivalent modulo n. If we had $au_k \equiv au_i \pmod{n}$ then by the cancellation property we would have $u_k \equiv u_l \pmod{n}$. So the list S' also consists of $\phi(n)$ distinct units modulo n. Since there are only $\phi(n)$ distinct units up to congruence modulo n, these lists contain the same members up to equivalence, perhaps in a different order. We multiply all the members of S together to get $u_1 \cdot u_2 \cdot \cdots \cdot u_{\phi(n)} := u$. We multiply all members of S' together to get $(au_1) \cdot (au_2) \cdot \cdots \cdot (au_{\phi(n)}) = a^{\phi(n)}u$. These products must be equivalent so we have

$$u \equiv ua^{\phi(n)}.$$

But u is a product of units and hence a unit, and so we can cancel it from both sides of the above equivalence to get

$$1 \equiv a^{\phi(n)} \pmod{n}.$$

□

Example 4.59. Suppose we want to find the last three digits of the number 13457^{89621}. First we note that $13457 \equiv 457 \pmod{1000}$. We also see that $\phi(1000) = 400$ so we know that since $\gcd(457, 1000) = 1$ that $457^{\phi(1000)} \equiv 1 \pmod{1000}$ or $457^{400} \equiv 1 \pmod{1000}$. Then $457^{89621} = (457^{400})^{224}(457^{21}) \equiv (1^{224})(457^{21}) = 457^{21} \pmod{1000}$. Now $457^2 \equiv 849 \equiv -151 \pmod{1000}, 457^4 \equiv (-151)^2 \equiv 801 \equiv -199 \pmod{1000}, 457^8 \equiv (-199)^2 \equiv 601 \equiv -399 \pmod{1000}$ and so $457^{16} \equiv (-399)^2 \equiv 201 \pmod{1000}$ giving us that:

$$457^{21} = (457^{16})(457^4)(457^1) \equiv (201)(801)(457) \equiv 457 \pmod{1000}.$$

Altogether we see that $12457^{89621} \equiv 457 \pmod{1000}$ and so the last three digits are 457. The number 12457^{89621} has 367036 digits.

5 Recursion and Strong Induction

The proof of the Fundamental Theorem of Arithmetic as mentioned earlier requires a form of induction a bit more flexible that our Principle of Mathematical Induction (first versions) from earlier. Sometimes this second version is called 'strong' induction but it is logically equivalent.

Theorem 5.1 (Principle of Mathematical Induction (version 2))**.** *Let P_n be a predicate statement that depends on $n \in \mathbb{N}$. Then to prove $\forall n \in \mathbb{N}\, P_n$ is suffices to establish:*

$$1. \quad P_0$$
$$2. \quad \forall k \in \mathbb{N} \wedge (0 \leq j \leq k), P_j \Longrightarrow P_{n+1}$$

Condition 2 can also be written as $\forall k \in \mathbb{N}$, $(P_0 \wedge P_1 \wedge \cdots \wedge P_k) \Longrightarrow P_{k+1}$. Notice we are allowed to assume not just P_k as inductive hypothesis but all previous statements making our inductive hypothesis stronger. If we want an even more general version that allows us to prove $\forall n \in \mathbb{Z}, n \geq a, P_n$ we can do so by first establishing P_a and then showing that for any $k \in \mathbb{Z}$ so that $a \leq k$ we have that $(P_a \wedge P_{a+1} \wedge \cdots \wedge P_k) \Longrightarrow P_{k+1}$.

As our first of many examples to come we prove the existence part of the Fundamental Theorem of Arithmetic at last.

Theorem 5.2 (Proof of Existence, FTA)**.** *Let n be a natural number greater than* 1*. Then n can be written as the product of primes.*

Proof. Let P_n be the statement that n can be written as a product of primes. Our base case is P_2, and we check that 2 can be written as a product of primes. Well 2 *is* prime and so itself is the product of a single prime. Assume that P_j is true for all $2 \leq j \leq k$ where k is an integer greater than or equal to 2. We look at $k+1$. Either $k+1$ is prime, in which case it is a product of primes, or it is of the form $k+1 = ab$ where a, b are natural numbers greater than 2 and hence less than k (why?). Well P_a tells us that a can be written as a product of primes and P_b tells us that b can be written as a product of primes, and so $ab = k+1$ is a product of a product of primes...and so is a product of primes itself. So P_{k+1} holds and we've established the existence portion of the FTA. □

Why did we need this second form of induction in the above proof, why wouldn't our first version of the PMI work here? The factors of $n+1$ generally have no relation to the factors of n: think of $n = 99$ which has prime factorization $(3^2)(11)$ and $n+1 = 100$ which has prime factorization $(2^2)(5^2)$.

Example 5.3. We use this second form of induction to prove the following statement (you should try a few examples for yourself): $\forall n \in \mathbb{N}^+$, n can be written as a sum of *distinct* powers of 2, i.e., $n = 2^{p_1} + 2^{p_2} + \cdots + 2^{p_t}$ where $0 \leq p_1 < p_2 < \cdots < p_t$. Here is the proof: we proceed by induction on $n \in \mathbb{N}^+$, the base case being that 1 can be written as a sum of distinct powers of 2; this is easy $1 = 2^0$. Assume that for some $k \in \mathbb{N}^+$ that each $1 \leq j \leq k$ can be written as a sum of distinct powers of 2 and look at $k+1$. If $k+1$ is even then $1 \leq \frac{k+1}{2} < k+1$ is

a natural number and so $\frac{k+1}{2} = 2^{t_1} + 2^{t_2} + \cdots + 2^{t_m}$ where $0 \leq t_1 < t_2 < \cdots < t_m$. But then $k+1 = 2(2^{t_1} + 2^{t_2} + \cdots + 2^{t_m}) = 2^{t_1+1} + 2^{t_2+1} + \cdots + 2^{t_m+1}$ is a sum of distinct powers of 2 since $1 \leq t_1 + 1 < t_2 + 1 < \cdots < t_m + 1$. Otherwise $k+1$ is odd and so k is even. Since $k < k+1$ we know that k can be written as a distinct sum of powers of 2: $k = 2^{a_1} + 2^{a_2} + \cdots + 2^{a_l}$ where $0 < a_1 < a_2 < \cdots < a_l$ where here we can specify that $a_1 \neq 0$ since k is even. But then $k+1 = 1 + 2^{a_1} + 2^{a_2} + \cdots + 2^{a_l} = 2^0 + 2^{a_1} + 2^{a_2} + \cdots + 2^{a_l}$ is a way of writing $k+1$ as a sum of distinct powers of 2. In either case we are done and so the statement holds $\forall n \in \mathbb{N}^+$.

We can also define sequences using a process very much related to the PMI, the concept of a recursive sequence which we will articulate with a few examples.

Example 5.4. Let $x_0 = 0$ and let $x_{n+1} = x_n + (2n+1)$ for $n \geq 0$. Let's compute the first few terms of this sequence: $x_1 = x_0 + 2(0) + 1 = 1, x_2 = x_1 + 2(1) + 1 = 4, x_3 = x_2 + 2(2) + 1 = 9, x_4 = x_2 + 2(4) + 1 = 16$. Can you guess a formula for x_n that only depends on n?

The sequence of numbers in the last example is an example of a *recursive* sequence: future terms of the sequence are defined using previously defined terms. Here's another example:

Collatz Conjecture: Let $x_0 = k \in \mathbb{N}^+$. Define

$$x_{n+1} = \begin{cases} \frac{x_n}{2} & x_n \text{ even} \\ 3x_n + 1 & x_n \text{ odd} \end{cases}$$

Say we start with $x_0 = 20$. Then $x_1 = 10, x_2 = 5, x_3 = 16, x_4 = 8, x_5 = 4, x_6 = 2, x_7 = 1, x_8 = 4, x_9 = 2, \ldots$ where we enter a cycle $4 \to 2 \to 1 \to 4$. Let's try starting value $x_0 = 56$. Then the sequence $x_0, x_1, x_2, \ldots$ is given by $56 \to 28 \to 14 \to 7 \to 22 \to 11 \to 34 \to 17 \to 52 \to 26 \to 13 \to 40 \to 20 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1 \to 4 \to 2 \to 1 \to 4 \ldots$ and we are right back at that same length three cycle. We encourage the reader to choose a few different x_0 and to make a conjecture. Currently it is not known whether all starting values x_0 lead to this cycle, this is an open conjecture know as the Collatz Conjecture or the '$3x+1$ problem'.

There are many common objects in mathematics that are defined recursively or have natural recursive definitions, we note a few of these now.

Definition 5.5. We define $0! := 1$ and for $k \in \mathbb{N}$ we define $(k+1)! := (k+1)k!$. This defines the sequence of *factorials* $1, 1, 2, 6, 24, 120, \ldots$

Definition 5.6. Let $(a_n)_{n=0}^{\infty}$ be an infinite sequence of real numbers, i.e. $\forall i \in \mathbb{N}, a_i \in \mathbb{R}$. We define $\sum_{k=0}^{0} a_k := a_0$ and define $\sum_{k=0}^{n+1} a_k := (a_{n+1}) + \sum_{k=0}^{n} a_k$. If our sequence starts at some $t \in \mathbb{Z}$, namely that our sequence is $a_t, a_{t+1}, a_{t+2}, \ldots$ we easily define $\sum_{k=t}^{n} a_k$ for $n \geq t$ by setting $\sum_{k=t}^{t} a_k := a_t$ and $\sum_{k=t}^{n+1} a_k := (a_{n+1}) + \sum_{k=t}^{n} a_k$.

The Division Algorithm is a special case of a more general theorem involving polynomials. Let

$$\mathbb{R}[x] = \{a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n : a_i \in \mathbb{R}, n \in \mathbb{N}\}$$

be the set of all polynomials in x with integer coefficients. Given some $p(x) \in \mathbb{R}[x]$ where $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ with $a_n \neq 0$ we write $deg(p) = n$ and call this the degree of p. We call a_n the *leading coefficient* of p and a_nx^n the *leading term* of p. We leave undefined the degree of the identically zero polynomial $0(x) = 0$. We define addition, subtraction, and multiplication of polynomials as follows:

Definition 5.7. Let $p(x) = \sum_{i=0}^{n} a_ix^i, q(x) = \sum_{j=0}^{m} b_jx^j$. Then

$$(p \pm q)(x) = \sum_{i=0}^{\max(n.m)} (a_i \pm b_i)x^i,$$

and

$$(pq)(x) = \sum_{i=0}^{n+m} c_ix^i,$$

where $c_i = \sum_{j=0}^{i} a_jb_{i-j} = a_0b_i + a_1b_{i-1} + \cdots + a_{i-1}b_1 + a_ib_0$.

Lemma 5.8. *Let $p, q \in \mathbb{R}[x]$ be non-zero. Then $deg(p \pm q) \leq \max\{deg(p), deg(q)\}$ and $deg(pq) = deg(p)deg(q)$.*

We leave the proof of this lemma to the reader. We are now ready to state and prove our extension of the Division Algorithm.

Theorem 5.9 (Division Algorithm for Polynomials)**.** *If $p, d \in \mathbb{R}[x]$ have $deg(p) = n, deg(d) = m$, $n \geq m$, and $d \neq 0$, then $\exists q, r \in \mathbb{R}[x]$ so that*

$$p(x) = d(x)q(x) + r(x)$$

where $deg(q) = n - m$ and either $deg(r) < m$ or $r = 0$. Moreover q, r satisfying these conditions are unique.

Proof. Fix m and proceed by induction on $n \geq 0$. The base case is when $n = 0$ and since $n \geq m$ we must have $m = 0$ as well. If we let $q = d$ and $r = 0$ we see that $p = dq + r$ holds. Assume the theorem is true for all $k \leq n$ some $n \geq 0$ and let $p \in \mathbb{R}[x]$ have $deg(p) = n+1$. Let $a_{n+1} \neq 0$ and $b_m \neq 0$ be the leading coefficients of p and d respectively so that $p(x) = a_0 + \cdots + a_{n+1}x^{n+1}$ and $d(x) = b_0 + \cdots + b_mx^m$. Define $g(x) = p(x) - \frac{a_{n+1}}{b_m}x^{n+1-m}d(x)$. Then

$$g(x) = (a_0 + \cdots + a_{n+1}x^{n+1}) - \frac{a_{n+1}}{b_m}x^{n+1-m}(b_0 + \cdots + b_mx^m) =$$

$$a_0 + \cdots + a_nx^n + a_{n+1}x^{n+1} - \frac{b_0a_{n+1}}{b_m}x^{n+1-m} - \cdots - a_{n+1}x^{n+1}$$

has degree less than $n + 1$ since the leading term of p gets cancelled out. Suppose first that $deg(g) \geq m$. Then we can apply our inductive hypothesis to g and d to get that there are $a, r \in \mathbb{R}[x]$ so that $g(x) = a(x)d(x) + r(x)$ with $deg(a) = deg(g) - m$ and either $deg(r) < m$ or

$r = 0$. But then $p(x) - c(x)d(x) = a(x)d(x) + r(x)$ where $c(x) = -\frac{a_{n+1}}{b_m}x^{n+1-m}$. But this means that

$$p(x) = (a(x) + c(x))d(x) + r(x)$$

where $deg(a+c) = n+1-m$ and $deg(r) < m$ or $r = 0$, and we are done. If we have $deg(g) < m$ we can just write $p(x) = c(x)d(x) + g(x)$ and note that $deg(c) = n + 1 - m$ and $deg(g) < m$. This completes our induction.

We now prove that q, r satisfying the conditions in the theorem are uniquely determined. Suppose that $p = qd+r = q'd+r'$ where r, r' have degree less than d or are the zero polynomial. Then $(qd + r) - (q'd + r') = 0$ and so $(q - q')d + (r - r') = 0$. Assume that $q - q' \neq 0$. Then $deg((q - q')d) \geq deg(d)$ by the preceding lemma, and either $deg(r - r') < deg(d)$ or $r - r' = 0$. But this cannot be the case since the sum of two polynomials of different degrees cannot be the zero polynomial. So $q - q' = 0$ which in turn implies $q = q'$ and hence $r = r'$. □

Corollary 5.9.1. *Let $p \in \mathbb{R}[x]$ and $a \in \mathbb{R}$. Then there exists $q \in \mathbb{R}[x]$ so that $p(x) = (x - a)q(x) + p(a)$.*

Proof. Apply the division algorithm to p and $d(x) = x - a$ to get q, r so that $p = qd + r$ and $deg(r) < deg(d) = 1$ or $r = 0$. So r is constant and it's easy to see that $r = p(a)$. □

Example 5.10. Let us prove that $\sum_{k=0}^{n} 2k + 1 = (n + 1)^2$ by using the PMI where $n \in \mathbb{N}$. We check that $1 = 2(0) + 1 = \sum_{0}^{0} 2k + 1 = (0 + 1)^2 = 1$ and so the base case holds. Assume that $\sum_{k=0}^{n} 2k + 1 = (n + 1)^2$ for some $n \in \mathbb{N}$. Then

$$\sum_{k=0}^{n+1} 2k + 1 = \sum_{k=0}^{n} 2k + 1 + 2(n + 1) + 1 = (n + 1)^2 + 2(n + 1) + 2 = (n + 2)^2,$$

and so we are done by the PMI. Note how natural it was to prove this statement by induction.

Example 5.11 (Sum Induction Proofs)**.** Proving a formula of the form $\sum_{i=1}^{n} f(i) = g(n)$ for some functions f, g by induction follows, for the most part, a template. One first needs to check the base case which involves checking that $\sum_{i=1}^{1} f(i) = f(1)$ equals $g(1)$. Once one has assumed that $\sum_{i=1}^{n} f(i) = g(n)$ holds for some $n \geq 1$ one looks at $\sum_{i=1}^{n+1} f(i)$.

1. Step 1: evaluate $i = n + 1$ to get $\sum_{i=1}^{n+1} f(i) = f(n + 1) + \sum_{i=1}^{n} f(i)$.
2. Step 2: substitute the inductive hypothesis to get $f(n+1)+\sum_{i=1}^{n} f(i) = f(n+1)+g(n)$.
3. Step 3: find a way to manipulate $f(n + 1) + g(n)$ so that it equals $g(n + 1)$.

Hence through steps $1 - 3$ one gets $\sum_{i=1}^{n+1} f(i) = g(n + 1)$ which completes the induction.

Theorem 5.12 (Difference Theorem)**.** *Let $a, b \in \mathbb{R}$ and $n \in \mathbb{N}^+$. Then*

$$a^n - b^n = (a - b)\sum_{j=0}^{n-1} a^j b^{n-1-j}.$$

Proof. Suppose that $a \neq b$ so that $(a-b) \neq 0$ (if $a-b$ the statement is obvious). We prove the statement $\frac{a^n-b^n}{a-b} = \sum_{j=0}^{n-1} a^j b^{n-1-j}$ is true for all $n \geq 1$ by induction on $n \geq 1$. We check the base case when $n = 1$ and see $\frac{a^1-b^1}{a-b} = 1$ and $\sum_{j=0}^{0} a^j b^{-1-j} = 1$ which are equal. Assume that $\frac{a^n-b^n}{(a-b)} = \sum_{j=0}^{n-1} a^j b^{n-1-j}$ for some $n \geq 1$. Then

$$\sum_{j=0}^{n} a^j b^{(n+1)-1-j} = a^n + \sum_{j=0}^{n-1} a^j b^{n-j} = a^n + b\sum_{j=0}^{n-1} a^j b^{n-1-j} =$$

$$a^n + b\frac{a^n-b^n}{a-b} = \frac{a^n(a-b)}{a-b} + \frac{ba^n - b^{n+1}}{a-b} = \frac{a^{n+1}-b^{n+1}}{a-b}.$$

and so we are done by the PMI □

Most students recognize a few summation formulas from when they see Riemann sums in first semester calculus (though these are often forgotten). Some examples are:

$$\sum_{k=1}^{n} 1 = n$$

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4} = \left(\sum_{k=1}^{n} k\right)^2$$

Each of these can be proven in an manner completely analogous to our proof that $\sum_{k=0}^{n} 2k+1 = (n+1)^2$. But where do you get these formulas in the first place? They can be obtained a few different ways, the following is due to the Bernoulli brothers, a pair of Swiss mathematicians, about 300 years ago. Let $S_p := \sum_{k=1}^{n} k^p$, we wish to see where the above formulas for S_0, S_1, S_2, S_3 come from and see a general way to find formulas for $p = 4, 5, 6 \ldots$ as well. The formula for S_0 is obvious, the formula for S_1 is also pretty easy to get:

$$S_1 = 1 + 2 + 3 + \cdots + nS_1 = n + (n-1) + \cdots + 2 + 1$$

Adding these together we get $2S_1 = n(n+1)$ and so $S_1 = \frac{n(n+1)}{2}$. To get S_2 we proceed to look at the following sum, which might appear a bit strange at first: $\sum_{k=1}^{n} (k+1)^3 - k^3$. On the one hand

$$\sum_{k=1}^{n} (k+1)^3 - k^3 = \sum_{k=1}^{n} 3k^2 + 3k + 1 =$$

$$3\sum_{k=1}^{n} k^2 + 3\sum_{k=1}^{n} k + \sum_{k=1}^{n} = 3S_2 + 3S_1 + S_0$$

But we can also see this is a telescoping sum in the sense that $\sum_{k=1}^{n} (k+1)^3 - k^3 = (2^3 - 1^3) + (3^3 - 2^3) + \cdots + ((n+1)^3 - n^3) = (n+1)^3 - 1 = n^3 + 3n^2 + 3n$. But then

$$n^3 + 3n^2 + 3n = 3S_2 + 3S_1 + S_0 = 3S_2 + 3\frac{n(n+1)}{2} + n$$

enables us to solve for S_2:

$$6S_2 = 2(n^3 + 3n^2 + 3n) - n(n+1) - 2n = 2n^3 + 5n^2 + 3n =$$
$$n(2n^2 + 5n + 3) = n(n+1)(2n+3)$$

and our formula for S_2 above follows. Now that we have this nice formula for S_2 to find a nice formula for S_3 we look at the sum $\sum_{k=1}^{n}(k+1)^4 - k^4$ and evaluate it in the above ways and solve for S_3. Continuing, once the formulas up to S_p are found for $p \in \mathbb{N}^+$ to find the formula for S_{p+1} one looks at the sum $\sum_{k=1}^{n}(k+1)^{p+2} - k^{p+2}$ in the same manner.

We encourage the reader to try to find some of these, it's great practice getting more comfortable with sigma notation and sums in general. One obstacle as one proceeds to look at larger and larger powers p in the above, one that the Bernoulli's were keenly aware of, is that a bit more elbow grease seems to be needed in expanding $(k+1)^p$ as p gets larger. No one wants to see an expansion of $(k+1)^{12}$ by repeatedly multiplying by the term $k+1$. Thankfully there is nice streamlined way to expand such expressions that is also a product, pun intended, of recursive definitions. We first write out some of these expansions to get some ideas:

$$(1+k)^0 = 1$$
$$(1+k)^1 = 1 + k$$
$$(1+k)^2 = 1 + 2k + k^2$$
$$(1+k)^3 = 1 + 3k + 3k^3 + k^3$$
$$(1+k)^4 = 1 + 4k + 6k^2 + 4k^3 + k^4$$
$$(1+k)^5 = 1 + 5k + 10k^2 + 10k^3 + 5k^4 + k^5$$

This isn't anarchy, there are some strong patterns here where we've carefully arranged the resulting polynomials from lowest degree to highest degree. One pattern is that the coefficients as we move from left to right have a palindromic symmetry: they read the same from left to right as they do right to left.

We can also just look at the coefficients:

$n=0$				1			
$n=1$			1	1			
$n=2$		1	2	1			
$n=3$	1	3	3	1			
$n=4$	1	4	6	4	1		
$n=5$	1	5	10	10	5	1	
$n=6$	1	6	15	20	15	6	1
	0	1	2	3	4	5	6

Wihch is known as **Pascal's Triangle** with the entries called *binomial coefficients*. We give another definition of these numbers and then show that it coincides.

Definition 5.13. Let $n \in \mathbb{N}$ and let $0 \leq k \leq n$. We define $\binom{n}{k} := \frac{n!}{(k!(n-k)!} \in \mathbb{Q}$ and read 'n choose k.

It's not immediately clear that this rational number is actually a positive integer, we'll prove that shortly. But a few things are readily apparent from this definition.

Lemma 5.14. *Let $n \in \mathbb{N}$ and let $0 \leq k \leq n$. Then*

$$\binom{n}{k} = \binom{n}{n-k}$$

and

$$\binom{n}{0} = \binom{n}{n} = 1.$$

Theorem 5.15 (Pascal's Identity)**.** *Let $n \in \mathbb{N}$ and let $1 \leq k \leq n$. Then*

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

Proof. This is just a calculation; $(n+1)! = (n+1)n! = n!(k+n-k+1) = n!k + n!(n-k+1)$. If we divide through by $k!$ we get

$$\frac{(n+1)!}{k!} = \frac{n!}{(k-1)!} + \frac{n!(n-k+1)}{k!},$$

and dividing this through by $(n-k+1)!$ gives us

$$\frac{(n+1)!}{k!(n+1-k)!} = \frac{n!}{(k-1)!(n-(k-1))!} + \frac{n!}{k!(n-k)!},$$

or

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

□

Theorem 5.16. *Let $n \in \mathbb{N}$ and let $0 \leq k \leq n$. Then $\binom{n}{k} \in \mathbb{N}$, namely $k!(n-k)!|n!$*

Proof: We proceed by induction on $n \in \mathbb{N}$ to show that the statement $k!(n-k)!|n!$ holds where $0 \leq k \leq n$. The base case is when $n = k = 0$ and this holds by the above lemma. Now assume that it holds for some fixed but arbitrary $n \in \mathbb{N}$. If $k = 0, n+1$ then clearly $k!(n+1-k)!|(n+1)!$ so assume $1 \leq k \leq n$. Then by assumption $\binom{n}{k}, \binom{n}{k-1} \in \mathbb{N}^+$ so $\exists s, t \in \mathbb{Z}$ so that $sk!(n-k)! = n!, t(k-1)!(n-k+1)! = n!$. So

$$(n+1)! = n!(n-k+1+k) = n!(n-k+1) + n!k =$$
$$sk!(n-k)!(n-k+1) + t(k-1)!(n-k+1)!k =$$
$$(s+t)[k!(n+1-k)!],$$

and we are done by the PMI.

We next show that the binomial coefficients are indeed the coefficients of $(1+k)^n$. For the sake of simplicity we define $r^0 := 1$ for any $r \in \mathbb{R}$.

Theorem 5.17 (Binomial Theorem)**.** *Let $n \in \mathbb{N}$ and $x \in \mathbb{R}$. Then*

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k.$$

Proof. We proceed by induction on $t \in \mathbb{N}$, the base case of the theorem being that $1 = (1+x)^0 = \sum_{k=0}^{0} \binom{0}{k} = \binom{0}{0}$ which is true. Assume that $(1+x)^t = \sum_{k=0}^{t} \binom{t}{k} x^k$ for some $t \in \mathbb{N}$. Then

$$\sum_{k=0}^{t+1} \binom{t+1}{k} x^k = 1 + x^{t+1} + \sum_{k=1}^{t} \binom{t+1}{k} x^k$$

$$1 + x^{t+1} + \sum_{k=1}^{t} \binom{t}{k} x^k + \sum_{k=1}^{t} \binom{t}{k-1} x^k =$$

$$1 + \sum_{k=1}^{t} \binom{t}{k} x^k + x^{t+1} + \sum_{k=0}^{t-1} \binom{t}{k} x^{k+1} =$$

$$\sum_{k=0}^{t} \binom{t}{k} x^k + \sum_{k=0}^{t} \binom{t}{k} x^{k+1} = \sum_{k=0}^{t} \binom{t}{k} x^k + x \sum_{k=0}^{t} \binom{t}{k} x^k =$$

$$(1+x)^t + x(1+x)^t = (1+x)(1+x)^{t+1} = (1+x)^{t+1},$$

and we are done by the PMI. □

Corollary 5.17.1. *$2^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$ for any $n \in \mathbb{N}$.*

Proof. Let $x = 1$ in the Binomial Theorem. □

Corollary 5.17.2. *$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0$ for any $n \in \mathbb{N}$.*

Proof. Let $x = -1$ in the Binomial Theorem. □

Corollary 5.17.3. *Let $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$. Then*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Proof. If $b = 0$ this is clearly true, so assume $b \neq 0$ and set $x = \frac{a}{b}$. Then by the Binomial Theorem we know that

$$\left(1 + \frac{a}{b}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{a}{b}\right)^k,$$

and multiplying both sides by b^n we get

$$b^n \left(1 + \frac{a}{b}\right)^n = b^n \sum_{k=0}^{n} \binom{n}{k} \frac{a^k}{b^k},$$

or

$$(b + a)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

as desired. □

Theorem 5.18. *Let f, g be n times differentiable functions. Then*

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}$$

where $f^{(0)} = f$.

Proof. This is essentially the same proof as the Binomial Theorem. One proceeds by induction on $n \geq 1$, the base case being straightforward. If the theorem holds for some $n \geq 1$ and f, g are $n + 1$ times differentiable, then

$$(fg)^{(n+1)} = [(fg)^{(n)}]' = \left[\sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}\right]' = \sum_{k=0}^{n} \binom{n}{k} [f^{(k)} g^{(n-k)}]' =$$

$$\sum_{k=0}^{n} \binom{n}{k} f^{(k+1)} g^{(n-k)} + f^{(k)} g^{(n-k+1)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k+1)} g^{(n-k)} + \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k+1)}.$$

We note that

$$\sum_{k=0}^{n} \binom{n}{k} f^{(k+1)} g^{(n-k)} = f^{(n+1)} g + \sum_{k=0}^{n-1} \binom{n}{k} f^{(k+1)} g^{(n-k)} = f^{(n+1)} g + \sum_{k=1}^{n} \binom{n}{k-1} f^{(k)} g^{(n-k+1)}$$

and

$$\sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k+1)} = f g^{(n+1)} + \sum_{k=1}^{n} \binom{n}{k} f^{(k)} g^{(n-k+1)},$$

and so altogether we have

$$(fg)^{(n+1)} = f g^{(n+1)} + f^{(n+1)} g + \sum_{k=1}^{n} \binom{n}{k-1} f^{(k)} g^{(n-k+1)} + \sum_{k=1}^{n} \binom{n}{k} f^{(k)} g^{(n-k+1)} =$$

$$fg^{(n+1)} + f^{(n+1)}g + \sum_{k=1}^{n} \binom{n+1}{k} f^{(k)} g^{(n-k+1)}$$

where we combined sums using Pascal's identity. This is in turn equal to

$$\sum_{k=0}^{n+1} \binom{n+1}{k} f^{(k)} g^{(n-k+1)}.$$

which establishes the $n+1$ case, and so we are done by the PMI. □

Example 5.19. If $n = 4$ and f, g are 4 times differentiable, we have

$$(fg)^{(4)} = fg^{(4)} + 4f'g^{(3)} + 6f''g'' + 4f^{(3)}g' + f^{(4)}g.$$

The Binomial Theorem can be used to give a different proof of Fermat's Little Theorem and we ask the reader to supply a proof of the following:

Theorem 5.20. *Let p be a prime and $0 \leq k \leq p$. Then $\binom{p}{k} \equiv 0 \pmod{p}$ iff $0 < k < p$.*

Corollary 5.20.1 (Fermat's Little Theorem)**.** *Let p be a prime and $n \in \mathbb{Z}$. Then $n^p \equiv n \pmod{p}$.*

Proof. We first use induction on n to prove that $n^p \equiv n \pmod{p}$ where $n \in \mathbb{N}$, the base case being clear. Assume that $n^p \equiv n \pmod{p}$ for some $n \in \mathbb{N}$. Then

$$(n+1)^p = \sum_{k=0}^{p} \binom{p}{k} n^k \equiv n^p + 1 \equiv n + 1 \pmod{p},$$

where the first equality is from the binomial theorem, the second equivalence is from our theorem on binomial coefficients modulo p, and the last is our inductive hypothesis. So we are done by the PMI. □

Example 5.21 (Arithmetic-Geometric Mean Inequality)**.** Given two real numbers a, b we can form their *arithmetic mean* as $\frac{a+b}{2}$. If $a, b \geq 0$ we can also form their *geometric mean* as $\sqrt{ab}$. Expanding the inequality $(\sqrt{a} - \sqrt{b})^2 \geq 0$ we get $a + b - 2\sqrt{ab} \geq 0$ or that $\frac{a+b}{2} \geq \sqrt{ab}$, in other words, the geometric mean never exceeds the arithmetic mean. It's also straightforward to see that these means are equal exactly when $a = b$. We define the arithmetic mean of n real numbers $a_1, a_2, \ldots, a_n$ as $\frac{a_1+a_2+\cdots+a_n}{n}$ and if each $a_i \geq 0$ we define the geometric mean to be $\sqrt[n]{a_1 a_2 \ldots a_n}$.

Theorem 5.22 (AM-GM Inequality)**.** *Suppose that $a_1, a_2, \ldots, a_n$ are n non-negative real numbers. Then*

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \ldots a_n}.$$

We prove this by means of two lemmas:

Lemma 5.23. *Suppose $a, b \in \mathbb{R}$ with $a < 1$ and $b > 1$. Then $a + b \geq ab + 1$.*

Proof. We multiply the inequality $b > 1$ by the positive real number $1-a$ to get $(1-a)b > (1-a)$ or $b - ab > 1 - a$ which is exactly what we want. □

Lemma 5.24. *Suppose $c_1, c_2, \ldots, c_n$ are positive real numbers so that $c_1 c_2 \ldots c_n = 1$. Then $c_1 + c_2 + \cdots + c_n \geq n$, and equality holds precisely when $c_1 = c_2 = \cdots = c_n = 1$.*

Proof. We proceed by strong induction on $n \geq 2$, where n is the number of positive real numbers in our product. The base case $n = 2$ is covered by the first lemma. Suppose that we know the lemma holds for all natural numbers k up to and including n where $n \geq 2$ let $c_1, c_2, \ldots, c_n, c_{n+1}$ be $n+1$ numbers so that $c_1 c_2 \ldots c_{n+1} = 1$. Unless each $c_i = 1$ we must have at least one $c_j > 1$ and at least one $c_i < 1$, and we can re-index so that $c_n < 1$ and $c_{n+1} > 1$. Then we have $1 = c_1 c_2 \ldots c_{n+1} = (c_1 c_2 \ldots c_{n-1})(c_n c_{n+1})$ and we look at

$$c_1 c_2 \ldots c_{n-1} + c_n c_{n+1} \geq (n-1) + 2 = n + 1,$$

where we've applied our inductive hypothesis to $k = 2, n-1$. But we know

$$c_1 + c_2 + \ldots c_n + c_{n+1} = (c_1 + c_2 + \ldots c_{n-1}) + (c_n + c_{n+1}) \geq$$
$$c_1 + c_2 + \ldots c_{n-1} + c_n c_{n+1} + 1 \geq (n+1) + 1 = n + 2 > n + 1,$$

where we used the first lemma applied to c_n, c_{n+1}. Thus the second lemma holds by the PMI. □

We now give a proof of the AM-GM inequality using the lemmas:

Proof of AM-GM inequality. Let $a_1, a_2, \ldots, a_n > 0$ and set $g = \sqrt[n]{a_1 a_2 \ldots a_n}$. Define $x_i = \frac{a_i}{g}$. Then $x_1 x_2 \ldots x_n = 1$ and so we can apply the second lemma to get $x_1 + x_2 + \cdots + x_n \geq n$ or

$$\frac{a_1}{\sqrt[n]{a_1 a_2 \ldots a_n}} + \frac{a_2}{\sqrt[n]{a_1 a_2 \ldots a_n}} + \cdots + \frac{a_n}{\sqrt[n]{a_1 a_2 \ldots a_n}} \geq n,$$

which immediately gives us the AM-GM inequality. □

We return to number theory with a discussion of the *arithmetic derivative*, which we first consider as the function $D : \mathbb{Z} \to \mathbb{Z}$ defined by the properties:

1. $D(0) = 0$
2. $D(p) = 1$ for any prime p
3. $D(ab) = aD(b) + D(a)b$ for any $a, b \in \mathbb{N}$

4. $D(-a) = -D(a)$ for any $a \in \mathbb{N}$

Since any natural number $n \geq 2$ can be written uniquely as a product of primes we'll see that the above conditions allow us to compute $D(n)$ fairly easily once we know how to factor n. We also see that $D(1) = D(1 \cdots 1) = 1D(1) + D(1)1 = 2D(1)$ and so $D(1) = 0$. Property 4 allows us to compute the arithmetic derivative of any integer if we know how to compute the arithmetic derivative of any natural number. To find $D(125) = D(5^3)$ we compute:

$$D(25 \cdot 5) = 25D(5) + D(25)5 = 25 + 5D(5 \cdot 5) = 25 + 5[5D(5) + D(5)5]$$

which gives us $D(125) = 75$ or $D(5^3) = 3 \cdot 5^2$, which looks a lot like the standard power rule of regular derivatives. It will be convenient for us to write a' for $D(a)$ just like we do for regular derivatives. We now show how to compute $D(a)$ when $a = p_1^{t_1} p_2^{t_2} \dots p_k^{t_k}$.

Theorem 5.25. *Let $a \geq 2$ have prime factorization $a = p_1^{t_1} p_2^{t_2} \dots p_k^{t_k}$. Then*

$$a' = a \sum_{i=1}^{k} \frac{t_i}{p_i}.$$

Proof. We prove this by induction on the number m of not necessarily distinct prime factors of a, the base case is when $m = 1$ so that $a = p$ for some prime p. Then $a' = 1 = a \sum_{i=1}^{1} \frac{1}{p}$ and so the base case holds. Assume we know the formula works for all natural numbers greater than one with m prime factors and let $a = bp_{m+1}$ have $m + 1$ prime factors. So b has m prime factors, say $b = p_1 p_2 \dots p_m$, and we see

$$a' = (bp_{m+1})' = bp'_{m+1} + p_{m+1}b' = b + p_{m+1}b \sum_{i=1}^{m} \frac{1}{p_i} =$$

$$\frac{a}{p_{m+1}} + a \sum_{i=1}^{m} \frac{1}{p_i} = a \sum_{i=1}^{m+1} \frac{1}{p_i}.$$

So our induction is complete. It's straightforward now to see that if $a = p_1^{t_1} p_2^{t_2} \dots p_k^{t_k}$ that $a' = a \sum_{i=1}^{k} \frac{t_i}{p_i}$. □

Corollary 5.25.1. *We have $(a^n)' = na^{n-1}a'$ for any natural numbers a and n.*

We leave the proof of this as an exercise for the reader. We make a note that even though we have an analogue of the product rule for the arithmetic derivative, it is not linear, in general we do not have $(a + b)' = a' + b'$. Take for instance $1 = 2'$; we can write $2 = 1 + 1$ and it does not follow that $1 = 2' = 1' + 1' = 0 + 0 = 0$. We leave it as an exercise to the reader to show that $a' = 1$ iff a is a prime number. Let $c \in \mathbb{Z}$. We ask the interested reader to solve the 'differential equation $a' = c$. When $c = 1$ we must have a being a prime number, what can you say about a if $a' = 2$? What if $a' = 6$? It's interesting to see how a' grows as a grows. We have the following bounds:

Theorem 5.26. *For every positive integer a we have*

$$a' \leq \frac{a \log_2(a)}{2},$$

and if a is composite then $a' \geq 2\sqrt{a}$. Furthermore if a is a product of k factors then $a' \geq ka^{\frac{k-1}{k}}$.

Proof. If $a = p_1^{t_1} p_2^{t_2} \dots p_m^{t_m}$ then

$$a \geq 2^{t_1} 2^{t_2} \dots 2^{t_m} = 2^{t_1+t_2+\dots+t_m}$$

implies that $\log_2(a) \geq t_1 + t_2 + \dots + t_m$. Since $a' = a \sum_{i=1}^{m} \frac{t_i}{p_i}$ we have

$$a' = a \sum_{i=1}^{m} \frac{t_i}{p_i} \leq \frac{a \sum_{i=1}^{m} t_i}{2} \leq \frac{a \log_2(a)}{2},$$

which proves the first part. If $a = a_1 a_2 \dots a_k$ then we know

$$a' = a_1' a_2 \dots a_k + a_1 a_2' \dots a_k + \dots + a_1 a_2 \dots a_k' =$$

$$a_2 a_3 \dots a_k + a_1 a_3 \dots a_k + \dots + a_1 a_2 \dots a_{k-1} = a \left(\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_k} \right) \geq$$

$$ak \left(\frac{1}{a_1} \cdot \frac{1}{a_2} \cdots \frac{1}{a_k} \right)^{\frac{1}{k}} = aka^{-\frac{1}{k}} = ka^{\frac{k-1}{k}},$$

where we used the $AM - GM$ inequality to complete our proof. □

One can read more about the arithmetic derivative and its deep connection to various conjectures in number theory such as the Twin Prime Conjecture and the Goldbach Conjecture by clicking HERE

We can extend our concept of recursion to define sequences where future terms depend on not just the previous term but multiple previous terms. Define a sequence of natural numbers by $f_0 := 0, f_1 := 1$ and for $n \geq 0$ define $f_{n+2} = f_{n+1} + f_n$. So $f_2 = f_1 + f_0 = 1 + 0 = 1, f_3 = f_2 + f_1 = 1 + 1 = 2$ etc... We list the first few terms of this sequence:

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \dots$$

$$f_0, f_1, f_2, f_3, f_4, f_5, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, \dots$$

The sequence $(f_k)_{k=0}^{\infty}$ is known as the *Fibonacci sequence* and is an example of a doubly recursive sequence; each new term of the sequence depends on the previous two, in this case it is their sum. A nice way to organize and work with a doubly recursive sequence is to use 2×2 matrices, so we give a brief introduction to them and some basic facts about them.

Definition 5.27. A 2×2 (real) matrix A is an array $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where $a, b, c, d \in \mathbb{R}$. We define the product of two matrices as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{pmatrix}$$

We define the determinant of a matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, written $det(A)$, as $ad - bc$ and leave it as exercises to the reader that matrix multiplication is associative and that $det(AB) = det(A)det(AB)$. We denote by I the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and note that $IA = A = AI$ for any matrix A.

Theorem 5.28. *Let* $F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} f_2 & f_1 \\ f_1 & f_0 \end{pmatrix}$. *Then for any* $n \in \mathbb{N}^+$ *we have* $F^n = \begin{pmatrix} f_{n+1} & f_n \\ f_n & f_{n-1} \end{pmatrix}$.

Proof. We prove this by induction on $n \in \mathbb{N}^+$ the base case $n = 1$ simply being the definition of F. Assume that $F^k = \begin{pmatrix} f_{k+1} & f_k \\ f_k & f_{k-1} \end{pmatrix}$ for some $k \in \mathbb{N}^+$. Then

$$F^{k+1} = (F^k)(F) = \begin{pmatrix} f_{k+1} & f_k \\ f_k & f_{k-1} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} f_{k+1} + f_k & f_{k+1} \\ f_k + f_{k-1} & f_k \end{pmatrix}$$

which is equal to $\begin{pmatrix} f_{k+2} & f_{k+1} \\ f_{k+1} & f_k \end{pmatrix}$ and so we are done by the PMI. □

We can find a wealth of interesting and useful identities involving the Fibonacci numbers simply by exploiting the above theorem. For instance, if $k, n \in \mathbb{N}^+$ then the fact that $F^{k+n} = F^k F^n$ or

$$\begin{pmatrix} f_{k+n+1} & f_{k+n} \\ f_{k+n} & f_{k+n-1} \end{pmatrix} = \begin{pmatrix} f_{k+1} & f_k \\ f_k & f_{k-1} \end{pmatrix} \begin{pmatrix} f_{n+1} & f_n \\ f_n & f_{n-1} \end{pmatrix} =$$
$$\begin{pmatrix} f_{k+1}f_{n+1} + f_n f_k & f_{k+1}f_n + f_k f_{n-1} \\ f_{n+1}f_k + f_n f_{k-1} & f_n f_k + f_{n-1}f_{k-1} \end{pmatrix}$$

so by equating corresponding matrix entries we get the following corollary:

Corollary 5.28.1. *For* $k, n \in \mathbb{N}^+$ *we have*

$$f_{k+n} = f_{k+1}f_n + f_k f_{n-1}.$$

In particular if we set $n = k + 1$ *we get* $f_{2k+1} = f_{k+1}^2 + f_k^2$.

We leave the following as an exercise to the reader (use induction and the above corollary):

Corollary 5.28.2. *$f_k | f_{nk}$ for any $k, n \in \mathbb{N}$.*

Corollary 5.28.3. *$\gcd(f_k, f_m) = f_g$ where $g = \gcd(k, m)$ where $k, m \in \mathbb{N}$.*

The Fibonacci sequence, although recursive, does have a non recursive description. Let $r = \frac{1+\sqrt{5}}{2}$ be the positive root of $x^2 = x + 1$ and $s = \frac{1-\sqrt{5}}{2}$ be the negative root. We know that $r^2 = r + 1, s^2 = s + 1$ and hence that $r = 1 + \frac{1}{r}$ and $s = 1 + \frac{1}{s}$.

We check that $det(F) = -1$ and so $det(F^n) = (det(F))^n = (-1)^n$ and this immediately gives us another well known identity:

Corollary 5.28.4 (Cassini's Identity)**.** *Let $n \in \mathbb{N}^+$. Then $f_{n+1} f_{n-1} - f_n^2 = (-1)^n$.*

Theorem 5.29 (Binet's Formula)**.** *For $n \in \mathbb{N}^+$ we have*

$$f_n = \frac{1}{\sqrt{5}}(r^n - s^n).$$

Proof. We prove $\sqrt{5} f_n = r^n - s^n$ by induction on $n \in \mathbb{N}^+$, the base case being that $\sqrt{5} = \sqrt{5} f_1 = r^1 - s^1$ which is true. Assume that $\sqrt{5} f_k = r^k - s^k$ holds for all $1 \leq k \leq n$ where $n \in \mathbb{N}^+$. Then

$$\sqrt{5} f_{n+1} = \sqrt{5} f_n + \sqrt{5} f_{n-1} = (r^n - s^n) + (r^{n-1} - s^{n-1})$$

where the first equality is from the definition of the Fibonacci sequence and the second is our inductive hypothesis applied when $k = n, n - 1$. This in turn equals

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

or $r^{n+1} - s^{n+1}$ and we are done by the PMI. □

Note that $0 < |s| < 1$ and so $s^n \to 0$ as $n \to \infty$ and so we can approximate f_n by $\frac{r^n}{\sqrt{5}} = \frac{(1+\sqrt{5})^n}{2^n \sqrt{5}}$ when n is large. The number r is very famous in its own right and is often denoted by $\phi = \frac{1+\sqrt{5}}{2}$ and is called the **Golden ratio**, an important mathematical and physical constant.

Theorem 5.30. *For each $n \in \mathbb{N}$ wive $n \geq 3$ we have $\frac{1.5^n}{\sqrt{5}} < f_n < \frac{2^n}{\sqrt{5}}$.*

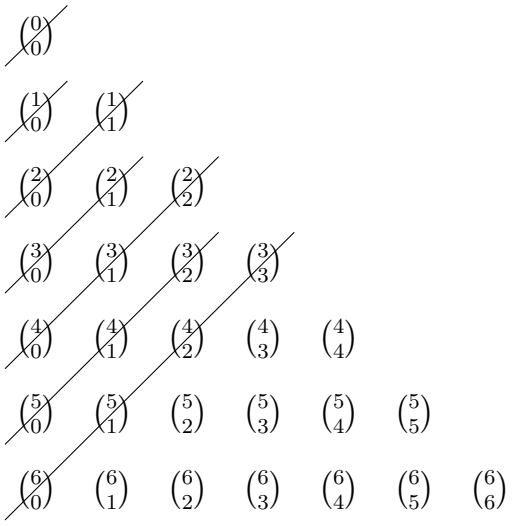
Proof. We first prove $(1.5)^n < \sqrt{5} f_n$ by induction on $n \in \mathbb{N}$ where $n \geq 1$ the base being that $(1.5)^3 < \sqrt{5} f_3 = 2\sqrt{5}$ which is true. Assume that $(1.5)^k < \sqrt{5} f_k$ for all $1 \leq k \leq n$. Then

$$\sqrt{5} f_{n+1} = \sqrt{5} f_n + \sqrt{5} f_{n-1} > \sqrt{5}((1.5)^n + (1.5)^{n-1}) = \sqrt{5}(1.5)^{n-1}(2.5) >$$
$$\sqrt{5}(1.5)^{n-1}(1.5)^2 = \sqrt{5}(1.5)^{n+1} > (1.5)^{n+1}.$$

So by the PMI the first inequality holds for all $n \in \mathbb{N}$ with $n \geq 3$.

Now we prove that $\sqrt{5} f_n < 2^n$ for $n \geq 3$ by induction on n where the the base case is $\sqrt{5} f_3 < 2^3 = 8$ which is true. Assume that $\sqrt{5} f_k < 2^k$ for all $1 \leq k \leq n$. Then $\sqrt{5} f_{n+1} = \sqrt{5} f_n + \sqrt{5} f_{n-1} < 2^n + 2^{n-1} < 2^n + 2^n = 2^{n+1}$ and so by the PMI the second inequality holds for all $n \in \mathbb{N}$ with $n \geq 3$. □

Let's revisit Pascal's triangle and see if we can find the Fibonacci numbers. We arrange the binomial coefficients $\binom{n}{k}$ so that we start with a vertical column with $\binom{n}{0}$ for $n \in \mathbb{N}$ and add binomial coefficients with a slightly different spacing from before as follows:



By drawing line segments with slope one as indicated and adding up any binomial coefficients that get touched we get the sequence:

$$\binom{0}{0}, \binom{1}{0}, \binom{1}{1} + \binom{2}{0}, \binom{2}{1} + \binom{3}{0}, \binom{2}{2} + \binom{3}{1} + \binom{4}{0},$$
$$\binom{3}{2} + \binom{4}{1} + \binom{5}{0}, \binom{3}{3} + \binom{4}{2} + \binom{5}{1} + \binom{6}{0}, \dots$$

or

$$1, 1, 2, 3, 5, 8, 13 \dots$$

which is very clearly the Fibonacci sequence. This can be expressed as

$$\sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} = f_{n+1},$$

where $\lfloor n/2 \rfloor$ is the largest integer less than or equal to $n/2$.

There are even more interesting relationships between the Fibonnaci numbers, the binomial coefficients, and even the golden ratio in the form of Binet's formula. Set $f_{-1} := 1$ and recall r, s the positive and negative roots of $x^2 = x + 1$. We recall that $r = 1 + \frac{1}{r}$, $-\frac{1}{r} = 1 - r, s = 1 + \frac{1}{s}, -\frac{1}{s} = 1 - s$.

Theorem 5.31. *For any $n \in \mathbb{N}$,*

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} f_{k-1} = f_{n+1},$$

Proof. Oddly this doesn't need a proof by induction; the results we're curated so far using induction will suffice here. We first note that Binet's formula can be extended to $n = -1$ by checking that $1 = f_{-1} = \frac{1}{\sqrt{5}}(\frac{1}{r} - \frac{1}{s}) = \frac{\sqrt{5}}{\sqrt{5}}$. Then

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}f_{k-1} = -\frac{1}{\sqrt{5}}\sum_{k=0}^{n}\binom{n}{k}((-r)^{k-1} - (-s)^{k-1}) =$$

$$\frac{r^{-1}}{\sqrt{5}}\sum_{k=0}^{n}\binom{n}{k}(-r)^k - \frac{s^{-1}}{\sqrt{5}}\sum_{k=0}^{n}\binom{n}{k}(-s)^k$$

By the Binomial Theorem we know that $\sum_{k=0}^{n}\binom{n}{k}(-r)^k = (1-r)^n$ and $\sum_{k=0}^{n}\binom{n}{k}(-s)^k = (1-s)^n$ and so our sum above is just

$$\frac{r^{-1}}{\sqrt{5}}(1-r)^n - \frac{s^{-1}}{\sqrt{5}}(1-s)^n$$

where we know $(1-r)^n = (-\frac{1}{r})^n$ and $(1-s)^n = (-\frac{1}{s})^n$ and so the above becomes

$$\frac{r^{-1}}{\sqrt{5}}\left(-\frac{1}{r}\right)^n - \frac{s^{-1}}{\sqrt{5}}\left(-\frac{1}{s}\right)^n = -\frac{1}{\sqrt{5}}(-r)^{-n-1} + \frac{1}{\sqrt{5}}(-s)^{-n-1}.$$

But $(-s)^{-1} = r$ and $(-r)^{-1} = s$ since $rs = -1$ and so this becomes

$$\frac{1}{\sqrt{5}}(r^{n+1} - s^{n+1}) = f_{n+1}$$

by Binet's formula once again. □

This formula might seem obtuse so let's look at what's going on for $n = 7$ for instance. We have $(-1)^k$, $\binom{n}{k}$, and f_{k-1} that we are multiplying together and then adding up:

$(-1)^k$:	1	−1	1	−1	1	−1	1	−1
$\binom{n}{k}$:	1	7	21	35	35	21	7	1
f_{k-1} :	1	0	1	1	2	3	5	8
$(-1)^k\binom{n}{k}f_{k-1}$:	1	0	21	−35	70	−63	35	−8

and then we add up the bottom row to get 21 which is indeed f_8. Patterns beget patterns.

Example 5.32. Define a sequence t_n recursively by $t_0 = 0, t_1 = 1, t_2 = 1$ and $t_{n+3} = t_{n+2} + t_{n+1} + t_n$ for $n \geq 0$. This is known as the *tribonacci sequence* and its first few terms are

$$0, 1, 1, 2, 4, 7, 13, 24, 44, 81, 149, 274, 504, 927\ldots$$

If one looks at the related sequence $q_n := \frac{t_{n+1}}{t_n}$ for $n \geq 1$ we see that the first few terms are

$$q_3 = 2, q_4 = 1.75, q_5 = 1.857.., q_6 = 1.846.., q_7 = 1.83\overline{3}, q_8 = 1.841.., q_9 = 1.839..$$

Just as the case with the Fibonacci sequence, this sequence of quotients converges in very much the same manner. The number T it converges to is approximately

$$T \sim 1.83928675521416113255185256465328660042417874609 75...$$

or

$$T = \frac{1 + \sqrt[3]{19 + 3\sqrt{33}} + \sqrt[3]{19 - 3\sqrt{33}}}{3}$$

T is one of the solutions to the cubic equation $x^3 = x^2 + x + 1$.

6 Rational and Real Numbers

There are many similarities between the real numbers $\mathbb{R}$ and its subset of rational numbers $\mathbb{Q}$, the most important similarity is that they are both **ordered fields**.

Definition 6.1. A set $\mathbb{F}$ along with binary operations $+$ and $\cdot$ and partial order $\leq$ is said to be an ordered field if it satisfies the following properties:

1. $\forall a, b, c \in \mathbb{F}$ we have $(a + b) + c = a + (b + c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
2. $\forall a, b \in \mathbb{F}$ we have $a + b = b + a$ and $a \cdot b = b \cdot a$
3. There exists elements $0 \neq 1$ of $\mathbb{F}$ so the $\forall a \in \mathbb{F}$ we have $0 + a = a$ and $1 \cdot a = a$
4. $\forall a \in \mathbb{F}\, \exists b \in \mathbb{F}$ so that $a + b = 0$
5. $\forall a \in \mathbb{F}\, (a \neq 0)$ implies $(\exists b \in \mathbb{F})$ so that $ab = 1$
6. $\forall a, b, c \in \mathbb{F}$ we have $a \cdot (b + c) = a \cdot b + a \cdot c$
7. $\forall a, b \in \mathbb{F}$ either $a \leq b$ or $b \leq a$
8. $\forall a, b, c \in \mathbb{F}$ if $a \leq b$ then $a + c \leq b + c$
9. $\forall a, b \in \mathbb{F}$ if $0 < a$ and $0 < b$ then $0 < a \cdot b$

[We recall here that $+$ and $\cdot$ being binary operations means they are both functions from $\mathbb{F} \times \mathbb{F}$ to $\mathbb{F}$, in other words, $\forall a, b \in \mathbb{F}$ we have $a + b, a \cdot b \in \mathbb{F}$; we also recall that $\leq$ being a partial order means it is reflexive, transitive, and antisymmetric, and $a < b$ means $a \leq b$ and $a \neq b$]

Properties 1 through 6 describe a structure that is a field, the addition of properties 7 through 9 make $\mathbb{F}$ an ordered field. We will write xy for $x \cdot y$. We often write $-a$ for the element b in property 4; we often write a^{-1} or $\frac{1}{a}$ for the element b in property 5; we leave it to the reader to show that these elements are uniquely defined. We also will often write $\frac{a}{b} := ab^{-1}$ when we have a non-zero element $b \in \mathbb{F}$. One immediate consequence of property 9 is that $x^2 > 0$ for any non-zero element x of an ordered field. We also have $0 < 1$ or else if $1 < 0$ we would have $0 < -1$ and hence $0 < (-1)^2 = 1$. This rules out certain fields from being ordered fields, in particular the field of complex numbers $\mathbb{C}$ of numbers of the form $a+bi$ where $i^2 = -1$, since if $\leq$ was such an order for $\mathbb{C}$ we could not have $0 < i$ or else $0 < i^2 = -1$ and so $1 < 0$; if $i < 0$ then $0 < -i$ and so $0 < (-i)^2 = -1$.

$\mathbb{Q} \subseteq \mathbb{R}$ are both ordered fields with the usual operations of $+$ and $\cdot$ and the usual left to right order. There are plenty of distinct proper subsets of $\mathbb{R}$ that are also fields, for instance $\mathbb{Q}[\sqrt[n]{r}] := \{a + b\sqrt[n]{r} \mid a, b \in \mathbb{Q}\}$ where r is a positive integer. These are distinct from $\mathbb{Q}$ if $n \geq 2$ and $r \neq s^2$ for any element $s \in \mathbb{N}$. It is a fact not proven here that if $\mathbb{F} \subseteq \mathbb{R}$ is an ordered field with the same operations and order as those in $\mathbb{R}$ that $\mathbb{Q} \subseteq \mathbb{F}$, and so in a very strong sense $\mathbb{Q}$ is the smallest ordered field inside of $\mathbb{R}$. If we let $m, n \in \mathbb{Z}$ with $n \neq 0$ we have $\frac{m}{n} \in \mathbb{Q}$ by definition. Let $g = \gcd(m, n)$, and set $m' = m/g, n' = n/g$ so that $\frac{m}{n} = \frac{m'g}{n'g} = \frac{m'}{n'}$. We see that $\gcd(m', n') = 1$ (why?) and say that the expression $\frac{m'}{n'}$ is in *lowest terms*.

Theorem 6.2. *Suppose* $\gcd(a,b)=1$ *and that* $\frac{a}{b}=\frac{c}{d}$. *Then* $a|c$ *and* $b|d$.

Proof. If $\frac{a}{b}=\frac{c}{d}$, we have $ad=bc$, and so any prime power p^t that shows up in the prime factorization of a must show up in the prime factorization of bc. The condition $\gcd(a,b)=1$ tells us that p^t must divide c and so every prime power dividing a divides c. So $a|c$. Then $c=as$ for some $s\in\mathbb{Z}$ and so $ad=b(as)$ or $d=bs$ and so $b|d$ as well. □

Definition 6.3. The set $\mathbb{R}-\mathbb{Q}$ is called the set of irrational numbers.

Theorem 6.4. $\sqrt{2}\in\mathbb{R}-\mathbb{Q}$; *in other words* $\sqrt{2}$ *is irrational.*

Proof 1. Assume to the contrary, that $\sqrt{2}\in\mathbb{Q}$. Then we can write $\sqrt{2}=\frac{m}{n}$ in lowest terms, i.e. $\gcd(m,n)=1$. So $2=\frac{m^2}{n^2}$ and so $2n^2=m^2$. Since $2|m^2$ we know by Euclid's Lemma that $2|m$ so $\exists k\in\mathbb{Z}$ so that $m=2k$. But then $2n^2=(2k)^2$ or $n^2=2k^2$ and similarly we get that $2|n$ contradicting the assumption $\gcd(m,n)=1$. So $\sqrt{2}$ is irrational. □

Proof 2. Assume to the contrary, that $\sqrt{2}\in\mathbb{Q}$. Then we can write $\sqrt{2}=\frac{a}{b}$ in lowest terms, i.e. $\gcd(a,b)=1$. So $2=\frac{a^2}{b^2}$ and we have $\frac{a}{b}=\frac{2b}{a}$. We can apply Theorem 6.2 to get that $b|a$ and so $\sqrt{2}=\frac{a}{b}\in\mathbb{N}$. This is a contradiction as $1<\sqrt{2}<2$ is clearly not an integer.
□

Corollary 6.4.1. *Let* p *be a prime. Then* $\sqrt[n]{p}$ *is irrational for any natural number* $n\geq 2$.

Proof. This is more or less the same proof that $\sqrt{2}$ is irrational: assume to the contrary and write $\sqrt[n]{p}=\frac{a}{b}$ in lowest terms, use Euclid's Lemma to get that $p|a,b$. We leave these details to the reader. □

Theorem 6.5. *Let* $d\in\mathbb{N}$ *so that* $\nexists k\in\mathbb{N}$ *so that* $d=k^2$. *Then* $\sqrt{d}$ *is irrational.*

Proof. There exists $n\in\mathbb{N}$ so that $n^2<d<(n+1)^2$ and so $0<\sqrt{d}-n<1$. Suppose $\sqrt{d}=\frac{a}{b}\in\mathbb{Q}$ where $a,b>0$ and choose b to be the smallest such denominator so that $\exists a\in\mathbb{N}$ with $\sqrt{d}=\frac{a}{b}$ (hence b is the smallest natural number so that $b\sqrt{d}=a\in\mathbb{N}$). Then $(\sqrt{d}-n)(b\sqrt{d})=bd-nb\sqrt{d}=bd-nba\in\mathbb{N}$. But since $0<\sqrt{d}-n<1$ we have $(\sqrt{d}-n)b<b$ and so $(\sqrt{d}-n)b$ is a natural number smaller than b so that $(\sqrt{d}-n)b\sqrt{d}$ is a natural number, a contradiction. So $\sqrt{d}$ is irrational. □

What distinguishes the real numbers from other ordered fields is that the real numbers are *complete*; we need a few additional definitions to make this concept precise. We write $t \leq A$ where $t \in \mathbb{R}$ and $A \subseteq \mathbb{R}$ to mean $\forall a \in A, l \leq a$. Analogously we define $B \leq s$ where $s \in \mathbb{R}$ and $B \subseteq \mathbb{R}$ to mean $\forall b \in B, b \leq s$. We note that given any $r \in \mathbb{R}$ we have both $r \leq \emptyset$ and $\emptyset \leq r$ since these definitions are satisfied vacuously, so we will in general not include the emptyset in definitions like the following:

Definition 6.6. We say $\emptyset \neq A \subseteq \mathbb{R}$ is bounded below if $\exists l \in \mathbb{R}$ so that $l \leq A$; likewise we say that $\emptyset \neq B \subseteq \mathbb{R}$ is bounded above if $\exists u \in \mathbb{R}$ so that $B \leq u$. We say that $\emptyset \neq A \subseteq \mathbb{R}$ is bounded if it is bounded below and bounded above. We say $u := \sup(A)$ is the *least upper bound* or *supremum* for $\emptyset \neq A$ if $A \leq u$ and if $A \leq u' \leq u$ implies $u = u'$. We say that $l := \inf(B)$ is the *greatest lower bound* or *infimimum* for $\emptyset \neq B$ if $l \leq B$ and if $l \leq l' \leq B$ implies $l = l'$. We leave it as a simple exercise to the reader to check that if a set has a least upper bound or a greatest lower bound, these are unique and well-defined.

Example 6.7. $-10 \leq [-1, 7)$ and $[-1, 7) < 23$ but $\inf([-1, 7)) = -1$ and $\sup([-1, 7)) = 7$. If $A = \{\frac{1}{n} \,|\, n \in \mathbb{N}^+\} = \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$ then $A \leq 10$ but $\sup(A) = 1$. We have $-7 \leq A$ of course but $\inf(A) = 0$. Note that in these cases $\inf([-1, 7)) \in [-1, 7)$ while $\sup([-1, 7)) \notin [-1, 7)$ while $\inf(A) \notin A$ but $\sup(A) \in A$.

Completeness Property of $\mathbb{R}$: Any non-empty subset A that is bounded above has a least upper bound.

Note that this is not a property that $\mathbb{Q}$ enjoys: if $A = \{r \in \mathbb{Q} \,|\, r < \sqrt{2}\}$, clearly A is non-empty and $A \leq 2$. But given an upper bound $u \in \mathbb{Q}$ for A we have that $u' = \frac{2u+2}{u+2}$ is a rational number that so that $u' < u$ since $u' < u$ is equivalent to saying $2u+2 < u^2+2u$ or $A \leq \sqrt{2} < u$, and $A \leq u'$ since $\sqrt{2} < u'$ is equivalent to $2(u+2)^2 < (2u+2)^2$ or $u^2 + 4u + 4 < 2u^2 + 4u + 2$ or $\sqrt{2} < u$. So there is no smallest rational upper bound for A. There is a smallest real upper bound for A namely $\sqrt{2}$.

Theorem 6.8. *Any non-empty subset $B \subseteq \mathbb{R}$ that is bounded below has a greatest lower bound.*

Proof 1. Let $L = \{l \in \mathbb{R} \,|\, l \leq B\}$. Then L is a non-empty subset of $\mathbb{R}$ that is bounded above by every element of B. So $t = \sup(L)$ exists by the Completeness Property and we claim that $t = \inf(B)$. Clearly $t \leq B$ and so let $t \leq t' \leq B$. Then $t' \in L$ and so $t' \leq t$ by the definition of $t = \sup(L)$ and so $t = t'$ and indeed $t = \inf B$ exists. □

Proof 2. Let $-B = \{-b \,|\, b \in \mathbb{B}\}$. B is bounded below by say x and so $-B \leq -x$ is bounded above and also non-empty, and so by the Completeness Property $s = \sup(-B)$ exists. We claim that $t := -s = \inf(B)$. We know $t \leq B$ since this is equivalent to $-B \leq s$ which if true. Suppose $t \leq t' \leq B$. Then $-B \leq -t' \leq -t = s$ and since $s = \sup(-B)$ we have $-t' = s = -t$ or $t = t'$ and we are done. □

Another fairly immediate consequence of the Completeness Property is the following:

Theorem 6.9 (Archimedean Property of $\mathbb{R}$)**.** *Given any $r \in \mathbb{R}$ there exists $n \in \mathbb{N}$ so that $r < n$.*

Proof. Suppose this is false, that $\mathbb{N} < r$ for some $r \in \mathbb{R}$. Then by the Completeness Property $s = \sup(\mathbb{N})$ exists. But then $s - 1$ is not an upper bound for $\mathbb{N}$ or else this would contradict the definition of s, so $\exists n \in \mathbb{N}$ so that $s - 1 < n$. But then $s < n + 1$ contradicts s being an upper bound for $\mathbb{N}$ since $n + 1 \in \mathbb{N}$. □

A useful consequence of the Archimedean Property is the following:

Corollary 6.9.1. *Given and real number $0 < \epsilon$, there is an $N \in \mathbb{N}^+$ so that $0 < \frac{1}{N} < \epsilon$.*

Proof. By the Archimedian property $\exists N \in \mathbb{N}^+$ so that $\frac{1}{\epsilon} < N$. Then $0 < \frac{1}{N} < \epsilon$. □

We use the last result to show that between any two real numbers there is a rational number:

Theorem 6.10. *Let $a, b \in \mathbb{R}$ with $a < b$. Then $\exists q \in \mathbb{Q}$ so that $a < q < r$.*

Proof. If $a < 0 < b$ we can set $q = 0$. First assume $0 \leq a < b$. If $a = 0$ we can use the above corollary to find $N \in \mathbb{N}$ so that $0 < \frac{1}{N} < b$, so let's assume $0 < a < b$. By the corollary above $\exists t \in \mathbb{N}$ so that $0 < \frac{1}{t} < b - a$. By the Archimedean property there exists $s \in \mathbb{N}$ so that $at < s$; choose s to be the smallest such natural number with this property so that $s - 1 \leq at < s$. Then $\frac{s}{t} - \frac{1}{t} \leq a < \frac{s}{t}$ and so $\frac{s}{t} \leq a + \frac{1}{t} < a + (b - a) = b$ gives us altogether $a < \frac{s}{t} < b$ as desired. If $a < b \leq 0$ then we can look at $0 \leq -b < -a$ and apply the previous argument. □

This theorem says that the rational numbers are *dense* in the real numbers, no matter how small the distance between two real numbers is we can always find a rational number hiding in between. Likewise, between any two real numbers we can find an irrational number as well:

Theorem 6.11. *Let $a, b \in \mathbb{R}$ with $a < b$. Then $\exists t \in \mathbb{R} - \mathbb{Q}$ so that $a < t < b$.*

Proof. Since the rational numbers are dense we can find a rational number p so that $a < p < b$ and likewise be can find another rational number q so that $p < q < b$. Let $t = \frac{q-p}{\sqrt{2}} + p$. We claim that $t \notin \mathbb{Q}$; if $t \in \mathbb{Q}$ then so is $t - p = \frac{q-p}{\sqrt{2}}$ and $\sqrt{2} = \frac{t-p}{q-p}$ contradicting the fact that $\sqrt{2}$ is irrational. We know $p < t$ and since $\sqrt{2} > 1$ we know that $\frac{q-p}{\sqrt{2}} < q - p$ so that $t < (q - p) + p = q$. □

A useful tool and concept when discussing properties of the rational vs real numbers is the concept of a limit of a sequence. A real sequence $(a_k)_{k=0}^{\infty}$ is simply a function $a : \mathbb{N} \to \mathbb{R}$ where we set $a(k) = a_k$. We want to make precise what it means to say $a_n \to L$ as $n \to \infty$. It's very likely you're never seen a precise formulation of this concept, more likely something vaguely intuitive yet very imprecise. The following definition was settled on in the early 19th century several hundred year after calculus had been discovered. We give the definition first when $L = 0$, for sequences that converge to 0, introduce examples and theorems for these, and then expand our focus to sequences that converse to an arbitrary real number L.

Definition 6.12. Let $(a_k)_{k=0}^{\infty}$ be a real sequence. We say $\lim_{n\to\infty} a_n = 0$ or write $a_n \to 0$ as $n \to \infty$ if for all $\epsilon > 0$ there is a natural number N so that

$$\forall n \geq N \Longrightarrow |a_n| < \epsilon$$

Definition 6.13. Let $(a_k)_{k=0}^{\infty}$ be a real sequence. We say $\lim_{n\to\infty} a_n = L$ or write $a_n \to L$ as $n \to \infty$ if:

$$(\forall \epsilon > 0)(\exists N \in \mathbb{N})(\forall n \geq N)\,, |a_n - L| < \epsilon.$$

These might look a bit intimidating on first glance, especially the second one, so we introduce some terminology and facts involving absolute value and inequalities so we can get a better understanding and focus first on sequences that converge to zero. We also note here that our sequences do not need to start with $n = 0$, we can begin indexing our sequences at any integer, we merely do so here for convenience.

Lemma 6.14. *Let $x, y \in \mathbb{R}$. Then*

1. $-|x| \leq x \leq |x|$ *and* $|xy| = |x||y|$
2. $|x|^2 = x^2$
3. $||x| \pm ||y|| \leq |x \pm y| \leq |x| + |y|$

Proof. Everything but part 3 is an immediate consequence of the definition of absolute value. We note that $|x+y|^2 = (x+y)^2 = x^2+y^2+2xy \leq |x|^2+|y|^2+2|x||y| = (|x|+|y|)^2$ and since $|x+y|$ and $|x| + |y|$ are both non-negative taking square roots gives us $|x + y| \leq |x| + |y|$. Replacing y with $-y$ gives us $|x - y| \leq |x| + |-y| = |x| + |y|$ and so $|x \pm y| \leq |x| + |y|$. Then we know $|x| = |y+(x-y)| \leq |y|+|x-y|$ or $|x|-|y| \leq |x-y|$ and also $|y| = |x+(y-x)| \leq |x|+|y-x|$ or $|y|-|x| \leq |x-y|$. So $||x|-|y|| \leq |x-y|$. Replacing y by $-y$ in this we get $||x|-|y|| \leq |x \pm y|$. □

The third part of the lemma is often called the triangle inequality: if one sets $x = a - b$ and $y = b - c$ it becomes $|a - c| = |(a - b) + (b - c)| \leq |a - b| + |b - c|$ and we can think of $|a - b|$ as the distance from a to b: for example if $a = -1$ and $b = 8$ then $|a - b| = |-1 - 8| = |-9| = 9$ is certainly the length of the interval $[-1, 8]$ in the real line. This idea of thinking of $|a - b|$ as the distance from a to b turns out to be pretty useful.

Definition 6.15. Let $r > 0$ be any positive real number and $a \in \mathbb{R}$. Define $D_r(a) := \{x \in \mathbb{R} \,|\, |x - a| < r\}$ to be the r-neighborhood of a.

So $D_r(a)$ is the set of all points that are no more than distance r from the point a: solving the inequality $|x - a| < r$ is the same as solving $-r < x - a < r$ or $a - r < x < a + r$ and so is the open interval $(a - r, a + r)$. So $D_r(a) = (a - r, a + r)$. Note that if $0 < r_1 < r_2$ we have $D_{r_1}(a) \subsetneq D_{r_2}(a)$ since the interval $(a - r_1, a + r_1)$ is properly contained in $(a - r_2, a + r_2)$. For instance $D_1(0) \subseteq D_2(0)$ is simply saying $(-1, 1) \subseteq (-2, 2)$.

Let us revisit our definition of $a_n \to 0$, a sequence converging to zero. What we are saying is that given any radius $\epsilon > 0$ there should exist some time, say $N = N(\epsilon)$, so that after we reach this time (so for all $n > N$) we have that $a_n \in D_\epsilon(0)$. It is not good enough for us to merely arrive in this ϵ-neighborhood of zero, we have to stay within this ϵ-neighborhood. A sequence might have some wild behaviour before it settles down and approaches zero.

Example 6.16. Let $(a_n)_{n=0}^{\infty}$ be the sequence given by

$$a_n = \begin{cases} 1 & n \quad \text{even} \\ \frac{1}{n} & n \quad \text{odd} \end{cases}$$

The first terms of this sequence are $1, 1, 1, \frac{1}{3}, 1, \frac{1}{5}, 1, \frac{1}{7}, 1, \frac{1}{9}, 1,$
$\frac{1}{11}, 1, \frac{1}{13}, 1, \frac{1}{15}, 1, \frac{1}{17}, 1, \frac{1}{19}, 1, \frac{1}{21}, \ldots$. There are terms in this sequence that do get arbitrarily small, so given any ϵ-neighborhood of zero $D_\epsilon(0)$ we can certainly find $n \in \mathbb{N}$ so that $a_{2n+1} = \frac{1}{2n+1} \in D_\epsilon(0)$. But unless $\epsilon \geq 1$ we would immediately have $a_{2n+2} = 1 \notin D_\epsilon(0)$, the sequence would bounce back out of the neighborhood. In this case it is true that the sequence would have infinitely many terms inside of any $D_\epsilon(0)$, but for any significantly small ϵ there would not be any consistency.

Example 6.17. Let $(x_n)_{n=1}^{\infty}$ be the sequence $x_n = \frac{4}{n}$. Then we do have $x_n \to 0$ since given $\epsilon > 0$ we can find $N \in \mathbb{N}$ so that $N > \frac{4}{\epsilon}$ by the Archimedean property. So $\frac{4}{N} < \epsilon$ and moreover, given any $n \geq N$ we have $\frac{4}{n} \leq \frac{4}{N}$ and so $x_n = \frac{4}{n} \in D_\epsilon(0)$ for every $n \geq N$. Here it was not good enough for us to merely find N so that $x_N \in D_\epsilon(0)$ we needed $x_N, x_{N+1}, x_{N+2}, x_{N+3}, x_{N+4}, \ldots$ are in this neighborhood.

Example 6.18. Let $y_n = \frac{1-n}{1+n+n^2}$ where $n = 1, 2, 3 \ldots$. We show that $y_n \to 0$. Before we give our proof, let us see what we need to accomplish. We need to find N first and foremost so that $y_N \in D_\epsilon(0)$ given an $\epsilon > 0$, namely we want $\left|\frac{1-N}{1+N+N^2}\right| < \epsilon$. Now $\left|\frac{1-N}{1+N+N^2}\right| = \left|\frac{N-1}{1+N+N^2}\right| < \frac{N}{N^2}$ since this last fraction has a bigger numerator and a smaller denominator. But this is just $\frac{1}{N}$ and know we can choose $N > \frac{1}{\epsilon}$ here. Let us put these ingredients into proof form: Let $\epsilon > 0$. Choose $N \in \mathbb{N}$ so that $N > \frac{1}{\epsilon}$. Then for all $n \geq N$ we have

$$\left|\frac{1-n}{1+n+n^2}\right| < \frac{1}{n} \leq \frac{1}{N} < \epsilon$$

and so $y_n \in D_\epsilon(0)$. Our proof is complete. Note that in our planning we focused on N and ϵ first and brought n later on; the reason we were able to do this is because of that very last part of the proof, that $n \geq N$ implied $\frac{1}{n} \leq \frac{1}{N}$.

Let us look at the negation of the statement $a_n \to 0$:

$$\neg((\forall \epsilon > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(a_n \in D_\epsilon(0))$$

The above negation is logically equivalent to

$$(\exists \epsilon > 0)(\forall N \in \mathbb{N})(\exists n \geq N)(a_n \notin D_\epsilon(0)).$$

In other words a sequence *doesn't* converge to zero if we can find a D_ϵ so that no matter what N is provided we are able to find an $n \geq N$ with $a_n \notin D_\epsilon(0)$.

Example 6.19. Let's show that that $b_n = \frac{n+1}{n+2}$ does not converge to zero (in fact it converges to 1): let $\epsilon = \frac{1}{4}$ and $N \in \mathbb{N}$. Then $\frac{N+1}{N+2} = 1 - \frac{1}{N+2} \geq 1 - \frac{1}{2} = \frac{1}{2} > \frac{1}{4}$ and so given any $n \geq N$ we would have $b_n \notin D_{\frac{1}{4}}(0)$ since $|b_n| \geq |b_N| > \frac{1}{4}$.

Lemma 6.20. *For any sequence $(a_n)_{n=0}^\infty$ we have*

$$(a_n \to 0) \iff (|a_n| \to 0) \iff (-a_n \to 0)$$

Proof. This is a consequence of $|a_n| = ||a_n|| = |-a_n|$. □

Theorem 6.21. *If $a_n \to 0$ then $\exists M > 0$ so that $\forall n$ we have $|a_n| < M$, i.e. (a_n) is* bounded.

Proof. We know that $\exists N \in \mathbb{N}$ so that $\forall n \geq N$ we have $a_n \in D_1(0)$. The finite set $\{|a_0|, |a_1|, \ldots, |a_N|\}$ is bounded by some $B > 0$. Let $M = B + 1$. Then for any $n \in \mathbb{N}$ we have $|a_n| < M$. □

Theorem 6.22 (Limit Laws)**.** *If (a_n) and (b_n) are sequences with $a_n, b_n \to 0$ and $c \in \mathbb{R}$ then:*

1. $a_n \pm b_n \to 0$
2. $a_n b_n \to 0$
3. $ca_n \to 0$

Proof. We prove that $a_n + b_n \to 0$, the case $a_n - b_n \to 0$ follows immediately from this. Let $\epsilon > 0$. Since $a_n \to 0$ there is an $N_1 \in \mathbb{N}$ so that $\forall n \geq N$ we have $a_n \in D_{\frac{1}{2}}(0)$. Since $b_n \to 0$ there is an $N_2 \in \mathbb{N}$ so that $\forall n \geq N$ we have $b_n \in D_{\frac{1}{2}}(0)$. Let $N = N_1 + N_2$. Then for any $n \geq N$ we have

$$|a_n + b_n| \leq |a_n| + |b_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and so $a_n + b_n \in D_\epsilon(0)$.

To show $a_n b_n \to 0$ we take $\epsilon > 0$ and then use the fact that since $b_n \to 0$ the sequence is bounded by some $M > 0$. Let $\epsilon' = \frac{\epsilon}{M}$. Then there exists $N \in \mathbb{N}$ so that for all $n \geq N$ we have $a_n \in D_{\epsilon'}(0)$. Then we also have

$$|a_n b_n| \leq |a_n| b_n| \leq |a_n| M < \epsilon' M = \epsilon$$

and so $n \geq N$ implies $a_n b_n \in D_\epsilon(0)$.

The third part is left to the reader. □

We add two more useful tools to our list of theorems about sequences converging to zero, proving the first one and leaving the second as an exercise for the reader to enjoy:

Theorem 6.23 (Squeeze Theorem)**.** *Let $a_n \leq b_n \leq c_n$ be three sequences so that $a_n, c_n \to 0$. Then $b_n \to 0$ as well.*

Proof. Let $\epsilon > 0$. Choose $N \in \mathbb{N}$ so that if $n \geq N$ we have $a_n, c_n \in D_\epsilon(0)$. Then we know that if $b_n \geq 0$ that $|b_n| \leq c_n < \epsilon$ and if $b_n < 0$ that $|b_n| - b_n \leq -a_n = |a_n| < \epsilon$ so either way $b_n \in D_\epsilon(0)$ as well. □

Theorem 6.24. *Suppose $a_n \to 0$ and b_n is any bounded sequence, namely $\exists M > 0$ so that $\forall n \in \mathbb{N}$ we have $|b_n| < M$. Then $a_n b_n \to 0$.*

Proof. Let $\epsilon > 0$. Choose $N \in \mathbb{N}$ so that if $n \geq N$ we have $a_n \in D_{\frac{\epsilon}{M}}(0)$. Then if $n \geq N$ we also have $|a_n b_n| = |a_n||b_n| < |a_n| M < \frac{\epsilon}{M} M = \epsilon$. So $a_n b_n \in D_\epsilon(0)$

$$|a_n b_n| = |a_n||b_n|$$

□

We note that the converse to this theorem is false: let $a_n = \frac{1}{n^2}$ and $b_n = n$. Then $a_n b_n \to 0$ but b_n is not bounded.

Example 6.25. Let $x \in \mathbb{R}$. We wish to show that if $|x| < 1$ then $x^n \to 0$. Let us brainstorm how we might prove this. We want at least to find some N so that $|x|^N < \epsilon$ for a given ϵ. Maybe one thinks of taking a logarithm on both sides, but have not proven anything about the logarithm, so it will be nice to avoid this at least for now. So we get more creative here: let $r > 0$. Then $(1+r)^N = 1 + Nr + \sum_{k=2}^{N} \binom{N}{k} r^k$ by the Binomial Theorem, and so $(1+r)^N < 1 + Nr$ since the terms in the sum are positive (this is called *Bernouilli's Inequality*). Let us focus on $0 < x < 1$. Then $\frac{1}{|x|} = 1 + \frac{1-|x|}{|x|}$ and so $\left(\frac{1}{x}\right)^N = \left(1 + \frac{1-x}{x}\right)^N > 1 + N\left(\frac{1-x}{x}\right) > N\left(\frac{1-x}{x}\right)$ and so $|x|^N = x^N < \frac{x}{N(1-x)}$. It's this last quantity that we want to have less than a given ϵ so we are ready for our proof in the case $0 < x < 1$: let $\epsilon > 0$. Choose $N \in \mathbb{N}$ so that $N > \frac{x}{\epsilon(1-x)}$. Then if $n \geq N$ we have:

$$|x^n| = x^n \leq x^N < \frac{x}{N(1-x)} < \epsilon$$

and so $x^n \in D_\epsilon(0)$ as needed.
We leave the proof that this holds for $-1 < x \leq 0$ to the reader.

We turn to looking at sequences converging to arbitrary limits. Our definition of $a_n \to L$ is that for any $\epsilon > 0$ there is an $N = N(\epsilon) \in \mathbb{N}$ so that if $n \geq N$ we have $a_n \in D_\epsilon(L)$, that given any neighborhood of L however small, our sequence after some point stays in this neighborhood. Another way of saying that $a_n \to L$ is to say that the sequence $b_n := a_n - L \to 0$ and so we can translate an awful lot of statements about limits converging to zero to statements about limits converging to arbitrary real numbers and vice a versa.

Example 6.26. We show that $z_n = \frac{n^2-1}{n^2+1} \to 1$ by first doing some preliminary work. We need that $|\frac{N^2-1}{N^2+1} - 1|$ is smaller than a given ϵ. $|\frac{N^2-1}{N^2+1} - 1| = |\frac{N^2-1}{N^2+1} - \frac{N^2+1}{N^2+1}| = \frac{2}{N^2+1} < \frac{2}{N}$ since the last fraction has a smaller denominator. This sets up our proof: let $\epsilon > 0$ and choose $N \in \mathbb{N}$ so that $N > \frac{2}{\epsilon}$. Then if $n \geq N$ we have

$$\left|\frac{n^2-1}{n^2+1} - 1\right| = \frac{2}{n^2+1} < \frac{2}{n} \leq \frac{2}{N} < \epsilon,$$

and so $\frac{n^2-1}{n^2+1} \in D_\epsilon(1)$.

Example 6.27. We show that $w_n = \sqrt{n^2+n} - n \to \frac{1}{2}$ by again first doing some preliminary work to determine how we should determine $N = N(\epsilon)$. We need that $|\sqrt{N^2+1} - N - \frac{1}{2}|$ is smaller than a given $\epsilon > 0$. Multiplying by $\frac{\sqrt{N^2+N}+N}{\sqrt{N^2+N}+N}$ we get that $|\sqrt{N^2+1} - N - \frac{1}{2}| = |\frac{2N}{2\sqrt{N^2+N}+2N} - \frac{\sqrt{N^2+N}+N}{2\sqrt{N^2+N}+2N}| = \frac{\sqrt{N^2+N}-N}{2\sqrt{N^2+N}+2N}$ which we again multiply by $\frac{\sqrt{N^2+N}+N}{\sqrt{N^2+N}+N}$ to get $|\sqrt{N^2+1} - N - \frac{1}{2}| = \frac{N}{2(\sqrt{N^2+N}+N)^2} < \frac{N}{2N^2} = \frac{1}{2N}$ since leaving off the term $\sqrt{N^2+N}$ makes the denominator smaller and hence the fraction bigger. This we can work with and give our proof: let $\epsilon > 0$ and choose $N > \frac{1}{2\epsilon}$. Then for $n \geq N$ we have that

$$\left|w_n - \frac{1}{2}\right| = \frac{n}{2(\sqrt{n^2+n}+n)^2} < \frac{1}{2n} \leq \frac{1}{2N} < \epsilon,$$

and so $w_n \in D_\epsilon(\frac{1}{2})$ as needed.

How do we know that a given sequence can't converge to two different real numbers? Why couldn't we also get $w_n \to 0$ in the preceding example?

Theorem 6.28 (Uniqueness of Limits)**.** *Suppose we have $a_n \to L$ and $a_n \to L'$. Then $L = L'$.*

Proof. Suppose that $L \neq L'$ and set $\epsilon = \frac{|L-L'|}{2} > 0$. Then there is an $N \in \mathbb{N}$ so that for every $n \geq N$ we have both $a_n \in D_\epsilon(L), a_n \in D_\epsilon(L')$. But then $|L - L'| = |L - N + a_N - L'| \leq |L - a_N| + |a_N - L'| < \frac{|L-L'|}{2} + \frac{|L-L'|}{2} = |L - L'|$ a contradiction. So $L = L'$. □

For some additional practice setting up (ϵ, N) proofs that various sequences converge we direct the reader to HERE. We encourage the reader to state and prove analogs of the Squeeze Theorem and the Limit Laws and to show that sequences that converge are bounded; we add some additional Limit Law theorems here:

Theorem 6.29 (Add'l Limit Laws)**.** *Let $a_n \to L, b_n \to M$. Then*

1. $a_n b_n \to LM$.

2. $\frac{a_n}{b_n} \to \frac{L}{M}$ *if* $b_n, L \neq 0$.

Proof. For 1, if $L = 0$ this follows from a previous theorem so assume $L \neq 0$ and let B be a bound for the sequence a_n, i.e., $|a_n| \leq B$ for all $n \in \mathbb{N}$. Let $\epsilon > 0$. Then there exists $N_a \in \mathbb{N}$ so that if $n \geq N_a$ we have $a_n \in D_{\frac{\epsilon}{2|L|}}(M)$. There also exists $N_b \in \mathbb{N}$ so that if $n \geq N_b$ we have $b_n \in D_{\frac{\epsilon}{2B}}(L)$. Then for $n \geq N := N_a + N_b$ we have

$$|a_n b_n - LM| = |a_n b_n - a_n L + a_n L - LM| \leq |a_n||b_n - L| + |L||a_n - M|$$

$$\leq B|b_n - L| + |L||a_n - M| < B\left(\frac{\epsilon}{2B}\right) + |L|\left(\frac{\epsilon}{2|L|}\right) = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

For 2, we first prove this when $a_n = 1$. Since b_n is bounded and does not converge to zero we know that there exists $0 < B$ so that for all $n \in \mathbb{N}$ we have $B < |b_n|$ (why?). Let $\epsilon > 0$ and choose $N \in \mathbb{N}$ so that if $n \geq N$ we have $b_n \in D_{B|M|\epsilon}(M)$. Then if $n \geq N$ we have

$$\left|\frac{1}{b_n} - \frac{1}{M}\right| = \frac{|b_n - M|}{|b_n||M|} < \frac{|b_n - M|}{B|M|} < \epsilon.$$

We could also have proven this using the Squeeze Theorem noting that

$$0 \leq \left|\frac{1}{b_n} - \frac{1}{M}\right| < \frac{|b_n - M|}{B|M|}$$

and noting that $|b_n - M| \to 0$ as $n \to \infty$.

For 2 in general combine this with part 1. □

Often the terms of a sequence grow bigger and bigger, so we define exactly what it means to say $a_n \to \infty$:

Definition 6.30. We say $a_n \to \infty$ if for any $M > 0$ there exists $N \in \mathbb{N}$ so that for all $n \geq N$ we have $a_n > M$. Like wise we say $b_n \to -\infty$ if $-b_n \to \infty$.

Example 6.31. Clearly the sequence $a_n = n$ has the property that $a_n \to \infty$; this is simply a restatement of the Archimedean Property. A more interesting example is given by $b_n = \sqrt{n + \sqrt{n + \sqrt{n + \sqrt{n}}}}$; we note that $\sqrt{n + a} > \sqrt{n}$ for any $a > 0$ and so $a_n > \sqrt{n + \sqrt{n + \sqrt{n}}} > \sqrt{n + \sqrt{n}} > \sqrt{n}$ and so given any $M > 0$ we choose $N \in \mathbb{N}$ so that $N > M^2$. Then if $n \geq N$ we know that $a_n > \sqrt{n} \geq \sqrt{N} > M$ as needed.

Theorem 6.32. *Suppose $a_n > 0$ is a sequence with positive terms. Then $a_n \to \infty$ iff $\frac{1}{a_n} - \to 0$.*

Proof. Let $\epsilon > 0$ and assume $a_n \to \infty$. Then we can find $N \in \mathbb{N}$ so that $n \geq N$ implies $a_n > \frac{1}{\epsilon}$. But then $|\frac{1}{a_n}| = \frac{1}{a_n} < \epsilon$ when $n \geq N$ and so $\frac{1}{a_n} - \to 0$. Likewise if $\frac{1}{a_n} \to 0$ and $M > 0$ we can find $N \in \mathbb{N}$ so that $n \geq N$ implies $\frac{1}{a_n} < \frac{1}{M}$ or equivalently $a_n > M$ and so $a_n \to \infty$. □

An important class of sequences are those that contain terms that are either increasing or decreasing; we call such sequences *monotone* and establish the following very useful theorem:

Theorem 6.33 (Monotone Convergence Theorem)**.** *A monotone sequence converges iff it is bounded.*

Proof. We give the proof in the case where we have an increasing sequence x_n, i.e. a sequence so that $\forall n\ x_n \leq x_{n+1}$. Assume that x_n is bounded. Let $X = \{x_n \,|\, n \in \mathbb{N}\}$ and note that X is bounded by definition and so $s := \sup(X)$ exists by the Completeness Property. We show that $x_n \to s$. Let $\epsilon > 0$. We know that $s - \epsilon$ is not an upper bound for X since s is the *least* upper bound. So there is some x_N so that $s - \epsilon < x_N$. But then for $n \geq N$ we have

$$|s - x_n| = s - x_n \leq s - x_N < \epsilon,$$

since x_n is increasing. So $x_n \in D_\epsilon(s)$ and we know that $x_n \to s$ converges. The other direction of the 'iff' is a consequence of any convergent sequence being bounded. □

Example 6.34. Let $x_0 = a \in (0, \frac{1}{2})$ and let $x_{n+1} = \frac{x_n}{2}(1 - x_n)$. We prove first that $\forall k \in \mathbb{N}$ that $x_{k+1} \leq x_k$ and that $x_k \geq 0$. We note first that $x_1 = \frac{a}{2}(1 - a)$ and $0 < a < \frac{1}{2}$ implies $\frac{1}{2} < 1 - a < 1$ and so $\frac{1}{4} < \frac{1-a}{2} < \frac{1}{2}$ and so $\frac{a}{2}(1 - a) = x_1 < \frac{a}{2} < x_0$. We note in general that $x_{k+1} - x_k = \frac{x_k}{2}(1 - x_k) - x_k = -\frac{x_k}{2} - \frac{x_k^2}{2} < 0$ and so x_n is decreasing. Since $x_0 > 0$ if we assume that $0 < x_k < \frac{1}{2}$ for some $k \in \mathbb{N}$ we know that $\frac{1}{2} < 1 - x_k < 1$ and so $\frac{1}{4} < \frac{1-x_k}{2} < \frac{1}{2}$ and so multiplying through by x_k we get $\frac{x_k}{4} < x_{k+1} < \frac{x_k}{2}$ and so since $0 < \frac{x_k}{4}$ we know $0 < x_{k+1}$ and since $\frac{x_k}{2} < \frac{1}{4}$ we know $x_{k+1} < \frac{1}{4}$ and so altogether $0 < x_{k+1} < \frac{1}{4} < \frac{1}{2}$. So we know that x_n is both decreasing and bounded and hence has a limit L by the Monotone Convergence Theorem, namely that $x_n \to L$. Then $x_{n+1} \to L$ as well and we know that the limit must satisfy:

$$L = \frac{L}{2}(1 - L) \quad \text{or} \quad L^2 - L = 0,$$

and so the possibilities for L are $L = 0$ or $L = 1$. $L = 1$ is nonsense since our sequence is decreasing from $x_0 = a < \frac{1}{2}$ and so $L = 0$.

Example 6.35. Let $y_0 = y_1 = 0$ and for $n \in \mathbb{N}$ define $y_{n+2} = \frac{1}{3}y_{n+1} + \frac{1}{6}y_n + 1$. We prove by induction that $y_n < 2$ for all $n \in \mathbb{N}$, the base case following from the definition $y_0 = y_1 = 0 < 2$. Assume that the bound holds for $y_0, y_1, \ldots y_k, y_{k+1}$ where $k \in \mathbb{N}$. Then $y_{k+2} = \frac{1}{3}y_{k+1} + \frac{1}{6}y_k + 1 < \frac{2}{3} + \frac{2}{6} + 1 = 2$ as well. Now we show that y_n is increasing by induction, again the base case following from the definition; we note that $y_2 = 1 > 0 = y_1$ as well. Suppose that $y_0 \leq y_1 \leq y_2 \leq \cdots \leq y_{k+1}$ for some $k \geq 1$. Then

$$y_{k+2} - y_{k+1} = \frac{1}{3}(y_{k+1} - y_k) + \frac{1}{6}(y_k - y_{k-1}) \geq 0,$$

so that $y_{k+2} \geq y_{k+1}$ showing that y_n is an increasing sequence. It therefore converges by the Monotone Convergence Theorem, say $y_n \to L$. then $L = \frac{L}{3} + \frac{L}{6} + 1$ or $L = 2$.

Example 6.36. Suppose we want to make sense of an expression like

$$\sqrt{t+\sqrt{t+\sqrt{t+\sqrt{t+\ldots}}}}$$

where $t \geq 1$.

We could define a sequence recursively by $a_1 = \sqrt{t}$ and for $k \geq 0$ set $a_{k+1} = \sqrt{t+a_k}$. So $a_2 = \sqrt{t+\sqrt{t}}, a_3 = \sqrt{t+\sqrt{t+\sqrt{t}}}$ etc... How do we determine if such a sequence has a limit? First we establish that a_n is an increasing sequence where $a_1 = \sqrt{t} \leq \sqrt{t+\sqrt{t}} = a_2$ follows from the fact that $t \geq 1$. If $a_k \leq a_{k+1}$ for some $k \in \mathbb{N}^+$ then we know that $a_k + \sqrt{t} \leq a_{k+1} + \sqrt{t}$ and hence $a_{k+1} = \sqrt{a_k + \sqrt{t}} \leq \sqrt{a_{k+1} + \sqrt{t}} = a_{k+2}$ as well. So we know that this sequence converges exactly when it is bounded...which is all the time. We claim that for all $k \geq 1$ that $a_k < 2t$, the base case being $a_1 = \sqrt{t} < 2t$ which is certainly true when $t \geq 1$. Suppose that $a_k < 2t$ for some $k \geq 1$. Then $\sqrt{t} + a_k < \sqrt{t} + 2t$ and so $a_{k+1} = \sqrt{a_k + \sqrt{t}} < \sqrt{\sqrt{t} + 2t} < \sqrt{4t} = 2\sqrt{t} \leq 2t$. So this sequence has a limit R satisfying $R = \sqrt{t+R}$ or $R^2 - R - t = 0$ and so by the quadratic formula $R = \frac{1 \pm \sqrt{1+4t}}{2}$. Since every term of this sequence is positive we know that $R = \frac{1+\sqrt{1+4t}}{2}$ is the correct limit. Note the fun case when $t = 1$:

$$\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\ldots}}}} = \frac{1+\sqrt{5}}{2},$$

where this last number is the Golden Ratio.

The Monotone Convergence Theorem is used extensively to establish theorems like the comparison tests for infinite series. We recall the definition of a convergent infinite series:

Definition 6.37. Let $(a_k)_{k=0}^{\infty}$. Define the n-th partial sum to be

$s_n := \sum_{k=0}^{n} a_k = a_0 + a_1 + \cdots + a_n$. If the sequence $s_n \to s$ then we write $\sum_{k=0}^{\infty} a_k = s = \lim_{n\to\infty} s_n$ and say that the *series* $\sum a_k$ converges. If $\lim s_n$ does not exists we say that the series $\sum a_k$ diverges.

Example 6.38. Let $a_k = r^k$ where $r \in \mathbb{R}$. We leave it as a simple proof by induction for the reader to show that $(1-r)\sum_{k=0}^{n} r^k = 1 - r^{n+1}$ for any $n \in \mathbb{N}$. If $|r| < 1$ we know that $\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r} \to \frac{1}{1-r}$ and so $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$. We can easily adapt this to get that $\sum_{k=t}^{\infty} ar^k = ar^t \sum_{k=t}^{\infty} r^{k-t} = ar^t \sum_{k'=0}^{\infty} r^{k'} = \frac{ar^t}{1-r}$ when $|r| < 1$. We call this last series the geometric series with first term ar^t and common ratio r.

Theorem 6.39 (Comparison Test)**.** *Suppose $0 \leq a_n \leq b_n$ and that $\sum_{k=0}^{\infty} b_k = B$ converges. Then $\sum_{k=0}^{\infty} a_k$ also converges and $\sum_{k=0}^{\infty} a_k \leq B$.*

Proof. Since $a_n, b_n \geq 0$ we have that the partial sums s_n, t_n respectively are increasing since:

$$s_n \leq s_n + a_{n+1} = s_{n+1}, \quad t_n \leq t_n + b_{n+1} = t_{n+1},$$

and one readily sees that $s_n \leq t_n$. We know that $s_n \leq t_n \leq B$ and so we have that s_n is an increasing sequence that is bounded above by B and so must converge by the Monotone Convergence Theorem to a limit no bigger than B. □

Theorem 6.40. *Suppose that $0 < a_n, b_n$ and that $\frac{b_n}{a_n} \to L$ where $L \neq 0$. Then $\sum_{k=0}^{\infty} a_k$ converges iff $\sum_{k=0}^{\infty} b_k$ converges.*

Proof. Suppose that $\sum a_n$ converges. Then since $\frac{b_n}{a_n} \to L$ there is an $N \in \mathbb{N}$ so that $n \geq N$ implies that $|\frac{b_n}{a_n} - L| < \frac{L}{2}$ or $-\frac{L}{2} < \frac{b_n}{a_n} - L < \frac{L}{2}$ so that $0 < b_n < \frac{3L}{2} a_n$. Since $\sum a_n$ converges $\sum \frac{3L}{2} a_n$ converges and by the Comparison Test $\sum b_n$ converges. The 'iff' comes from the fact that $\frac{a_n}{b_n} \to \frac{1}{L}$ by the Limit Laws. □

Theorem 6.41 (Ratio Test)**.** *Suppose that $a_1, a_2, a_3, \ldots$ is a sequence of positive numbers so that $a_{n+1}/a_n \to L$ with $0 \leq L < 1$. Then $\sum_{n=1}^{\infty} a_n$ converges.*

Proof. Let us take $\epsilon > 0$ so that $L + \epsilon < 1$ (for instance $\epsilon = \frac{1-L}{10}$ or something similar). Since $a_{n+1}/a_n \to L$ there exists $N \in \mathbb{N}$ so that if $n \geq N$ we have $|\frac{a_{n+1}}{a_n} - L| < \epsilon$ and so $\frac{a_{n+1}}{a_n} < \epsilon + L$. We have $a_n = \frac{a_n}{a_{n-1}} \frac{a_{n-1}}{a_{n-2}} \ldots \frac{a_{N+1}}{a_N} a_N < (\epsilon + L)^{n-N} a_N$ for any $n \geq N$ and so $\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{N-1} a_n + \sum_{n=N}^{\infty} a_n \leq \sum_{n=1}^{N-1} a_n + \sum_{n=N}^{\infty} (\epsilon+L)^{n-N} a_N$ where $\sum_{n=N}^{\infty} (\epsilon+L)^{n-N} a_N = a_N \sum_{k=0}^{\infty} (\epsilon + L)^k$ is a convergent geometric series. Therefore by the Comparison Test we have $\sum a_n$ converges. □

We make a note here that a series $\sum a_n$ where $\sum |a_n|$ converges is said to be *absolutely convergent*. We get the following simple corollary of the Comparison Test:

Theorem 6.42. *If a series is absolutely convergent it is convergent.*

Proof. Suppose that $\sum_{n=1}^{\infty} |a_n|$ is convergent. Then since $\forall x \in \mathbb{R}$ we know $0 \leq x + |x| \leq 2|x|$ we know that $\sum_{n=1}^{\infty} a_n + |a_n|$ is convergent by the comparing it to the convergent series $\sum_{n=1}^{\infty} 2|a_n|$ via the Comparison Test. But $\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} a_n + |a_n| - \sum_{n=1}^{\infty} |a_n|$ must also converge being the difference of convergent series (we leave it to the reader to show that the sum and difference of convergent series are convergent). □

Example 6.43. We will show that $\sum_{k=1}^{\infty} \frac{1}{k^2}$ converges. Note that $\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$ and so for $n > 1$ we have $\sum_{k=2}^{n} \frac{1}{k^2} < \sum_{k=2}^{n} \frac{1}{k-1} - \frac{1}{k} = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) = 1 - \frac{1}{n} < 1$ and so $\sum_{k=2}^{n} \frac{1}{k^2}$ is an increasing sequence that is bounded above and so converges to a limit no bigger than 1 by the Monotone Convergence Theorem and so necessarily $\sum_{k=1}^{\infty} \frac{1}{k^2}$ does as well to a limit no bigger than 2. The exact value of this limit is $\frac{\pi^2}{6}$.

Example 6.44. We show that the series $\sum_{k=1}^{\infty} \frac{1}{k}$ does not converge (this series is called the **harmonic series**). We note that $e^x \geq 1 + x$ for all $x \in \mathbb{R}$ since e^x is concave up and $1 + x$ is its tangent line at $x = 0$ (we are assuming quite a bit here). Then for any $n \in \mathbb{N}^+$ we know that $e^{\frac{1}{n}} \geq 1 + \frac{1}{n} = \frac{n+1}{n}$. Assume that $s_n = \sum_{k=1}^{n} \frac{1}{k} \to L$. Then $e^{s_n} = (e^1)(e^{\frac{1}{2}}) \ldots (e^{\frac{1}{n}}) \geq (\frac{2}{1})(\frac{3}{2})(\frac{4}{3}) \ldots (\frac{n+1}{n}) = n + 1$. Since $s_n \to L$ and the function e^x is continuous we have that $e^{s_n} \to e^L$. But e^{s_n} is unbounded so cannot converge. Here is another proof: look at $s_{2^n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \cdots + \frac{1}{2^n} \geq 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \cdots + \frac{1}{2^n}) = 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} = 1 + \frac{n}{2}$. So s_{2^n} grows without bound and hence so does the sequence s_n.

There are many proofs that the harmonic series does not converge, one might think that the first proof above contains a bit of cheating in the sense that it relied on properties of continuous functions as well as the fact that $f(x) = e^x$ is indeed continuous, things we've certainly not established in these notes; we will not discuss continuity of functions rigorously here. But we would like to be rigorous and give a rigorous definition of the number e and the function e^x and furthermore show that the number e is irrational. We first note that for $n > 3$ we have that $n! > n^2$ (provide a simple proof by induction). So $\frac{1}{n!} < \frac{1}{n^2}$ in this range and so $\sum_{k=4}^{\infty} \frac{1}{k!}$ converges by the Comparison Test.

Definition 6.45. Define $e := \sum_{k=0}^{\infty} \frac{1}{k!}$ and for any $x \in \mathbb{R}$ define $e^x := \sum_{k=0}^{\infty} \frac{x^k}{k!}$. We leave it as an exercise to the reader to show that $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ converges for any $x \in \mathbb{R}$ by using the Ratio Test. By earlier comments it's easy to see that $2 < e < 3$.

Theorem 6.46. *e is irrational.*

Proof (due to Fourier). Suppose to the contrary that $e = \frac{a}{b}$ where $a, b \in \mathbb{N}^+$. Define $x := b!\left(e - \sum_{k=0}^{b} \frac{1}{k!}\right) = b!\left(\frac{a}{b} - \sum_{k=0}^{b} \frac{1}{k!}\right) = a(b-1)! - \sum_{k=0}^{b} \frac{b!}{k!}$ which is an integer since $k!|b!$ for every $0 \leq k \leq b$.

We see that $x = b!\left(\sum_{k=0}^{\infty} \frac{1}{k!} - \sum_{k=0}^{b} \frac{1}{k!}\right) = \sum_{k=b+1}^{\infty} \frac{b!}{k!} > 0$. In the sum $\sum_{k=b+1}^{\infty} \frac{b!}{k!}$ each term $\frac{b!}{k!} = \frac{1}{(b+1)(b+2)...(b+(k-b)} \leq \frac{1}{(b+1)^{k-b}}$. So $x = \sum_{k=b+1}^{\infty} \frac{b!}{k!} \leq \sum_{k=b+1}^{\infty} \frac{1}{(b+1)^{k-b}} = \sum_{j=1}^{\infty} \left(\frac{1}{b+1}\right)^j$ where the latter is a convergent geometric series with limit $\left(\frac{1}{b+1}\right)\left(\frac{1}{1-\frac{1}{b+1}}\right) = \frac{1}{b}$. But clearly $b \geq 2$ since $2 < e < 3$ is not an integer, and so we have that $0 < x \leq \frac{1}{2}$. But x is an integer, a contradiction. □

In the above we defined the number e somewhat conveniently so that we could give a proof that it is irrational. There are several other equivalent 'definitions' of the quantity e^x besides $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ some of which we state here:

Theorem 6.47. *The following definitions are equivalent to our definition of e^x as $\sum_{n=0}^{\infty} \frac{x^n}{n!}$:*

1. *$e^x := y$ is the unique number y so that $\int_1^y \frac{dt}{t} = x$.*
2. *$e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$.*
3. *$e^x := z$ is the unique function $z(x)$ so that $z'(x) = z(x)$ and $z(0) = 1$.*

The proofs of these equivalences can be found HERE.

We finish this section with a brief look at the relationship between real numbers and their decimal representations.

Theorem 6.48. *Given $r \in \mathbb{R}$ $\exists d_0 \in \mathbb{Z}$ and a sequence $(d_i)_{i=1}^{\infty}$ where $d_i \in \{0, 1, 2 \ldots, 9\}$ so that $r = \sum_{i=0}^{\infty} d_i 10^{-i}$.*

Proof. Choose $d_0 \in \mathbb{Z}$ so that $d_0 \leq r < d_0 + 1$; d_0 exists by the Archimedean property. Next choose $d_1 \in \{0, 1, 2 \ldots, 9\}$ so that $\frac{d_1}{10} \leq r - d_0 < \frac{d_1+1}{10}$; d_1 exists because $0 \leq 10(r - d_0) < 10$. Note that $0 \leq r - d_0 - \frac{d_1}{10} < \frac{1}{10}$. Choose $d_2 \in \{0, 1, 2 \ldots, 9\}$ so that $d_0 + \frac{d_1}{10} + \frac{d_2}{10^2} \leq r < d_0 + \frac{d_1}{10} + \frac{d_2+1}{10^2}$; d_2 exists because $0 \leq 10^2(r - d_0 - \frac{d_1}{10}) < 10$. Note that $0 \leq r - \sum_{i=0}^{2} d_i 10^{-i} < \frac{1}{10^2}$. Continue defining the sequence d_k recursively so that for each $k \in \mathbb{N}^+$ d_k is chosen so that $\frac{d_k}{10^k} + \sum_{i=0}^{k-1} d_i 10^{-i} \leq r < \frac{d_k+1}{10^k} + \sum_{i=0}^{k-1} d_i 10^{-i}$ and hence $0 \leq r - \sum_{i=0}^{k} d_i 10^{-i} < \frac{1}{10^k}$. This defines a sequence d_k of the required form and we immediately get that $r = \sum_{i=0}^{\infty} d_i 10^{-i}$ by the Squeeze Theorem. □

Note that decimal representations are far from being unique (for instance $0.9999999\ldots$ is equal to 1 and so is $1.0000\ldots$), something we don't get into detail here (we encourage the reader to check out HERE). Note in the above proof that there was nothing special about the number 10, we could have easily adapted the proof and replaced 10 by any base $b \in \mathbb{N}$ where $b \geq 2$ and so that the $d_i \in \{0, 1, 2 \ldots, b - 1\}$.

We've seen that the rational numbers don't satisfy the Completeness Property and that there are infinitely many irrational numbers. In the next section we make sense of what it means to say there are much more real numbers than there are rational numbers but that there are as many rational numbers as there are natural numbers.

7 Cardinality of Sets

We start by defining what it means to say that a set is finite. Define $[[0]] := \emptyset$ and for $n \in \mathbb{N}^+$ define $[[n]] := \{1, 2, 3 \ldots, n\}$.

Definition 7.1. We say that a set X is finite of size $n \in \mathbb{N}$ if there is a bijection from $[[n]]$ to X. In particular we have that $[[n]]$ is finite of size n.

Example 7.2. The set $Y = \{b, l, o, n, d, i, e\}$ is a set of size 7 since the function $f : [[7]] \to Y$ given by $f(1) = b, f(2) = l, f(3) = o, f(4) = n, f(5) = d, f(6) = i, f(7) = e$ is clearly a bijection. In a very strong sense f is simply counting the elements of Y.

What might seem obvious but requires a proof is the fact that if a set X is finite of size n and finite of size m then $m = n$; what is clear is that if there is a bijection $f : [[n]] \to X$ and a bijection $g : [[m]] \to X$ then there is a bijection $g^{-1} \circ f : [[n]] \to [[m]]$. So we need to establish the following:

Theorem 7.3. *If $m \neq n$ are natural numbers, there is no bijection from $[[m]]$ to $[[n]]$.*

Proof. Assume this is false and let $n \in \mathbb{N}$ be the smallest element of the natural numbers where there is an $m \in \mathbb{N}$ with $m \neq n$ so that there is a bijection $f : [[n]] \to [[m]]$. We first note that $m > n$ since if $m < n$ then m would be smaller and satisfy the above condition. We then note that $n \geq 1$ since otherwise f would be a bijection between a non-empty set and the empty set. So $1 \in [[n]]$ and so $f(1) \in [[m]]$. Define a new function $g : [[n]] - \{1\} \to [[m]] - \{f(1)\}$ by $g(t) = f(t)$ for all $t \in [[n]] - \{1\}$. It's easy to see this is also a bijection. Let $h : [[n-1]] \to [[n]] - \{1\}$ be given by $h(t) = t + 1$; it's easy to check that h is a bijection (it has inverse $h^{-1} : [[n]] - \{1\} \to [[n-1]]$ given by $h^{-1}(s) = s - 1$). Lastly let $p : [[m]] - \{f(1)\} \to [[m-1]]$ be given by $p(t) = t$ if $1 \leq t < f(1)$ and $p(t) = t - 1$ if $f(1) < t \leq m$; p is also a bijection. Then set $q = p \circ g \circ h : [[n-1]] \to [[m-1]]$ being a composition of bijections is a bijection with $n - 1 \neq m - 1$ contradicting our assumption that n was the smallest natural number with this property. □

We ask the reader to indeed carefully verify that the functions p, g, h in the last proof are bijections.

Theorem 7.4. *Let X and Y be finite sets of the same size $n \in \mathbb{N}$ and let $f : X \to Y$. Then f is $1-1$ iff f is a surjection iff f is a bijection.*

Proof. We proceed by induction to prove the statement "if a function from a set with n elements to a set with n elements is injective, then it is bijective (where $n \in \mathbb{N}$)," the base case being clear (check that this is true for the empty function). Assume it is true for some fixed but arbitrary $n \in \mathbb{N}$ and let $h : A \to B$ be a $1-1$ function between two sets A, B of size $n + 1$. Let $x \in A$ and consider the function $h' : A - \{x\} \to B - \{h(x)\}$ defined by $h'(a) = h(a)$ for

every $a \in A - \{x\}$. Both $A - \{x\}$ and $B - \{h(x)\}$ have size n (check this!) and h' is certainly $1-1$ and so by our inductive hypothesis is a bijection. Let $y \in B$. If $y = h(x)$ then $y \in im(h)$ otherwise $y \in B - \{h(x)\}$ and since h' is onto there exists $a \in A - \{x\}$ so that $h'(a) = y$. But then $h(a) = h'(a) = y$ and so $y \in im(h)$. So h is onto and hence a bijection and we are done by the PMI.

We now proceed by induction to prove the statement "if a function from a set with n elements to a set with n elements is surjective then it is bijective (where $n \in \mathbb{N}$)," the base case being clear (check that this is true for the empty function). Assume it is true for some fixed but arbitrary $n \in \mathbb{N}$ and let $g : C \to D$ be an onto function between two sets of size $n+1$. Since C has size $n+1$ it is not empty and so $\exists c_0 \in C$. Let $g' : C - \{c_0\} \to D - \{g(c_0)\}$ be defined by $g'(c) = g(c)$ for each $c \in C - \{c_0\}$. Then since $C - \{c_0\}$ and $D - \{g(c_0)\}$ are finite sets of size n (check this!) and certainly g' is still onto we have by our inductive hypothesis that g' is a bijection. Suppose $g(x_1) = g(x_2)$ for some $x_1, x_2 \in C$. Then if neither x_1 nor x_2 is c_0 we would have $g'(x_1) = g(x_1) = g(x_2) = g'(x_2)$ and so $x_1 = x_2$ since g' is $1-1$. Otherwise one of x_1, x_2 is equal to c_0, let's say $x_1 = c_0$. If $x_2 \neq c_0$ then $g'(x_2)$ exists with $g'(x_2) = g(x_2) = g(x_1) = g(c_0)$ which is nonsense. So $x_1 = x_2$ and g is $1-1$ and hence a bijection and we are done by the PMI. □

Going back to our function $f : X \to Y$ assume f is $1-1$. Then by the above we get that f is a bijection and hence onto. If f is onto, then by the above we get that f is a bijection and hence $1-1$.

Theorem 7.5 (Pigeonhole Principle)**.** *If $m > n \geq 1$ are natural numbers then there does not exists a $1-1$ function from $[[m]]$ to $[[n]]$.*

Proof. Assume this is false and let $f : [[m]] \to [[n]]$ be $1-1$. We know that the function $g : [[n]] \to [[m]]$ given by $g(x) = x$ is $1-1$ and so $g \circ f : [[m]] \to [[m]]$ is also $1-1$ and hence a bijection by the previous theorem. But $m \notin im(g \circ f)$ since if $\exists x \in [[m]]$ so that $g(f(x)) = m$ we'd have $f(x) = m$ which makes no sense since the codomain of f doesn't include m. □

For some interesting applications of the seemingly obvious Pigeonhole Principle click HERE.

We make an important note, that if X and Y are finite disjoint sets of size m and n then the size of $X \cup Y$ is $m+n$; if we have bijections $f : [[m]] \to X$ and $g : [[n]] \to Y$ the function $h : X \cup Y \to X \cup Y$ given by $h(a) = f(a)$ if $a \in X$ and $h(a) = g(a)$ if $a \in Y$ is clearly a bijection. This simple fact gives rise to some important counting methods such as the Inclusion-Exclusion Principle:

Theorem 7.6 (Inclusion-Exclusion Principle)**.** *Let X have size n and Y have size m. If $X \cap Y$ has size k then $X \cup Y$ has size $n+m-k$. Moreover if we let $|X|$ be the size of a finite set X then $|X_1 \cup X_2 \cup \cdots \cup X_t|$ equals*

$$\sum_{i=1}^{n} |X_i| - \sum_{1 \leq i \leq j \leq n} |X_i \cap X_j| + \sum_{1 \leq i \leq j \leq k} |X_i \cap X_j \cap X_k| - \cdots + (-1)^{n+1} |X_1 \cap X_2 \cap \cdots \cap X_n|$$

Proof. First note that $X \cup Y = (X - (X \cap Y)) \cup (X \cap Y) \cup (Y - (X \cap Y))$ where the latter three sets are pairwise disjoint. So by the comment preceding the theorem we know that

$$|X \cup Y| = |X - (X \cap Y)| + |X \cap Y| + |Y - (X \cap Y)|.$$

But $|X - (X \cap Y)| = |X| - |X \cap Y|$ (why?) and $|Y - (X \cap Y)| = |Y| - |X \cap Y|$, and so altogether we have

$$|X \cup Y| = (n - k) + k + (m - k) = n + m - k.$$

The general case can be proven a variety of ways, we direct the interested reader to the link HERE □

Definition 7.7. Let X and Y be sets. We say X has *cardinality* less than or equal to Y if there exists a $1-1$ function $f : X \to Y$ and write $|X| \leq |Y|$. We write $|X| = |Y|$ if there is a bijection from X to Y. If there is a $1-1$ function from X to Y but no bijection from X to Y we will write $|X| < |Y|$.

An immediate consequence of the left/right inverse theorems is that $f : X \to Y$ is $1-1$ iff it has a right inverse $g : Y \to X$ so that $g \circ f = id_X$; such a g is necessarily onto (we ask the reader to provide a proof of this), and so we can also write $|X| \leq |Y|$ is there is a surjection $g : Y \to X$.

Example 7.8. Let $n \leq m$ be natural numbers. Then $[[n]] \subseteq [[m]]$ and so the function $\iota : [[n]] \to [[m]]$ given by $\iota(t) = t$ is $1-1$ and so $|[[n]]| \leq |[[m]]|$. If X is a set of size n and Y is a set of size m then there are bijections $f : [[n]] \to X, g : [[m]] \to Y$ and so $g \circ \iota \circ f^{-1} : X \to Y$ is $1-1$ and so $|X| \leq |Y|$. So the above definition of cardinality agrees with our idea of size with respect to finite sets.

Example 7.9. Let $f : \mathbb{Z} \to \mathbb{N}$ be given by

$$f(n) = \begin{cases} 2n & n \geq 0 \\ -2n - 1 & n < 0 \end{cases}$$

and $g : \mathbb{N} \to \mathbb{Z}$ be given by

$$g(n) = \begin{cases} \frac{n}{2} & n \quad even \\ -\frac{n+1}{2} & n \quad odd \end{cases}$$

We leave it to the reader to verify that $g \circ f = id_{\mathbb{Z}}$ and $f \circ g = id_{\mathbb{N}}$ and so both f and g are bijections. So $|\mathbb{N}| = |\mathbb{Z}|$. Note that $\mathbb{N} \subsetneq \mathbb{Z}$ and so just because there are elements of $\mathbb{Z}$ that are not in $\mathbb{N}$ does not mean that the cardinality of $\mathbb{Z}$ is bigger. In many ways sets that have the same cardinality as $\mathbb{N}$ are the smallest infinite sets.

Definition 7.10. If $|X| \leq |\mathbb{N}|$ we say that X is countable; if $|X| = |\mathbb{N}|$ we say that X is countably infinite.

Example 7.11. Let $r : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be given by

$$r(m, n) = 2^m(2n + 1) - 1.$$

We claim that r is a bijection. Let $k \in \mathbb{N}$. If $k = 0$ we have $r(0, 0) = k$ and if $k = 1$ then $r(1, 0) = k$. If k is even then $r(0, \frac{k}{2}) = k$ and if k is odd then $k + 1$ is even and so there is a highest power of two dividing $k + 1$ say $k + 1 = 2^t(2s + 1)$. Then $r(t, s) = k$, and so r is onto. Suppose that $r(m_1, n_1) = r(m_2, n_2)$ or $2^{m_1}(2n_1 + 1) - 1 = 2^{m_2}(2n_2 + 1) - 1$ which implies $2^{m_1-m_2}(2n_1 + 1) = (2n_2 + 1)$ and since the integer on the right hand side is odd we must have $m_1 = m_2$ and so $2n_1 + 1 = 2n_2 + 1$ and hence $n_1 = n_2$. So r is $1 - 1$ and hence a bijection. So $\mathbb{N} \times \mathbb{N}$ is countable. We leave it to the reader to show that $\mathbb{N}^+$ and $\mathbb{Z} \times \mathbb{N}^+$ are also countable.

Example 7.12. Let $d : \mathbb{Z} \times \mathbb{N}^+ \to \mathbb{Q}$ be given by $d(a, b) = \frac{a}{b}$. Then clearly d is onto. Since $\mathbb{Z} \times \mathbb{N}^+$ is countably infinite there is a bijection $p : \mathbb{N} \to \mathbb{Z} \times \mathbb{N}^+$ and so $d \circ p : \mathbb{N} \to \mathbb{Q}$ is also a surjection. So $|\mathbb{Q}| \leq |\mathbb{N}|$ and $\mathbb{Q}$ is countable as well.

We say that a set is *infinite* if it is not a finite set. We show now that the natural numbers have the smallest cardinality of any infinite set.

Theorem 7.13. *Let X be an infinite set. Then $|\mathbb{N}| \leq |X|$.*

Proof. Since X is infinite it is non-empty and so contains some x. Define the following elements of X recursively: $x_0 = x$ and x_{k+1} is an element of $X - \{x_0, x_1, \ldots, x_k\}$, where we know x_{k+1} exists since otherwise $X = \{x_0, x_1, \ldots, x_k\}$ would be finite of size $k + 1$. So this defines a sequence $(x_i)_{i=1}^{\infty}$ of distinct elements of X and a sequence is exactly a function from $\mathbb{N} \to X$. The fact that these elements are distinct implies that this function is $1 - 1$. □

There is a philosophical point that is getting somewhat sidestepped in the above argument, something called the Axiom of Choice; this was also sidestepped much earlier in the lecture notes when we proved that every surjection has a right inverse. But we won't get into this here other than it is well worth a Wikipedia search. We mention another way we can distinguish finite sets from infinite sets (this property is called being *Dedekind infinite*).

Theorem 7.14. *A set X is infinite implies it has the same cardinality as a proper subset of itself.*

Proof. If X is infinite then there is a $1 - 1$ function $f : \mathbb{N} \to X$. Define a function $h : X \to X - \{f(0)\}$ by:

$$h(x) = \begin{cases} f(n + 1) & \exists n \in \mathbb{N}, x = f(n) \\ x & x \notin im(f) \end{cases}$$

We leave the details that this is a bijection as an exercise to the reader, the idea being that h leaves the elements of the set X that are not in $im(f)$ alone and then just sends $h(f(0)) = f(1), h(f(1)) = f(2), \ldots$. □

Lemma 7.15. *If $|X| \leq |Y|$ and $|Y| \leq |Z|$ then $|X| \leq |Z|$.*

Proof. This follows from the fact that the composition of injective functions is injective. □

Suppose we know $|X| \leq |Y|$ and $|Y| \leq |X|$. It's not immediately clear at all that one can conclude that $|X| = |Y|$ since $|X| \leq |Y|$ and $|Y| \leq |X|$ simply means there exists $1-1$ functions $f : X \to Y, g : Y \to X$, these might not be bijections, and it is careful work to see that given this information that a bijection exists. It turns out that there always does exist a bijection given these injections, this is a result known as the Cantor-Bernstein Theorem and we will give a proof of this towards the end of this section. We've seen that an interesting variety of sets are countably infinite: $\mathbb{N}, \mathbb{Z}, \mathbb{N} \times \mathbb{N}, \mathbb{Q}$ for instance. But are there infinite sets that are not countably infinite? Is for instance $\mathbb{R}$ countably infinite? The answer to that is a very profound no. We call an infinite set that is not countable an *uncountable* set.

Lemma 7.16. *Let $a < b$ and $c < d$ be real numbers. Then $|(a, b)| = |(c, d)|$ and $|(a, b)| = |\mathbb{R}|$.*

Proof. $p : (0, 1) \to (c, d)$ by $p(x) = (c - d)x + d$ is clearly a bijection with inverse $p{-}1 : (c, d) \to (0, 1)$ given by $p^{-1}(x) = \frac{x-d}{c-d}$ and so $|(0, 1)| = |(c, d)|$. Likewise $|(0, 1)| = |(a, b)|$ and so $|(a, b)| = |(c, d)|$. The function $t : (-\pi/2, \pi/2) \to \mathbb{R}$ given by $t(x) = \tan(x)$ is a bijection from $(-\pi/2, \pi/2)$ to $\mathbb{R}$ and so $|(0, 1)| = |(-\pi/2, \pi/2)| = |\mathbb{R}|$. □

Theorem 7.17 (Cantor)**.** *$\mathbb{R}$ is uncountable.*

Proof. We show that $(0, 1)$ is uncountable and then the theorem follows from the preceding lemma. Suppose to the contrary that $(0, 1)$ was countably infinite, i.e, there exists a bijection $c : \mathbb{N}^+ \to (0, 1)$. We write out this bijection as a list representing each element of $(0, 1)$ as a decimal; we choose representatives that do not have repeating $9's$, so we choose the representation $0.005000000\ldots$ over $0.0049999\ldots$ and so on:

$$c(1) = 0.d_{11}d_{12}d_{13}d_{14}d_{15}\ldots$$
$$c(2) = 0.d_{21}d_{22}d_{23}d_{24}d_{25}\ldots$$
$$c(3) = 0.d_{31}d_{32}d_{33}d_{34}d_{35}\ldots$$
$$c(4) = 0.d_{41}d_{42}d_{43}d_{44}d_{45}\ldots$$

etc....where d_{ij} is the j-th decimal place of the i-th number on the list. Define a sequence $(a_k)_{k=1}^{\infty}$ by the rule $a_k = 9 - d_{kk}$ for each $k \geq 1$. Note that we have $a_k \neq d_{kk}$ otherwise we would have $a_k + d_{kk} = 2d_{kk} = 9$ which can't happen since 9 is odd.

Let $a = 0.a_1a_2a_3a_4\ldots$ where the k-th decimal place of a is a_k. Then $a \notin im(c)$ since otherwise we would have $0.d_{j1}d_{j2}d_{j3}d_{j4}d_{j5}\cdots = 0.a_1a_2a_3a_4\ldots$ for some $j \in \mathbb{N}^+$ which is impossible since $d_{jj} \neq a_j$ by definition of the sequence a_k. So c is not onto giving us a contradiction. So $(0, 1)$ is uncountable. □

Theorem 7.18. *Let X and Y be countably infinite. Then $X \cup Y$ is countably infinite.*

Proof. Let $Y' = Y - X$ and note that $X \cup Y = X \cup Y'$ and $X \cap Y' = \emptyset$. We show that $X \cup Y'$ is countably infinite. If Y' is finite then we leave it as an exercise to show that $X \cup Y'$ is countably infinite. Otherwise assume that X and Y' are both countably infinite and let $f : X \to \mathbb{N}, g : Y' \to \mathbb{N}$ be bijections. Define $h : X \cup Y' \to \mathbb{N}$ as follows:

$$h(x) = \begin{cases} 2f(x) & x \in X \\ 2g(x) + 1 & x \in Y' \end{cases}$$

which is well-defined since $X \cap Y' = \emptyset$. If $h(x_1) = h(x_2)$ either $x_1, x_2 \in X$ or $x_1, x_2 \in Y'$; either way we would get $x_1 = x_2$ since f and g are $1 - 1$. If $n \in \mathbb{N}$ then if $n = 2k$ for some $k \in \mathbb{N}$ then since f is onto $\exists x \in X$ so that $f(x) = k$. But then $h(x) = 2f(x) = 2k = n$. If $n = 2k + 1$ for some $k \in \mathbb{N}$ then since g is onto $\exists y \in Y'$ so that $g(y) = k$. But then $h(y) = 2g(y) + 1 = 2k + 1 = n$. So h is onto and hence a bijection. □

Theorem 7.19. *The set of irrational numbers $\mathbb{R} - \mathbb{Q}$ is uncountable.*

Proof. Assume to the contrary that $\mathbb{R} - \mathbb{Q}$ is countable. Then $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} - \mathbb{Q})$ would be countable being the union of two countable sets. So $\mathbb{R} - \mathbb{Q}$ is uncountable. □

Given a set X there is always a way to find another set Y so that $|X| < |Y|$. We first look examine how this is done with a finite set. We recall from earlier in the course the definition of the power set of a set: given X we define $P(X) := \{Y \mid Y \subseteq X\}$.

Example 7.20. Let $X = [[4]] = \{1, 2, 3, 4\}$. Then $P([[4]])$ contains the following elements arranged by size:

$$\begin{gathered} \emptyset \qquad \text{size } 0 \\ \{1\}, \{2\}, \{3\}, \{4\} \quad \text{size } 1 \\ \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\} \quad \text{size } 2 \\ \{2,3,4\}, \{1,3,4\}, \{1,2,4\}, \{1,2,3\} \quad \text{size } 3 \\ [[4]] = \{1,2,3,4\} \qquad \text{size } 4 \end{gathered}$$

Altogether we have $1 + 4 + 6 + 4 + 1$ subsets for a total of 2^4 subsets. If we take $[[n]]$ where $n \in \mathbb{N}$ we see that there are $\binom{n}{k}$ subsets of $[[n]]$ of size k and so in total there are

$$\binom{n}{0} + \binom{n}{0} + \cdots + \binom{n}{n} = \sum_{k=0}^{n} \binom{n}{k} = 2^n$$

subsets and so $|P([[n]])| = |[[2^n]]|$. We note here that $n < 2^n$ is true for any $n \in \mathbb{N}$.

We can say a similar thing about every set regardless of whether it is finite or infinite:

Theorem 7.21 (Cantor)**.** *Let X be a set. Then $|X| < |P(X)|$.*

Proof. Clearly $|X| \leq |P(X)|$ since the function $s : X \to P(X)$ given by $s(a) = \{a\}$ is $1-1$: $s(a_1) = s(a_2)$ means $\{a_1\} = \{a_2\}$ and hence $a_1 = a_2$. Assume to the contrary that there is a bijection $c : X \to P(X)$. Define

$$D = \{a \in X \mid a \notin c(a)\}.$$

Then $D \subseteq X$ or $D \in P(X) = im(c)$ and so $\exists a \in X$ so that $c(a) = D$. But then we have

$$a \in D = c(a) \iff a \notin c(a) = D,$$

which is clearly absurd. So there is no such bijection and $|X| < |P(X)|$. □

If we start with $\mathbb{N}$ and keep forming successive power sets we get:

$$|\mathbb{N}| < |P(\mathbb{N})| < |P(P(\mathbb{N}))| < |P(P(P(\mathbb{N})))| < \dots$$

Since $|\mathbb{N}| < |\mathbb{R}|$ it's natural to ask whether $|P(\mathbb{N})| \leq |\mathbb{R}|$ or whether $|\mathbb{R}| \leq |P(\mathbb{N})|$. We will see that $|P(\mathbb{N})| = |\mathbb{R}|$ but we'll need a few tools first.

Theorem 7.22 (Tarski Fixed Point Theorem)**.** *If a function $h : P(X) \to P(X)$ satisfies $A \subseteq B \Longrightarrow h(A) \subseteq h(B)$ then $\exists S \subseteq X$ so that $h(S) = S$.*

Proof. Let $L = \{A \subseteq X \mid A \subseteq h(A)\}$ and let $S = \cup L$. Then any $B \in L$ is a subset of S and so $B \subseteq h(B) \subseteq h(S)$ and we have $S = \cup L \subseteq h(S)$ and then $S \in L$. $h(S) \subseteq h(h(S))$ implies that $h(S) \in L$ and so $h(S) \subseteq S$ or $h(S) = S$ as desired. □

Theorem 7.23 (Cantor-Bernstein Theorem)**.** *If $|X| \leq |Y|$ and $|Y| \leq |X|$ then $X| = |Y|$.*

Proof. Let $f : X \to Y, g : Y \to X$ be $1-1$ and define $h : P(X) \to P(X)$ by

$$h(A) = X - g(Y - f(A)).$$

Suppose $A \subseteq B \subseteq X$. Then $f(A) \subseteq f(B)$ and $Y - f(B) \subseteq Y - f(B)$ and $g(Y - f(B)) \subseteq g(Y - f(A))$ and lastly $X - g(Y - f(A)) \subseteq X - g(Y - f(B))$, so $h(A) \subseteq h(B)$. So by the lemma there is an S so that $S = h(S) = X - g(Y - f(S))$ or $X - S = g(Y - f(S))$. The function $p : X \to Y$ given by

$$p(x) = \begin{cases} f(x) & x \in S \\ g^{-1}(x) & x \in X - S \end{cases}$$

is a bijection since f maps S bijectively onto its image $f(S)$ and g maps $Y - f(S)$ bijectively onto its image $g(Y - f(S)) = X - S$ (i.e. g^{-1} maps $X - S$ bijectively onto its image $Y - f(S)$). □

Theorem 7.24. $|\mathbb{R}| = |P(\mathbb{N})|$.

Proof. Let $t : P(\mathbb{N}) \to \mathbb{R}$ be given by $t(S) = \sum_{n \in S} 10^{-n}$. Then if $t(S_1) = t(S_2)$ these real numbers agree at every place since they consist of decimals with 0's and 1's in exactly the same places and so $S_1 = S_2$ and so t is $1-1$ and $|P(\mathbb{N})| \leq |\mathbb{R}|$.

Since $\mathbb{Q}$ is countable $\mathbb{Q} = \{q_0, q_1, q_2, q_3, \dots\}$. Define $s : \mathbb{R} \to P(\mathbb{N})$ be given by

$$s(x) = \{n \in \mathbb{N} \mid q_n < x\}.$$

If $s(x_1) = s(x_2)$ then any $q_i < x_1$ satisfies $q_i < x_2$ and vice a versa and so $x_1 = x_2$. So $|\mathbb{R}| \leq |P(\mathbb{N})|$ and so by the Cantor-Bernstein Theorem we have $|\mathbb{R}| = |P(\mathbb{N})|$. □

Earlier in the course notes we looked at an example of a partial order on the set of infinite binary sequences $B = \{f \mid f : \mathbb{N} \to \{0,1\}\}$. We look now to show that this set has the same cardinality as $\mathbb{R}$.

Theorem 7.25. $|B| = |\mathbb{R}| = |P(\mathbb{N})|$.

Proof. For each $A \subseteq N$ define a function $\chi_A : \mathbb{N} \to \{0,1\}$ as follows:

$$\chi_A(n) = \begin{cases} 1 & n \in A \\ 0 & n \notin A \end{cases}$$

Let $\alpha : P(\mathbb{N}) \to B$ be given by $\alpha(A) = \chi_A$. Define $\beta : B \to P(\mathbb{N})$ be given by $\beta(f) = \{n \in \mathbb{N} \mid f(n) = 1\}$. Then it's simple to verify that $\alpha \circ \beta = id_B$ and $\beta \circ \alpha = id_{P(\mathbb{N})}$ and so both are bijections. □

There was absolutely nothing special about the natural numbers in this last theorem, we could replace $\mathbb{N}$ with any set and still have a theorem:

Corollary 7.25.1. *Let X be any set and let $2^X = \{f \mid f : X \to \{0,1\}\}$ be the set of all functions from X to $\{0,1\}$. Then $|2^X| = |P(X)|$.*

Note that when X is finite this is already something we've observed via the Binomial Theorem.

We can think of the rational numbers $\mathbb{Q}$ as the set of all solutions to linear equations with integer coefficients, that $\mathbb{Q} = \{x \in \mathbb{R} \mid \exists a, b \in \mathbb{Z}\ (ax + b = 0)\}$. If we look at all real numbers that show up as roots of polynomial equations with integer coefficients we get what are called the *real algebraic numbers*.

Definition 7.26. Let $\mathbb{K} \subseteq \mathbb{R}$ be the following subset:

$$\{x \in \mathbb{R} \mid \exists a_i \in \mathbb{Z}, a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0\},$$

and call the set $\mathbb{K}$ the (real) algebraic numbers. The set $\mathbb{R} - \mathbb{K}$ is called the set of (real) transcendental numbers.

Clearly we have $\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{K} \subseteq \mathbb{R}$.

Example 7.27. This might seem like a somewhat esoteric definition but $\mathbb{K}$ contains a lot of numbers we are already very familiar with like $\sqrt{5}$ since it is a solution to $x^2 - 5 = 0$, $\sqrt[11]{32}$ since it is a solution to $x^{11} - 32 = 0$, and $\sqrt{2} + \sqrt{3}$. It might not be clear that the last quantity is an algebraic number even though we can see that $\sqrt{2}$ and $\sqrt{3}$ are algebraic, so we find an appropriate polynomial by being pretty algebraic: if $r = \sqrt{2} + \sqrt{3}$ then $r^2 = 5 + 2\sqrt{6}$ and so $(r^2 - 5)^2 = 24$ or $r^4 - 10r^2 + 1 = 0$ and so one polynomial with integer coefficients that has $\sqrt{2} + \sqrt{3}$ as a root of $x^4 - 10x^2 + 1$. A more complicated example is that $\cos(\frac{\pi}{7})$ is algebraic; it is a root of $8x^3 - 4x^2 - 4x + 1$. There are lots of algebraic numbers that can't be written in terms of radicals, for instance the real root of $x^5 - x + 1 = 0$, although this is extremely non-trivial to show. In fact $\mathbb{K}$ is an ordered field, a fact that we won't show in these notes but you can read about HERE.

What is not immediately clear is that there are transcendental numbers and it is beyond the scope of these notes to actual prove that a given number is transcendental. It has been shown that π and e are both transcendental. The first explicit transcendental numbers were written down and proven to be transcendental numbers by Liouville in 1844; the interested reader can read about them HERE. In particular Liouville showed that the number $\sum_{k=0}^{\infty} 10^{-k!}$ is transcendental, an elementary proof of which can be found HERE. We will however show that there exists countably many algebraic numbers and hence there exists uncountably many transcendental numbers, a result due to Cantor in 1874. We need a few more tools regarding countable sets in order to do this.

Let $C = \{A_0, A_1, A_2, A_3, \dots\}$ be a countable (finite or countably infinite) collection of countable (finite or countably infinite) sets. We recall that $\cup C = \{x \,|\, \exists A_k, x \in A_k\}$ is the union of all of the sets in C. For example if $C = \{\{1,2,3\},\{3,4,5\},\{5,6,7\}\}$ then $\cup C = \{1,2,3,4,5,6,7\}$.

Lemma 7.28. *Given any countable collection of sets*
$C = \{A_0, A_1, A_2, A_3, \dots\}$ *we can find another countable collection of pairwise disjoint sets* $C' = \{B_0, B_1, B_2, B_3, \dots\}$ *so that* $\cup C = \cup C'$.

Proof. We define the sequence $B_0, B_1, \dots$ recursively as follows:
Set $B_0 := A_0$; $B_1 := A_1 - A_0$; $B_2 := A_2 - (A_0 \cup A_1)$ $B_3 := A_3 - (A_0 \cup A_1 \cup A_2)$ and for $n \geq 1$ we set $B_n := A_n - (A_0 \cup A_1 \cup \cdots \cup A_{n-1}) \subseteq A_n$. Let $C' = \{B_0, B_1, \dots\}$. Suppose $s < t$ are natural numbers. Then $x \in B_s$ implies that $x \in A_s$ and so $x \notin B_t = A_t - (A_0 \cup A_1 \cup \dots A_{t-1})$ since A_s is in the list $A_0, A_1, \dots A_{t-1}$. So $B_s \cap B_t = \emptyset$. Since $B_k \subseteq A_k$ we get immediately that $\cup C' \subseteq \cup C$. Let $x \in \cup C$ so that that $\exists n$ so that $x \in A_n$. Pick n_0 to be the smallest natural number so that $x \in A_{n_0}$. Then $x \in B_{n_0}$ since $x \notin A_k$ for $k < n_0$ and so $x \in \cup C'$ and $\cup C' = \cup C$. □

For example if $C = \{\{1,2,3\},\{3,4,5\},\{5,6,7\}\}$ where $A_0 = \{1,2,3\}$, $A_1 = \{3,4,5\}, A_2 = \{5,6,7\}$ then $C' = \{\{1,2,3\},\{4,5\},\{6,7\}\}$. We use this to establish the following generalization that the union of two countable sets is countable:

Theorem 7.29. *Suppose $C = \{A_0, A_1, A_2, A_3, \dots\}$ is a countable collection of countable sets. Then $\cup C$ is countable.*

Proof. If we apply the construction of the collection C' from C as in the lemma, each $B_k \subseteq A_k$ where now we are assuming A_k is countable. Hence the B_k are countable as well (why?). Since $\cup C' = \cup C$ we show that $\cup C'$ is countable. List the prime numbers by $p_0 = 2, p_1 = 3, p_2 = 5, \dots$. Since each B_k is countable there is by definition a $1-1$ function $f_k : B_k \to \mathbb{N}$. Define $f : C' \to \mathbb{N}$ as follows: if $x \in B_k$ then set $f(x) = p_k^{f_k(x)+1}$. Since the B_k are disjoint an $x \in C'$ can only belong to exactly one of them so this is a well defined function. To show f is $1-1$ let $x \neq y$ be elements of C'. Then if $x, y \in B_m$ for some m then we have $f_m(x) \neq f_m(y)$ since f_m is $1-1$ and so $f(x) \neq f(y)$; if $x \in B_m$ and $y \in B_n$ with $n \neq m$ then $f(x)$ is a positive power of p_m while $f(y)$ is a positive power of p_n, and these are different by the Fundamental Theorem of Arithmetic. So f is $1-1$ and $\cup C' = \cup C$ is countable. □

We are now ready to prove that the set of algebraic numbers is countable; it is an immediate consequence that the set of transcendental numbers is uncountable since $\mathbb{R}$ being uncountable is not the union of two countable sets. We will assume as true that a real polynomial of degree n has at most n real roots.

Theorem 7.30. *The set $\mathbb{K}$ of algebraic numbers is countable.*

Proof. For each $k \in \mathbb{N}$ set $P_k := \{p = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \,|\, |a_n| + |a_{n_1}| + \cdots + |a_0| + n = k, a_i \in \mathbb{Z}\}$, i.e., the set of all polynomials with integer coefficients so that the sum of the degree and the absolute value of the coefficients is exactly k. Each P_k is finite (why?) and hence countable. Let $R_k := \{r \in \mathbb{R} \,|\, p(r) = 0, p \in P_k\}$ be the set of all real roots of polynomials in P_k which since each polynomial in this set has finitely many roots must also be a finite set and hence countable. Let $C = \{R_0, R_1, R_2 \dots\}$ and observe that $\mathbb{K} = \cup C$ is a countable union of countable sets and so is countable. □