Combinatorics and Number Theory

Joseph R. Mileti

April 30, 2020

Contents

Chapter 1

Introduction

1.1 Sets, Set Construction, and Subsets

Sets and Set Construction

We begin by reviewing the fundamental structure that mathematicians use to package objects together.

Definition 1.1.1. *A* set *is a collection of elements without regard to repetition or order.*

Intuitively, a set is a box where the only thing that matters are the objects that are inside it, and furthermore the box does not have more than one of any given object. We use $\{$ and $\}$ as delimiters for sets. For example, $\{3, 5\}$ is a set with two elements. Since all that matters are the elements, we define two sets to be equal if they have the same elements, regardless of how the sets themselves are defined or described.

Definition 1.1.2. *Given two sets A and B, we say that $A = B$ if A and B have exactly the same elements.*

Since only the actual elements matter, not their order, we have $\{3, 7\} = \{7, 3\}$ and $\{1, 2, 3\} = \{3, 1, 2\}$. Also, although we typically would not even write something like $\{2, 5, 5\}$, if we choose to do so then we would have $\{2, 5, 5\} = \{2, 5\}$ because both have the same elements, namely 2 and 5.

Notation 1.1.3. *Given an object x and a set A, we write $x \in A$ to mean that x is an element of A, and we write $x \notin A$ to mean that x is not an element of A.*

For example, we have $2 \in \{2, 5\}$ and $3 \notin \{2, 5\}$. Since sets are mathematical objects, they may be elements of other sets. For example, we can form the set $S = \{1, \{2, 3\}\}$. Notice that we have $1 \in S$ and $\{2, 3\} \in S$, but $2 \notin S$ and $3 \notin S$. As a result, S has only 2 elements, namely 1 and $\{2, 3\}$. Thinking of a set as a box, one element of S is the number 1, and the other is a different box (which happens to have the two elements 2 and 3 inside it).

The empty set is the unique set with no elements. We can write it as $\{\}$, but instead we typically denote it by $\emptyset$. There is only *one* empty set, because if both A and B have no elements, then they have exactly the same elements for vacuous reasons, and hence $A = B$. Notice that $\{\emptyset\}$ does not equal $\emptyset$. After all, $\{\emptyset\}$ has one element! You can think of $\{\emptyset\}$ as a box that has one empty box inside it.

Notice that sets can be either finite or infinite. At this point, our standard examples of infinite sets are the various universes of numbers:

- $\mathbb{N} = \{0, 1, 2, 3, \dots\}$.
- $\mathbb{N}^+ = \{1, 2, 3, \dots\}$.
- $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$.

- $\mathbb{Q}$ is the set of rational numbers.
- $\mathbb{R}$ is the set of real numbers.

Beyond these fundamental sets, there are various ways to define new sets. In some cases, we can simply list the elements as we did above. Although this often works for small finite sets, it is almost never a good idea to list the elements of a set with 20 or more elements, and it rarely works for infinite sets (unless there is an obvious pattern like $\{5, 10, 15, 20, \dots\}$). One of the standard ways to define a set S is to carve it out of some bigger set A by describing a certain property that may or may not be satisfied by an element of A. For example, we could define

$$S = \{n \in \mathbb{N} : 5 < n < 13\}.$$

We read this line by saying that S is defined to be the set of all $n \in \mathbb{N}$ such that $5 < n < 13$. Thus, in this case, we are taking $A = \mathbb{N}$, and forming a set S by carving out those elements of A that satisfy the condition that $5 < n < 13$. In other words, think about going through each of element n, checking if $5 < n < 13$ is a true statement, and collecting those $n \in \mathbb{N}$ that make it true into a set that we call S. In more simple terms, we can also describe S as follows:

$$S = \{6, 7, 8, 9, 10, 11, 12\}.$$

It is important that we put the "$\mathbb{N}$" in the above description, because if we wrote $\{n : 5 < n < 13\}$ then it would be unclear what n we should consider. For example, should $\frac{11}{2}$ be in this set? How about $\sqrt{17}$? Sometimes the "universe" of numbers (or other mathematical objects) that we are working within is clear, but typically it is best to write the global set that we are picking elements from in order to avoid such ambiguity. Notice that when we define a set, there is no guarantee that it has any elements. For example, $\{q \in \mathbb{N} : q^2 = 2\} = \emptyset$ because $\sqrt{2}$ is irrational (we will prove this, and much more general facts, later). Keep in mind that we can also use words in our description of sets, such as $\{n \in \mathbb{N} : n \text{ is an even prime}\}$. As mentioned above, two sets that have quite different descriptions can be equal. For example, we have

$$\{n \in \mathbb{N} : n \text{ is an even prime}\} = \{n \in \mathbb{N} : 3 < n^2 < 8\}$$

because both sets equal $\{2\}$. Always remember the structure of sets formed in this way. We write

$$\{x \in A : \mathsf{P}(x)\}$$

where A is a known set and $\mathsf{P}(x)$ is a "property" such that given a particular $y \in A$, the statement $\mathsf{P}(y)$ is either true or false.

Another way to describe a set is through a "parametric" description. Rather than carving out a certain subset of a given set by describing a property that the elements must satisfy, we can instead form all the elements one obtains by varying a value through a particular set. For example, consider the following description of a set:

$$S = \{3x^2 + 1 : x \in \mathbb{R}\}.$$

Although the notation looks quite similar to the above (in both case we have curly braces, with a : in the middle), this set is described differently. Notice that instead of having a set that elements are coming from on the left of the colon, we now have a set that elements are coming from on the right. Furthermore, we now have a formula on the left rather than a property on the right. The difference is that for a property, when we plug in an element from the given set, we either obtain a true or false value, but that isn't the case for a formula like $3x^2 + 1$. The idea here is that instead of carving out a subset of $\mathbb{R}$ by using a property (i.e. taking those elements that make the property *true*), we let x vary through all real numbers, plug each of these real numbers x into $3x^2 + 1$, and form the set of all possible outputs. For example, we have $4 \in S$ because $4 = 3 \cdot 1^2 + 1$. In other words, when $x = 1$, the left hand side gives the value 4, so we should put $4 \in S$. Notice also that $4 = 3 \cdot (-1)^2 + 1$, so we can also see that $4 \in S$ because of the "witness" -1. Of course, we are forming a set, so we do not repeat the number 4. We also have $1 \in S$ because $1 = 3 \cdot 0^2 + 1$, and we have $76 \in S$ because $76 = 3 \cdot 5^2 + 1$. Notice also that $7 \in S$ because $7 = 3 \cdot (\sqrt{2})^2 + 1$.

In a general parametric set description, we will have a set A and a *function* $f(x)$ that allows inputs from A, and we write

$$\{f(x) : x \in A\}$$

for the set of all possible outputs of the function as we vary the inputs through the set A. We will discuss the general definition of a function in the next section, but for the moment you can think of them as given by formulas.

Now it is possible and indeed straightforward to turn any parametric description of a set into one where we carve out a subset by a property. In our case of $S = \{3x^2 + 1 : x \in \mathbb{R}\}$ above, we can alternatively write it as

$$S = \{y \in \mathbb{R} : \text{There exists } x \in \mathbb{R} \text{ with } y = 3x^2 + 1\}.$$

Notice how we flipped the way we described the set by introducing a "there exists" quantifier in order to form a property. This is always possible for a parametric description. For example, we have

$$\{5n + 4 : n \in \mathbb{N}\} = \{m \in \mathbb{N} : \text{There exists } n \in \mathbb{N} \text{ with } m = 5n + 4\}.$$

Thus, these parametric descriptions are not essentially new ways to describe sets, but they can often be more concise and clear.

By the way, we can use multiple parameters in our description. For example, consider the set

$$S = \{18m + 33n : m, n \in \mathbb{Z}\}.$$

Now we are simply letting m and n vary through all possible values in $\mathbb{Z}$ and collecting all of the values $18m + 33n$ that result. For example, we have $15 \in S$ because $15 = 18 \cdot (-1) + 33 \cdot 1$. We also have $102 \in S$ because $102 = 18 \cdot 2 + 33 \cdot 2$. Notice that we are varying m and n independently, so they might take different values, or the same value (as in the case of $m = n = 2$). Don't be fooled by the fact that we used different letters! As above, we can flip this description around by writing

$$S = \{k \in \mathbb{Z} : \text{There exists } m, n \in \mathbb{Z} \text{ with } k = 18m + 33n\}.$$

Subsets and Set Equality

Definition 1.1.4. *Given two sets A and B, we write $A \subseteq B$ to mean that every element of A is an element of B. More formally, $A \subseteq B$ means that for all x, if $x \in A$, then $x \in B$.*

Written more succinctly, $A \subseteq B$ means that for all $a \in A$, we have that $a \in B$. To prove that $A \subseteq B$, one takes a completely arbitrary $a \in A$, and argues that $a \in B$. For example, let $A = \{6n : n \in \mathbb{Z}\}$ and let $B = \{2n : n \in \mathbb{Z}\}$. Since both of these sets are infinite, we can't show that $A \subseteq B$ by taking each element of A in turn and showing that it is an element of B. Instead, we take an *arbitrary* $a \in A$, and show that $a \in B$. Here's the proof.

Proposition 1.1.5. *Let $A = \{6n : n \in \mathbb{Z}\}$ and $B = \{2n : n \in \mathbb{Z}\}$. We have $A \subseteq B$.*

Proof. Let $a \in A$ be arbitrary. By definition of A, this means that we can fix an $m \in \mathbb{Z}$ with $a = 6m$. Notice then that $a = 2 \cdot (3m)$. Since $3m \in \mathbb{Z}$, it follows that $a \in B$. Since $a \in A$ we arbitrary, we conclude that $A \subseteq B$. $\square$

As usual, pause to make sure that you understand the logic of the argument above. First, we took an arbitrary element a from the set A. Now since $A = \{6n : n \in \mathbb{Z}\}$ and this is a parametric description with an implicit "there exists" quantifier, there must be one fixed integer value of n that puts a into the set A.

In our proof, we chose to call that one fixed integer m. Now in order to show that $a \in B$, we need to exhibit a $k \in \mathbb{Z}$ with $a = 2k$. In order to do this, we hope to manipulate $a = 6m$ to introduce a 2, and ensure that the element we are multiplying by 2 is an integer.

What would go wrong if we tried to prove that $B \subseteq A$? Let's try it. Let $b \in B$ be arbitrary. Since $b \in B$, we can fix $m \in \mathbb{Z}$ with $b = 2m$. Now our goal is to try to prove that we can find an $n \in \mathbb{Z}$ with $b = 6n$. It's not obvious how to obtain a 6 from that 2, but we can try to force a 6 in the following way. Since $b = 2m$ and $2 = \frac{6}{3}$, we can write $b = 6 \cdot \frac{m}{3}$. We have indeed found a number n such that $b = 6n$, but we have not checked that this n is an integer. In general, dividing an integer by 3 does not result in an integer, so this argument currently has a hole in it.

Although that argument has a problem, we can not immediately conclude that $B \not\subseteq A$. Our failure to find an argument does not mean that an argument does not exist. So how can we show that $B \not\subseteq A$? All that we need to do is find just *one example* of an element of B that is not an element of A (because the negation of the "for all" statement $A \subseteq B$ is a "there exists" statement). We choose 2 as our example. However, we need to convince everybody that this choice works. So let's do it! First, notice that $2 = 2 \cdot 1$, so $2 \in B$ because $1 \in \mathbb{Z}$. We now need to show that $2 \notin A$, and we'll do this using a proof by contradiction. Suppose instead that $2 \in A$. Then, by definition, we can fix an $m \in \mathbb{Z}$ with $2 = 6m$. We then have that $m = \frac{2}{6} = \frac{1}{3}$. However, this a contradiction because $\frac{1}{3} \notin \mathbb{Z}$. Since our assumption that $2 \in A$ led to a contradiction, we conclude that $2 \notin A$. We found an example of an element that is in B but not in A, so we conclude that $B \not\subseteq A$.

Recall that two sets A and B are defined to be equal if they have the same elements. Therefore, we have $A = B$ exactly when both $A \subseteq B$ and $B \subseteq A$ are true. Thus, given two sets A and B, we can prove that $A = B$ by performing two proofs like the one above. Such a strategy is called a *double containment* proof. We give an example of such an argument now.

Proposition 1.1.6. *Let $A = \{7n - 3 : n \in \mathbb{Z}\}$ and $B = \{7n + 11 : n \in \mathbb{Z}\}$. We have $A = B$.*

Proof. We prove that $A = B$ by showing that both $A \subseteq B$ and also that $B \subseteq A$.

- We first show that $A \subseteq B$. Let $a \in A$ be arbitrary. By definition of A, we can fix an $m \in \mathbb{Z}$ with $a = 7m - 3$. Notice that

$$\begin{aligned} a &= 7m - 3 \\ &= 7m - 14 + 11 \\ &= 7(m - 2) + 11. \end{aligned}$$

  Now $m - 2 \in \mathbb{Z}$ because $m \in \mathbb{Z}$, so it follows that $a \in B$. Since $a \in A$ was arbitrary, we conclude that $A \subseteq B$.

- We now show that $B \subseteq A$. Let $b \in B$ be arbitrary. By definition of B, we can fix an $m \in \mathbb{Z}$ with $a = 7m + 11$. Notice that

$$\begin{aligned} a &= 7m + 11 \\ &= 7m + 14 - 3 \\ &= 7(m + 2) - 3. \end{aligned}$$

  Now $m + 2 \in \mathbb{Z}$ because $m \in \mathbb{Z}$, so it follows that $a \in B$. Since $a \in A$ was arbitrary, we conclude that $A \subseteq B$.

We have shown that both $A \subseteq B$ and $B \subseteq A$ are true, so it follows that $A = B$. $\square$

Here is a more interesting example. Consider the set

$$S = \{9m + 15n : m, n \in \mathbb{Z}\}.$$

For example, we have $9 \in S$ because $9 = 9 \cdot 1 + 15 \cdot 0$. We also have $3 \in S$ because $3 = 9 \cdot 2 + 15 \cdot (-1)$ (or alternatively because $3 = 9 \cdot (-3) + 15 \cdot 2$). We can always generate new values of S by simply plugging in values for m and n, but is there another way to describe the elements of S in an easier way? We now show that an integer is in S exactly when it is a multiple of 3.

Proposition 1.1.7. *We have* $\{9m + 15n : m, n \in \mathbb{Z}\} = \{3m : m \in \mathbb{Z}\}$.

Proof. We give a double containment proof.

- We first show that $\{9m + 15n : m, n \in \mathbb{Z}\} \subseteq \{3m : m \in \mathbb{Z}\}$. Let $a \in \{9m + 15n : m, n \in \mathbb{Z}\}$ be arbitrary. By definition, we can fix $k, \ell \in \mathbb{Z}$ with $a = 9k + 15\ell$. Notice that

$$\begin{aligned} a &= 9k + 15\ell \\ &= 3 \cdot (3k + 5\ell). \end{aligned}$$

  Now $3k+5\ell \in \mathbb{Z}$ because $k, \ell \in \mathbb{Z}$, so it follows that $a \in \{3m : m \in \mathbb{Z}\}$. Since $a \in \{9m+15n : m, n \in \mathbb{Z}\}$ was arbitrary, we conclude that $\{9m + 15n : m, n \in \mathbb{Z}\} \subseteq \{3m : m \in \mathbb{Z}\}$.

- We now show that $\{3m : m \in \mathbb{Z}\} \subseteq \{9m + 15n : m, n \in \mathbb{Z}\}$. Let $a \in \{3m : m \in \mathbb{Z}\}$ be arbitrary. By definition, we can fix $k \in \mathbb{Z}$ with $a = 3k$. Notice that

$$\begin{aligned} a &= 3k \\ &= (9 \cdot (-3) + 15 \cdot 2) \cdot k \\ &= 9 \cdot (-3k) + 15 \cdot 2k. \end{aligned}$$

  Now $-3k, 2k \in \mathbb{Z}$ because $k \in \mathbb{Z}$, so it follows that $a \in \{9m+15n : m, n \in \mathbb{Z}\}$. Since $a \in \{3m : m \in \mathbb{Z}\}$ was arbitrary, we conclude that $\{3m : m \in \mathbb{Z}\} \subseteq \{9m + 15n : m, n \in \mathbb{Z}\}$.

We have shown that both $\{9m+15n : m, n \in \mathbb{Z}\} \subseteq \{3m : n \in \mathbb{Z}\}$ and $\{3m : m \in \mathbb{Z}\} \subseteq \{9m+15n : m, n \in \mathbb{Z}\}$ are true, so it follows that $\{9m + 15n : m, n \in \mathbb{Z}\} = \{3m : m \in \mathbb{Z}\}$. $\square$

Ordered Pairs and Sequences

In contrast to sets, we define *ordered pairs* in such a way that order and repetition *do* matter. We denote an ordered pair using normal parentheses rather than curly braces. For example, we let $(2, 5)$ be the ordered pair whose first element is 2 and whose second element is 5. Notice that we have $(2, 5) \neq (5, 2)$ despite the fact that $\{2, 5\} = \{5, 2\}$. Make sure to keep a clear distinction between the ordered pair $(2, 5)$ and the set $\{2, 5\}$. We *do* allow the possibility of an ordered pair such as $(2, 2)$, and here the repetition of 2's is meaningful. Furthermore, we do not use $\in$ in ordered pairs, so we would **not** write $2 \in (2, 5)$. We'll talk about ways to refer to the two elements of an ordered pair later.

We can generalize ordered pairs to the possibility of having more than 2 elements. In this case, we have an ordered list of n elements, like $(5, 4, 5, -2)$. We call such an object an *n-tuple*, a *list* with n elements, or a finite *sequence* of length n. Thus, for example, we could call $(5, 4, 5, -2)$ a 4-tuple. It is also possible to have infinite sequences (i.e. infinite lists), but we will wait to discuss these until the time comes.

Operations on Sets and Sequences

Aside from listing elements, carving out subsets of a given set using a given property, and giving a parametric description (which as mentioned above is just a special case of the previous type), there are other ways to build sets.

Definition 1.1.8. *Given two sets A and B, we define $A \cup B$ to be the set consisting of those elements that are in A or B (or both). In other words, we define*

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

We call this set the union *of A and B.*

Here, as in mathematics generally, we use *or* to mean "inclusive or". In other words, if x is an element of both A and B, then we still put x into $A \cup B$. Here are a few examples (we leave the proofs of the latter results until we have more theory):

- $\{1, 2, 7\} \cup \{4, 9\} = \{1, 2, 4, 7, 9\}$.
- $\{1, 2, 3\} \cup \{2, 3, 5\} = \{1, 2, 3, 5\}$.
- $\{2n : n \in \mathbb{N}\} \cup \{2n + 1 : n \in \mathbb{N}\} = \mathbb{N}$.
- $\{2n : n \in \mathbb{N}^+\} \cup \{2n + 1 : n \in \mathbb{N}^+\} = \{2, 3, 4, \dots\}$.
- $\{2n : n \in \mathbb{N}^+\} \cup \{2n - 1 : n \in \mathbb{N}^+\} = \{1, 2, 3, 4, \dots\} = \mathbb{N}^+$.
- $A \cup \emptyset = A$ for every set A.

Definition 1.1.9. *Given two sets A and B, we define $A \cap B$ to be the set consisting of those elements that are in both of A and B. In other words, we define*

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

We call this set the intersection *of A and B.*

Here are a few examples (again we leave some proofs until later):

- $\{1, 2, 7\} \cap \{4, 9\} = \emptyset$.
- $\{1, 2, 3\} \cap \{2, 3, 5\} = \{2, 3\}$.
- $\{1, \{2, 3\}\} \cap \{1, 2, 3\} = \{1\}$.
- $\{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\} = \{6n : n \in \mathbb{Z}\}$.
- $\{3n + 1 : n \in \mathbb{N}^+\} \cap \{3n + 2 : n \in \mathbb{N}^+\} = \emptyset$.
- $A \cap \emptyset = \emptyset$ for every set A.

Definition 1.1.10. *Given two sets A and B, we define $A \backslash B$ to be the set consisting of those elements that are in A, but not in B. In other words, we define*

$$A \backslash B = \{x : x \in A \text{ and } x \notin B\}.$$

We call this set the (relative) complement *of B (in A).*

In many cases where we consider $A\backslash B$, we will have that $B \subseteq A$, but we will occasionally use it even when $B \not\subseteq A$. Here are a few examples:

- $\{5,6,7,8,9\}\backslash\{5,6,8\} = \{7,9\}$.
- $\{1,2,7\}\backslash\{4,9\} = \{1,2,7\}$.
- $\{1,2,3\}\backslash\{2,3,5\} = \{1\}$.
- $\{2n : n \in \mathbb{Z}\}\backslash\{4n : n \in \mathbb{Z}\} = \{4n+2 : n \in \mathbb{Z}\}$.
- $A\backslash\emptyset = A$ for every set A.
- $A\backslash A = \emptyset$ for every set A.

Definition 1.1.11. *Given two sets A and B, we let $A \times B$ be the set of all ordered pairs (a,b) such that $a \in A$ and $b \in B$, and we call this set the* Cartesian product *of A and B.*

For example, we have

$$\{1,2,3\} \times \{6,8\} = \{(1,6),(1,8),(2,6),(2,8),(3,6),(3,8)\}$$

and

$$\mathbb{N} \times \mathbb{N} = \{(0,0),(0,1),(1,0),(2,0),\dots,(4,7),\dots\}.$$

Notice that elements of $\mathbb{R} \times \mathbb{R}$ correspond to points in the plane.

We can also generalize the concept of a Cartesian product to more than 2 sets. If we are given n sets $A_1, A_2, \dots, A_n$, we let $A_1 \times A_2 \times \cdots \times A_n$ be the set of all n-tuples $(a_1, a_2, \dots, a_n)$ such that $a_i \in A_i$ for each i. For example, we have

$$\{1,2\} \times \{3\} \times \{4,5\} = \{(1,3,4),(1,3,5),(2,3,4),(2,3,5)\}$$

In the special case when $A_1, A_2, \dots, A_n$ are all the same set A, we use the notation A^n to denote the set $A \times A \times \cdots \times A$ (where we have n copies of A). Thus, A^n is the set of all finite sequences of elements of A of length n. For example, $\{0,1\}^n$ is the set of all finite sequences of 0's and 1's of length n. Notice that this notation fits in with the notation $\mathbb{R}^n$ that we are used to in Calculus and Linear Algebra.

Definition 1.1.12. *Given a set A, we let $\mathcal{P}(A)$ be the set of all subsets of A, and we call $\mathcal{P}(A)$ the* power set *of A.*

For example, we have

$$\mathcal{P}(\{1,2\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$$

and

$$\mathcal{P}(\{4,5,7\}) = \{\emptyset, \{4\}, \{5\}, \{7\}, \{4,5\}, \{4,7\}, \{5,7\}, \{4,5,7\}\}.$$

Notice that it is can be tricky to write out the power set of even small finite sets. We'll see ways to both generate and count the number of elements of $\mathcal{P}(A)$ for a given set A a bit later.

Definition 1.1.13. *Given a set A, we let A^* be the set of all finite sequences of elements of A of any length, including the empty sequence (the unique sequence of length 0).*

Thus, for example, the set $\{0,1\}^*$ is the set of all finite sequences of 0's and 1's. If we use λ to denote the empty sequence and write things like 010 in place of the more precise $(0,1,0)$, then we have

$$\{0,1\}^* = \{\lambda, 0, 1, 00, 01, 10, 11, 000, 001, \dots\}.$$

Notice that if $A \neq \emptyset$, then A^* is an infinite set.

Definition 1.1.14. *Given two finite sequences σ and τ, we let $\sigma\tau$ be the* concatenation *of σ and τ, i.e. if $\sigma = (a_1, a_2, \dots, a_m)$ and $\tau = (b_1, b_2, \dots, b_n)$, then $\sigma\tau = (a_1, a_2, \dots, a_m, b_1, b_2, \dots, b_n)$.*

1.2 The Cardinality of Sets

We will spend a significant amount of time trying to count the number of elements in certain sets. For now, we will study some simple properties that will eventually become extremely useful when employed in clever ways.

Definition 1.2.1. *Given a set A, we let $|A|$ be the number of elements of A, and we call $|A|$ the cardinality of A. If A is infinite, then we write $|A| = \infty$.*

Of course, if we list the elements of a set A, then it's usually quite easy to determine $|A|$. For example, we trivially have $|\{1, \sqrt{2}, \frac{5}{2}, 18\}| = 4$. However, it can be very hard to determine the cardinality of a set. For example, consider the set

$$A = \{(x, y) \in \mathbb{Z}^2 : y^2 = x^3 - 1\}.$$

Determining the elements of A is not easy. It is easy to see that $(1, 0) \in A$, but it is not clear whether there are any other elements. Using some sophisticated number theory, it is possible to show that $A = \{(1, 0)\}$, and hence $|A| = 1$.

We start with one of the most basic, yet important, rules about the cardinality of sets.

Definition 1.2.2. *We say that two sets A and B are disjoint if $A \cap B = \emptyset$.*

Fact 1.2.3 (Sum Rule)**.** *If A and B are finite disjoint sets, then $|A \cup B| = |A| + |B|$.*

We won't give a formal proof of this fact, because it is so basic that it's hard to know what to assume (although if one goes through the trouble of carefully axiomatizing math with something like set theory, then it's possible to give a formal proof using a technique called mathematical induction that we will discuss in Chapter 2). At any rate, the key fact is that since A and B are disjoint, they have no elements in common. Therefore, each element of $A \cup B$ is in exactly one of A or B. Notice that the assumption that A and B are disjoint is essential. If $A = \{1, 2\}$ and $B = \{2, 3\}$, then $|A| = 2 = |B|$, but $|A \cup B| = 3$ because $A \cup B = \{1, 2, 3\}$.

Although the next result is again very intuitive, we show how to prove it using the Sum Rule.

Proposition 1.2.4 (Complement Rule)**.** *If A and B are finite sets and $B \subseteq A$, then $|A \backslash B| = |A| - |B|$.*

Proof. Notice that $A \backslash B$ and B are disjoint sets and that $(A \backslash B) \cup B = A$. Using the Sum Rule, it follows that $|A \backslash B| + |B| = |A|$. Subtracting $|B|$ from both sides, we conclude that $|A \backslash B| = |A| - |B|$. □

We can now easily generalize this to the case where B might not be a subset of A.

Proposition 1.2.5 (General Complement Rule)**.** *If A and B are finite sets, then $|A \backslash B| = |A| - |A \cap B|$.*

Proof. We have $A \backslash B = A \backslash (A \cap B)$. Since $A \cap B \subseteq A$, we can now apply the Complement Rule. □

We can generalize the Sum Rule to the following.

Definition 1.2.6. *A collection of sets $A_1, A_2, \dots, A_n$ is* pairwise disjoint *if $A_i \cap A_j = \emptyset$ whenever $i \neq j$.*

Fact 1.2.7 (General Sum Rule)**.** *If $A_1, A_2, \dots, A_n$ are finite sets that are pairwise disjoint, then $|A_1 \cup A_2 \cup \cdots \cup A_n| = |A_1| + |A_2| + \cdots + |A_n|$.*

Again, we won't give a formal proof of this fact (although it it possible to do so from the Sum Rule by induction on n). Notice that the pairwise disjoint assumption is again key, and it's not even enough to assume that $A_1 \cap A_2 \cap \cdots \cap A_n = \emptyset$ (see the homework).

Proposition 1.2.8. *If A and B are finite sets, we have $|A \cup B| = |A| + |B| - |A \cap B|$.*

Proof. Consider the three sets $A\backslash B$, $B\backslash A$, and $A\cap B$. These three sets are pairwise disjoint, and their union is $A\cup B$. Using the General Sum Rule, we conclude that

$$|A\cup B| = |A\backslash B| + |B\backslash A| + |A\cap B|.$$

Now $|A\backslash B| = |A| - |A\cap B|$ and $|B\backslash A| = |B| - |A\cap B|$ by the General Complement Rule. Plugging these in, we conclude that

$$|A\cup B| = |A| - |A\cap B| + |B| - |A\cap B| + |A\cap B|,$$

and hence

$$|A\cup B| = |A| + |B| - |A\cap B|.$$

□

Proposition 1.2.9 (Product Rule)**.** *If A and B are finite sets, then $|A\times B| = |A|\cdot|B|$.*

Proof. Let $n = |A|$ and let $m = |B|$. List the elements of A so that $A = \{a_1, a_2, \dots, a_n\}$. Similarly, list the elements of B so that $B = \{b_1, b_2, \dots, b_m\}$. For each i, let

$$A_i = \{(a_i, b_j) : 1 \le j \le m\} = \{(a_i, b_1), (a_i, b_2), \dots, (a_i, b_m)\}.$$

Thus, A_i is the subset of $A\times B$ consisting only of those pairs whose first element is a_i. Notice that the sets $A_1, A_2, \dots, A_n$ are pairwise disjoint and that

$$A\times B = A_1\cup A_2\cup\cdots\cup A_n.$$

Furthermore, we have that $|A_i| = m$ for all i. Using the General Sum Rule, we conclude that

$$\begin{aligned}
|A\times B| &= |A_1| + |A_2| + \cdots + |A_n| \\
&= m + m + \cdots + m \\
&= n\cdot m \\
&= |A|\cdot|B|.
\end{aligned}$$

□

Using induction (again, see Chapter 2), one can prove the following generalization.

Proposition 1.2.10 (General Product Rule)**.** *If $A_1, A_2, \dots, A_n$ are finite sets, then $|A_1\times A_2\times\cdots\times A_n| = |A_1|\cdot|A_2|\cdots|A_n|$.*

Corollary 1.2.11. *If A is a finite set and $n\in\mathbb{N}^+$, then $|A^n| = |A|^n$.*

Corollary 1.2.12. *For any $n\in\mathbb{N}^+$, we have that $|\{0,1\}^n| = 2^n$, i.e. there are 2^n many sequences of 0's and 1's of length n.*

1.3 Relations and Equivalence Relations

Definition 1.3.1. *Let A and B be sets. A* (binary) relation *between A and B is a subset $R\subseteq A\times B$. If $A = B$, then we call a subset of $A\times A$ a* (binary) relation on *A.*

For example, let $A = \{1, 2, 3\}$ and $B = \{6, 8\}$ as above. We saw above that

$$\{1, 2, 3\} \times \{6, 8\} = \{(1, 6), (1, 8), (2, 6), (2, 8), (3, 6), (3, 8)\}.$$

The set

$$R = \{(1, 6), (1, 8), (3, 8)\}$$

is a relation between A and B, although certainly not a very interesting one. However, we'll use it to illustrate a few facts. First, in a relation, it's possible for an element of A to be related to multiple elements of B, as in the case for $1 \in A$ for our example R. Also, it's possible that an element of A is related to no elements of B, as in the case of $2 \in A$ for our example R.

For a more interesting example, consider the binary relation on $\mathbb{Z}$ defined by $R = \{(a, b) \in \mathbb{Z}^2 : a < b\}$. Notice that $(4, 7) \in R$ and $(5, 5) \notin R$.

By definition, relations are sets. However, it is typically cumbersome to use set notation to write things like $(1, 6) \in R$. Instead, it usually makes much more sense to use infix notation and write $1R6$. Moreover, we can use better notation for the relation by using a symbol like $\sim$ instead of R. In this case, we would write $1 \sim 6$ instead of $(1, 6) \in \sim$ or $2 \not\sim 8$ instead of $(2, 8) \notin \sim$.

With this new notation, we give a few examples of binary relations on $\mathbb{R}$:

- Given $x, y \in \mathbb{R}$, we let $x \sim y$ if $x^2 + y^2 = 1$.
- Given $x, y \in \mathbb{R}$, we let $x \sim y$ if $x^2 + y^2 \leq 1$.
- Given $x, y \in \mathbb{R}$, we let $x \sim y$ if $x = \sin y$.
- Given $x, y \in \mathbb{R}$, we let $x \sim y$ if $y = \sin x$.

Again, notice from these examples that given $x \in \mathbb{R}$, there might be 0, 1, 2, or even infinitely many $y \in \mathbb{R}$ with $x \sim y$.

If we let $A = \{0, 1\}^*$ be the set of all finite sequences of 0's and 1's, then the following are binary relations on A:

- Given $\sigma, \tau \in A$, we let $\sigma \sim \tau$ if σ and τ have the same number of 1's.
- Given $\sigma, \tau \in A$, we let $\sigma \sim \tau$ if σ occurs as a consecutive subsequence of τ (for example, we have $010 \sim 001101011$ because 010 appears in positions 5-6-7 of 001101011).

For a final example, let A be the set consisting of the 50 states. Let R be the subset of $A \times A$ consisting of those pairs of states that have a common letter in the second position of their postal codes. For example, we have (Iowa,California) $\in R$ and and (Iowa, Virginia) $\in R$ because the postal codes of these sets are IA, CA, VA. We also have (Minnesota, Tennessee) $\in R$ because the corresponding postal codes are MN and TN. Now (Texas, Texas) $\in R$, but there is no $a \in A$ with $a \neq$ Texas such that (Texas, a) $\in R$, because no other state has X as the second letter of its postal code. Texas stands alone.

Recall that a binary relation on a set A is *any* subset of $A \times A$. As a result, a given relation might have very few nice properties. However, there are many special classes of relations, and one of the most important types is the following.

Definition 1.3.2. *An* equivalence relation *on a set A is a binary relation $\sim$ on A having the following three properties:*

- *$\sim$ is reflexive: $a \sim a$ for all $a \in A$.*
- *$\sim$ is symmetric: Whenever $a, b \in A$ satisfy $a \sim b$, we have $b \sim a$.*
- *$\sim$ is transitive: Whenever $a, b, c \in A$ satisfy $a \sim b$ and $b \sim c$, we have $a \sim c$.*

Consider the binary relation $\sim$ on $\mathbb{Z}$ where $a \sim b$ means that $a \leq b$. Notice that $\sim$ is reflexive because $a \leq a$ for all $a \in \mathbb{Z}$. Also, $\sim$ is transitive because if $a \leq b$ and $b \leq c$, then $a \leq c$. However, $\sim$ is not symmetric because $3 \sim 4$ but $4 \not\sim 3$. Thus, although $\sim$ satisfies two out of the three requirements, it is not an equivalence relation.

A simple example of an equivalence relation is where $A = \mathbb{R}$ and $a \sim b$ means that $|a| = |b|$. In this case, it is straightforward to check that $\sim$ is an equivalence relation. We now move on to some more interesting examples which we treat more carefully.

Example 1.3.3. *Let A be the set of all $n \times n$ matrices with real entries. Let $M \sim N$ mean that there exists an invertible $n \times n$ matrix P such that $M = PNP^{-1}$. We then have that $\sim$ is an equivalence relation on A.*

Proof. We need to check the three properties.

- Reflexive: Let $M \in A$ be arbitrary. The $n \times n$ identity matrix I is invertible and satisfies $I^{-1} = I$, so we have $M = IMI^{-1}$. Therefore, $\sim$ is reflexive.

- Symmetric: Let $M, N \in A$ be arbitrary with $M \sim N$. Fix an $n \times n$ invertible matrix P with $M = PNP^{-1}$. Multiplying on the left by P^{-1} we get $P^{-1}M = NP^{-1}$, and now multiplying on the right by P we conclude that $P^{-1}MP = N$. We know from linear algebra that P^{-1} is also invertible and $(P^{-1})^{-1} = P$, so $N = P^{-1}M(P^{-1})^{-1}$ and hence $N \sim M$.

- Transitive: Let $L, M, N \in A$ be arbitrary with $L \sim M$ and $M \sim N$. Since $L \sim M$, we may fix an $n \times n$ invertible matrix P with $L = PMP^{-1}$. Since $M \sim N$, we may fix an $n \times n$ invertible matrix Q with $M = QNQ^{-1}$. We then have

  $$L = PMP^{-1} = P(QNQ^{-1})P^{-1} = (PQ)N(Q^{-1}P^{-1}).$$

  Now by linear algebra, we know that the product of two invertible matrices is invertible, so PQ is invertible and furthermore we know that $(PQ)^{-1} = Q^{-1}P^{-1}$. Therefore, we have

  $$L = (PQ)N(PQ)^{-1},$$

  so $L \sim N$.

Putting it all together, we conclude that $\sim$ is an equivalence relation on A. □

Example 1.3.4. *Let A be the set $\mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$, i.e. A is the set of all pairs $(a, b) \in \mathbb{Z}^2$ with $b \neq 0$. Define a relation $\sim$ on A as follows. Given $a, b, c, d \in \mathbb{Z}$ with $b, d \neq 0$, we let $(a, b) \sim (c, d)$ mean $ad = bc$. We then have that $\sim$ is an equivalence relation on A.*

Proof. We check the three properties.

- Reflexive: Let $a, b \in \mathbb{Z}$ be arbitrary with $b \neq 0$. Since $ab = ba$, it follows that $(a, b) \sim (a, b)$.

- Symmetric: Let $a, b, c, d \in \mathbb{Z}$ be arbitrary with $b, d \neq 0$, and $(a, b) \sim (c, d)$. We then have that $ad = bc$. From this, we conclude that $cb = da$ so $(c, d) \sim (a, b)$.

- Transitive: Let $a, b, c, d, e, f \in \mathbb{Z}$ be arbitrary with $b, d, f \neq 0$ where $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. We then have that $ad = bc$ and $cf = de$. Multiplying the first equation by f we see that $adf = bcf$. Multiplying the second equation by b gives $bcf = bde$. Therefore, we know that $adf = bde$. Now $d \neq 0$ by assumption, so we may cancel it to conclude that $af = be$. It follows that $(a, b) \sim (e, f)$

Therefore, $\sim$ is an equivalence relation on A. □

Let's analyze the above situation more carefully. We have $(1,2) \sim (2,4)$, $(1,2) \sim (4,8)$, $(1,2) \sim (-5,-10)$, etc. If we think of (a,b) as representing the fraction $\frac{a}{b}$, then the relation $(a,b) \sim (c,d)$ is saying exactly that the fractions $\frac{a}{b}$ and $\frac{c}{d}$ are equal. You may never have thought about equality of fractions as the result of imposing an equivalence relation on pairs of integers, but that is exactly what it is. We will be more precise about this below.

Definition 1.3.5. *Let $\sim$ be an equivalence relation on a set A. Given $a \in A$, we let*

$$\overline{a} = \{b \in A : a \sim b\}.$$

The set $\overline{a}$ is called the equivalence class *of a.*

Some sources use the notation $[a]$ instead of $\overline{a}$. The former notation helps emphasize that the equivalence class of a is a *subset* of A rather than an element of A. However, it is cumbersome notation to use when we begin working with equivalence classes. We will stick with our notation, although it might take a little time to get used to it. Notice that by the reflexive property of $\sim$, we have that $a \in \overline{a}$ for all $a \in A$.

For example, let's return to where A is the set consisting of the 50 states and R is the subset of $A \times A$ consisting of those pairs of states that have a common letter in the second position of their postal codes. It's straightforward to show that R is an equivalence relation on A. We have

$$\overline{\text{Iowa}} = \{\text{California}, \text{Georgia}, \text{Iowa}, \text{Louisiana}, \text{Massachusetts}, \text{Pennsylvania}, \text{Virginia}, \text{Washington}\},$$

while

$$\overline{\text{Minnesota}} = \{\text{Indiana}, \text{Minnesota}, \text{Tennessee}\}$$

and

$$\overline{\text{Texas}} = \{\text{Texas}\}.$$

Notice that each of these are sets, even in the case of $\overline{\text{Texas}}$.

For another example, suppose we are working with $A = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ where $(a,b) \sim (c,d)$ means that $ad = bc$. As discussed above, some elements of $\overline{(1,2)}$ are $(1,2)$, $(2,4)$, $(4,8)$, $(-5,-10)$, etc. So

$$\overline{(1,2)} = \{(1,2), (2,4), (4,8), (-5,-10), \dots\}.$$

Again, I want to emphasize that $\overline{(a,b)}$ is a subset of A.

The following proposition is hugely fundamental. It says that if two equivalence classes overlap, then they must in fact be equal. In other words, if $\sim$ is an equivalence on A, then the equivalence classes *partition* the set A into pieces.

Proposition 1.3.6. *Let $\sim$ be an equivalence relation on a set A and let $a, b \in A$. If $\overline{a} \cap \overline{b} \neq \emptyset$, then $\overline{a} = \overline{b}$.*

Proof. Suppose that $\overline{a} \cap \overline{b} \neq \emptyset$. Since this set is nonempty, we can fix some $c \in \overline{a} \cap \overline{b}$. We then have $a \sim c$ and $b \sim c$. By symmetry, we know that $c \sim b$, and using transitivity we get that $a \sim b$. Using symmetry again, we conclude that $b \sim a$. We now show that $\overline{a} = \overline{b}$ by showing each containment:

- We first show that $\overline{a} \subseteq \overline{b}$. Let $x \in \overline{a}$ be arbitrary. We then have that $a \sim x$. Since $b \sim a$ from above, we can use transitivity to conclude that $b \sim x$, hence $x \in \overline{b}$.
- We next show that $\overline{b} \subseteq \overline{a}$. Let $x \in \overline{b}$ be arbitrary. We then have that $b \sim x$. Since $a \sim b$ from above, we can use transitivity to conclude that $a \sim x$, hence $x \in \overline{a}$.

Putting this together, we conclude that $\overline{a} = \overline{b}$. $\square$

With that proposition in hand, we are ready for the foundational theorem about equivalence relations.

Theorem 1.3.7. *Let $\sim$ be an equivalence relation on a set A and let $a, b \in A$.*

1. *$a \sim b$ if and only if $\overline{a} = \overline{b}$.*
2. *$a \not\sim b$ if and only if $\overline{a} \cap \overline{b} = \emptyset$.*

Proof.
1. Suppose first that $a \sim b$. We then have that $b \in \overline{a}$. Now we know that $b \sim b$ because $\sim$ is reflexive, so $b \in \overline{b}$. Thus, $b \in \overline{a} \cap \overline{b}$, so $\overline{a} \cap \overline{b} \neq \emptyset$. Using Proposition 1.3.6, we conclude that $\overline{a} = \overline{b}$.

   Suppose conversely that $\overline{a} = \overline{b}$. Since $b \sim b$ because $\sim$ is reflexive, we have that $b \in \overline{b}$. Therefore, $b \in \overline{a}$ and hence $a \sim b$.

2. Suppose that $a \not\sim b$. Since we just proved (1), it follows that $\overline{a} \neq \overline{b}$, so by Proposition 1.3.6 we must have $\overline{a} \cap \overline{b} = \emptyset$.

   Suppose conversely that $\overline{a} \cap \overline{b} = \emptyset$. We then have $\overline{a} \neq \overline{b}$ (because $a \in \overline{a}$ so $\overline{a} \neq \emptyset$), so $a \not\sim b$ by part (1). □

Therefore, given an equivalence relation $\sim$ on a set A, the equivalence classes partition A into pieces. Working out the details in our postal code example, one can show that $\sim$ has 1 equivalence class of size 8 (namely $\overline{\text{Iowa}}$, which is the same set as $\overline{\text{California}}$ and 6 others), 3 equivalence classes of size 4, 4 equivalence classes of size 3, 7 equivalence classes of size 2, and 4 equivalence classes of size 1.

Let's revisit the example of $A = \mathbb{Z} \times (\mathbb{Z} \backslash \{0\})$ where $(a, b) \sim (c, d)$ means $ad = bc$. The equivalence class of $(1, 2)$, namely the set $\overline{(1,2)}$ is the set of all pairs of integers which are ways of representing the fraction $\frac{1}{2}$. In fact, this is how once can "construct" the rational numbers from the integers. We simply *define* the rational numbers to be the set of equivalence classes of A under $\sim$. In other words, we let

$$\frac{a}{b} = \overline{(a,b)}.$$

So when we write something like

$$\frac{1}{2} = \frac{4}{8},$$

we are simply saying that

$$\overline{(1,2)} = \overline{(4,8)},$$

which is true because $(1, 2) \sim (4, 8)$.

1.4 Functions

We're all familiar with functions from Calculus. In that context, a function is often given by a "formula", such as $f(x) = x^4 - 4x^3 + 2x - 1$. However, we also encounter piecewise-defined functions, such as

$$f(x) = \begin{cases} x^2 + 1 & \text{if } x \geq 2, \\ x - 1 & \text{if } x < 2, \end{cases}$$

and the function $g(x) = |x|$, which is really piecewise defined as

$$g(x) = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

For a more exotic example of a piecewise defined function, consider

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

Despite these examples, only the most basic functions in mathematics are defined through formulas on pieces. For instance, the function $f(x) = \sin x$ is *not* given by a formula, and it is difficult to compute values of this function with any accuracy using only basic operations like $+$ and $\cdot$. In fact, we give this function the strange new name of "sine" because we can not express it easily using more basic operations. The function $f(x) = 2^x$ is easy to compute for $x \in \mathbb{Z}$, but it's actually nontrivial to compute and even define in general (after all, do you remember the definition of 2^π?). Even more fundamentally, the function $f(x) = \sqrt{x}$ is also not really given by a formula, because the definition, i.e. $f(x)$ is the unique positive y with the property that $y^2 = x$, does not give us an easy way to compute it.

Beyond these fundamental functions that you encounter before Calculus, you learn more exotic ways to define functions in Calculus. Given a function f, you learn how to define a new function f', called the derivative of f, using a certain limit operation. Now in many cases, you can compute f' more easily using facts like the Product Rule and the Chain Rule, but these rules do not always apply. Moreover, given any continuous function g, we can define a new function f by letting

$$f(x) = \int_0^x g(t)\ dt.$$

In other words, f is defined as the "(signed) area of g so far" function, in that $f(x)$ is defined to be the (signed) area between the graph of g and the x-axis over the interval from 0 to x. Formally, f is defined as a limit of Riemann sums. Again, in Calculus you learn ways to compute $f(x)$ more easily in many special cases using the Fundamental Theorem of Calculus. For example, if

$$f(x) = \int_0^x (3t^2 + t)\ dt,$$

then we can also compute f as

$$f(x) = x^3 + \frac{x^2}{2},$$

while if

$$f(x) = \int_0^x \sin t\ dt,$$

then we can also compute f as

$$f(x) = 1 - \cos x.$$

However, not all integrals can be evaluated so easily. In fact, it turns out that the perfectly well-defined function

$$f(x) = \int_0^x e^{-t^2}\ dt$$

can not be expressed through polynomials, exponentials, logs, and trigonometric functions using only operations like $+$, $\cdot$, and function composition. Of course, we can still approximate it using Riemann sums (or Simpson's Rule), and this is important for us to be able to do since this function represents the area under a normal curve, which is essential in statistics.

If we move away from functions whose inputs and outputs are real numbers, we can think about other interesting ways to define functions. For example, suppose we define a function whose inputs and outputs are elements of $\mathbb{R}^2$ by letting $f(\vec{u})$ be the result of rotating $\vec{u}$ by 27^o clockwise around the origin. This seems to be a well-defined function despite the fact that it is not clear how to compute it (though you likely learned how to compute it in Linear Algebra).

Alternatively, consider a function whose inputs and outputs are natural numbers by letting $f(n)$ be the number of primes less than or equal to n. For example, we have $f(3) = 2$, $f(4) = 2$, $f(9) = 4$, and $f(30) = 10$. Although it is possible to compute this function, it's not clear whether we can compute it quickly. In other words, it's not obvious if we can compute something like $f(2^{50})$ without a huge amount of work.

Perhaps you have some exposure to the concept of a function as it is used in computer programming. From this perspective, a function is determined by a sequence of imperative statements or function compositions as defined by a precise programming language. Since a computer is doing the interpreting, of course all such functions can be computed in principle (or if your computations involve real numbers, then at least up to good approximations). However, if you take this perspective, an interesting question arises. If we write two different functions f and g that do not follow the same steps, and perhaps even act qualitatively differently in structure, but they always produce the same output on the same input, should we consider them to be the same function? We can even ask this question outside of the computer science paradigm. For example, if we define $f(x) = \sin^2 x + \cos^2 x$ and $g(x) = 1$, then should we consider f and g be the same function?

We need to make a choice about how to define a function in general. Intuitively, given two sets A and B, a function $f\colon A \to B$ is an input-output "mechanism" that produces a *unique* output $b \in B$ for any given input $a \in A$. As we've seen, the vast majority of functions that we have encountered so far can be computed in principle, so up until this point, we could interpret "mechanism" in an algorithmic and computational sense. However, we want to allow as much freedom as possible in this definition so that we can consider new ways to define functions in time. In fact, as you might see in later courses (like Automata, Formal Languages, and Computational Complexity), there are some natural functions that are not computable even in theory. As a result, we choose to abandon the notion of computation in our definition. By making this choice, we will be able to sidestep some of the issues in the previous paragraph, but we still need to make a choice about whether to consider the functions $f(x) = \sin^2 x + \cos^2 x$ and $g(x) = 1$ to be equal.

With all of this background, we are now in a position to define functions as certain special types of sets. Thinking about functions from this more abstract point of view eliminates the vague "mechanism" concept because they will simply be sets. With this perspective, we'll see that functions can be defined in any way that a set can be defined. Our approach both clarifies the concept of a function and also provides us with some much needed flexibility in defining functions in more interesting ways. Here is the formal definition.

Definition 1.4.1. *Let A and B be sets. A* function *from A to B is a subset f of $A \times B$ with the property that for all $a \in A$, there exists a unique $b \in B$ with $(a, b) \in f$. Also, instead of writing "f is a function from A to B", we typically use the shorthand notation "$f\colon A \to B$".*

For example, let $A = \{2, 3, 5, 7\}$ and let $B = \mathbb{N} = \{0, 1, 2, 3, 4, \dots\}$. An example of a function $f\colon A \to B$ is the set

$$f = \{(2, 71), (3, 4), (5, 9382), (7, 4)\}.$$

Notice that in the definition of a function from A to B, we know that for every $a \in A$, there is a unique $b \in B$ such that $(a, b) \in f$. However, as this example shows, it may not be the case that for every $b \in B$, there is a unique $a \in A$ with $(a, b) \in f$. Be careful with the order of quantifiers!

We can also convert the typical way of defining a function into this formal set theoretic way. For example, consider the function $f\colon \mathbb{R} \to \mathbb{R}$ by letting $f(x) = x^2$. We can instead define f by the set

$$\{(x, y) \in \mathbb{R} \times \mathbb{R} : y = x^2\},$$

or parametrically as

$$\{(x, x^2) : x \in \mathbb{R}\}.$$

One side effect of our definition of a function is that we immediately obtain a nice definition for when two functions $f\colon A \to B$ and $g\colon A \to B$ are equal because we have defined when two sets are equal. Given two function $f\colon A \to B$ and $g\colon A \to B$, if we unwrap our definition of set equality, we see that $f = g$ exactly when f and g have the same elements, which is precisely the same thing as saying that $f(a) = g(a)$ for all $a \in A$. In particular, the *manner* in which we describe functions does not matter so long as the functions behave the same on all inputs. For example, if we define $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ by letting $f(x) = \sin^2 x + \cos^2 x$ and $g(x) = 1$, then we have that $f = g$ because $f(x) = g(x)$ for all $x \in \mathbb{R}$.

Thinking of functions as special types of sets is helpful to clarify definitions, but is often awkward to work with in practice. For example, writing $(2, 71) \in f$ to mean that f sends 2 to 71 quickly becomes annoying. Thus, we introduce some new notation matching up with our old experience with functions.

Notation 1.4.2. *Let A and B be sets. If $f : A \to B$ and $a \in A$, we write $f(a)$ to mean the unique $b \in B$ such that $(a, b) \in f$.*

For instance, in the above example of f, we can instead write

$$f(2) = 71, \qquad f(3) = 4, \qquad f(5) = 9382, \qquad \text{and} \qquad f(7) = 4.$$

Definition 1.4.3. *Let $f : A \to B$ be a function.*

- *We call A the* domain *of f.*
- *We call B the* codomain *of f.*
- *We define range$(f) = \{b \in B :$ There exists $a \in A$ with $f(a) = b\}$.*

Notice that given a function $f : A \to B$, we have $\text{range}(f) \subseteq B$, but it is possible that $\text{range}(f) \neq B$. For example, in the above case, we have that the codomain of f is $\mathbb{N}$, but $\text{range}(f) = \{4, 71, 9382\}$. In general, given a function $f : A \to B$, it may be very difficult to determine $\text{range}(f)$ because we may need to search through all $a \in A$.

For an interesting example of a function with a mysterious looking range, fix $n \in \mathbb{N}^+$ and define $f : \{0, 1, 2, \dots, n-1\} \to \{0, 1, 2, \dots, n-1\}$ by letting $f(a)$ be the remainder when dividing a^2 by n. For example, if $n = 10$, then we have the following table of values of f:

$$\begin{array}{lllll} f(0) = 0 & f(1) = 1 & f(2) = 4 & f(3) = 9 & f(4) = 6 \\ f(5) = 5 & f(6) = 6 & f(7) = 9 & f(8) = 4 & f(9) = 1. \end{array}$$

Thus, for $n = 10$, we have $\text{range}(f) = \{0, 1, 4, 5, 6, 9\}$. This simple but strange looking function has many interesting properties. Given a reasonably large number $n \in \mathbb{N}$, it looks potentially difficult to determine whether an element is in $\text{range}(f)$ because we might need to search through a huge number of inputs to see if a given output actually occurs. If n is prime, then it turns out that there are much faster ways to determine if a given element is in $\text{range}(f)$ (see Algebraic Number Theory). However, it is widely believed (although we do not currently have a proof!) that there is no efficient method to do this when n is the product of two large primes, and this is the basis for some cryptosystems (Goldwasser-Micali) and pseudo-random number generators (Blum-Blum-Shub).

Definition 1.4.4. *Suppose that $f : A \to B$ and $g : B \to C$ are functions. The* composition *of g and f, denoted $g \circ f$, is the function $g \circ f : A \to C$ defined by letting $(g \circ f)(a) = g(f(a))$ for all $a \in A$.*

Notice that in general we have $f \circ g \neq g \circ f$ even when both are defined! If $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = x + 1$ and $g : \mathbb{R} \to \mathbb{R}$ is $g(x) = x^2$, then

$$\begin{aligned} (f \circ g)(x) &= f(g(x)) \\ &= f(x^2) \\ &= x^2 + 1 \end{aligned}$$

while

$$\begin{aligned} (g \circ f)(x) &= g(f(x)) \\ &= g(x + 1) \\ &= (x + 1)^2 \\ &= x^2 + 2x + 1. \end{aligned}$$

Notice then that $(f\circ g)(1) = 1^2+1 = 2$ while $(g\circ f)(1) = 1^2+2\cdot 1+1 = 4$. Since we have found one example of an $x \in \mathbb{R}$ with $(f\circ g)(x) \neq (f\circ g)(x)$, we conclude that $f\circ g \neq g\circ f$. It does not matter that there do exist some values of x with $(f\circ g)(x) = (f\circ g)(x)$ (for example, this is true when $x = 0$). Remember that two functions are equal precisely when they agree on *all* inputs, so to show that the two functions are not equal it suffices to find just one value where they disagree (again remember that the negation of a "for all" statement is a "there exists" statement).

Proposition 1.4.5. *Let A, B, C, D be sets. Suppose that $f\colon A \to B$, that $g\colon B \to C$, and that $h\colon C \to D$ are functions. We then have that $(h\circ g)\circ f = h\circ(g\circ f)$. Stated more simply, function composition is associative whenever it is defined.*

Proof. Let $a \in A$ be arbitrary. We then have

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

where each step follows by definition of composition. Therefore $((h\circ g)\circ f)(a) = (h\circ(g\circ f))(a)$ for all $a \in A$. It follows that $(h\circ g)\circ f = h\circ(g\circ f)$. □

1.5 Divisibility

Definition 1.5.1. *Let $a, b \in \mathbb{Z}$. We say that a* divides *b, and write $a \mid b$, if there exists $m \in \mathbb{Z}$ with $b = am$.*

For example, we have $2 \mid 6$ because $2\cdot 3 = 6$ and $3 \mid -21$ because $3\cdot(-7) = 21$. On the other hand, we have $2 \nmid 5$. To see this, we argue as follows.

- For any $m \in \mathbb{Z}$ with $m \leq 2$, we have $2m \leq 4$.
- For any $m \in \mathbb{Z}$ with $m > 2$, we have $m \geq 3$, so $2m \geq 6$.

Therefore, for every $m \in \mathbb{Z}$, we have $2m \neq 5$. It follows that $2 \nmid 5$. We will see less painful ways to prove this later.

Notice that $a \mid 0$ for every $a \in \mathbb{Z}$ because $a\cdot 0 = 0$ for all $a \in \mathbb{Z}$. In particular, we have $0 \mid 0$ because as noted we have $0\cdot 0 = 0$. Of course we also have $0\cdot 3 = 0$ and in fact $0\cdot m = 0$ for all $m \in \mathbb{Z}$, so every integer serves as a "witness" that $0 \mid 0$. Our definition says nothing about the $m \in \mathbb{Z}$ being unique.

For example, we have $2 \mid 6$ because $2\cdot 3 = 6$ and $-3 \mid 21$ because $-3\cdot 7 = 21$. We also have that $2 \nmid 5$ since it is "obvious" that no such integer exists. If you are uncomfortable with that (and you should be!), we will give methods to prove such statements in the next couple of sections.

Proposition 1.5.2. *If $a \mid b$ and $b \mid c$, then $a \mid c$.*

Proof. Since $a \mid b$, there exists $m \in \mathbb{Z}$ with $b = am$. Since $b \mid c$, there exists $n \in \mathbb{Z}$ with $c = bn$. We then have

$$c = bn = (am)n = a(mn)$$

Since $mn \in \mathbb{Z}$, it follows that $a \mid c$. □

Proposition 1.5.3.

1. *If $a \mid b$, then $a \mid kb$ for all $k \in \mathbb{Z}$.*

2. *If $a \mid b$ and $a \mid c$, then $a \mid (b + c)$.*

3. *If $a \mid b$ and $a \mid c$, then $a \mid (mb + nc)$ for all $m, n \in \mathbb{Z}$.*

Proof.

1. Let $a, b, k \in \mathbb{Z}$ be arbitrary with $a \mid b$. Since $a \mid b$, we can fix $m \in \mathbb{Z}$ with $b = am$. We then have

$$kb = k(am) = a(mk).$$

Since $mk \in \mathbb{Z}$, it follows that $a \mid kb$.

2. Let $a, b, c \in \mathbb{Z}$ be arbitrary with both $a \mid b$ and $a \mid c$. Since $a \mid b$, we can fix $m \in \mathbb{Z}$ with $b = am$. Since $a \mid c$, we can fix $n \in \mathbb{Z}$ with $c = an$. Notice that

$$b + c = am + an = a(m + n).$$

Since $m, n \in \mathbb{Z}$, we know that $m + n \in \mathbb{Z}$, it follows that $a \mid b + c$.

3. Let $m, n \in \mathbb{Z}$ be arbitrary. Since $a \mid b$, we conclude from part 1 that $a \mid mb$. Since $a \mid c$, we conclude from part 1 again that $a \mid nc$. Using part 2, it follows that $a \mid (bm + cn)$.

□

Proposition 1.5.4. *Suppose that $a, b \in \mathbb{Z}$. If $a \mid b$ and $b \neq 0$, then $|a| \leq |b|$.*

Proof. Suppose that $a \mid b$ with $b \neq 0$. Fix $d \in \mathbb{Z}$ with $ad = b$. Since $b \neq 0$, we have $d \neq 0$. Thus, $|d| \geq 1$, and so

$$\begin{aligned} |b| &= |ad| \\ &= |a| \cdot |d| \\ &\geq |a| \cdot 1 \\ &= |a|. \end{aligned}$$

□

Corollary 1.5.5. *Suppose that $a, b \in \mathbb{Z}$. If $a \mid b$ and $b \mid a$, then either $a = b$ or $a = -b$.*

Proof. We handle three cases:

- *Case 1:* Suppose that $a \neq 0$ and $b \neq 0$. By the Proposition 1.5.4, we know that both $|a| \leq |b|$ and $|b| \leq |a|$. It follows that $|a| = |b|$, and hence either $a = b$ or $a = -b$.

- *Case 2:* Suppose that $a = 0$. Since $a \mid b$, we may fix $m \in \mathbb{Z}$ with $b = am$. We then have $b = am = 0m = 0$ as well. Therefore, $a = b$.

- *Case 3:* Suppose that $b = 0$. Since $b \mid a$, we may fix $m \in \mathbb{Z}$ with $a = bm$. We then have $a = bm = 0m = 0$ as well. Therefore, $a = b$.

Thus, in all cases, we have that either $a = b$ or $a = -b$. □

Given an integer $a \in \mathbb{Z}$, we introduce the following notation for the set of all divisors of a.

Definition 1.5.6. *Given $a \in \mathbb{Z}$, we let $Div(a) = \{d \in \mathbb{Z} : d \mid a\}$.*

Notice that if $a \neq 0$, then $|d| \leq |a|$ for all $d \in Div(a)$ by Proposition 1.5.4. Thus, we need only check finitely many values to determine $Div(a)$. For instance, we have $Div(7) = \{1, -1, 7, -7\}$ (don't forget the negatives!), which we can write more succinctly as $\{\pm 1, \pm 7\}$, while $Div(6) = \{\pm 1, \pm 2, \pm 3, \pm 6\}$ (we'll soon see better ways to compute these sets that do not require an exhaustive search). For a more interesting example, we have $Div(0) = \mathbb{Z}$.

Proposition 1.5.7. *For any $a \in \mathbb{Z}$, we have $Div(a) = Div(-a)$.*

Proof. Exercise. $\square$

Chapter 2

Induction and Well-Ordering

2.1 Mathematical Induction

Suppose that we want to prove that a certain statement is true for all natural numbers. In other words, we want to do the following:

- Prove that the statement is true for 0.
- Prove that the statement is true for 1.
- Prove that the statement is true for 2.
- Prove that the statement is true for 3.
- $\ldots$.

Of course, since there are infinitely many natural numbers, going through each one in turn does not work because we will never handle them all this way. How can we get around this? Suppose that when we examine the first few proofs above that they look the same except that we replace 0 by 1 everywhere, or 0 by 2 everywhere, etc. In this case, one is tempted to say that "the pattern continues" or something similar, but that is not convincing because we can't be sure that the pattern does not break down when we reach 5419. One way to argue that the "the pattern continues" and handle all of the infinitely many possibilities at once is to take an arbitrary natural number n, and prove that the statement is true for n using *only* the fact that n is a natural number (but *not* any particular natural number).

The method of taking an arbitrary $n \in \mathbb{N}$ and proving that the statement is true for n is the standard way of proving a statement involving a "for all" quantifier. This technique also works to prove that a statement is true for all real numbers or for all matrices, as long as we take an *arbitrary* such object. However, there is a different method one can use to prove that every natural number has a certain property, and this one does not carry over to other settings like the real numbers. The key fact is that the natural numbers start with 0 and proceed in discrete steps forward. With this in mind, consider what would happen if we could accomplish each of the following:

- Prove that the statement is true for 0.
- Prove that if the statement is true for 0, then the statement is true for 1.
- Prove that if the statement is true for 1, then the statement is true for 2.
- Prove that if the statement is true for 2, then the statement is true for 3.

- ....

Suppose that we are successful in proving each of these. From the first line, we then know that the statement is true for 0. Since we now know that it's true for 0, we can use the second line to conclude that the statement is true for 1. Since we now know that it's true for 1, we can use the second line to conclude that the statement is true for 2. And so on. In the end, we are able to conclude that the statement is true for all natural numbers.

Let's examine this situation more closely. On the fact of it, each line looks more complicated than the corresponding line for for a direct proof. However, the key fact is that from the second line onward, we now have an additional assumption! Thus, instead of proving that the statement is true for 3 without any help, we can now use the assumption that the statement is true for 2 in that argument. Extra assumptions are always welcome because we have more that we can use in the actual argument.

Of course, as in our discussion at the beginning of this section, we can't hope to prove each of these infinitely many things one at a time. In an ideal world, the arguments from the second line onward all look exactly the same with the exception of replacing the number involved. Thus, the idea is to prove the following:

- Prove that the statement is true for 0.
- Prove that if the statement is true for n, then the statement is true for $n + 1$.

Notice that for the second line, we would need to prove that it is true for an arbitrary $n \in \mathbb{N}$, just like we would have to in a direct argument. An argument using these method is called a proof by (mathematical) *induction*, and it is an extremely useful and common technique in mathematics. We can also state this approach formally in terms of sets, allowing us to bypass the vague notion of "statement" that we employed above.

Fact 2.1.1 (Principle of Mathematical Induction on $\mathbb{N}$)**.** *Let $X \subseteq \mathbb{N}$. Suppose that the following are true:*

- $0 \in X$ *(the* base case*).*
- $n + 1 \in X$ *whenever* $n \in X$ *(the* inductive step*).*

We then have that $X = \mathbb{N}$.

Once again, here's the intuitive argument for why induction is valid. By the first assumption, we know that $0 \in X$. Since $0 \in X$, the second assumption tells us that $1 \in X$. Since $1 \in X$, the second assumption again tells us that $2 \in X$. By repeatedly applying the second assumption in this manner, each element of $\mathbb{N}$ is eventually determined to be in X. Notice that a similar argument works if we start with a different base case, i.e. if we start by proving that $3 \in X$ and then prove the inductive step, then it follows that $n \in X$ for all $n \in \mathbb{N}$ with $n \geq 3$.

Although we have stated induction with a base case of 0, it is also possible to give an inductive proof that starts at a different natural. For example, if we prove a base case the $4 \in X$, and we prove the usual inductive step that $n + 1 \in X$ whenever $n \in X$, then we can conclude that $n \in X$ for all $n \in \mathbb{N}$ with $n \geq 4$, i.e. that $\{n \in \mathbb{N} : n \geq 4\} \subseteq X$.

We now give many examples of proofs by induction. For our first example, we establish a formula for the sum of the first n positive natural numbers.

Proposition 2.1.2. *For any $n \in \mathbb{N}^+$, we have*

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

i.e.

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

We give two proofs. The first is a clever argument that avoids induction, while the second is a typical application of induction.

Proof 1. We first give a proof without induction. Let $n \in \mathbb{N}^+$ be arbitrary. Let $S = 1 + 2 + \cdots + (n-1) + n$. We also have $S = n + (n-1) + \cdots + 2 + 1$. Adding both of these equalities, we conclude that

$$2S = (n+1) + (n+1) + \cdots + (n+1) + (n+1),$$

and hence

$$2S = n(n+1).$$

Dividing both sides by 2, we conclude that

$$S = \frac{n(n+1)}{2}$$

so $1 + 2 + \cdots + (n-1) + n = \frac{n(n+1)}{2}$. □

While elegant, the previous argument required the creative insight of rewriting the sum in a different way, and finding the resulting clever pairing. We now give an inductive proof, which replaces the creative leap with some algebraic manipulations.

Proof 2. We give a proof using induction. Since we are proving something about all elements of $\mathbb{N}^+$, we start with a base case of 1.

- *Base Case:* For $n = 1$, the sum on the left-hand side is 1, and the right-hand side is $\frac{1 \cdot 2}{2} = 1$. Thus, that statement is true when $n = 1$.
- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. suppose that n is a number for which we know that

  $$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

  We then have

  $$\begin{aligned} 1 + 2 + \cdots + n + (n+1) &= \frac{n(n+1)}{2} + (n+1) && \text{(by the inductive hypothesis)} \\ &= \frac{n^2 + n + 2n + 2}{2} \\ &= \frac{n^2 + 3n + 2}{2} \\ &= \frac{(n+1)(n+2)}{2} \\ &= \frac{(n+1)((n+1)+1)}{2}. \end{aligned}$$

  Thus, the statement is true for $n + 1$.

By induction, we conclude that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

for all $n \in \mathbb{N}^+$. □

In the previous proof, we could have written it using the set-theoretic form of induction by letting

$$X = \left\{ n \in \mathbb{N}^+ : \sum_{k=1}^{n} i = \frac{n(n+1)}{2} \right\},$$

and then used the principle of induction to argue that $X = \mathbb{N}^+$. Typically, we will avoid formally writing the set, and working in this way, but it is always possible to translate arguments into the corresponding set-theoretic approach.

Proposition 2.1.3. *For any $n \in \mathbb{N}^+$, we have*

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

i.e.

$$1 + 3 + 5 + 7 + \cdots + (2n-1) = n^2.$$

Proof. We give a proof by induction.

- *Base Case:* Suppose that $n = 1$. We have

  $$\sum_{k=1}^{1} (2k-1) = 2 \cdot 1 - 1 = 1,$$

  so the left hand-side is 1. The right-hand side is $1^2 = 1$. Thus, the statement is true when $n = 1$.

- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. suppose that n is a number for which we know that

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

  Notice that $2(n+1) - 1 = 2n + 2 - 1 = 2n + 1$, hence

  $$\begin{aligned} \sum_{k=1}^{n+1} (2k-1) &= [\sum_{k=1}^{n} (2k-1)] + [2(n+1) - 1] \\ &= [\sum_{k=1}^{n} (2k-1)] + (2n+1) \\ &= n^2 + (2n+1) && \text{(by induction)} \\ &= (n+1)^2. \end{aligned}$$

  Thus, the statement is true for $n + 1$.

By induction, we conclude that

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

for all $n \in \mathbb{N}^+$. □

Although induction is a useful tool for proving certain equalities, it can also be used in much more flexible ways. We now give several examples of proving divisibility and inequalities by induction.

Proposition 2.1.4. *For all $n \in \mathbb{N}$, we have $3 \mid (4^n - 1)$.*

Proof. We give a proof by induction.

- *Base Case:* Suppose that $n = 0$. We have $4^0 - 1 = 1 - 1 = 0$, hence $3 \mid (4^0 - 1)$ because $3 \cdot 0 = 0$. Thus, the statement is true when $n = 0$.
- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. suppose that n is a number for which we know that $3 \mid (4^n - 1)$. Fix $k \in \mathbb{Z}$ with $3k = 4^n - 1$. Adding 1 to both sides, it follows that $4^n = 3k + 1$, and hence

$$\begin{aligned} 4^{n+1} - 1 &= 4 \cdot 4^n - 1 \\ &= 4 \cdot (3k + 1) - 1 \\ &= 12k + 3 \\ &= 3 \cdot (4k + 1). \end{aligned}$$

  Since $4k + 1 \in \mathbb{Z}$, we conclude that $3 \mid (4^{n+1} - 1)$. Thus, the statement is true for $n + 1$.

By induction, we conclude that $3 \mid (4^n - 1)$ for all $n \in \mathbb{N}$. $\square$

Proposition 2.1.5. *We have $2n + 1 < n^2$ for all $n \in \mathbb{N}$ with $n \geq 3$.*

Proof. We give a proof by induction.

- *Base Case:* Suppose that $n = 3$. We have $2 \cdot 3 + 1 = 7$ and $3^2 = 9$, so $2 \cdot 3 + 1 < 3^2$. Thus, the statement is true when $n = 3$.
- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}$ with $n \geq 3$, i.e. suppose that $n \geq 3$ is a number for which we know that $2n + 1 < n^2$. Since $2n + 1 \geq 2 \cdot 3 + 1 = 7 > 2$, we then have

$$\begin{aligned} 2(n + 1) + 1 &= 2n + 3 \\ &= (2n + 1) + 2 \\ &< n^2 + 2 \\ &< n^2 + 2n + 1 \\ &= (n + 1)^2. \end{aligned}$$

  Thus, the statement is true for $n + 1$.

By induction, we conclude that $2n + 1 < n^2$ for all $n \in \mathbb{N}$ with $n \geq 3$. $\square$

Proposition 2.1.6. *We have $n^2 < 2^n$ for all $n \geq 5$.*

Proof. We give a proof by induction.

- *Base Case:* Suppose that $n = 5$. We have $5^2 = 25$ and $2^5 = 32$, so $5^2 < 2^5$. Thus, the statement is true when $n = 5$.
- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}$ with $n \geq 5$, i.e. suppose that $n \geq 5$ is a number for which we know that $n^2 < 2^n$. Since $n^2 = n \cdot n \geq 3n = 2n + n > 2n + 1$, we have then have

$$\begin{aligned} (n + 1)^2 &= n^2 + 2n + 1 \\ &< n^2 + n^2 \\ &= 2n^2 \\ &< 2 \cdot 2^n \\ &= 2^{n+1}. \end{aligned}$$

Thus, the statement is true for $n+1$.

By induction, we conclude that $n^2 < 2^n$ for all $n \geq 5$. $\square$

Proposition 2.1.7. *For all $x \in \mathbb{R}$ with $x \geq -1$ and all $n \in \mathbb{N}^+$, we hve $(1+x)^n \geq 1+nx$.*

On the face of it, this problem looks a little different because we are also quantifying over infinitely many real numbers x. Since x is coming from $\mathbb{R}$, we can't induct on x. However, we *can* take an arbitrary $x \in \mathbb{R}$ with $x \geq -1$, and then induct on n for this particular x. We now carry out that argument.

Proof. Let $x \in \mathbb{R}$ be arbitrary with $x \geq -1$. For this x, we show that $(1+x)^n \geq 1+nx$ for all $n \in \mathbb{N}^+$ by induction.

- *Base Case:* Suppose that $n = 1$. We then have that $(1+x)^1 = 1+x = 1+1x$, so certainly $(1+x)^1 \geq 1+1x$.
- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. suppose that n is a number for which we know that $(1+x)^n \geq 1+nx$. Since $x \geq -1$, we have $1+x \geq 0$, so we can multiply both sides of this inequality by $(1+x)$ to conclude that

$$(1+x)^n \cdot (1+x) \geq (1+nx) \cdot (1+x).$$

We then have

$$\begin{aligned}
(1+x)^{n+1} &= (1+x)^n \cdot (1+x) \\
&\geq (1+nx) \cdot (1+x) && \text{(from above)} \\
&= 1+nx+x+nx^2 \\
&= 1+(n+1)x+nx^2 \\
&\geq 1+(n+1)x. && \text{(since } nx^2 \geq 0\text{)}
\end{aligned}$$

Hence, we have shown that $(1+x)^{n+1} \geq 1+(n+1)x$, i.e. that the statement is true for $n+1$.

By induction, we conclude that $(1+x)^n \geq 1+nx$ for all $n \in \mathbb{N}^+$. Since $x \in \mathbb{R}$ with $x \geq -1$ was arbitrary, the result follows. $\square$

Proposition 2.1.8. *For all $n \in \mathbb{N}^+$, we have*

$$\sum_{k=1}^{n} \frac{1}{k^2} \leq 2 - \frac{1}{n}.$$

Proof. We prove the statement by induction.

- *Base Case:* Suppose that $n = 1$. In this case, we have

$$\sum_{k=1}^{1} \frac{1}{k^2} = \frac{1}{1^2} = 1$$

and

$$2 - \frac{1}{1} = 2 - 1 = 1$$

hence

$$\sum_{k=1}^{1} \frac{1}{k^2} \leq 2 - \frac{1}{1}.$$

- *Inductive Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. suppose that n is a number for which we know that
$$\sum_{k=1}^{n} \frac{1}{k^2} \leq 2 - \frac{1}{n}.$$
We then have
$$\begin{aligned} \sum_{k=1}^{n+1} \frac{1}{k^2} &= \left(\sum_{k=1}^{n} \frac{1}{k^2}\right) + \frac{1}{(n+1)^2} \\ &\leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2} \\ &= 2 - \left(\frac{1}{n} - \frac{1}{(n+1)^2}\right) \\ &= 2 - \frac{(n+1)^2 - n}{n(n+1)^2} \\ &= 2 - \frac{n^2+n+1}{n(n+1)^2} \\ &\leq 2 - \frac{n^2+n}{n(n+1)^2} \\ &= 2 - \frac{n(n+1)}{n(n+1)^2} \\ &= 2 - \frac{1}{n+1}. \end{aligned}$$
Thus, the statement is true for $n+1$.

By induction, we conclude we conclude that
$$\sum_{k=1}^{n} \frac{1}{k^2} \leq 2 - \frac{1}{n}$$
for all $n \in \mathbb{N}^+$. □

2.2 Strong Induction and Well-Ordering

Remember our original model for induction:

- Prove that the statement is true for 0.
- Prove that if the statement is true for 0, then the statement is true for 1.
- Prove that if the statement is true for 1, then the statement is true for 2.
- Prove that if the statement is true for 2, then the statement is true for 3.
- Prove that if the statement is true for 3, then the statement is true for 4.
- ....

In the previous section, we argued why this model was sound and gave many examples. However, upon closer inspection, it appears that we can assume more. In the second line, when proving that the statement is true for 1 we are allowed to assume that the statement is true for 0. Now in the third line, when proving that the statement is true for 2, we only assume that it is true for 1. If we are knocking down the natural numbers in order, then we've already proved that it's true for 0, so why can't we assume that as well? The answer is that we can indeed assume it! In general, when working to prove that the statement is true for a natural number n, we can assume that the statement is true for all smaller natural numbers. In other words, we do the following:

- Prove that the statement is true for 0.
- Prove that if the statement is true for 0, then the statement is true for 1.
- Prove that if the statement is true for 0 and 1, then the statement is true for 2.
- Prove that if the statement is true for 0, 1, and 2, then the statement is true for 3.
- Prove that if the statement is true for 0, 1, 2, and 3, then the statement is true for 4.
- ....

Suppose that we are successful in doing this. From the first line, we then know that the statement is true for 0. Since we now know that it's true for 0, we can use the second line to conclude that the statement is true for 1. Since we now know that it's true for both 0 and 1, we can use the second line to conclude that the statement is true for 2. And so on. In the end, we are able to conclude that the statement is true for all natural numbers.

As usual, we can't hope to prove each of these infinitely many things one at a time. In an ideal world, the arguments from the second line onward all look exactly the same with the exception of replacing the number involved. Thus, the idea is to prove the following.

- Prove that the statement is true for 0.
- Prove that if the statement is true for each of $0, 1, 2, \dots, n$, then the statement is true for $n + 1$.

Alternatively, we can state this as follows:

- Prove that the statement is true for 0.
- Prove that if the statement is true for each of $0, 1, 2, \dots, n - 1$, then the statement is true for n (for $n \geq 1$).

An argument using these method is called a proof by *strong induction*. As we will see in the examples below, sometimes we need to modify this simple structure to include several base cases in order to get the argument going. Rather than going through a theoretical discussion of how and why one would do this, it's easier to illustrate the technique by example.

We start with an example where we verify a simple closed formed formula for a recursively defined sequence. Since the sequence uses the past two values to define the current value, regular induction does not give enough power to complete the proof.

Proposition 2.2.1. *Define a sequence a_n recursively by letting $a_0 = 0$, $a_1 = 1$, and*

$$a_n = 3a_{n-1} - 2a_{n-2}$$

for $n \geq 2$. Show that $a_n = 2^n - 1$ for all $n \in \mathbb{N}$.

Proof. We prove that $a_n = 2^n - 1$ for all $n \in \mathbb{N}$ by strong induction.

- *Base Case:* We handle two bases where $n = 0$ and $n = 1$ because our inductive step will use the result for two steps back. When $n = 0$, we have $a_0 = 0$ and $2^0 - 1 = 1 - 1 = 0$, so $a_0 = 2^0 - 1$. When $n = 1$, we have $a_1 = 1$ and $2^1 - 1 = 2 - 1 = 1$, so $a_1 = 2^1 - 1$.

- *Inductive Step:* Let $n \geq 2$ and assume that the statement is true for $0, 1, 2, \dots, n-1$, i.e. assume that $a_m = 2^m - 1$ for all $m \in \{0, 1, 2, \dots, n-1\}$. We prove that the statement is true for n. Notice that since $n \geq 2$, we have $0 \leq n - 1 < n$ and $0 \leq n - 2 < n$, so we know that $a_{n-1} = 2^{n-1} - 1$ and $a_{n-2} = 2^{n-2} - 1$. Now

$$\begin{aligned}
a_n &= 3a_{n-1} - 2a_{n-1} && \text{(by definition since } n \geq 2\text{)} \\
&= 3 \cdot (2^{n-1} - 1) - 2 \cdot (2^{n-2} - 1) && \text{(by the inductive hypothesis)} \\
&= 3 \cdot 2^{n-1} - 3 - 2 \cdot 2^{n-2} + 2 \\
&= 3 \cdot 2^{n-1} - 2^{n-1} - 1 \\
&= (3 - 1) \cdot 2^{n-1} - 1 \\
&= 2 \cdot 2^{n-1} - 1 \\
&= 2^n - 1.
\end{aligned}$$

  Thus, $a_n = 2^n - 1$ and so the statement is true for n.

Using strong induction, we conclude that $a_n = 2^n - 1$ for all $n \in \mathbb{N}$. □

We now turn to an interesting example of using strong induction to establish when we can solve an equation in the natural numbers.

Proposition 2.2.2. *If $n \in \mathbb{N}$ and $n \geq 12$, then there exist $k, \ell \in \mathbb{N}$ with $n = 3k + 7\ell$.*

Proof. We give a proof by strong induction.

- *Base Case:* We first prove that the statement is true for all $n \in \{12, 13, 14\}$ (we will see why we need so many base cases in the inductive step below). We have the following cases:

  - $12 = 3 \cdot 4 + 7 \cdot 0$.
  - $13 = 3 \cdot 2 + 7 \cdot 1$.
  - $14 = 3 \cdot 0 + 7 \cdot 2$.

  Thus, the statement is true for all $n \in \{12, 13, 14\}$.

- *Inductive Step:* Let $n \geq 15$ and assume that the statement is true for all $k \in \mathbb{N}$ with $12 \leq k < n$, i.e. assume that the statement is true for $12, 13, 14, \dots, n-1$. We prove that the statement is true for n. Since $n \geq 15$, we have $12 \leq n - 3 < n$, so we can use the inductive hypothesis to fix $k, \ell \in \mathbb{N}$ with

$$n - 3 = 3k + 7\ell.$$

  Adding 3 to both sides, we see that

$$\begin{aligned}
n &= 3k + 7\ell + 3 \\
&= 3(k + 1) + 5\ell.
\end{aligned}$$

  Since $k + 1, \ell \in \mathbb{N}$, we conclude that the statement is true for n.

By strong induction, we conclude that for all $n \in \mathbb{N}$ with $n \geq 12$, there exist $k, \ell \in \mathbb{N}$ with $n = 3k + 7\ell$. □

We can also use strong induction to establish bounds for recursively defined sequences.

Proposition 2.2.3. *Define a sequence recursively by letting $f_0 = 0$, $f_1 = 1$, and $f_n = f_{n-1} + f_{n-2}$ for all $n \geq 2$. We have*

$$f_n \leq 2^n$$

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

Proof. We prove the result by strong induction.

- *Base Case:* We first handle the cases when $n = 0$ and $n = 1$.
  - Notice that $2^0 = 1 > 0$, so $f_0 \leq 2^0$.
  - Notice that $2^1 = 2 > 1$, so $f_1 \leq 2^1$.

  Thus, the statement is true for $n = 0$ and $n = 1$.

- *Inductive Step:* Suppose that $n \geq 2$ and the statement is true for all $k \in \mathbb{N}$ with $k < n$. In particular, we have $0 \leq n-2 < n$ and $0 \leq n-1 < n$, so the statement is true for both $n-2$ and $n-1$, and hence $f_{n-2} \leq 2^{n-2}$ and $f_{n-1} \leq 2^{n-1}$. We then have

$$\begin{aligned} f_n &= f_{n-1} + f_{n-2} && \text{(since } n \geq 2\text{)} \\ &\leq 2^{n-1} + 2^{n-2} && \text{(from above)} \\ &\leq 2^{n-1} + 2^{n-1} \\ &= 2 \cdot 2^{n-1} \\ &= 2^n. \end{aligned}$$

  Therefore, $f_n \leq 2^n$, i.e. the statement is true for n.

By strong induction, we conclude that $f_n \leq 2^n$ for all $n \in \mathbb{N}$. □

Once we have such a proof, it is natural to ask how it could be improved. A nearly identical argument shows that $f_n \leq 2^{n-1}$ for all $n \in \mathbb{N}$. However, if we try to show that $f_n \leq 2^{n-2}$ for all $n \in \mathbb{N}$, then the inductive step goes through without a problem, but the base case of $n = 1$ does not work. As a result, the argument fails.

Can we obtain a significantly better upper bound for f_n than 2^{n-1}? In particular, can we use an exponential whose base is less than 2? If we replace 2 with a number $\alpha > 1$, i.e. try to prove that $f_n \leq \alpha^n$ (or $f_n \leq \alpha^{n-1}$), then the base case goes through without a problem. In the inductive step, the key fact that we used was that $2^{n-1} + 2^{n-2} \leq 2^n$ for all $n \in \mathbb{N}$. If we replace 2 by an $\alpha > 1$ with the property that $\alpha^{n-1} + \alpha^{n-2} \leq \alpha^n$ for all $n \in \mathbb{N}$, then we can carry out the argument. Dividing through by α^{n-2}, we want to find the smallest possible $\alpha > 1$ such that $\alpha + 1 \leq \alpha^2$, which is equivalent to $\alpha^2 - \alpha - 1 \geq 0$. Using the quadratic formula, the solutions to $x^2 - x - 1 = 0$ are

$$x = \frac{1 \pm \sqrt{5}}{2}.$$

Now $\frac{1+\sqrt{5}}{2} > 1$, so we now go back and check that we can use it in an inductive argument. In fact, we can use it as a lower bound too (due to the fact that we get *equality* at the necessary step), so long as we change the exponent slightly and start with f_1.

Proposition 2.2.4. *Define a sequence recursively by letting $f_0 = 0$, $f_1 = 1$, and $f_n = f_{n-1} + f_{n-2}$ for all $n \geq 2$. Let $\phi = \frac{1+\sqrt{5}}{2}$ and notice that $\phi^2 = \phi + 1$ (either from above, or by direct calculation). We have*

$$\phi^{n-2} \leq f_n \leq \phi^{n-1}$$

for all $n \in \mathbb{N}^+$.

Proof. We prove the result by strong induction.

- *Base Case:* We first handle the cases when $n = 1$ and $n = 2$. Notice that

$$\phi = \frac{1+\sqrt{5}}{2} > \frac{1+2}{2} = \frac{3}{2},$$

hence

$$\phi^{-1} < \frac{2}{3}.$$

We also have

$$\phi = \frac{1+\sqrt{5}}{2} < \frac{1+3}{2} = 2.$$

Since $f_1 = 1 = f_2$, we have

$$\phi^{-1} < f_1 = \phi^0$$

and

$$\phi^0 = f_2 < \phi^1.$$

Therefore, the statement is true for $n = 1$ and $n = 2$.

- *Inductive Step:* Suppose that $n \geq 3$ and the statement is true for all $k \in \mathbb{N}^+$ with $k < n$. In particular, we have $1 \leq n-2 < n$ and $1 \leq n-1 < n$, so the statement is true for both $n-2$ and $n-1$, and hence

$$\phi^{n-4} \leq f_{n-2} \leq \phi^{n-3} \qquad \text{and} \qquad \phi^{n-3} \leq f_{n-1} \leq \phi^{n-2}$$

We have

$$\begin{aligned} f_n &= f_{n-1} + f_{n-2} && \text{(since } n \geq 3\text{)} \\ &\geq \phi^{n-3} + \phi^{n-4} && \text{(from above)} \\ &= \phi^{n-4}(\phi+1) \\ &= \phi^{n-4} \cdot \phi^2 \\ &= \phi^{n-2}, \end{aligned}$$

and also

$$\begin{aligned} f_n &= f_{n-1} + f_{n-2} && \text{(since } n \geq 3\text{)} \\ &\leq \phi^{n-2} + \phi^{n-3} && \text{(from above)} \\ &= \phi^{n-3}(\phi+1) \\ &= \phi^{n-3} \cdot \phi^2 \\ &= \phi^{n-1}. \end{aligned}$$

Therefore, $\phi^{n-2} \leq f_n \leq \phi^{n-1}$, i.e. the statement is true for n.

By strong induction, we conclude that $\phi^{n-2} \leq f_n \leq \phi^{n-1}$ for all $n \in \mathbb{N}^+$. □

Closely related to strong induction, the following is a core fact about the ordering of the natural numbers:

Fact 2.2.5 (Well-Ordering of $\mathbb{N}$)**.** *Every nonempty set $X \subseteq \mathbb{N}$ has a smallest element. That is, for all nonempty $X \subseteq \mathbb{N}$, there exists $m \in X$ such that $m \leq n$ for all $n \in X$.*

Why is this statement true? Suppose that $X \subseteq \mathbb{N}$ is nonempty. If $0 \in X$, then 0 is clearly the smallest element of X, and we are done. Suppose then that $0 \notin X$. If $1 \in X$, then 1 is the smallest element of X, and we are done. Suppose then that $1 \notin X$. If $2 \in X$, then 2 is the smallest element of X, and we are done. Continuing this process, we must eventually reach a point where we encounter an element X, because otherwise we would eventually argue that each fixed $n \in \mathbb{N}$ is not an element of X, which would then imply that $X = \emptyset$.

This argument, like the arguments for induction and strong induction, is intuitively reasonable and convincing. However, it is not particularly formal. It is possible to formally prove each of induction, strong induction, and well-ordering from any of the others, so in a certain precise sense the three statements are equivalent. If you're interested, think about how to prove well-ordering using induction (along with some of the other implications). However, since all three are intuitively very reasonable, and it's beyond the scope of the course to construct the natural numbers and articulate exactly what we are allowed to use in the proofs of these equivalences, we will omit the careful arguments.

Notice that the given statement is false if we consider subsets of $\mathbb{Z}$ or $\mathbb{R}$ (rather than subsets of $\mathbb{N}$). For example, $\mathbb{Z}$ is trivially a nonempty subset of $\mathbb{Z}$, but it does not have a smallest element. Even if we consider only subsets of the nonnegative reals $\{x \in \mathbb{R} : x \geq 0\}$, we can find nonempty subsets with no smallest element (for example, the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ does not have a smallest element).

We can often write an inductive proof as a proof using well-ordering, by considering a smallest potential counterexample. For example, here is a proof of Proposition 2.2.2 (if $n \in \mathbb{N}$ and $n \geq 12$, then there exist $k, \ell \in \mathbb{N}$ with $n = 3k + 7\ell$) using a well-ordering argument.

Proof of Proposition 2.2.2. Consider the set

$$X = \{n \in \mathbb{N} : n \geq 12 \text{ and there does not exist } k, \ell \in \mathbb{N} \text{ with } n = 3k + 7\ell\}$$

of counterexamples to the given statement. It suffices to show that $X = \emptyset$. Suppose instead that $X \neq \emptyset$. By well-ordering, we can let m be the smallest element of X. Notice that $m \notin \{12, 13, 14\}$ because we have the following:

- $12 = 3 \cdot 4 + 7 \cdot 0$.
- $13 = 3 \cdot 2 + 7 \cdot 1$.
- $14 = 3 \cdot 0 + 7 \cdot 2$.

Therefore, we must have $m \geq 15$, and hence $12 \leq m - 3 < 15$. Now m is the smallest element of X, so we must have $m - 3 \notin X$, and hence we can fix $k, \ell \in \mathbb{N}$ with $m - 3 = 3k + 7\ell$. Adding 3 to both sides, we see that

$$\begin{aligned} m &= 3k + 7\ell + 3 \\ &= 3(k + 1) + 5\ell. \end{aligned}$$

Since $k + 1, \ell \in \mathbb{N}$, we conclude that $m \notin X$, which is a contradiction. Therefore, it must be the case that $X = \emptyset$, giving the result. $\square$

2.3 Division with Remainder

The primary goal of this section is to prove the following deeply fundamental result.

Theorem 2.3.1. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. There exist unique $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < |b|$. Uniqueness here means that if $a = q_1 b + r_1$ with $0 \leq r_1 < |b|$ and $a = q_2 b + r_2$ with $0 \leq r_2 < |b|$, then $q_1 = q_2$ and $r_1 = r_2$.*

Here are a bunch of examples illustrating existence:

- If $a = 5$ and $b = 2$, then we have $5 = 2 \cdot 2 + 1$.
- If $a = 135$ and $b = 45$, then we have $135 = 3 \cdot 45 + 0$.
- If $a = 60$ and $b = 9$, then we have $60 = 6 \cdot 9 + 6$.
- If $a = 29$ and $b = -11$, then we have $29 = (-2)(-11) + 7$.
- If $a = -45$ and $b = 7$, then we have $-45 = (-7) \cdot 7 + 4$.
- If $a = -21$ and $b = -4$, then we have $-21 = 6 \cdot (-4) + 3$.

We begin by proving existence via a sequence of lemmas, starting in the special case where a and b are both natural numbers.

Lemma 2.3.2. *Let $a, b \in \mathbb{N}$ with $b > 0$. There exist $q, r \in \mathbb{N}$ such that $a = qb + r$ and $0 \leq r < b$.*

We give three separate proofs (induction, strong induction, and well-ordering) to illustrate the different perspectives.

Proof 1 of Lemma 2.3.2 - By Induction. Let $b \in \mathbb{N}$ with $b > 0$ be arbitrary. For this fixed b, we prove the existence of both q and r for all $a \in \mathbb{N}$ by induction. That is, for this fixed b, we define

$$X = \{a \in \mathbb{N} : \text{ There exist } q, r \in \mathbb{N} \text{ with } a = qb + r\},$$

and show that $X = \mathbb{N}$ by induction.

- *Base Case:* Suppose that $a = 0$. We then have $a = 0 \cdot b + 0$ and clearly $0 < b$, so we may take $q = 0$ and $r = 0$.
- *Inductive Step:* Let $a \in \mathbb{N}$ be arbitrary such that $a \in X$. We show that $a + 1 \in X$. Since $a \in X$, we can fix $q, r \in \mathbb{N}$ with $0 \leq r < b$ such that $a = qb + r$. We then have $a + 1 = qb + (r + 1)$. Since $b, r \in \mathbb{N}$ and $r < b$, we know that $r + 1 \leq b$. If $r + 1 < b$, then we are done. Otherwise, we have $r + 1 = b$, hence

$$\begin{aligned} a + 1 &= qb + (r + 1) \\ &= qb + b \\ &= (q + 1)b \\ &= (q + 1)b + 0, \end{aligned}$$

  so we may take $q + 1$ and 0.

By induction, we conclude that $X = \mathbb{N}$. Since b was arbitrary, the result follows. □

Proof 2 of Lemma 2.3.2 - By Strong Induction. Let $b \in \mathbb{N}$ with $b > 0$ be arbitrary. For this fixed b, we prove the existence of both q and r for all $a \in \mathbb{N}$ by strong induction. That is, for this fixed b, we define

$$X = \{a \in \mathbb{N} : \text{ There exist } q, r \in \mathbb{N} \text{ with } a = qb + r\},$$

and show that $X = \mathbb{N}$ by strong induction.

- *Base Case:* Let $a \in \mathbb{N}$ with $a < b$ be arbitrary. We then have $a = 0 \cdot b + a$ and clearly $a < b$, so we may take $q = 0$ and $r = a$.

- *Inductive Step:* Let $a \in \mathbb{N}$ with $a \geq b$ be arbitrary, and assume that $c \in X$ for all $c \in \mathbb{N}$ with $c < a$. We show that $a \in X$. Since $a \geq b$, we can subtract b from both sides to conclude that $a - b \geq 0$, and hence $a - b \in \mathbb{N}$. Also, since $b > 0$, we know that $a - b < a$. Since we have $0 \leq a - b < a$, we know that $a - b \in X$, so we can fix $q, r \in \mathbb{N}$ with $0 \leq r < b$ such that $a - b = qb + r$. Adding b to both sides, it follows that
$$\begin{aligned} a &= qb + r + b \\ &= qb + b + r \\ &= (q+1)b + r, \end{aligned}$$
so we may take $q + 1$ and r.

By strong induction, we conclude that $X = \mathbb{N}$. Since b was arbitrary, the result follows. □

Proof 3 of Lemma 2.3.2 - By Well-Ordering. Let $a, b \in \mathbb{N}$ with $b > 0$ be arbitrary. Consider the set
$$S = \{a - kb : k \in \mathbb{N}\} \cap \mathbb{N}.$$
Notice that $a \in S$ (by taking $k = 0$ and recalling that $a \in \mathbb{N}$), so $S \neq \emptyset$. By well-ordering, S has a smallest element $r \in \mathbb{N}$. Since $r \in S$, we can fix $q \in \mathbb{N}$ with $r = a - qb$. We then have that $a = qb + r$, so we need only show that $r < b$. Notice that
$$\begin{aligned} r - b &= a - qb - b \\ &= a - (q+1)b, \end{aligned}$$
so as $q + 1 \in \mathbb{N}$, it follows that $r - b \in \{a - kb : k \in \mathbb{N}\}$. Now $r - b < r$ because $b > 0$, so as r is the smallest element of S, it must be the case that $r - b \notin S$. As a result, we conclude that $r - b \notin \mathbb{N}$, so must have $r - b < 0$ (because clearly $r - b \in \mathbb{Z}$). Adding b to both sides, it follows that $r < b$. □

With this in hand, we now extend to the case where $a \in \mathbb{Z}$.

Lemma 2.3.3. *Let $a, b \in \mathbb{Z}$ with $b > 0$. There exist $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < b$.*

Proof. If $a \geq 0$, we are done by the previous lemma. Suppose that $a < 0$. We then have $-a > 0$, so by the previous lemma we may fix $q, r \in \mathbb{N}$ with $0 \leq r < b$ such that $-a = qb + r$. We then have $a = -(qb + r) = (-q)b + (-r)$. If $r = 0$, then $-r = 0$ and we are done. Otherwise we $0 < r < b$ and
$$\begin{aligned} a &= (-q)b + (-r) \\ &= (-q)b - b + b + (-r) \\ &= (-q-1)b + (b - r). \end{aligned}$$
Now since $0 < r < b$, we have $0 < b - r < b$, so this gives existence. □

And now we can extend to the case where $b < 0$.

Lemma 2.3.4. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. There exist $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < |b|$.*

Proof. If $b > 0$, we are done by the previous lemma. Suppose that $b < 0$. We then have $-b > 0$, so by the previous lemma we can fix $q, r \in \mathbb{N}$ with $0 \leq r < -b$ and $a = q(-b) + r$. We then have $a = (-q)b + r$ and we are done because $|b| = -b$. □

With that sequence of lemmas building to existence now in hand, we finish off the proof of the theorem.

Proof of Theorem 2.3.1. The final lemma above gives us existence, so we need only prove uniqueness. Let $q_1, r_1, q_2, r_2 \in \mathbb{Z}$ be arbitrary with

$$q_1 b + r_1 = a = q_2 b + r_2,$$

and where $0 \leq r_1 < |b|$ and $0 \leq r_2 < |b|$. We need to show that $q_1 = q_2$ and $r_1 = r_2$. Manipulating the above equality, we have

$$b(q_2 - q_1) = r_1 - r_2,$$

hence $b \mid (r_2 - r_1)$. Now $-|b| < -r_1 \leq 0$, so adding this to $0 \leq r_2 < |b|$, we conclude that

$$-|b| < r_2 - r_1 < |b|,$$

and therefore

$$|r_2 - r_1| < |b|.$$

Now if $r_2 - r_1 \neq 0$, then since $b \mid (r_2 - r_1)$, we can use Proposition 1.5.4 to conclude that $|b| \leq |r_2 - r_1|$, a contradiction. It follows that $r_2 - r_1 = 0$, and hence $r_1 = r_2$. Since

$$q_1 b + r_1 = q_2 b + r_2$$

and $r_1 = r_2$, we conclude that $q_1 b = q_2 b$. Now $b \neq 0$, so we can divide both sides by b to conclude that $q_1 = q_2$. □

Now that we have established the core facts about division with remainder, we can use them to give a simple check for divisibility.

Proposition 2.3.5. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Write $a = qb + r$ for the unique choice of $q, r \in \mathbb{Z}$ with $0 \leq r < |b|$. We then have that $b \mid a$ if and only if $r = 0$.*

Proof. If $r = 0$, then $a = qb + r = bq$, so $b \mid a$. Suppose conversely that $b \mid a$ and fix $m \in \mathbb{Z}$ with $a = bm$. We then have $a = mb + 0$ and $a = qb + r$, so by the uniqueness part of the above theorem, we must have $r = 0$. □

For example, we can now easily verify that $2 \nmid 5$ without any work as follows. Simply notice that $2 = 2 \cdot 2 + 1$ and $0 \leq 1 < 2$, so since the unique remainder is $1 \neq 0$, it follows that $2 \nmid 5$.

Chapter 3

GCDs, Primes, and the Fundamental Theorem of Arithmetic

3.1 The Euclidean Algorithm

Definition 3.1.1. *Suppose that $a, b \in \mathbb{Z}$. We say that $d \in \mathbb{Z}$ is a* common divisor *of a and b if both $d \mid a$ and $d \mid b$.*

We can write the set of common divisors of a and b as an intersection, i.e. given $a, b \in \mathbb{Z}$, the set of common divisors of a and b is the set $Div(a) \cap Div(b)$. For example, the set of common divisors of 120 and 84 is the set $\{\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12\}$. One way to determine the values in this set is to exhaustively determine each of the sets $Div(120)$ and $Div(84)$, and then comb through them both to find the common elements. However, we will work out a much more efficient way to solve such problems in this section.

The set of common divisors of 10 and 0 is $\{\pm 1, \pm 2, \pm 5, \pm 10\}$ because $Div(0) = \mathbb{Z}$, and hence the set of common divisors of 10 and 0 is just $Div(10) \cap Div(0) = Div(10) \cap \mathbb{Z} = Div(10)$. In contrast, every element of $\mathbb{Z}$ is a common divisor of 0 and 0, because $Div(0) \cap Div(0) = \mathbb{Z} \cap \mathbb{Z} = \mathbb{Z}$. The following little proposition is fundamental to this entire section.

Proposition 3.1.2. *Suppose that $a, b, q, r \in \mathbb{Z}$ and $a = qb + r$ (we do* not *assume that $0 \leq r < |b|$). We then have $Div(a) \cap Div(b) = Div(b) \cap Div(r)$, i.e.*

$$\{d \in \mathbb{Z} : d \text{ is a common divisor of } a \text{ and } b\} = \{d \in \mathbb{Z} : d \text{ is a common divisor of } b \text{ and } r\}.$$

Proof. Let $d \in Div(b) \cap Div(r)$ be arbitrary. Since $d \mid b$, $d \mid r$, and $a = qb + r = q \cdot b + 1 \cdot r$, we may use Proposition 1.5.3 to conclude that $d \mid a$. Therefore, $d \in Div(a) \cap Div(b)$.

Conversely, let $d \in Div(a) \cap Div(b)$ be arbitrary. Since $d \mid a$, $d \mid b$, and $r = a - qb = 1 \cdot a + (-q) \cdot b$, we may use Proposition 1.5.3 to conclude that $d \mid r$. Therefore, $d \in Div(b) \cap Div(r)$. ☐

For example, suppose that we are trying to find the set of common divisors of 120 and 84, i.e. we want to understand the elements of the set $Div(120) \cap Div(84)$ (we wrote them above, but now want to justify it). We repeatedly perform division with remainder to reduce the problem as follows:

$$\begin{aligned} 120 &= 1 \cdot 84 + 36 \\ 84 &= 2 \cdot 36 + 12 \\ 36 &= 3 \cdot 12 + 0. \end{aligned}$$

The first line tells us that
$$Div(120) \cap Div(84) = Div(84) \cap Div(36).$$
The next line tells us that
$$Div(84) \cap Div(36) = Div(36) \cap Div(12).$$
The last line tells us that
$$Div(36) \cap Div(12) = Div(12) \cap Div(0).$$
Now $Div(0) = \mathbb{Z}$, so
$$Div(12) \cap Div(0) = Div(12).$$
Putting it all together, we conclude that
$$Div(120) \cap Div(84) = Div(12),$$
which is a more elegant way to determine the set of common divisors of 120 and 84 than the exhaustive process we alluded to above.

The above arguments illustrates the idea behind the following very general and important fact:

Theorem 3.1.3. *For all $a, b \in \mathbb{Z}$, there exists a unique $m \in \mathbb{N}$ such that $Div(a) \cap Div(b) = Div(m)$. In other words, for any $a, b \in \mathbb{Z}$, we can always find a (unique) natural number m such that the set of common divisors of a and b equals the set of divisors of m.*

We first sketch the idea of the proof of existence in the case where $a, b \in \mathbb{N}$. If $b = 0$, then since $Div(0) = \mathbb{Z}$, we can simply take $m = a$. Suppose then that $b \neq 0$. Fix $q, r \in \mathbb{N}$ with $a = qb + r$ and $0 \leq r < b$. Now the idea is to inductively assert the existence of an m that works for the pair of numbers (b, r) because this pair is "smaller" than the pair (a, b). The only issue is how to make this intuitive idea of "smaller" precise. There are several ways to do this, but perhaps the most straightforward is to only induct on b. Thus, our base case handles all pairs of form $(a, 0)$. Next, we handle all pairs of the form $(a, 1)$ and in doing this we can use the fact the we know the result for all pairs of the form $(a', 0)$. Notice that we can we even change the value of the first coordinate here, which is why we used the notation a'. Then, we handle all pairs of the form $(a, 2)$ and in doing this we can use the fact that we know the result for all pairs of the form $(a', 0)$ and $(a', 1)$. We now carry out the formal argument.

Proof. We begin by proving existence only in the special case where $a, b \in \mathbb{N}$. We use (strong) induction on b to prove the result. That is, we let
$$X = \{b \in \mathbb{N} : \text{For all } a \in \mathbb{N}, \text{ there exists } m \in \mathbb{N} \text{ with } Div(a) \cap Div(b) = Div(m)\}$$
and prove that $X = \mathbb{N}$ by strong induction.

- *Base Case:* Suppose that $b = 0$. Let $a \in \mathbb{N}$ be arbitrary. We then have that $Div(b) = \mathbb{Z}$, so
$$Div(a) \cap Div(b) = Div(a) \cap \mathbb{Z} = Div(a),$$
and hence we may take $m = a$. Since $a \in \mathbb{N}$ was arbitrary, we showed that $0 \in X$.

- *Inductive Step:* Let $b \in \mathbb{N}^+$ be arbitrary, and suppose that we know that the statement is true for all smaller natural numbers. In other words, we are assuming that $c \in X$ whenever $0 \leq c < b$. We prove that $b \in X$. Let $a \in \mathbb{N}$ be arbitrary. From above, we may fix $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < b$. Since $0 \leq r < b$, we know by strong induction that $r \in X$, so we can fix $m \in \mathbb{N}$ with
$$Div(b) \cap Div(r) = Div(m).$$
By Proposition 3.1.2, we have that $Div(a) \cap Div(b) = Div(b) \cap Div(r)$. Therefore, $Div(a) \cap Div(b) = Div(m)$. Since $a \in \mathbb{N}$ was arbitrary, we showed that $b \in X$.

Therefore, we have shown that $X = \mathbb{N}$, which implies that whenever $a, b \in \mathbb{N}$, there exists $m \in \mathbb{N}$ such that $Div(a) \cap Div(b) = Div(m)$.

To prove the existence statement more generally when $a, b \in \mathbb{Z}$, we use Proposition 1.5.7. So, for example, if $a < 0$ but $b \geq 0$, we can fix $m \in \mathbb{N}$ with $Div(-a) \cap Div(b) = Div(m)$, and then use the fact that $Div(a) = Div(-a)$ to conclude that $Div(a) \cap Div(b) = Div(m)$. A similar argument works if $a \geq 0$ and $b < 0$, or if both $a < 0$ and $b < 0$.

For uniqueness, suppose that $m, n \in \mathbb{N}$ are such that both $Div(a) \cap Div(b) = Div(m)$ and also $Div(a) \cap Div(b) = Div(n)$. We then have that $Div(m) = Div(n)$. Since $m \in Div(m)$ trivially, we have that $m \in Div(n)$, so $m \mid n$. Similarly, we have $n \mid m$. Therefore, by Corollary 1.5.5, either $m = n$ or $m = -n$. Since $m, n \in \mathbb{N}$, we have $m \geq 0$ and $n \geq 0$, so it must be the case that $m = n$. □

With this great result in hand, we now turn our attention to a fundamental concept: greatest common divisors. Given $a, b \in \mathbb{Z}$, one might be tempted to define the greatest common divisor of a and b to be the *largest* natural number that divides both a and b (after all, the name *greatest* surely suggests this!). However, it turns out that this is a poor definition for several reasons:

1. Consider the case $a = 120$ and $b = 84$ from above. We saw that the set of common divisors of 120 and 84 is $Div(120) \cap Div(84) = Div(12)$. Thus, the *largest* natural number that divides both 120 and 84 is 12, but in fact 12 has a much stronger property: *every* common divisor of 120 and 84 is also a divisor of 12. This stronger property is surprising and much more fundamental that simply being the largest common divisor.

2. There is one pair of integers where no largest common divisors exists! In the trivial case where $a = 0$ and $b = 0$, *every* integer is a common divisor of a and b. Although this is a somewhat silly edge case, we would ideally like a definition that handles all cases elegantly.

3. The integers have a natural ordering associated with them, but we will eventually want to generalize the idea of a greatest common divisor to settings where there is no analogue of $\leq$ (see Abstract Algebra).

With all of this background in mind, we now give our formal definition.

Definition 3.1.4. *Let $a, b \in \mathbb{Z}$. We say that an element $m \in \mathbb{Z}$ is a* greatest common divisor *of a and b if the following are all true:*

- *$m \geq 0$.*
- *m is a common divisor of a and b.*
- *Whenever $d \in \mathbb{Z}$ is a common divisor of a and b, we have $d \mid m$.*

In other words, a greatest common divisor is a natural number, is a common divisor, and has the property that every common divisors happens to divide it. In terms of point 3 above, it is a straightforward matter to check that 0 is in fact a greatest common divisor of 0 and 0, because every element of $Div(0) \cap Div(0) = \mathbb{Z}$ is a divisor of 0.

Since we require more of a greatest common divisor than just picking the largest, we first need to check that they do indeed exist. However, the next proposition reduces this task to our previous work.

Proposition 3.1.5. *Let $a, b \in \mathbb{Z}$ and let $m \in \mathbb{N}$. The following are equivalent:*

1. *$Div(a) \cap Div(b) = Div(m)$.*
2. *m is a greatest common divisor of a and b.*

Proof. We first prove $1 \rightarrow 2$: Suppose then that $Div(a) \cap (b) = Div(m)$. Since we are assuming that $m \in \mathbb{N}$, we have that $m \geq 0$. Since $m \mid m$, we have $m \in Div(m)$, so $m \in Div(a) \cap Div(b)$, and hence both $m \mid a$ and $m \mid b$. Now let $d \in \mathbb{Z}$ be an arbitrary common divisor of a and b. We then have that both $d \mid a$ and $d \mid b$, so $d \in Div(a) \cap Div(b)$, hence $d \in Div(m)$, and therefore $d \mid m$. Putting it all together, we conclude that m is a greatest common divisor of a and b.

We now prove $2 \rightarrow 1$: Suppose that m is a greatest common divisor of a and b. We need to prove that $Div(a) \cap Div(b) = Div(m)$.

- We first show that $Div(a) \cap Div(b) \subseteq Div(m)$. Let $d \in Div(a) \cap Div(b)$ be arbitrary. We then have both $d \mid a$ and $d \mid b$, so since m is a greatest common divisor of a and b, we conclude that $d \mid m$. Therefore, $d \in Div(m)$.

- We now show that $Div(m) \subseteq Div(a) \cap Div(b)$. Let $d \in Div(m)$ be arbitrary, so $d \mid m$. Now we know that m is a common divisor of a and b, so both $m \mid a$ and $m \mid b$. Using Proposition 1.5.2, we conclude that both $d \mid a$ and $d \mid b$, so $d \in Div(a) \cap Div(b)$.

Since we have shown both $Div(a) \cap Div(b) \subseteq Div(m)$ and $Div(m) \subseteq Div(a) \cap Div(b)$, we conclude that $Div(a) \cap Div(b) \subseteq Div(m)$. □

Corollary 3.1.6. *Every pair of integers $a, b \in \mathbb{Z}$ has a unique greatest common divisor.*

Proof. Immediate from Theorem 3.1.3 and Proposition 3.1.5. □

Definition 3.1.7. *Let $a, b \in \mathbb{Z}$. We let $\gcd(a, b)$ be the unique greatest common divisor of a and b.*

For example we have $\gcd(120, 84) = 12$ and $\gcd(0, 0) = 0$. The following corollary now follows from Proposition 3.1.2.

Corollary 3.1.8. *Suppose that $a, b, q, r \in \mathbb{Z}$ and $a = qb + r$. We have $\gcd(a, b) = \gcd(b, r)$.*

The method of using repeated division with remainder, together with this corollary, to reduce the problem of calculating greatest common divisors is known as the *Euclidean Algorithm*. We saw it in action of above with 120 and 84. Here is another example where we are trying to compute $\gcd(525, 182)$. We have:

$$\begin{aligned}
525 &= 2 \cdot 182 + 161 \\
182 &= 1 \cdot 161 + 21 \\
161 &= 7 \cdot 21 + 14 \\
21 &= 1 \cdot 14 + 7 \\
14 &= 2 \cdot 7 + 0,
\end{aligned}$$

so $\gcd(525, 182) = \gcd(7, 0) = 7$.

Let $a, b \in \mathbb{Z}$. Consider the set

$$\{ka + \ell b : k, \ell \in \mathbb{Z}\}.$$

This looks something like the *span* that you saw in Linear Algebra, but here we are only using integer coefficients, so we could describe this as the set of all *integer* combinations of a and b. Notice that if d is a common divisor of a and b, then $d \mid (ka + \ell b)$ for all $k, \ell \in \mathbb{Z}$ by Proposition 1.5.3, and hence d divides every element of this set. Applying this fact in the most interesting case where $d = \gcd(a, b)$ (since all other common divisors of a and b will divide $\gcd(a, b)$), we conclude that every element of $\{ka + \ell b : k, \ell \in \mathbb{Z}\}$ is a multiple of $\gcd(a, b)$. In other words, we have

$$\{ka + \ell b : k, \ell \in \mathbb{Z}\} \subseteq \{n \cdot \gcd(a, b) : n \in \mathbb{Z}\}.$$

What about the reverse containment? In particular, is $\gcd(a,b)$ always an element of $\{ka + \ell b : k, \ell \in \mathbb{Z}\}$? For example, is
$$12 \in \{k \cdot 120 + \ell \cdot 84 : k, \ell \in \mathbb{Z}\}?$$

We can attempt to play around to try to find a suitable value of k and ℓ, but there is a better way. Let's go back and look at the steps of the Euclidean Algorithm:

$$\begin{aligned} 120 &= 1 \cdot 84 + 36 \\ 84 &= 2 \cdot 36 + 12 \\ 36 &= 3 \cdot 12 + 0. \end{aligned}$$

Notice that the middle line can be manipulated to write 12 as an integer combination of 84 and 36:

$$12 = 1 \cdot 84 + (-2) \cdot 36.$$

With this in hand, we can work our way toward our goal by using the first line, which lets us write 36 as an integer combination of 120 and 84:
$$36 = 1 \cdot 120 + (-1) \cdot 84.$$

Now we can plug this expression of 36 in terms of 120 and 84 into the previous equation:

$$12 = 1 \cdot 84 + (-2) \cdot [1 \cdot 120 + (-1) \cdot 84].$$

From here, we can manipulate this equation (performing only additions and multiplications on the coefficients, not on 84 and 120 themselves!) to obtain

$$12 = (-2) \cdot 120 + 3 \cdot 84.$$

We now generalize this idea and prove that it is always possible to express $\gcd(a,b)$ as an integer combination of a and b. The proof is inductive, and follows a similar strategy to the proof of Theorem 3.1.3. Given $a, b \in \mathbb{N}$, here is the idea. To express $\gcd(a,b)$ as an integer combination of a and b, we first fix $q, r \in \mathbb{N}$ with $a = qb + r$ and $0 \leq r < b$. Now since (b, r) is "smaller" than (a, b), we inductively write $\gcd(b, r)$ as an integer combination of b and r. We then use this combination together with the equation $a = qb + r$ to write $\gcd(a,b)$ as an integer combination of a and b. Notice the similarity to the above argument where we have $120 = 1 \cdot 84 + 36$, and we used a known way to write 12 as an integer combination of 84 and 36 in order to write 12 as an integer combination of 120 and 84.

Theorem 3.1.9. *For all $a, b \in \mathbb{Z}$, there exist $k, \ell \in \mathbb{Z}$ with $\gcd(a,b) = ka + \ell b$.*

Proof. We begin by proving existence in the special case where $a, b \in \mathbb{N}$. We use induction on b to prove the result. That is, we let

$$X = \{b \in \mathbb{N} : \text{For all } a \in \mathbb{N}, \text{ there exist } k, \ell \in \mathbb{Z} \text{ with } \gcd(a,b) = ka + \ell b\}$$

and prove that $X = \mathbb{N}$ by strong induction.

- *Base Case:* Suppose that $b = 0$. Let $a \in \mathbb{N}$ be arbitrary. We then have that

  $$\gcd(a,b) = \gcd(a,0) = a$$

  Since $a = 1 \cdot a + 0 \cdot b$, so we may let $k = 1$ and $\ell = 0$. Since $a \in \mathbb{N}$ was arbitrary, we conclude that $0 \in X$.

- *Inductive Step:* Suppose then that $b \in \mathbb{N}^+$ and we know the result for all smaller nonnegative values. In other words, we are assuming that $c \in X$ whenever $0 \leq c < b$. We prove that $b \in X$. Let $a \in \mathbb{N}$ be arbitrary. From above, we may fix $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < b$. We also know from above that $\gcd(a, b) = \gcd(b, r)$. Since $0 \leq r < b$, we know by strong induction that $r \in X$, hence there exist $k, \ell \in \mathbb{Z}$ with

$$\gcd(b, r) = kb + \ell r$$

Now $r = a - qb$, so

$$\begin{aligned}\gcd(a, b) &= \gcd(b, r) \\ &= kb + \ell r \\ &= kb + \ell(a - qb) \\ &= kb + \ell a - qb\ell \\ &= \ell a + (k - q\ell)b.\end{aligned}$$

Since $a \in \mathbb{N}$ was arbitrary, we conclude that $b \in X$.

Therefore, we have shown that $X = \mathbb{N}$, which implies that whenever $a, b \in \mathbb{N}$, there exists $k, \ell \in \mathbb{Z}$ with $\gcd(a, b) = ka + \ell b$.

To prove the result more generally when $a, b \in \mathbb{Z}$, we again use Proposition 1.5.7. For example, if $a < 0$ but $b \geq 0$. Let $m = \gcd(a, b)$, so that $Div(m) = Div(a) \cap Div(b)$ by Proposition 3.1.5. Since $Div(-a) = Div(a)$, we also have $Div(m) = Div(-a) \cap Div(b)$, hence $m = \gcd(-a, b)$. Since $-a, b \in \mathbb{N}$, we can fix $k, \ell \in \mathbb{Z}$ with $\gcd(-a, b) = k(-a) + \ell b$. Using the fact that $\gcd(-a, b) = \gcd(a, b)$, we have $\gcd(a, b) = k(-a) + \ell b$, hence $\gcd(a, b) = (-k)a + \ell b$. Since $-k, \ell \in \mathbb{Z}$, we are done. A similar argument works if $a \geq 0$ and $b < 0$, or if both $a < 0$ and $b < 0$. □

Notice the basic structure of the above proof. If $a = qb + r$, and we happen to know $k, \ell \in \mathbb{Z}$ such that

$$\gcd(b, r) = kb + \ell r,$$

then we have

$$\gcd(a, b) = \ell a + (k - q\ell)b.$$

Given $a, b \in \mathbb{Z}$, this argument provides a recursive procedure in order to find an integer combination of a and b that gives $\gcd(a, b)$. Although the recursive procedure can be nicely translated to a computer program, we can carry it out directly by "winding up" the work created from the Euclidean Algorithm. For example, we saw above that $\gcd(525, 182) = 7$ by calculating:

$$\begin{aligned}525 &= 2 \cdot 182 + 161 \\ 182 &= 1 \cdot 161 + 21 \\ 161 &= 7 \cdot 21 + 14 \\ 21 &= 1 \cdot 14 + 7 \\ 14 &= 2 \cdot 7 + 0.\end{aligned}$$

We now use these steps in reverse to calculate:

$$\begin{aligned}
7 &= 1 \cdot 7 + 0 \cdot 0 \\
&= 1 \cdot 7 + 0 \cdot (14 - 2 \cdot 7) \\
&= 0 \cdot 14 + 1 \cdot 7 \\
&= 0 \cdot 14 + 1 \cdot (21 - 1 \cdot 14) \\
&= 1 \cdot 21 + (-1) \cdot 14 \\
&= 1 \cdot 21 + (-1) \cdot (161 - 7 \cdot 21) \\
&= (-1) \cdot 161 + 8 \cdot 21 \\
&= (-1) \cdot 161 + 8 \cdot (182 - 1 \cdot 161) \\
&= 8 \cdot 182 + (-9) \cdot 161 \\
&= 8 \cdot 182 + (-9) \cdot (525 - 2 \cdot 182) \\
&= (-9) \cdot 525 + 26 \cdot 182.
\end{aligned}$$

This wraps everything up perfectly, but it is easier to simply start at the fifth line.

Now that we've showed that $\gcd(a,b) \in \{ka + \ell b : k, \ell \in \mathbb{Z}\}$ for all $a, b \in \mathbb{Z}$, we can now completely characterize the set of all integer combinations of a and b.

Corollary 3.1.10. *For all $a, b \in \mathbb{Z}$, we have $\{ka + \ell b : k, \ell \in \mathbb{Z}\} = \{n \cdot \gcd(a,b) : n \in \mathbb{Z}\}$.*

Proof. Let $a, b \in \mathbb{Z}$ be arbitrary. Let $m = gcd(a,b)$. We give a double containment proof.

- $\{ka + \ell b : k, \ell \in \mathbb{Z}\} \subseteq \{nm : n \in \mathbb{Z}\}$: Let $c \in \{ka + \ell b : k, \ell \in \mathbb{Z}\}$ be arbitrary, and fix $k, \ell \in \mathbb{Z}$ with $c = ka + \ell b$. Since $m = \gcd(a,b)$, we have both $m \mid a$ and $m \mid b$. Using Proposition 1.5.3, we conclude that $m \mid c$. Therefore, we can fix $n \in \mathbb{Z}$ with $c = mn$, and hence $c \in \{nm : n \in \mathbb{Z}\}$.

- $\{nm : n \in \mathbb{Z}\} \subseteq \{ka + \ell b : k, \ell \in \mathbb{Z}\}$: Let $c \in \{nm : n \in \mathbb{Z}\}$ be arbitrary, and fix $n \in \mathbb{Z}$ with $c = nm$. Since $m = \gcd(a,b)$, we can use Theorem 3.1.9 to fix $k, \ell \in \mathbb{Z}$ with $m = ka + \ell b$. Multiplying both sides of this equation by n, we have $nm = nka + n\ell b$, so $c = (nk)a + (n\ell)b$. Since $nk, n\ell \in \mathbb{Z}$, it follows that $\{nm : n \in \mathbb{Z}\} \subseteq \{ka + \ell b : k, \ell \in \mathbb{Z}\}$.

Since we have shown both containments, it follows that $\{ka + \ell b : k, \ell \in \mathbb{Z}\} = \{n \cdot \gcd(a,b) : n \in \mathbb{Z}\}$. □

Before moving on, we work through another proof of the existence of greatest common divisors, along with the fact that we can write $\gcd(a,b)$ as an integer combination of a and b. This proof also works because of Theorem 2.3.1, but it uses well-ordering and establishes existence without a method of computation. One may ask why we bother with another proof. One answer is that this result is so fundamental and important that two different proofs help to reinforce its value. Another reason is that each proof generalizes in different ways in more abstract settings (see Abstract Algebra).

Theorem 3.1.11. *Let $a, b \in \mathbb{Z}$ with at least one of a and b nonzero. The set*

$$\{ka + \ell b : k, \ell \in \mathbb{Z}\}$$

has positive elements, and the least positive element is a greatest common divisor of a and b. In particular, for any $a, b \in \mathbb{Z}$, there exist $k, \ell \in \mathbb{Z}$ with $\gcd(a,b) = ka + \ell b$.

Proof. Let

$$S = \{ka + \ell b : k, \ell \in \mathbb{Z}\} \cap \mathbb{N}^+.$$

We first claim that $S \neq \emptyset$. If $a > 0$, then $a = 1 \cdot a + 0 \cdot b \in S$. Similarly, if $b > 0$, then $b \in S$. If $a < 0$, then $-a > 0$ and $-a = (-1) \cdot a + 0 \cdot b \in S$. Similarly, if $b < 0$, then $-b \in S$. Since at least one of a and b is nonzero, it follows that $S \neq \emptyset$. By the Well-Ordering property of $\mathbb{N}$, we know that S has a least element. Let $m = \min(S)$. Since $m \in S$, we may fix $k, \ell \in \mathbb{Z}$ with $m = ka + \ell b$. We claim that m is a greatest common divisor of a and b.

First, we need to check that m is a common divisor of a and b. We begin by showing that $m \mid a$. Fix $q, r \in \mathbb{Z}$ with $a = qm + r$ and $0 \leq r < m$. We want to show that $r = 0$. We have

$$\begin{aligned} r &= a - qm \\ &= a - m(ak + b\ell) \\ &= (1 - qk) \cdot a + (-q\ell) \cdot b. \end{aligned}$$

Now if $r > 0$, then we have shown that $r \in S$, which contradicts the choice of m as the least element of S. Hence, we must have $r = 0$, and so $m \mid a$.

We next show that $m \mid b$. Fix $q, r \in \mathbb{Z}$ with $b = qm + r$ and $0 \leq r < m$. We want to show that $r = 0$. We have

$$\begin{aligned} r &= b - qm \\ &= b - q(ak + b\ell) \\ &= (-qk) \cdot a + (1 - q\ell) \cdot b. \end{aligned}$$

Now if $r > 0$, then we have shown that $r \in S$, which contradicts the choice of m as the least element of S. Hence, we must have $r = 0$, and so $m \mid b$.

Finally, we need to check the last condition for m to be the greatest common divisor. Let d be a common divisor of a and b. Since $d \mid a$, $d \mid b$, and $m = ka + \ell b$, we may use Proposition 1.5.3 to conclude that $d \mid m$. □

3.2 Primes and Relatively Prime Integers

We start by defining prime numbers. We choose to only consider positive natural numbers, and we also rule out the number 1 for reasons that we will explain later.

Definition 3.2.1. *An element $p \in \mathbb{Z}$ is prime if $p > 1$ and the only positive divisors of p are 1 and p. If $n \in \mathbb{Z}$ with $n > 1$ is not prime, we say that n is* composite.

Since $a \mid b$ if and only if $-a \mid b$, we can equivalently say that an integer $p > 1$ is prime if $Div(p) = \{\pm 1, \pm p\}$, which is also equivalent to saying that $|Div(p)| = 4$ (or that $|Div(p) \cap \mathbb{N}| = 2$). Notice that 2 is prime, because if $d \in \mathbb{N}^+$ is such that $d \mid 2$, then $1 \leq d \leq 2$ by Proposition 1.5.4.

Proposition 3.2.2. *If $n \in \mathbb{Z}$ and $n \notin \{1, -1\}$, then there exists a prime $p \in \mathbb{N}$ with $p \mid n$.*

Proof. First notice that 2 is prime, and that $2 \mid 0$, so the statement is true for $n = 0$. Thus, by Problem 6 on Homework 2 (that $d \mid n$ if and only if $d \mid -n$), it suffices to prove the result for $n \in \mathbb{N}$ with $n \geq 2$. We do this by strong induction.

- *Base Case:* When $n = 2$, we have that 2 is prime and $2 \mid n$ trivially.
- *Inductive Step:* Let $n \in \mathbb{N}$ with $n \geq 2$ be arbitrary such that statement is true for all natural numbers k with $2 \leq k < n$. We have two cases:
  - *Case 1:* Suppose that n is prime. We have that $n \mid n$ trivially, so in this case we can just take n.
  - *Case 2:* Suppose instead that n is not prime. By definition, we can fix $d \in \mathbb{N}$ with $d \notin \{1, n\}$ such that $d \mid n$. By Proposition 1.5.4, we have $1 \leq d \leq n$, and hence $2 \leq d < n$. By induction, we can fix a prime p with $p \mid d$. By transitivity of divisibility (Proposition 1.5.2), it follows that $p \mid n$.

Thus, in either case, we have shown that there exists a prime p with $p \mid n$.

By induction, we conclude that every $n \in \mathbb{N}$ with $n \geq 2$ is divisible by some prime, which suffices by the above discussion. □

Proposition 3.2.3. *There are infinitely many primes.*

Proof. We know that 2 is a prime, so there is at least one prime. We will take an arbitrary given finite list of primes and show that there exists a prime which is omitted. Suppose then that $p_1, p_2, \dots, p_k$ is an arbitrary finite list of prime numbers with $k \geq 1$. We show that there exists a prime not in the list. Let

$$n = p_1 p_2 \cdots p_k + 1$$

We have $n \geq 3$, so by Proposition 3.2.2 we can fix a prime q with $q \mid n$. Suppose, for the sake of obtaining a contradiction, that $q = p_i$ for some i. We then have that $q \mid n$ and also $q \mid p_1 p_2 \cdots p_k$, so $q \mid (n - p_1 p_2 \cdots p_k)$. However, this implies that $q \mid 1$, so $|q| \leq 1$ by Proposition 1.5.4, a contradiction. Therefore $q \neq p_i$ for all i, and we have succeeded in finding a prime not in the list. □

While primality is a property of certain numbers, there is another closely related property of pairs of numbers.

Definition 3.2.4. *Two integers $a, b \in \mathbb{Z}$ are* relatively prime *if* $\gcd(a, b) = 1$.

For example, we have that 40 and 33 are relatively prime (despite neither number itself being prime), either by exhaustively checking divisors, or using the Euclidean Algorithm:

$$\begin{aligned}
40 &= 1 \cdot 33 + 7 \\
33 &= 4 \cdot 7 + 5 \\
7 &= 1 \cdot 5 + 2 \\
5 &= 2 \cdot 2 + 1 \\
2 &= 2 \cdot 1 + 0.
\end{aligned}$$

Thus, $\gcd(40, 33) = \gcd(1, 0) = 1$.

Notice that if $a \mid bc$, then it might not be the case that either $a \mid b$ or $a \mid c$. For example, we have $6 \mid 10 \cdot 9$, but $6 \nmid 10$ and $6 \mid 9$. The next result says that if $a \mid bc$, and a and b are relatively prime, then we can eliminate the b to conclude that $a \mid c$. To prove this fundamental and powerful result, we will make use of all of our hard work from the last section.

Proposition 3.2.5. *Let $a, b, c \in \mathbb{Z}$. If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.*

Proof. Since $a \mid bc$, we may fix $n \in \mathbb{Z}$ with $bc = an$. Since $\gcd(a, b) = 1$, we can use Theorem 3.1.9 to fix $k, \ell \in \mathbb{Z}$ with $ak + b\ell = 1$. Multiplying this equation through by c we conclude that $akc + b\ell c = c$, so

$$\begin{aligned}
c &= akc + \ell(bc) \\
&= akc + \ell(an) \\
&= a(kc + \ell n).
\end{aligned}$$

Since $kc + \ell n \in \mathbb{Z}$, it follows that $a \mid c$. □

We can quickly obtain the following consequence, which is one of the most useful facts about prime numbers.

Corollary 3.2.6. *Let $p, a, b \in \mathbb{Z}$. If $p \in \mathbb{Z}$ is prime and $p \mid ab$, then either $p \mid a$ or $p \mid b$.*

Proof. Suppose that $p \mid ab$ and $p \nmid a$. Since $\gcd(a,p)$ divides p and we know that $p \nmid a$, we have $\gcd(a,p) \neq p$. As p is prime, the only other positive divisor of p is 1, so $\gcd(a,p) = 1$. Therefore, by the Proposition 3.2.5, we conclude that $p \mid b$. □

Now that we've handled the product of two numbers, we get the following corollary about finite products by a trivial induction.

Corollary 3.2.7. *Let $p, a_1, a_2, \dots, a_n \in \mathbb{Z}$. If p is prime and $p \mid a_1 a_2 \cdots a_n$, then $p \mid a_i$ for some i.*

Considering the special case when all of the a_i are equal, we obtain the following.

Corollary 3.2.8. *Let $p, a \in \mathbb{Z}$ and $n \in \mathbb{N}^+$. If p is prime and $p \mid a^n$, then $p \mid a$.*

Let $a, b \in \mathbb{Z}$. We know from Theorem 3.1.9 that there exists $k, \ell \in \mathbb{Z}$ with $ka + \ell b = \gcd(a,b)$. However, be careful to note that if we find $d, k, \ell \in \mathbb{Z}$ with $ak + b\ell = d$, then it need not be the case that $d = \gcd(a,b)$. Using Corollary 3.1.10, all that we can conclude in this case is that d is a *multiple* of $\gcd(a,b)$. Nonetheless, since 1 is only a multiple of 1 and -1, if we do happen to find $k, \ell \in \mathbb{Z}$ with $ak + b\ell = 1$, then we can indeed conclude that $\gcd(a,b) = 1$. We include this equivalent condition, along with another than only looks at common prime divisors, in the following result.

Proposition 3.2.9. *Let $a, b \in \mathbb{Z}$. The following are equivalent:*

1. $\gcd(a,b) = 1$*, i.e. a and b are relatively prime.*
2. *There exist $k, \ell \in \mathbb{Z}$ with $ak + b\ell = 1$.*
3. *There is no prime $p \in \mathbb{N}$ with both $p \mid a$ and $p \mid b$.*

Proof. We prove that $(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)$.

- $(1) \Rightarrow (2)$: This follows immediately from Theorem 3.1.9.
- $(2) \Rightarrow (3)$: Suppose that statement (2) is true, and fix $k, \ell \in \mathbb{Z}$ with $ak + b\ell = 1$. Let $d \in \mathbb{Z}$ be arbitrary such that both $d \mid a$ and $d \mid b$. Using Proposition 1.5.3, we can immediately conclude that $d \mid ak + b\ell$, so $d \mid 1$, and hence $|d| \leq 1$ by Proposition 1.5.4. Thus, every common divisor of a and b has absolute value less than or equal to 1. Since all primes are greater than 1 by definition, we conclude that there is no prime $p \in \mathbb{N}$ with both $p \mid a$ and $p \mid b$.
- $(3) \Rightarrow (1)$: We prove the contrapositive. Suppose then that statement (1) is false, i.e. that $\gcd(a,b) \neq 1$. Since $\gcd(a,b) \geq 0$ by definition, we then have that $\gcd(a,b) \notin \{1, -1\}$. Therefore, by Proposition 3.2.2, we can fix a prime $p \in \mathbb{N}$ with $p \mid \gcd(a,b)$. Since $\gcd(a,b)$ is a common divisor of a and b, we can use Proposition 1.5.2 to conclude that p is a common divisor of a and b. Thus, we have shown that statement (3) is false.

□

Corollary 3.2.10. *Let $a, b \in \mathbb{Z}$ and let $k, \ell \in \mathbb{N}^+$. If $\gcd(a,b) = 1$, then $\gcd(a^k, b^\ell) = 1$.*

Proof. Suppose that $\gcd(a,b) = 1$. If $p \in \mathbb{Z}$ was a prime such that both $p \mid a^k$ and $p \mid b^\ell$, then we would have both $p \mid a$ and $p \mid b$ by Corollary 3.2.8, contradicting the fact that $\gcd(a,b) = 1$. Therefore, there is no common prime divisor of a^k and b^ℓ, so $\gcd(a^k, b^\ell) = 1$ by Proposition 3.2.9. □

Corollary 3.2.11. *If $p_1, \dots, p_m, q_1, \dots, q_n \in \mathbb{Z}$ are distinct primes, and $k_1, \dots, k_m, \ell_1, \dots, \ell_m \in \mathbb{N}$, then $\gcd(p_1^{k_1} \cdots p_m^{k_m}, q_1^{\ell_1} \cdots q_n^{\ell_n}) = 1$.*

Proof. Suppose that $r \in \mathbb{Z}$ is a common prime divisor of $p_1^{k_1} \cdots p_m^{k_m}$ and $q_1^{\ell_1} \cdots q_n^{\ell_n}$. By Corollary 3.2.7, we can fix an i with $r \mid p_i$ and we can fix a j with $r \mid q_j$ As p_i is prime, the only positive divisors of p_i are 1 and p_i, so since $r > 1$, we must have $r = p_i$. Similarly, as q_j is prime, we must have $r = q_j$. Therefore, $p_i = q_j$, contradicting the fact that $p_1, \dots, p_m, q_1, \dots, q_n \in \mathbb{Z}$ are distinct primes. Hence, there is no common prime divisor of $p_1^{k_1} \cdots p_m^{k_m}$ and $q_1^{\ell_1} \cdots q_n^{\ell_n}$, so $\gcd(p_1^{k_1} \cdots p_m^{k_m}, q_1^{\ell_1} \cdots q_n^{\ell_n}) = 1$ by Proposition 3.2.9. □

3.3 Determining the Set of Divisors

Let $a \in \mathbb{Z}\backslash\{0\}$. We know from Proposition 1.5.4 that $Div(a) \subseteq \{b \in \mathbb{Z} : |b| \leq |a|\}$, so $Div(a)$ is finite. Suppose that we want to determine $Div(a)$. Of course, we could do an exhaustive search, checking each element of $\{b \in \mathbb{Z} : |b| \leq |a|\}$ individually to determine whether it belongs to $Div(a)$. However, we can do better in several ways. First, since $0 \nmid a$ and $d \mid a$ if and only if $-d \mid a$, it suffices to determine the set $Div(a) \cap \mathbb{N}^+$, i.e. the set of *positive* divisors of a. Now if p is prime, then $Div(p) \cap \mathbb{N}^+ = \{1, p\}$, and we are done. With a bit more work, we can determine $Div(a) \cap \mathbb{N}^+$ whenever a is a power of a prime.

Proposition 3.3.1. *For all primes $p \in \mathbb{N}$ and all $n \in \mathbb{N}^+$, we have $Div(p^n) \cap \mathbb{N}^+ = \{p^k : 0 \leq k \leq n\}$.*

Proof. Let $p \in \mathbb{N}$ be an arbitrary prime. For this fixed prime p, we prove the statement by induction on n.

- *Base Case:* Suppose that $n = 1$. Since p is prime, we know by definition that the positive divisors of p are exactly $1 = p^0$ and $p = p^1$. Therefore, $Div(p^n) \cap \mathbb{N}^+ = \{p^0, p^1\} = \{p^k : 0 \leq k \leq 1\}$.

- *Inductive Step:* Assume that the statement is true for a given $n \in \mathbb{N}^+$, i.e. assume that $Div(p^n) \cap \mathbb{N}^+ = \{p^k : 0 \leq k \leq n\}$. First notice that for any $k \in \mathbb{N}$ with $0 \leq k \leq n+1$, we have $n+1-k \geq 0$ and $p^{n+1} = p^k \cdot p^{n+1-k}$, so $p^k \mid p^{n+1}$. Thus, $\{p^k : 0 \leq k \leq n+1\} \subseteq Div(p^{n+1}) \cap \mathbb{N}^+$. We now prove the reverse containment. Let $a \in Div(p^{n+1}) \cap \mathbb{N}^+$ be arbitrary. By definition, we can fix $b \in \mathbb{Z}$ with $p^{n+1} = ab$. We have two cases:

  - *Case 1:* Suppose that $p \mid a$. By definition, we can fix $c \in \mathbb{Z}$ with $a = pc$. Notice that $c \in \mathbb{N}^+$ because $a, p \in \mathbb{N}^+$. Since $p^{n+1} = ab$ and $a = pc$, we have $p^{n+1} = pcb$, so dividing both sides by p, we conclude that $p^n = cb$. Since $b \in \mathbb{Z}$, it follows that $c \mid p^n$. Therefore, $c \in Div(p^n) \cap \mathbb{N}^+$, so by induction, we know that $c \in \{p^k : 0 \leq k \leq n\}$. Fix $\ell \in \mathbb{N}$ with $0 \leq \ell \leq n$ such that $c = p^\ell$. We then have $a = pc = pp^\ell = p^{\ell+1}$, so $a \in \{p^k : 1 \leq k \leq n+1\}$, and hence $a \in \{p^k : 0 \leq k \leq n+1\}$.
  - *Case 2:* Suppose then that $p \nmid a$. Since $\gcd(a, p)$ is a nonnegative common divisor of p and a, and the only nonnegative divisors of p are 1 and p, it follows that $\gcd(a, p) = 1$. Now we have that $a \mid p^{n+1}$, so $a \mid p \cdot p^n$. Since $\gcd(a, p) = 1$, we can use Proposition 3.2.5 to conclude that $a \mid p^n$. Therefore, $a \in Div(p^n) \cap \mathbb{N}^+$, so by induction, we know that $a \in \{p^k : 0 \leq k \leq n\}$, and hence $a \in \{p^k : 0 \leq k \leq n+1\}$

  Thus, in either case, we have shown that $a \in \{p^k : 0 \leq k \leq n+1\}$. Since $a \in Div(p^n) \cap \mathbb{N}^+$ was arbitrary, we conclude that $Div(p^n) \cap \mathbb{N}^+ \subseteq \{p^k : 0 \leq k \leq n+1\}$, which completes the inductive step.

The result follows by induction. □

For example, since 3 is prime and $243 = 3^4$, we immediately conclude that

$$\begin{aligned} Div(243) \cap \mathbb{N}^+ &= \{3^k : 0 \leq k \leq 4\} \\ &= \{3^0, 3^1, 3^2, 3^3, 3^4\} \\ &= \{1, 3, 9, 27, 243\}. \end{aligned}$$

Although this result allows us to determine the set of divisors of a power of a prime, it does not allow us to handle numbers like $36 = 2^2 \cdot 3^2$. We can determine the positive divisors of each of 2^2 and 3^2, but it's not immediately clear how to "combine" them. More generally, suppose that we have two numbers $a_1, a_2 \in \mathbb{N}^+$. Assume that the we know the set of positive divisors of each of a_1 and a_2 individually. Now if $b_1 \mid a_1$ and $b_2 \mid a_2$, then it turns out that we must have $b_1b_2 \mid a_1a_2$. Moreover, *every* positive divisor of a_1a_2 arises in this way.

Proposition 3.3.2. *Let $a_1, a_2 \in \mathbb{N}^+$.*

1. *If $b_1, b_2 \in \mathbb{N}^+$ satisfy $b_1 \mid a_1$ and $b_2 \mid a_2$, then $b_1 b_2 \mid a_1 a_2$.*
2. *For all $m \in \mathbb{N}^+$ with $m \mid a_1 a_2$ there exists $b_1, b_2 \in \mathbb{N}^+$ such that $b_1 \mid a_1$, $b_2 \mid a_2$, and $m = b_1 b_2$.*

Proof.

1. Let $b_1, b_2 \in \mathbb{N}^+$ be arbitrary with $b_1 \mid a_1$ with $b_2 \mid a_2$. Since $b_1 \mid a_1$, we can fix $c_1 \in \mathbb{Z}$ with $a_1 = b_1 c_1$. Since $b_2 \mid a_2$, we can fix $c_2 \in \mathbb{Z}$ with $a_2 = b_2 c_2$. We then have
$$\begin{aligned} a_1 a_2 &= b_1 c_1 \cdot b_2 c_2 \\ &= c_1 c_2 \cdot b_1 b_2. \end{aligned}$$
Since $c_1 c_2 \in \mathbb{Z}$, it follows that $b_1 b_2 \mid a_1 a_2$.

2. Let $m \in \mathbb{N}^+$ be arbitrary with $m \mid a_1 a_2$. Since $m \mid a_1 a_2$, we may fix $n \in \mathbb{Z}$ with $mn = a_1 a_2$. Let $b_1 = \gcd(m, a_1)$ and notice that $b_1 > 0$ (since $a_1 > 0$), so $b_1 \in \mathbb{N}^+$. Since b_1 is a common divisor of m and a_1, we can fix $b_2, k \in \mathbb{Z}$ with $m = b_1 b_2$ and $a_1 = b_1 k$. By Problem 5 on Homework 4, we know that $\gcd(b_2, k) = 1$. Now plugging these expressions for m and a_1 into $mn = a_1 a_2$, we see that
$$b_1 b_2 n = b_1 k a_2,$$
so dividing both sides by $b_1 > 0$, it follows that
$$b_2 n = k a_2.$$
Thus, $b_2 \mid k a_2$, so as $\gcd(b_2, k) = 1$, we can use Proposition 3.2.5 to conclude that $b_2 \mid a_2$.

□

Thus, we do obtain all of the (positive) divisors of $a_1 a_2$ by multiplying together the (positive) divisors of a_1 and a_2. However, one other natural question arises. Are the resulting divisors unique (i.e. is it impossible to obtain the same divisor in two separate ways using this process)? It turns out that uniqueness can fail. For example, if $a_1 = 6$ and $a_2 = 9$, then 18 is a divisor of $a_1 a_2 = 54$ that arises from both $18 = 6 \cdot 3$ and $18 = 2 \cdot 9$. However, if a_1 and a_2 are relatively prime, then we obtain uniqueness as well.

Proposition 3.3.3. *Let $a_1, a_2 \in \mathbb{N}^+$ be relatively prime integers. For all $m \in \mathbb{N}^+$ with $m \mid a_1 a_2$ there exists unique $b_1, b_2 \in \mathbb{N}^+$ such that $b_1 \mid a_1$, $b_2 \mid a_2$, and $m = b_1 b_2$.*

Proof. Let $m \in \mathbb{N}^+$ be arbitrary with $m \mid a_1 a_2$. The existence of b_1 and b_2 follow immediately from Proposition 3.3.2. We now prove uniqueness. Suppose that $b_1, b_2, c_1, c_2 \in \mathbb{N}^+$ satisfy $m = b_1 b_2 = c_1 c_2$, $b_1 \mid a_1$, $b_2 \mid a_2$, $c_1 \mid a_1$, and $c_2 \mid a_2$. We need to show that $b_1 = b_2$ and also that $c_1 = c_2$. Notice that any common divisor of b_1 and c_2 is a common divisor of a_1 and a_2 (by transitivity of divisibility), so must divide 1 because $\gcd(a_1, a_2) = 1$, and hence must be an element of $\{1, -1\}$. Thus, b_1 and c_2 are relatively prime. Similarly, b_2 and c_1 are relatively prime. Now we have
$$b_1 b_2 = m = c_1 c_2,$$
so $b_1 \mid c_1 c_2$. Since $\gcd(b_1, c_2) = 1$ from above, we can use Proposition 3.2.5 to conclude that $b_1 \mid c_1$. Similarly, we have $c_1 \mid b_1 b_2$, so as $\gcd(c_1, b_2) = 1$ from above, we conclude that $c_1 \mid b_1$. Since $b_1 \mid c_1$ and $c_1 \mid b_1$, we can Corollary 1.5.5 to deduce that $b_1 = \pm c_1$. Now $b_1, c_1 \in \mathbb{N}^+$, so we must have $b_1 = c_1$. Canceling this common term in the above displayed formula, we see that $b_2 = c_2$. This gives uniqueness. □

As an example, consider $36 = 2^2 3^2$. Using Proposition 3.3.1, we know that
$$Div(2^2) \cap \mathbb{N}^+ = \{2^0, 2^1, 2^2\} = \{1, 2, 4\}$$
and
$$Div(3^2) \cap \mathbb{N}^+ = \{3^0, 3^1, 3^2\} = \{1, 3, 9\}.$$

Since 2 and 3 are distinct primes, we can apply Corollary 3.2.11 to conclude that 2^2 and 3^2 are relatively prime. Now Proposition 3.3.2 says that we can obtain every (positive) divisor of $36 = 2^2 \cdot 3^2$ by multiplying together one (positive) divisor of 2^2 and one (positive) divisor of 3^2, and that furthermore, every way of doing this results in a different divisor of 36 by Proposition 3.3.3. Therefore,

$$\begin{aligned} Div(36) \cap \mathbb{N}^+ &= \{1 \cdot 1, 2 \cdot 1, 4 \cdot 1, 1 \cdot 3, 2 \cdot 3, 4 \cdot 3, 1 \cdot 9, 2 \cdot 9, 4 \cdot 9\} \\ &= \{1, 2, 4, 3, 6, 12, 9, 18, 36\} \\ &= \{1, 2, 3, 4, 6, 9, 12, 18, 36\}. \end{aligned}$$

We can also use these results to count the cardinality of the set of positive divisors of a given number, without having to enumerate the divisors. We first introduce a definition.

Definition 3.3.4. *Define a function* $d \colon \mathbb{N}^+ \to \mathbb{N}$ *by letting* $d(a)$ *be the number of positive divisors of* a*, i.e.* $d(a) = |Div(a) \cap \mathbb{N}^+|$.

Corollary 3.3.5. *For all primes* $p \in \mathbb{N}$ *and all* $n \in \mathbb{N}^+$*, we have* $d(p^n) = n + 1$.

Proof. Immediate from Proposition 3.3.1, together with the fact that $p^k \neq p^\ell$ whenever $0 \leq k < \ell \leq n$. $\square$

Corollary 3.3.6. *If* $a_1, a_2 \in \mathbb{N}^+$ *are relatively prime, then* $d(a_1 a_2) = d(a_1) \cdot d(a_2)$.

Proof. Let $m = d(a_1)$ and let $n = d(a_2)$. List the distinct elements of $Div(a_1) \cap \mathbb{N}^+$ as $b_1, b_2, \dots, b_m$, and list the distinct elements of $Div(a_2) \cap \mathbb{N}^+$ as $c_1, c_2, \dots, c_n$. By Proposition 3.3.2, we have that

$$Div(a_1 a_2) \cap \mathbb{N}^+ = \{b_i c_j : 1 \leq i \leq m \text{ and } 1 \leq j \leq n\},$$

Furthermore, since a_1 and a_2 are relatively prime, the elements $b_i c_j$ are distinct (i.e. if $b_i c_j = b_k c_\ell$, then $i = j$ and $k = \ell$) by Proposition 3.3.3. Therefore, we have $d(a_1 a_2) = mn$. $\square$

For example, we can now quickly compute that

$$\begin{aligned} d(36) &= d(2^2 \cdot 3^2) \\ &= d(2^2) \cdot d(3^2) \\ &= (2 + 1) \cdot (2 + 1) \\ &= 3 \cdot 3 \\ &= 9. \end{aligned}$$

More generally, we can use these results together with repeated applications of Corollary 3.2.11 to compute $d(a)$ whenever we have written a as a product of powers of distinct primes. For example,

$$\begin{aligned} d(720) &= d(2^4 \cdot 3^2 \cdot 5^1) \\ &= d(2^4) \cdot d(3^2 \cdot 5^1) \\ &= d(2^4) \cdot d(3^2) \cdot d(5^1) \\ &= (4 + 1) \cdot (2 + 1) \cdot (1 + 1) \\ &= 5 \cdot 3 \cdot 2 \\ &= 30. \end{aligned}$$

3.4 The Fundamental Theorem of Arithmetic

At the end of the previous section, we determined a way to compute $d(a)$, provided that we can write a as a product of powers of distinct primes. Can we always accomplish such a task? If so, is there always a unique such product? We begin by answering the first question. Note that when we say "product of primes", we are allowing the degenerate possibility of a 1-term product. That is, we still say that 2 is a product of primes, simply because 2 itself is prime.

Proposition 3.4.1. *Every $n \in \mathbb{N}$ with $n > 1$ can be written as a product of primes.*

Proof. We prove the result by strong induction on $\mathbb{N}$. If $n = 2$, we are done because 2 itself is prime. Suppose that $n > 2$ and we have proven the result for all k with $1 < k < n$. If n is prime, we are done. Suppose that n is not prime and fix a divisor $c \mid n$ with $1 < c < n$. Fix $d \in \mathbb{N}$ with $cd = n$. We then have that $1 < d < n$, so by induction, both c and d are products of primes, say $c = p_1p_2 \cdots p_k$ and $d = q_1q_2 \cdots q_\ell$ with each p_i and q_j prime. We then have

$$n = cd = p_1p_2 \cdots p_kq_1q_2 \cdots q_\ell,$$

so n is a product of primes. The result follows by induction. □

Corollary 3.4.2. *Every $a \in \mathbb{Z}$ with $a \notin \{-1, 0, 1\}$ can be written as either a product of primes, or -1 times a product of primes.*

We now have all the tools necessary to prove the uniqueness of prime factorizations.

Theorem 3.4.3 (Fundamental Theorem of Arithmetic)**.** *Every natural number greater than 1 factors uniquely (up to order) into a product of primes. In other words, if $n \geq 2$ and*

$$p_1p_2 \cdots p_k = n = q_1q_2 \cdots q_\ell$$

with $p_1 \leq p_2 \leq \cdots \leq p_k$ and $q_1 \leq q_2 \leq \cdots \leq q_\ell$ all primes, then $k = \ell$ and $p_i = q_i$ for $1 \leq i \leq k$.

Proof. Existence follows from Proposition 3.4.1. We prove uniqueness by (strong) induction on n. Let $n \in \mathbb{N}$ with $n \geq 2$, and assume that every $m \in \mathbb{N}$ with $2 \leq m < n$ factors uniquely into a product of primes. We prove that n factors uniquely into primes. Let $p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_\ell \in \mathbb{N}$ be primes with

$$p_1p_2 \cdots p_k = n = q_1q_2 \cdots q_\ell,$$

and where $p_1 \leq p_2 \leq \cdots \leq p_k$ and $q_1 \leq q_2 \leq \cdots \leq q_\ell$. We need to show that $k = \ell$ and that $p_i = q_i$ for all i. We have two cases:

- *Case 1:* Suppose that n is prime. Notice that $p_i \mid n$ for all i and $q_j \mid n$ for all j. Since the only positive divisors of n are 1 and n, and 1 is not prime, we conclude that $p_i = n$ for all i and $q_j = n$ for all j. If $k \geq 2$, then $p_1p_2 \cdots p_k = n^k > n$, a contradiction, so we must have $k = 1$. Similarly we must have $\ell = 1$.

- *Case 2:* Suppose now that n is composite. We then must have $k \geq 2$ and $\ell \geq 2$. Now $p_1 \mid q_1q_2 \cdots q_\ell$, so by Corollary 3.2.7, we can fix a j such that $p_1 \mid q_j$. Since q_j is prime and $p_1 \neq 1$, we must have $p_1 = q_j$. Similarly, we must have $q_1 = p_i$ for some i. We then have

$$p_1 = q_j \geq q_1 = p_i \geq p_1,$$

hence all inequalities must be equalities, and we conclude that $p_1 = q_1$. Canceling, we conclude that

$$p_2 \cdots p_k = q_2 \cdots q_\ell,$$

and this common value is some natural number m with $2 \leq m < n$. By induction, it follows that $k = \ell$ and $p_i = q_i$ for all i with $2 \leq i \leq k$.

□

Given a natural number $n \in \mathbb{N}$ with $n \geq 2$, when we write its prime factorization, we typically group together like primes and write

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k},$$

where the p_i are distinct primes. We often allow the insertion of "extra" primes in the factorization of n by permitting some α_i to equal to 0. This convention is particularly useful when comparing prime factorization of two numbers so that we can assume that both factorizations have the same primes occurring. It also allows us to write 1 in such a form by choosing all α_i to equal 0. Here is one example.

Proposition 3.4.4. *Suppose that $n, d \in \mathbb{N}^+$. Write the prime factorizations of n and d as*

$$\begin{aligned} n &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} \\ d &= p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}, \end{aligned}$$

where the p_i are distinct primes and possibly some α_i and β_j are 0. We then have that $d \mid n$ if and only if $0 \leq \beta_i \leq \alpha_i$ for all i.

Proof. Suppose first that $0 \leq \beta_i \leq \alpha_i$ for all i. We then have that $\alpha_i - \beta_i \geq 0$ for all i, so we may let

$$c = p_1^{\alpha_1-\beta_1} p_2^{\alpha_2-\beta_2} \cdots p_k^{\alpha_k-\beta_k} \in \mathbb{N}.$$

Notice that

$$\begin{aligned} dc &= p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k} \cdot p_1^{\alpha_1-\beta_1} p_2^{\alpha_2-\beta_2} \cdots p_k^{\alpha_k-\beta_k} \\ &= (p_1^{\beta_1} p_1^{\alpha_1-\beta_1})(p_2^{\beta_2} p_2^{\alpha_2-\beta_2}) \cdots (p_n^{\beta_n} p_n^{\alpha_1-\beta_n}) \\ &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n} \\ &= n, \end{aligned}$$

hence $d \mid n$.

Conversely, suppose that $d \mid n$ and fix $c \in \mathbb{Z}$ with $dc = n$. Notice that $c > 0$ because $d, n > 0$. Now we have $dc = n$, so $c \mid n$. If q is prime and $q \mid c$, then $q \mid n$ by transitivity of divisibility (Proposition 1.5.2), so $q \mid p_i$ for some i by Corollary 3.2.7, and hence $q = p_i$ for some i because each p_i is prime. Thus, we can write the prime factorization of c as

$$c = p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k},$$

where again we may have some γ_i equal to 0. We then have

$$\begin{aligned} n &= dc \\ &= (p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k})(p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}) \\ &= (p_1^{\beta_1} p_1^{\gamma_1})(p_2^{\beta_2} p_2^{\gamma_2}) \cdots (p_k^{\beta_k} p_k^{\gamma_k}) \\ &= p_1^{\beta_1+\gamma_1} p_2^{\beta_2+\gamma_2} \cdots p_k^{\beta_k+\gamma_k}. \end{aligned}$$

By the Fundamental Theorem of Arithmetic, we have $\beta_i + \gamma_i = \alpha_i$ for all i. Since $\beta_i, \gamma_i, \alpha_i \geq 0$ for all i, we conclude that $\beta_i \leq \alpha_i$ for all i. □

Corollary 3.4.5. *Let $a, b \in \mathbb{N}^+$ with and write*

$$\begin{aligned} a &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} \\ b &= p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}, \end{aligned}$$

where the p_i are distinct primes. We then have

$$\gcd(a, b) = p_1^{\min\{\alpha_1,\beta_1\}} p_2^{\min\{\alpha_2,\beta_2\}} \cdots p_k^{\min\{\alpha_k,\beta_k\}}.$$

Proof. Let $m = p_1^{\min\{\alpha_1,\beta_1\}} p_2^{\min\{\alpha_2,\beta_2\}} \cdots p_k^{\min\{\alpha_k,\beta_k\}}$ and notice that $m \geq 1$ trivially. Since $\min\{\alpha_i, \beta_i\} \leq \alpha_i$ for all i, it follows from Proposition 3.4.4 that $m \mid a$. Similarly, since $\min\{\alpha_i, \beta_i\} \leq \beta_i$ for all i, it follows that $m \mid b$. Therefore, m is a common divisor of a and b.

Now let d be an arbitrary common divisor of a and b. By Proposition 3.4.4, we can write $d = p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}$ with $\gamma_i \leq \alpha_i$ and $\gamma_i \leq \beta_i$ for all i. Since $\gamma_i \leq \alpha_i$ and $\gamma_i \leq \beta_i$ for all i, it follows that $\gamma_i \leq \min\{\alpha_i, \beta_i\}$ for all i. Therefore, we have both $d \mid m$ by Proposition 3.4.4.

Putting these facts together, we conclude that $m = \gcd(a, b)$. □

We can also obtain the formula for $d(n)$ that we derived in the previous section by appealing to these results.

Corollary 3.4.6. *Suppose that $n > 1$ and $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ where the p_i are distinct primes. We then have*

$$d(n) = \prod_{i=1}^{k} (\alpha_i + 1).$$

Proof. Using Proposition 3.4.4, we know that a given $d \in \mathbb{N}^+$ is a divisor of n if and only if it can be written as

$$d = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$

where $0 \leq \beta_i \leq \alpha_i$ for all i. Notice that we have $\alpha_i + 1$ many choices for each β_i. Furthermore, different choices of β_i give rise to different values of d by the Fundamental Theorem of Arithmetic. □

Theorem 3.4.7. *Let $m, n \in \mathbb{N}$ with $m, n \geq 2$. If the unique prime factorization of m does not have the property that every prime exponent is divisible by n, then $\sqrt[n]{m}$ is irrational.*

Proof. We prove the contrapositive. Suppose that $\sqrt[n]{m}$ is rational and fix $a, b \in \mathbb{N}^+$ with $\sqrt[n]{m} = \frac{a}{b}$ (we may assume that $a, b > 0$ because $\sqrt[n]{m} > 0$). We then have

$$\frac{a^n}{b^n} = \left(\frac{a}{b}\right)^n = m,$$

hence

$$a^n = b^n m.$$

Write a, b, m in their unique prime factorizations as

$$\begin{aligned} a &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} \\ b &= p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k} \\ m &= p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}, \end{aligned}$$

where the p_i are distinct (and possibly some $\alpha_i, \beta_i, \gamma_i$ are equal to 0). Since $a^n = b^n m$, we have

$$p_1^{n\alpha_1} p_2^{n\alpha_2} \cdots p_k^{n\alpha_k} = p_1^{n\beta_1+\gamma_1} p_2^{n\beta_2+\gamma_2} \cdots p_k^{n\beta_k+\gamma_k}.$$

By the Fundamental Theorem of Arithmetic, we conclude that $n\alpha_i = n\beta_i + \gamma_i$ for all i. Therefore, for each i, we have $\gamma_i = n\alpha_i - n\beta_i = n(\alpha_i - \beta_i)$, and so $n \mid \gamma_i$ for each i. □

Chapter 4

Injections, Surjections, and Bijections

4.1 Definitions and Examples

Recall that the defining property of a function $f\colon A \to B$ is that every input element from A produces a unique output element from B. However, this does not work in reverse. Given $b \in B$, it may be the case that b is the output of zero, one, or many elements from A. We give special names to the types of functions where we have limitations for how often elements $b \in B$ actually occur as an output.

Definition 4.1.1. *Let $f\colon A \to B$ be a function.*

- *We say that f is* injective *(or* one-to-one*) if whenever $a_1, a_2 \in A$ satisfy $f(a_1) = f(a_2)$, we have $a_1 = a_2$.*
- *We say that f is* surjective *(or* onto*) if for all $b \in B$, there exists $a \in A$ such that $f(a) = b$.*
- *We say that f is* bijective *if f is both injective and surjective.*

Let's take a moment to unpack these definitions. First, saying that function $f\colon A \to B$ is surjective is simply saying that every $b \in B$ is hit at least once by an element $a \in A$. We can rephrase this using Definition 1.4.3 by saying that $f\colon A \to B$ is surjective exactly when $\text{range}(f) = B$.

The definition of injective is slightly more mysterious at first. Intuitively, a function $f\colon A \to B$ is injective if every $b \in B$ is hit by at most one $a \in A$. Now saying this precisely takes a little bit of thought. After all, how can we say "there exists at most one" because our "there exists" quantifier is used to mean that there is at least one! The idea is to turn this around and not directly talk about $b \in B$ at all. Instead, we want to say that we never have a situation where we have two distinct elements $a_1, a_2 \in A$ that go to the same place under f. Thus, we want to say

"**Not** (There exists $a_1, a_2 \in A$ with $a_1 \neq a_2$ and $f(a_1) = f(a_2)$)".

We can rewrite this statement as

"For all $a_1, a_2 \in A$, we have **Not**($a_1 \neq a_2$ and $f(a_1) = f(a_2)$)",

which is equivalent to

"For all $a_1, a_2 \in A$, we have either $a_1 = a_2$ or $f(a_1) \neq f(a_2)$"

(notice that the negation of the "and" statement turned into an "or" statement). Finally, we can rewrite this as the following "if...then..." statement:

"For all $a_1, a_2 \in A$, if $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$".

Looking at our statement here, it captures what we want to express perfectly because it says that distinct inputs always go to distinct outputs, which exactly says no element of B is hit by 2 or more elements, and hence that every element of B is hit by at most 1 element. Thus, we could indeed take this as our definition of injective. The problem is that this definition is difficult to use in practice. To see why, think about how we would argue that a given function $f\colon A \to B$ is injective. It appears that we would want to take arbitrary $a_1, a_2 \in A$ with $a_1 \neq a_2$, and argue that under this assumption we must have that $f(a_1) \neq f(a_2)$. Now the problem with this is that is very difficult to work with an expression involving $\neq$ in ways that preserve truth. For example, we have that $-1 \neq 1$, but $(-1)^2 = 1^2$, so we can not square both sides and preserve non-equality. To get around this problem, we instead take the contrapositive of the statement in question, which turns into our formal definition of injective:

"For all $a_1, a_2 \in A$, if $f(a_1) = f(a_2)$, then $a_1 = a_2$".

Notice that in our definition above, we simply replace the "for all... if... then..." construct with a "whenever...we have..." for clarity, but these are saying precisely the same thing, i.e. that whenever we have two elements of A that happen to be sent to the same element of B, then in fact those two elements of A must be the same. Although our official definition is slightly harder to wrap one's mind around, it is *much* easier to work with in practice. To prove that a given $f\colon A \to B$ is injective, we take arbitrary $a_1, a_2 \in A$ with $f(a_1) = f(a_2)$, and use this equality to derive the conclusion that $a_1 = a_2$.

To recap the colloquial ways to understand these concepts, a function $f\colon A \to B$ is injective if every $b \in B$ is hit by at most one $a \in A$, and is surjective if every $b \in B$ is hit by at least one $a \in A$. It follows that a function $f\colon A \to B$ is bijective if every $b \in B$ is hit by exactly one $a \in A$. These ways of thinking about injective and surjective are great, but we need to be careful when proving that a function is injective or surjective. Given a function $f\colon A \to B$, here is the general process for proving that it has one or both of these properties:

- In order to prove that f is injective, you should start by taking arbitrary $a_1, a_2 \in A$ that satisfy $f(a_1) = f(a_2)$, and then work forward to derive that $a_1 = a_2$. In this way, you show that whenever two elements of A happen to go to the same output, then they must have been the same element all along.
- In order to prove that f is surjective, you should start by taking an arbitrary $b \in B$, and then show how to build an $a \in A$ with $f(a) = b$. In other words, you want to take an arbitrary $b \in B$ and fill in the blank in $f(\underline{\quad\quad}) = b$ with an element of A.

Here is an example.

Proposition 4.1.2. *The function $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = 2x$ is bijective.*

Proof. We need to show that f is both injective and surjective.

- We first show that f is injective. Let $x_1, x_2 \in \mathbb{R}$ be arbitrary with $f(x_1) = f(x_2)$. We then have that $2x_1 = 2x_2$. Dividing both sides by 2, we conclude that $x_1 = x_2$. Since $x_1, x_2 \in \mathbb{R}$ were arbitrary with $f(x_1) = f(x_2)$, it follows that f is injective.
- We next show that f is surjective. Let $y \in \mathbb{R}$ be arbitrary. Notice that $\frac{y}{2} \in \mathbb{R}$ and that

$$f\left(\frac{y}{2}\right) = 2 \cdot \frac{y}{2} = y.$$

  Thus, we have shown the existence of an $x \in \mathbb{R}$ with $f(x) = y$. Since $y \in \mathbb{R}$ was arbitrary, it follows that f is surjective

Since f is both injective and surjective, it follows that f is bijective. $\square$

Notice that if we define $g\colon \mathbb{Z} \to \mathbb{Z}$ by letting $g(x) = 2x$, then g is injective by the same proof, but g is not surjective because there does not exist $m \in \mathbb{Z}$ with $f(m) = 1$ (since this would imply that $2m = 1$, so $2 \mid 1$, a contradiction). Thus, changing the domain or codomain of a function can change the properties of that function.

Proposition 4.1.3. *The function $d\colon \mathbb{N}^+ \to \mathbb{N}^+$, where $d(n)$ is the number of positive divisors of n, is surjective but not injective.*

Proof. Both 3 and 5 are prime, so $d(3) = 2 = d(5)$. Since $3 \neq 5$, it follows that d is not injective. To show that d is surjective, first notice that $d(1) = 1$, so $1 \in \text{range}(d)$. Now given an arbitrary $m \in \mathbb{N}^+$ with $m \geq 2$, we have that $m - 1 \in \mathbb{N}^+$, so

$$d(2^{m-1}) = (m-1) + 1 = m$$

by Corollary 3.3.5 (since 2 is prime). Therefore, $\text{range}(d) = \mathbb{N}^+$, and hence d is surjective. $\square$

Here are several more examples, where $|\sigma|$ is defined to be the length of the sequence σ:

- $f\colon \{0,1\}^* \to \mathbb{N}$ defined by $f(\sigma) = |\sigma|$ is surjective but not injective.
- $f\colon \{0,1\}^* \to \mathbb{Z}$ defined by $f(\sigma) = |\sigma|$ is neither surjective nor injective.
- $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \sin x$ is neither injective nor surjective.

Proposition 4.1.4. *Let A, B, C be sets and let $f\colon A \to B$ and $g\colon B \to C$ be functions*

1. *If f and g are both injective, then $g \circ f$ is injective.*
2. *If f and g are both surjective, then $g \circ f$ is surjective.*
3. *If f and g are both bijective, then $g \circ f$ is bijective.*
4. *If $g \circ f$ is injective, then f is injective.*
5. *If $g \circ f$ is surjective, then g is surjective.*

Proof.

1. Suppose that f and g are both injective. Let $a_1, a_2 \in A$ be arbitrary with $(g \circ f)(a_1) = (g \circ f)(a_2)$. By definition of composition, we then have $g(f(a_1)) = g(f(a_2))$. Using the fact that g is injective, we conclude that $f(a_1) = f(a_2)$. Now we use the fact that f is injective to conclude that $a_1 = a_2$. Therefore, $g \circ f$ is injective.

2. Suppose that f and g are both surjective. Let $c \in C$ be arbitrary. Since g is surjective, we can fix $b \in B$ with $g(b) = c$. Since f is surjective, we can fix $a \in A$ with $f(a) = b$. We then have

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

   Since $c \in C$ was arbitrary, we conclude that $g \circ f$ is surjective.

3. This follows from combining 1 and 2.

4. Suppose that $g \circ f$ is injective. Let $a_1, a_2 \in A$ be arbitrary with $f(a_1) = f(a_2)$. Applying g to both sides, we then have that $g(f(a_1)) = g(f(a_2))$, so $(g \circ f)(a_1) = (g \circ f)(a_2)$. Using the fact that $g \circ f$ is injective, it follows that $a_1 = a_2$. Therefore, f is injective.

5. Suppose that $g \circ f$ is surjective. Let $c \in C$ be arbitrary. Since $g \circ f$ is surjective, we can fix $a \in A$ with $(g \circ f)(a) = c$. By definition of composition, we then have $g(f(a)) = c$. Since $f(a) \in B$, we have succeeded in finding a b with $g(b) = c$ (namely $b = f(a)$). Since $c \in C$ was arbitrary, we conclude that g is surjective.

□

Definition 4.1.5. *Let A be a set. The function $id_A \colon A \to A$ defined by $id_A(a) = a$ for all $a \in A$ is called the* identity function *on A.*

We call this function the identity function because it leaves other functions alone when we compose with it. However, we have to be careful that we compose with the identity function on the correct set and the correct side.

Proposition 4.1.6. *For any function $f \colon A \to B$, we have $f \circ id_A = f$ and $id_B \circ f = f$.*

Proof. Let $f \colon A \to B$ be an arbitrary function.

- We first show that $f \circ id_A = f$. Let $a \in A$ be arbitrary. We have

$$\begin{aligned} (f \circ id_A)(a) &= f(id_A(a)) && \text{(by definition of composition)} \\ &= f(a) \end{aligned}$$

  Since $a \in A$ was arbitrary, it follows that $f \circ id_A = f$.

- We now show that $id_B \circ f = f$. Let $a \in A$ be arbitrary. We have

$$\begin{aligned} (id_B \circ f)(a) &= id_B(f(a)) && \text{(by definition of composition)} \\ &= f(a) && \text{(because } f(a) \in B\text{)} \end{aligned}$$

  Since $a \in A$ was arbitrary, it follows that $id_B \circ f = f$.

□

Suppose that $f \colon A \to B$ is a function. We want to think about what an *inverse function* of f would even mean. Naturally, an inverse should "undo" what f does. Since $f \colon A \to B$, we should think about functions $g \colon B \to A$. In order for g to undo f, it seems that we would want $g(f(a)) = a$ for all $a \in A$. Similarly, we might want this to work in the other direction so that $f(g(b)) = b$ for all $b \in B$. Notice that we can write the statement "$g(f(a)) = a$ for all $a \in A$" in a more elegant fashion by saying that $g \circ f = id_A$, where $id_A \colon A \to A$ is the identity function on A (i.e. $id_A(a) = a$ for all $a \in A$). Similarly, we can write "$f(g(b)) = b$ for all $b \in B$" as $f \circ g = id_B$. We codify these ideas in a definition.

Definition 4.1.7. *Let $f \colon A \to B$ be a function.*

- *A* left inverse *for f is a function $g \colon B \to A$ such that $g \circ f = id_A$.*
- *A* right inverse *for f is a function $g \colon B \to A$ such that $f \circ g = id_B$.*
- *An* inverse *for f is a function $g \colon B \to A$ such that both $g \circ f = id_A$ and $f \circ g = id_B$.*

Let's consider an example. Let $A = \{1, 2, 3\}$ and $B = \{5, 6, 7, 8\}$, and consider the function $f \colon A \to B$ defined as follows:

$$f(1) = 7 \qquad f(2) = 5 \qquad f(3) = 8.$$

As a set, we can write $f = \{(1, 7), (2, 5), (3, 8)\}$. Notice that f is injective but not surjective because $6 \notin \text{range}(f)$. Does f have a left inverse or a right inverse? A guess would be to define $g \colon B \to A$ as follows:

$$g(5) = 2 \qquad g(6) = \,? \qquad g(7) = 1 \qquad g(8) = 3.$$

Notice that it is unclear how to define $g(6)$ because 6 is not hit by f. Suppose that we pick a random $c \in A$ and let $g(6) = c$. We have the following:

$$\begin{aligned} g(f(1)) &= g(7) = 1 \\ g(f(2)) &= g(5) = 2 \\ g(f(3)) &= g(8) = 3. \end{aligned}$$

Thus, we have $g(f(a)) = a$ for all $a \in A$, so $g \circ f = id_A$ regardless of how we choose to define $g(6)$. We have shown that f has a left inverse (in fact, we have shown that f has at least 3 left inverses because we have 3 choices for $g(6)$). Notice that the value of $g(6)$ never came up in the above calculation because $6 \notin \text{range}(f)$. What happens when we look at $f \circ g$? Ignoring 6 for the moment, we have the following:

$$\begin{aligned} f(g(5)) &= f(2) = 5 \\ f(g(7)) &= f(1) = 7 \\ f(g(8)) &= f(3) = 8. \end{aligned}$$

Thus, we have $f(g(b)) = b$ for all $b \in \{5, 7, 8\}$. However, notice that no matter how we choose $c \in A$ to define $g(6) = c$, it doesn't work. For example, if let $g(6) = 1$, then $f(g(6)) = f(1) = 7$. You can work through the other two possibilities directly, but notice that no matter how we choose c, we will have $f(g(c)) \in \text{range}(f)$, and hence $f(g(c)) \neq 6$ because $6 \notin \text{range}(f)$. In other words, it appears that f does not have a right inverse. Furthermore, this problem seems to arise whenever we have a function that is not surjective.

Let's see an example where f is not injective. Let $A = \{1, 2, 3\}$ and $B = \{5, 6\}$, and consider the function $f \colon A \to B$ defined as follows:

$$f(1) = 5 \qquad f(2) = 6 \qquad f(3) = 5.$$

As a set, we can write $f = \{(1, 5), (2, 6), (3, 5)\}$. Notice that f is surjective but not injective (since $f(1) = f(3)$ but $1 \neq 3$). Does f have a left inverse or a right inverse? The guess would be to define $g \colon B \to A$ by letting $g(6) = 2$, but it's unclear how to define $g(5)$. Should we let $g(5) = 1$ or should we let $g(5) = 3$? Suppose that we choose $g \colon B \to A$ as follows:

$$g(5) = 1 \qquad g(6) = 2.$$

Let's first look at $f \circ g$. We have the following:

$$\begin{aligned} f(g(5)) &= f(1) = 5 \\ f(g(6)) &= f(2) = 6. \end{aligned}$$

We have shown that $f(g(b)) = b$ for all $b \in B$, so $f \circ g = id_B$. Now if we instead choose $g(5) = 3$, then we would have

$$\begin{aligned} f(g(5)) &= f(3) = 5 \\ f(g(6)) &= f(2) = 6, \end{aligned}$$

which also works. Thus, we have shown that f has a right inverse, and in fact it has at least 2 right inverses. What happens if we look at $g \circ f$ for these functions g? If we define $g(5) = 1$, then we have

$$g(f(3)) = g(5) = 1,$$

which does not work. Alternatively, if we define $g(5) = 3$, then we have

$$g(f(1)) = g(5) = 3,$$

which does not work either. It seems that no matter how we choose $g(5)$, we will obtain the wrong result on some input to $g \circ f$. In other words, it appears that f does not have a left inverse. Furthermore, this problem seems to arise whenever we have a function that is not injective.

Now if $f: A \to B$ is bijective, then it seems reasonable that if we define $g: B \to A$ by simply "flipping all of the arrows", then g will be an inverse for f (on both sides), and that this is the only possible way to define an inverse for f. We now prove all of these results in general, although feel free to skim the next couple of results for now (we won't need them for a little while).

Proposition 4.1.8. *Let $f: A \to B$ be a function.*

1. *f is injective if and only if there exists a left inverse for f.*
2. *f is surjective if and only if there exists a right inverse for f.*
3. *f is bijective if and only if there exists an inverse for f.*

Proof.

1. Suppose first that f has a left inverse, and fix a function $g: B \to A$ with $g \circ f = id_A$. Suppose that $a_1, a_2 \in A$ satisfy $f(a_1) = f(a_2)$. Applying the function g to both sides we see that $g(f(a_1)) = g(f(a_2))$, and hence $(g \circ f)(a_1) = (g \circ f)(a_2)$. We now have

$$\begin{aligned} a_1 &= id_A(a_1) \\ &= (g \circ f)(a_1) \\ &= (g \circ f)(a_2) \\ &= id_A(a_2) \\ &= a_2 \end{aligned}$$

so $a_1 = a_2$. It follows that f is injective.

Suppose conversely that f is injective. If $A = \emptyset$, then $f = \emptyset$, and we are done by letting $g = \emptyset$ (if the empty set as a function annoys you, just ignore this case). Let's assume then that $A \neq \emptyset$ and fix $a_0 \in A$. We now define $g: B \to A$. Given $b \in B$, we define $g(b)$ as follows:

- If $b \in \text{range}(f)$, then there exists a unique $a \in A$ with $f(a) = b$ (because f is injective), and we let $g(b) = a$ for this unique choice.
- If $b \notin \text{range}(f)$, then we let $g(b) = a_0$.

This completes the definition of $g: B \to A$. In terms of sets, g is obtained from f by flipping all of the pairs, and adding (b, a_0) for all $b \notin \text{range}(f)$. We need to check that $g \circ f = id_A$. Let $a \in A$ be arbitrary. We then have that $f(a) \in B$, and furthermore $f(a) \in \text{range}(f)$ trivially. Therefore, in the definition of g on the input $f(a)$, we defined $g(f(a)) = a$, so $(g \circ f)(a) = id_A(a)$. Since $a \in A$ was arbitrary, it follows that $g \circ f = id_A$. Therefore, f has a left inverse.

2. Suppose first that f has a right inverse, and fix a function $g: B \to A$ with $f \circ g = id_B$. Let $b \in B$ be arbitrary. We then have that

$$\begin{aligned} b &= id_B(b) \\ &= (f \circ g)(b) \\ &= f(g(b)) \end{aligned}$$

hence there exists $a \in A$ with $f(a) = b$, namely $a = g(b)$. Since $b \in B$ was arbitrary, it follows that f is surjective.

Suppose conversely that f is surjective. We define $g\colon B \to A$ as follows. For every $b \in B$, we know that there exists (possibly many) $a \in A$ with $f(a) = b$ because f is surjective. Given $b \in B$, we then define $g(b) = a$ for some (any) $a \in A$ for which $f(a) = b$. Now given any $b \in B$, notice that $g(b)$ satisfies $f(g(b)) = b$ by definition of g, so $(f \circ g)(b) = b = id_B(b)$. Since $b \in B$ was arbitrary, it follows that $f \circ g = id_B$.

3. The right to left direction is immediate from parts 1 and 2. For the left to right direction, we need only note that if f is a bijection, then the function g defined in the left to right direction in the proof of 1 equals the function g defined in the left to right direction in the proof of 2.

□

Proposition 4.1.9. *Let $f\colon A \to B$ be a function. If $g\colon B \to A$ is a left inverse of f and $h\colon B \to A$ is a right inverse of f, then $g = h$.*

Proof. By definition, we have that that $g \circ f = id_A$ and $f \circ h = id_B$. The key function to consider is the composition $(g \circ f) \circ h = g \circ (f \circ h)$ (notice that these are equal by Proposition 1.4.5). We have

$$\begin{aligned} g &= g \circ id_B \\ &= g \circ (f \circ h) \\ &= (g \circ f) \circ h && \text{(by Proposition 1.4.5)} \\ &= id_A \circ h \\ &= h. \end{aligned}$$

Therefore, we conclude that $g = h$. □

Corollary 4.1.10. *If $f\colon A \to B$ is a function, then there exists at most one function $g\colon B \to A$ that is an inverse of f.*

Proof. Suppose that $g\colon B \to A$ and $h\colon B \to A$ are both inverses of f. In particular, we then have that g is a left inverse of f and h is a right inverse of f. Therefore, $g = h$ by Proposition 4.1.9. □

Corollary 4.1.11. *If $f\colon A \to B$ is a bijective function, then there exists a unique inverse for f.*

Proof. Immediate from Proposition 4.1.8 and Corollary 4.1.10. □

Notation 4.1.12. *Suppose that $f\colon A \to B$ is bijective. We let $f^{-1}\colon B \to A$ be the the unique inverse for f. More concretely, f^{-1} is defined as follows. Given $b \in B$, we define $f^{-1}(b)$ to equal the unique $a \in A$ with $f(a) = b$.*

Notice that by definition, we have both $f^{-1} \circ f = id_A$ and $f \circ f^{-1} = id_B$. In other words, we have $f^{-1}(f(a)) = a$ for all $a \in A$, and $f(f^{-1}(b)) = b$ for all $b \in B$.

4.2 The Bijection Principle

Perhaps somewhat surprisingly, we can use functions to help us determine the cardinality of a set. The following fact connects up the concepts introduced in the previous section with cardinalities of the domain and codomain of functions.

Fact 4.2.1. *Let A and B be finite sets.*

- *There exists an injective function $f\colon A \to B$ if and only if $|A| \le |B|$.*

- *There exists a surjective function* $f \colon A \to B$ *if and only if* $|B| \leq |A|$.
- *There exists a bijective function* $f \colon A \to B$ *if and only if* $|A| = |B|$.

It is reasonably straightforward to provide intuitive arguments for each of these facts. Let's look at the first one. Suppose first that there exists an injective function $f \colon A \to B$. In this case, every element of B is hit by at most one element of A via f, so there must be at least as many elements in B as there are in A. For the converse, if $A = \{a_1, a_2, \dots, a_m\}$ and $B = \{b_1, b_2, \dots, b_n\}$ with $m \leq n$, then we can define an injective function $f \colon A \to B$ by letting $f(a_i) = b_i$ for all i. Although each of these arguments is convincing, the first one is not terribly precise. If desired, it is possible to prove the first one by induction on the cardinalities of A and B, but just as for the Sum Rule we will avoid being so formal. The other two results can be argued similarly.

The third fact listed above is very helpful when trying to determine the cardinality of a set, and is sometimes called the "Bijection Principle". Perhaps surprisingly, we've already used this type of argument informally on several occasions. On the first homework, we showed that if A is a set with $|A| = n$ and $D = \{(a,a) : a \in A\}$, then we had $|D| = n$. Intuitively, if $A = \{a_1, a_2, \dots, a_n\}$, then $D = \{(a_1,a_1), (a_2,a_2), \dots, (a_n,a_n)\}$, so $|D| = n$. Using the Bijection Principle, we can argue this more formally by exhibiting the following function: define $f \colon A \to A^2$ by letting $f(a) = (a,a)$. It is then straightforward to check that f is injective and $\text{range}(f) = D$, so f is a bijection, and hence $|D| = |A| = n$.

We were also implicitly using the bijection principle in the proof of Corollary 3.3.6, which said that $d(a_1 a_2) = d(a_1) \cdot d(a_2)$ whenever $a_1, a_2 \in \mathbb{N}^+$ were relatively prime. The idea behind the argument was that in this case, every (positive) divisor of $a_1 a_2$ can be decomposed uniquely as a product of a (positive) divisor of a_1 and a (positive) divisor of a_2. More formally, we can use the first part of Proposition 3.3.2 to define the function

$$f \colon (Div(a_1) \cap \mathbb{N}^+) \times (Div(a_2) \cap \mathbb{N}^+) \to Div(a_1 a_2) \cap \mathbb{N}^+$$

given by letting $f(d_1, d_2) = d_1 d_2$. Notice that f is surjective by the second part of Proposition 3.3.2, and f is injective by Proposition 3.3.3 (which is the only place where we use that a_1 and a_2 are relatively prime). Therefore, by the Bijection Principle, we know that

$$|(Div(a_1) \cap \mathbb{N}^+) \times (Div(a_2) \cap \mathbb{N}^+)| = |Div(a_1 a_2) \cap \mathbb{N}^+|.$$

Now the Product Rule tells us that

$$|(Div(a_1) \cap \mathbb{N}^+) \times (Div(a_2) \cap \mathbb{N}^+)| = |(Div(a_1) \cap \mathbb{N}^+)| \cdot |(Div(a_2) \cap \mathbb{N}^+)|,$$

so

$$|Div(a_1) \cap \mathbb{N}^+| \cdot |Div(a_2) \cap \mathbb{N}^+| = |Div(a_1 a_2) \cap \mathbb{N}^+|,$$

and hence $d(a_1) \cdot d(a_2) = d(a_1 a_2)$.

Finally, we can see the Bijection Principle at work in the proof of Corollary 3.4.6. Suppose that $n > 1$ and $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ where the p_i are distinct primes. Consider the set

$$S = \{0, 1, \dots, \alpha_1\} \times \{0, 1, \dots, \alpha_2\} \times \cdots \times \{0, 1, \dots, \alpha_k\}.$$

Since $|\{0, 1, \dots, \alpha_k\}| = \alpha_i + 1$ for all i, we can use the General Product Rule to conclude that

$$|S| = (\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_k + 1) = \prod_{i=1}^{k} (\alpha_i + 1).$$

Define $f \colon S \to \mathbb{N}^+$ by letting $f(\beta_1, \beta_2, \dots, \beta_k) = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$. The Fundamental Theorem of Arithmetic tells us that f is injective, and Proposition 3.4.4 allows us to conclude that $\text{range}(f) = Div(n) \cap \mathbb{N}^+$. Thus, if we restrict the codomain to view $f\colon S \to Div(n) \cap \mathbb{N}^+$, then f is bijective. Using the Bijection Principle, we conclude that $|S| = |Div(n) \cap \mathbb{N}^+|$, so

$$d(n) = |S| = \prod_{i=1}^{k} (\alpha_i + 1).$$

In general, if we have a set A and want to determine $|A|$, the idea is to create a bijection between A and another set B, where $|B|$ is easier to determine. Our first new example of this technique is the following fundamental result:

Proposition 4.2.2. *Given a finite set A with $|A| = n \in \mathbb{N}^+$, there exists a bijection $f\colon \{0,1\}^n \to \mathcal{P}(A)$.*

Before jumping into the proof, we first illustrate with a special case. Let $A = \{1, 2, 3\}$, so that $|A| = 3$. Notice that

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

Looking at the elements of $\mathcal{P}(\{1,2,3\})$, we see that to create a subset of $\{1,2,3\}$, we need to decide which elements to keep, and which elements to omit. In other words, when building a subset of $\{1,2,3\}$, we ask ourselves whether to include 1 in our subset, whether to include 2 in our subset, and whether to include 3 in our subset. Each of these is a binary choice, and can be coded by either a 0 or a 1. Given a 3-tuple of 0's and 1's, say $(1,0,1)$, we can think the first 1 as coding the information that we should include 1 in our set, the 0 as coding that we should omit 2, and the 1 as coding that we should include 3. Thus, we associate $(1,0,1)$ with the subset $\{1,3\}$ of $\{1,2,3\}$. In this way, we establish the following bijection between $\{0,1\}^3$ and $\mathcal{P}(\{1,2,3\})$:

$$\begin{aligned}
(0,0,0) &\mapsto \emptyset \\
(0,0,1) &\mapsto \{3\} \\
(0,1,0) &\mapsto \{2\} \\
(1,0,0) &\mapsto \{1\} \\
(0,1,1) &\mapsto \{2,3\} \\
(1,0,1) &\mapsto \{1,3\} \\
(1,1,0) &\mapsto \{1,2\} \\
(1,1,1) &\mapsto \{1,2,3\}.
\end{aligned}$$

We now write the general proof.

Proof. Let $A = \{a_1, a_2, \dots, a_n\}$ where the a_i are distinct. Define a function $f\colon \{0,1\}^n \to \mathcal{P}(A)$ by letting $f(b_1, b_2, \dots, b_n) = \{a_i : b_i = 1\}$. In other words, given a finite sequence $(b_1, b_2, \dots, b_n)$ of 0's and 1's, we send it to the subset of A obtained by including a_i precisely when the i^{th} element of the sequence is a 1. Notice that if $(b_1, b_2, \dots, b_n) \neq (c_1, c_2, \dots, c_n)$, then we can fix an i with $b_i \neq c_i$, and in this case we have $f(b_1, b_2, \dots, b_n) \neq f(c_1, c_2, \dots, c_n)$ because a_i is one of the sets but not the other. Furthermore, given any $S \subseteq A$, if we let $(b_1, b_2, \dots, b_n) \in \{0,1\}^n$ be defined by letting

$$b_i = \begin{cases} 1 & \text{if } a_i \in S \\ 0 & \text{if } a_i \notin S, \end{cases}$$

then $f(b_1, b_2, \dots, b_n) = S$, so f is surjective. Therefore, f is a bijection. □

Corollary 4.2.3. *If $|A| = n \in \mathbb{N}^+$, then $|\mathcal{P}(A)| = 2^n$.*

Proof. This is immediate from the Proposition 4.2.2, the Bijection Principle, and Corollary 1.2.11. □

Since this result is so fundamental, we give another proof that uses both induction and the Bijection Principle. As above, we first provide some intuition by considering $\mathcal{P}(\{1,2,3\})$. Notice that we can break up $\mathcal{P}(\{1,2,3\})$ into the union of two disjoint subsets, consisting of those elements that do not contain 3, and those that do contain 3:

$$\mathcal{P}(\{1,2,3\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\} \cup \{\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

Notice that first of these subsets is just $\mathcal{P}(\{1,2\})$, and the second can be created from the first by inserting 3 into each of the subsets. In other words, we can use induction to determine the cardinality of the first set, and then notice that there is a bijection between the subsets that do not contain 3, and those that do contain 3. Here is the general argument:

Proof 2 of Corollary 4.2.3. We prove the result by induction on $n \in \mathbb{N}^+$.

- *Base Case:* Suppose that $n = 1$. Let A be a set with $|A| = 1$, say $A = \{a\}$. We then have that $\mathcal{P}(A) = \{\emptyset, \{a\}\}$, so $|\mathcal{P}(A)| = 2 = 2^1$.

- *Induction Step:* Assume that the statement is true for some fixed $n \in \mathbb{N}^+$, i.e. assume that for some fixed $n \in \mathbb{N}^+$, we know that $|\mathcal{P}(A)| = 2^n$ for all sets A with $|A| = n$. Consider an arbitrary set A with $|A| = n+1$. Fix some (any) element $a_0 \in A$. Let $\mathcal{S} \subseteq \mathcal{P}(A)$ be the collection of subsets of A not having a_0 as an element, and let $\mathcal{T} \subseteq \mathcal{P}(A)$ be the collection of subsets of A having a_0 as an element. Notice then that $\mathcal{S}$ and $\mathcal{T}$ are disjoint sets with $\mathcal{P}(A) = \mathcal{S} \cup \mathcal{T}$, so by the Sum Rule we know that

  $$|\mathcal{P}(A)| = |\mathcal{S}| + |\mathcal{T}|.$$

  Now consider the function $f \colon \mathcal{S} \to \mathcal{T}$ defined by letting $f(B) = B \cup \{a_0\}$, i.e. given $B \in \mathcal{S}$, we have that B is a subset of A not having a_0 as an element, and we send it to the subset of A obtained by throwing a_0 in as a new element. Notice that f is a bijection, so $|\mathcal{S}| = |\mathcal{T}|$. Therefore, we have

  $$|\mathcal{P}(A)| = |\mathcal{S}| + |\mathcal{S}|.$$

  Finally, notice that $\mathcal{S} = \mathcal{P}(A \backslash \{a_0\})$, so since $|A \backslash \{a_0\}| = n$, we can use induction to conclude that $|A \backslash \{a_0\}| = 2^n$. Therefore,

  $$\begin{aligned} |\mathcal{P}(A)| &= 2^n + 2^n \\ &= 2 \cdot 2^n \\ &= 2^{n+1}. \end{aligned}$$

  Thus, the statement is true for $n+1$.

By induction, we conclude that if $|A| = n \in \mathbb{N}^+$, then $|\mathcal{P}(A)| = 2^n$. □

By the way, Corollary 4.2.3 is also true in the case $n = 0$. When $n = 0$, we have $A = \emptyset$ and $\mathcal{P}(\emptyset) = \{\emptyset\}$, so $|\mathcal{P}(\emptyset)| = 1 = 2^0$.

We can also use the Bijection Principle when we do not know the size of either set. For an illustrative example, consider the set $\mathcal{P}(\{1,2,3,4,5\})$. We know from Corollary 4.2.3 that $|\mathcal{P}(\{1,2,3,4,5\})| = 2^5 = 32$. What if we only wanted to consider the subsets of $\{1,2,3,4,5\}$ that have exactly 2 elements? If we let $\mathcal{S}$ be this subset of $\mathcal{P}(\{1,2,3,4,5\})$, then

$$\mathcal{S} = \{\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \{3,4\}\}.$$

Similarly, if we let $\mathcal{T}$ be the subsets of $\{1,2,3,4,5\}$ having exactly 3 elements, then

$$\mathcal{T} = \{\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{1,4,5\}, \{2,3,4\}, \{2,3,5\}, \{2,4,5\}, \{3,4,5\}\}.$$

In this case, you can directly check that $|\mathcal{S}| = 10 = |\mathcal{T}|$, but we can also argue that $|\mathcal{S}| = |\mathcal{T}|$ without performing an exhaustive count. The key idea is that the relative complement of a size 2 subset of $\{1,2,3,4,5\}$ is a size 3 subset of $\{1,2,3,4,5\}$. In other words, we have the following bijection between $\mathcal{S}$ and $\mathcal{T}$:

$$\begin{aligned}
\{1,2\} &\mapsto \{3,4,5\} \\
\{1,3\} &\mapsto \{2,4,5\} \\
\{1,4\} &\mapsto \{2,3,5\} \\
\{1,5\} &\mapsto \{2,3,4\} \\
\{2,3\} &\mapsto \{1,4,5\} \\
\{2,4\} &\mapsto \{1,3,5\} \\
\{2,5\} &\mapsto \{1,3,4\} \\
\{3,4\} &\mapsto \{1,2,5\} \\
\{3,5\} &\mapsto \{1,2,4\} \\
\{4,5\} &\mapsto \{1,2,3\}.
\end{aligned}$$

Generalizing this argument leads to following result.

Proposition 4.2.4. *Let A be a set with $|A| = n \in \mathbb{N}^+$ and let $k \in \mathbb{N}$ be such that $0 \le k \le n$. The number of subsets of A having cardinality k equals the number of subsets of A having cardinality $n-k$.*

Proof. Let $\mathcal{S}$ be the collection of all subsets of A having cardinality k, and let $\mathcal{T}$ be the collection of all subsets of A having cardinality $n-k$. Define $f\colon \mathcal{S} \to \mathcal{T}$ by letting $f(B) = A \backslash B$, i.e. given $B \subseteq A$ with $|B| = k$, send it to the complement of B in A (notice that if $|B| = k$, then $|A \backslash B| = n-k$ by the complement rule). Notice that f is a bijection (it is surjective because if $C \subseteq A$ is such that $|C| = n-k$, then $|A \backslash C| = k$ and $f(A \backslash C) = C$). Therefore, $|\mathcal{S}| = |\mathcal{T}|$. □

Thus, despite the fact that we do not (yet) have a formula for the number of subsets of a certain size, we do know that the number of subsets of size k must equal the number of subsets of size $n-k$.

4.3 The Pigeonhole Principle

Although we have thus far focused on bijections, we also know from the previous section that if A and B are finite sets, and $f\colon A \to B$ is an injective function, then $|A| \le |B|$. Taking the contrapositive of this fact, we obtain the following:

Corollary 4.3.1 (Pigeonhole Principle)**.** *If A and B are finite sets with $|A| > |B|$, and $f\colon A \to B$ is a function, then there exist $a_1, a_2 \in A$ with $a_1 \neq a_2$ such that $f(a_1) = f(a_2)$.*

Stated informally, the Pigeonhole Principle says that if $n > k$ and we place n balls into k boxes, then (at least) one box will contain at least 2 balls. For a very simple example, in any group of 13 people, there must exist (at least) 2 people in the group who were born in the same month. Here is a more interesting example:

Proposition 4.3.2. *Given $n+1$ integers, it is always possible to find two whose difference is divisible by n.*

Proof. Let A be a set of $n+1$ integers, so $A = \{a_0, a_1, \dots, a_n\}$. For each i, we can use division with remainder to fix $q_i, r_i \in \mathbb{Z}$ with

$$a_i = nq_i + r_i$$

and $0 \le r_i < n$. Notice then that $r_i \in \{0,1,2,\dots,n-1\}$ for all i. Define $f\colon A \to \{0,1,2,\dots,n-1\}$ by letting $f(a_i) = r_i$ for each i. Since $|A| = n+1$ and $|\{0,1,2,\dots,n-1\}| = n$, we know by the Pigeonhole

Principle that f is not injective. Thus, we can fix $i \neq j$ with $r_i = r_j$. We then have

$$\begin{aligned} a_i - a_j &= (nq_i + r_i) - (nq_j + r_j) \\ &= n(q_i - q_j) + (r_i - r_j) \\ &= n(q_i - q_j) && \text{(since } r_i - r_j = 0\text{)}, \end{aligned}$$

so $n \mid (a_i - a_j)$. □

Proposition 4.3.3. *For each $n \in \mathbb{N}^+$, let $a_n = 333\cdots 3$ where there are n many 3s. There exists an $n \leq 1492$ such that $1491 \mid a_n$.*

Proof. For each n with $1 \leq n \leq 1492$, we can use division with remainder to fix $q_i, r_i \in \mathbb{Z}$ with

$$a_i = 1491q_i + r_i$$

and $0 \leq r_i < 1491$. Since we have 1491 many possible distinct r_i, it follows that there exists $i < j$ with $r_i = r_j$. We then have

$$1491 \mid (a_j - a_i).$$

as in the proof of the previous proposition. The problem is that $a_j - a_i$ does not equal any of the a_n. However, notice that

$$a_j - a_i = 333\cdots 300\cdots 0 = a_{j-i} \cdot 10^i,$$

so as $1491 \mid (a_j - a_i)$, it follows that

$$1491 \mid a_{j-i} \cdot 10^i.$$

Now the prime factorization of 10^i is $2^i 5^i$, so the any prime divisor of 10^i must divide either 2 or 5 by Proposition 3.2.7, so must be either 2 or 5 because 2 and 5 are prime. Now we have $1491 = 2 \cdot 745 + 1$ and $1491 = 5 \cdot 298 + 1$, so $2 \nmid 1491$ and $5 \nmid 1491$ by Proposition 2.3.5. Using Proposition 3.2.9, it follows that $\gcd(1491, 10^i) = 1$. As a result, we can apply Proposition 3.2.5 to conclude that $1491 \mid a_{j-i}$, completing the proof. □

Notice that the above argument works if we replace 1491 by any number that ends in 1, 3, 7, or 9 (since such a number is not divisible by 2 or 5), and we if replace the 3 in $333\cdots 3$ by any nonzero digit.

Proposition 4.3.4. *Suppose we have a gathering of $n \geq 2$ people, and at the beginning of the gathering some pairs of people shake hands. There always must exist (at least) two people who have shaken the same number of hands.*

Proof. Label the people with the numbers $1, 2, 3, \dots, n$. We can then define a function $f \colon \{1, 2, 3, \dots, n\} \to \{0, 1, 2, \dots, n-1\}$ by letting $f(k)$ be the number of people that person k shook hands with. On the face of it, this looks bad because both sets have n elements. However, it is impossible that both 0 and $n - 1$ are elements of $\text{range}(f)$ because if somebody shook hands with all of the other $n - 1$ people, then everybody shook hands with a least one person, so $0 \notin \text{range}(f)$. Thus, we can either view f as a function $f \colon \{1, 2, 3, \dots, n\} \to \{0, 1, 2, \dots, n-2\}$ or as a function $f \colon \{1, 2, 3, \dots, n\} \to \{1, 2, \dots, n-1\}$. In either case, f is not injective by the Pigeonhole Principle, so there exist two people who have shaken the same number of hands. □

Proposition 4.3.5. *Let $f \colon \{0,1\}^* \to \{0,1\}^*$ be injective. For every $n \in \mathbb{N}^+$, there exists $\sigma \in \{0,1\}^n$ with $|f(\sigma)| \geq |\sigma|$ (here $|\tau|$ is the length of the finite sequence τ).*

Proof. Let $f\colon \{0,1\}^* \to \{0,1\}^*$ be injective. Let $n \in \mathbb{N}^+$ be arbitrary. Suppose instead that $|f(\sigma)| < |\sigma|$ for all $\sigma \in \{0,1\}^n$. Notice that $|\{0,1\}|^n = 2^n$ and the number of sequences of length strictly less than n is $1+2+4+\cdots+2^{n-1}$ because we can write it as the union $\{0,1\}^0 \cup \{0,1\}^1 \cup \{0,1\}^2 \cup \cdots \cup \{0,1\}^{n-1}$ where the sets are pairwise disjoint. Now the key fact is that

$$1+2+2^2+\cdots+2^{n-1} = \frac{2^n-1}{2-1} = 2^n - 1$$

by the homework. Alternatively, this can be argued by noting that

$$\begin{aligned}
1+2+2^2+\cdots+2^{n-1} &= (1+2+2^2+\cdots+2^{n-1})\cdot 1 \\
&= (1+2+2^2+\cdots+2^{n-1})\cdot(2-1) \\
&= (2+2^2+2^3+\cdots+2^n) - (1+2+2^2+\cdots+2^{n-1}) \\
&= 2^n - 1.
\end{aligned}$$

Since $|\{0,1\}^n| = 2^n$ and the set of sequences of length strictly less than n is $2^n - 1$, we may use the Pigeonhole Principle to conclude that there exists distinct $\sigma_1, \sigma_2 \in \{0,1\}^n$ with $f(\sigma_1) = f(\sigma_2)$, which contradicts the fact that f is injective. Therefore, there must exist $\sigma \in \{0,1\}^n$ with $|f(\sigma)| \geq |\sigma|$. □

We can interpret the previous proposition as follows. Suppose that we have a compression algorithm, i.e. a program that takes a sequence of 0's and 1's and tries to compress it down to a shorter sequence (think of any standard zip program). If we look at how the function behaves on every input, we obtain a function $f\colon \{0,1\}^* \to \{0,1\}^*$. Of course, for this compression algorithm to be at all useful, we would need to be able to uncompress any file back to its original. In order to do this, the function f must be injective (otherwise, if two files compress to the same thing, we would have no way to know which file to return). This proposition says that any purported compression scheme must fail to actually shrink the size of some file, and in fact for *every* length n, there is a file of length n that is not actually made smaller.

Proposition 4.3.6. *Let $n \in \mathbb{N}^+$. Given a set $S \subseteq \{1,2,3,\dots,2n\}$ with $|S| \geq n+1$, there always exists a pair of distinct elements $a, b \in S$ with $a \mid b$.*

Before proving this proposition, we examine some special cases in order to get some intuition. First, consider the case when $n = 5$ so that $2n = 10$. We want to prove that whenever we have at least 6 numbers from the set $\{1,2,3,\dots,10\}$, we can find two distinct numbers a and b such that $a \mid b$. The idea is to build five "boxes" of numbers with the following properties:

- Every number from $\{1,2,3,\dots,10\}$ is in a box.
- Given any two distinct numbers from the same box, one divides the other.

Suppose that we are successful in doing this. Then given any set of at least six numbers, we can find two of the numbers in the same box (because we only five boxes), and then we will be done. So let's build five boxes with the above properties in the case where $n = 5$:

- Box 1: $\{1,2,4,8\}$.
- Box 2: $\{3,6\}$.
- Box 3: $\{5,10\}$.
- Box 4: $\{7\}$.
- Box 5: $\{9\}$.

We now want to generalize this argument. The key idea behind the above boxes was as follows: Given a natural number, keep dividing by 2 until we reach an odd number, and put two numbers in the same box if we arrive at the same odd number. In order to formalize this, we prove the following lemma.

Lemma 4.3.7. *Let $n \in \mathbb{N}^+$. There exist unique $k, \ell \in \mathbb{N}$ such that ℓ is odd and $n = 2^k \ell$.*

Although it is possible to deduce this result from the Fundamental Theorem of Arithmetic, we give a direct proof.

Proof. We first prove the existence of k and ℓ by strong induction on n.

- When $n = 1$, we can write $1 = 2^0 \cdot 1$, so we can take $k = 0$ and $\ell = 1$.
- Let $n \in \mathbb{N}$ with $n \geq 2$, and assume that we know the existence part is true for all m with $1 \leq m < n$. We prove it for n. First, notice that if n is odd, then we can simply write $n = 2^0 n$, and we are done. Suppose then that n is even. Fix $m \in \mathbb{Z}$ with $n = 2m$ and notice that $1 \leq m < n$. By induction, we can fix $k, \ell \in \mathbb{N}$ such that ℓ is odd and $m = 2^k \ell$. We then have $n = 2m = 2^{k+1}\ell$, hence the result holds for n.

The existence of k and ℓ for all n follows by induction.

We now prove uniqueness. Suppose that $k_1, k_2, \ell_1, \ell_2 \in \mathbb{N}$ are such that ℓ_1 and ℓ_2 are both odd and $2^{k_1}\ell_1 = 2^{k_2}\ell_2$. If $k_1 < k_2$, then dividing both sides by 2^{k_1}, we would be able to conclude that $\ell_1 = 2^{k_2-k_1}\ell_2$, which contradicts the fact that ℓ_1 is odd (since $k_2 - k_1 \geq 1$). A similar contradiction occurs if $k_1 > k_2$. Therefore, we must have that $k_1 = k_2$. Diving both sides by $2^{k_1} = 2^{k_2}$, we then conclude that $\ell_1 = \ell_2$. This gives uniqueness. □

Proof of Proposition 4.3.6. Let $S \subseteq \{1, 2, 3, \dots, 2n\}$ with $|S| \geq n + 1$ be arbitrary. Let X be the set of all odd integers ℓ with $1 \leq \ell \leq 2n$, and notice that $|X| = n$ (because $g\colon \{0, 1, 2, \dots, n-1\} \to X$ given by $g(k) = 2k + 1$ is a bijection). Define a function $f\colon S \to X$ as follows. Given $a \in S$, write $a = 2^k \ell$ for the unique k and ℓ from the previous lemma, and define $f(a) = \ell$ (notice that $\ell \leq 2n$ because $a \leq 2n$). Intuitively, we associate to each given $n \in S$ the unique odd number obtained by repeatedly dividing by 2 until we reach an odd number. Since $|S| \geq n + 1$ and $|X| = n$, the Pigeonhole Principle tells us that we can find distinct $a, b \in S$ with $a < b$ such that $f(a) = f(b)$. Call this common value ℓ, i.e. let $\ell = f(a) = f(b)$, and fix $k_1, k_2 \in \mathbb{N}$ with $a = 2^{k_1}\ell$ and $b = 2^{k_2}\ell$. Since $a < b$, we have $k_1 < k_2$. Now

$$\begin{aligned} b &= 2^{k_2}\ell \\ &= 2^{k_2-k_1} \cdot 2^{k_1} \cdot \ell \\ &= 2^{k_2-k_1} \cdot a, \end{aligned}$$

so $a \mid b$. This completes the proof. □

We end with a more sophisticated example. Suppose that we have a finite sequence of (possibly real) numbers. For example, consider the following sequence of 10 numbers:

$$3 \quad 1 \quad 6 \quad 9 \quad 0 \quad 2 \quad 8 \quad 5 \quad 7 \quad 4$$

Although these numbers are not sorted in any sense, one can find a decently long decreasing subsequence by pulling out the $9, 8, 7, 4$. It turns out that no matter what sequence of length n one looks at, it always possible to pull out an increasing or decreasing subsequence of length about $\sqrt{n}$. We first provide the necessary definitions:

Definition 4.3.8. *Suppose that $a_1, a_2, \dots, a_n$ is a finite sequence of real numbers. Suppose that we have a sequence of indices with $1 \leq i_1 < i_2 < \dots < i_k \leq n$. We then call $a_{i_1}, a_{i_2}, \dots, a_{i_k}$* a subsequence *of $a_1, a_2, \dots, a_n$.*

Definition 4.3.9. *Suppose that $a_1, a_2, \dots, a_n$ is a finite sequence of real numbers.*

- *We call the sequence* increasing *if $a_1 \leq a_2 \leq \cdots \leq a_n$.*
- *We call the sequence* decreasing *if $a_1 \geq a_2 \geq \cdots \geq a_n$.*
- *We call the sequence* monotonic *if it is either increasing or decreasing.*

For example, suppose that $a_1, a_2, \dots, a_{10}$ is our original sequence

$$3 \quad 1 \quad 6 \quad 9 \quad 0 \quad 2 \quad 8 \quad 5 \quad 7 \quad 4$$

Notice that $9, 8, 7, 4$ is a decreasing subsequence of this sequence (with $i_1 = 4$, $i_2 = 7$, $i_3 = 9$, and $i_4 = 10$).

Theorem 4.3.10. *Let $n \in \mathbb{N}^+$. Given a sequence of $(n-1)^2 + 1$ real numbers, it is always possible to find a monotonic subsequence of length n.*

Proof. Consider an arbitrary sequence

$$a_1, a_2, a_3, \dots, a_{(n-1)^2+1}$$

of $(n-1)^2+1$ many real numbers. Associate to each i the pair $(k, \ell) \in \mathbb{N}^+ \times \mathbb{N}^+$, where k is the length of the longest increasing subsequence ending with (and including) a_i and ℓ is the length of the longest decreasing subsequence ending with (and including) a_i. If any one of these pairs has a coordinate that is at least n, then we are done. Otherwise, every pair (k, ℓ) is such that $1 \leq k \leq n-1$ and $1 \leq \ell \leq n-1$. There are only $(n-1)^2$ many possible pairs, so since we have $(n-1)^2 + 1$ many numbers, some pair must be repeated by the Pigeonhole Principle. Fix $i < j$ with $(k_i, \ell_i) = (k_j, \ell_j)$. Now if $a_j \geq a_i$, then we can add a_j onto the end of the longest increasing subsequence ending in a_i to form an increasing subsequence of length $k_i + 1 > k_j$, a contradiction. Similarly, if $a_j \leq a_i$, then we can add a_j onto the end of the longest decreasing subsequence ending in a_i to form an idecreasing subsequence of length $\ell_i + 1 > \ell_j$, a contradiction. □

For example, for our sequence

$$3 \quad 1 \quad 6 \quad 9 \quad 0 \quad 2 \quad 8 \quad 5 \quad 7 \quad 4$$

we would assign the values

$$(1,1) \quad (1,2) \quad (2,1) \quad (3,1) \quad (1,3) \quad (2,2) \quad (3,2) \quad (3,3) \quad (4,3) \quad (3,4)$$

Thus, we can take either $0, 2, 5, 7$ as an increasing subsequence or $9, 8, 7, 4$ as a decreasing subsequence.

4.4 Countability and Uncountability

Recall that if A and B are finite sets, then $|A| = |B|$ if and only if there exists a bijection $f \colon A \to B$. Now if A and B are infinite sets, then we have no obvious way to define the cardinality of A and B like we do for finite sets. However, it still makes sense to talk about bijections, and so one can simply *define* two (possibly infinite) sets A and B to have the same size if there is a bijection $f \colon A \to B$.

With this in mind, think about $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ and the subset $\mathbb{N}^+ = \{1, 2, 3, 4, \dots\}$. Although $\mathbb{N}^+$ is a proper subset of $\mathbb{N}$ and "obviously" has one fewer element, the function $f \colon \mathbb{N} \to \mathbb{N}^+$ given by $f(n) = n + 1$ is a bijection, and so $\mathbb{N}$ and $\mathbb{N}^+$ have the same "size". For another even more surprising example, let $A = \{2n : n \in \mathbb{N}\} = \{0, 2, 4, 6, \dots\}$ be the set even natural numbers, and notice that the function $f \colon \mathbb{N} \to A$ given by $f(n) = 2n$ is a bijection from $\mathbb{N}$ to A. Hence, even though A intuitively seems to only have "half" of the elements of $\mathbb{N}$, there is still a bijection between $\mathbb{N}$ and A.

The next proposition shows that $\mathbb{N}$ is the "smallest" infinite set, in the sense that we can injectively embed it into any infinite set.

Proposition 4.4.1. *If A is an infinite set, then there is an injective function $f\colon \mathbb{N} \to A$.*

Proof. We define $f\colon \mathbb{N} \to A$ recursively. Pick some (any) $a_0 \in A$, and define $f(0) = a_0$. Suppose that $n \in \mathbb{N}$ and we have defined the values $f(0), f(1), \dots, f(n)$, all of which are elements of A. Since A is infinite, we have that $\{f(0), f(1), \dots, f(n)\} \neq A$. Thus, we can pick some $a_{n+1} \in A \backslash \{f(0), f(1), \dots, f(n)\}$, and define $f(n+1) = a_{n+1}$. With this recursive definition, we have defined a function $f\colon \mathbb{N} \to A$. Notice that if $m < n$, then $f(n)$ was chosen to be distinct from $f(m)$ by definition, so $f(m) \neq f(n)$. Therefore, f is injective. $\square$

With this in mind, we introduce a name for those infinite sets for which we can find a bijection with $\mathbb{N}$, and think of them as the "smallest" types of infinite sets.

Definition 4.4.2. *Let A be a set.*

- *We say that A is* countably infinite *if there exists a bijection $f\colon \mathbb{N} \to A$.*
- *We say that A is* countable *if it is either finite or countably infinite.*
- *If A is not countable, we say that A is* uncountable.

Suppose that A is countably infinite. We then have a bijection $f\colon \mathbb{N} \to A$, so we can arrange its elements in a list without repetition by listing out $f(0), f(1), f(2), f(3), \dots$ to get:

$$a_0 \quad a_1 \quad a_2 \quad a_3 \quad \cdots$$

Conversely, writing out such a list without repetition shows how to build a bijection $f\colon \mathbb{N} \to A$. Since working with such lists is more intuitively natural (although perhaps a little less rigorous), we'll work with countable sets in this way. What about lists that allow repetitions?

Proposition 4.4.3. *Let A be a set. The following are equivalent.*

1. *It is possible to list A, possibly with repetitions, as $a_0, a_1, a_2, a_3, \dots$.*
2. *There is a surjection $g\colon \mathbb{N} \to A$.*
3. *A is countable, i.e. either finite or countably infinite.*

Proof. (1) $\Leftrightarrow$ (2): This is essentially the same as the argument just given. If we can list A, possibly with repetitions, as $a_0, a_1, a_2, a_3, \dots$, then the function $g\colon \mathbb{N} \to A$ given by $g(n) = a_n$ is a surjection. Conversely, if there is a surjection $g\colon \mathbb{N} \to A$, then $g(0), g(1), g(2), g(3), \dots$ is a listing of A.

(1) $\Rightarrow$ (3): Assume (1), and fix a listing $a_0, a_1, a_2, a_3, \dots$ of A, possibly with repetition. If A is finite, then A is countable by definition, so we may assume that A is infinite. We define a new list as follows. Let $b_0 = a_0$. If we have defined $b_0, b_1, \dots, b_n$, let $b_{n+1} = a_k$, where k is chosen as the least value such that $a_k \notin \{b_0, b_1, \dots, b_n\}$ (such a k exists because A is infinite). Then

$$b_0 \quad b_1 \quad b_2 \quad b_3 \quad \cdots$$

is a listing of A without repetitions. Therefore, A is countably infinite.

(3) $\Rightarrow$ (1): Suppose that A is countable. If A is countably infinite, then there is a bijection $f\colon \mathbb{N} \to A$, in which case

$$f(0) \quad f(1) \quad f(2) \quad f(3) \quad \cdots$$

is a listing of A (even without repetition). On the other hand, if A is finite, say $A = \{a_0, a_1, a_2, \dots, a_n\}$, then

$$a_0 \quad a_1 \quad a_2 \quad \cdots \quad a_n \quad a_n \quad a_n \quad \cdots$$

is a listing of A (with repetition). $\square$

Our first really interesting result is that $\mathbb{Z}$, the set of integers, is countable. Of course, some insight is required because if we simply start to list the integers as

$$0 \quad 1 \quad 2 \quad 3 \quad 4 \quad \cdots$$

we won't ever get to the negative numbers. We thus use the sneaky strategy of bouncing back-and-forth between positive and negative integers.

Proposition 4.4.4. *$\mathbb{Z}$ is countable.*

Proof. We can list $\mathbb{Z}$ as

$$0 \quad 1 \quad -1 \quad 2 \quad -2 \quad \cdots$$

More formally, we could define $f \colon \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} -\frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases}$$

and check that f is a bijection. $\square$

The key idea used in previous proof can be abstracted into the following result.

Proposition 4.4.5. *If A and B are countable, then $A \cup B$ is countable.*

Proof. Since A is countable, we may list it as $a_0, a_1, a_2, a_3, \dots$. Since B is countable, we may list it as $b_0, b_1, b_2, b_3, \dots$. We therefore have the following two lists:

$$\begin{matrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ b_0 & b_1 & b_2 & b_3 & \cdots \end{matrix}$$

We can list $A \cup B$ by going back-and-forth between the above lists as

$$a_0 \quad b_0 \quad a_1 \quad b_1 \quad a_2 \quad b_2 \quad \cdots$$

$\square$

A slightly stronger result is now immediate.

Corollary 4.4.6. *If $A_0, A_1, \dots, A_n$ are countable, then $A_0 \cup A_1 \cup \cdots \cup A_n$ is countable.*

Proof. This follows from Proposition 4.4.5 by induction. Alternatively, we can argue as follows. For each fixed k with $0 \leq k \leq n$, we know that A_k is countable, so we may list it as $a_{k,0}, a_{k,1}, a_{k,2}, \dots$. We can visualize the situation with the following table.

$$\begin{matrix} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} & \cdots \\ a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ a_{n,0} & a_{n,1} & a_{n,2} & a_{n,3} & \cdots \end{matrix}$$

We now list $A_0 \cup A_1 \cup \cdots \cup A_n$ by moving down each column in order, to obtain:

$$a_{0,0} \quad a_{1,0} \quad \cdots \quad a_{n,0} \quad a_{0,1} \quad a_{1,1} \quad \cdots \quad a_{n,1} \quad \cdots \quad \cdots$$

$\square$

In fact, we can prove quite a significant extension of the above results. The next proposition is usually referred to by saying that "the countable union of countable sets is countable".

Proposition 4.4.7. *If $A_0, A_1, A_2, \dots$ are all countable, then $\bigcup_{k=0}^{\infty} A_k = A_0 \cup A_1 \cup A_2 \cup \cdots$ is countable.*

Proof. For each $n \in \mathbb{N}$, we know that A_n is countable, so we may list it as $a_{k,0}, a_{k,1}, a_{k,2}, a_{k,3}, \dots$. We now have the following table.

$$\begin{array}{ccccc} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} & \cdots \\ a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} & \cdots \\ a_{2,0} & a_{2,1} & a_{2,2} & a_{2,3} & \cdots \\ a_{3,0} & a_{3,1} & a_{3,2} & a_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Now we can't list this by blindly walking down the rows or columns. We thus need a new, much more clever, strategy. The idea is to list the elements of the table by moving between rows and columns. One nice approach which works is to step along certain diagonals and obtain the following listing of $\bigcup_{n=0}^{\infty} A_n$:

$$a_{0,0} \quad a_{0,1} \quad a_{1,0} \quad a_{0,2} \quad a_{1,1} \quad a_{2,0} \quad \cdots$$

The pattern here is that we are walking along the diagonals in turn, each of which is finite. Alternatively, we can describe this list as follows. For each $m \in \mathbb{N}$, there are only finitely many pairs $(i, j) \in \mathbb{N} \times \mathbb{N}$ with $i + j = m$. We first list the finitely many $a_{i,j}$ with $i + j = 0$, followed by those finitely many $a_{i,j}$ with $i + j = 1$, then those finitely many $a_{i,j}$ with $i + j = 2$, etc. This gives a listing of $\bigcup_{k=0}^{\infty} A_k$. $\square$

Theorem 4.4.8. $\mathbb{Q}$ *is countable.*

Proof. For each $k \in \mathbb{N}^+$, let $A_k = \{\frac{a}{k} : a \in \mathbb{Z}\}$. Notice that each A_k is countable because we can list it as

$$\tfrac{0}{k} \quad \tfrac{1}{k} \quad \tfrac{-1}{k} \quad \tfrac{2}{k} \quad \tfrac{-2}{k} \quad \cdots$$

Since

$$\mathbb{Q} = \bigcup_{k=1}^{\infty} A_k = A_1 \cup A_2 \cup A_3 \cup \cdots$$

we can use Proposition 4.4.7 to conclude that $\mathbb{Q}$ is countable. $\square$

With all of this in hand, it is natural to ask whether uncountable sets exist.

Theorem 4.4.9. $\mathbb{R}$ *is uncountable.*

Proof. We need to show that there is no list of real numbers that includes every element of $\mathbb{R}$. Suppose then that $r_1, r_2, r_3, \dots$ is an arbitrary list of real numbers. We show that there exists $x \in \mathbb{R}$ with $x \neq r_n$ for every $n \in \mathbb{N}$. For each $n \in \mathbb{N}$, we write out the (nonterminating) decimal expansion of r_n as

$$a_n \quad . \quad d_{n,1} \quad d_{n,2} \quad d_{n,3} \quad d_{n,4} \quad \cdots$$

where $a_n \in \mathbb{Z}$ and each $d_{n,i} \in \mathbb{Z}$ satisfies $0 \leq d_{n,i} \leq 9$. We arrange our list of reals $r_1, r_2, r_3, \dots$ as a table

$$\begin{array}{ccccccc} a_1 & . & d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & \cdots \\ a_2 & . & d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} & \cdots \\ a_3 & . & d_{3,1} & d_{3,2} & d_{3,3} & d_{3,4} & \cdots \\ a_4 & . & d_{4,1} & d_{4,2} & d_{4,3} & d_{4,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

For each $n \in \mathbb{N}$, let

$$e_n = \begin{cases} 3 & \text{if } d_{n,n} \neq 3 \\ 7 & \text{if } d_{n,n} = 3 \end{cases}$$

Let x be the real number with decimal expansion

$$. \quad e_1 \quad e_2 \quad e_3 \quad e_4 \quad \cdots$$

We claim that $x \neq r_n$ for every $n \in \mathbb{N}$. Let $n \in \mathbb{N}$ be arbitrary. Since $e_n \neq d_{n,n}$ by construction, it follows x and r_n disagree in the n^{th} decimal position. Therefore, since the (nonterminating) decimal expansions of x and r_n are different, it follows that $x \neq r_n$. $\square$

Chapter 5

Counting

5.1 Arrangements, Permutations, and Combinations

Let A be a finite set with $|A| = n$. Given $k \in \mathbb{N}^+$, the set A^k is the set of all finite sequences of length k whose elements are all from A. Occasionally, and especially in computer science, such a finite sequence is called a *string* over A of length k. Using Corollary 1.2.11, we already know that $|A^k| = |A|^k = n^k$, so we can count the number of finite sequences of length k. For example, if $A = \{a, b, c, d\}$, then there are exactly $4^2 = 16$ many two letter strings over A. Similarly, there are exactly $128^2 = 16,384$ many two character long ASCII sequences, and there are 10^7 many potential phone numbers.

Recall that a finite sequence might contain repetition. For example, if $A = \{1, 2, 3\}$, then $(1, 1, 3) \in A^3$ and $(3, 1, 2, 3) \in A^4$. Suppose that A is a set with $|A| = n$, and we want to count the number of sequences of length 2 where there is no repetition, i.e. we want to determine the cardinality of the set

$$B = \{(a, b) \in A^2 : a \neq b\}.$$

There are (at least) two straightforward ways to do this:

- *Method 1:* As on the first homework, we use the complement rule. Let $D = \{(a, a) : a \in A\}$ and notice that $|D| = n$ because $|A| = n$ (and $f \colon A \to D$ given by $f(a) = (a, a)$ is a bijection). Since $B = A^2 \backslash D$, it follows that $|B| = |A^2| - |D| = n^2 - n = n(n - 1)$.

- *Method 2:* We use a modified version of the product rule as follows. Think about constructing an element of B in two stages. First, we need to pick the first coordinate of our pair, and we have n choices here. Now once we fix the first coordinate of our pair, we have $n - 1$ choices for the second coordinate because we can choose any element of A other than the one that we chose in the first round. By making these two choices in succession, we determine an element of B, and furthermore, every element of B is obtained via a unique sequence of such choices. Therefore, we have $|B| = n(n - 1)$.

Notice that in the argument for Method 2 above, we are not directly using the Product Rule. The issue is that we can not write B in the form $B = X \times Y$ where $|X| = n$ and $|Y| = n - 1$ because the choice of second coordinates depends upon the choice of first component. For example, if $A = \{1, 2, 3\}$, then if we choose 1 as our first coordinate, then we can choose any element of $\{2, 3\}$ for the second, while if we choose 3 as our first coordinate, then we can choose any element of $\{1, 2\}$ for the second. However, the key fact is that the *number* of choices for the second coordinate is the same no matter what we choose for the first. If you want to be more formal in the example with $A = \{1, 2, 3\}$, we are setting up a bijection between $\{1, 2, 3\} \times \{1, 2\}$

and B as follows:

$$\begin{aligned}(1,1) &\mapsto (1,2)\\(1,2) &\mapsto (1,3)\\(2,1) &\mapsto (2,1)\\(2,2) &\mapsto (2,3)\\(3,1) &\mapsto (3,1)\\(3,2) &\mapsto (3,2)\end{aligned}$$

In other words, if we input (k,ℓ), then the output will have first coordinate k, but the second coordinate will be the ℓ^{th} element smallest element of $\{1,2,3\}\backslash\{k\}$. Moreover, if $A = \{1,2,3,\ldots,n\}$, then function from $\{1,2,\ldots,n\} \times \{1,2,\ldots,n-1\}$ that sends (k,ℓ) to the pair whose first coordinate is k, and whose second coordinate is the ℓ^{th} smallest element of $\{1,2,\ldots,n\}\backslash\{k\}$ is a bijection, so $|B| = n(n-1)$.

Suppose more generally that we are building a set of objects in stages, in such a way that each *sequence* of choices throughout the stages determines a unique object, and no two distinct sequences determine the same object. Suppose also that we have the following number of choices at each stage:

- There are n_1 many choices at the first stage.
- For each choice in the first stage, there are n_2 many objects to pair with it in the second stage.
- For each pair of choices in the first two stages, there are n_3 many choices to append at the third stage.
- ...
- For each sequence of choices in the first $k-1$ stages, there are n_k many choices to append at the k^{th} stage.

In this situation, there are $n_1 n_2 n_3 \cdots n_k$ many total objects in the set. The formal argument is similar to the above argument, where we set up a bijection from the set

$$\{1,2,\ldots,n_1\} \times \{1,2,\ldots,n_2\} \times \cdots \times \{1,2,\ldots,n_k\}$$

to the set that we are counting. However, we will avoid such formalities in the future when we use it. With this new rule in hand, we can count a new type of object.

Definition 5.1.1. *Let A be a finite set with $|A| = n$. A* permutation *of A is an element of A^n without repeated elements.*

For example, consider $A = \{1,2,3\}$. One example of a permutation of A is $(3,1,2)$. The set of all permutations of A is:

$$\{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\}.$$

Thus, there are 6 permutations of the set $\{1,2,3\}$. In order to count the number of permutations of a set with n elements, we use our new technique.

Proposition 5.1.2. *If A is a finite set with $n \in \mathbb{N}^+$ elements, then there are $n!$ many permutations of A.*

Proof. We can build a permutation of A through a sequence of choices.

- We being by choosing the first element, and we have n choices.
- Once we've chosen the first element, we have $n-1$ choices for the second because we can choose any element of A other than the one chosen in the first stage.

- Next, we have $n-2$ many choices for the third element.
- ...
- At stage $n-1$, we have chosen $n-2$ distinct elements so far, so we have 2 choices here.
- Finally, we have only one choice remaining for the last position.

Since every such sequence of choices determines a permutation of A, and distinct choices given distinct permutations, it follows that there are $n(n-1)(n-2)\cdots 2\cdot 1 = n!$ many permutations of A. $\square$

Alternatively, we can give a recursive description of the number of permutations of a set with n-elements, and use that to derive the above result. Define $f\colon \mathbb{N}^+ \to \mathbb{N}^+$ by letting $f(n)$ be the number of permutations of $\{1,2,\dots,n\}$. Notice that $f(1)=1$. Suppose that we know the value of $f(n)$. We show how to build all permutations of $\{1,2,\dots,n+1\}$ from the $f(n)$ many permutations of $\{1,2,\dots,n\}$ along with an element of the set $\{1,2,\dots,n,n+1\}$. Given a permutation of $\{1,2,\dots,n\}$ together with a number k with $1\le k\le n+1$, we form a permutation of $\{1,2,\dots,n,n+1\}$ by taking our permutation of $\{1,2,\dots,n\}$, and inserting $n+1$ into the sequence in position k (and then shifting all later numbers to the right). For example, when $n=4$ and we have the permutation $(4,1,3,2)$ together with $k=2$, then we insert 5 into the second position to form the permutation $(4,5,1,3,2)$.

In this way, we form all permutations of $\{1,2,\dots,n,n+1\}$ in a unique way. More formally, if we let $\mathcal{R}_n$ is the set of all permutations of $\{1,2,\dots,n\}$, then this rule provides a bijection from $\mathcal{R}_n \times \{1,2,\dots,n,n+1\}$ to $\mathcal{R}_{n+1}$. Therefore, we have $f(n+1) = (n+1)\cdot f(n)$ for all $n\in\mathbb{N}^+$. Combining this with the fact that $f(1)=1$, we conclude that $f(n)=n!$ for all $n\in\mathbb{N}^+$.

Definition 5.1.3. *Let A be a finite set with $|A|=n$, and let $k\in\mathbb{N}$ with $1\le k\le n$. A* partial permutation *of A of length k is an element of A^k with no repeated element. A partial permutation of length k is also called a k-*permutation *of A.*

Proposition 5.1.4. *If A is a finite set with $n\in\mathbb{N}^+$ elements and $k\in\mathbb{N}^+$ is such that $1\le k\le n$, then there are*

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

many k-permutations of A.

Proof. The proof is the same as for permutations, except we stop after k stages. Notice that the last term in the product, corresponding to the number of choices at stage k, is $n-(k-1) = n-k+1$ because at stage k we have chosen the first $k-1$ many element. Finally, notice that

$$\begin{aligned}\frac{n!}{(n-k)!} &= \frac{n(n-1)(n-2)\cdot(n-k+1)(n-k)(n-k-1)\cdots 1}{(n-k)(n-k-1)\cdots 1}\\ &= n(n-1)(n-2)\cdots(n-k+1)\end{aligned}$$

giving the last equality. $\square$

For example, using the standard 26-letter alphabet, there are $26\cdot 25\cdot 24 = \frac{26!}{23!} = 15,600$ many three-letter strings of letters having no repetition.

Notation 5.1.5. *If $k,n\in\mathbb{N}^+$ with $1\le k\le n$, we use the notation $(n)_k$ or $P(n,k)$ for the number of k-permutations of a set with n elements, i.e. we define*

$$(n)_k = P(n,k) = \frac{n!}{(n-k)!}.$$

We now count the number of functions between two finite sets, as well as the number of injections (it turns out that the number of surjections is much harder).

Proposition 5.1.6. *Suppose that A and B are finite sets with $|A| = m$ and $|B| = n$.*

1. *The number of functions from A to B is n^m.*
2. *If $m \leq n$, then the number of injective functions from A to B is $P(n,m) = \frac{n!}{(n-m)!}$.*

Proof.

1. To see this, first list the elements of A in some order as $a_1, a_2, \dots, a_m$. A function assigns a unique value in B to each a_i, so we go through the a_i in order. For a_1, we have n possible images because we can choose any element of B. Once we've chosen this, we now have n possible images for a_2. As we go along, we always have n possible images for each of the a_i. Therefore, the number of functions from A to B is $n \cdot n \cdots n = n^m$.

2. Notice that if $n < m$, then there are no injective functions $f \colon A \to B$ by the Pigeonhole Principle. Suppose instead that $m \leq n$. The argument here is similar to the one for general functions, but we get fewer choices as we progress through A. As above, list the elements of A in some order as $a_1, a_2, \dots, a_m$. A function assigns a unique value in B to each a_i, so we go through that a_i in order. For a_1, we have n possible images because we can choose any element of B. Once we've chosen this, we now have $n-1$ possible images for a_2 because we can choose any value of B other than the one we sent a_1 to. Then we have $n-2$ many choices for a_3, etc. Once we arrive at a_m, we have already used up $m-1$ many elements of B, so we have $n-(m-1) = n-m+1$ many choices for where to send a_m. Therefore, the number of functions from A to B is

$$n \cdot (n-1) \cdot (n-2) \cdots (n-m+1) = \frac{n!}{(n-m)!},$$

which is $P(n,m)$.

□

Suppose we ask the following question: Given $A = \{1,2,3,4,5,6,7\}$, how many elements of A^4 contain the number 7 at least once? In other words, how many four digit numbers are there such that each digit is between 1 and 7 (inclusive), and 7 occurs at least once? A natural guess is that the answer is $4 \cdot 7^3$ by the following argument:

- First, pick one of the 4 positions to place the 7.
- Now we have three positions open. Going through them in order, we have 7 choices for what to put in each of these three positions.

This all looks great, but unfortunately, there is a problem. It is indeed true that such a sequence of four choices does create one of the numbers we are looking for. If we choose the sequence $(3,1,5,1)$, saying that we put a 7 in the third position and then place $(1,5,1)$ in order in the remaining positions, then we obtain the number 1571. However, the sequence of choices $(2,7,3,4)$ and the sequence of choices $(1,7,3,4)$ both produce the same string, namely 7734. More formally, the function that takes a position (for the first 7) together with a sequence of 3 digits, and produces the corresponding 4-digit sequence that contains a 7, is surjective but not injective. As a result, we do not have a bijection, and so can not count the set in this way.

In order to get around this problem, the key idea is to count the complement. That is, instead of counting the number of elements of A^4 that *do* contain the number 7 at least once, we count the number of elements of A^4 that *do not* contain the number 7 at all, and subtract this amount from the total number of elements in A^4. Now since $|A| = 7$, we have that $|A^4| = 7^4$ because we have 7 choices for each of the 4 spots. To count the number of elements of A^4 that do not contain a 7, we simply notice that we have 6 choices for each of the 4 spots, so there are 6^4 of these. Therefore, by the Complement Rule, the number of elements of A^4 that do contain the number 7 at least once is $7^4 - 6^4$.

We next move on to a fundamental question that will guide a lot of our later work. Let $n \in \mathbb{N}^+$. We know that there are 2^n many subsets of $\{1, 2, \ldots, n\}$. However, what if we ask how many subsets there are of a certain size? For instance, how many subsets are there of $\{1, 2, 3, 4, 5\}$ that have exactly 3 elements? The intuitive idea is to make 3 choices: First, pick one of the 5 elements to go into our set. Next, pick one of the 4 remaining elements to add to it. Finally, finish off the process by picking one of the 3 remaining elements. For example, if we choose the number $1, 3, 5$ then we get the set $\{1, 3, 5\}$. Thus, a natural guess is that there are $5 \cdot 4 \cdot 3$ many subsets with 3 elements. However, recall that a set has neither repetition nor order, so just as in the previous example we count the same set multiple times. For example, picking the sequence $3, 5, 1$ would also give the set $\{1, 3, 5\}$. In fact, we arrive at the set $\{1, 3, 5\}$ in the following six ways:

$$\begin{array}{ccc} (1,3,5) & (3,1,5) & (5,1,3) \\ (1,5,3) & (3,5,1) & (5,3,1) \end{array}$$

In hindsight, we realize that were just counting the number of 3-permutations of $\{1, 2, 3, 4, 5\}$, since the order matters there.

At this point, we may be tempted to throw our hands in the air as we did above. However, there is one crucial difference. In our previous example, some sequences of 4 numbers including a 7 were counted once (like 1571), some were counted twice (like 7712), and others were counted three or four times. However, in our current situation, *every* subset is counted exactly 6 times because given a set with 3 elements, we know that there are $3! = 6$ many permutations of that set (i.e. ways to arrange the elements of the set in order). The fact that we count each element 6 times means that the total number of subsets of $\{1, 2, 3, 4, 5\}$ having exactly 3 elements equals $\frac{5 \cdot 4 \cdot 3}{6} = 10$. The general principle that we are applying is the following:

Proposition 5.1.7 (Quotient Rule)**.** *Suppose that A is a finite set with $|A| = n$. Suppose that $\sim$ is an equivalence relation on A, and that every equivalence class has exactly k elements. In this case there are $\frac{n}{k}$ many equivalence classes.*

Proof. Let ℓ be the number of equivalence classes. To obtain an element of A, we can first pick one of the ℓ equivalence classes, and then pick one of the k many elements from that class. Since the equivalence classes partition A, it follows that this sequence of choices produces each element of A in a unique way. Thus, $n = k \cdot \ell$ by the Product Rule, and hence $\ell = \frac{n}{k}$. $\square$

Another way to state the Quotient Rule is in terms of surjective functions.

Proposition 5.1.8 (Quotient Rule - Alternative Form)**.** *Suppose that A and B are a finite sets with $|A| = n$. Suppose that $f \colon A \to B$ is a surjective function and that $k \in \mathbb{N}^+$ has the property that*

$$|\{a \in A : f(a) = b\}| = k$$

for all $b \in B$ (i.e. every $b \in B$ is hit by exactly k elements of A). We then have $|B| = \frac{n}{k}$.

Proof. The argument is similar to previous proof. Let $\ell = |B|$. To obtain an element of A, we can first pick one of the ℓ elements of B, and then pick one of the k many elements from the set $\{a \in A : f(a) = b\}$. Notice that this sequence of choices produces each element of A in a unique way. Thus, $n = k \cdot \ell$ by the Product Rule, and hence $\ell = \frac{n}{k}$. $\square$

In fact, these two versions of the Quotient Rule are really different aspects of the same underlying phenomena. To see this, given a surjective function $f \colon A \to B$, and define relation $\sim$ on A defined by letting $a_1 \sim a_2$ if $f(a_1) = f(a_2)$. It is then straightforward to check that $\sim$ is an equivalence relation on A, and that for all $c \in A$, we have

$$\overline{c} = \{a \in A : f(a) = f(c)\}.$$

In other words, the sets $\{a \in A : f(a) = b\}$ given in the second version are equivalence classes of $\sim$ (since f is surjective). If we also assume that each of these sets have the same size (as we do in the second version), then we are just saying that the equivalence classes have the same size, and hence we can apply the first version of the Quotient Rule.

Proposition 5.1.9. *Let $n, k \in \mathbb{N}^+$ and with $1 \leq k \leq n$. Suppose that A is a finite set with $|A| = n$. The number of subsets of A having exactly k elements equals*

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{k! \cdot (n-k)!}.$$

Proof. We generalize the above argument. We know that the number of k-permutations of A equals

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

Now a k-permutation of A picks k distinct elements of A, put also assigns an order to the elements. Now every subset of A of size k is coded by exactly $k!$ many such k-permutations because we can order the subset in $k!$ many ways. Therefore, by the Quotient Rule, the number of subsets of A having exactly k elements equals

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{k! \cdot (n-k)!}.$$

□

Notice also that if $k = 0$, then there is one subset of any set having zero elements (namely $\emptyset$). Thus, by defining $0! = 1$, the above formula works in the case when $k = 0$ as well.

Definition 5.1.10. *Let $n, k \in \mathbb{N}$ and with $0 \leq k \leq n$. We define the notations $\binom{n}{k}$ and $C(n, k)$ by*

$$\binom{n}{k} = C(n,k) = \frac{n!}{k! \cdot (n-k)!}.$$

We call this the number of k-combinations *of an n-element set, and pronounce $\binom{n}{k}$ as "n choose k".*

For example, the number of 5-card poker hands from a standard 52-card deck is:

$$\binom{52}{5} = \frac{52!}{5! \cdot 48!} = 2,598,960.$$

We now give a number of examples of counting problems:

- Over the standard 26-letter alphabet, how many "words" of length 8 have exactly 5 consonants and 3 vowels? We build every such word in a unique way via a sequence of choices:
  - First, we pick out a subset of 3 of the 8 positions to house the vowels, and we have $\binom{8}{3}$ many possibilities.
  - Next, we pick 3 vowels in order allowing repetition to fill in these positions. Since we have 5 vowels, there are 5^3 many possibilities.
  - Finally, we pick 5 consonants in order, allowing repetition, to fill in remaining 5 positions. Since we have 21 consonants, there are 21^5 many possibilities.

  Since every word is uniquely determined by this sequence of choices, the number of such words is

  $$\binom{8}{3} \cdot 5^3 \cdot 21^6 = 56 \cdot 5^3 \cdot 21^5.$$

- How many ways are there to seat n people around a circular table (so the only thing that matters is the relative position of people with respect to each other)? To count this, we use the Quotient Rule. We first consider each of the chairs as distinct. List the people in some order, and notice that we have n choices for where to seat the first person, then $n-1$ for where to seat the second, then $n-2$ for the third, and so forth. Thus, if the seats are distinct, then we have $n!$ many ways to seat the people. However, two such seating arrangements are equivalent if we can get one from the other via a rotation of the seats. Since there are n possible rotations, each seating arrangement occurs n times in this count, so the total number of such seatings is $\frac{n!}{n} = (n-1)!$.

  More formally, we can think about this as following. Consider all permutations of an n-element set (the people): we know that there are $n!$ of these. Now given two permutations, which are just sequences of length n without repetition, we consider two of these sequences equivalent exactly when every pair of numbers is the same distance apart where we allow "wrap around" (since the seating is circular). We then have that two such sequences are equivalent precisely when one is a cyclic shift of the other. Thus, every equivalence class has exactly n elements, and hence there are $\frac{n!}{n} = (n-1)!$ many circular arrangements.

- Suppose that we are in a city where all streets are straight and either east-west or north-south. Suppose that we are at one corner, and want to travel to a corner that is m blocks east and n blocks north, but we want to do it efficiently. More formally, we want to count the number of ways to get from the point $(0,0)$ to the point (m,n) where at each stage we either increase the x-coordinate by 1 or we increase the y-coordinate by 1. At first sight, it appears that we at each intersection, we have 2 choices: Either go east or go north. However, this is not really the case, because if we can east m times, then we are forced to go north the rest of the way. The idea for how to count this is that such a path is uniquely determined by a sequence of $m+n$ many E's and N's (representing east and north) having exactly m many E's. To determine such a sequence, we need only choose the positions of the m many E's, and there are
$$\binom{m+n}{m}$$
  many choices. Of course, we could instead choose the positions of n many N's to count it as
$$\binom{m+n}{n}$$
  which is the same number.

- How many anagrams (i.e. rearrangements of the letters) are there of MISSISSIPPI? Here is one approach. Notice that MISSISSIPPI has one M, four I's, four S's, and two P's, for a total of eleven letters. First pick the position of the M and notice that we have 11 choices. Once that is done, pick the position of the four I's and notice that this amount to picking a 4 element subset of the remaining 10 positions. There are $\binom{10}{4}$ many such choices. Once that is done, pick the position of the four S's and notice that this amount to picking a 4 element subset of the remaining 6 positions. There are $\binom{6}{4}$ many such choices. Once this is done, the position of the two P's is fixed. This gives a total number of anagrams equal to
$$11\binom{10}{4}\binom{6}{4} = 11 \cdot \frac{10!}{4! \cdot 6!} \cdot \frac{6!}{4! \cdot 2!} = \frac{11!}{4! \cdot 4! \cdot 2!} = 34,650.$$
  Another argument is as follows. Think of distinguishing common letters with different colors. We then have $11!$ many ways to rearrange the letters, but this number overcounts the numbers of anagrams. Each actual anagram comes about in $4! \cdot 4! \cdot 2!$ many ways because we can permute the currently distinct four I's amongst each other in $4!$ ways, we can permute the currently distinct four S's amongst each

other in 4! ways, and we can permute the the currently distinct two P's amongst each other in 2! many ways. Thus, since each actual anagram is counted $4! \cdot 4! \cdot 2!$ many times in the 11! count, it follows that there are

$$\frac{11!}{4! \cdot 4! \cdot 2!} = 34,650$$

many anagrams of MISSISSIPPI.

As mentioned above, there are a total of

$$\binom{52}{5} = 2,598,960$$

many (unordered) 5-card poker hands from a standard 52-card deck. Using this, we now count the number of special hands of each type, as well as the probability of being dealt such a hand (this probability is calculated by dividing the number of such hands by the the total number $2,598,960$). We use the fact that each card has one of four suits (clubs, diamonds, hearts, and spades) and one of thirteen ranks (2, 3, 4, 5, 6, 7, 8, 9,10, jack, queen, king, ace). We follow the common practice of allowing the ace to be either a low card or a high card for a straight, but we do not allow "wrap around" straights such as king, ace, 2, 3, 4.

- Straight Flush: There are

  $$4 \cdot 10 = 40$$

  many of these because they are determined by picking the suit and then picking the rank of the lowest card (from ace through 10). The probability is about .00154%.

- Four of a kind: There are

  $$13 \cdot 48 = 624$$

  of these because we choose a rank (and take all four cards of that rank), and then choose one of the remaining 48 cards. The probability is about .0256%.

- Full House: There are

  $$13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2} = 3,744$$

  many, which can be seen by making the following sequence of choices:

  - Choose one of the 13 ranks for the three of a kind.
  - Choose 3 of the 4 suits for the three of a kind.
  - Choose one of the 12 remaining ranks for the pair.
  - Choose 2 of the 4 suits for the pair.

  The probability is about .14406%.

- Flush: There are

  $$4 \cdot \binom{13}{5} = 5,148$$

  many because we need to choose 1 of the 4 suits, and then 5 of the 13 ranks. However, 40 of these are actually straight flushes, so we really have $5,108$ many flushes that are not stronger hands. The probability is about .19654%

- Straight: There are
$$10 \cdot 4^5 = 10,240$$
many because we need to choose the rank of the lowest card, and the suits for the five cards in increasing order of rank. However, we again have that 40 of these are straight flushes, so we really have 10,200 many straights that are not stronger hands. The probability is about .39246%.

- Three of a kind: There are
$$13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4^2 = 54,912$$
many, which can be seen by making the following sequence of choices:
  - Choose one of the 13 ranks for the three of a kind.
  - Choose 3 of the 4 suits for the three of a kind.
  - Choose two of the other ranks for the remaining two cards (they are different because we do not want to include full houses).
  - Choose the suit of the lower ranked card not in the three of a kind.
  - Choose the suit of the higher ranked card not in the three of a kind.

  (Alternatively, we can choose the last two cards in different ranks in $48 \cdot 44$ many ways, but then we need to divide by 2 because the order of choosing these does not matter.) The probability is about 2.1128%.

- Two Pair: There are
$$\binom{13}{2} \cdot \binom{4}{2}^2 \cdot 44 = 123,552$$
many, which can be seen by making the following sequence of choices:
  - Choose the two ranks for the two pairs.
  - Choose the two suits for the lower ranked pair.
  - Choose the two suits for the higher ranked pair.
  - Choose one of the 44 cards not in these two ranks.

  The probability is about 4.7539%.

- One pair: There are
$$13 \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot 4^3 = 1,098,240$$
many, which can be seen by making the following sequence of choices:
  - Choose the rank for the pair.
  - Choose the two suits for the pair.
  - Choose three distinct ranks for the other three cards (which are not the same rank as the pair).
  - Choose the suit of the lowest ranked card not in the pair.
  - Choose the suit of the middle ranked card not in the pair.
  - Choose the suit of the highest ranked card not in the pair.

  The probability is about 42.257%.

5.2 The Binomial Theorem and Properties of Binomial Coefficients

Recall that if $n, k \in \mathbb{N}$ with $k \leq n$, then we defined

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Notice that when $k = n = 0$, then $\binom{n}{k} = 1$ because we define $0! = 1$, and indeed there is a unique subset of $\emptyset$ having 0 elements, namely $\emptyset$. When $n, k \in \mathbb{N}$ with $n < k$, then we define

$$\binom{n}{k} = 0$$

because there are no subsets of an n-element set with cardinality k (notice that the above formula doesn't make sense because $n - k < 0$).

Using Proposition 4.2.4, we know that whenever $k, n \in \mathbb{N}$ are such that $k \leq n$, then

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

because the function that takes the relative complement is a bijection between subsets of cardinality k and subsets of cardinality $n - k$. Of course, one can prove this directly from the formulas because

$$\begin{aligned}
\binom{n}{n-k} &= \frac{n!}{(n-k)! \cdot (n-(n-k))!} \\
&= \frac{n!}{(n-k)! \cdot k!} \\
&= \frac{n!}{k! \cdot (n-k)!} \\
&= \binom{n}{k}.
\end{aligned}$$

Although the algebraic manipulations here are easy, the bijective proof feels more satisfying because it "explains" the formula. Proving that two numbers are equal by showing that the both count the numbers of elements in one common set, or by proving that there is a bijection between a set counted by the first number and a set counted by the second, is called either a *combinatorial proof* or a *bijective proof*.

Proposition 5.2.1. *Let $n, k \in \mathbb{N}^+$ with $0 < k < n$. We have*

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

Proof. One extremely unenlightening proof is to expand out the formula on the right and do terrible algebraic manipulations on it. If you haven't done so, I encourage you to do it. However, we use the combinatorial description of $\binom{n}{k}$ to give a more meaningful combinatorial argument. Let $n, k \in \mathbb{N}$ with $k \leq n$. Consider a set A with n many elements. To determine $\binom{n}{k}$, we need to count the number of subsets of A of size k. We do this as follows. Fix an arbitrary $a \in A$. Now an arbitrary subset of A of size k fits into exactly one of the following types:

- The subset has a as an element. In this case, to completely determine the subset, we need to pick the remaining $k - 1$ elements of the subset from $A \backslash \{a\}$. Since $A \backslash \{a\}$ has $n - 1$ elements, the number of ways to do this is $\binom{n-1}{k-1}$.

- The subset does not have a as an element. In this case, to completely determine the subset, we need to pick all k elements of the subset from $A\backslash\{a\}$. Since $A\backslash\{a\}$ has $n-1$ elements, the number of ways to do this is $\binom{n-1}{k}$.

Since every subset of A of size k fits into exactly one these types, we have written the collection of all such subsets as a disjoint union (of those satisfying the first condition, and those satisfying the second). By the Sum Rule, we conclude that the number of subsets of A of size k equals $\binom{n-1}{k-1}+\binom{n-1}{k}$. □

Using this proposition, together with the fact that

$$\binom{n}{0}=1 \qquad \text{and} \qquad \binom{n}{n}=1$$

for all $n \in \mathbb{N}$, we can compute $\binom{n}{k}$ recursively to obtain the following table. The rows are labeled by n and the columns by k. To determine the number that belongs in a given square, we simply add the number above it and the number above and to the left. This table is known as *Pascal's Triangle*:

$\binom{n}{k}$	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
2	1	2	1	0	0	0	0	0
3	1	3	3	1	0	0	0	0
4	1	4	6	4	1	0	0	0
5	1	5	10	10	5	1	0	0
6	1	6	15	20	15	6	1	0
7	1	7	21	35	35	21	7	1

There are many curious properties of Pascal's Triangle that we will discover in time. One of the first things to note is that these numbers seem to appear in other places. For example, if $x, y \in \mathbb{R}$, then we have:

- $(x+y)^1 = x+y$
- $(x+y)^2 = x^2+2xy+y^2$
- $(x+y)^3 = x^3+3x^2y+3xy^2+y^3$
- $(x+y)^4 = x^4+4x^3y+6x^2y^2+4xy^3+y^4$

Looking at these, it appears that the coefficients are exactly the corresponding elements of Pascal's Triangle. What is the connection here? Notice that if we do not use commutativity and do not collect like terms (so just use distributivity repeatedly), we have

$$\begin{aligned}(x+y)^2 &= (x+y)(x+y)\\ &= x(x+y)+y(x+y)\\ &= xx+xy+yx+yy,\end{aligned}$$

and so

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)^2\\ &= (x+y)(xx+xy+yx+yy)\\ &= x(xx+xy+yx+yy)+y(xx+xy+yx+yy)\\ &= xxx+xxy+xyx+xyy+yxx+yxy+yyx+yyy.\end{aligned}$$

In other words, it looks like when we fully expand $(x+y)^n$, without using commutativity or collecting x's and y's, then we are getting a sum of all sequences of x's and y's of length n. Thus, if we want to know the coefficient of $x^{n-k}y^k$, then we need only ask how many such sequences have exactly k many y's (or equivalently exactly $n-k$ many x's), and the answer is $\binom{n}{k} = \binom{n}{n-k}$ because we need only pick out the position of the y's (or the x's). More formally, we can prove this by induction.

Theorem 5.2.2 (Binomial Theorem)**.** *Let $x, y \in \mathbb{R}$ and let $n \in \mathbb{N}^+$. We have*

$$\begin{aligned}(x+y)^n &= \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n \\ &= \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k \\ &= \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\end{aligned}$$

Proof. We prove the result by induction. When $n = 1$, we trivially have

$$(x+y)^1 = x + y = \binom{1}{0}x + \binom{1}{1}y$$

Suppose then that we have an $n \in \mathbb{N}^+$ for which we know that the statement is true. We then have

$$\begin{aligned}(x+y)^{n+1} &= (x+y)^n \cdot (x+y) \\ &= \left(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n\right) \cdot (x+y) \\ &= \left(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n\right) \cdot x \\ &\qquad \left(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n\right) \cdot y \\ &= \binom{n}{0}x^{n+1} + \binom{n}{1}x^n y + \binom{n}{2}x^{n-1}y^2 + \cdots + \binom{n}{n-1}x^2y^{n-1} + \binom{n}{n}xy^n \\ &\qquad + \binom{n}{0}x^n y + \binom{n}{1}x^{n-1}y^2 + \cdots + \binom{n}{n-2}x^2y^{n-1} + \binom{n}{n-1}xy^n + \binom{n}{n}y^{n+1} \\ &= x^{n+1} + \left(\binom{n}{1} + \binom{n}{0}\right) \cdot x^n y + \left(\binom{n}{2} + \binom{n}{1}\right) \cdot x^{n-1}y^2 + \cdots + \left(\binom{n}{n} + \binom{n}{n-1}\right) \cdot xy^n + y^{n+1} \\ &= \binom{n+1}{0}x^{n+1} + \binom{n+1}{1}x^n y + \binom{n+1}{2}x^{n-1}y^2 + \cdots + \binom{n+1}{n}xy^n + \binom{n+1}{n+1}y^{n+1},\end{aligned}$$

where we have used Proposition 5.2.1 to combine each of the sums to get the last line. □

Corollary 5.2.3. *For any $n \in \mathbb{N}^+$, we have*

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n.$$

Proof 1. We use the Binomial Theorem in the special case where $x = 1$ and $y = 1$ to obtain

$$\begin{aligned} 2^n &= (1+1)^n \\ &= \sum_{k=0}^{n} \binom{n}{k} \cdot 1^{n-k} \cdot 1^k \\ &= \sum_{k=0}^{n} \binom{n}{k} \\ &= \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}. \end{aligned}$$

This completes the proof. □

Proof 2. Let $n \in \mathbb{N}^+$ be arbitrary. We give a combinatorial proof by arguing that both sides count the number of subsets of an n-element set. Suppose then that A is a set with $|A| = n$. On the one hand, we know that $|\mathcal{P}(A)| = 2^n$ by Corollary 4.2.3.

We now argue that

$$|\mathcal{P}(A)| = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}.$$

For each $k \in \mathbb{N}$ with $0 \leq k \leq n$, let $\mathcal{P}_k(A)$ be the subset of $\mathcal{P}(A)$ consisting of those subsets of A having exactly k elements. We then have that

$$\mathcal{P}(A) = \mathcal{P}_0(A) \cup \mathcal{P}_1(A) \cup \mathcal{P}_2(A) \cup \cdots \cup \mathcal{P}_n(A)$$

and furthermore that the $\mathcal{P}_k(A)$ are pairwise disjoint (i.e. if $k \neq \ell$, then $\mathcal{P}_k(A) \cap \mathcal{P}_\ell(A) = \emptyset$). Therefore,

$$|\mathcal{P}(A)| = |\mathcal{P}_0(A)| + |\mathcal{P}_1(A)| + |\mathcal{P}_2(A)| + \cdots + |\mathcal{P}_n(A)|$$

by the General Sum Rule. Now for each k with $0 \leq k \leq n$, we know that

$$|\mathcal{P}_k(A)| = \binom{n}{k},$$

so it follows that

$$|\mathcal{P}(A)| = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}.$$

Hence

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}$$

because both sides count the number of elements of $\mathcal{P}(A)$. □

Corollary 5.2.4. *For any $n \in \mathbb{N}^+$, we have*

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0$$

Thus

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1} = \binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \ldots$$

Proof 1. We use the Binomial Theorem in the special case where $x = 1$ and $y = -1$ to obtain

$$\begin{aligned} 0 &= 0^n \\ &= (1 + (-1))^n \\ &= \sum_{k=0}^{n} \binom{n}{k} \cdot 1^{n-k} \cdot (-1)^k \\ &= \sum_{k=0}^{n} (-1)^k \binom{n}{k} \\ &= \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n}. \end{aligned}$$

This gives the first claim. Adding $\binom{n}{k}$ to both sides for each odd k, we conclude that

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = \binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \cdots.$$

Since

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$$

by the previous result, it follows that

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1} = \binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \cdots.$$

□

Proof 2. Let $n \in \mathbb{N}^+$ be arbitrary. We begin by giving a combinatorial proof for the second claim. We first show that

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1}.$$

Let A be an arbitrary set with $|A| = n$, and list the elements of A as $A = \{a_1, a_2, \dots, a_n\}$. Recall that we know that $|\mathcal{P}(A)| = 2^n$ because for each i, we have 2 choices for whether or not to include a_i in our subset. Now in our case, the sum on the left

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \dots$$

counts the numbers of subset of A having an even number of elements. We argue that 2^{n-1} also counts the number of subsets of A having an even number of elements. To build these subsets, we make the following sequence of choices:

- Determine whether to include a_1 in our subset: We have 2 choices.
- Determine whether to include a_2 in our subset: We have 2 choices.
- ...
- Determine whether to include a_{n-1} in our subset: We have 2 choices.
- Finally, examine the first $n - 1$ choices, and determine whether we have included an even number of a_i. If so, do not include a_n in our subset. If not, include a_n in our subset.

Notice that in the last step, we do not make any choices, but do one of two things that are completely determined by the previous choices. Now no matter what sequence of choices we make, we end up with a subset of A having an even number of elements, and furthermore every subset with an even number of elements arrises in a unique way. Since there are 2 choices in each of the opening $n-1$ stages, it follows that there are 2^{n-1} many subsets of A with an even number of elements. Therefore,

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1}$$

Now the proof that

$$\binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \cdots = 2^{n-1}$$

is completely analogous except for changing the last stage (or alternatively comes from the complement rule). Finally, since both of these sums equals 2^{n-1}, we conclude that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0.$$

□

Proposition 5.2.5. *For any $n, k \in \mathbb{N}^+$ with $k \leq n$, we have*

$$k \cdot \binom{n}{k} = n \cdot \binom{n-1}{k-1}$$

hence

$$\binom{n}{k} = \frac{n}{k} \cdot \binom{n-1}{k-1}.$$

Proof. We claim that each side counts the number of ways of selecting a committee consisting of k people, including a distinguished president of the committee, from a group of n people. On the one hand, we can do this as follows:

- First pick the committee of k people from the total group of all n people. We have $\binom{n}{k}$ many ways to do this.
- Within this committee, choose one of the k people to serve as president. We have k options here.

Therefore, the number of possibilities is $k \cdot \binom{n}{k}$. On the other hand, we can count it as follows.

- First pick one of the n people to be the president.
- Next pick the remaining $k-1$ many people to serve on the committee amongst the remaining $n-1$ people. We have $\binom{n-1}{k-1}$ many ways to do this.

Therefore, the number of possibilities is $n \cdot \binom{n-1}{k-1}$.

Since each side counts the number of elements of one set, the values must be equal. Therefore,

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

□

Proposition 5.2.6. *For any n, we have*

$$\sum_{k=1}^{n} k \cdot \binom{n}{k} = n \cdot 2^{n-1}.$$

Proof 1. We have

$$\begin{aligned}
\sum_{k=1}^{n} k \cdot \binom{n}{k} &= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} && \text{(by Proposition 5.2.5)} \\
&= n \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} \\
&= n \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} \\
&= n \cdot 2^{n-1} && \text{(by Corollary 5.2.3)}
\end{aligned}$$

□

Proof 2. We give a direct combinatorial proof by arguing that both sides count the number of ways of building a committee, including a distinguished president of that committee, of any size from a group of n people.

One the one hand we can count the number of such committees as follows. We break up the situation into cases based on the size of the committee. For a committee of size k including a distinguished president, we know from Proposition 5.2.5 that there are $k \cdot \binom{n}{k}$ many ways to do this. Since we can break up the collection of all such committees into the pairwise disjoint union of those committees of size 1, those of size 2, etc. Therefore, by the Sum Rule, the number of ways to do this is $\sum_{k=1}^{n} k \cdot \binom{n}{k}$.

On the other hand, we can count the number of such committees differently. First, pick the president of the committee, and notice that we have n choices. Once we pick the president, we need to pick the rest of the committee. Thus, we need to pick a subset (of any size) from the remaining $n-1$ people to fill out the committee, and we know that there are 2^{n-1} many subsets of a set of size $n-1$. Therefore, there are $n \cdot 2^{n-1}$ many such committees.

Since each side counts the number of elements of one set, the values must be equal. Therefore,

$$\sum_{k=1}^{n} k \cdot \binom{n}{k} = n \cdot 2^{n-1}.$$

□

Proof 3. We give another proof using the Binomial Theorem, which tells us that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

for all $x, y \in \mathbb{R}$. Plugging in $y = 1$, we conclude that

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

for all $x \in \mathbb{R}$. Now each side is a function of the real variable x, so taking the derivative of each side, it follows that

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

for all $x \in \mathbb{R}$. Plugging in $x = 1$, we conclude that

$$n \cdot 2^{n-1} = \sum_{k=1}^{n} k \cdot \binom{n}{k}$$

This completes the proof. □

Proposition 5.2.7. *If $k \leq n$, then*

$$\sum_{m=k}^{n} \binom{m}{k} = \binom{k}{k} + \binom{k+1}{k} + \binom{k+2}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1}$$

and since $\binom{m}{k} = 0$ if $m < k$, it follows that

$$\sum_{m=0}^{n} \binom{m}{k} = \binom{n+1}{k+1}.$$

Proof 1. Using Proposition 5.2.1 repeatedly, we have:

$$\begin{aligned}
\binom{n+1}{k+1} &= \binom{n}{k} + \binom{n}{k+1} \\
&= \binom{n}{k} + \binom{n-1}{k} + \binom{n-1}{k+1} \\
&= \binom{n}{k} + \binom{n-1}{k} + \binom{n-2}{k} + \binom{n-2}{k+1} \\
&= \binom{n}{k} + \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k+2}{k+1} \\
&= \binom{n}{k} + \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k+1}{k} + \binom{k+1}{k+1} \\
&= \binom{n}{k} + \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k+1}{k} + \binom{k}{k}.
\end{aligned}$$

where the last line follows from the fact that

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

□

Proof 2. We can also give a combinatorial proof by arguing that

$$\binom{k}{k} + \binom{k+1}{k} + \binom{k+2}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1}$$

counts the number of subsets of $[n+1] = \{1, 2, 3, \dots, n, n+1\}$ having cardinality $k+1$. To see this, first notice that if $A \subseteq [n+1]$ with $|A| = k+1$, then the largest element of A must be at least $k+1$. For each $m \in \mathbb{N}$ with $k+1 \leq m \leq n+1$, let

$$\mathcal{F}_m = \{A \in \mathcal{P}([n+1]) : |A| = k+1 \text{ and } \max(A) = m\}.$$

We then have that the $\mathcal{F}_m$ are pairwise disjoint, and that

$$\mathcal{F}_{k+1} \cup \mathcal{F}_{k+2} \cup \cdots \cup \mathcal{F}_n \cup \mathcal{F}_{n+1}$$

equals the collection of subsets of $[n+1]$ having cardinality $k+1$. Using the General Sum Rule, it follows that

$$|\mathcal{F}_{k+1}| + |\mathcal{F}_{k+2}| + \cdots + |\mathcal{F}_n| + |\mathcal{F}_{n+1}| = \binom{n+1}{k+1}.$$

Now notice that for any $m \in \mathbb{N}$ with $k+1 \leq m \leq n+1$, we have

$$|\mathcal{F}_m| = \binom{m-1}{k}$$

because to determine any $A \in \mathcal{F}_m$, we need only choose the k elements of A that are less than the maximum value m. Therefore,

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

□

Plugging in $k = 1$, we get

$$\binom{1}{1} + \binom{2}{1} + \binom{3}{1} + \cdots + \binom{n}{1} = \binom{n+1}{2}.$$

for all $n \in \mathbb{N}^+$. Since $\binom{m}{1} = m$ for all $m \in \mathbb{N}^+$, it follows that

$$1 + 2 + 3 + \cdots + n = \binom{n+1}{2} = \frac{n(n+1)}{2}.$$

for all $n \in \mathbb{N}^+$. Notice that letting $k = 2$, we conclude that that

$$\binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \cdots + \binom{n}{2} = \binom{n+1}{3}$$

for all $n \in \mathbb{N}^+$. Since $\binom{1}{2} = 0$, we can also write this as

$$\binom{1}{2} + \binom{2}{2} + \binom{3}{2} + \cdots + \binom{n}{2} = \binom{n+1}{3}.$$

Now we can use these to find a formula for the sum of the first n squares:

$$1^2 + 2^2 + 3^2 + \cdots + n^2.$$

The idea is to find $A, B \in \mathbb{R}$ such that

$$m^2 = A \cdot \binom{m}{1} + B \cdot \binom{m}{2}$$

is true for all $m \in \mathbb{N}^+$, because if we can do this, then we can use the above summation formulas for the two sums that appear on the right. Since $\binom{m}{1} = m$ for all $m \in \mathbb{N}^+$, and

$$\binom{m}{2} = \frac{m(m-1)}{2}$$

for all $m \in \mathbb{N}^+$ (even for $m = 1$ because then both sides are 0), we want to find A and B such that:

$$m^2 = A \cdot m + B \cdot \frac{m(m-1)}{2}$$

for all $m \in \mathbb{N}^+$. Now

$$\begin{aligned} A \cdot m + B \cdot \frac{m(m-1)}{2} &= A \cdot m + B \cdot \frac{m^2 - m}{2} \\ &= \left(A - \frac{B}{2}\right) \cdot m + \frac{B}{2} \cdot m^2 \end{aligned}$$

so equating coefficients with $m^2 = 0 \cdot m + 1 \cdot m^2$, we want to solve the linear system:

$$\begin{array}{rcrcl} A & - & \frac{1}{2} \cdot B & = & 0 \\ & & \frac{1}{2} \cdot B & = & 1 \end{array}$$

Now $A = 1$ and $B = 2$ as the unique solution to this system, so it follows that

$$m^2 = \binom{m}{1} + 2 \cdot \binom{m}{2}$$

is true for all $m \in \mathbb{N}^+$. Thus, using Proposition 5.2.7, we conclude that

$$\begin{aligned} 1^2 + 2^2 + \cdots + n^2 &= \left[\binom{1}{1} + 2 \cdot \binom{1}{2}\right] + \left[\binom{2}{1} + 2 \cdot \binom{2}{2}\right] + \cdots + \left[\binom{n}{1} + 2 \cdot \binom{n}{2}\right] \\ &= \left[\binom{1}{1} + \binom{2}{1} + \cdots + \binom{n}{1}\right] + 2 \cdot \left[\binom{1}{2} + \binom{2}{2} + \cdots + \binom{n}{2}\right] \\ &= \binom{n+1}{2} + 2 \cdot \binom{n+1}{3} \\ &= \frac{(n+1)n}{2} + 2 \cdot \frac{(n+1)n(n-1)}{6} \\ &= \frac{3(n+1)n}{6} + \frac{2(n+1)n(n-1)}{6} \\ &= \frac{n(n+1)(3+2n-2)}{6} \\ &= \frac{n(n+1)(2n+1)}{6}. \end{aligned}$$

One can generalize these techniques to get the sum of the first n cubes. Doing so would require finding $A, B, C \in \mathbb{R}$ such that

$$m^3 = A \cdot \binom{m}{1} + B \cdot \binom{m}{2} + C \cdot \binom{m}{3}$$

for all $m \in \mathbb{N}^+$. Although it's not too onerous to do the algebra in order to set up the linear system, and then solve for A, B, C, we will see more unified ways to determine these coefficients (along with for fourth powers, etc.) soon.

Suppose that we want to pick out 5 days from the month of February (having 28 days) in such a way that we do not pick two consecutive days. How can we count it? Although we want to pick out an unordered subset, one idea is to first count the number of *ordered* choices, and then divide by 5!. The idea then is to pick out one day, and we have 28 choices. Once we've picked that day out, we then pick out a second day. It may appear that we have 25 choices here because we've eliminated one day and it's two neighbors. However, that it is only true if we did not pick out the first or last days of February in our first choice. Thus, the number of options in round two depends on our choice from round one. You might think about counting those sets including the first and/or last days of February as special cases, but this doesn't solve all of the problems. For example, if we choose 11 and 18 in our first two rounds, then we've eliminated 6 days and have 22 choices for the third round. However, if we choose 11 and 13 in our first two rounds, then we've only eliminated 5 days and so have 23 choices for the third round. In other words, we need a new way to count this.

Let's attack the problem from a different angle. Instead of trying to avoid bad configurations directly, we think about picking out an arbitrary subset of 5 days and "spreading" them to guarantee that the result will not have any consecutive days. To do this, we will leave the lowest numbered day alone, but add 1 to the second lowest day (to ensure we have a "gap" between the first two), and then add 2 to the middle day, etc. More formally, given an arbitrary subset $\{a_1, a_2, a_3, a_4, a_5\}$ of [28] with $a_1 < a_2 < a_3 < a_4 < a_5$, we turn it into the subset $\{a_1, a_2 + 1, a_3 + 2, a_4 + 3, a_5 + 4\}$ which does not have any consecutive days. The only problem is that now we might "overflow". For example, although

$$\{3, 4, 15, 16, 21\} \mapsto \{4, 6, 17, 19, 25\}$$

works out just fine, we also have

$$\{1, 10, 21, 26, 27\} \mapsto \{1, 11, 23, 30, 31\}$$

which is not allowed. However, there's an easy fix. Instead of picking our original subset from [28], we pick it from [24], for a total of $\binom{24}{5}$ many possibilities. In general, we have the following:

Proposition 5.2.8. *The number of subsets of* $[n] = \{1, 2, 3, \dots, n\}$ *of size* k *having no two consecutive numbers equals* $\binom{n-k+1}{k}$.

Proof. We establish a bijection between the k-element subsets of $[n - k + 1]$ and the sets we want. Given a subset $\{a_1, a_2, a_3, \dots, a_k\}$ of $[n - k + 1]$ with $a_1 < a_2 < a_3 < \dots < a_k$, we map it to the set $\{a_1, a_2 + 1, a_3 + 2, \dots, a_k + (k - 1)\}$, i.e. the i^{th} element of the new set equals $a_i + (i - 1)$. Now since $a_i < a_{i+1}$ for each i, we have that $a_{i+1} - a_i \geq 1$ for each i, and hence

$$\begin{aligned} a_{i+1} + ((i+1) - 1) - (a_i + (i - 1)) &= a_{i+1} + i - a_i - i + 1 \\ &= a_{i+1} - a_i + 1 \\ &\geq 1 + 1 \\ &= 2 \end{aligned}$$

for all i, so there are no consecutive elements in the resulting set. Furthermore, since $a_k \leq n-k+1$, we have $a_k + (k-1) \leq n-k+1+(k-1) = n$, so the resulting subset is indeed a subset of $[n]$ of size k having no two consecutive elements. Notice that this function is injective because if $\{a_1, a_2 + 1, a_3 + 2, \dots, a_k + (k - 1)\} = \{b_1, b_2+1, b_3+2, \dots, b_k+(k-1)\}$, then $a_i+(i-1) = b_i+(i-1)$ for all i, hence $a_i = b_i$ for all i. Furthermore, given a subset $\{c_1, c_2, c_3, \dots, c_k\}$ of $[n]$ with $c_1 < c_2 < c_3 < \dots < c_n$ and $c_{i+1} - c_i \geq 2$ for all i, we have that $\{c_1, c_2 - 1, c_3 - 2, \dots, c_k - (k - 1)\}$ is a subset of $[n - k + 1]$ that maps to $\{c_1, c_2, c_3, \dots, c_k\}$, so it is surjective. The result follows. □

What if we just wanted to count the number number of subsets of $[n]$ having no two consecutive numbers, without any size restrictions? One approach is to sum over all possible sizes to obtain:

$$\sum_{k=0}^{n} \binom{n-k+1}{k} = \binom{n+1}{0} + \binom{n}{1} + \binom{n-1}{2} + \binom{n-2}{3} + \cdots + \binom{1}{n}$$

Of course, many of the terms on the right equal 0 because if $k > n-k+1$, i.e. if $k > \frac{n+1}{2}$, then $\binom{n-k+1}{k} = 0$. Thus, if we let $\lfloor m \rfloor$ be the greatest integer less than or equal to m, then we have

$$\sum_{k=0}^{\lfloor \frac{n+1}{2} \rfloor} \binom{n-k+1}{k}.$$

For example, the number of subsets of $[6] = \{1,2,3,4,5,6\}$ having no two consecutive numbers is

$$\begin{aligned} \sum_{k=0}^{3} \binom{7-k}{k} &= \binom{7}{0} + \binom{6}{1} + \binom{5}{2} + \binom{4}{3} \\ &= 1+6+10+4 \\ &= 21 \end{aligned}$$

while the number of subsets of $[7] = \{1,2,3,4,5,6,7\}$ having no two consecutive numbers is

$$\begin{aligned} \sum_{k=0}^{4} \binom{8-k}{k} &= \binom{8}{0} + \binom{7}{1} + \binom{6}{2} + \binom{5}{3} + \binom{4}{4} \\ &= 1+7+15+10+1 \\ &= 34. \end{aligned}$$

Notice that we are summing up diagonals of Pascal's triangle, and we are seeing Fibonacci numbers. You will prove that this holds true generally on the homework.

Returning to the Binomial Theorem, what happens if we look powers of $x+y+z$ instead of $x+y$? For example, we have

$$\begin{aligned} (x+y+z)^2 &= (x+y+z)(x+y+z) \\ &= x(x+y+z) + y(x+y+z) + z(x+y+z) \\ &= xx + xy + xz + yx + yy + yz + zx + zy + zz \end{aligned}$$

Thus, we obtain a sum of $9 = 3 \cdot 3$ terms, where each term is an ordered product of two elements (with repetition) from $\{x,y,z\}$. If we work out $(x+y+z)^3$, we see a sum of $27 = 3 \cdot 3 \cdot 3$ terms, where each possible ordered sequence of 3 elements (with repetition) from $\{x,y,z\}$ appears exactly once. In general, one expects that we expand $(x+y+z)^n$, then we obtain a sequence of 3^n many terms where each possible ordered sequence of n elements (with repetition) from $\{x,y,z\}$ appears exactly once. What happens when we use collapse these sums by using commutativity, so write $xxz + xzx + zxx$ as $3x^2z$? In general, we are asking what the coefficient of $x^a y^b z^c$ will be in the result? Notice that we need only examine the coefficients where $a+b+c=n$ because each term involves a product of n of the variables. Suppose then that $a+b+c=n$. To know the coefficient of $x^a y^b z^c$, we want to know the number of sequences of x's, y's, and z's of length n having exactly a many x's, b many y's, and c many z's. To count these, we can first pick out that a positions in which to place the x's in $\binom{n}{a}$ many ways. Next, we have $n-a$ open positions, and need to pick out b positions to place the y's in $\binom{n-a}{b}$ many ways. Finally, we have $n-a-b=c$ many positions for the c many

z's, so they are completely determined. Thus, if $a+b+c=n$, then the coefficient of $x^a y^b z^c$ in $(x+y+z)^n$ equals

$$\begin{aligned}\binom{n}{a}\cdot\binom{n-a}{b} &= \frac{n!}{a!\cdot(n-a)!}\cdot\frac{(n-a)!}{b!\cdot(n-a-b)!}\\ &= \frac{n!}{a!\cdot b!\cdot(n-a-b)!}\\ &= \frac{n!}{a!\cdot b!\cdot c!}.\end{aligned}$$

More generally, suppose that we expand $(x_1+x_2+\cdots+x_k)^n$. In the result, we will have a sum of term of the form $x_1^{a_1}x_2^{a_2}\cdots x_k^{a_k}$ where the $a_i\in\mathbb{N}$ and $a_1+a_2+\cdots+a_k=n$. To determine the coefficient of such a term, we need only determine the number of sequences of x_i of length n such that there are exactly a_1 many x_1's, exactly a_2 many x_2's, $\dots$, and exactly a_k many x_k's. Following the above template, the number of such sequences equals

$$\begin{aligned}&\binom{n}{a_1}\cdot\binom{n-a_1}{a_2}\cdot\binom{n-a_1-a_2}{a_3}\cdots\binom{n-a_1-a_2-\cdots-a_{k-2}}{a_{k-1}}\cdot\binom{n-a_1-a_2-\cdots-a_{k-1}}{a_k}\\ &\quad=\binom{n}{a_1}\cdot\binom{n-a_1}{a_2}\cdot\binom{n-a_1-a_2}{a_3}\cdots\binom{n-a_1-a_2-\cdots-a_{k-2}}{a_{k-1}}\cdot\binom{a_k}{a_k}\\ &\quad=\frac{n!}{a_1!\cdot(n-a_1)!}\cdot\frac{(n-a_1)!}{a_2!\cdot(n-a_1-a_2)!}\cdot\frac{(n-a_1-a_2)!}{a_3!\cdot(n-a_1-a_2-a_3)!}\cdots\frac{(n-a_1-a_2-\cdots-a_{k-1})!}{a_{k-1}!\cdot(n-a_1-a_2-\cdots-a_{k-2}-a_{k-1})!}\\ &\quad=\frac{n!}{a_1!\cdot a_2!\cdot a_3!\cdots a_{k-1}!\cdot(n-a_1-a_2-\cdots-a_{k-2}-a_{k-1})!}\\ &\quad=\frac{n!}{a_1!\cdot a_2!\cdot a_3!\cdots a_{k-1}!\cdot a_k!}\end{aligned}$$

Notice that this is just like our problem with anagrams of MISSISSIPPI. Instead of doing the above count, we could have treated all the x_1 as different (and x_2 as different, etc.), rearranged them in $n!$ many ways, and then divided by the overcount from the permuting the x_i within themselves in $a_1!$ ways, the x_2 within themselves in $a_2!$ many ways, etc.

Definition 5.2.9. *If $n,a_1,a_2,\dots,a_k\in\mathbb{N}$ and $a_1+a_2+\cdots+a_k=n$, we define*

$$\binom{n}{a_1,a_2,\dots,a_k}=\frac{n!}{a_1!\cdot a_2!\cdots a_k!}$$

We call this a multinomial coefficient*.*

The above argument proves the generalization of the Binomial Theorem:

Theorem 5.2.10 (Multinomial Theorem)**.** *For all $n,k\in\mathbb{N}^+$, we have*

$$(x_1+x_2+\cdots+x_k)^n=\sum\binom{n}{a_1,a_2,\dots,a_k}x_1^{a_1}x_2^{a_2}\cdots x_k^{a_k}$$

where the sum is taken over all k-tuples of nonnegative integers $(a_1,a_2,\dots,a_k)$ such that $a_1+a_2+\cdots+a_k=n$.

5.3 Compositions and Partitions

Compositions

There are six different M&M colors: Red, Yellow, Blue, Green, Orange, Brown. Suppose that we want to pick out 13 total M&M's. How ways can you do it, if all that matters is how many of each color we take?

Notice that we can model this as follows: if we let a_i be the number that you choose with color i, then we need $a_1 + a_2 + a_3 + a_4 + a_5 + a_6 = 13$.

Definition 5.3.1. *Let $n, k \in \mathbb{N}$. A sequence of nonnegative integers $(a_1, a_2, \ldots, a_k)$ such that $a_1 + a_2 + \cdots + a_k = n$ is called a* weak composition *of n into k parts. If all the a_i are positive, then it is called a* composition *of n into k parts.*

For example $(1, 3, 5, 3)$ is a composition of 12 into 4 parts and $(2, 0, 5, 1, 0, 0)$ is a weak composition of 8 into 6 parts.

One can view the number of weak compositions of n into k parts as the number of ways to distribute n identical balls into k distinct boxes. In this interpretation, the value a_i is the number of balls that we put into box i. We are treating the balls as identical because all that matters are the number of balls in each box, but the boxes are distinct because $(5, 2, 1)$ is different than $(2, 5, 1)$.

We can also view these another way. Recall that a k-permutation of n distinct objects is a way to pick out k of those objects, where order matters and repetition is not allowed. Also, a k-combination of n distinct objects is a way to pick out k of those objects, where order does not matter and repetition is not allowed. A different way to interpret a weak composition of n into k parts is as a way to pick out n objects from k distinct objects, where order doesn't matter but repetition *is* allowed (yes, the n and k have switched, and this is incredibly annoying). The value a_i is the number of times that we pick out object i. Due to the fact that order doesn't matter but repetition is allowed, some sources think about something they call *multisets*. The idea is to allow one to write something like "$\{1, 1, 4\}$" and think about it as different from "$\{1, 4\}$", but the same as "$\{1, 4, 1\}$". Since, by definition, two sets are equal exactly when they have the same elements, we should introduce new notation rather than $\{$ and $\}$ used in sets. Instead of dealing with all of these issues, we avoid the situation entirely by writing $(2, 0, 0, 1)$ to represent that we picked the number 1 twice and the number 4 once.

The number of weak compositions of n into k parts is the number of nonnegative integer solutions to the equation
$$x_1 + x_2 + \cdots + x_k = n,$$
while the number of compositions of n into k parts is the number of positive integer solutions to the equation
$$x_1 + x_2 + \cdots + x_k = n.$$

How do we count the number of weak compositions of n into k parts? In the M&M case, think about lining them up in order of color, so red first, then yellow, etc. If we eliminate the colors from the M&M's themselves, then we only need some kind of "marker" to distinguish when we change colors. If we represent the M&M's as dots, then we can place 5 bars to denote the dividing line as to when we switch colors. Since we have 5 bars and 13 M&M's that we have to put into a line, we have 18 positions and need to choose the positions for the 5 bars. Therefore, there are $\binom{18}{5}$ many possibilities.

Proposition 5.3.2. *Let $n, k \in \mathbb{N}$. The number of weak compositions of n into k parts is*
$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}.$$

Proof. As above, there is a bijection between arrangements of n dots and $k-1$ bars into a line and weak compositions of n into k parts (the number of dots before the first bar is a_1, then number of dots between the first and second is a_2, etc.). We want to place $n + k - 1$ many objects, and we need only choose the $k - 1$ positions for the bars (or alternatively the n positions for the dots). Therefore, the number of weak compositions of n into k parts equals
$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}.$$
$\square$

Another way to visualize this is as follows: Consider the following bijection between subsets of $[n+k-1]$ of size n and weak compositions of n into k parts: Given a subset $\{a_1, a_2, \ldots, a_n\}$ of $[n+k-1]$ with $a_1 < a_2 < \cdots < a_n$, consider the multiset "$\{a_1, a_2 - 1, a_3 - 2, \ldots, a_n - (n-1)\}$" and form the corresponding weak composition. For example if $k = 5$ and $n = 3$, then $n + k - 1 = 7$ and we do the following:

$$\{1,2,3\} \mapsto \text{“}\{1,1,1\}\text{”} \mapsto (3,0,0,0,0)$$
$$\{1,3,7\} \mapsto \text{“}\{1,2,5\}\text{”} \mapsto (1,1,0,0,1)$$
$$\{3,4,6\} \mapsto \text{“}\{3,3,4\}\text{”} \mapsto (0,0,2,1,0)$$

More formally, given a subset $\{a_1, a_2, \ldots, a_n\}$ of $[n+k-1]$ with $a_1 < a_2 < \cdots < a_n$, we send it to the sequence $(b_1, b_2, \ldots, b_k)$ where b_ℓ equals the number of i such that $a_i - (i-1) = \ell$, i.e. the cardinality of the set $\{i : a_i = i + \ell - 1\}$.

Now that we've determined the number of *weak* compositions of n into k parts, we can count the number of compositions of n into k parts. The idea is that if $k \leq n$, then the number of positive integer solutions to the equation

$$x_1 + x_2 + \cdots + x_k = n$$

equals to the number of nonnegative solutions to

$$x_1 + x_2 + \cdots + x_k = n - k.$$

Corollary 5.3.3. *Let $n, k \in \mathbb{N}$ with $k \leq n$. The number of compositions of n into k parts equals*

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

Proof. First distribute one ball to each of the k boxes . We now have $n-k$ balls to put into k boxes with no restrictions, and so we want to count the number of weak compositions of $n-k$ into k parts. The answer to this is:

$$\binom{(n-k)+k-1}{k-1} = \binom{n-1}{k-1}$$

Since $(n-1)-(k-1) = n-k$, this also equals

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

More formally, given a weak composition $(a_1, a_2, \ldots, a_k)$ of $n-k$ into k parts, the sequence $(a_1+1, a_2+1, \ldots, a_k+1)$ is composition of n into k parts, and this mapping is a bijection. □

Another way to visualize the previous corollary with a direct bijection is as follows: Consider the function

$$(a_1, a_2, \ldots, a_k) \mapsto \{a_1, a_1 + a_2, \ldots, a_1 + a_2 + \cdots + a_{k-1}\}$$

from compositions of n into k parts to $(k-1)$-element subsets of $[n-1]$. For example if $n = 10$ and $k = 4$, then

$$(1,2,3,4) \mapsto \{1,3,6\}$$
$$(6,1,1,2) \mapsto \{6,7,8\}$$
$$(2,1,1,6) \mapsto \{2,3,4\}$$

Notice that since $a_i \geq 1$ for all i, we have $a_1 < a_1 + a_2 < \cdots < a_1 + a_2 + \cdots + a_{k-1}$. Now since $a_1 + a_2 + \cdots + a_k = n$ and $a_k \geq 1$, it follows that $a_1 + a_2 + \cdots + a_{k-1} \leq n-1$, and hence the set on the right is an element of $[n-1]$. Finally, one must check that this is a bijection, but I'll leave that to you (since we already have a proof of the result).

What happens if we try to count *all* compositions of a number n without specifying the number of parts? For example, we have 4 compositions of 3 given by (3), (1, 2), (2, 1), and (1, 1, 1). The compositions of 4 are (4), (1, 3), (3, 1), (2, 2), (2, 1, 1), (1, 2, 1), (1, 1, 2), and (1, 1, 1, 1), so we have 8 of those.

Theorem 5.3.4. *The number of compositions of n is 2^{n-1}.*

Proof. We give two proofs. The first is to notice that a composition of n must be a composition of n into k parts for some unique k with $1 \leq k \leq n$. Therefore, the number of compositions of n equals

$$\begin{aligned}
\sum_{k=1}^{n} \binom{n-1}{k-1} &= \binom{n-1}{0} + \binom{n-1}{1} + \binom{n-1}{2} + \cdots + \binom{n-1}{n-1} \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} \\
&= 2^{n-1}. && \text{(by Corollary 5.2.3)}
\end{aligned}$$

Alternatively, we can give a direct combinatorial proof. Write down n dots. Notice that we can not put a bar before the first dot or after the last one, and we also can not put two bar in the same place because in a composition all numbers must be positive. Therefore, a composition arises by picking a subset of the $n-1$ spaces between the dots to serve as bars (i.e. the dividers). Since there are 2^{n-1} many subsets of a set with $n-1$ many elements, it follows that there are 2^{n-1} many compositions of n. ☐

Set Partitions

Above we considered the case where the balls were identical and the boxes were distinct. Now consider the case where the balls are distinct but the boxes are identical.

Definition 5.3.5. *A* (set) partition *of a set A is a set $\{B_1, B_2, \ldots, B_k\}$ where the B_i are nonempty pairwise disjoint subsets of A with*

$$A = B_1 \cup B_2 \cup \cdots \cup B_k.$$

In this case, we call this a partition of A into k nonempty parts.

Definition 5.3.6. *Given $n, k \in \mathbb{N}^+$ with $k \leq n$, we define $S(n, k)$ to be the number of partitions of $[n]$ into k nonempty parts. The numbers $S(n, k)$ are called the* Stirling numbers of the second kind *and are sometimes denoted by:*

$$S(n,k) = \left\{ {n \atop k} \right\}.$$

We also define $S(0, 0) = 1$, $S(n, 0) = 0$ if $n \geq 1$, and $S(n, k) = 0$ if $k > n$.

For example, we have $S(3, 2) = 3$ because the following are all possible partitions of $[3] = \{1, 2, 3\}$ into 2 parts:

- $\{\{1\}, \{2, 3\}\}$
- $\{\{2\}, \{1, 3\}\}$
- $\{\{3\}, \{1, 2\}\}$

Notice that these are all of them because if we partition [3] into 2 parts, then one must have size 1 and the other have size 2, so the partition is completely determined by the choice of the the element that is in its own block (and hence there are $\binom{3}{1} = 3$ many choices).

Here are few more examples:

- If $n \geq 1$, then
$$S(n,1) = 1 = S(n,n)$$
because the only partition on $[n]$ into one part is $\{\{1,2,3,\dots,n\}\}$ and the only partition into n parts is $\{\{1\},\{2\},\dots,\{n\}\}$.

- We have $S(4,3) = \binom{4}{2} = 6$ because such a partition must have one set of size 2 and the others of size 1, so we need only choose the subset of size 2.

- More generally, for any $n \geq 2$, we have
$$S(n,n-1) = \binom{n}{2}$$
because a partition of $[n]$ into $n-1$ many blocks must have one block of size 2 and $n-2$ of size 1, so we need to pick the two unique elements for the block of size 2.

- The number $S(4,2)$ is more interesting. We can partition $\{1,2,3,4\}$ into a set of size 3 and a set of size 1, or into two sets of size 2. There are $\binom{4}{1} = 4$ ways to do the former because we need only pick the element in the set of size 1. For the latter, there are 3 possibilities:
  - $\{\{1,2\},\{3,4\}\}$
  - $\{\{1,3\},\{2,4\}\}$
  - $\{\{1,4\},\{2,3\}\}$

  Therefore, $S(4,2) = 4 + 3 = 7$.

In general, the numbers $S(n,k)$ are difficult to compute. Recall that
$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$
whenever $k,n \in \mathbb{N}^+$. We get a similar recurrence here.

Theorem 5.3.7. *For all $k,n \in \mathbb{N}^+$ with $k \leq n$, we have*
$$S(n,k) = S(n-1,k-1) + k \cdot S(n-1,k).$$
In other words, if $k \leq n$, then
$$\left\{ n \atop k \right\} = \left\{ n-1 \atop k-1 \right\} + k \cdot \left\{ n-1 \atop k \right\}.$$

Proof. A partition of $[n]$ into k parts is of one of two possible types:

- *Case 1:* The number n is in a block by itself. If we remove the block $\{n\}$, then we are left with a partition of $[n-1]$ into $k-1$ parts, so there are $S(n-1,k-1)$ many such possibilities. Notice that every partition of $[n]$ into k blocks having $\{n\}$ as one of the blocks arises in a unique way from such a partition of $[n-1]$ into $k-1$ parts. Thus, there are $S(n-1,k-1)$ many partitions of this type.

- *Case 2:* The number n is not in its own block. If we remove n from its block, we then obtain a partition of $[n-1]$ into k parts, and there are $S(n-1,k)$ many possible outcomes. Notice that each of these outcomes arise in k many ways because given a partition of $[n-1]$ into k blocks, we can add n into any of the blocks to obtain a partition of $[n]$ into k parts. Therefore, there are $k \cdot S(n-1,k)$ many partitions of this type.

It follows that $S(n,k) = S(n-1,k-1) + k \cdot S(n-1,k)$. $\square$

We now get a triangle like Pascal's triangle, but with $S(n,k) = \left\{ {n \atop k} \right\}$ in place of $C(n,k) = \binom{n}{k}$.

$\left\{ {n \atop k} \right\}$	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0
3	0	1	3	1	0	0	0	0
4	0	1	7	6	1	0	0	0
5	0	1	15	25	10	1	0	0
6	0	1	31	90	65	15	1	0
7	0	1	63	301	350	140	21	1

Given $n, k \in \mathbb{N}^+$, recall that we have the following from Proposition 5.1.6:

- The number of functions $f\colon [n] \to [k]$ equals k^n because for each $i \in [n]$, we have k possibilities for the value of $f(i)$.
- If $k < n$, then there are no injective functions $f\colon [n] \to [k]$ by the Pigeonhole Principle.
- If $k > n$, then the number of injective functions $f\colon [n] \to [k]$ equal $k(k-1)(k-2)\cdots(k-n+1) = (k)_n = \frac{k!}{(k-n)!}$ because we have k choices for the value of $f(1)$, then $k-1$ for the value of $f(2)$, $\dots$, and finally $k-(n-1)$ for the value of $f(n)$.

We now give one way to count the number of surjective functions from $[n]$ to $[k]$ in terms of Stirling numbers.

Proposition 5.3.8. *Given $n, k \in \mathbb{N}^+$, there are exactly $k! \cdot S(n,k)$ many surjective functions $f\colon [n] \to [k]$.*

Proof. If $k > n$, then there are no surjective functions $f\colon [n] \to [k]$, and $k! \cdot S(n,k) = k! \cdot 0 = 0$. Suppose then that $k \leq n$. Consider a surjective $f\colon [n] \to [k]$. For each $c \in [k]$, let $B_c = \{a \in [n] : f(a) = c\}$, i.e. B_c is the set of all elements of $[n]$ than map to c. Since f is surjective, we know that $B_c \neq \emptyset$ for all $c \in [k]$. Furthermore, since f is a function, the sets $B_1, B_2, \dots, B_k$ are pairwise disjoint, and $[n] = B_1 \cup B_2 \cup \cdots \cup B_k$. Therefore, $\{B_1, B_2, \dots, B_k\}$ is a partition of $[n]$ into k nonempty parts. Notice that each of these partitions arise in $k!$ many ways because we can reorder the B_i in terms of their outputs, i.e. if $n = 4$ and $k = 2$ then $\{\{1,4\},\{2,3\}\}$ is a partition arising from both the function

$$f(1) = 1 \quad f(2) = 2 \quad f(3) = 2 \quad f(4) = 1$$

and the function

$$f(1) = 2 \quad f(2) = 1 \quad f(3) = 1 \quad f(4) = 2.$$

In other words, every surjective function arises uniquely from a partition of $[n]$ into k nonempty parts, together with a permutation of $[k]$. Therefore, the number of surjective functions $f\colon [n] \to [k]$ equals $k! \cdot S(n,k)$. $\square$

Theorem 5.3.9. *For all $m, n \in \mathbb{N}^+$, we have*

$$m^n = \sum_{k=1}^{n} k! \cdot S(n,k) \cdot \binom{m}{k},$$

i.e.

$$m^n = \sum_{k=1}^{n} k! \cdot \left\{ {n \atop k} \right\} \cdot \binom{m}{k}.$$

Proof. The left-hand side m^n is simply the number of functions from $[n]$ to $[m]$ (by Proposition 5.1.6). The key fact is that given any function $f\colon A \to B$, if we let $C = \operatorname{range}(f)$, then we can view f as a surjective function $f\colon A \to C$. Thus, every function $f\colon [n] \to [m]$ can be viewed as a surjective function onto some nonempty subset of $[m]$. Now given a subset $X \subseteq [m]$ with $|X| = k$, we know from the previous proposition that there are $k! \cdot S(n,k)$ many surjections from $[n]$ to X. For a fixed k, there are $\binom{m}{k}$ many subsets of $[m]$ of size k, so there are $\binom{m}{k} \cdot k! \cdot S(n,k)$ many functions from $[n]$ to $[m]$ whose range has size k. Summing over all possible sizes for the range, we conclude that the number of functions from $[n]$ to $[m]$ equals

$$\sum_{k=0}^{n} \binom{m}{k} \cdot k! \cdot S(n,k)$$

Therefore,

$$m^n = \sum_{k=0}^{n} k! \cdot S(n,k) \cdot \binom{m}{k}.$$

□

In particular, we have

$$\begin{aligned} m^2 &= 1 \cdot 1 \cdot \binom{m}{1} + 2 \cdot 1 \cdot \binom{m}{2} \\ &= 1 \cdot \binom{m}{1} + 2 \cdot \binom{m}{2} \end{aligned}$$

for all $m \in \mathbb{N}$ as we learned above. We also have

$$\begin{aligned} m^3 &= 1 \cdot 1 \cdot \binom{m}{1} + 2 \cdot 3 \cdot \binom{m}{2} + 6 \cdot 1 \cdot \binom{m}{3} \\ &= 1 \cdot \binom{m}{1} + 6 \cdot \binom{m}{2} + 6 \cdot \binom{m}{3} \end{aligned}$$

and

$$\begin{aligned} m^4 &= 1 \cdot 1 \cdot \binom{m}{1} + 2 \cdot 7 \cdot \binom{m}{2} + 6 \cdot 6 \cdot \binom{m}{3} + 24 \cdot 1 \cdot \binom{m}{4} \\ &= 1 \cdot \binom{m}{1} + 14 \cdot \binom{m}{2} + 36 \cdot \binom{m}{3} + 24 \cdot \binom{m}{4} \end{aligned}$$

for all $m \in \mathbb{N}$. Using these formulas together with Proposition 5.2.7, we can now develop formulas for the sum of the first n cubes, the first n fourth powers, etc. For example, we have

$$\begin{aligned}
\sum_{m=1}^{n} m^3 &= \sum_{m=1}^{n} \left(1 \cdot \binom{m}{1} + 6 \cdot \binom{m}{2} + 6 \cdot \binom{m}{3}\right) \\
&= \sum_{m=1}^{n} \binom{m}{1} + 6 \cdot \sum_{m=1}^{n} \binom{m}{2} + 6 \cdot \sum_{m=1}^{n} \binom{m}{3} \\
&= \binom{n+1}{2} + 6 \cdot \binom{n+1}{3} + 6 \cdot \binom{n+1}{4} \\
&= \frac{(n+1)n}{2} + 6 \cdot \frac{(n+1)n(n-1)}{6} + 6 \cdot \frac{(n+1)n(n-1)(n-2)}{24} \\
&= \frac{(n+1)n}{2} + (n+1)n(n-1) + \frac{(n+1)n(n-1)(n-2)}{4} \\
&= \frac{(n+1)n}{4} \cdot (2 + 4(n-1) + (n-1)(n-2)) \\
&= \frac{(n+1)n}{4} \cdot (2 + 4n - 4 + n^2 - 3n + 2) \\
&= \frac{(n+1)n}{4} \cdot (n^2 + n) \\
&= \frac{(n+1)^2 n^2}{4} \\
&= \left(\frac{n(n+1)}{2}\right)^2
\end{aligned}$$

Therefore, we obtain the surprising result that

$$\sum_{m=1}^{n} m^3 = \left(\sum_{m=1}^{n} m\right)^2$$

for all $n \in \mathbb{N}^+$.

Definition 5.3.10. *Let $n \in \mathbb{N}$. The number of all partitions of $[n]$ into nonempty parts is denoted by $B(n)$ and is called the n^{th} Bell number. We also define $B(0) = 0$. Notice that*

$$B(n) = \sum_{k=0}^{n} S(n,k) = \sum_{k=0}^{n} \left\{ {n \atop k} \right\}$$

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

Recall than an equivalence relation on A induces a partition of A into nonempty parts through the equivalence classes. Conversely, it's not hard to show that if $\{B_1, B_2, \ldots, B_k\}$ is a partition of A with each $B_i \neq \emptyset$, then then the relation $a \sim b$ if there exists an i with $a, b \in B_i$ is an equivalence relation on A whose equivalence classes are the B_i. Therefore, $B(n)$ equals the number of equivalence relations on a set of size n.

Adding up the rows of the above table, we obtain the following values for the first few Bell numbers:

n	$B(n)$
0	1
1	1
2	2
3	5
4	15
5	52
6	203
7	877

Theorem 5.3.11. *For any $n \in \mathbb{N}$, we have*

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

Proof. We need to argue that the right-hand side counts the number of partitions of $[n+1]$. We look at the block containing $n+1$. We examine how many elements are *not* in the block containing $n+1$. If there are k such elements, then there are $\binom{n}{k}$ many ways to choose these elements (and hence choose the $n-k$ many elements of $[n]$ grouped with $n+1$) and then $B(k)$ many ways to partition them. Thus,

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

Alternatively, we can count as follows. If the block containing $n+1$ has exactly k elements, then there are $\binom{n}{k-1}$ many ways to choose the other elements in the block, and then $B(n+1-k)$ many ways to partition the rest. Thus

$$\begin{aligned} B(n+1) &= \sum_{k=1}^{n+1} \binom{n}{k-1} \cdot B(n+1-k) \\ &= \sum_{k=0}^{n} \binom{n}{k} \cdot B(n-k) \\ &= \sum_{k=0}^{n} \binom{n}{n-k} \cdot B(k) \\ &= \sum_{k=0}^{n} \binom{n}{k} \cdot B(k). \end{aligned}$$

□

Integer Partitions

We've seen that compositions correspond to ways to distribute n identical balls to k distinct boxes in such a way that each box receives at least one ball. Also, (set) partitions correspond to ways to distribute n distinct balls to k identical boxes in such a way that each box receives at least one ball. We now introduce (integer) partitions that correspond to ways to distribute n identical balls to k identical boxes in such a way that each box receives at least one ball.

Definition 5.3.12. *An* (integer) partition *of an $n \in \mathbb{N}$ into k parts is a composition $(a_1, a_2, \ldots, a_k)$ of n where $a_1 \geq a_2 \geq \cdots \geq a_k$. The number of partitions of n into k parts is denoted by $p(n,k)$. We also define $p(0,0) = 1$.*

Notice that $p(n,0) = 0$ if $n \geq 1$ and $p(n,k) = 0$ if $k > n$. We have $p(4,2) = 2$ because $(2,2)$ and $(3,1)$ are the only partitions of 4 into 2 parts. Notice that $p(7,3) = 1$ because the partitions are $(5,1,1)$, $(4,2,1)$, $(3,3,1)$, and $(3,2,2)$.

Definition 5.3.13. *The number of partitions of n (into any number of parts) is denoted by $p(n)$, so*

$$p(n) = \sum_{k=0}^{n} p(n,k).$$

In order to calculate p, we first establish a simple recurrence like we did for for $\binom{n}{k}$ and $S(n,k)$.

Theorem 5.3.14. *For all $n, k \in \mathbb{N}$ with $0 < k < n$, we have*

$$\begin{aligned} p(n,k) &= \sum_{i=1}^{k} p(n-k,i) \\ &= p(n-k,1) + p(n-k,2) + \cdots + p(n-k,k). \end{aligned}$$

Proof. Given a partition of n into k parts, if we subtract 1 from each part, then we obtain a partition of $n-k$ into some number (at most k) parts. Notice that we might have fewer parts, because any 1 in the original partition will become 0. However, since $k < n$, and we are subtracting k, at least one part will remain. Moreover, it's straightforward to check that this is a bijection (i.e. that it is injective and that every partition of $n-k$ into at most k parts arise). $\square$

Thus, we have

$$p(7,3) = p(4,1) + p(4,2) + p(4,3)$$

and

$$\begin{aligned} p(7,4) &= p(3,1) + p(3,2) + p(3,3) + p(3,4) \\ &= p(3,1) + p(3,2) + p(3,3). \end{aligned}$$

By starting with some simple values, we can use this recurrence to fill in a table of values.

*	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0
3	0	1	1	1	0	0	0	0
4	0	1	2	1	1	0	0	0
5	0	1	2	2	1	1	0	0
6	0	1	3	3	2	1	1	0
7	0	1	3	4	3	2	1	1

Adding up the rows, we obtain the following values:

n	$p(n)$
0	1
1	1
2	2
3	3
4	5
5	7
6	11
7	15

The question of how fast $p(n)$ grows is extremely interesting and subtle. It turns out that

$$p(n) \sim \frac{1}{4\sqrt{3}} \cdot \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$$

where $\exp(x) = e^x$ and $f(n) \sim g(n)$ means that

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1.$$

5.4 Inclusion-Exclusion

Recall that if A and B are any finite sets, then

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

What about three sets, i.e. if we wanted to count $|A \cup B \cup C|$? A natural guess would be that we need to subtract off the various intersection, so one might hope that $|A \cup B \cup C|$ equals

$$|A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C|.$$

Let's examine if this is correct. Notice that if $x \in A$, but $x \notin B$ and $x \notin C$, then x contributes 1 to $|A \cup B \cup C|$, and in the formula it contributes

$$1 + 0 + 0 - 0 - 0 - 0 = 1.$$

Similar arguments work if x is in only B, or x is in only C. Let's examine what happens if x is in two of the sets, say $x \in A$, $x \in B$, but $x \notin C$. Again, x contributes 1 to $|A \cup B \cup C|$, and in the formula it contributes

$$1 + 1 + 0 - 1 - 0 - 0 = 1.$$

Again, everything looks good so far. Finally, suppose that x is an element of each of A, B, and C. As usual, x contributes 1 to $|A \cup B \cup C|$, but in the formula it contributes

$$1 + 1 + 1 - 1 - 1 - 1 = 0.$$

Thus, elements that are if $A \cap B \cap C$ are not counted at all on the right-hand side. To correct this, we need to add it back in. We then claim that the correct formula is

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Working though each of the possibilities, one can check that this count is correct to matter where x lies in the Venn diagram of sets. How does this generalize? For four sets, one can show by working through all of the cases that

$$\begin{aligned}|A_1 \cup A_2 \cup A_3 \cup A_4| = |A_1| &+ |A_2| + |A_3| + |A_4| \\ &- |A_1 \cap A_2| - |A_1 \cap A_3| - |A_1 \cap A_4| - |A_2 \cap A_3| - |A_2 \cap A_4| - |A_3 \cap A_4| \\ &+ |A_1 \cap A_2 \cap A_3| + |A_1 \cap A_2 \cap A_4| + |A_1 \cap A_3 \cap A_4| + |A_2 \cap A_3 \cap A_4| \\ &- |A_1 \cap A_2 \cap A_3 \cap A_4|.\end{aligned}$$

It's extremely tedious to check all the possibilities here, so we would like a way to prove that this works in general. We'll do that below, but first we'll demonstrate how to use these formulas to count an interesting set. For our example, we will count the number of primes less than or equal to 120. Before jumping into this, we prove a few small but important facts.

Proposition 5.4.1. *Let $n \in \mathbb{N}^+$ with $n \geq 2$. If n is not prime, then there is a prime p such that $p \mid n$ and $p \leq \sqrt{n}$.*

Proof. Suppose that $n \geq 2$ is not prime. Since n is not prime, we can fix $d \in \mathbb{N}$ with $1 < d < n$ such that $d \mid n$. Fix $c \in \mathbb{Z}$ with $cd = n$. Notice that $c > 0$ because both $d > 0$ and $n > 0$, and moreover we must have $1 < c < n$ (if $c = 1$ then $d = n$, and if $c = n$ then $d = 1$). Now at least one of $c \leq \sqrt{n}$ or $d \leq \sqrt{n}$ must be true, because otherwise $n = cd > \sqrt{n} \cdot \sqrt{n} = n$. In either case, this number has a prime divisor (by Proposition 3.2.2) less than or equal to it, so by transitivity of divisibility, n has a prime divisor p with $p \leq \sqrt{n}$. □

Proposition 5.4.2. *If $a, b, c \in \mathbb{Z}$ are such that $a \mid c$, $b \mid c$, and $\gcd(a, b) = 1$, then $ab \mid c$.*

Proof. See Problem 6 on Homework 4. □

Proposition 5.4.3. *Let $a \in \mathbb{Z}$ and let $p_1, p_2, \dots, p_k$ be distinct primes. If $p_i \mid a$ for all i, then $p_1 p_2 \cdots p_k \mid a$.*

Proof. We prove the result by induction on k. Notice that if $k = 1$, then the statement is trivial. Suppose that we know the statement is true for a fixed $k \in \mathbb{N}$. Let $p_1, p_2, \dots, p_k, p_{k+1}$ be distinct primes with the property that $p_i \mid a$ for all i. By induction, we know that $p_1 p_2 \cdots p_k \mid a$. We also have that $p_{k+1} \mid a$ by assumption. Using Corollary 3.2.11, we know that $\gcd(p_1 p_2 \cdots p_k, p_{k+1}) = 1$, so Proposition 5.4.2 allows us to conclude that $p_1 p_2 \cdots p_k p_{k+1} \mid a$. This completes the induction. □

We now return to counting the number of primes in $[120]$. By Proposition 5.4.1, if $a \in [120]$ is not prime and $a \geq 2$, then a is divisible by some prime less than or equal to $\sqrt{a} \leq \sqrt{120}$. Now $\sqrt{120} < 11$, so the only primes less than or equal to $\sqrt{120}$ are 2, 3, 5, and 7. We thus let $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, and $p_4 = 7$. For each i, let A_i be the set of numbers in $[120]$ divisible by p_i. We count

$$|A_1 \cup A_2 \cup A_3 \cup A_4|,$$

which is the number of elements of $[120]$ that are divisible by at least one 2, 3, 5, or 7. We have

$$|A_1| = \frac{120}{2} = 60 \qquad |A_2| = \frac{120}{3} = 40 \qquad |A_3| = \frac{120}{5} = 24 \qquad |A_4| = \left\lfloor \frac{120}{7} \right\rfloor = 17.$$

To determine the cardinalities of intersections, we use Proposition 5.4.3. For example, the numbers divisible by both 2 and 3 are the numbers divisible by 6. Working these out, we conclude that

$$|A_1 \cap A_2| = \frac{120}{6} = 20 \qquad |A_1 \cap A_3| = \frac{120}{10} = 12 \qquad |A_1 \cap A_4| = \left\lfloor \frac{120}{14} \right\rfloor = 8$$

$$|A_2 \cap A_3| = \frac{120}{15} = 8 \qquad |A_2 \cap A_4| = \left\lfloor \frac{120}{21} \right\rfloor = 5 \qquad |A_3 \cap A_4| = \left\lfloor \frac{120}{35} \right\rfloor = 3.$$

Next we compute

$$|A_1 \cap A_2 \cap A_3| = \frac{120}{30} = 4 \qquad |A_1 \cap A_2 \cap A_4| = \left\lfloor \frac{120}{42} \right\rfloor = 2$$

$$|A_1 \cap A_3 \cap A_4| = \frac{120}{70} = 1 \qquad |A_2 \cap A_3 \cap A_4| = \left\lfloor \frac{120}{105} \right\rfloor = 1$$

and

$$|A_1 \cap A_2 \cap A_3 \cap A_4| = \left\lfloor \frac{120}{210} \right\rfloor = 0.$$

Thus

$$|A_1 \cup A_2 \cup A_3 \cup A_4| = (60 + 40 + 24 + 17) - (20 + 12 + 8 + 8 + 5 + 3) + (4 + 2 + 1 + 1) - 0 = 93.$$

By the Complement Rule, it follows that there are

$$120 - 93 = 27$$

many numbers in [120] that are not divisible by any of 2, 3, 5, or 7. All of these except 1 are prime, so this gives 26 new primes in [120]. Adding back in the primes 2, 3, 5, and 7, we see that there are 30 primes in [120].

Theorem 5.4.4 (Inclusion-Exclusion). *Let $A_1, A_2, \dots, A_n$ be finite sets. We then have*

$$\begin{aligned} |A_1 \cup A_2 \cup \dots \cup A_n| &= \sum_{S \in \mathcal{P}([n]) \setminus \{\emptyset\}} (-1)^{|S|-1} \cdot |\bigcap_{i \in S} A_i| \\ &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq [n], |S|=k} |\bigcap_{i \in S} A_i|. \end{aligned}$$

Less formally, this says that

$$|A_1 \cup A_2 \cup \dots \cup A_n| = \sum_i |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \dots.$$

Proof. Let $x \in A_1 \cup A_2 \cup \dots \cup A_n$ be arbitrary. Let $T = \{i \in [n] : x \in A_i\}$, i.e. T is the nonempty set of indices i such that $x \in A_i$. Let $k = |T|$ and notice that $k \geq 1$. We examine the number of times that x is counted on each side. On the left, x contributes 1 to the cardinality. On the right, it contributes

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \dots + (-1)^{k-1}\binom{k}{k}$$

to the sum. Now from Corollary 5.2.4, we know that

$$\binom{k}{0} - \binom{k}{1} + \binom{k}{2} - \binom{k}{3} + \dots - (-1)^{k}\binom{k}{k} = 0.$$

Hence

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \dots + (-1)^{k-1}\binom{k}{k} = \binom{k}{0} = 1$$

Therefore, every $x \in A_1 \cup A_2 \cup \dots \cup A_n$ contributes 1 to both sides. The result follows. □

We next count the number of surjections $f : [n] \to [k]$. Of course we know that the answer is $k! \cdot S(n, k)$ from Proposition 5.3.8, but we count it in a different way using Inclusion-Exclusion (from which we will be able to derive a formula for $S(n, k)$). We first illustrate the general argument in the special case where $n = 7$ and $k = 4$, i.e. we count the number of surjections $f : [7] \to [4]$. The idea is to count the complement. We know that there are 4^7 many total functions $f : [7] \to [4]$, so we count the number of functions that are *not* surjective. Now a function can fail to be a surjective by missing 1, missing 2, missing 3, or missing 4. Thus, given $i \in [4]$, we let A_i be the set of functions $f : [7] \to [4]$ such that $i \notin \text{range}(f)$. Then the set of functions $f : [7] \to [4]$ that are not surjective equals $A_1 \cup A_2 \cup A_3 \cup A_4$. Now we know that:

$$\begin{aligned} |A_1 \cup A_2 \cup A_3 \cup A_4| = |A_1| &+ |A_2| + |A_3| + |A_4| \\ &- |A_1 \cap A_2| - |A_1 \cap A_3| - |A_1 \cap A_4| - |A_2 \cap A_3| - |A_2 \cap A_4| - |A_3 \cap A_4| \\ &+ |A_1 \cap A_2 \cap A_3| + |A_1 \cap A_2 \cap A_4| + |A_1 \cap A_3 \cap A_4| + |A_2 \cap A_3 \cap A_4| \\ &- |A_1 \cap A_2 \cap A_3 \cap A_4|. \end{aligned}$$

To count $|A_1|$, we need to count the number of functions $f\colon [7] \to [4]$ such that $1 \notin \text{range}(f)$. This is just the number of functions $f\colon [7] \to \{2,3,4\}$, which equals 3^7. Similarly, $|A_2| = |A_3| = |A_4| = 3^7$. To count $|A_1 \cap A_2|$, we just need to count the number of functions $f\colon [7] \to [4]$ such that $1, 2 \notin \text{range}(f)$. This is just the number of functions $f\colon [7] \to \{3,4\}$, which equals 2^7. Following through on this, we conclude that

$$\begin{aligned} |A_1 \cup A_2 \cup A_3 \cup A_4| = 3^7 &+ 3^7 + 3^7 + 3^7 \\ &- 2^7 - 2^7 - 2^7 - 2^7 - 2^7 - 2^7 \\ &+ 1^7 + 1^7 + 1^7 + 1^7 + 1^7 \\ &- 0, \end{aligned}$$

so

$$|A_1 \cup A_2 \cup A_3 \cup A_4| = 4 \cdot 3^7 - 6 \cdot 2^7 + 4 \cdot 1^7.$$

Notice that coefficients are $\binom{4}{1} = 4$, $\binom{4}{2} = 6$, and $\binom{4}{3} = 4$ because $\binom{4}{m}$ is the number of ways to pick out m elements from $[4]$. It follows that the number of surjective functions $f\colon [7] \to [4]$ equals

$$4^7 - (4 \cdot 3^7 - 6 \cdot 2^7 + 4 \cdot 1^7) = 4^7 - 4 \cdot 3^7 + 6 \cdot 2^7 - 4 \cdot 1^7 = 8,400.$$

We now generalize this argument.

Theorem 5.4.5. *Let $n, k \in \mathbb{N}^+$ with $k \leq n$. The number of surjections $f\colon [n] \to [k]$ is*

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

Proof. The total number of functions $f\colon [n] \to [k]$ is k^n. For each $i \in [k]$, let A_i be the set of all functions $f\colon [n] \to [k]$ such that $i \notin \text{range}(f)$. We then have that

$$A_1 \cup A_2 \cup \cdots \cup A_k$$

is the set of all functions which are *not* surjective, and we count

$$|A_1 \cup A_2 \cup \cdots \cup A_k|$$

using Inclusion-Exclusion. Let $S \subseteq [k]$ be arbitrary, and let $m = |S|$. We then have that

$$\bigcap_{i \in S} A_i$$

is the set of functions whose range is contained in $[k]\backslash S$, so since $|[k]\backslash S| = k - m$, it follows that

$$|\bigcap_{i \in S} A_i| = (k - |S|)^n = (k-m)^n.$$

Therefore

$$\begin{aligned} |A_1 \cup A_2 \cup \cdots \cup A_k| &= \sum_{S \subseteq [k] \backslash \{\emptyset\}} (-1)^{|S|-1} \cdot |\bigcap_{i \in S} A_i| \\ &= \sum_{m=1}^{k} (-1)^{m-1} \sum_{S \subseteq [k], |S|=m} |\bigcap_{i \in S} A_i| \\ &= \sum_{m=1}^{k} (-1)^{m-1} \sum_{S \subseteq [k], |S|=m} (k - |S|)^n \\ &= \sum_{m=1}^{k} (-1)^{m-1} \binom{k}{m} (k-m)^n, \end{aligned}$$

where the last line follows from the fact that $\binom{k}{m}$ is the number of subsets of $[k]$ of cardinality m. Thus, the number of surjections $f\colon [n] \to [k]$ is

$$\begin{aligned} k^n - \sum_{m=1}^{k} (-1)^{m-1} \binom{k}{m} (k-m)^n &= k^n + \sum_{m=1}^{k} (-1)^m \binom{k}{m} (k-m)^n \\ &= \sum_{m=0}^{k} (-1)^m \binom{k}{m} (k-m)^n. \end{aligned}$$

□

Corollary 5.4.6. *Let $n, k \in \mathbb{N}^+$ with $k \leq n$. We have*

$$\begin{aligned} S(n,k) &= \frac{1}{k!} \sum_{m=0}^{k} (-1)^m \binom{k}{m} (k-m)^n \\ &= \sum_{m=0}^{k} (-1)^m \frac{(k-m)^n}{m! \cdot (k-m)!}. \end{aligned}$$

Proof. We know that the number of surjections $f\colon [n] \to [k]$ equals $k! \cdot S(n,k)$ by Proposition 5.3.8, and it also equals

$$\sum_{m=0}^{k} (-1)^m \binom{k}{m} (k-m)^n$$

by Theorem 5.4.5. Therefore,

$$k! \cdot S(n,k) = \sum_{m=0}^{k} (-1)^m \binom{k}{m} (k-m)^n,$$

and hence

$$\begin{aligned} S(n,k) &= \frac{1}{k!} \sum_{m=0}^{k} (-1)^m \binom{k}{m} (k-m)^n \\ &= \sum_{m=0}^{k} (-1)^m \frac{(k-m)^n}{m! \cdot (k-m)!}. \end{aligned}$$

□

For example, since

$$\sum_{m=0}^{4} (-1)^m \binom{4}{m} (4-m)^7 = 8,400$$

form above, we have

$$S(7,4) = \frac{8,400}{24} = 350.$$

Definition 5.4.7. *A* derangement *of $[n]$ is a permutation $(a_1, a_2, \ldots, a_n)$ of $[n]$ such that $a_i \neq i$ for all i.*

For example, $(3,1,4,2)$ is a derangement of $[4]$, but $(3,2,4,1)$ is not (because $a_2 = 2$).

Theorem 5.4.8. *Let $n \in \mathbb{N}^+$. The number of derangements of $[n]$ is*

$$n! \cdot \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

Proof. We know that there are $n!$ many permutations of $[n]$. For each $i \in [n]$, let A_i be the set of all permutations $(a_1, a_2, \ldots, a_n)$ of $[n]$ such that $a_i = i$. We then have that

$$A_1 \cup A_2 \cup \cdots \cup A_n$$

is the set of all functions which are *not* derangements. We count

$$|A_1 \cup A_2 \cup \cdots \cup A_n|$$

using Inclusion-Exclusion. Let $S \subseteq [n]$ be arbitrary, and let $k = |S|$. We then have that

$$\bigcap_{i \in S} A_i$$

is the set of permutations of $[n]$ such that $a_i = i$ for all $i \in S$. To count this, notice that k of the elements are determined, and the remaining $n-k$ elements can be permuted in the remaining $n-k$ spots arbitrarily, so

$$|\bigcap_{i \in S} A_i| = (n - |S|)! = (n-k)!.$$

We then have

$$\begin{aligned}
|A_1 \cup A_2 \cup \cdots \cup A_n| &= \sum_{S \subseteq [n] \setminus \{\emptyset\}} (-1)^{|S|-1} \cdot |\bigcap_{i \in S} A_i| \\
&= \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq [n], |S|=k} |\bigcap_{i \in S} A_i| \\
&= \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq [n], |S|=k} (n-k)! \\
&= \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} (n-k)! \\
&= \sum_{k=1}^{n} (-1)^{k-1} \frac{n!}{k!} \\
&= n! \cdot \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k!}.
\end{aligned}$$

Thus, the number of derangements of $[n]$ is

$$n! - n! \cdot \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k!} = n! \cdot \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

□

Notice that the fraction of permutations that are derangements equals

$$\begin{aligned}
\sum_{k=0}^{n} \frac{(-1)^k}{k!} &= 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} \\
&= \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}.
\end{aligned}$$

For example, when $n = 6$, we have
$$\frac{1}{2} - \frac{1}{6} + \frac{1}{24} - \frac{1}{120} + \frac{1}{720} = \frac{53}{144} \approx .36806,$$
so approximately 36.8% of the permutations are derangements. When $n = 7$, we have
$$\frac{53}{144} - \frac{1}{5040} = \frac{1854}{5040} = \frac{103}{280} \approx .36786,$$
so again about 36.8% of the permutations are derangements. Now if you've seen infinite series, then you know that
$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} - \dots$$
for all $x \in \mathbb{R}$. In particular, when $x = -1$, we have
$$\begin{aligned} e^{-1} &= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \\ &= 1 - (-1) + \frac{(-1)^2}{2!} - \frac{(-1)^3}{3!} + \frac{(-1)^4}{4!} - \dots \\ &= \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots. \end{aligned}$$
Therefore, as n gets large, the percentage of permutations of $[n]$ that are derangements approaches the number
$$1/e \approx .36788.$$

5.5 Permutations

Recall that given a finite set A with $|A| = n$, we defined a permutation of A to be an element of A^n without repeated elements. Consider the case where $A = [n]$. In this situation, we can view a permutation of A differently. Instead of thinking about the finite sequence $(a_1, a_2, \dots, a_n)$, we can think about the function $\sigma\colon [n] \to [n]$ defined by letting $\sigma(i) = a_i$ for all i. For example, if $n = 6$, then instead of writing the permutation $(5, 6, 3, 1, 4, 2)$ as an element of $[n]^n$ without any repetition, we can think about the function $\sigma\colon [6] \to [6]$ defined by:

- $\sigma(1) = 5$.
- $\sigma(2) = 6$.
- $\sigma(3) = 3$.
- $\sigma(4) = 1$.
- $\sigma(5) = 4$.
- $\sigma(6) = 2$.

Notice that since a permutation of $[n]$ does not have any repeated elements, it follows that every element of $[n]$ appears at most once as an output of σ, i.e. that $\sigma\colon [n] \to [n]$ is injective. However, since the domain and codomain are the same finite set, we know that any σ must then be bijective. Conversely, given a bijection $\sigma\colon [n] \to [n]$, the sequence $(\sigma(1), \sigma(2), \dots, \sigma(n))$ is a permutation of $[n]$. In other words, permutations of $[n]$ and bijections from $[n]$ to $[n]$ are really the same thing.

Rather than list out the values of the function as we did in our example above, we can instead write out the values in a table. For example, for our σ above, we can write it as:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 3 & 1 & 4 & 2 \end{pmatrix}.$$

In this table, entries on the top row are input values and corresponding entries on the bottom row are the output values. Notice that bottom row is simply our original sequence. We call $(5, 6, 3, 1, 4, 2)$ (or 563142 if we want to be even more compact) the *one-line notation* of σ and we call the above table the *two-line notation* of σ.

At this point, you may wonder why we care about viewing permutations as functions. The primary answer is that functions can be *composed*. Recall that the composition of two bijections is a bijection by Proposition 4.1.4, so the composition of two permutations of $[n]$ is again a permutation of $[n]$. For example, consider the following two permutations of $[6]$:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 3 & 1 & 4 & 2 \end{pmatrix} \qquad \text{and} \qquad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 6 & 2 & 4 \end{pmatrix}.$$

Let's compute $\sigma \circ \tau$. Remember that function composition happens from right to left. That is, the composition $\sigma \circ \tau$ is obtained by performing τ first and following after by performing σ. For example, we have

$$(\sigma \circ \tau)(2) = \sigma(\tau(2)) = \sigma(1) = 5.$$

Working through the 6 inputs, we obtain:

$$\sigma \circ \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 4 & 2 & 6 & 1 \end{pmatrix}.$$

On the other hand, we have

$$\tau \circ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 3 & 6 & 1 \end{pmatrix}.$$

Notice that $\sigma \circ \tau \neq \tau \circ \sigma$. Remember that function composition is not commutative in general!

Given a permutation σ of $[n]$, we can define $\sigma^2 = \sigma \circ \sigma$, $\sigma^3 = \sigma \circ \sigma \circ \sigma$, etc. We also define $\sigma^0 = id_{[n]}$, which matches up with the exponent rules. Notice that since function composition is associative by Proposition 1.4.5, we do not need to insert parentheses in things like σ^3 because we know that $\sigma \circ (\sigma \circ \sigma) = (\sigma \circ \sigma) \circ \sigma$. We now show that if we start with any $i \in [n]$, and repeatedly apply σ in this way, then we eventually cycle back around to i.

Proposition 5.5.1. *Let $\sigma \colon [n] \to [n]$ be a permutation and let $i \in [n]$. There exists $k \in \mathbb{N}^+$ with $1 \leq k \leq n$ such that $\sigma^k(i) = i$. Moreover, if k is the least positive integer with $\sigma^k(i) = i$, then the numbers*

$$i \quad \sigma(i) \quad \sigma^2(i) \quad \sigma^3(i) \quad \dots \quad \sigma^{k-1}(i)$$

are distinct.

Proof. We first show that there exists $k \in \mathbb{N}^+$ with $1 \leq k \leq n$ such that $\sigma^k(i) = i$. Consider the first $n + 1$ many numbers that we obtain by staring with i and iterating σ:

$$i \quad \sigma(i) \quad \sigma^2(i) \quad \sigma^3(i) \quad \dots \quad \sigma^n(i).$$

Since we have a list of $n+1$ numbers, and only n possible values for those numbers, there must exist $\ell, m \in \mathbb{N}$ with $0 \leq \ell < m \leq n$ and $\sigma^\ell(i) = \sigma^m(i)$ by the Pigeonhole Principle. Since $\sigma^m(i) = \sigma^\ell(\sigma^{m-\ell}(i))$, it follows that $\sigma^\ell(\sigma^{m-\ell}(i)) = \sigma^\ell(i)$. Now using the fact that σ^ℓ is injective (because it is a permutation as mentioned above), it follows that $\sigma^{m-\ell}(i) = i$. Since $1 \leq m - \ell \leq n$, we have shown the existence of a $k \in \mathbb{N}^+$ with $\sigma^k(i) = i$.

Suppose now that k is the least positive integer with $\sigma^k(i) = i$ (such a k exists by well-ordering). Assume that there is a repeat in the list:

$$i \quad \sigma(i) \quad \sigma^2(i) \quad \sigma^3(i) \quad \dots \quad \sigma^{k-1}(i).$$

We may then fix $0 \leq \ell < m \leq k$ with $\sigma^\ell(i) = \sigma^m(i)$. As above, this implies that $\sigma^{m-\ell}(i) = i$. Since $0 < m - \ell < k$, this would contradict the minimality of k. Therefore, we must have that the above values are distinct. □

With this proposition in mind, we now develop a new notation to represent permutations called *cycle notation*. The basic idea is to take an element of $[n]$ and follow its path through σ. For example, let's work with our permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 3 & 1 & 4 & 2 \end{pmatrix}.$$

We begin by starting with 1, and notice that $\sigma(1) = 5$. Now instead of moving on to deal with 2, let's continue this thread and determine the value $\sigma(5)$. Looking above, we see that $\sigma(5) = 4$. If we continue on this path to investigate 4, we see that $\sigma(4) = 1$, and we have found a "cycle" $1 \to 5 \to 4 \to 1$ hidden inside σ. We will denote this cycle with the notation $(1\ 5\ 4)$. Now that those numbers are taken care of, we start again with the smallest number not yet claimed, which in this case is 2. We have $\sigma(2) = 6$ and following up gives $\sigma(6) = 2$. Thus, we have found the cycle $2 \to 6 \to 2$ and we denote this by $(2\ 6)$. We have now claimed all numbers other than 3, and when we investigate 3 we see that $\sigma(3) = 3$, so we form the sad lonely cycle (3). Putting this all together, we write σ in cycle notation as

$$\sigma = (1\ 5\ 4)(2\ 6)(3).$$

Notice that Proposition 5.5.1 justifies why we never get "stuck" when trying to build these cycles. When we start with 1 and follow the path, we can not repeat a number before coming back to 1. For example, we will never see $1 \to 3 \to 6 \to 2 \to 6$ because then the purported permutation must send both 3 and 2 to 6, which would violate the fact that the purported permutation is injective. Also, if we finish a few cycles and start up a new one, then it is not possible that our new cycle has any elements in common with previous ones. For example, if we already have the cycle $1 \to 3 \to 2 \to 1$ and we start with 4, we can't find $4 \to 5 \to 3$ because then both 1 and 5 would map to 3.

Our conclusion is that this process of writing down a permutation in cycle notation never gets stuck and results in writing the given permutation as a product of disjoint cycles. Working through the same process with the permutation

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 6 & 2 & 4 \end{pmatrix},$$

we see that in cycle notation we have

$$\tau = (1\ 3\ 5\ 2)(4\ 6).$$

Now we can determine $\sigma \circ \tau$ in cycle notation directly from the cycle notations of σ and τ. For example, suppose we want to calculate the following:

$$(1\ 2\ 4)(3\ 6)(5) \circ (1\ 6\ 2)(3\ 5\ 4).$$

We want to determine the cycle notation of the resulting function, so we first need to determine where it sends 1. Again, function composition happens from right to left. Looking at the function represented on the right, we see the cycle containing 1 is $(1\ 6\ 2)$, so the right function sends 1 to 6. We then go to the function on the left and see where it sends 6 The cycle containing 6 there is (3 6), so it takes 6 and sends it to 3. Thus, the composition sends 1 to 3. Thus, our result starts out as

$$(1\ 3.$$

Now we need to see what happens to 3. The function on the right sends 3 to 5, and the function on the left takes 5 and leave it alone, so we have

$$(1\ 3\ 5.$$

When we move on to see what happens to 5, we notice that the right function sends it to 4 and then the left function takes 4 to 1. Since 1 is the first element the cycle we started, we now close the loop and have

$$(1\ 3\ 5).$$

We now pick up the least element not in the cycle and continue. Working it out, we end with:

$$(1\ 2\ 4)(3\ 6)(5) \circ (1\ 6\ 2)(3\ 5\ 4) = (1\ 3\ 5)(2)(4\ 6).$$

Notice that cycle notation is not unique. For example, if $n = 4$, then (1 2 3 4) and (2 3 4 1) both represent the same function, namely the function that sends 1 to 2, 2 to 3, 3 to 4, and 4 to 1. In general, we can always "cyclically shift" a cycle without changing the actual function. Also, notice that $(1\ 2)(3\ 4) = (3\ 4)(1\ 2)$, so we can also swap the ordering of the disjoint cycles.

Let's examine the possible cycle types for permutations of [4], along with the number of permutations of each type.

- One 4-cycle, such as (1 2 3 4): There are two ways to count the number of 4-cycles. One approach is to list the elements of [4] in order in 4! ways, but realize that we are over counting because we can cyclically shift each result in 4 ways to arrive at the same permutation. Thus, there are $\frac{4!}{4} = 6$ many 4-cycles. Alternatively, we can say that any 4-cycle can be shifted uniquely to put the 1 first, at which point we have $3! = 6$ many ways to arrange the three numbers after it.

- One 3-cycle and one 1-cycle, such as (1 2 3)(4): There are $4 \cdot 2 = 8$ many such permutations because we need to choose the unique element that is in the 1-cycle in 4 possible ways, and then choose the 3-cycle in $\frac{3!}{3} = 2$ ways as in the argument for 4-cycles.

- Two 2-cycles, such as (1 2)(3 4): This one is a bit tricky. We can pick two element to go in one of the cycles in $\binom{4}{2} = 6$ many ways, and once we pick this the other cycle is completely determined. However, notice that we count each of these permutations twice with this method, because if we pick $\{1, 2\}$ then we are describing the permutation (1 2)(3 4), while if we pick $\{3, 4\}$, then we are describing the permutation $(3\ 4)(1\ 2) = (1\ 2)(3\ 4)$ as well. In other words, we can't pick the "first" 2-cycle because we can list the cycles in either order. Therefore, we need to divide by 2 to handle the overcount, and so there are 3 possibilities here. Alternatively, one can notice that such a permutation is completely determined by the element that is in the cycle with 1, and we have 3 choices.

- One 2-cycle and two 1-cycles, such as (1 2)(3)(4): In this case, we need only pick the two elements of the 2-cycle (noting that order does not matter), and there are $\binom{4}{2} = 6$ many possibilities.

- Four 1-cycles, such as (1)(2)(3)(4): There is only 1 of these.

Notice that

$$6 + 8 + 3 + 6 + 1 = 24,$$

as we should expect because there are $4! = 24$ many permutations of [4].

Definition 5.5.2. *Let $k, n \in \mathbb{N}$ with $k \leq n$. The number of permutations of $[n]$ with exactly k total cycles is denoted by $c(n,k)$ and is called the* signless (or unsigned) Stirling numbers of the first kind. *Alternatively, these numbers are sometimes denoted by:*

$$c(n,k) = \begin{bmatrix} n \\ k \end{bmatrix}.$$

We also define $c(0,0) = 1$, $c(n,0) = 0$ if $n \geq 1$, and $c(n,k) = 0$ if $k > n$.

For example, our above calculations show the following:

- $c(4,1) = 6$.
- $c(4,2) = 8 + 3 = 11$.
- $c(4,3) = 6$.
- $c(4,4) = 1$.

In general, we have the following values:

- $c(n,n) = 1$ for all $n \in \mathbb{N}^+$ because the only permutation of $[n]$ with n many cycles is the one where all elements of $[n]$ are fixed.
- $c(n,1) = (n-1)!$ for all $n \in \mathbb{N}^+$ because a permutation of $[n]$ with only 1 cycles must be a cycle of length n, and we can count this by looking at all $n!$ many ways to list the elements, and then divide by n for the n many cyclic shifts. Alternatively, we can place 1 at the front of the cycle, and then order the other $n-1$ elements in all possible $(n-1)!$ many ways afterwards.
- $c(n, n-1) = \binom{n}{2}$ (which also equals $S(n, n-1)$) for all $n \geq 2$. To see this, simply notice that a permutation of $[n]$ has exactly $n-1$ many cycles if and only if it consists of $n-2$ many 1-cycles and 2-cycles. Such a permutation is completely determined by the 2 elements in the 2-cycle.

Although we were able to directly calculate $c(4,k)$ for each k, it becomes more difficult to compute values like $c(9,3)$ because such a permutation may have three 3-cycles, or one 7-cycle and two 1-cycles, or a 5-cycle and two 2-cycles, or a 2-cycle, 3-cycle, and 4-cycle, etc. Rather than attempting to calculate these values directly by looking at all possible cases, we now develop a recurrence similar to the one for the binomial coefficients and Stirling numbers of the second kind.

Theorem 5.5.3. *Let $k, n \in \mathbb{N}^+$ with $k \leq n$. We have*

$$c(n,k) = c(n-1, k-1) + (n-1) \cdot c(n-1, k).$$

In other words, if $k \leq n$, then

$$\begin{bmatrix} n \\ k \end{bmatrix} = \begin{bmatrix} n-1 \\ k-1 \end{bmatrix} + (n-1) \cdot \begin{bmatrix} n-1 \\ k \end{bmatrix}.$$

Proof. We need to show that $c(n-1, k-1) + (n-1) \cdot c(n-1, k)$ counts the number of permutations of $[n]$ with exactly k cycles. We do this by considering two cases.

- Consider those permutations of $[n]$ with exactly k cycles in which n forms a 1-cycle by itself, i.e. where n is a fixed point of the permutation. Since n forms its own cycle, if we remove it, then the rest of permutation must be a permutation of $[n-1]$ with exactly $k-1$ cycles. Furthermore, every permutation of $[n-1]$ with exactly $k-1$ cycles arises uniquely in this way. Therefore, the number of permutations of $[n]$ with exactly k cycles in which n forms a 1-cycle by itself is $c(n-1, k-1)$.

- Now consider those permutations of $[n]$ with exactly k cycles in which n does not form a 1-cycle by itself, i.e. where n is not a fixed point of the permutation. If we simply delete n from the cycle notation, we obtain a permutation of $[n-1]$ with exactly k cycles. The key fact is that that every permutation of $[n-1]$ into exactly k cycles arises in $n-1$ ways from this process, because given a permutation of $[n-1]$ into exactly k cycles, we can insert n into the permutation after any of the numbers in cycle notation. Therefore, the number of permutations of $[n]$ with exactly k cycles in which n does not form a 1-cycle by itself equals $(n-1)\cdot c(n-1,k)$.

Since we have broken up the set of all permutations of $[n]$ with exactly k cycles into the disjoint union of two sets, it follows that $c(n,k) = c(n-1,k-1) + (n-1)\cdot c(n-1,k)$. □

As an illustration of the second part of the proof, consider the case where $n = 8$ and $k = 4$, and we have the permutation

$$(1\ 3\ 4)(2\ 8)(5\ 7)(6).$$

By deleting 8 we arrive at the permutation

$$(1\ 3\ 4)(2)(5\ 7)(6).$$

Notice that we can also arrive at this latter permutation by deleting 8 from

$$(1\ 3\ 8\ 4)(2)(5\ 7)(6).$$

In fact, there are a total of 7 ways to arrive at the permutation

$$(1\ 3\ 4)(2)(5\ 7)(6)$$

by starting with a permutation of [8] where 8 is not in its own cycle.

Using this recurrence, we can compute the following table of values:

$c(n,k)$	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0
3	0	2	3	1	0	0	0	0
4	0	6	11	6	1	0	0	0
5	0	24	50	35	10	1	0	0
6	0	120	274	225	85	15	1	0
7	0	720	1764	1624	735	175	21	1

The recurrence does indeed allow us to compute the values of $c(n,k)$ quickly, but we need more work to compute the permutations of a certain "cycle structure". For example, suppose that we want to count how many permutations of [20] consist of four 2-cycles and three 4-cycles. Notice that this value will occur as one summand in the value $c(20,7)$ (as will those permutations consisting of one 14-cycle and six 1-cycles, etc.). To count this, we think as follows. Arrange the 20 elements of [20] in sequence without repetition, and build a permutation from it as follows: Put the first two elements in a 2-cycle, then the 3^{rd} and 4^{th} elements in a 2-cycle, as well as the 5^{th} and 6^{th}, and 7^{th} and 8^{th}. Next, put the the 9^{th} through 12^{th} elements in a 4-cycle, and then the 13^{th} through 16^{th}, and 17^{th} through 20^{th} into four cycles as well. For example, if we write out our 20 numbers as

$$5\ \ 19\ \ 3\ \ 11\ \ 16\ \ 17\ \ 1\ \ 9\ \ 7\ \ 12\ \ 2\ \ 4\ \ 10\ \ 14\ \ 18\ \ 8\ \ 13\ \ 6\ \ 15\ \ 20,$$

then we view this as representing the permutation

$$(5\ \ 19)(3\ \ 11)(16\ \ 17)(1\ \ 9)(7\ \ 12\ \ 2\ \ 4)(10\ \ 14\ \ 18\ \ 8)(13\ \ 6\ \ 15\ \ 20).$$

Notice that every permutation of [20] with four 2-cycles and three 4-cycles can be written with the four 2-cycles in the front (because we can always reorder the cycles), so we do get every permutation we are looking for in this way. However, this is a lot of overcount in this method. Notice that in each of the four 2-cycles in front, we can swap the order of the two 2 entries without changing the permutation. Thus, we get an overcount of 2^4 with these swappings. Furthermore, for each of the three 4-cycles, we can cyclically shift them in 4 ways without changing the actual permutation, so we get an overcount of 4^3 here. Finally, notice that we can rearrange the four 2-cycles up front in 4! ways, and also rearrange the three 4-cycles in 3! ways, without affecting the underlying permutation. It follows that the number of permutations of [20] that consist of four 2-cycles and three 4-cycles equals.

$$\frac{20!}{2^4 \cdot 4! \cdot 4^3 \cdot 3!}.$$

Theorem 5.5.4. *Let $n \in \mathbb{N}$. Suppose that we have a sequence $(a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$ with $\sum_{i=1}^n i \cdot a_i = n$. The number of permutations of $[n]$ with exactly a_i many i-cycles for each i is:*

$$\frac{n!}{a_1! \cdot a_2! \cdots a_n! \cdot 1^{a_1} \cdot 2^{a_2} \cdots n^{a_n}}.$$

Proof. Write the numbers from $[n]$ in order in $n!$ many possible ways, then insert parentheses to form a_1 many 1-cycles at the front, then a_2 many 2-cycles, etc. We get all possible permutations in this way, but there is some overcount. Within each cycle, we can rotate it, so the same permutation occurs with the same cycle order in $1^{a_1} \cdot 2^{a_2} \cdots n^{a_n}$ many ways. Furthermore, we can permute the cycles of length i amongst themselves in $a_i!$ many ways. □

Definition 5.5.5. *Let σ be a permutation of $[n]$. An* inversion *of σ is an ordered pair (i, j) with $i < j$ but $\sigma(i) > \sigma(j)$. We let $Inv(\sigma)$ be the set of all inversions of σ.*

For example, consider the following permutations in one-line notation:

$$\sigma = 312546 \qquad \tau = 315246 \qquad \pi = 342516.$$

In two-line notation, these are:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 2 & 5 & 4 & 6 \end{pmatrix} \qquad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 2 & 4 & 6 \end{pmatrix} \qquad \pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 2 & 5 & 1 & 6 \end{pmatrix},$$

while in cycle notation, these are

$$\sigma = (1\ 3\ 2)(4\ 5)(6) \qquad \tau = (1\ 3\ 5\ 4\ 2)(6) \qquad \pi = (1\ 3\ 2\ 4\ 5)(6).$$

Notice that τ and π are obtained by swapping just two elements in the one-line notation, i.e. by swapping two elements in the bottom row of the two-line notation. In terms of functions, τ and π are obtained by composing σ with a permutation consisting of one 2-cycle and four 1-cycles: we have $\tau = \sigma \circ (3\ 4)(1)(2)(5)(6)$ and $\pi = \sigma \circ (2\ 5)(1)(3)(4)(6)$.

We now examine the inversions in each of these permutations. Notice that it is typically easier to determine these in the first representations rather than in cycle notation:

$$\begin{aligned} Inv(\sigma) &= \{(1,2),(1,3),(4,5)\} \\ Inv(\tau) &= \{(1,2),(1,4),(3,4),(3,5)\} \\ Inv(\pi) &= \{(1,3),(1,5),(2,3),(2,5),(3,5),(4,5)\}. \end{aligned}$$

From this example, it may seem puzzling to see how the inversions are related. However, there is something quite interesting that is happening. Let's examine the relationship between $Inv(\sigma)$ and $Inv(\tau)$. By swapping the third and fourth positions in the second row, the inversion $(1,3)$ in σ became the inversion $(1,4)$ in τ, and the inversion $(4,5)$ in σ became the inversion $(3,5)$ in τ, so those match up. However, we added a new inversion by this swap, because although originally we had $\sigma(3) < \sigma(4)$, but the swapping made $\tau(3) > \tau(4)$. This accounts for the one additional inversion in τ. If instead we had $\sigma(3) > \sigma(4)$, then this swap would have lost an inversion. However, in either case, this example illustrates that a swapping of two adjacent numbers either increases or decreases the number of inversions by 1.

Lemma 5.5.6. *Suppose that σ is a permutation of $[n]$, and suppose τ is obtained from σ by swapping two adjacent entries in the one-line notation of σ. In other words, suppose that there is a k with $1 \leq k < n$ such that*

$$\tau(i) = \begin{cases} \sigma(i) & \text{if } i \neq k \text{ and } i \neq k+1 \\ \sigma(k+1) & \text{if } i = k \\ \sigma(k) & \text{if } i = k+1. \end{cases}$$

We then have that $|Inv(\sigma)|$ and $|Inv(\tau)|$ differ by 1.

Proof. Suppose that τ is obtained from σ by swapping the entries k and $k+1$. Notice that if $i, j \notin \{k, k+1\}$, then

$$(i,j) \in Inv(\sigma) \Longleftrightarrow (i,j) \in Inv(\tau).$$

Now given any i with $i < k$, we have

$$\begin{aligned} (i,k) \in Inv(\sigma) &\Longleftrightarrow (i,k+1) \in Inv(\tau) \\ (i,k+1) \in Inv(\sigma) &\Longleftrightarrow (i,k) \in Inv(\tau). \end{aligned}$$

Similarly, given any j with $j > k+1$, we have

$$\begin{aligned} (k,j) \in Inv(\sigma) &\Longleftrightarrow (k+1,j) \in Inv(\tau) \\ (k+1,j) \in Inv(\sigma) &\Longleftrightarrow (k,j) \in Inv(\tau). \end{aligned}$$

The final thing to notice is that

$$(k,k+1) \in Inv(\sigma) \Longleftrightarrow (k,k+1) \notin Inv(\tau),$$

because if $\sigma(k) > \sigma(k+1)$ then $\tau(k) < \tau(k+1)$, while if $\sigma(k) < \tau(k+1)$ then $\tau(k) > \tau(k+1)$. Since we have a bijection between $Inv(\sigma)\backslash\{(k,k+1)\}$ and $Inv(\tau)\backslash\{(k,k+1)\}$, while $(k,k+1)$ is exactly one of the sets $Inv(\sigma)$ and $Inv(\tau)$, it follows that $|Inv(\sigma)|$ and $|Inv(\tau)|$ differ by 1. ☐

A similar analysis is more difficult to perform on π because the swapping involved two non-adjacent numbers. As a result, elements in the middle had slightly more complicated interactions, and the above example shows that a swap of this type can sizably increase the number of inversions. Although it is possible to handle it directly, the key idea is to realize we can perform this swap through a sequence of adjacent swaps. This leads to the following result.

Corollary 5.5.7. *Suppose that σ is a permutation of $[n]$, and suppose τ is obtained from σ by swapping two entries in the one-line notation of σ. We then have that $|Inv(\sigma)|$ and $|Inv(\tau)|$ have different parities, i.e. one is even while the other is odd.*

Proof. Suppose that τ is obtained from σ by swapping positions k and ℓ, where $k < \ell$. We can assume that $\ell \geq k+2$ because otherwise $|Inv(\sigma)|$ and $|Inv(\tau)|$ differ by 1 and we are done. The key fact is that we can obtain this swap by performing an odd number of adjacent swaps. To see this, start by swapping k and

$k+1$, then $k+1$ and $k+2$, then $k+2$ and $k+3$, etc. until we end by swapping $\ell-1$ and ℓ. Notice that there are $\ell-k$ many swaps here, and we end by shifting the entries in positions $k+1$ through ℓ by one to the left, and moving the entry in position k to the entry in position ℓ. Now we swap positions $\ell-2$ and $\ell-1$, then $\ell-3$ and $\ell-2$, etc. until we end by swapping k and $k+1$. Notice that there are $\ell-k-1$ many swaps here, and in the final product we have swapped the entries in positions k and ℓ of σ and left the rest in place. We have a total of $2k+2\ell-1 = 2(k+\ell)-1$ many swaps. Now Lemma 5.5.6 says each of these adjacent swaps changes the number of inversions by 1 (either increasing or decreasing by 1), so each of these inversions changes the parity of the number inversions. Since there are an odd number of such swaps, we conclude that $|Inv(\sigma)|$ and $|Inv(\tau)|$ have different parities. $\square$

The above corollary has interesting implications. Suppose that we have a permutation σ, and think about its one-line notation. Suppose that we want to repeatedly swap two entries at a time until we arrive at the permutation $123\dots n$. This corresponds to trying to sort a list of numbers by applying a sequence of simple swaps. In general, there are many ways to carry this out, but one consequence of the above result is that the *parity* of the number of swaps in any such process will be an invariant. To see why this is, suppose that I accomplish the task using 7 swaps. Since $123\dots n$ has no inversions, and 0 is even, it follows that the initial permutation must have an odd number of inversions (because each swap changes the parity of the number of inversions). Therefore, no matter how you choose to perform the swaps, you must use an odd number of swaps as well. Although this might seem like a useless (but cute!) fact, it plays an essential role in linear algebra when working with determinants (recall that swapping the rows of a matrix switches the sign by -1), and in abstract algebra (the so-called *alternating group* is the collection of permutations that have an even number of inversions).

5.6 Relationship Between Stirling Numbers

The two types of Stirling numbers do have some commonalities (they both count the number of ways to break up $[n]$ into k nonempty parts, but the first kind involves some "order" in the cycles). However, there is a very deep connection between them that arise from looking at coefficients of certain special polynomials. In fact, Stirling numbers first arose in this context, rather than from counting problems.

Definition 5.6.1. *Given a polynomial $p(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ with $a_n \neq 0$, we define $\deg(p(x)) = n$. We leave $\deg(0)$ undefined.*

For example, we have $\deg(x^2 + 5x - 1) = 2$ and $\deg(5) = 0$. Notice that

$$\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x))$$

for all nonzero polynomials $p(x)$ and $q(x)$, and this is one of the reasons why we leave $\deg(0)$ undefined. A fundamental fact about polynomials is the following.

Fact 5.6.2. *A polynomial of degree n has at most n roots.*

You will see a proof of this result in Abstract Algebra. Essentially, the key idea is that if a is a root of a polynomial $p(x)$, then it is possible to factor out $x-a$ from $p(x)$. Notice also that this is another reason why we leave $\deg(0)$ undefined, because *every* element of $\mathbb{R}$ is a root of the zero polynomial. Furthermore, from this fact, we obtain the following important result.

Proposition 5.6.3. *Let $p(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ and $q(x) = b_nx^n + b_{n-1}x^{n-1} + \cdots + b_1x + a_0$ be polynomials of degree at most n. Suppose that $p(c) = q(c)$ for at least $n+1$ many $c \in \mathbb{R}$. We then have that $a_i = b_i$ for all i, so $p(x)$ and $q(x)$ are equal polynomials (and hence agree on all possible inputs).*

Proof. Consider the polynomial

$$p(x) - q(x) = (a_n - b_n)x^n + (a_{n-1} - b_{n-1})x^{n-1} + \cdots + (a_1 - b_1)x + (a_0 - b_0).$$

Notice that this polynomial has degree at most n but has at least $n+1$ many roots (because if $c \in \mathbb{R}$ is such that $p(c) = q(c)$, then c is a root of $p(x) - q(x)$). Since a polynomial of degree n has at most n roots, this is only possible if $p(x) - q(x)$ is the zero polynomial, i.e. if $a_i - b_i = 0$ for all i. We conclude that $a_i = b_i$ for all i. □

Since $\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x))$ for all nonzero polynomials $p(x)$ and $q(x)$, it follows that $\deg((x+1)^n) = n$ for all $n \in \mathbb{N}$. The Binomial Theorem tells us the coefficients of the resulting polynomial:

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

In other words, if we expand out

$$(x+1)(x+1)(x+1)\cdots(x+1),$$

and collect terms to form a polynomial of degree n, then the coefficient of x^k in the result is $\binom{n}{k}$. We next work to determine the coefficients of slightly more complicated polynomials.

Definition 5.6.4. *Given $n \in \mathbb{N}^+$, we define the following two polynomials:*

- $x^{\overline{n}} = x(x+1)(x+2)\cdots(x+n-1)$.
- $x^{\underline{n}} = x(x-1)(x-2)\cdots(x-n+1)$.

We also define $x^{\overline{0}} = 1 = x^{\underline{0}}$. Notice that $\deg(x^{\overline{n}}) = n = \deg(x^{\underline{n}})$ for all $n \in \mathbb{N}$.

For example, we have the following:

- $x^{\overline{0}} = 1$.
- $x^{\overline{1}} = x = 0 + x$.
- $x^{\overline{2}} = x(x+1) = 0 + x + x^2$.
- $x^{\overline{3}} = x(x+1)(x+2) = 0 + 2x + 3x^2 + x^3$.
- $x^{\overline{4}} = x(x+1)(x+2)(x+3) = 0 + 6x + 11x^2 + 6x^3 + x^4$.

We also have:

- $x^{\underline{0}} = 1$.
- $x^{\underline{1}} = x = 0 + x$.
- $x^{\underline{2}} = x(x-1) = 0 - x + x^2$.
- $x^{\underline{3}} = x(x-1)(x-2) = 0 + 2x - 3x^2 + x^3$.
- $x^{\underline{4}} = x(x-1)(x-2)(x-3) = 0 - 6x + 11x^2 - 6x^3 + x^4$.

These numbers look familiar! In the first collection of examples, it looks like we are seeing rows of Stirling numbers of the first kind. In second, we are obtaining these rows with some alternation of signs. We now go about proving these results.

Theorem 5.6.5. *For every $n \in \mathbb{N}$, we have*

$$x^{\overline{n}} = \sum_{k=0}^{n} c(n,k) \cdot x^k.$$

Proof. We prove the result by induction on $n \in \mathbb{N}$.

- *Base Cases:* We prove the result for $n = 0$ and $n = 1$ (we will need two base cases because we assume $n \geq 1$ in Theorem 5.5.3).

  - When $n = 0$, we have

$$\begin{aligned} x^{\overline{0}} &= 1 \\ &= c(0,0) \\ &= c(0,0) \cdot x^0 \\ &= \sum_{k=0}^{0} c(n,k) \cdot x^k. \end{aligned}$$

  - When $n = 1$, we have

$$\begin{aligned} x^{\overline{1}} &= 0 + 1x \\ &= c(1,0) \cdot x^0 + c(1,1) \cdot x^1 \\ &= \sum_{k=0}^{1} c(n,k) x^k. \end{aligned}$$

  Thus, the statement is true when $n = 0$ and $n = 1$.

- *Inductive Step:* Let $n \in \mathbb{N}^+$ be arbitrary and assume that the statement is true for n, i.e. assume that

$$x^{\overline{n}} = \sum_{k=0}^{n} c(n,k) \cdot x^k.$$

  Using the fact that $c(n,0) = 0 = c(n+1,0)$ and $c(n,n) = 1 = c(n+1,n+1)$, along with Theorem

5.5.3, we then have

$$\begin{aligned}
x^{\overline{n+1}} &= x^{\overline{n}}(x+n) \\
&= \left(\sum_{k=0}^{n} c(n,k)\cdot x^k\right)\cdot(x+n) \\
&= x\cdot\left(\sum_{k=0}^{n} c(n,k)\cdot x^k\right) + n\cdot\left(\sum_{k=0}^{n} c(n,k)\cdot x^k\right) \\
&= \left(\sum_{k=0}^{n} c(n,k)\cdot x^{k+1}\right) + \left(\sum_{k=0}^{n} n\cdot c(n,k)\cdot x^k\right) \\
&= \left(\sum_{k=1}^{n+1} c(n,k-1)\cdot x^k\right) + \left(\sum_{k=1}^{n} n\cdot c(n,k)\cdot x^k\right) \\
&= \left(\sum_{k=1}^{n} c(n,k-1)\cdot x^k\right) + c(n,n)\cdot x^{n+1} + \left(\sum_{k=1}^{n} n\cdot c(n,k)\cdot x^k\right) \\
&= 0\cdot x^0 + \left(\sum_{k=1}^{n}[c(n,k-1)+n\cdot c(n,k)]\cdot x^k\right) + c(n,n)\cdot x^{n+1} \\
&= c(n+1,0)\cdot x^0 + \left(\sum_{k=1}^{n} c(n+1,k)\cdot x^k\right) + c(n+1,n+1)\cdot x^{n+1} \\
&= \sum_{k=0}^{n+1} c(n+1,k)\cdot x^k.
\end{aligned}$$

Thus, the statement is true for $n+1$.

The result follows by induction. $\square$

Definition 5.6.6. *Let $k, n \in \mathbb{N}$. We define*

$$s(n,k) = (-1)^{n+k}c(n,k) = (-1)^{n+k}\begin{bmatrix} n \\ k \end{bmatrix},$$

and call $s(n,k)$ the (signed) Stirling numbers of the first kind.

Notice that $s(0,0) = 1$, $s(n,0) = 0$ if $n \geq 1$, and $s(n,k) = 0$ if $k > n$ because the same are true of $c(n,k)$ by definition.

Corollary 5.6.7. *For every $n \in \mathbb{N}^+$, we have*

$$x^{\underline{n}} = \sum_{k=0}^{n} s(n,k)\cdot x^k.$$

Proof. One can prove this by induction as in the previous theorem, but there is another more clever method. Let $n \in \mathbb{N}^+$ be arbitrary. We know from the previous theorem that

$$x(x+1)(x+2)\cdots(x+(n-1)) = \sum_{k=0}^{n} c(n,k)\cdot x^k.$$

Since this is a polynomial equality, we can plug in any real value to obtain an equality of real numbers. Thus, for any $a \in \mathbb{R}$, we can plug $-a$ into the above polynomials to conclude that

$$(-a)((-a)+1)((-a)+2)\cdots((-a)+(n-1)) = \sum_{k=0}^{n} c(n,k)\cdot(-a)^k,$$

which implies that

$$(-1)^n \cdot a(a-1)(a-2)\cdots(a-(n-1)) = \sum_{k=0}^{n}(-1)^k c(n,k)\cdot a^k.$$

Multiplying both sides by $(-1)^n$ it follows that

$$a(a-1)(a-2)\cdots(a-(n-1)) = \sum_{k=0}^{n}(-1)^{n+k} c(n,k)\cdot a^k,$$

and hence

$$a(a-1)(a-2)\cdots(a-(n-1)) = \sum_{k=0}^{n} s(n,k)\cdot a^k$$

is true for all $a \in \mathbb{R}$. Since the polynomials $x^{\underline{n}} = x(x-1)(x-2)\cdots(x-(n-1))$ and $\sum_{k=0}^{n} s(n,k)\cdot x^k$ agree for all real numbers, we may use Proposition 5.6.3 to conclude that

$$x^{\underline{n}} = \sum_{k=0}^{n} s(n,k)\cdot x^k.$$

This completes the proof. $\square$

We can interpret our two polynomial equalities in the following way. Let $n \in \mathbb{N}^+$ and consider the vector space V of all polynomials of degree at most n (as well as the zero polynomial). We know that $\{x^0, x^1, x^2, \dots, x^n\}$ is a basis for V. For each $\ell \in \mathbb{N}$ with $0 \leq \ell \leq n$, we have

$$x^{\overline{\ell}} = \sum_{k=0}^{\ell} c(\ell,k)\cdot x^k$$

and

$$x^{\underline{\ell}} = \sum_{k=0}^{\ell} s(\ell,k)\cdot x^k,$$

so the (unsigned/signed) Stirling numbers of the first kind show how to express $x^{\overline{\ell}} \in V$ and $x^{\underline{\ell}} \in V$ as linear combinations of the standard basis vectors in $\{x^0, x^1, x^2, \dots, x^n\}$. Can we reverse this process? In other words, can we express $1, x, x^2, \dots, x^n$ in terms of the vectors $\{x^{\underline{0}}, x^{\underline{1}}, x^{\underline{2}}, \dots, x^{\underline{n}}\}$? If this latter set is a basis for V, then this is indeed possible. One can show directly that $\{x^{\underline{0}}, x^{\underline{1}}, x^{\underline{2}}, \dots, x^{\underline{n}}\}$ is a linearly independent set of size $n+1$, so it must be a basis. Hence, it is at least theoretically possible. However, we can just directly prove that is possible, while also finding the coefficients, using a few previous results.

Theorem 5.6.8. *For every $n \in \mathbb{N}$, we have*

$$x^n = \sum_{k=0}^{n} S(n,k)\cdot x^{\underline{k}}.$$

Proof. When $n = 0$, we have $x^0 = 1$ and

$$\begin{aligned}\sum_{k=0}^{0} S(n,k) \cdot x^{\underline{k}} &= S(0,0) \cdot x^{\underline{0}} \\ &= 1 \cdot x^0 \\ &= 1\end{aligned}$$

as well, so the statement is true in this case. Suppose now that $n \geq 1$. Recall that Theorem 5.3.9 tells us that

$$m^n = \sum_{k=1}^{n} k! \cdot S(n,k) \cdot \binom{m}{k}$$

for all $m \in \mathbb{N}^+$ (because both sides count the number of functions from $[n]$ to $[m]$). Therefore, for any $m \in \mathbb{N}^+$, we have

$$\begin{aligned}m^n &= \sum_{k=1}^{n} k! \cdot S(n,k) \cdot \binom{m}{k} \\ &= \sum_{k=1}^{n} k! \cdot S(n,k) \cdot \frac{m!}{k! \cdot (m-k)!} \\ &= \sum_{k=1}^{n} S(n,k) \cdot \frac{m!}{(m-k)!} \\ &= \sum_{k=1}^{n} S(n,k) \cdot m(m-1)(m-2)\cdots(m-k+1) \\ &= \sum_{k=0}^{n} S(n,k) \cdot m(m-1)(m-2)\cdots(m-k+1),\end{aligned}$$

where the last line follows from the fact that $S(n,0) = 0$. Thus, the polynomial x^n and the polynomial

$$\sum_{k=0}^{n} S(n,k) \cdot x^{\underline{k}}$$

agree at every natural number m. Since these two polynomials have degree at most n and agree at infinitely many points, we may use Proposition 5.6.3 to conclude that

$$x^n = \sum_{k=0}^{n} S(n,k) \cdot x^{\underline{k}}.$$

This completes the proof. $\square$

We now know that

$$x^{\underline{\ell}} = \sum_{k=0}^{\ell} s(\ell,k) \cdot x^k$$

and

$$x^{\ell} = \sum_{k=0}^{\ell} S(\ell,k) \cdot x^{\underline{k}}$$

for all $\ell \in \mathbb{N}$. Let's return to the above setting, i.e. let $n \in \mathbb{N}^+$ and consider the vector space V of all polynomials of degree at most n (as well as the zero polynomial). Since $\{x^0, x^1, x^2, \dots, x^n\}$ is a basis for V, and we've just seen that each x^ℓ is in the span of $\{x^{\underline{0}}, x^{\underline{1}}, x^{\underline{2}}, \dots, x^{\underline{n}}\}$, it follows that $\{x^{\underline{0}}, x^{\underline{1}}, x^{\underline{2}}, \dots, x^{\underline{n}}\}$ spans V. Since this is a spanning set of $n+1$ many vectors, it follows that this set is also basis of V. Furthermore, the above equalities show that the Stirling numbers give the change of basis matrices between these two bases. Thus, if we cut off the Stirling matrices $S(n,k)$ and $s(n,k)$ at some finite point, then the matrices must be inverses of each other.

$S(n,k)$	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0
3	0	1	3	1	0	0	0	0
4	0	1	7	6	1	0	0	0
5	0	1	15	25	10	1	0	0
6	0	1	31	90	65	15	1	0
7	0	1	63	301	350	140	21	1

$s(n,k)$	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	-1	1	0	0	0	0	0
3	0	2	-3	1	0	0	0	0
4	0	-6	11	-6	1	0	0	0
5	0	24	-50	35	-10	1	0	0
6	0	-120	274	-225	85	-15	1	0
7	0	720	-1764	1624	-735	175	-21	1

Alternatively, instead of appealing to linear algebra, we can also determine this inverse relationship directly. For any $n \in \mathbb{N}$, we have

$$\begin{aligned}
x^n &= \sum_{\ell=0}^{n} S(n,\ell) \cdot x^{\underline{\ell}} \\
&= \sum_{\ell=0}^{n} [S(n,\ell) \cdot (\sum_{k=0}^{\ell} s(\ell,k) \cdot x^k)] \\
&= \sum_{\ell=0}^{n} \sum_{k=0}^{\ell} [S(n,\ell) \cdot s(\ell,k) \cdot x^k] \\
&= \sum_{\ell=0}^{n} \sum_{k=0}^{n} [S(n,\ell) \cdot s(\ell,k) \cdot x^k] \\
&= \sum_{k=0}^{n} \sum_{\ell=0}^{n} [S(n,\ell) \cdot s(\ell,k) \cdot x^k] \\
&= \sum_{k=0}^{n} [\sum_{\ell=0}^{n} S(n,\ell) \cdot s(\ell,k)] \cdot x^k.
\end{aligned}$$

Therefore

$$\sum_{\ell=0}^{n} S(n,\ell) \cdot s(\ell,k) = \begin{cases} 1 & \text{if } n = k \\ 0 & \text{if } n \neq k. \end{cases}$$

Similarly, we have

$$\sum_{\ell=0}^{n} s(n,\ell)\cdot S(\ell,k) = \begin{cases} 1 & \text{if } n=k \\ 0 & \text{if } n\neq k. \end{cases}$$

In other words, the Stirling matrices $S(n,k)$ and $s(n,k)$ are inverses of each other.

Notice that we do not even need to cut off the matrices at some point to be $n\times n$ matrices (the fact that every row is eventually zero means the the matrix products make sense even for infinite matrices). In linear algebra terminology, the sets $\{x^0, x^1, x^2, \dots\}$ and $\{x^{\underline{0}}, x^{\underline{1}}, x^{\underline{2}}, \dots\}$ are both bases for the infinite-dimensional vector space of *all* polynomials (without degree restrictions), and the matrices $S(n,k)$ and $s(n,k)$ form the change of basis matrices for these two bases.

Chapter 6

Congruences and Modular Arithmetic

6.1 Definitions and Fundamental Results

We now embark on a study of congruences. Given an $m \in \mathbb{N}^+$, there are m possible remainders that occur when we divide a general integer by m. Intuitively, congruences are a way of grouping together integers that leave the same remainder. Formally, we will define them in terms of a certain divisibility condition. Despite the fact that congruences can always be translated into divisibility relationships, it turns out that this change in perspective opens many new doors. We start with the formal definition.

Definition 6.1.1. *Let $m \in \mathbb{N}^+$. We define a relation $\equiv_m$ on $\mathbb{Z}$ by letting $a \equiv_m b$ mean that $m \mid (a-b)$. When $a \equiv_m b$, we say that a is* congruent *to b* modulo *m.*

Proposition 6.1.2. *Let $m \in \mathbb{N}^+$. The relation $\equiv_m$ is an equivalence relation on $\mathbb{Z}$.*

Proof. We need to check the three properties:

- Reflexive: Let $a \in \mathbb{Z}$. Since $a - a = 0$ and $m \mid 0$, we have that $m \mid (a-a)$, hence $a \equiv_m a$.
- Symmetric: Let $a, b \in \mathbb{Z}$ with $a \equiv_m b$. We then have that $m \mid (a-b)$. Thus $m \mid (-1)(a-b)$ by Proposition 1.5.3, which says that $m \mid (b-a)$, and so $b \equiv_m a$.
- Transitive: Let $a, b, c \in \mathbb{Z}$ with $a \equiv_m b$ and $b \equiv_m c$. We then have that $m \mid (a-b)$ and $m \mid (b-c)$. Using Proposition 1.5.3, it follows that $m \mid [(a-b)+(b-c)]$, which is to say that $m \mid (a-c)$. Therefore, $a \equiv_m c$.

Putting it all together, we conclude that $\equiv_m$ is an equivalence relation on $\mathbb{Z}$. □

By our general theory of equivalence relations, we know that $\equiv_m$ partitions $\mathbb{Z}$ into equivalence classes. We next determine the number of such equivalence classes, together with a canonical choice of representatives from the equivalence classes.

Proposition 6.1.3. *Let $m \in \mathbb{N}^+$ and let $a \in \mathbb{Z}$. There exists a unique $b \in \{0, 1, \dots, m-1\}$ such that $a \equiv_m b$. In fact, if we write $a = qm + r$ for the unique choice of $q, r \in \mathbb{Z}$ with $0 \leq r < m$, then $b = r$ is the unique such choice.*

Proof. As in the statement, fix $q, r \in \mathbb{Z}$ with $a = qm + r$ and $0 \leq r < m$. We then have $a - r = mq$, so $m \mid (a-r)$. It follows that $a \equiv_m r$, so we have proven existence.

Suppose now that $b \in \{0, 1, \dots, n-1\}$ and $a \equiv_m b$. We then have that $m \mid (a-b)$, so we may fix $k \in \mathbb{Z}$ with $mk = a - b$. This gives $a = km + b$. Since $0 \leq b < m$, we may use the uniqueness part of Theorem 2.3.1 to conclude that $k = q$ (which is unnecessary) and also that $b = r$. This proves uniqueness. □

In other words, given $m \in \mathbb{N}^+$, the proposition says that every integer is related to something in the set $\{0, 1, 2, \dots, m-1\}$, and furthermore that no two distinct elements of the set $\{0, 1, 2, \dots, m-1\}$ are related to each other. Thus, $\equiv_m$ spits up $\mathbb{Z}$ into exactly m equivalence classes. To help understand the equivalence classes, we prove the following simple result.

Proposition 6.1.4. *Let $m \in \mathbb{N}^+$. Under the equivalent relation $\equiv_m$, we have $\overline{a} = \{a + mk : k \in \mathbb{Z}\}$ for all $a \in \mathbb{Z}$.*

Proof. Let $a \in \mathbb{Z}$ be arbitrary. We give a double containment proof that $\overline{a} = \{a + mk : k \in \mathbb{Z}\}$.

- $\overline{a} \subseteq \{a + mk : k \in \mathbb{Z}\}$: Let $b \in \overline{a}$ be arbitrary. We then have that $a \equiv_m b$, so $m \mid (a - b)$. By definition, we can fix $\ell \in \mathbb{Z}$ with $a - b = m\ell$. We then have $b = a - m\ell = m(-\ell) + a$, so since $-\ell \in \mathbb{Z}$, it follows that $b \in \{a + mk : k \in \mathbb{Z}\}$.
- $\{a + mk : k \in \mathbb{Z}\} \subseteq \overline{a}$: Let $b \in \{a + mk : k \in \mathbb{Z}\}$ be arbitrary. By definition, we can fix $k \in \mathbb{Z}$ with $b = a + mk$. We then have $a - b = -mk = m(-k)$, so $m \mid (a - b)$. Therefore, $a \equiv_m b$, and hence $b \in \overline{a}$.

Putting the containments together, we conclude that $\overline{a} = \{a + mk : k \in \mathbb{Z}\}$. □

For example, if $m = 5$, then we have the following equivalence classes:

- $\overline{0} = \{\dots, -10, -5, 0, 5, 10, \dots\} = \{5k : k \in \mathbb{Z}\}$.
- $\overline{1} = \{\dots, -9, -4, 1, 6, 11, \dots\} = \{1 + 5k : k \in \mathbb{Z}\}$.
- $\overline{2} = \{\dots, -8, -3, 2, 7, 12, \dots\} = \{2 + 5k : k \in \mathbb{Z}\}$.
- $\overline{3} = \{\dots, -7, -2, 3, 8, 13, \dots\} = \{3 + 5k : k \in \mathbb{Z}\}$.
- $\overline{4} = \{\dots, -6, -1, 4, 9, 14, \dots\} = \{4 + 5k : k \in \mathbb{Z}\}$.

Notice that

$$\overline{6} = \{6 + 5k : k \in \mathbb{Z}\}$$

by Proposition 6.1.4, but since $1 \equiv_5 6$, we know from Theorem 1.3.7 that $\overline{1} = \overline{6}$. Thus, we can also write

$$\overline{6} = \{1 + 5k : k \in \mathbb{Z}\}.$$

Of course, it's also possible to just prove $\{6 + 5k : k \in \mathbb{Z}\} = \{1 + 5k : k \in \mathbb{Z}\}$ directly, but why not use our theory? Always keep in mind that $\equiv_m$ breaks up $\mathbb{Z}$ into m different equivalence classes, but we can choose infinitely many different representatives from each equivalence class.

So far, we've used the notation $a \equiv_m b$ to denote the relation because it fits in with our general infix notation for relations. However, for both historical reasons and because the subscript can be annoying, one typically uses the following notation.

Notation 6.1.5. *Given $a, b \in \mathbb{Z}$ and $m \in \mathbb{N}^+$, we write $a \equiv b \pmod{m}$ to mean that $a \equiv_m b$.*

Definition 6.1.6. *Let $m \in \mathbb{N}^+$. Given distinct $b_1, b_2, \dots, b_m \in \mathbb{Z}$, we say that $\{b_1, b_2, \dots, b_m\}$ is a* complete residue system modulo m *if for every $a \in \mathbb{Z}$, there exists a unique i with $1 \leq i \leq m$ such that $a \equiv b_i \pmod{m}$.*

In other words, a complete residue system modulo m is a way of choosing exactly one representative from each of the equivalence classes modulo m. Since we know that there are m equivalence classes, we know that every complete residue system modulo m has exactly m elements, which explains why we put m many b_i in the definition. As we saw above, $\{0, 1, 2, 3, 4\}$ is a complete residue system modulo 5. In general, Proposition 6.1.3 tells us that $\{0, 1, 2, \dots, m-1\}$ is complete residue system modulo m. This is often the most natural choice, but it is sometimes extremely useful to pick use different complete residue systems. For example, $\{-2,-1,0,1,2\}$ is also a complete residue system modulo 5, and it has an appealing symmetry that is lacking in $\{0,1,2,3,4\}$. The following simple result gives another way to check that a given set is a complete residue system.

Proposition 6.1.7. *Let $m \in \mathbb{N}^+$, and let $b_1, b_2, \dots, b_m \in \mathbb{Z}$. If $b_i \not\equiv b_j \pmod{m}$ whenever $i \neq j$, then $\{b_1, b_2, \dots, b_m\}$ is a complete residue system modulo m.*

Proof. Let $a \in \mathbb{Z}$ be arbitrary. If $a \not\equiv b_i \pmod{m}$ for all i, then that $b_1, b_2, \dots, b_m, a$ would be a list of $m+1$ numbers with the property that m does not divide the difference of any two, contrary to Proposition 4.3.2. Therefore, there must exist an i with $a \equiv b_i \pmod{m}$. Now if there existed $i \neq j$ with both $a \equiv b_i \pmod{m}$ and $a \equiv b_j \pmod{m}$, then we would have $b_i \equiv b_j \pmod{m}$ by Proposition 6.1.2, a contradiction. Therefore, we have uniqueness as well. $\square$

Recall that equivalence relations give a kind of generalization of equality. With this in mind, we would like to establish what operations we can perform on both sides of a congruence without affecting its truth. Our first result along this line is the following.

Proposition 6.1.8. *Let $m \in \mathbb{N}^+$ and let $a, b, k \in \mathbb{Z}$. We have $a \equiv b \pmod{m}$ if and only if $a + k \equiv b + k \pmod{m}$.*

Proof. Notice that $(a+k) - (b+k) = a - b$, so $m \mid (a-b)$ if and only if $m \mid [(a+k) - (b+k)]$. $\square$

When we turn to multiplication, we encounter an interesting wrinkle.

Proposition 6.1.9. *Let $m \in \mathbb{N}^+$ and let $a, b \in \mathbb{Z}$.*

1. *If $k \in \mathbb{Z}$ and $a \equiv b \pmod{m}$, then $ka \equiv kb \pmod{m}$.*
2. *If $k \in \mathbb{Z}$ is such that $ka \equiv kb \pmod{m}$ and $\gcd(k,m) = 1$, then $a \equiv b \pmod{m}$.*

Proof.

1. Suppose that $a \equiv b \pmod{m}$ and $k \in \mathbb{Z}$. We then have $m \mid (a-b)$, so $m \mid k(a-b)$ by Proposition 1.5.3. It follows that $m \mid (ka - kb)$, and therefore $ka \equiv kb \pmod{m}$.
2. Let $k \in \mathbb{Z}$, and suppose that $ka \equiv kb \pmod{m}$ and $\gcd(k,m) = 1$. We then have $m \mid (ka - kb)$, so $m \mid k(a-b)$. Since $\gcd(k,m) = 1$, we can use Proposition 3.2.5 to conclude that $m \mid (a-b)$. Therefore, $a \equiv b \pmod{m}$.

$\square$

Notice that $3 \cdot 4 \equiv 3 \cdot 2 \pmod{6}$ because $6 \mid (12-6)$, but $4 \not\equiv 2 \pmod{6}$. Therefore, we can not drop the assumption that $\gcd(k,m) = 1$ in the second part of this proposition. The fact that we can multiply both sides of a congruence by a common value, but in general can not divide both sides by a common nonzero value, is a fascinating new feature of congruences that distinguish them from our usual number systems. We will have a great deal more to say about this in time.

Although we can not naively "divide" both sides of a congruence by a common factor, we can "divide" all *three* values (including the modulus) by a common positive factor, as we now show.

Proposition 6.1.10. *Let $m, k \in \mathbb{N}^+$ and let $a, b \in \mathbb{Z}$. We have $a \equiv b \pmod{m}$ if and only if $ka \equiv kb \pmod{km}$.*

Proof. Suppose first that $a \equiv b \pmod{m}$. We then have $m \mid (a-b)$, so we can fix $c \in \mathbb{Z}$ with $mc = a-b$. Multiplying both sides by k, we have $kmc = ka - kb$, so since $c \in \mathbb{Z}$, it follows that $km \mid (ka-kb)$. Therefore, $ka \equiv kb \pmod{km}$.

Suppose conversely that $ka \equiv kb \pmod{km}$. We then have $km \mid (ka-kb)$, so we can fix $c \in \mathbb{Z}$ with $kmc = ka - kb$. Now $k \in \mathbb{N}^+$, so $k \neq 0$. Thus, we can divide both sides by k to conclude that $mc = a-b$. Since $c \in \mathbb{Z}$, it follows that $m \mid (a-b)$, so $a \equiv b \pmod{m}$. □

We can also "reduce" the modulus to a divisor.

Proposition 6.1.11. *Let $m, d \in \mathbb{N}^+$ and let $a, b \in \mathbb{Z}$. If $a \equiv b \pmod{m}$ and $d \mid m$, then $a \equiv b \pmod{d}$.*

Proof. Suppose that $a \equiv b \pmod{m}$ and that $d \mid m$. Since $d \mid m$ and $m \mid (a-b)$, we can use Proposition 1.5.2 to conclude that $d \mid (a-b)$. Therefore, $a \equiv b \pmod{d}$. □

With all of this background in hand, we now ask a more subtle question. In Proposition 6.1.8 and Proposition 6.1.9, we showed that if $a \equiv b \pmod{m}$, then we can add/multiply both sides of the congruence by the same number (although we can not necessarily "divide"). Can we generalize this to adding/multiplying both sides by congruent numbers, rather than the same number?

For example, consider the case where $m = 7$. Notice that $2 \equiv 23 \pmod{7}$ and that $10 \equiv 3 \pmod{7}$. Can we add/multiply these two congruences to arrive at another correct congruence relation? In other words, are $2 + 10 \equiv 23 + 3 \pmod{7}$ and $2 \cdot 10 \equiv 23 \cdot 3 \pmod{7}$ both true? In this simple case, you can directly check that both $12 \equiv 26 \pmod{7}$ and $20 \equiv 69 \pmod{7}$ are true. But can we do this generally? Consider:

$$\begin{aligned}\overline{2} &= \{\dots, -5, 2, 9, 16, 23, 30, \dots\} \\ \overline{3} &= \{\dots, -4, 3, 10, 17, 24, 31, \dots\}.\end{aligned}$$

Now the question before us is the following. If we pick $a, c \in \overline{2}$ and we pick $b, d \in \overline{3}$, do we get the "same answer" if we add $a + b$ as we do if we add $c + d$? Now by "same answer", we do not mean the exact same number, but only that we land in the same equivalence class. We now check that this is indeed always the case.

Proposition 6.1.12. *Suppose that $m \in \mathbb{N}^+$ and $a, b, c, d \in \mathbb{Z}$ are such that $a \equiv c \pmod{m}$ and $b \equiv d \pmod{m}$. We then have the following:*

1. *$a + b \equiv c + d \pmod{m}$.*
2. *$ab \equiv cd \pmod{m}$.*

Proof. Since $a \equiv c \pmod{m}$ and $b \equiv d \pmod{m}$, we have $m \mid (a-c)$ and $m \mid (b-d)$.

1. Notice that
$$(a+b) - (c+d) = (a-c) + (b-d).$$
Since $m \mid (a-c)$ and $m \mid (b-d)$, it follows from Proposition 1.5.3 that $m \mid [(a-c) + (b-d)]$ and so $m \mid [(a+b) - (c+d)]$. Therefore, $a + b \equiv c + d \pmod{m}$.

2. Notice that
$$\begin{aligned}ab - cd &= ab - bc + bc - cd \\ &= (a-c) \cdot b + (b-d) \cdot c.\end{aligned}$$
Since $m \mid (a-c)$ and $m \mid (b-d)$, it follows from Proposition 1.5.3 that $m \mid [(a-c) \cdot b + (b-d) \cdot c]$ and so $m \mid (ab - cd)$. Therefore, $ab \equiv cd \pmod{m}$.

□

Corollary 6.1.13. *If $a \equiv b \pmod{m}$, then $a^k \equiv b^k \pmod{m}$ for all $k \in \mathbb{N}$.*

Proof. By induction on k. For the base case when $k = 0$, the statement is trivial because $a^0 = 1 = b^0$. Suppose know that the statement is true for k, i.e. assume that $a^k \equiv b^k \pmod{m}$. Since $a^k \equiv b^k \pmod{m}$ and $a \equiv b \pmod{m}$, we can use Proposition 6.1.12 to conclude that $a^k \cdot a \equiv b^k \cdot b \pmod{m}$, so $a^{k+1} \equiv b^{k+1} \pmod{m}$. The result follows by induction. □

For example, notice that $4 \equiv 1 \pmod{3}$, so by the corollary we know that $4^k \equiv 1^k \pmod{3}$ for all $k \in \mathbb{N}$, from which we conclude that $4^k \equiv 1 \pmod{3}$ for all $k \in \mathbb{N}$. Therefore, $3 \mid 4^k - 1$ for all $k \in \mathbb{N}$, which gives another proof of Proposition 2.1.4.

By combining all of these results, we obtain the following, which says that if we plug equivalent values into a polynomial with integer coefficients, then the outputs of the polynomial are equivalent.

Corollary 6.1.14. *If $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ where each $a_i \in \mathbb{Z}$, and $b \equiv c \pmod{m}$, then $f(b) \equiv f(c) \pmod{m}$.*

Proof. For each k with $0 \leq k \leq n$, we have $b^k \equiv c^k \pmod{m}$ by Corollary 6.1.13. Thus, for each k with $0 \leq k \leq n$, we have $a_k b^k \equiv a_k c^k \pmod{m}$ by Proposition 6.1.9. We can now apply Proposition 6.1.12 repeatedly to conclude that

$$a_n b^n + a_{n-1}b^{n-1} + \cdots + a_1 b + a_0 \equiv a_n c^n + a_{n-1}c^{n-1} + \cdots + a_1 c + a_0 \pmod{m},$$

so $f(b) \equiv f(c) \pmod{m}$. □

For another nice consequence of these results, let's establish a simple test for determining divisibility by 3: A positive natural number n is divisible by 3 if and only if the sum of its decimal digits is divisible by 3. To see this, let $n \in \mathbb{N}^+$ be arbitrary, and write $n = a_m a_{m-1} \dots a_1 a_0$ in decimal notation, where each $a_i \in \mathbb{N}$ and $0 \leq a_i \leq 9$. By definition of decimal notation, we have

$$n = a_m \cdot 10^m + a_{m-1} \cdot 10^{m-1} + \cdots + a_1 \cdot 10 + a_0.$$

Now since $10 \equiv 1 \pmod{3}$, we know that $10^k \equiv 1^k \pmod{3}$, and hence $10^k \equiv 1 \pmod{3}$ for all $k \in \mathbb{N}$ by Corollary 6.1.13. Therefore,

$$a_m \cdot 10^m + a_{m-1} \cdot 10^{m-1} + \cdots + a_1 \cdot 10 + a_0 \equiv a_m \cdot 1 + a_{m-1} \cdot 1 + \cdots + a_1 \cdot 1 + a_0 \pmod{3},$$

and so

$$n \equiv a_m + a_{m-1} + \cdots + a_1 + a_0 \pmod{3}.$$

Alternatively, we can view this argument as looking at the polynomial $f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$, and noting that since $10 \equiv 1 \pmod{3}$, we have that $f(10) \equiv f(1) \pmod{3}$ by Corollary 6.1.14. Regardless of which path you take, we have shown that

$$3 \mid (n - (a_m + a_{m-1} + \cdots + a_1 + a_0)),$$

from which it follows that

$$3 \mid n \Longleftrightarrow 3 \mid (a_m + a_{m-1} + \cdots + a_1 + a_0).$$

A completely analogous argument gives the same test for divisibility by 9. That is, a positive natural number n is divisible by 9 if and only if the sum of its decimal digits is divisible by 9. We need only notice that $10 \equiv 1 \pmod{9}$, and then carry out the above argument with 3 replaced by 9.

We can also obtain a slightly twisted version of this idea to develop a test for divisibility by 11. Although $10 \not\equiv 1 \pmod{11}$, we do have that $10 \equiv -1 \pmod{11}$. Now if we follows the above argument with $n = a_m a_{m-1} \dots a_1 a_0$ in decimal notation, then we obtain

$$11 \mid n \Longleftrightarrow 11 \mid ((-1)^m a_m + (-1)^{m-1} a^{m-1} + \cdots - a_1 + a_0),$$

i.e. that 11 divides a positive natural number if and only if it divides the *alternating* sum of its decimal digits.

We now turn to the idea of solving simple linear equations in the world of congruences. Notice that given two numbers a and c, then solving a linear equation $ax = c$ for x is equivalent to finding roots of the degree 1 polynomial $ax - c$. Now if we working with numbers in $\mathbb{Q}$ or $\mathbb{R}$ (or any place where we can always divide by nonzero numbers), then as long as $a \neq 0$, there always exists a unique x with $ax = c$, namely $x = \frac{c}{a}$. In $\mathbb{Z}$, the situation is more interesting. For example, there does not exist $x \in \mathbb{Z}$ with $2x = 1$. In fact, we defined divisibility as a way to denote when we can solve such equations in $\mathbb{Z}$. That is, given $a, c \in \mathbb{Z}$, we defined $a \mid c$ to mean that there exists $x \in \mathbb{Z}$ with $ax = c$.

What happens when we move to the world of modular arithmetic? That is, given $a, c \in \mathbb{Z}$ and $m \in \mathbb{N}^+$, we want to know whether there exists $x \in \mathbb{Z}$ with $ax \equiv c \pmod{m}$. For example, does there exist $x \in \mathbb{Z}$ with $5x \equiv 4 \pmod{7}$? That is, can we find an $x \in \mathbb{Z}$ with $7 \mid (5x - 4)$? A little bit of trial and error shows that 5 works, i.e. $5 \cdot 5 \equiv 4 \pmod{7}$. Now 5 is not the only integer that works, because we know that if $a \equiv b \pmod{m}$ and $k \in \mathbb{Z}$, then $ka \equiv kb \pmod{m}$. Thus, if $b \equiv 5 \pmod{7}$, then $5b \equiv 5 \cdot 5 \pmod{7}$, so $5b \equiv 4 \pmod{7}$ as well. In other words, every element of $\overline{5}$ is also a solution to $5x \equiv 4 \pmod{7}$. It is also possible to show that these are the only solutions (we will develop general theory to help here later). Thus, although there are infinitely many $x \in \mathbb{Z}$ such that $5x \equiv 4 \pmod{7}$, there is only one solution up to equivalence.

By working through the possibilities, it turns out that we can always solve $5x \equiv c \pmod{7}$ for each choice of $c \in \mathbb{Z}$:

- $5 \cdot 0 \equiv 0 \pmod{7}$.
- $5 \cdot 3 \equiv 1 \pmod{7}$.
- $5 \cdot 6 \equiv 2 \pmod{7}$.
- $5 \cdot 2 \equiv 3 \pmod{7}$.
- $5 \cdot 5 \equiv 4 \pmod{7}$.
- $5 \cdot 1 \equiv 5 \pmod{7}$.
- $5 \cdot 4 \equiv 6 \pmod{7}$.

For $c < 0$ or $c > 6$, we know that c is congruent to an element of $\{0, 1, 2, 3, 4, 5, 6\}$, so the above choices will cover these cases by transitivity of $\equiv_7$.

Although we were able (surprisingly) to solve all of the above examples, we can not solve every linear congruence $ax \equiv c \pmod{m}$. A trivial example is when $m \mid a$ but $m \nmid c$. In this case, we will have $ax \equiv 0 \pmod{m}$ for all $x \in \mathbb{Z}$, so $ax \not\equiv c \pmod{m}$ for all $x \in \mathbb{Z}$. This is completely analogous to letting $a = 0$ and $c \neq 0$ when trying to solve $ax = c$ in $\mathbb{Q}$ or $\mathbb{R}$. However, there are also much more interesting examples. For example, suppose that we are trying to solve $4x \equiv 1 \pmod{6}$. We have the following:

- $4 \cdot 0 \equiv 0 \pmod{6}$.
- $4 \cdot 1 \equiv 4 \pmod{6}$.
- $4 \cdot 2 \equiv 2 \pmod{6}$.
- $4 \cdot 3 \equiv 0 \pmod{6}$.

- $4 \cdot 4 \equiv 4 \pmod{6}$.
- $4 \cdot 5 \equiv 2 \pmod{6}$.

Since every $x \in \mathbb{Z}$ is congruent to one of $0, 1, 2, 3, 4, 5$ modulo 6, it follows that there is no $x \in \mathbb{Z}$ with $4x \equiv 1 \pmod{6}$. Alternatively, we can simply notice that there is no $x \in \mathbb{Z}$ with $6 \mid (4x - 1)$ because 6 is even and $4x - 1$ is always odd.

The next theorem completely classifies when we can solve linear congruences.

Theorem 6.1.15. *Let $a, c \in \mathbb{Z}$ and $m \in \mathbb{N}^+$. The following are equivalent:*

1. *There exists $x \in \mathbb{Z}$ with $ax \equiv c \pmod{m}$.*
2. *There exists $x, y \in \mathbb{Z}$ with $ax + my = c$.*
3. $\gcd(a, m) \mid c$.

Proof.

- $(1) \Rightarrow (2)$: Suppose that (1) is true, and fix $k \in \mathbb{Z}$ with $ak \equiv c \pmod{m}$. By definition, we then have that $m \mid ak - c$, so we can fix $\ell \in \mathbb{Z}$ with $m\ell = ak - c$. We then have that $ak - m\ell = c$, and hence $ak + m(-\ell) = c$. Since $k, \ell \in \mathbb{Z}$, we have shown that (2) is true.
- $(2) \Rightarrow (3)$: Suppose that (2) is true. We then have that $c \in \{ak + m\ell : k, \ell \in \mathbb{Z}\}$, so $c \in \{n \cdot \gcd(a, m) : n \in \mathbb{Z}\}$ by Corollary 3.1.10. Fixing $n \in \mathbb{Z}$ with $c = n \cdot \gcd(a, m)$, we conclude that $\gcd(a, m) \mid c$.
- $(3) \Rightarrow (1)$: Suppose that (3) is true, and fix $n \in \mathbb{Z}$ with $n \cdot \gcd(a, m) = c$. By Corollary 3.1.10, we can fix $k, \ell \in \mathbb{Z}$ with $ak + m\ell = c$. We then have $ak - c = m \cdot \ell$, so $m \mid (ak - c)$, and hence $ak \equiv c \pmod{m}$.

□

We now immediately obtain the following result, telling us which numbers have multiplicative inverses modulo m.

Corollary 6.1.16. *Let $a \in \mathbb{Z}$ and $m \in \mathbb{N}^+$. There exists $x \in \mathbb{Z}$ with $ax \equiv 1 \pmod{m}$ if and only if $\gcd(a, m) = 1$.*

This gives another "explanation" for why we can cancel a common $k \in \mathbb{Z}$ when $ka \equiv kb \pmod{m}$ and $\gcd(k, m) = 1$. In that case, our corollary tells us that we can fix $\ell \in \mathbb{Z}$ with $k\ell \equiv 1 \pmod{m}$, and we can then multiply both sides of $ka \equiv kb \pmod{m}$ by ℓ and use the fact that $k\ell \equiv 1 \pmod{m}$ to conclude that $a \equiv b \pmod{m}$. In other words, if k has a multiplicative inverse modulo m, then we can use that inverse in the exact same way that we would "divide" by k in something like $\mathbb{Q}$ or $\mathbb{R}$. Of course, in $\mathbb{Q}$ and $\mathbb{R}$, the number 0 does not have a multiplicative inverse, but every other element does. In our setting, there can be other nonzero elements without multiplicative inverse. For example, if we are working modulo 6, then 1 and 5 have multiplicative inverses because they are relatively prime to 6, but 0, 2, 3, and 4 do not have multiplicative inverses.

How do we find these multiplicative inverses? For small values of m, we can simply perform an exhaustive check. However, for large values, the answer lies in the proof of the above theorem, where we found an x with $ax \equiv c \pmod{m}$ by finding $x, y \in \mathbb{Z}$ with $ax + my = c$. Recall that given $a, b \in \mathbb{Z}$, we can always find $k, \ell \in \mathbb{Z}$ with $ak + b\ell = \gcd(a, b)$ by "winding up" the Euclidean Algorithm. For example, in Problem 2 on Homework 4, you showed that $\gcd(471, 562) = 1$, so by the above theorem, we know that there exists $x \in \mathbb{Z}$ with $471x \equiv 562 \pmod{1}$. To find such an x, we look back to the solution to that homework problem, where we showed that

$$(-105) \cdot 471 + 88 \cdot 562 = 1.$$

Therefore,

$$562 \cdot (-88) = 471 \cdot (-105) - 1,$$

so

$$562 \mid 471 \cdot (-105) - 1,$$

and hence

$$471 \cdot (-105) \equiv 1 \pmod{562}.$$

Therefore, $x = -105$ is a multiplicative inverse for 471 modulo 562. Now any integer that is equivalent to -105 modulo 562 will also work, so if you want your integer to be from the standard complete residue system $\{0, 1, 2, 3, \dots, 561\}$, then we can choose $-105 + 562 = 457$. We then have

$$471 \cdot 457 \equiv 1 \pmod{562},$$

so 457 is a multiplicative inverse for 471 modulo 562.

Notice that when the modulus is prime, we are in a particularly nice situation.

Corollary 6.1.17. *Let p be prime. For each $a, c \in \mathbb{Z}$ with $p \nmid a$, there exists $x \in \mathbb{Z}$ with $ax \equiv c \pmod{p}$. In particular, for each $a \in \mathbb{Z}$ with $p \nmid a$, there exists $x \in \mathbb{Z}$ with $ax \equiv 1 \pmod{p}$, so every "nonzero" element has a multiplicative inverse.*

Proof. Let $a, c \in \mathbb{Z}$ be arbitrary with $p \nmid a$. We know that $\gcd(a, p)$ is a common divisor of p and a. Now the only positive divisors of p are 1 and p, so since $p \nmid a$, we must have $\gcd(a, p) = 1$. Thus, we trivially have $\gcd(a, p) \mid c$, and so there exists $x \in \mathbb{Z}$ with $ax \equiv c \pmod{p}$ by Theorem 6.1.15. $\square$

To see how we can visualize this theorem, consider the prime $p = 5$. Suppose that we choose a complete residue system modulo 5, and for simplicity here, suppose that we choose $\{0, 1, 2, 3, 4\}$. We can form a multiplication table of these residues, where we always return to this residue system upon multiplication. For example, we have $2 \cdot 3 = 6$, which is not in our residue system, but $2 \cdot 3 \equiv 1 \pmod{5}$, so in our table we will put 1 in the position corresponding to $2 \cdot 3$. With this in mind, we obtain the following table:

$\cdot$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Now the above corollary says that every element of $\{0, 1, 2, 3, 4\}$ will appear in each nonzero row (because the row corresponding to 2, say, is obtained by multiplying 2 by each x). Since multiplication is commutative, every element of $\{0, 1, 2, 3, 4\}$ will also appear in each nonzero column. Moreover, in each case, a number will appear exactly once in each nonzero row and column, because every surjective function from a finite set to itself must be bijective (or alternatively we can use Proposition 6.1.9 to argue that cancel the common factor to argue that we can not have any repeats in a given nonzero row). We can form a similar table for $p = 7$, using the complete residue system $\{0, 1, 2, 3, 4, 5, 6\}$ as follows:

$\cdot$	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Now if our modulus is not prime, then the tables are less pretty. For example, consider when $m = 6$, using the complete residue system $\{0, 1, 2, 3, 4, 5\}$:

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

In this case, since 1 and 5 are relatively prime to 6, every residue does appear in each of their rows and columns, but this is not the case for the other residues.

6.2 Modular Powers and Fermat's Little Theorem

We begin this section by asking a simple, but at first seemingly pointless, question. What is the remainder when we divide 2^{239} by 31? We certainly do *not* want to actually compute 2^{239}, because it is an incredibly large number. Instead, the idea is to look for a pattern in the remainders when we divide successively larger powers of 2 by 31. To determine the remainder when dividing 2^k by 31, we want to the find the unique natural number r with $0 \leq r < 31$ such that $2^k \equiv r \pmod{31}$. We start as follows:

- $2^0 = 1$.
- $2^1 = 2$.
- $2^2 = 4$.
- $2^3 = 8$.
- $2^4 = 16$.
- $2^5 = 32$, so $2^5 \equiv 1 \pmod{31}$.
- Instead of computing 2^6 directly, we multiply both sides of the previous congruence by 2 to conclude that $2^6 \equiv 2 \pmod{31}$.

Now that we have found a repeated remainder, it looks like we will continually repeat the cycle $1, 2, 4, 8, 16$ of remainders as we continue to multiply by 2. It is possible to prove this by induction in order to answer our original question, but we can instead argue directly as follows. Notice that we $2^5 = 32$, so $2^5 \equiv 1 \pmod{31}$. Using Division with Remainder, we have

$$239 = 5 \cdot 47 + 4.$$

Since $2^5 \equiv 1 \pmod{31}$, we can use Corollary 6.1.13 to conclude that

$$(2^5)^{47} \equiv 1^{47} \pmod{31},$$

and so it follows that

$$2^{235} \equiv 1 \pmod{31}.$$

Multiplying both sides by 2^4, we conclude that

$$2^{235} \cdot 2^4 \equiv 2^4 \pmod{31},$$

so

$$2^{239} \equiv 16 \pmod{31}.$$

Therefore, the remainder when we divide 2^{239} by 31 is 16.

In the above example, we started with 1, and repeatedly multiplied by 2, reducing to a residue in $\{0, 1, 2, \dots, 30\}$ at each step. In this process, we eventually reached 1 again. Does this always happen? If we are working modulo m, then we will certainly have a repeat within the first $m + 1$ steps, because there are only m possible residues. Furthermore, once we find a repeat, we will forever stay in the resulting cycle. However, it turns out that we might not return to 1. Here are some examples:

- Suppose that we are looking at powers of 2 modulo 12. We start with $1, 2, 4, 8$ as above. However, we have $8 \cdot 2 \equiv 4 \pmod{12}$, which is our first repetition. We then have $2 \cdot 4 = 8$, so we return to 8, and then forever will bounce between 4 and 8. Thus, the sequence of residues of powers of 2 modulo 12 looks like $1, 2, 4, 8, 4, 8, 4, 8, \dots$.

- Suppose that we are looking at powers of 6 modulo 15. We start with 1 and 6, but then a funny thing happens. We have $6 \cdot 6 = 36$, and so $6 \cdot 6 \equiv 6 \pmod{15}$. Thus, after the first 6 we see another 6, and thereafter will keep getting 6's. Thus, the sequence of residues of powers of 6 modulo 15 looks like $1, 6, 6, 6, \dots$.

The process of "following the output" should hopefully remind you of cycle notation in permutations. In that case, when we started with an $i \in [n]$, and repeatedly applied a permutation $\sigma \colon [n] \to [n]$, we argued in Proposition 5.5.1 that we must eventually return to i. The key fact there was that σ is a bijection, so we could not wrap back around at a later point, because otherwise we would violate the injectivity of σ. The same situation is happening here. In our original example, if we work with the set of residues $\{0, 1, 2, \dots, 30\}$ modulo 31, then the function that multiplies by 2 (and reduces to the corresponding residue) is injective by Proposition 6.1.9 because $\gcd(2, 31) = 1$, so if $2 \cdot b \equiv 2 \cdot c \pmod{31}$, then $b \equiv c \pmod{31}$. In other words, the row corresponding to 2 in the multiplication table modulo 31 is a permutation of the residues. As a result, we must eventually return to 1, rather than fall into a different pattern of repetition.

In fact, we can say even more about when we reach 1. In the remainder of this section, we consider the case when our modulus is a prime p. In this situation, we know that given any $a \in \mathbb{Z}$ with $p \nmid a$, we have that a has a multiplicative inverse modulo p, and so the corresponding row will be a bijection. Thus, we expect that some positive power of a will be congruent to 1 modulo p, and moreover this should occur in the first $p - 1$ powers (since no power will be congruent 0 modulo p because $p \nmid a$, and so we must reach a repeat in the p many powers from the set $\{0, 1, 2, 3, \dots, p - 1\}$). Our first major theorem confirms this, and says that $p - 1$ will serve as one such power that works.

Theorem 6.2.1 (Fermat's Little Theorem)**.** *Let $p \in \mathbb{N}^+$ be prime.*

1. *For all $a \in \mathbb{Z}$ with $p \nmid a$, we have $a^{p-1} \equiv 1 \pmod{p}$.*
2. *For all $a \in \mathbb{Z}$, we have $a^p \equiv a \pmod{p}$.*

Notice that Fermat's Little Theorem does *not* claim that $p - 1$ is the *smallest* positive power of a that is congruent to 1 modulo p. For example, in the case where $a = 2$ and $p = 31$ given above, we have $2^5 \equiv 1 \pmod{31}$. Of course, we also have $2^{30} \equiv 1 \pmod{p}$, which follows from raising both sides to the 6^{th} power.

We will give several proofs of this fundamental result below, but we first argue that the two versions of Fermat's Little Theorem are equivalent.

Proof of Equivalence.

- $(1) \Rightarrow (2)$: Suppose that (1) is true. Let $a \in \mathbb{Z}$ be arbitrary. We have two cases:

  - Suppose that $p \nmid a$. By (1), we then have $a^{p-1} \equiv 1 \pmod{p}$. Multiplying both sides by a, we conclude that $a^p \equiv a \pmod{p}$.
  - Suppose that $p \mid a$. We then have that $p \mid a^p$ trivially, so $p \mid (a^p - a)$, and hence $a^p \equiv a \pmod{p}$.

  Thus, $a^p \equiv a \pmod{p}$ in either case.

- (2) $\Rightarrow$ (1): Suppose that (2) is true. Let $a \in \mathbb{Z}$ be arbitrary with $p \nmid a$. By (2), we know that $a^p \equiv a \pmod{p}$, so $a \cdot a^{p-1} \equiv a \cdot 1 \pmod{p}$. Since $p \nmid a$ and p is prime, we know that $\gcd(a, p) = 1$, so we can apply Proposition 6.1.9 to conclude that $a^{p-1} \equiv 1 \pmod{p}$.

□

Our first argument proves the first version of Fermat's Little Theorem by thinking about multiplication of a as a bijection of the (nonzero) residues. As alluded to above, the key idea is that if $p \mid a$, then the nonzero elements of the row corresponding to a in the multiplication table modulo p is a permutation of the residues $\{1, 2, 3, \dots, p-1\}$.

Proof 1 of Fermat's Little Theorem. We prove the first version. Let $a \in \mathbb{Z}$ be arbitrary with $p \nmid a$. We know that $\{0, 1, 2, \dots, p-1\}$ is a complete residue system modulo p. Since $p \nmid a$ and p is prime, we have that $\gcd(a, p) = 1$. Therefore, by Proposition 6.1.9, if $b, c \in \{0, 1, 2, \dots, p-1\}$ with $b \neq c$, then $a \cdot b \not\equiv a \cdot c \pmod{p}$. Using Proposition 6.1.7, we conclude that $\{a \cdot 0, a \cdot 1, a \cdot 2, \dots, a \cdot (p-1)\}$ is also a complete residue system. Now we trivially have $a \cdot 0 = 0$, so every element of $\{1, 2, \dots, p-1\}$ is congruent to a unique element of $\{a, 2a, 3a, \dots, (p-1)a\}$. Applying Proposition 6.1.12, it follows that

$$a \cdot 2a \cdots (p-1)a \equiv 1 \cdot 2 \cdots (p-1) \pmod{p},$$

and hence

$$(p-1)! \cdot a^{p-1} \equiv (p-1)! \pmod{p}.$$

Now p is prime, and p does not divide any of the $p-1$ factors in $(p-1)!$, so $p \nmid (p-1)!$ by Corollary 3.2.7. Since p is prime, we have $\gcd(p, (p-1)!) = 1$, and hence we can apply Proposition 6.1.9 to conclude that

$$a^{p-1} \equiv 1 \pmod{p}.$$

□

Our second proof will be by induction. In order to make it work, we will need to express $(a+1)^p$ in terms of a^p. Of course, we can expand the former using the Binomial Theorem. The key fact that will make the induction work is the following interesting observation. It does not hold generally because, for example, $\binom{4}{2} = 6$ but $4 \nmid 6$.

Lemma 6.2.2. *Let $p \in \mathbb{N}^+$ be prime. For all $k \in \mathbb{N}$ with $1 \leq k \leq p-1$, we have $p \mid \binom{p}{k}$.*

Proof. Using our formula for $\binom{p}{k}$, notice that

$$p! = \binom{p}{k} \cdot k! \cdot (p-k)!.$$

Now $p \mid p!$ trivially, so p must divide the right-hand side. However, p does not divide any factor of $k!$, and p also does not divide any factor of $(p-k)!$. Therefore, by Corollary 3.2.7, it must be the case that $p \mid \binom{p}{k}$. □

Proof 2 of Fermat's Little Theorem. We prove the second version for all $a \in \mathbb{N}$ by induction. For the base case, we trivially have $0^p \equiv 0 \pmod{p}$. Now assume that $a \in \mathbb{N}$ has the property that $a^p \equiv a \pmod{p}$. By the Binomial Theorem, we have

$$\begin{aligned}(a+1)^p &= \sum_{k=0}^{p} \binom{p}{k} \cdot a^k \cdot 1^{p-k} \\ &= \sum_{k=0}^{p} \binom{p}{k} \cdot a^k \\ &= a^p + 1 + \sum_{k=1}^{p-1} \binom{p}{k} \cdot a^k.\end{aligned}$$

Since $p \mid \binom{p}{k}$ for all k with $1 \leq k \leq p-1$ by the lemma, we have

$$\binom{p}{k} \cdot a^k \equiv 0 \pmod{p}$$

for all k with $1 \leq k \leq p-1$. Therefore, we have

$$(a+1)^p \equiv a^p + 1 \pmod{p},$$

and since $a^p \equiv a \pmod{p}$ by induction, we conclude that

$$(a+1)^p \equiv a + 1 \pmod{p}.$$

Thus, the statement is true for $a+1$. By induction, we conclude that $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{N}$. Since every negative number is congruent to some nonnegative number modulo p, it follows that $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{Z}$ as well. □

We next give a combinatorial argument. The idea is take an arbitrary $a \in \mathbb{N}^+$, and think about coloring a p-spoke wheel with a many colors, but we will consider two colorings to be the same if we can rotate one coloring into the other. Since we have p spokes and a possible colors, there are a^p many total colorings. Now if we color all of the spokes the same, then that coloring is only equivalent to itself. But it turns out that every other coloring will be equivalent to exactly p many colorings. Using this idea, we can argue directly that $p \mid a^p - a$.

Proof 3 of Fermat's Little Theorem. We prove the second version for all $a \in \mathbb{N}^+$ by a combinatorial argument. Let $a \in \mathbb{N}^+$ be arbitrary. Consider the set $[a]^p$ of all sequences $(c_0, c_1, \dots, c_{p-1})$, where each c_i is a natural number with $1 \leq c_i \leq a$. Notice that there a^p many such sequences. Consider two such sequences to be equivalent if we can cyclically shift one to the other. It is straightforward to check that this is an equivalence relation. Notice that for each $c \in [a]$, the sequence $(c, c, \dots, c)$ is only equivalent to itself, giving a many equivalence classes of size 1.

Now consider an arbitrary sequence $(c_0, c_1, \cdots, c_{p-1})$ such that there exists $i < j$ with $c_i \neq c_j$. Our key claim is that all p of its cyclic shifts are distinct. To do this, we need to show that if two nontrivial cyclic shifts of $(c_0, c_1, \cdots, c_{p-1})$ are equal, then $c_0 = c_1 = \cdots = c_{p-1}$. Suppose then that two nontrivial cyclic shifts, say by d_1 and d_2 with $0 \leq d_1 < d_2 \leq p-1$, of the sequence $(c_0, c_1, \cdots, c_{p-1})$ are equal. We then have that $(c_0, c_1, \dots, c_{p-1})$ cyclically shifted by $d_2 - d_1$ must equal the original sequence $(c_0, c_1, \dots, c_{p-1})$. Let $d = d_2 - d_1$, and notice that $1 \leq d \leq p-1$. We then have that

$$(c_0, c_1, \dots, c_{p-1}) = (c_d, c_{d+1}, \dots, c_{p-1}, c_0, \dots, c_{d-1}).$$

From here, we can immediately conclude that $c_d = c_0$. Next, notice that $c_d = c_{2d}$, so $c_0 = c_{2d}$. By continuing this logic (and "wrapping around" as necessary), we see that $c_0 = c_i$ whenever there exists $k \in \mathbb{N}$ with $kd \equiv i \pmod{p}$. But notice that since $\gcd(d,p) = 1$ (because p is prime), there exists such a k for each i with $0 \leq i \leq p-1$. Thus, $c_i = c_0$ for all i.

We have shown that if $(c_0, c_1, \cdots, c_{p-1})$ is not a constant sequence, then it is equivalent to p distinct sequences. Let ℓ be the number of equivalence classes of size p. Now the equivalence classes partition $[a]^p$, and we have a total equivalence classes of size 1, so the set $[a]^p$ is partitioned into a equivalence classes of size 1 and ℓ equivalence classes of size p. Since $|[a]^p| = a^p$, it follows that $a^p = a + \ell p$. Thus, $p\ell = a^p - a$, so $p \mid a^p - a$, and hence $a^p \equiv a \pmod{p}$. □

Let's go back and look to examine our first proof of Fermat's Little Theorem. In that argument, a factor of $(p-1)!$ appeared on both sides, and we are argued that $p \nmid (p-1)!$, allowing us to cancel it from both sides. It turns out that we can say more about what $(p-1)!$ is modulo p. We will need the following result.

Proposition 6.2.3. *If $p \in \mathbb{N}^+$ is prime and $a^2 \equiv 1 \pmod{p}$, then either $a \equiv 1 \pmod{p}$ or $a \equiv -1 \pmod{p}$.*

Proof. Exercise (see Homework). □

Theorem 6.2.4 (Wilson's Theorem)**.** *If $p \in \mathbb{Z}$ is prime, then $(p-1)! \equiv -1 \pmod{p}$.*

Proof. Consider the complete residue system $\{0, 1, 2, \ldots, p-1\}$. Notice that $(p-1)!$ is the product of the nonzero elements of this residue system. From Corollary 6.1.17, we know that whenever $a \not\equiv 0 \pmod{p}$, there exists $b \in \mathbb{Z}$ with $ab \equiv 1 \pmod{p}$. Thus, for each $a \in \{1, 2, \ldots, p-1\}$, there exists $b \in \{1, 2, \ldots, p-1\}$ with $ab \equiv 1 \pmod{p}$. Furthermore, the choice of b is unique for each such a by Proposition 6.1.9. Now if we try to pair each element of $\{1, 2, \ldots, p-1\}$ with its unique such inverse, then the elements 1 and $p-1$ are the only elements that pair with themselves by Proposition 6.2.3 (since $p - 1 \equiv -1 \pmod{p}$). Therefore, when we perform this pairing on the remaining elements, we see that

$$\begin{aligned} (p-1)! &\equiv 1 \cdot (p-1) \pmod{p} \\ &\equiv (p-1) \pmod{p} \\ &\equiv -1 \pmod{p}. \end{aligned}$$

□

We now use these results to solve the following question: Given a prime p, when does there exist $a \in \mathbb{Z}$ with $a^2 \equiv -1 \pmod{p}$? That is, if we work with the world of arithmetic modulo a prime p, when does that world have a square root of -1? We start by noticing that $1^2 \equiv -1 \pmod{2}$, that $2^2 \equiv -1 \pmod{5}$, that $5^2 \equiv -1 \pmod{13}$, and that $4^2 \equiv -1 \pmod{17}$. However, it is possible to check by exhaustive search that there is no $a \in \mathbb{Z}$ with $a^2 \equiv -1 \pmod{p}$ when $p \in \{3, 7, 11, 19\}$. It turns out that we can classify the odd primes that have this property in a very elegant way (the only even prime is 2, which we handled above). The key idea is to use Wilson's Theorem to construct the element a. For example, consider the case when $p = 13$. By Wilson's Theorem, we know that

$$12! \equiv -1 \pmod{13},$$

so

$$12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \equiv -1 \pmod{13}.$$

Now the left-hand side might not look like a square, but we can view it in another way. Intuitively, when we are working modulo a number m, we "wrap around" when we reach m. With this in mind, we have $12 \equiv -1 \pmod{13}$, that $11 \equiv -2 \pmod{13}$, etc. Thus, we can rewrite the above line as

$$(-1) \cdot (-2) \cdot (-3) \cdot (-4) \cdot (-5) \cdot (-6) \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \equiv -1 \pmod{13},$$

and hence

$$(-1)^6 \cdot (6!)^2 \equiv -1 \pmod{13}.$$

Since 6 is even, it follows that

$$(6!)^2 \equiv -1 \pmod{13},$$

and thus we can take $a = 6!$. Notice that this argument will work whenever $\frac{p-1}{2}$ is even, because $\frac{p-1}{2}$ is the number of pairs that we have, and thus will be the exponent of -1. We now carry out the general argument.

Theorem 6.2.5. *Let $p \in \mathbb{N}^+$ be an odd prime. There exists an $a \in \mathbb{Z}$ with $a^2 \equiv -1 \pmod{p}$ if and only if $p \equiv 1 \pmod{4}$.*

Proof. Suppose first that there exists an $a \in \mathbb{Z}$ with $a^2 \equiv -1 \pmod{p}$, and fix such an a. Notice that $p \nmid a$ because otherwise $a^2 \equiv 0 \not\equiv -1 \pmod{p}$. By Fermat's Little Theorem, we have $a^{p-1} \equiv 1 \pmod{p}$. Thus

$$\begin{aligned}
1 &\equiv a^{p-1} \pmod{p} \\
&\equiv (a^2)^{\frac{p-1}{2}} \pmod{p} \\
&\equiv (-1)^{\frac{p-1}{2}} \pmod{p}.
\end{aligned}$$

Since $p \geq 3$, we have $1 \not\equiv -1 \pmod{p}$ and thus $\frac{p-1}{2}$ must be even. Fixing $k \in \mathbb{Z}$ with $2k = \frac{p-1}{2}$, we then have $4k = p - 1$, so $4 \mid (p-1)$ and hence $p \equiv 1 \pmod{4}$.

Suppose conversely that $p \equiv 1 \pmod{4}$. We then have that $4 \mid (p-1)$, so $\frac{p-1}{2}$ is even. Let $\ell = \frac{p-1}{2}$. By Wilson's Theorem we have

$$(p-1)! \equiv -1 \pmod{p}.$$

Now $-k \equiv p - k \pmod{p}$ for all k, so in the product $(p-1)!$ we can replace the latter half of the elements in the product $(p-1)!$ by the first ℓ negative numbers, i.e. the numbers in the list

$$\ell + 1, \ell + 2, \dots, p - 2, p - 1$$

are each equivalent modulo p to exactly one number in the following list:

$$-\ell, -(\ell - 1), \dots, -2, -1$$

Therefore, working modulo p, we have

$$\begin{aligned}
-1 &\equiv (p-1)! \pmod{p} \\
&\equiv 1 \cdot 2 \cdots (\ell - 1) \cdot \ell \cdot (\ell + 1) \cdot (\ell + 2) \cdots (p - 2) \cdot (p - 1) \pmod{p} \\
&\equiv 1 \cdot 2 \cdots (\ell - 1) \cdot \ell \cdot (-\ell) \cdot (-(\ell - 1)) \cdots (-2) \cdot (-1) \pmod{p} \\
&\equiv (-1)^{\ell} \cdot 1 \cdot 2 \cdots (\ell - 1) \cdot \ell \cdot \ell \cdot (\ell - 1) \cdots 2 \cdot 1 \pmod{p} \\
&\equiv (-1)^{\ell} \cdot (\ell!)^2 \pmod{p} \\
&\equiv (\ell!)^2 \pmod{p},
\end{aligned}$$

where the last line follows because $\ell = \frac{p-1}{2}$ is even. Therefore, there exists $a \in \mathbb{Z}$ with $a^2 \equiv -1 \pmod{p}$, namely $a = \ell! = (\frac{p-1}{2})!$. □

6.3 The Euler Function

When we were working modulo a prime p in the last section, we relied extensively on the fact that every a with $0 < a < p$ had a multiplicative inverse modulo p, which followed from Corollary 6.1.16 together with the fact that $\gcd(a, p) = 1$ whenever $0 < a < p$. When we move to a modulus m that is not prime, it is no longer the case that $\gcd(a, m) = 1$ whenever $0 < a < m$. Our first order of business is to determine how many elements of $\{1, 2, \dots, m - 1\}$ do have a multiplicative inverse.

Definition 6.3.1. *We define a function $\varphi\colon \mathbb{N}^+ \to \mathbb{N}^+$ as follows. For each $m \in \mathbb{N}^+$, we let*

$$\varphi(m) = |\{a \in [m] : \gcd(a, m) = 1\}|.$$

The function φ is called the Euler φ-function*, or* Euler totient function*.*

Notice that if $m > 1$, then $\gcd(m, m) = m > 1$, so $\varphi(m)$ is also just the number of elements of $[m-1]$ that are relatively prime to m. For example, we have the following:

- $\varphi(1) = 1$ because $\gcd(1, 1) = 1$.
- $\varphi(4) = 2$ because 1 and 3 are the only elements in $[4]$ that are relatively prime with 4.
- $\varphi(5) = 4$ because $1, 2, 3, 4$ are all relatively prime with 5, but $\gcd(5, 5) \neq 1$.
- $\varphi(6) = 2$ because 1 and 5 are the only elements in $[6]$ that are relatively prime with 6.
- $\varphi(p) = p - 1$ for all primes p because if $1 \leq a < p$, then $\gcd(a, p) = 1$.

In general, it is difficult to compute $\varphi(m)$, although we will eventually derive a formula in terms of the prime factorization of m. As we've just noted, the value of $\varphi(m)$ is easy to determine when m is prime. The next target is when m is a power of a prime. For some intuition, consider $125 = 5^3$. To determine $\varphi(125)$, we count the complement. That is, we determine which numbers in $\{1, 2, 3, \ldots, 125\}$ are *not* relatively prime to 125. Now we know from Proposition 3.3.1 that $Div(125) = \{1, 5, 25, 125\}$, so if $\gcd(a, 125) \neq 1$, then we must have $5 \mid a$. From here, it follows that the set of numbers in $[125]$ that are *not* relatively prime to 125 equals

$$\{5, 10, 15, 20, \ldots, 125\} = \{5k : k \in [25]\}.$$

Since we are counting the complement, we conclude that $\varphi(125) = 125 - 25 = 100$. We now generalize this example.

Proposition 6.3.2. *If $p \in \mathbb{N}$ is prime and $k \in \mathbb{N}^+$, then*

$$\begin{aligned} \varphi(p^k) &= p^k - p^{k-1} \\ &= p^{k-1}(p-1) \\ &= p^k \cdot \left(1 - \frac{1}{p}\right). \end{aligned}$$

Proof. Let

$$A = \{a \in [p^k] : \gcd(a, p^k) = 1\}.$$

By definition, we have that $\varphi(p^k) = |A|$, so we need to count how many elements are in A. To this, we count the complement. In other words, we determine the cardinality of

$$B = [p^k] \backslash A = \{a \in [p^k] : \gcd(a, p^k) \neq 1\}.$$

Our claim is that

$$B = \{pm : 1 \leq m \leq p^{k-1}\}.$$

We first show that $\{pm : 1 \leq m \leq p^{k-1}\} \subseteq B$. Let $a \in \{pm : 1 \leq m \leq p^{k-1}\}$ be arbitrary, and fix $m \in \mathbb{N}$ with $1 \leq m \leq p^{k-1}$ such that $a = pm$. Notice that since $1 \leq m \leq p^{k-1}$, we have $p \leq pm \leq pp^{k-1}$, which is to say that $p \leq a \leq p^k$, so $a \in [p^k]$. Now we clearly have that $p \mid a$ because $a = pm$, and we also have $p \mid p^k$ because $p = pp^{k-1}$ and $k - 1 \geq 0$, so $\gcd(n, p^k) \neq 1$ (as $p > 1$ is a common divisor). It follows that $a \in B$, and since a was arbitrary we conclude that $\{pm : 1 \leq m \leq p^{k-1}\} \subseteq B$.

Conversely, let $a \in B$ be arbitrary. Let $d = \gcd(a, p^k)$, so since $a \in B$, we know that $d > 1$. Now we know from Proposition 3.3.1 that $Div(p^k) = \{1, p, p^2, \dots, p^k\}$, so since $d \mid p$ and $d \neq 1$, we know that $d \in \{p, p^2, \dots, p^k\}$. Since p divides every element of this set, it follows that $p \mid d$. Now we also know that $d \mid a$, so by transitivity of the divisibility relation it follows that $p \mid a$. Thus, we fix $m \in \mathbb{Z}$ with $a = pm$. Notice that $m > 0$ because $a > 0$ and $p > 0$. Finally, we must have $m \leq p^{k-1}$ because otherwise $m > p^{k-1}$ and so $a = pm > p^k$, contradicting our assumption that $a \in B$. Therefore, $B \subseteq \{pm : 1 \leq m \leq p^{k-1}\}$.

We have shown that $B = \{pm : 1 \leq m \leq p^{k-1}\}$. Now the set on the right has p^{k-1} many elements (one for each choice of m), so $|B| = p^{k-1}$. It follows that

$$\varphi(p^k) = |[p^k] \backslash B| = p^k - p^{k-1}.$$

The latter two formulas are now just simple algebra. □

With prime powers in hand, we next turn to the simplest example of a number that is divisible by more that one prime. Suppose then that $n = pq$, where p and q are distinct primes. Notice that if $a \in [n]$ with $\gcd(a, n) \neq 1$, then a must be divisible by either p or q (or both) because $Div(pq) = \{1, p, q, pq\}$. There are $\frac{pq}{p} = q$ many elements of $[pq]$ divisible by p (namely $p, 2p, 3p, \dots, qp$), and $\frac{pq}{q} = p$ many elements of $[pq]$ divisible by q (namely $q, 2q, 3q, \dots, pq$). Also, there is one element, namely pq, that is divisible by both. Thus, there are $p + q - 1$ many elements divisible by at least one of p or q, and hence

$$\begin{aligned}\varphi(pq) &= pq - (p + q - 1) \\ &= pq - p - q + 1 \\ &= (p - 1)(q - 1).\end{aligned}$$

Notice that we used a simple version of Inclusion-Exclusion here, and that $\varphi(pq) = \varphi(p) \cdot \varphi(q)$.

Before moving on, we recall the following result.

Proposition 6.3.3. *Let $a, b, c \in \mathbb{Z}$. If $a \mid c$, $b \mid c$, and $\gcd(a, b) = 1$, then $ab \mid c$.*

Proof. See Problem 6 on Homework 4. □

For a more interesting example, consider

$$504 = 2^3 \cdot 3^2 \cdot 7.$$

Thus, the prime divisors of 504 are exactly 2, 3, and 7. Let

$$S = \{a \in [504] : \gcd(a, 504) = 1\}.$$

To count the number of elements in S, we instead count the number of elements in the complement, i.e. we determine the cardinality of $[504] \backslash S$. Now given $a \in [504]$, we claim that $a \notin S$ if and only if at least one of the primes 2, 3, or 7 divides a. The right-to-left direction is clear. For the left-to-right direction, let $a \notin S$ be arbitrary, and let $d = \gcd(a, 504) > 1$. By Proposition 3.2.2, we can fix a prime divisor p of d. We then have that both $p \mid a$ and $p \mid 504$. Since $504 = 2^3 \cdot 3^2 \cdot 7$, we can use Corollary 3.2.7 to conclude that p divides one of 2, 3, or 7, and hence must equal one of 2, 3, or 7. Therefore, at least one of 2, 3, or 7 divides a.

With this is mind, we define the following sets:

- Let $A_1 = \{a \in [504] : 2 \mid a\}$.
- Let $A_2 = \{a \in [504] : 3 \mid a\}$.
- Let $A_3 = \{a \in [504] : 7 \mid a\}$.

Now an element of [504] is *not* in S exactly when it is in $A_1 \cup A_2 \cup A_3$, so

$$[504] \backslash S = A_1 \cup A_2 \cup A_3.$$

We have

$$|A_1| = \frac{504}{2} = 252 \qquad |A_2| = \frac{504}{3} = 168 \qquad |A_3| = \frac{504}{7} = 72,$$

because, for example, $A_2 = \{3k : 1 \leq k \leq 168\}$. To determine $A_1 \cap A_2$, notice that if $2 \mid a$ and $3 \mid a$, then $6 \mid a$ by Proposition 6.3.3 (since $\gcd(2,3) = 1$). Since the converse is trivially true, it follows that

$$A_1 \cap A_2 = \{a \in [504] : 6 \mid a\}$$

and so $A_1 \cap A_2 = \{6k : k \in 84\}$. Following the same argument for the other double intersections, we see that

$$|A_1 \cap A_2| = \frac{504}{6} = 84 \qquad |A_1 \cap A_3| = \frac{504}{14} = 36 \qquad |A_2 \cap A_3| = \frac{504}{21} = 24.$$

Finally, we can apply Proposition 6.3.3 twice together with Corollary 3.2.11 to conclude that $A_1 \cap A_2 \cap A_3 = \{a \in [504] : 42 \mid a\}$, and hence

$$|A_1 \cap A_2 \cap A_3| = \frac{504}{42} = 12.$$

Using Inclusion-Exclusion, it follows that

$$\begin{aligned} |A_1 \cup A_2 \cup A_3| &= |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3| \\ &= 252 + 168 + 72 - 84 - 36 - 24 + 12 \\ &= 360. \end{aligned}$$

Subtracting these 360 many elements from the 504 many elements in [504], we conclude that

$$|S| = 504 - 360 = 144.$$

We now generalize this proof to give a formula for $\varphi(n)$ in terms of its prime factorization.

Theorem 6.3.4. *Suppose that $n \in \mathbb{N}$ with $n \geq 2$. Write $n = p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}$ where the p_i are distinct primes and each $k_i > 0$. We then have*

$$\begin{aligned} \varphi(n) &= n \cdot \left(1 - \frac{1}{p_1}\right) \cdot \left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_\ell}\right) \\ &= p_1^{k_1}\left(1 - \frac{1}{p_1}\right) \cdot p_2^{k_2}\left(1 - \frac{1}{p_2}\right) \cdots p_\ell^{k_\ell}\left(1 - \frac{1}{p_\ell}\right) \\ &= p_1^{k_1-1}(p_1 - 1) \cdot p_2^{k_2-1}(p_2 - 1) \cdots p_\ell^{k_\ell-1}(p_\ell - 1) \\ &= \varphi(p_1^{k_1}) \cdot \varphi(p_2^{k_2}) \cdots \varphi(p_\ell^{k_\ell}). \end{aligned}$$

Proof. For each $i \in [\ell]$, let

$$A_i = \{a \in [n] : p_i \mid a\}.$$

We first claim that

$$A_1 \cup A_2 \cup \cdots \cup A_\ell = \{a \in [n] : \gcd(a, n) \neq 1\}.$$

To see this, first let $a \in A_1 \cup A_2 \cup \cdots \cup A_\ell$ be arbitrary. Fix $i \in [\ell]$ with $a \in A_i$. We then have that $p_i \mid a$ by definition of A_i. We also trivially have that $p_i \mid n$, so p_i is a common divisor of a and n. Therefore, $\gcd(a, n) \neq 1$.

Conversely, let $a \in [n]$ with $\gcd(a,n) \neq 1$ be arbitrary. Let $d = \gcd(a,n) > 1$. By Proposition 3.2.2, we can fix a prime q with $q \mid d$. We then have that $q \mid a$ and $q \mid n$. Since $n = p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}$, we know that $q \mid p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}$, so we use Corollary 3.2.7 to fix an i with $q \mid p_i$. Now p_i is prime, so must have $q = p_i$. Since $q \mid a$, it follows that $a \in A_i$.

We now calculate

$$|A_1 \cup A_2 \cup \cdots \cup A_\ell|$$

using Inclusion-Exclusion. For each i, we have

$$|A_i| = \frac{n}{p_i}$$

because $A_i = \{kp_i : 1 \leq k \leq \frac{n}{p_i}\}$. Whenever $i < j$ we have

$$A_i \cap A_j = \{a \in [n] : p_i p_j \mid a\}$$

by Proposition 6.3.3 and Corollary 3.2.11, so

$$|A_i \cap A_j| = \frac{n}{p_i p_j}.$$

Similarly, if $i < j < k$, then

$$A_i \cap A_j \cap A_k = \{a \in [n] : p_i p_j p_k \mid a\}$$

and hence

$$|A_i \cap A_j \cap A_k| = \frac{n}{p_i p_j p_k}.$$

Therefore

$$|A_1 \cup A_2 \cup \cdots \cup A_k| = \sum_i \frac{n}{p_i} - \sum_{i<j} \frac{n}{p_i p_j} + \sum_{i<j<k} \frac{n}{p_i p_j p_k} - \dots$$

It follows that

$$\begin{aligned}
\varphi(n) &= n - \sum_i \frac{n}{p_i} + \sum_{i<j} \frac{n}{p_i p_j} - \sum_{i<j<k} \frac{n}{p_i p_j p_k} + \dots \\
&= n \cdot \left(1 - \sum_i \frac{1}{p_i} + \sum_{i<j} \frac{1}{p_i p_j} - \sum_{i<j<k} \frac{1}{p_i p_j p_k} + \dots\right) \\
&= n \cdot \left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_\ell}\right)
\end{aligned}$$

□

We can now use this formula directly to calculate $\varphi(504)$. For example, we have

$$\begin{aligned}
\varphi(504) &= \varphi(2^3 \cdot 3^2 \cdot 7) \\
&= \varphi(2^3) \cdot \varphi(3^2) \cdot \varphi(7) \\
&= 4(2-1) \cdot 3(3-1) \cdot (7-1) \\
&= 4 \cdot 6 \cdot 6 \\
&= 144.
\end{aligned}$$

Although we have established a formula for $\varphi(m)$, the fact that we can break down φ to its values on prime powers suggests that φ has the following property: Whenever $m, n \in \mathbb{N}^+$ with $\gcd(m,n) = 1$, we have

$\varphi(mn) = \varphi(m) \cdot \varphi(n)$. In other words, it appears that φ has the same property as the divisor function d that we established in Corollary 3.3.6.

In fact, we can prove that $\varphi(mn) = \varphi(m) \cdot \varphi(n)$ whenever $\gcd(m, n) = 1$ directly, and then use this result to give another derivation of the above formula. Before working out the general details, we look at a few examples that illustrate an interesting general phenomenon.

Suppose that we want to compute $\varphi(12)$. Of course, we can just compute that

$$\{a \in [12] : \gcd(a, 12) = 1\} = \{1, 5, 7, 11\},$$

so $\varphi(12) = 4$. But let's think about factoring 12 into the product of two relatively prime natural numbers. We have $12 = 4 \cdot 3$ where $\gcd(4, 3) = 1$. Suppose that we arrange the first 12 positive natural numbers in a the following 4×3 table:

$$\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{array}$$

Now let's highlight only those numbers that are relatively prime to 12:

$$\begin{array}{ccc} 1 & - & - \\ - & 5 & - \\ 7 & - & - \\ - & 11 & - \end{array}$$

Notice that in the last column, each of the numbers is divisible by 3, so none of the numbers is relatively prime to 12. In the other two columns, we see that exactly 2 of the four numbers survive, and that $\varphi(4) = 2$. In other words, exactly $\varphi(3)$ many columns have elements in them, and in each of those, there are exactly $\varphi(4)$ many such elements.

Alternatively, we could have instead switched the role of 3 and 4. In this case, we have the following 3×4 table

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{array}$$

and if we only keep the elements that are relatively prime to 12, then we are left with

$$\begin{array}{cccc} 1 & - & - & - \\ 5 & - & 7 & - \\ - & - & 11 & - \end{array}$$

In this case, notice that nothing appears in the second and fourth columns, which correspond to numbers that are *not* relatively prime to 4. In the other two columns, corresponding to numbers that are relatively prime 4, we see that exactly $\varphi(3) = 2$ many numbers survive. In other words, exactly $\varphi(4)$ many columns have elements in them, and in each of those, there are exactly $\varphi(3)$ many such elements.

For another example, consider $20 = 4 \cdot 5$. If we arrange the first 20 natural numbers in a 5×4 table, and only keep those relatively prime to 20, we obtain the table

$$\begin{array}{cccc} 1 & - & 3 & - \\ - & - & 7 & - \\ 9 & - & 11 & - \\ 13 & - & - & - \\ 17 & - & 19 & - \end{array}$$

Note that exactly $\varphi(4) = 2$ columns have elements, and in each of those, there are exactly $\varphi(5) = 4$ many such elements. Furthermore, notice that columns that do not have elements are the second and fourth, which correspond to the numbers that are not relatively prime to 4.

If we instead arrange the first 20 natural numbers in a 5×4 table, and only keep those relatively prime to 20, we obtain

$$\begin{array}{ccccc} 1 & - & 3 & - & - \\ - & 7 & - & 9 & - \\ 11 & - & 13 & - & - \\ - & 17 & - & 19 & - \end{array}$$

Again, exactly $\varphi(5) = 4$ columns have elements, and in each of those, there are exactly $\varphi(4) = 2$ many such elements. Furthermore, notice that column that does not have elements is the fifth, which correspond to the number that is not relatively prime to 5.

Before jumping into the general proof, we establish a few results.

Proposition 6.3.5. *If $a, b \in \mathbb{Z}$ and $m \in \mathbb{N}^+$ are such that $a \equiv b \pmod{m}$, then $\gcd(a, m) = \gcd(b, m)$.*

Proof. Exercise (see Homework). □

Proposition 6.3.6. *Suppose that $a, b \in \mathbb{Z}$, that $m \in \mathbb{N}^+$, and that $\gcd(a, m) = 1$. We then have that $\{b, a + b, 2a + b, \ldots, (m - 1)a + b\}$ is a complete residue system modulo m.*

Proof. Let $k, \ell \in \{0, 1, 2, \ldots, m - 1\}$ be arbitrary such that $b + ka \equiv b + \ell a \pmod{m}$. Subtracting b from both sides, we have $ka \equiv \ell a \pmod{m}$. Since $\gcd(a, m) = 1$, we can cancel a from both sides to conclude that $k \equiv \ell \pmod{m}$. Since $k, \ell \in \{0, 1, 2, \ldots, m - 1\}$, it follows that $k = \ell$.

Therefore, if $k, \ell \in \{0, 1, 2, \ldots, m - 1\}$ with $k \neq \ell$, then $b + ka \not\equiv b + \ell a \pmod{m}$. Since $|\{b, a + b, 2a + b, \ldots, (m - 1)a + b\}| = m$, we can apply Proposition 6.1.7 to conclude that $\{b, a + b, 2a + b, \ldots, (m - 1)a + b\}$ is a complete residue system. □

Proposition 6.3.7. *Let $a, b, c \in \mathbb{Z}$. If $\gcd(a, b) = 1$ and $\gcd(a, c) = 1$, then $\gcd(a, bc) = 1$.*

Proof. Exercise (see Homework) □

Theorem 6.3.8. *If $m, n \in \mathbb{N}^+$ and $\gcd(m, n) = 1$, then $\varphi(mn) = \varphi(m) \cdot \varphi(n)$.*

Proof. Let $m, n \in \mathbb{N}^+$ be arbitrary with $\gcd(m, n) = 1$. We consider the first mn positive natural numbers, and arrange them in the following table:

$$\begin{array}{ccccc} 1 & 2 & 3 & \ldots & n \\ n+1 & n+2 & n+3 & \ldots & 2n \\ 2n+1 & 2n+2 & 2n+3 & \ldots & 3n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (m-1)n+1 & (m-1)n+2 & (m-1)n+3 & \ldots & mn \end{array}$$

Let $k \in \{1, 2, 3, \ldots, n\}$ be arbitrary, and consider the k^{th} column of this table, i.e. consider the set $\{k, n + k, 2n + k, \ldots (m - 1)n + k\}$. We have two possibilities:

- *Case 1:* Suppose that $\gcd(k, n) \neq 1$: Let $d = \gcd(k, n)$. Since $d \mid n$, we have $d \mid \ell n$ for all $\ell \in \mathbb{Z}$, and hence $d \mid \ell n + k$ for all $\ell \in \mathbb{Z}$. Also, since $d \mid n$, we have $d \mid mn$. Therefore, $d > 1$ is a common divisor of mn and each element of the this column. It follows that no element of the k^{th} column is relatively prime to mn.

- *Case 2:* Suppose that $\gcd(k, n) = 1$: First notice that for all $\ell \in \mathbb{Z}$, we have $\ell n + k \equiv k \pmod{n}$, so $\gcd(\ell n + k, n) = \gcd(k, n) = 1$ by Proposition 6.3.5. In other words, all elements of the k^{th} column are relatively prime to n. Now since $\gcd(m, n) = 1$, we know that $\{k, n + k, 2n + k, \dots (m-1)n + k\}$ is a complete residue system modulo m by Proposition 6.3.6. Recall that $\{1, 2, 3, \dots, m\}$ is also a complete residue system modulo m, so using Proposition 6.3.5 together with the fact that exactly $\varphi(m)$ of the elements of $\{1, 2, 3, \dots, m\}$ are relatively prime to m, we conclude that exactly m elements of the set $\{k, n + k, 2n + k, \dots (m-1)n + k\}$ are relatively prime to m. Now the elements of this column that are not relatively prime to m are also not relatively prime to mn (because a common divisor of such an element and m is certainly a common divisor of such an element and mn). However, the $\varphi(m)$ many elements of this column that are relatively prime to m are also relatively prime to n (since every element of the column is), and hence are relatively prime to mn by Proposition 6.3.7. Thus, exactly $\varphi(m)$ many elements of the k^{th} column are relatively prime to mn.

Putting it all together, the elements of $\{a \in [mn] : \gcd(a, m) = 1\}$ appear in only $\varphi(n)$ many columns (since there are $\varphi(n)$ many k that belong to Case 2), and each of those columns contains exactly $\varphi(m)$ many elements. Therefore, $\varphi(mn) = \varphi(m) \cdot \varphi(n)$. $\square$

We can now use Theorem 6.3.8, Proposition 6.3.2, and Corollary 3.2.11 to give another proof of our formula for $\varphi(n)$ in terms of the prime factorization of a natural number $n \geq 2$. That is, given $n \in \mathbb{N}$ and $n = p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}$ where the p_i are distinct primes and each $k_i > 0$, then

$$\begin{aligned}
\varphi(n) &= \varphi(p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}) \\
&= \varphi(p_1^{k_1}) \cdot \varphi(p_2^{k_2}) \cdots \varphi(p_\ell^{k_\ell}) \\
&= p_1^{k_1 - 1}(p_1 - 1) \cdot p_2^{k_2 - 1}(p_2 - 1) \cdots p_\ell^{k_\ell - 1}(p_\ell - 1).
\end{aligned}$$

Recall that a complete residue system modulo m is a choice of exactly one element out of each equivalence class of $\equiv_m$. By Proposition 6.3.5, we know that the elements of any one equivalence class all have the same greatest common divisor with m. By definition of φ, we know that exactly $\varphi(m)$ of the elements of $\{1, 2, 3, \dots, m\}$ are relatively prime to m, so since $\{1, 2, 3, \dots, m\}$ has exactly one element in each equivalence class, we conclude that exactly $\varphi(m)$ equivalence classes have at least one (and hence every) element that is relatively prime to m. We give a special name to sets that choose exactly one element out of each of these special equivalence classes.

Definition 6.3.9. *Let $m \in \mathbb{N}^+$. Given distinct $b_1, b_2, \dots, b_{\varphi(m)} \in \mathbb{Z}$, we say that $\{b_1, b_2, \dots, b_{\varphi(m)}\}$ is a* reduced residue system modulo m *if $\gcd(b_i, m) = 1$ for all i, and if for every $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, there exists a unique i with $1 \leq i \leq \varphi(m)$ such that $a \equiv b_i \pmod{m}$.*

Proposition 6.3.10. *Let $m \in \mathbb{N}^+$, and let $b_1, b_2, \dots, b_{\varphi(m)} \in \mathbb{Z}$. If $\gcd(b_i, m) = 1$ for all i, and $b_i \not\equiv b_j \pmod{m}$ whenever $i \neq j$, then $\{b_1, b_2, \dots, b_{\varphi(m)}\}$ is a reduced residue system modulo m.*

Proof. Assume that $\gcd(b_i, m) = 1$ for all i, and that $b_i \not\equiv b_j \pmod{m}$ whenever $i \neq j$. The former assumption implies that the b_i each come from an equivalence class where least one (and hence every) element that is relatively prime to m. The latter assumption implies that the b_i come from distinct equivalence classes. Since there are exactly $\varphi(m)$ many equivalence classes that contain elements that are relatively prime to m, it follows that every $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$ must be equivalent to some b_i modulo m. $\square$

With these concepts in hand, we can now generalize Fermat's Little Theorem to the any modulus. Notice that we make two changes to Fermat's Little Theorem. One change is that in the first version of Fermat's Little Theorem, we assumed that $p \nmid a$. Since p was prime, this is equivalent to saying that $\gcd(a, p) = 1$, and in the general setting we need to adopt this latter version. The second change is that we replace the exponent $p - 1$ by $\varphi(m)$. Of course, when m is prime, then $\varphi(m) = m - 1$, so everything matches up when m is prime.

Theorem 6.3.11 (Euler's Theorem)**.** *Let $m \in \mathbb{N}^+$. For all $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, we have $a^{\varphi(m)} \equiv 1 \pmod{m}$.*

Proof. Let $a \in \mathbb{Z}$ be arbitrary with $\gcd(a, m) = 1$. Let $\{c_1, c_2, \dots, c_{\varphi(m)}\}$ be the subset of $\{1, 2, 3, \dots, m\}$ consisting of those elements that are relatively prime to m, and notice that $\{c_1, c_2, \dots, c_{\varphi(m)}\}$ is a reduced residue system modulo m. Since $\gcd(a, m) = 1$ and $\gcd(c_i, m) = 1$ for all i, we can apply Proposition 6.3.7 to conclude that $\gcd(ac_i, m) = 1$ for all i. Also, notice that if $i, j \in \{1, 2, \dots, \varphi(m)\}$ and $ac_i \equiv ac_j \pmod{m}$, then since $\gcd(a, m) = 1$, we can use Proposition 6.1.9 to conclude that $c_i \equiv c_j \pmod{m}$, and hence $i = j$. Therefore, $ac_i \not\equiv ac_j \pmod{m}$ whenever $i \neq j$. Using Proposition 6.3.10, it follows that $\{ac_1, ac_2, \dots, ac_{\varphi(m)}\}$ is also a reduced residue system modulo m. Thus, each element of $\{ac_1, ac_2, \dots, ac_{\varphi(m)}\}$ is congruent modulo m to a unique element of $\{c_1, c_2, \dots, c_{\varphi(m)}\}$, so

$$ac_1 \cdot ac_2 \cdot \dots \cdot ac_{\varphi(m)} \equiv c_1 \cdot c_2 \cdot \dots \cdot c_{\varphi(m)} \pmod{m}.$$

and hence

$$a^{\varphi(m)} c_1 c_2 \cdots c_{\varphi(m)} \equiv c_1 c_2 \cdots c_{\varphi(m)} \pmod{m}.$$

Now we know that $\gcd(c_i, m) = 1$ for each i, so we can repeatedly apply Proposition 6.3.7 to conclude that $\gcd(c_1 c_2 \cdots c_{\varphi(m)}, m) = 1$. Finally, using Proposition 6.1.9, we cancel $c_1 c_2 \cdots c_{\varphi(m)}$ from both sides to arrive at

$$a^{\varphi(m)} \equiv 1 \pmod{m}.$$

□

Given $n \in \mathbb{N}^+$, recall that $\varphi(n)$ counts the number of elements $a \in [n]$ such that $\gcd(a, n) = 1$. Now if we look at the set $\{\gcd(a, n) : a \in [n]\}$, then we notice that every element of this set is a divisor of n, so

$$\{\gcd(a, n) : a \in [n]\} \subseteq Div(n) \cap \mathbb{N}^+.$$

In fact, we have

$$\{\gcd(a, n) : a \in [n]\} = Div(n) \cap \mathbb{N}^+$$

because for any $d \in \mathbb{N}^+$ with $d \mid n$, we have $d \in [n]$ and $\gcd(d, n) = d$. Given $d \in \mathbb{N}^+$ with $d \mid n$, what if we wanted to count the number of elements $a \in [n]$ such that $\gcd(a, n) = d$? For example, consider $n = 20$. We saw about that $\varphi(20) = 8$, so

$$|\{a \in [20] : \gcd(a, 20) = 1\}| = 8.$$

Now $2 \mid 20$, so suppose we want to determine

$$|\{a \in [20] : \gcd(a, 20) = 2\}|.$$

Of course, in this small example, we can simply work through each value to determine that

$$\{a \in [20] : \gcd(a, 20) = 2\} = \{2, 6, 14, 18\},$$

and hence

$$|\{a \in [20] : \gcd(a, 20) = 1\}| = 4.$$

However, if we look at the above set more closely, we notice something interesting. Of course, every element is divisible by 2, and it we divide each of the elements by 2, we end up with the set $\{1, 3, 7, 9\}$, which is just the set of elements relatively prime to 10. In other words, there is a natural bijection between

$$\{a \in [10] : \gcd(a, 10) = 1\} \qquad \text{and} \qquad \{a \in [20] : \gcd(a, 20) = 2\},$$

obtained by multiplying every element of the first set by 2, and thus $|\{a \in [20] : \gcd(a, 20) = 2\}| = \varphi(10) = 4$. The key fact that makes this work is that $\gcd(2a, 2b) = 2 \cdot \gcd(a, b)$ for all $a, b \in \mathbb{N}^+$. In fact, this generalizes to the following.

Proposition 6.3.12. *For all $a, b, c \in \mathbb{N}^+$, we have* $\gcd(ca, cb) = c \cdot \gcd(a, b)$.

Proof. Exercise (see Exam 1). $\square$

Proposition 6.3.13. *For all $d, n \in \mathbb{N}^+$ with $d \mid n$, we have* $|\{a \in [n] : \gcd(a, n) = d\}| = \varphi(n/d)$.

Proof. Let
$$S = \{b \in [n/d] : \gcd(b, n/d) = 1\},$$
and notice that $|S| = \varphi(\frac{n}{d})$ by definition. We claim that
$$\{a \in [n] : \gcd(a, n) = d\} = \{db : b \in S\}.$$

We now prove each containment:

- Let $a \in [n]$ be arbitrary with $\gcd(a, n) = d$. We then have that $d \mid a$, so we can fix $b \in \mathbb{Z}$ with $a = db$. Notice that $b \geq 1$ because $a \geq 1$ and $d \geq 1$. Also, since $a \leq n$, we must have $b \leq \frac{n}{d}$, and hence $b \in [n/d]$. Now
$$\begin{aligned} d &= \gcd(a, n) \\ &= \gcd(db, d \cdot (n/d)) \\ &= d \cdot \gcd(b, n/d) && \text{(by Proposition 6.3.12).} \end{aligned}$$
Dividing both sides by $d \geq 1$, it follows that $\gcd(b, n/d) = 1$, so $b \in S$. Thus, $a \in \{db : b \in S\}$.

- Conversely, let $b \in S$ be arbitrary. Since $b \in S$, we know that $1 \leq b \leq n/d$, so $d \leq db \leq n$, and hence $db \in [n]$. We also have
$$\begin{aligned} \gcd(db, n) &= \gcd(db, d \cdot (n/d)) \\ &= d \cdot \gcd(b, n/d) && \text{(by Proposition 6.3.12).} \\ &= d \cdot 1 \\ &= d, \end{aligned}$$
Thus, $db \in \{a \in [n] : \gcd(a, n) = d\}$.

Since $\{a \in [n] : \gcd(a, n) = d\} = \{db : b \in S\}$, and since $db_1 \neq db_2$ whenever $b_1, b_2 \in S$ with $b_1 \neq b_2$, it follows that
$$\{a \in [n] : \gcd(a, n) = d\} = |S| = \varphi(n/d).$$
$\square$

Corollary 6.3.14. *For all $n \in \mathbb{N}^+$, we have*
$$n = \sum_{d \mid n} \varphi(d).$$

Proof. We know that the greatest common divisor of each element of $[n]$ with n is a divisor of n. Furthermore, for each d such that $d \mid n$, we know that there are exactly $\varphi(n/d)$ many elements of $[n]$ whose gcd with n is d. Since $|[n]| = n$, it follows that
$$n = \sum_{d \mid n} \varphi(n/d).$$

Now we just notice that
$$Div(n) \cap \mathbb{N}^+ = \{n/d : d \in Div(n) \cap \mathbb{N}^+\},$$

(which can be shown by a simple double containment argument), so

$$\sum_{d|n} \varphi(n/d) = \sum_{d|n} \varphi(d).$$

Therefore,

$$n = \sum_{d|n} \varphi(d).$$

□

Notice that the above proposition gives a recursive way to calculate $\varphi(n)$ because

$$\varphi(n) = n - \sum_{\substack{d|n \\ d<n}} \varphi(d).$$

For example, suppose that we want to compute $\varphi(12)$, and that we have already computed $\varphi(k)$ for all k with $1 \leq k \leq 19$. Now $Div(20) \cap \mathbb{N}^+ = \{1, 2, 4, 5, 10, 20\}$, and we're assumging that we already computed the following:

$$\varphi(1) = 1 \qquad \varphi(2) = 1 \qquad \varphi(4) = 2 \qquad \varphi(5) = 4 \qquad \varphi(10) = 4.$$

Then since

$$20 = \varphi(1) + \varphi(2) + \varphi(4) + \varphi(5) + \varphi(10) + \varphi(20),$$

it follows that

$$\begin{aligned} \varphi(20) &= 20 - \varphi(1) - \varphi(2) - \varphi(4) - \varphi(5) - \varphi(10) \\ &= 20 - 1 - 1 - 2 - 4 - 4 \\ &= 20 - 12 \\ &= 8. \end{aligned}$$

6.4 Chinese Remainder Theorem

Can we find an integer $x \in \mathbb{Z}$ that leaves a remainder of 1 when dividing by 6, and leaves a remainder of 8 when divided by 10? Stated in terms of congruences, this question is asking whether there exists an $x \in \mathbb{Z}$ such that both $x \equiv 1 \pmod{6}$ and $x \equiv 8 \pmod{10}$. In other words, does the *system* of congruences

$$x \equiv 1 \pmod{6} \qquad \text{and} \qquad x \equiv 8 \pmod{10}$$

have a solution. In this case, the answer is no. To see this, suppose instead that there was a solution. We can then fix an $x \in \mathbb{Z}$ such both $6 \mid (x-1)$ and $10 \mid (x-8)$. Since 2 is a common divisor of 6 and 10, we then have that both $2 \mid (x-1)$ and $2 \mid (x-8)$, and therefore 2 would divide $(x-1)-(x-8) = 7$, which is a contradiction. Thus, no such x exists. Notice that we used the fact that 6 and 10 had a common divisor larger than 1 in order to gain extra information from the existence of a solution, which led to a contradiction.

Suppose instead that we want to try find an integer $x \in \mathbb{Z}$ that leaves a reminder of 13 when divided by 28, and leaves a remainder of 7 when divided by 33. That is, we want to know if the system of congruences

$$x \equiv 13 \pmod{28} \qquad \text{and} \qquad x \equiv 7 \pmod{33}$$

has a solution. Since $\gcd(28, 33) = 1$, we can not find a contradiction using the same logic as in the previous example. Now through an exhaustive search, it's possible to determine that $x = 601$ is a solution. However, it turns out that there will *always* exist a solution when the two moduli are relatively prime.

Theorem 6.4.1 (Chinese Remainder Theorem - Two Moduli). *Let $m, n \in \mathbb{Z}$ are relatively prime, and let $a, b \in \mathbb{Z}$. There exists $x \in \mathbb{Z}$ such that*

$$x \equiv a \pmod{m} \qquad \textit{and} \qquad x \equiv b \pmod{n}.$$

Furthermore, if $x_0 \in \mathbb{Z}$ is one solution to the above congruences, then an arbitrary $x \in \mathbb{Z}$ is also a solution if and only if $x \equiv x_0 \pmod{mn}$.

Proof 1 of Theorem 6.4.1 - Nonconstructive. Notice that since congruence is transitive, that $\{0, 1, 2, \dots, m-1\}$ is a complete residue system modulo m, and that $\{0, 1, 2, \dots, n-1\}$ is a complete residue system modulo n, we can assume that $a \in \{0, 1, 2, \dots, m-1\}$ and that $b \in \{0, 1, 2, \dots, n-1\}$.

Let $S = \{0, 1, 2, \dots, mn-1\}$. We define a function $f \colon S \to \{0, 1, 2, \dots, m-1\} \times \{0, 1, 2, \dots, n-1\}$ as follows. Given $c \in S$, let $f(c) = (i, j)$ for the unique choice of $i \in \{0, 1, 2, \dots, m-1\}$ and $j \in \{0, 1, 2, \dots, n-1\}$ with $c \equiv i \pmod{m}$ and $c \equiv j \pmod{n}$. We claim that f is injective. To see this, let $c, d \in S$ be arbitrary with $f(c) = f(d)$. Fix $(i, j) \in \{0, 1, 2, \dots, m-1\} \times \{0, 1, 2, \dots, n-1\}$ such that $(i, j) = f(c) = f(d)$. We then have that $c \equiv i \pmod{m}$ and $d \equiv i \pmod{m}$, so since congruence is an equivalence relation, it follows that $c \equiv d \pmod{m}$. Similarly, we have $c \equiv d \pmod{n}$. Thus, we have both $m \mid (c-d)$ and $n \mid (c-d)$, so as $\gcd(m, n) = 1$, we can apply Problem 6 on Homework 4 to conclude that $mn \mid (c-d)$. Now since $c, d \in \{0, 1, 2, \dots, mn-1\}$, we know that $-mn < c - d < mn$. Therefore, it must be the case that $c - d = 0$, and hence $c = d$.

Since f is injective and we have $|S| = mn = |\{0, 1, 2, \dots, m-1\} \times \{0, 1, 2, \dots, n-1\}|$, it follows that f must be surjective as well. Thus, there must exist $x \in S$ with $f(x) = (a, b)$. We then have that $x \equiv a \pmod{m}$ and $x \equiv b \pmod{m}$.

We now verify the last statement. Suppose that x_0 is one solution. Suppose first that $x \equiv x_0 \pmod{mn}$. Since $m \mid mn$ and $n \mid mn$, we then have that $x \equiv x_0 \equiv a \pmod{m}$ and $x \equiv x_0 \equiv b \pmod{n}$, so x is also a solution. Suppose conversely that $x \equiv a \pmod{m}$ and $x \equiv a \pmod{n}$. We then have that $x \equiv x_0 \pmod{m}$ and $x \equiv x_0 \pmod{n}$, so $m \mid (x - x_0)$ and $n \mid (x - x_0)$. Since $\gcd(m, n) = 1$, we can apply Problem 6 on Homework 4, to conclude that $mn \mid (x - x_0)$, and hence $x \equiv x_0 \pmod{mn}$. □

Thus, without exhaustively searching, we know that there does exist an $x_0 \in \mathbb{Z}$ with

$$x_0 \equiv 13 \pmod{28} \qquad \text{and} \qquad x_0 \equiv 7 \pmod{33},$$

and moreover we know that such a unique such x_0 exists with $0 \leq x_0 < 924$ (since $28 \cdot 33 = 924$). Now we could exhaustively search to find that $x_0 = 601$. We can now use the theorem to conclude that any $x \in \mathbb{Z}$ with $x \equiv 601 \pmod{924}$ is also a solution. In other words, any $x \in \mathbb{Z}$ of the form $x = 601 + 924k$ with $k \in \mathbb{Z}$ is also a solution.

Of course, it would be nice to be able to find a solution like 601 without resorting to an exhaustive search. One approach is the following. We know that the solution set of the congruence $x \equiv 13 \pmod{28}$ is the set $\{13 + 28c : c \in \mathbb{Z}\}$. Since each number in this set is a solution to the first congruence, it suffices to find a $c \in \mathbb{Z}$ with

$$13 + 28c \equiv 7 \pmod{33}.$$

Subtracting 13 from both sides (and adding 13 from both sides to go backwards), this is equivalent to finding $c \in \mathbb{Z}$ with

$$28c \equiv -6 \pmod{33},$$

which is equivalent to finding $c \in \mathbb{Z}$ with

$$28c \equiv 27 \pmod{33}.$$

Now $\gcd(28, 33) = 1$, so since $1 \mid 27$, we can use Theorem 6.1.15 to conclude that there exists $c \in \mathbb{Z}$ making this last congruence true. How would we actually compute it? We would first use the Euclidean Algorithm to find $k, \ell \in \mathbb{Z}$ with $28k + 33\ell = 1$, and then since we want 27 instead of 1, we could multiply through by 27 to find that $28 \cdot (27k) + 33 \cdot (27\ell) = 27$. Thus, we have $28 \cdot (27k) \equiv 27 \pmod{33}$, so we can take $c = 27k$, and hence can take $x_0 = 13 + 28 \cdot 27 \cdot k$.

Although this certainly works, it turns out that we can streamline the core ideas above so that we can do one computation based on m and n, and then use it to solve the system for any a and b. The key idea is to think about solving the system with the simplest possible choices of a and b. Of course, $a = 0 = b$ is the easiest, but it is too trivial. Suppose then that $\gcd(m, n) = 1$, and we're trying to solve

$$x \equiv 1 \pmod{m} \qquad \text{and} \qquad x \equiv 0 \pmod{n}.$$

As in the previous example, since $\gcd(m, n) = 1$, we can use the Euclidean Algorithm to find $k, \ell \in \mathbb{Z}$ with $km + \ell n = 1$. Notice then $m \mid \ell n - 1$, so $\ell n \equiv 1 \pmod{m}$. We also have $n \mid \ell n$ trivially, so $\ell n \equiv 0 \pmod{n}$. Thus, ℓn is a solution to the above system. Notice that km is a solution to

$$x \equiv 0 \pmod{m} \qquad \text{and} \qquad x \equiv 1 \pmod{n}$$

by a completely analogous argument. Thus, we've solved the two most basic systems using just one computation involving the Euclidean Algorithm on m and n. What if we wanted to solve the system

$$x \equiv 3 \pmod{m} \qquad \text{and} \qquad x \equiv 0 \pmod{n}?$$

Since we can multiply both sides of a congruence by 3, we should of course just take 3 times the solution to our original congruence! In other words, $3\ell n$ should be a solution. And if we wanted to solve

$$x \equiv 0 \pmod{m} \qquad \text{and} \qquad x \equiv 7 \pmod{n},$$

then we would just take 7 times the solution to our second, i.e. $7km$ is a solution. Finally, it seems reasonable to hope that if we wanted to solve

$$x \equiv 3 \pmod{m} \qquad \text{and} \qquad x \equiv 7 \pmod{n},$$

then we can just take 3 times our first solution plus 7 times our second. Thus, we might hope that $3\ell n + 7km$ is a solution, and we now check that this is indeed the case in our second proof.

Proof 2 of Theorem 6.4.1 - Constructive. Since $\gcd(m, n) = 1$, we may fix $k, \ell \in \mathbb{Z}$ with $km + \ell n = 1$. We then have $n \mid mk - 1$ and $m \mid n\ell - 1$, so $km \equiv 1 \pmod{n}$ and $\ell n \equiv 1 \pmod{m}$. Let $x_0 = a\ell n + bmk$. We check that x_0 satisfies the above congruences:

- Since $\ell n \equiv 1 \pmod{m}$, we have $a\ell n \equiv a \pmod{m}$. Now $bkm \equiv 0 \pmod{m}$, and by adding these congruences we conclude that $x_0 \equiv a \pmod{m}$.
- Since $km \equiv 1 \pmod{n}$, we have $bkm \equiv b \pmod{n}$. Now $a\ell n \equiv 0 \pmod{n}$, and by adding these congruences we conclude that $x_0 \equiv b \pmod{n}$.

The verification of the last statement is identical to the proof above. □

For example, suppose that we want to find all $x \in \mathbb{Z}$ which simultaneously satisfy

$$x \equiv 3 \pmod{14} \qquad \text{and} \qquad x \equiv 8 \pmod{9}.$$

Notice that $2 \cdot 14 + (-3) \cdot 9 = 1$ which can be found by inspection or by the Euclidean Algorithm:

$$\begin{aligned} 14 &= 1 \cdot 9 + 5 \\ 9 &= 1 \cdot 5 + 4 \\ 5 &= 1 \cdot 4 + 1 \\ 4 &= 4 \cdot 1 + 0. \end{aligned}$$

Therefore, working backwards, we have

$$\begin{aligned}1 &= 5 - 4\\ &= 5 - (9 - 5)\\ &= (-1) \cdot 9 + 2 \cdot 5\\ &= (-1) \cdot 9 + 2 \cdot (14 - 9)\\ &= 2 \cdot 14 + (-3) \cdot 9.\end{aligned}$$

The proof of the Chinese Remainder Theorem lets $x_0 = 3 \cdot (-3) \cdot 9 + 8 \cdot 2 \cdot 14 = 143$. Thus, the complete solution set is the set of all $x \in \mathbb{Z}$ that satisfy $x \equiv 143 \pmod{126}$, i.e. all $x \in \mathbb{Z}$ such that $x \equiv 17 \pmod{126}$.

We can extend the Chinese Remainder Theorem to the case of more than two congruences. However, we first need to make an important distinction. Given a triple of natural numbers (a, b, c), it is natural to call the triple relatively prime if the only positive natural number that divides all three is 1. With this definition, the triple $(6, 10, 15)$ is relatively prime. However, notice that any pair of distinct numbers in this triple is *not* relatively prime, as we have $\gcd(6, 10) = 2$, $\gcd(6, 15) = 3$, and $\gcd(10, 15) = 5$. We can not extend the Chinese Remainder Theorem to the situation where the whole triple is relatively prime, as congruence modulo 6 will affect congruence modulo 10 (for example, there is no number congruent to 0 modulo 6 and congruent to 1 modulo 10). Thus, we need to adopt the stronger assumption that our moduli are *pairwise* relatively prime, i.e. that any pair of distinct moduli are relatively prime.

Theorem 6.4.2 (Chinese Remainder Theorem)**.** *Suppose that $m_1, m_2, \dots, m_\ell \in \mathbb{Z}$ are pairwise relatively prime (i.e. $\gcd(m_i, m_j) = 1$ whenever $i \neq j$) and $a_1, a_2, \dots, a_\ell \in \mathbb{Z}$. There exists $x \in \mathbb{Z}$ such that*

$$x \equiv a_i \pmod{m_i}$$

for all i. Furthermore, if $x_0 \in \mathbb{Z}$ is one solution to the above congruences, then an arbitrary $x \in \mathbb{Z}$ is also a solution if and only if $x \equiv x_0 \pmod{m_1 m_2 \cdots m_\ell}$.

Proof. There are many possible arguments. One is to use induction together with the two moduli version. Another is to adapt the first nonconstructive proof by defining a corresponding function from

$$\{0, 1, \dots, m_1 m_2 \cdots m_\ell - 1\}$$

to

$$\{0, 1, \dots, m_1 - 1\} \times \{0, 1, \dots, m_2 - 1\} \times \cdots \times \{0, 1, \dots, m_\ell - 1\},$$

and use the pairwise relatively prime assumption to argue that the function is injective. Finally, it is also possible to adapt the last constructive proof. I leave the details to you. □

6.5 Primality Testing

Suppose that we have a number $m \in \mathbb{N}$ with $m \geq 2$, and we want to determine whether it is prime. The naive method is to exhaustively check each number $2, 3, 4, \dots, m - 1$ in turn to see if any of them divide m. If we find such a number, then m is not prime, while if we exhaust the entire list without finding such a divisor, then m is prime. Although this certainly works, the procedure is very slow on moderately sized inputs. For example, if m has about 10 decimal digits, then this procedure checks about 10^{10}, i.e. about 10 billion, numbers. We can improve this brute-force search somewhat by using Proposition 5.4.1, which says that it suffices to check each number $2, 3, 4, \dots, \lfloor \sqrt{m} \rfloor$ in turn to see if any of them divide m. If m has a about 10 decimal digits, then this procedure checks about $\sqrt{10^{10}} = (10^{10})^{1/2} = 10^5$ many numbers, which is certainly feasible with modern computers. We can even do a bit better by cutting out every even number except 2, bringing it down to about $\frac{1}{2} \cdot 10^5$.

Even with these improvements, such a method is still quite slow. For example, if m has 40 decimal digits, then we need to check about 10^{20} (or maybe $\frac{1}{2} \cdot 10^{20}$ if we omit the evens) many numbers, and such a procedure can not be carried out in any reasonable amount of time on any modern computer.

To start thinking about other approaches, we turn to some of the theoretical results in this section. We start with Fermat's Little Theorem, which implies that if m is prime, and if $a \in \mathbb{N}$ with $0 < a < m$ (so $m \nmid a$), then $a^{m-1} \equiv 1 \pmod{m}$. Thus, if we have a natural number m, and we can find an $a \in \mathbb{N}$ with $0 < a < m$ such that $a^{m-1} \not\equiv 1 \pmod{m}$, then m is not prime. In order to turn this into an efficient test, we need to be able to compute powers of a modulo m quickly. In the past, we looked for a repeat in the early powers to find a pattern, or used Fermat's Little Theorem/Euler's Theorem to know that powers will eventually cycle back around to 1. However, in those cases, the power was significantly larger than the modulus. Here, the power is one less than the modulus, and there is no guarantee that patterns will emerge quickly (and in fact they might not).

So if m is large, and we have an a with $0 < a < m$, how can we determine whether $a^{m-1} \equiv 1 \pmod{m}$ quickly? If we just repeatedly multiply by a, then we have to do $m-2$ many multiplications, which is prohibitive. However, it turns out that we can do better. Let's start by thinking about how to compute a^4. Of course,

$$a^4 = a \cdot (a \cdot (a \cdot a)),$$

so we can compute a^4 with 3 multiplications. However, there is a simple way to improve this. We first compute a^2 using one multiplication, and then we notice that $a^4 = a^2 \cdot a^2$, so we only need one more multiplication. From here, we can compute a^8 with just one more multiplication. In general, by repeated squaring, we can compute a^{2^n} using just n multiplications. But what if we do not have a power of 2 in the exponent? For example, suppose that we want to compute a^{41}? By repeated squaring, we compute the following:

$$a^1, a^2, a^4, a^8, a^{16}, a^{32}.$$

Now we notice that $41 = 32 + 8 + 1$, so

$$a^{41} = a^{32} \cdot a^8 \cdot a^1,$$

and hence we can obtain a^{41} using just two more multiplications once we have the powers of 2. In general, to compute a^n, we can write n in base 2 (where $41 = 101001$), square repeatedly until we reach the largest power of 2 less than n, and multiply the powers of 2 that correspond to the positions of the 1's in the base 2 representation. Instead of using $n-1$ many multiplications, this process uses roughly $2 \cdot \log_2(n)$ many multiplications, which is exponentially better.

Although we can compute a^n with many fewer multiplications, notice that the numbers a^n become large very quickly. However, if we want to know whether $a^{m-1} \equiv 1 \pmod{m}$, we can always keep reducing the values that we compute to an equivalent value in $\{0, 1, 2, \dots, m-1\}$. In other words, we know that $\{0, 1, 2, \dots, m-1\}$ is a compute residue system modulo m, so we can define a function $f_m \colon \mathbb{Z} \to \{0, 1, 2, \dots, m-1\}$ by letting $f_m(b)$ equal the unique element of $\{0, 1, 2, \dots, m-1\}$ that is congruent to b modulo m. With this notation, asking whether $a^{m-1} \equiv 1 \pmod{m}$ is equivalent to asking whether $f_m(a^{m-1}) = 1$. Notice that f_m has the following properties:

- $f_m(b) = b$ for all $b \in \{0, 1, 2, \dots, m-1\}$
- $f_m(b) = f_m(f_m(b))$ for all $b \in \mathbb{Z}$.
- $f_m(bc) = f_m(f_m(b) \cdot f_m(c))$ for all $b, c \in \mathbb{Z}$ (by using Proposition 6.1.12).

Intuitively, as we are computing powers of a, we can "reduce" any value that is greater than $m-1$ to a value in the set $\{0, 1, 2, \dots, m-1\}$. In this way, we can avoid dealing with the very large numbers a^n directly.

Using all of these ideas, we can describe a very efficient recursive algorithm for computing $f_m(a^n)$ that avoids dealing with base 2 representations of n, and avoids working with the large values a^n directly. We assume that we have an efficient way to compute the function f_m, which takes an input, and outputs the remainder upon division by m (the naive algorithm is sufficiently fast for our purposes). With that in hand, given $m \in \mathbb{N}^+$, $a \in \mathbb{Z}$, and $n \in \mathbb{N}$, here is how we compute $f_m(a^n)$:

- If $n = 0$, then output 1.
- If n is even, then find $k \in \mathbb{N}$ with $n = 2k$, recursively determine $f_m(a^k)$, and output $f_m(f_m(a^k) \cdot f_m(a^k))$.
- If n is odd, then recursively determine $f_m(a^{n-1})$, and output $f_m(f_m(a) \cdot f_m(a^{n-1}))$.

Using the above properties of f_m, it is straightforward to show that this algorithm correctly computes $f_m(a^n)$ by induction on n. Notice that after 2 steps of the algorithm, we have reduced n by at least a factor of 2, so the process will terminate in at most $2 \cdot \log_2(n)$ many steps.

Now that we have this method in hand, we can return to thinking about a primality test. Fermat's Little Theorem tells us that if m is prime, and $0 < a < m$, then $f_m(a^{m-1}) = 1$. Thus, given an $m \in \mathbb{N}^+$, the idea is to pick some value of a with $0 < a < m$, compute $f_m(a^{m-1})$ using the above algorithm, and check whether the output is 1. If not, then m is not prime. For example, we can compute that $f_{91}(2^{90}) = 64$ (i.e. that $2^{90} \equiv 64 \pmod{91}$), so 91 is not prime. Notice that although this computation verifies that 91 is not prime, it does *not* give us a way to factor 91. Moreover, we need to be very careful about how we interpret these computations. *If* we find an a with $0 < a < m$ such that $f_m(a^{m-1}) \neq 1$, *then* m is not prime. However, if we find an a with $0 < a < m$ such that $f_m(a^{m-1}) = 1$, then we can *not* conclude that a is prime. For example, it turns out that $f_{91}(3^{90}) = 1$, but 91 is not prime (as we just saw).

By Problem 4 on Homework 13, we know that if m is not prime, then there must exist *some* choice of a with $0 < a < m$ such that $f_m(a^{m-1}) \neq 1$. Thus, a naive idea is to try all possible a with $0 < a < m$. However, there are $m - 2$ many such a, and so this procedure is no better than exhaustively searching for divisors. Instead, we can think about trying several *random* values of a. Now if any of these test values of a come back with $f_m(a^{m-1}) \neq 1$, then we know that m is not prime. But what if they all come back equal to 1? We might think that we have strong evidence that m might be prime, but in order to quantify this, we are led to the following question: If m is not prime, what fraction of the $a \in \{1, 2, 3, \dots, m-1\}$ have the property that $f_m(a^{m-1}) = 1$? Unfortunately, it turns out that there are some composite numbers, known as Carmichael numbers, where the vast majority of the values of $a \in \{1, 2, 3, \dots, m-1\}$ have the property that $f_m(a^{m-1}) = 1$. For these numbers, it turns out that $f_m(a^{m-1}) = 1$ for all values of a with $0 < a < m$ that satisfy $\gcd(a, m) = 1$. And, in general, the fraction of values of a with $0 < a < m$ that satisfy $\gcd(a, m) \neq 1$ can be quite small.

With that in mind, we try to enhance our above procedure. Fermat's Little Theorem provided the structure for our previous test, and now we look to another result about prime numbers. Recall that if p is prime, then Proposition 6.2.3 says that $b^2 \equiv 1 \pmod{p}$ if and only if either $b \equiv 1 \pmod{p}$ or $b \equiv -1 \pmod{p}$. Suppose then that $m \geq 3$ is odd, and we have found an $a \in \{1, 2, 3, \dots, m-1\}$ with $f_m(a^{m-1}) = 1$, i.e. with $a^{m-1} \equiv 1 \pmod{m}$. Since m is odd, we know that $m - 1$ is even, so $2 \mid m - 1$. Consider the value $a^{\frac{m-1}{2}}$. Notice that

$$\left(a^{\frac{m-1}{2}}\right)^2 \equiv 1 \pmod{m}.$$

Now if m is prime, then Proposition 6.2.3 says that we must have either

$$a^{\frac{m-1}{2}} \equiv 1 \pmod{m} \quad \text{ or } \quad a^{\frac{m-1}{2}} \equiv -1 \pmod{m},$$

i.e. we must have either $f_m(a^{\frac{m-1}{2}}) = 1$ or $f_m(a^{\frac{m-1}{2}}) = m-1$. Thus, instead of only checking that $f_m(a^{m-1}) = 1$, we can perform another check in this case to test whether $f_m(a^{\frac{m-1}{2}}) \in \{1, m-1\}$. If not, then m is not prime.

For example, suppose that we are trying to determine whether 91 is prime. If we choose $a = 3$ as our test case, then we noted above that $f_{91}(3^{90}) = 1$, so 3 passes this simple test. However, if we compute $f_{91}(3^{45})$, we see that $f_{91}(3^{45}) = 27$. From here we can immediately conclude that $27^2 \equiv 1 \pmod{91}$, so since $27 \not\equiv 1 \pmod{91}$ and $27 \not\equiv -1 \pmod{91}$, we can use Proposition 6.2.3 to conclude that 91 is not prime.

Now if $f_m(a^{\frac{m-1}{2}}) = 1$, and $\frac{m-1}{2}$ is even, then we can continue the process of dividing by 2. For example, consider $m = 561$, which happens to be the first Carmichael number. We have $f_{561}(2^{560}) = 1$, and if we divide 560 by 2, we can also compute that $f_{561}(2^{280}) = 1$. At this point, we have no evidence from either Fermat's Little Theorem or Proposition 6.2.3 that 561 is not prime. However, if we divide the exponent in 2 again, we can compute that $f_{561}(2^{140}) = 67$. As a result, we must have $67^2 \equiv 1 \pmod{561}$, but $67 \not\equiv 1 \pmod{561}$ and $67 \not\equiv -1 \pmod{561}$, and hence 561 is not prime. Therefore, if we divide the exponent by enough factors of 2, we can sometimes find a violation of Proposition 6.2.3 using a value of a that passes the test imposed by Fermat's Little Theorem.

Instead of computing $f_m(a^{m-1})$ and working backwards by dividing the exponent repeatedly by 2, we can instead work in the other direction and repeatedly square up to the exponent $m-1$. We first need to figure out how many times we can divide $m-1$ by 2. To that end, we perform the following procedure. Suppose that $m \geq 3$ is odd. Fix $k, d \in \mathbb{N}$ with $m - 1 = 2^k \cdot d$, where d is odd (see Lemma 4.3.7). Notice that since m is odd, we know that $m-1$ is even, and hence $k \geq 1$. Furthermore, we can compute k and d by repeatedly dividing $m-1$ by 2, as long as the result is even. Now given any $a \in \{1, 2, 3, \dots, m-1\}$, we first compute $f_m(a^d)$, and then repeatedly square the result, reducing to a value in $\{0, 1, 2, \dots, m-1\}$ as necessary, to form the sequence:

$$f_m(a^d),\ f_m(a^{2d}),\ f_m(a^{2^2 d}),\ \dots\ ,\ f_m(a^{2^{k-1} d}),\ f_m(a^{2^k d}).$$

Notice that the last value in this sequence is simply $f_m(a^{m-1})$.

Suppose now that $m \geq 3$ is an odd prime, and $a \in \{1, 2, 3, \dots, m-1\}$. By Fermat's Little Theorem, the last element of this sequence will be 1. Now previous elements of this sequence might also be 1. However, if there is an element that does not equal 1, then the *last* element of the sequence that is not a 1 must be $m-1$ by Proposition 6.2.3.

With this in mind, we form the following primality test. Let $m \in \mathbb{N}$ with $m \geq 3$ be odd (it is trivial to check if an even number is prime). Compute $k, d \in \mathbb{N}$ with $m - 1 = 2^k \cdot d$, where d is odd. Pick some value $a \in \{1, 2, 3, \dots, m-1\}$ to use as a test, and start computing the sequence:

$$f_m(a^d),\ f_m(a^{2d}),\ f_m(a^{2^2 d}),\ \dots\ ,\ f_m(a^{2^{k-1} d})$$

(notice that we cut off the last value). We proceed down the sequence with the following rules. We begin by examining the first element:

- If we find that $f_m(a^d) = 1$, then a passes the test (notice that all elements of the original sequence will then equal 1 because $1^2 \equiv 1 \pmod{m}$).
- If we find that $f_m(a^d) = m - 1$, then a passes the test (notice that the rest of the elements of the original sequence will then equal 1 because $(m-1)^2 \equiv (-1)^2 \equiv 1 \pmod{m}$).

Suppose then that $f_m(a^d) \notin \{1, m-1\}$. Work down the sequence in order by repeated squaring, reducing to a value in $\{0, 1, 2, \dots, m-1\}$ as necessary, while looking for either $m-1$ or 1.

- If we find the value $m-1$ at any point in the truncated sequence, then a passes the test (notice that the rest of the elements of the sequence will then equal 1 because $(m-1)^2 \equiv (-1)^2 \equiv 1 \pmod{m}$).
- If we instead find the value 1 before discovering $m-1$, then a fails the test, as we have found a value $b \in \{2, 3, \dots, m-2\}$ with $b^2 \equiv 1 \pmod{m}$, so m is not prime by Proposition 6.2.3.

Finally, if we make it through the entire truncated sequence up to $f_m(a^{2^{k-1} d})$ without finding the value 1 or $m-1$, then a fails the test. Notice that in this situation, either $f_m(a^{2^k d}) \neq 1$, demonstrating that m is not prime by Fermat's Little Theorem, or $f_m(a^{2^k d}) = 1$, in which case $f_m(a^{2^{k-1} d})$ serves as a value distinct from 1 and $m-1$ whose square is 1 modulo m, so m is not prime by Proposition 6.2.3.

Although it takes a bit of thought to work through all of the cases, notice that if m is prime, then *every* $a \in \{1, 2, 3, \ldots, m-1\}$ will pass this test. Thus, if we find a value of a that fails this test, then we can conclude with certainty that m is not prime. Also, if a fails the Fermat test, then it is also fails this test. The question is whether *enough* values of a will fail the test when m is not prime. That is, are there analogues of Carmichael numbers here? The answer lies in the following result.

Theorem 6.5.1 (Monier, Rabin)**.** *For each odd composite $m \in \mathbb{N}^+$ with $m > 9$, the number of values $a \in \{1, 2, 3, \ldots, m-1\}$ that pass the above test is at most $\frac{1}{4}\varphi(m)$.*

In particular, since we trivially have $\varphi(m) \leq m$, at most 25% of the test values will pass when $m > 9$ is an odd composite number. How do we find one of the (at least) 75% that work? The idea is pick several test values of *a at random*, and try them. If any fail the test, then we know with certainty that m is composite. If they all pass, then we are very confident that m is prime. Moreover, it is possible to precisely quantify how confident we are, assuming that we are really are using random test values. In fact, by using just 20 test values, we can be (much) more that 99.999999999% confident that the the number is prime. This randomized algorithm is known as the *Miller-Rabin primality test*.

6.6 Cryptography

The basic goal of cryptography is to encode messages in such a way that only the intended recipients are able to decipher the original contents. Mathematically, we can view an encoding as a function $h \colon A \to B$, where A is the set of possible messages that we might send, and B is the set of possible encodings. Typically, we use the same "alphabet", so we work in the case where $A = B$. For example, we might let A and B both be the set $\{\text{a,b,c}, \ldots, \text{z}\}^*$ of all finite sequences from our usual collection of letters. We can include numbers or special characters as well, or we might do something as simple as letting $A = B = \{0, 1\}^*$ be the set of all finite sequences of 0's and 1's (since we can code any reasonable object in such a way - see modern computers). We might also choose to work with finite sets, like $\{\text{a,b,c}, \ldots, \text{z}\}^{100}$ or $\{0, 1\}^{256}$ for simplicity.

Suppose then that we have a set A, and we want to devise an encoding function $h \colon A \to A$. What properties should h have? We certainly need h to be injective, because it must be possible to unambiguously decipher an encoded message. Now if A is a finite set, then any injective $h \colon A \to A$ is bijective. Assuming then that we work with a finite set A, the following properties of h would be desirable in any encoding scheme:

1. h should be reasonably fast to compute.

2. h^{-1} should be reasonably fast to compute for the intended recipient.

3. h^{-1} should be very difficult to determine (or very hard to compute) for any eavesdropper.

4. h should look "random". Ideally, if we change the input a little, the output might change dramatically.

As an initial basic example, let $C = \{\text{a,b,c}, \ldots, \text{z}\}$ be the letters of the standard alphabet, and let $A = C^{100}$ be the set of sequences of length 100 from this alphabet. Fix a bijection $g \colon C \to C$, to serve as a way of mixing up the letters. Define $h \colon A \to A$ by letting $h(c_1 c_2 \ldots c_{100}) = g(c_1)g(c_2) \cdots g(c_{100})$, i.e. we simply apply the fixed bijection of the characters to each character in turn. For instance, we might choose $g \colon C \to C$ to be a cyclic shift by 3, so $g(\text{a}) = \text{d}$, $g(\text{b}) = \text{e}$, $g(\text{c}) = \text{f} \ldots$, $g(\text{x}) = \text{a}$, $g(\text{y}) = \text{b}$, and $g(\text{z}) = \text{c}$. For example, we have

$$\begin{aligned} h(\text{mathrulez}) &= g(\text{m})g(\text{a})g(\text{t})g(\text{h})g(\text{r})g(\text{u})g(\text{l})g(\text{e})g(\text{z}) \\ &= \text{pdwkuxohc}. \end{aligned}$$

This simple encryption scheme is known as the Caesar cipher. Although it works against your average 3-year old, it is not terribly secure.

More generally, we might start with a random looking bijection $g\colon C \to C$ (not just a simple shift). Since there are 26! many such bijections, it might seem reasonable to believe that such a random looking bijection might provide a robust encryption scheme. If you code g and g^{-1} as simple look up tables, then with those in hand, both h and h^{-1} are reasonably fast to compute. Notice that the fourth condition above fails, as changing one letter of the input will only change one letter of the output, but you might not find that essential. However, there is a much more serious problem with this system. Although one might naively believe that somebody intercepting the encoded message would have to try all 26! many bijections to determine which one was used, it turns out that there are other attacks on this system. In fact, these encryption schemes are typically used as puzzles in newspapers. The basic idea is to use *frequency analysis*. In natural languages like English, some letters occur more frequently than others. Thus, with a sufficiently long encoded message, or with many short encoded messages, it is possible to mount an attack by noticing the most frequently occurring letter, and guessing that it might be the encoding of the letter e. From here, an attacker can make other educated guesses, and slowly learn the function g.

Historically, the response to this frequency analysis was to change the function g based on the position. For example, we might have three different bijections $g_1, g_2, g_3\colon C \to C$, and then we would define $h\colon A \to A$ by letting

$$h(c_1c_2c_3c_4c_5c_6 \dots c_{100}) = g_1(c_1)g_2(c_2)g_3(c_3)g_1(c_4)g_2(c_5)g_3(c_6)\cdots g_3(c_{99})g_1(c_{100}).$$

Although this makes frequency analysis more difficult, if an eavesdropper knows (or guesses) that the bijections cycle with a period of 3, then it is possible to mount a frequency analysis attack by pulling out the corresponding letters. Of course, it's conceivable that one could use 100 different bijections, but then it becomes difficult to keep track of it all, and hence to compute h and h^{-1} quickly. In order to accomplish this task, machines like the Enigma were built that mechanically cycled through thousands of bijections. In a given state, the Enigma machine encoded a bijection g, but once a key was pressed, certain rotors in the machine turned so that the machine now encoded a different bijection. In order to make this scheme work, it was helpful to set up the machine in such a way that each encoded g was its own inverse, so that if somebody typed the encoded message into a machine with the same initial setup, then the original message would be produced. Now all that is needed is a way for the intended recipient to know the initial state of the message, and then they would be able to type in the encoded message in order to compute the inverse.

Although the above system seems reasonably secure, it does have some issues. The fourth condition above still fails, as changing one letter of the input will still only change one letter of the output. Moreover, a system would need to be devised in order for the intended recipient to know the initial state of the encoding machine (so that they can decode the message). Elaborate books of such initial states of machines were produced during World War II in order to accomplish this task. So one way to defeat the system is to steal the codebooks that describe the initial states of the machines on a given day. Beyond that, there are other sophisticated attacks that one can mount by exploiting some weaknesses (like the fact that each g was its own inverse). Unfortunately, working out the cryptanalysis would take us too far afield.

Modern encryption schemes avoid working character by character in isolation in order to bypass these issues. In this way, we can develop schemes that can satisfy the fourth condition above. Thus, we want to define a bijection $h\colon A \to A$ all at once, without breaking it into small pieces. At first, it seems difficult to think about how to define such a bijection that can be computed quickly, but which is also nontrivial. The fundamental idea is to use modular arithmetic. That is, instead of working with a set like $A = \{\text{a,b,c}, \dots, \text{z}\}^{100}$, we will work with a set like $A = \{0, 1, 2, \dots, m-1\}$ for some large natural number m. For example, if we pick $m = 2^{500}$, then we can code code any short to medium length message using a number. For instance, if we use a simple binary encryption scheme by letting $a \mapsto 00000$, $b \mapsto 00001$, etc., then any element of $\{\text{a,b,c}, \dots, \text{z}\}^{100}$ can be coded by a sequence of 500 zeros and ones, which in binary codes a number less than 2^{500}.

Suppose then that we have settled on a way to code our messages as numbers, and that we have picked a large enough value of $m \in \mathbb{N}^+$ so that all of our possible messages are in the set $A = \{0, 1, 2, \dots, m-1\}$. We now want to define a bijection $h \colon A \to A$ to serve as our encryption scheme. As above, define $f_m \colon \mathbb{Z} \to \{0, 1, 2, \dots, m-1\}$ by letting $f_m(b)$ equal the unique element of $\{0, 1, 2, \dots, m-1\}$ that is congruent to b modulo m. The simplest idea is to adapt to Caeser cipher, except operating on the entire message at once (rather than character by character). Thus, we pick some $k \in A$ to be our shift amount, and define $h \colon A \to A$ by letting $h(a) = f_m(a + k)$, i.e. we shift by k, and cycle back around as necessary. We then have that h is a bijection since its inverse is given by $g \colon A \to A$ defined by $g(a) = f_m(a - k)$. Although this simple scheme bypasses an attack based on frequency analysis, it is terrible, so we won't say more about it.

How can we define a more interesting bijection? Instead of performing modular addition, we can perform modular multiplication. That is, we some $k \in \mathbb{Z}$, and define $h \colon A \to A$ by letting $h(a) = f_m(ka)$. Unlike modular addition, not all such functions are bijections. For example, if $m = 6$ and $k = 2$, then we have

$$h(0) = 0 \qquad h(1) = 2 \qquad h(2) = 4 \qquad h(3) = 0 \qquad h(4) = 2 \qquad h(5) = 4.$$

Thus, we need to think about what values of k will produce a bijection. For injectivity, we can rephrase the question in terms of modular arithmetic: Which values of $k \in \mathbb{Z}$ have the property that whenever $ka \equiv kb \pmod{m}$, we have $a \equiv b \pmod{m}$? From this perspective, Proposition 6.1.9 immediately tells us that whenever $\gcd(k, m) = 1$, the function h is injective. Suppose then that we choose $m, k \in \mathbb{N}^+$ with $\gcd(k, m) = 1$, and use this as our encryption scheme. How do we decrypt? That is, given ka, how do we determine the original a? The answer lies in Corollary 6.1.16 and the discussion afterwards. Since $\gcd(k, m) = 1$, we can fix $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{m}$, and we can compute such an ℓ using the Euclidean Algorithm (and shifting it to be in $\{0, 1, 2, \dots, m-1\}$, if necessary). The function $g \colon A \to A$ defined by $g(b) = \ell b$ is an inverse for h because for all $a \in \mathbb{Z}$, we have

$$\begin{aligned} g(h(a)) &\equiv \ell k a \pmod{m} \\ &\equiv 1 \cdot a \pmod{m} \\ &\equiv a \pmod{m} \end{aligned}$$

and

$$\begin{aligned} h(g(a)) &\equiv k \ell a \pmod{m} \\ &\equiv 1 \cdot a \pmod{m} \\ &\equiv a \pmod{m}. \end{aligned}$$

Taken together, here is the total encryption scheme:

- Fix some $m \in \mathbb{N}^+$ that is large enough to code any message you want to send. Then pick some $k \in \mathbb{N}^+$ such that $\gcd(k, m) = 1$. We can find such a k through trial and error using the Euclidean Algorithm.
- To encode a message, compute $h(a) = f_m(ka)$, i.e. $h(a)$ is the unique element of A that is congruent to ka modulo m.
- For decoding, we first compute an $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{m}$, which we can find through the Euclidean Algorithm. Then, given an encoded message, we compute $g(b) = f_m(\ell b)$ to decode the message.

This scheme works reasonably well. It doesn't "jumble" the contents particularly well (if you increase a by 1, then the output always increases by k), but it does a decent job in other regards. Of course, as in previous schemes, the two parties need to agree on both m and k in advance. If anybody is able to get their hands on m and k, then they can perform the Euclidean Algorithm to determine ℓ, enabling them to decode messages.

In order to finally deal with the fourth condition, we turn to functions defined by modular exponentiation. As above, our first question is the following: given m and k, when is the function $h(a) = f_m(a^k)$ a bijection? In order to get a handle on this question, we start by looking at the simplest case, which is when m is prime. Looking at the possible values in the case where $m = 11$, we see that the values of $k \in \{1, 2, \ldots, 11\}$ that give a bijection are precisely the values of k in the set $\{1, 3, 7, 9\}$. Notice that these are the precisely the elements of $[10]$ that are relatively prime to 10. Moreover, it turns that the inverse of $h(a) = f_m(a^3)$ is the function $g(a) = f_m(a^7)$. To see this, notice that for any $a \in \mathbb{Z}$ with $11 \nmid a$, we have

$$\begin{aligned} g(h(a)) &\equiv (a^3)^7 \pmod{11} \\ &\equiv a^{21} \pmod{11} \\ &\equiv a \cdot a^{20} \pmod{11} \\ &\equiv a \cdot (a^{10})^2 \pmod{11} \\ &\equiv a \cdot 1^2 \pmod{11} \\ &\equiv a \pmod{11}, \end{aligned}$$

where the second to last line follows from Fermat's Little Theorem. Moreover, if $a \in \mathbb{Z}$ and $11 \mid a$, then we trivially have $11 \mid a^{21}$, and so $a^{21} \equiv a \pmod{11}$. We now generalize this example.

Proposition 6.6.1. *Let $p \in \mathbb{N}^+$ be prime and let $A = \{0, 1, 2, \ldots, p-1\}$. Let $k \in \mathbb{N}^+$ with $\gcd(k, p-1) = 1$. The function $h\colon A \to A$ given by $h(a) = f_p(a^k)$ is a bijection. Moreover, if $\ell \in \mathbb{N}^+$ is such that $k\ell \equiv 1 \pmod{p-1}$, then the function $g\colon A \to A$ given by $g(b) = f_p(b^\ell)$ is an inverse of h.*

Proof. Suppose that $\gcd(k, p-1) = 1$. By Corollary 6.1.16, we can fix $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{p-1}$ (notice that we choose a positive such ℓ by finding an equivalent value in the set $\{1, 2, \ldots, p-1\}$). By definition, have $p-1 \mid k\ell - 1$, so we can fix $n \in \mathbb{Z}$ with $(p-1)n = k\ell - 1$. Since $k, \ell \in \mathbb{N}^+$, we have $k\ell - 1 \geq 0$, and hence $n \geq 0$. Notice that $k\ell = 1 + (p-1)n$. Now for any $a \in \mathbb{Z}$ with $p \nmid a$, we have

$$\begin{aligned} g(h(a)) &\equiv (a^k)^\ell \pmod{p} \\ &\equiv a^{k\ell} \pmod{p} \\ &\equiv a^{1+(p-1)n} \pmod{p} \\ &\equiv a \cdot a^{(p-1)n} \pmod{p} \\ &\equiv a \cdot (a^{p-1})^n \pmod{p} \\ &\equiv a \cdot 1^n \pmod{p} \\ &\equiv a \pmod{p}, \end{aligned}$$

where the second to last line follows from Fermat's Little Theorem. Moreover, for any $a \in \mathbb{Z}$ with $p \mid a$, we trivially have $p \mid a^{k\ell}$, so $a^{k\ell} \equiv a \pmod{p}$, and hence $g(h(a)) \equiv a \pmod{p}$. A completely analogous argument shows that $h(g(a)) = a$. Therefore, g is an inverse of h, and so h is bijective. □

Taken together, here is the total encryption scheme based on modular exponentiation relative to a prime modulus:

- Fix some prime $p \in \mathbb{N}^+$ that is large enough to code any message you want to send. We can find such a large prime using the methods of the previous section. Then pick some $k \in \mathbb{N}^+$ such that $\gcd(k, p-1) = 1$. We can find such a k through trial and error using the Euclidean Algorithm.
- To encode a message, compute $h(a) = f_p(a^k)$, i.e. $h(a)$ is the unique element of A that is congruent to a^k modulo p. Notice that we can do exponentiation modulo p quickly (as described in the previous section).

- For decoding, we first compute an $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{p-1}$, which we can find through the Euclidean Algorithm. Then, given an encoded message, we compute $g(b) = f_p(b^\ell)$ to decode the message.

This encryption scheme works nicely, and does a good job of looking "random". Notice that small changes in the input can produce dramatically different outputs. However, as above, both parties need to agree on both p and k in advance. If anybody is able to get their hands on p and k, then they can perform the Euclidean Algorithm to determine ℓ, enabling them to decode messages.

Can we generalize the above construction to moduli that are not prime? At first, such a generalization might just appear to be an unnecessary complication. However, we will see that the added generality makes an enormous qualitative difference. When we move to a modulus that is not prime, it is natural to try to use Euler's Theorem in place of Fermat's Little Theorem. By doing so, we can mimic the above argument with one caveat: We will only be able to apply Euler's Theorem to those $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, and this could be a much smaller fraction of the integers. We start by stating this simplified version.

Proposition 6.6.2. *Let $m \in \mathbb{N}^+$ and let $A = \{0, 1, 2, \dots, m-1\}$. Let $k \in \mathbb{N}^+$ with $\gcd(k, \varphi(m)) = 1$, and let $\ell \in \mathbb{N}^+$ be such that $k\ell \equiv 1 \pmod{\varphi(m)}$. Define $h, g \colon A \to A$ by letting $h(a) = f_m(a^k)$ and $g(b) = f_m(b^\ell)$. For all $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, we have $g(h(a)) = a$ and $h(g(a)) = a$.*

Proof. Suppose that $\gcd(k, \varphi(m)) = 1$. By Corollary 6.1.16, we can fix $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{\varphi(m)}$ (notice that we choose a positive such ℓ by finding an equivalent value in the set $\{1, 2, \dots, m-1\}$). By definition, have $\varphi(m) \mid k\ell - 1$, so we can fix $n \in \mathbb{Z}$ with $\varphi(m) \cdot n = k\ell - 1$. Since $k, \ell \in \mathbb{N}^+$, we have $k\ell - 1 \geq 0$, and hence $n \geq 0$. Notice that $k\ell = 1 + \varphi(m) \cdot n$. Now for any $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, we have

$$\begin{aligned}
g(h(a)) &\equiv (a^k)^\ell \pmod{m} \\
&\equiv a^{k\ell} \pmod{m} \\
&\equiv a^{1+\varphi(m)\cdot n} \pmod{m} \\
&\equiv a \cdot a^{\varphi(m)\cdot n} \pmod{m} \\
&\equiv a \cdot (a^{\varphi(m)})^n \pmod{m} \\
&\equiv a \cdot 1^n \pmod{m} \\
&\equiv a \pmod{m},
\end{aligned}$$

where the second to last line follows from Euler's Theorem. A completely analogous argument shows that $h(g(a)) = a$. □

Again, note that this argument only works for those $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$. However, if m is the product of two primes, then it is possible to show that *every* a works.

Proposition 6.6.3. *Let $m = pq$ for two distinct primes $p, q \in \mathbb{N}^+$, and let $A = \{0, 1, 2, \dots, m-1\}$. Let $k \in \mathbb{N}^+$ with $\gcd(k, \varphi(m)) = 1$. The function $h \colon A \to A$ given by $h(a) = f_m(a^k)$ is a bijection. Moreover, if $\ell \in \mathbb{N}^+$ is such that $k\ell \equiv 1 \pmod{\varphi(m)}$, then the function $g \colon A \to A$ given by $g(b) = f_m(b^\ell)$ is an inverse of h.*

Proof. By the previous result, we need only check the values of $a \in \mathbb{Z}$ such that $\gcd(a, m) \neq 1$. The proof in this case is an exercise (see the Homework). □

In fact, this proposition is true whenever m is a product of distinct primes, which is same as saying that m is *square-free* (i.e. m is not divisible by the square of any natural number greater than or equal to 2). However, it is not true for all $m \in \mathbb{N}^+$.

To summarize, here is the encryption scheme in this case:

- Fix two distinct primes $p, q \in \mathbb{N}^+$ so that $m = pq$ is large enough to code any message you want to send. We can find such large prime using the methods of the previous section. Then pick some $k \in \mathbb{N}^+$ such that $\gcd(k, \varphi(m)) = 1$. We can find such a k through trial and error using the Euclidean Algorithm, so long as we know $\varphi(m)$.

- To encode a message, compute $h(a) = f_m(a^k)$, i.e. $h(a)$ is the unique element of A that is congruent to a^k modulo m. Notice that we can do exponentiation modulo m quickly (as described in the previous section).

- For decoding, we first compute an $\ell \in \mathbb{N}^+$ with $k\ell \equiv 1 \pmod{\varphi(m)}$, which we can find through the Euclidean Algorithm (so long as we know $\varphi(m)$). Then, given an encoded message, we compute $g(b) = f_m(b^\ell)$ to decode the message.

At first sight, everything seems completely analogous to the previous situation. However, there is one major difference, which is why we added the phrase "so long as we know $\varphi(m)$" in two places above. Notice that since p and q are distinct primes, we have

$$\begin{aligned}\varphi(m) &= \varphi(pq) \\ &= \varphi(p) \cdot \varphi(q) \\ &= (p-1)(q-1),\end{aligned}$$

so we can indeed compute $\varphi(m)$ if we know both p and q. Thus, if we know both p and q, then we can compute such a k and ℓ as described. Suppose that somebody, let's call her Alice, is setting up this encryption scheme, and does the hard work of generating two primes p and q in order to form m. Of course, Alice has access to p and q, so can compute both k and ℓ. Now if Alice wants somebody, let's call him Bob, to send her encrypted messages, then Alice can tell Bob the values of m and k. Notice that with only knowledge of m and k, Bob is able to encrypt and send the result to Alice. Since Alice knows ℓ, she can perform the necessary decryption.

However, think about what an eavesdropper would try to do in order to decrypt. In the past, any eavesdropper who knew both m and k would be able to compute ℓ using the Euclidean Algorithm. However, in this scenario, it seems that an eavesdropper would first want to compute $\varphi(m) = (p-1)(q-1)$. In order to do that, the eavesdropper apparently needs to *factor* m in order to recover both p and q. But that seems difficult! We do not know of any efficient way to take a large number and factor it into primes!

Thus, even if Alice broadcasts both m and k to the world, thus providing anybody a means to encrypt messages to her, it seems computationally infeasible for anybody other than Alice to figure out ℓ. Think about how this can be used in practice. If you want to transmit sensitive information over the internet (say your credit card number or other financial information), then the organization that wants that sensitive information can play the role of Alice, and transmit instructions to you about to send an encrypted message. Even if somebody is monitoring the line and sees these instructions, that eavesdropper will not be able to determine ℓ in any reasonable amount of time, and so apparently can not decrypt the sensitive information! This is amazing!

The above encryption scheme is known as RSA, which is named after Rivest, Shamir, and Adleman, the first people to publicly discover and popularize the ideas (it turns out that GCHQ in England had discovered the essential ideas several years earlier, but the information was classified). RSA is part of a family of so-called public-key encryption schemes. In all the encryption methods we discussed earlier, the sender and the recipient needed to keep certain "keys" completely private. For example, they could not tell anybody the values of m and k, as anybody with knowledge of those values (and sufficient knowledge of number theory) could then determine how to decrypt messages. In contrast, Alice can shout her "keys" m and k publicly, and just keep the value of ℓ private. As mentioned, it seems that anybody who wants to recover ℓ would need to factor m, which is computationally infeasible (to the best of our knowledge).

Although we have covered the fundamental theoretical background for basic RSA encryption, there are many subtleties and pitfalls that arise when implementing it, and there other practical considerations and issues. However, those interesting aspects are for another course and time.

Chapter 7

Growth Rates of Number-Theoretic Functions

In this chapter, we begin a study of the the growth rate of some simple functions. Our first task will be to define what we mean when we say that two functions grow at the same rate. Notice that if $f(x) = x^2$, and $g(x) = x$, then we trivially have $f(x) > g(x)$ for all $x \geq 1$, but we mean more than that when we say that $f(x)$ grows much faster than $g(x)$. For example, we have $x^2 > x^2 - 1$ for all x, but $f(x) = x^2$ and $g(x) = x^2 - 1$ are "essentially the same" for large values of x. In other words, simple inequalities will not suffice.

7.1 The Factorial Function

Define $f\colon \mathbb{N} \to \mathbb{N}$ by letting $f(n) = n!$. How fast does this factorial function grow? The short answer is "very fast". But what if we try to approximate it by another function, such as an exponential function? We have the simple lower bound that $n! \geq 2^{n-1}$ for all $n \geq 2$ because

$$\begin{aligned} n! &= n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 \\ &\geq 2 \cdot 2 \cdot 2 \cdots 2 \cdot 2 \cdot 1 \\ &= 2^{n-1} \end{aligned}$$

for all $n \in \mathbb{N}^+$. In fact, for $n \geq 4$, it's easy to check that $n! \geq 2^n$ because we get an extra factor of 2 from the 4. On the upper bound side, the best that we can immediately conclude is that

$$\begin{aligned} n! &= n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 \\ &\leq n \cdot n \cdot n \cdots n \cdot n \cdot n \\ &= n^n. \end{aligned}$$

Therefore, for all $n \geq 4$, we have

$$2^n \leq n! \leq n^n.$$

Are either of these bounds "close"? No. For example, given any $a \in \mathbb{N}$, it seems that $n! \geq a^n$ for any sufficiently large n, because eventually the leading terms in the product of $n!$ will dominate the repeated factors of a.

How do we formalize the intuition that the above bounds are bad? In order to compare the growth rates of the functions f and g whose domain is some unbounded subset of $\mathbb{R}$ (like $\mathbb{N}$ or $\mathbb{R}$), and whose codomain

$\mathbb{R}$, the idea is to look at the quotient $\frac{f(x)}{g(x)}$. For example, when comparing $f(x) = x^2$ and $g(x) = x$, we have that

$$\frac{f(x)}{g(x)} = \frac{x^2}{x} = x,$$

so for large values of x, the quotient $\frac{f(x)}{g(x)}$ is also large. More precisely, we have

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} x = \infty$$

and similarly

$$\lim_{x\to\infty} \frac{g(x)}{f(x)} = \lim_{x\to\infty} \frac{1}{x} = 0.$$

Notice instead that if we consider $f(x) = x^2 - 1$ and $g(x) = x^2$, then we have

$$\begin{aligned}
\lim_{x\to\infty} \frac{f(x)}{g(x)} &= \lim_{x\to\infty} \frac{x^2-1}{x^2} \\
&= \lim_{x\to\infty} \left(1 - \frac{1}{x}\right) \\
&= \lim_{x\to\infty} 1 - \lim_{x\to\infty} \frac{1}{x} \\
&= 1 - 0 \\
&= 1.
\end{aligned}$$

In order to compare two functions and f and g, the idea then is to look at the value of $\frac{f(x)}{g(x)}$ for large inputs x. We can then categorize growth rates with the following ideas:

- If $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = 0$, then g grows much faster than f.
- If $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = \infty$, then f grows much faster than g.
- If $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = c$ for some *positive* number c, then g grows at roughly the same rate as g, i.e. within a multiplicative constant.
- If $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = 1$, then f and g grow at the same rate.

It is also possible that $\frac{f(x)}{g(x)}$ does not approach a limit. For example, the values might forever oscillate between 1 and 3. However, we will focus on the simpler cases here (see Analysis for tools that allow one to understand more complicated situations).

Definition 7.1.1. *Given two functions f and g with domain some unbounded subset of $\mathbb{R}$, and codomain $\mathbb{R}$, we write $f \sim g$ to mean that*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

Let's return to our above example, and compare the function $f(n)$ with the function $g(n) = n^n$. For any $n \in \mathbb{N}^+$, we have

$$\begin{aligned}\frac{n!}{n^n} &= \frac{1 \cdot 2 \cdot 3 \cdots n}{n \cdot n \cdot n \cdots n} \\ &= \frac{1}{n} \cdot \frac{2}{n} \cdot \frac{3}{n} \cdots \frac{n}{n} \\ &\leq \frac{1}{n} \cdot 1 \cdot 1 \cdots 1 \\ &= \frac{1}{n}.\end{aligned}$$

Therefore,

$$0 \leq \frac{n!}{n^n} \leq \frac{1}{n}$$

for every $n \in \mathbb{N}$. Since

$$\lim_{n\to\infty} 0 = 0$$

and

$$\lim_{n\to\infty} \frac{1}{n} = 0,$$

it follows from the Squeeze Theorem that $\lim\limits_{n\to\infty} \frac{n!}{n^n} = 0$. Thus, using our above characterization, $g(n)$ grows faster than $f(n)$.

It's desirable to get a better idea of just how quickly $n!$ grows (for both theoretical and computational reasons, and also just for kicks). Since $n!$ is the product of the first n natural numbers, it seems that $(\frac{n}{2})^n$ is a plausible guess, since we are somehow "averaging out" the n factors $1, 2, \dots, n$ and replacing them all by $\frac{n}{2}$. However, this doesn't work. It turns out that

$$\lim_{n\to\infty} \frac{(n/2)^n}{n!} = \infty,$$

so $(\frac{n}{2})^n$ grows too fast. It is possible to argue this carefully once we obtain some better bounds on $n!$. However, the intuition is as follows. Consider

$$20! = 20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot \dots \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1.$$

If we reorder the product so that we pair the largest with the smallest factor, then the second largest with second smallest, etc., we obtain

$$20! = (20 \cdot 1) \cdot (19 \cdot 2) \cdot (18 \cdot 3) \cdot \dots \cdot (13 \cdot 8) \cdot (12 \cdot 9) \cdot (11 \cdot 10).$$

Now if compare this to $(\frac{20}{2})^{20} = 10^{20} = (10^2)^{10} = 100^{10}$, we see that although a few of the pairings are slightly greater than 100 (the last 3), several of the early pairing are much less than 100, and these drag down the product. However, it's also true that

$$\lim_{n\to\infty} \frac{(n/3)^n}{n!} = 0,$$

so $(\frac{n}{3})^n$ grows too slow. Interesting... Again, it is possible to argue this carefully using some of the estimates below.

In order to get some actual bounds, perhaps we should try to find some strategy that is better than blind guessing. To that end, notice that $n!$ is a product of n numbers, and we have few direct tools (at this point) to estimate such products. It is often easier to approximate the value of a sum than that of a product, so the idea is to use logarithms to turn the product into a sum. And if we can get a handle on $\ln(n!)$, then we use the exponential function to help us understand $n!$. We start by noticing that

$$\ln(n!) = \ln(\prod_{k=1}^{n} k) = \sum_{k=1}^{n} \ln k.$$

Now the sum on the right can be viewed as a Riemann sum approximating $\int_1^n \ln x \; dx$.

In order to obtain inequalities, rather than approximations, we notice that the function $f(x) = \ln x$ is an increasing function on $(0, \infty)$. Let $n \in \mathbb{N}^+$. We have

$$\begin{aligned} \ln(n!) &= \ln 1 + \ln 2 + \ln 3 + \cdots + \ln n \\ &= \ln 2 + \ln 3 + \cdots + \ln n && \text{(since } \ln 1 = 0) \\ &= 1 \cdot \ln 2 + 1 \cdot \ln 3 + \cdots + 1 \cdot \ln n, \end{aligned}$$

and we can view this last sum as using the right-hand endpoints in a Riemann sum with $\Delta x = 1$, we can use the fact that $f(x) = \ln x$ is increasing to conclude that

$$\begin{aligned} \int_1^n \ln x \; dx &\leq 1 \cdot \ln 2 + 1 \cdot \ln 3 + \cdots + 1 \cdot \ln n \\ &= \ln(n!). \end{aligned}$$

On the other hand, we can also notice that

$$\begin{aligned} \ln(n!) &= \ln 1 + \ln 2 + \ln 3 + \cdots + \ln n \\ &= \ln 1 + \ln 2 + \ln 3 + \cdots + \ln(n-1) + \ln n \\ &= (1 \cdot \ln 1 + 1 \cdot \ln 2 + 1 \cdot \ln 3 + \cdots + 1 \cdot \ln(n-1)) + \ln n. \end{aligned}$$

Viewing the sum in parentheses as using the left-hand endpoints in our Riemann sum, and using the fact that $f(x) = \ln x$ is increasing on $(0, \infty)$, it follows that

$$\begin{aligned} \ln(n!) &= (1 \cdot \ln 1 + 1 \cdot \ln 2 + 1 \cdot \ln 3 + \cdots + 1 \cdot \ln(n-1)) + \ln n \\ &\leq \left(\int_1^n \ln x \; dx \right) + \ln n. \end{aligned}$$

Combining this two inequalities, it follows that

$$\int_1^n \ln x \; dx \leq \ln(n!) \leq \ln n + \int_1^n \ln x \; dx.$$

Now using integration by parts with $u = \ln x$ and $dv = dx$, we have $du = \frac{1}{x} \; dx$ and $v = x$, so

$$\begin{aligned} \int_1^n \ln x \; dx &= x \ln x|_1^n - \int_1^n 1 \; dx \\ &= n \ln n - 1 \ln 1 - (x|_1^n) \\ &= n \ln n - (n-1) \\ &= n \ln n - n + 1. \end{aligned}$$

Therefore,

$$n \ln n - n + 1 \leq \ln(n!) \leq \ln n + n \ln n - n + 1.$$

Since the function $g(x) = e^x$ is increasing, it follows that

$$e^{n \ln n - n + 1} \leq n! \leq e^{\ln n + n \ln n - n + 1}.$$

Using basic exponent rules, it follows that

$$(e^{\ln n})^n \cdot e^{-n} \cdot e^1 \leq n! \leq e^{\ln n} \cdot (e^{\ln n})^n \cdot e^{-n} \cdot e^1,$$

so

$$n^n \cdot e^{-n} \cdot e \leq n! \leq n \cdot n^n \cdot e^{-n} \cdot e,$$

and hence

$$e \cdot \left(\frac{n}{e}\right)^n \leq n! \leq en \cdot \left(\frac{n}{e}\right)^n.$$

Therefore, it looks like $(\frac{n}{e})^n$ is a much better approximation to $n!$ than either $(\frac{n}{2})^n$ or $(\frac{n}{3})^n$. Notice that the upper bound grows much faster than the lower bound, because

$$\lim_{n \to \infty} \frac{en \cdot \left(\frac{n}{e}\right)^n}{e \cdot \left(\frac{n}{e}\right)^n} = \lim_{n \to \infty} n = \infty.$$

Thus, although we've sandwiched $n!$ between these two values, it's unclear if $n!$ is closer to the lower estimate, to the upper estimate, or if it is somewhere in the middle. It turns out that the lower bound grows slower than $n!$, and the upper bound grows faster. The core of the problem with our approximations is that Riemann sums use flat-topped rectangles to approximate the area under the curve $f(x) = \ln x$. We can do better with the Trapezoid Rule, which approximates $\int_1^n \ln x \, dx$ using the sum

$$\frac{1}{2} \ln 1 + \ln 2 + \ln 3 + \cdots + \ln(n-1) + \frac{1}{2} \ln n = \ln 2 + \ln 3 + \cdots + \ln(n-1) + \frac{1}{2} \ln n.$$

Using the fact that $f(x) = \ln x$ is concave down on $(0, \infty)$, it turns out that the straight lines at the top of the trapezoids are below the graph of $f(x) = \ln x$, so

$$\ln 2 + \ln 3 + \cdots + \ln(n-1) + \frac{1}{2} \ln n \leq \int_1^n \ln x \, dx.$$

Thus

$$\begin{aligned}
\ln(n!) &= \ln 1 + \ln 2 + \ln 3 + \cdots + \ln(n-1) + \ln n \\
&= \ln 2 + \ln 3 + \cdots + \ln(n-1) + \frac{1}{2} \ln n + \frac{1}{2} \ln n \\
&\leq \frac{1}{2} \ln n + \int_1^n \ln x \, dx \\
&= \frac{1}{2} \ln n + n \ln n - n + 1.
\end{aligned}$$

With this in hand, we can take the exponential of both sides, and refine our upper bound to be

$$\begin{aligned}
n! &\leq e^{(1/2) \ln n + n \ln n - n + 1} \\
&= (e^{\ln n})^{(1/2)} \cdot (e^{\ln n})^n \cdot e^{-n} \cdot e^1 \\
&= n^{\frac{1}{2}} \cdot n^n \cdot e^{-n} \cdot e \\
&= e\sqrt{n} \cdot \left(\frac{n}{e}\right)^n.
\end{aligned}$$

Combining this with the above lower bound, it follows that

$$e \cdot \left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n} \cdot \left(\frac{n}{e}\right)^n.$$

Since the trapezoidal approximation appears to be losing much less information, it is reasonable to hope that this upper bound grows at the same rate as $n!$. In fact, it does turn out that

$$\lim_{n \to \infty} \frac{e\sqrt{n} \cdot \left(\frac{n}{e}\right)^n}{n!}$$

does exist and is a positive constant. However, the constant is not 1, so $n! \not\sim e\sqrt{n} \cdot \left(\frac{n}{e}\right)^n$. In order to remedy that, we just need to replace the constant e in front of $\sqrt{n}$ by the appropriate constant, and shockingly it turns out the correct value is $\sqrt{2\pi}$.

Theorem 7.1.2 (Stirling's Approximation to $n!$)**.** *We have*

$$\lim_{n \to \infty} \frac{\sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n}{n!} = 1,$$

so

$$n! \sim \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n.$$

Unfortunately, we have neither the tools nor the time to prove this remarkable fact.

7.2 Average Number of Divisors

Although we had some success approximating $n!$, the exact growth rate that appears in Stirling approximation is complicated. To completely solve a problem of this type, we turn to a different number-theoretic function. Recall the function $d\colon \mathbb{N}^+ \to \mathbb{N}^+$ defined by letting $d(n)$ be the number of positive divisors of n. Now the function d oscillates quite a bit. For example, since there are infinitely many primes, we know that $\{n \in \mathbb{N}^+ : d(n) = 2\}$ is infinite. Since $d(2^n) = n + 1$ for all $n \in \mathbb{N}$ (see Corollary 3.3.5), the set $\{n \in \mathbb{N}^+ : d(n) = \log_2(n) + 1\}$ is also infinite.

Instead of talking about the growth rate of the function d, we examine the function $g\colon \mathbb{N}^+ \to \mathbb{R}$ defined by letting

$$g(n) = \frac{1}{n} \sum_{k=1}^{n} d(k).$$

That is, $g(n)$ is the *average* number of divisors of the elements up to, and including n. By performing this averaging, we "smooth out" the function d, making it have some more regular properties. Here is table of some values of these functions:

n	$d(n)$	$\sum_{k=1}^{n} d(k)$	$g(n)$
1	1	1	1
2	2	3	1.5
3	2	5	≈ 1.6667
4	3	8	2
5	2	10	2
10	4	27	2.7
50	6	207	4.14
100	9	482	4.82
1000	16	7069	7.069
10000	25	93668	9.3668

Look closely at the value $g(10)$, $g(100)$, $g(1000)$, and $g(10000)$. Between each of these, when we multiply the input by 10, we add a bit more than 2 to the output of g. Since multiplicative factors turn into additive factors, it is reasonable to suspect that the growth rate of g is logarithmic. In fact, it turns out that $g(n) \sim \ln n$. In order to prove this, we start by finding another way to evaluate the third column of the above table.

Proposition 7.2.1. *For any $n \in \mathbb{N}^+$, we have*

$$\sum_{k=1}^{n} d(k) = \sum_{k=1}^{n} \left\lfloor \frac{n}{k} \right\rfloor.$$

Proof. Let

$$P = \{(a,b) \in \mathbb{N}^+ \times \mathbb{N}^+ : ab \leq n\}.$$

We count $|P|$ in two different ways as follows:

- For each k with $1 \leq k \leq n$, let $S_k = \{(a,b) \in P : ab = k\}$. Notice that the S_k are pairwise disjoint, and that

$$P = S_1 \cup S_2 \cup \cdots \cup S_n.$$

  Now given any k with $1 \leq k \leq n$, the function $f_k \colon Div(k) \cap \mathbb{N}^+ \to S_k$ defined by letting $f_k(a) = (a, \frac{k}{a})$ is a bijection, so $|S_k| = |Div(k) \cap \mathbb{N}^+| = d(k)$. Therefore, by the Sum Rule, we have

$$|P| = \sum_{k=1}^{n} |S_k| = \sum_{k=1}^{n} d(k).$$

- For each k with $1 \leq k \leq n$, let $T_k = \{(a,b) \in P : a = k\}$. Notice that the T_k are pairwise disjoint, and that $|T_k| = \left\lfloor \frac{n}{k} \right\rfloor$ for each k (since we can pair k with any value in the set $[\lfloor \frac{n}{k} \rfloor]$ to obtain a product less than or equal to n). It follows that

$$|P| = \sum_{k=1}^{n} |T_k| = \sum_{k=1}^{n} \left\lfloor \frac{n}{k} \right\rfloor.$$

Therefore, we have

$$\sum_{k=1}^{n} d(k) = |P| = \sum_{k=1}^{n} \left\lfloor \frac{n}{k} \right\rfloor.$$

□

The idea now is to approximate

$$\sum_{k=1}^{n} \left\lfloor \frac{n}{k} \right\rfloor$$

using an integral. However, the floor function is a nuisance in this regard. In order to get rid of it, notice that we trivially have $x - 1 \leq \lfloor x \rfloor \leq x$ for all $x \in \mathbb{R}$. Hence, we have

$$\frac{n}{k} - 1 \leq \left\lfloor \frac{n}{k} \right\rfloor \leq \frac{n}{k}$$

for all $n, k \in \mathbb{N}$. It follows that

$$\sum_{k=1}^{n} \left(\frac{n}{k} - 1\right) \leq \sum_{k=1}^{n} \left\lfloor \frac{n}{k} \right\rfloor \leq \sum_{k=1}^{n} \frac{n}{k}$$

for all $n \in \mathbb{N}^+$. Using the above proposition, we conclude that

$$\sum_{k=1}^{n} \left(\frac{n}{k} - 1\right) \leq \sum_{k=1}^{n} d(k) \leq \sum_{k=1}^{n} \frac{n}{k}.$$

By pulling out the constant n and distributing the sum on the left, we see that

$$-n + n \cdot \sum_{k=1}^{n} \frac{1}{k} \leq \sum_{k=1}^{n} d(k) \leq n \cdot \sum_{k=1}^{n} \frac{1}{k},$$

so

$$-1 + \sum_{k=1}^{n} \frac{1}{k} \leq \frac{1}{n} \sum_{k=1}^{n} d(k) \leq \sum_{k=1}^{n} \frac{1}{k},$$

i.e.

$$-1 + \sum_{k=1}^{n} \frac{1}{k} \leq g(n) \leq \sum_{k=1}^{n} \frac{1}{k},$$

We now obtain bounds on the sum $\sum_{k=1}^{n} \frac{1}{k}$ by computing the integral $\int_1^n \frac{1}{x}\,dx$. We begin by noticing that the function $f(x) = \frac{1}{x}$ is a decreasing function on $(0, \infty)$. Since

$$\begin{aligned}\sum_{k=1}^{n} \frac{1}{k} &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &= 1 + \left(\frac{1}{2} \cdot 1 + \frac{1}{3} \cdot 1 + \cdots + \frac{1}{n} \cdot 1\right),\end{aligned}$$

and we can view the sum in parentheses as using the right-hand endpoints in a Riemann sum with $\Delta x = 1$, we can use the fact that $f(x) = \frac{1}{x}$ is decreasing to conclude that

$$\begin{aligned}\sum_{1}^{n} \frac{1}{k} &= 1 + \left(\frac{1}{2} \cdot 1 + \frac{1}{3} \cdot 1 + \cdots + \frac{1}{n} \cdot 1\right) \\ &\leq 1 + \int_1^n \frac{1}{x}\,dx \\ &= 1 + \ln n.\end{aligned}$$

On the other hand, we can also notice that

$$\begin{aligned}\sum_{k=1}^{n} \frac{1}{k} &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &= \left(1 \cdot 1 + \frac{1}{2} \cdot 1 + \frac{1}{3} \cdot 1 + \cdots + \frac{1}{n-1} \cdot 1\right) + \frac{1}{n}.\end{aligned}$$

Viewing the sum in parentheses as using the left-hand endpoints in our Riemann sum, and using the fact that $f(x) = \frac{1}{x}$ is decreasing on $(0, \infty)$, it follows that

$$\begin{aligned}\sum_{1}^{n} \frac{1}{k} &= \left(1 \cdot 1 + \frac{1}{2} \cdot 1 + \frac{1}{3} \cdot 1 + \cdots + \frac{1}{n-1} \cdot 1\right) + \frac{1}{n} \\ &\geq \left(\int_1^n \frac{1}{x}\,dx\right) + \frac{1}{n} \\ &= \frac{1}{n} + \ln n.\end{aligned}$$

Combining this two inequalities, it follows that

$$\frac{1}{n} + \ln n \leq \sum_{k=1}^{n} \frac{1}{k} \leq 1 + \ln n.$$

Recalling that

$$-1 + \sum_{k=1}^{n} \frac{1}{k} \leq g(n) \leq \sum_{k=1}^{n} \frac{1}{k},$$

we conclude that

$$(-1) + \frac{1}{n} + \ln n \leq g(n) \leq 1 + \ln n.$$

If $n \geq 2$, we can divide through by $\ln n > 0$ to conclude that

$$-\frac{1}{\ln n} + \frac{1}{n \ln n} + 1 \leq \frac{g(n)}{\ln n} \leq \frac{1}{\ln n} + 1.$$

Since

$$\lim_{n \to \infty} \frac{1}{\ln n} = 0$$

and

$$\lim_{n \to \infty} \frac{1}{n \ln n} = 0,$$

we have both

$$\lim_{n \to \infty} \left(-\frac{1}{\ln n} + \frac{1}{n \ln n} + 1 \right) = 1$$

and

$$\lim_{n \to \infty} \left(\frac{1}{\ln n} + 1 \right) = 1.$$

Using the Squeeze Theorem, we establish the following result.

Theorem 7.2.2. *Define $g \colon \mathbb{N}^+ \to \mathbb{N}^+$ by letting*

$$g(n) = \sum_{k=1}^{n} d(k).$$

We then have that $g(n) \sim \ln n$, i.e. that

$$\lim_{n \to \infty} \frac{g(n)}{\ln n} = 1.$$

7.3 Counting the Number of Primes

Definition 7.3.1. *We define a function $\pi \colon \mathbb{R} \to \mathbb{N}$ by letting $\pi(x)$ be the number of primes less than or equal to x.*

For example, we have $\pi(0) = 0$, $\pi(2) = 1$, $\pi(3) = 2$, $\pi(10) = 4$, and $\pi(\pi) = 2$. Initially, it might seem strange to extend the domain of π to be all real numbers (rather than just $\mathbb{N}$ and $\mathbb{N}^+$), but it turns out to be useful. We know that there are infinitely many primes, so

$$\lim_{x \to \infty} \pi(x) = \infty.$$

However, it is not at all clear how quickly $\pi(x)$ grows. To get a sense of the growth rate, it might be easier to think about the function $f(x) = \frac{\pi(x)}{x}$. Now given $n \in \mathbb{N}^+$, we have that $f(n)$ is the fraction of elements in $[n]$ that are prime. Notice that we trivially have that $0 \leq f(n) \leq 1$ for all $n \in \mathbb{N}^+$.

Since there is only one even prime, we expect that $f(n) \leq \frac{1}{2}$ for all moderately sized values of n (notice that $f(3) = \frac{3}{2}$ and $f(5) = \frac{3}{5}$ are both greater than $\frac{1}{2}$). Indeed, since 9 is the first odd number (other than 1) that is not prime, it balances out the 2, and so it straightforward to show that $f(n) \leq \frac{1}{2}$ for all $n \in \mathbb{N}$ with $n \geq 8$. Now as n becomes larger, we expect $f(n)$ to become smaller still. For example, we know that 3 is the only prime that is a multiple of 3. So we might think that we can subtract $\frac{1}{3}$ from our upper bound on f. However, since we have already eliminated the even numbers, we now only eliminate the *odd* multiples of 3. Using this idea, it's possible to show that $f(n) \leq 1 - \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$ for all sufficiently large n. Playing this game a little more, it seems reasonable to expect that $f(n)$ approaches 0 for large values of n. Stated in terms of the function $\pi(x)$, it is natural to conjecture that

$$\lim_{x \to \infty} \frac{\pi(x)}{x} = 0.$$

Indeed, this is true, and it will follow easily from some more sophisticated estimates that we carry out below. In order to determine the growth rate of $\pi(x)$, we need to ask how quickly $\frac{\pi(x)}{x}$ goes to 0. The answer to this difficult question is given by the celebrated Prime Number Theorem, which says that

$$\frac{\pi(x)}{x} \sim \frac{1}{\ln x}.$$

The theorem is usually stated in the following form.

Theorem 7.3.2 (Prime Number Theorem - Hadamard, de la Vallée Poussin). *We have*

$$\lim_{x \to \infty} \frac{\pi(x)}{x/\ln x} = 1,$$

so

$$\pi(x) \sim \frac{x}{\ln x} \qquad \textit{and} \qquad \frac{\pi(x)}{x} \sim \frac{1}{\ln x}.$$

The original proofs of the Prime Number Theorem occurred in 1896, and used sophisticated ideas from complex analysis. As a result, we will not be able to prove it here. However, we will prove the following weaker version:

Theorem 7.3.3. *There exist $c, d \in \mathbb{R}$ with $0 < c < 1 < d$ such that*

$$c \cdot \frac{n}{\ln n} \leq \pi(n) \leq d \cdot \frac{n}{\ln n}$$

for all $n \in \mathbb{N}^+$.

In other words, we will be able to bound the quotient $\frac{\pi(n)}{n/\ln n}$ between two positive numbers. Notice that this weaker result does not even suffice to argue that $\frac{\pi(n)}{n/\ln n}$ approaches a limit (it could conceivably oscillate forever), but it does show that $\frac{n}{\ln n}$ is the correct general order of growth.

As you will show on the homework, there are arbitrary large gaps in the primes. Although these gaps grow as we move further out, it is reasonable to expect that there are no large gaps too "early". One way to make this precise is the following fundamental result.

Theorem 7.3.4 (Bertrand's Postulate). *For all $n \in \mathbb{N}^+$, there exists a prime p with $n < p \leq 2n$.*

Notice that in the language of the function $\pi(x)$, Bertrand's Postulate is equivalent to showing that $\pi(2n) - \pi(n) > 0$ for all $n \in \mathbb{N}^+$. It turns out that the ideas that go into the proof of Bertrand's Postulate correspond with the ideas that we use to prove Theorem 7.3.3. Suppose then that we want to think about how we could attack Bertrand's Postulate. Start by thinking about the smallest number you can create that is divisible by all of the primes between n and $2n$. Of course, $(2n)!$ is such a number, but it is divisible by *all* primes less than or equal to $2n$. If we only want the primes between n and $2n$, we would do better to consider the number

$$\frac{(2n)!}{n!},$$

which is the product of all numbers in this range. In fact, we can do even better. We know that

$$\binom{2n}{n} = \frac{(2n)!}{n! \cdot n!}$$

is a natural number, and it is reasonable to believe that it is divisible by all the primes between n and $2n$, because nothing in the denominator will be able to cancel these primes. We will soon argue that this belief is in fact true (see Lemma 6.2.2 for an analogous result). In fact, we'll soon examine the prime factorization of $\binom{2n}{n}$ in much more detail. To obtain a sense of these factorizations, consider the following examples:

$$\begin{aligned}
\binom{10}{5} &= 2^2 \cdot 3^2 \cdot 7^1 \\
\binom{20}{10} &= 2^2 \cdot 11^1 \cdot 13^1 \cdot 17^1 \cdot 19^1 \\
\binom{30}{15} &= 2^4 \cdot 3^2 \cdot 5^1 \cdot 17^1 \cdot 19^1 \cdot 23^1 \cdot 29^1 \\
\binom{40}{20} &= 2^2 \cdot 3^2 \cdot 5^1 \cdot 7^1 \cdot 11^1 \cdot 13^1 \cdot 23^1 \cdot 29^1 \cdot 31^1 \cdot 37^1 \\
\binom{50}{25} &= 2^3 \cdot 3^2 \cdot 7^2 \cdot 13^1 \cdot 29^1 \cdot 31^1 \cdot 37^1 \cdot 41^1 \cdot 43^1 \cdot 47^1 \\
\binom{60}{30} &= 2^4 \cdot 7^1 \cdot 11^1 \cdot 17^1 \cdot 19^1 \cdot 31^1 \cdot 37^1 \cdot 41^1 \cdot 43^1 \cdot 47^1 \cdot 53^1 \cdot 59^1
\end{aligned}$$

There is a great deal to unpack here, but the fundamental takeaway is that $\binom{2n}{n}$ looks like the product of the primes between n and $2n$, with some extra small primes thrown in. Why is this helpful? The fundamental reason is that we can obtain some simple bounds on the size of $\binom{2n}{n}$ using some combinatorial reasoning.

Proposition 7.3.5. *For all $n \in \mathbb{N}^+$, we have*

$$2^n \leq \frac{4^n}{2n} \leq \binom{2n}{n} \leq 4^n.$$

Proof. Let $n \in \mathbb{N}^+$ be arbitrary. We first show that

$$\binom{2n}{n} \leq 4^n.$$

Notice that $4^n = 2^{2n}$, so 4^n counts the number of all subsets of $[2n]$. Since $\binom{2n}{n}$ counts the number of subsets of $[2n]$ with exactly n elements, this inequality follows immediately. Alternatively, we can notice that $\binom{2n}{n}$ is just one of the positive summands when we expand $(1+1)^{2n}$ using the Binomial Theorem.

To see that

$$\frac{4^n}{2n} \leq \binom{2n}{n},$$

we begin by recalling that

$$4^n = 2^{2n} = \sum_{k=0}^{2n} \binom{2n}{k}$$

by Corollary 5.2.3. By the homework, we know that $\binom{2n}{k} \leq \binom{2n}{n}$ for all k. Now there are $2k+1$ many summands, but we know that

$$\binom{2n}{0} + \binom{2n}{n} = 1 + 1 = 2 \leq \binom{2n}{n}$$

because $\{1, 2, \dots, n\}$ and $\{n+1, n+2, \dots, 2n\}$ are both subsets of $[2n]$ with n elements. Combining these two terms into one, we have a sum of $2n$ many terms on the right-hand side above, each of which is at most $\binom{2n}{n}$. Therefore,

$$4^n = \sum_{k=0}^{2n} \binom{2n}{k} \leq 2n \cdot \binom{2n}{n}.$$

Dividing both sides by $2n$, we obtain our inequality.

Finally, we need only prove that $2^n \leq \frac{4^n}{2n}$ for all $n \in \mathbb{N}^+$. We start by noticing that $m \leq 2^{m-1}$ for all $m \in \mathbb{N}^+$ (which is easily shown by a simple induction). Since $n \leq 2^{n-1}$, we can multiply both sides by 2 to conclude that $2n \leq 2^n$. Now $2^n = (\frac{4}{2})^n = \frac{4^n}{2^n}$, so $2n \leq \frac{4^n}{2^n}$. Multiplying both sides by $\frac{2^n}{2n} > 0$, we conclude that $2^n \leq \frac{4^n}{2n}$. $\square$

In order to determine properties of the the prime factorization of $\binom{2n}{n}$, we need to understand the prime factorization of $n!$. Given a prime p, how do we determine the power of p in the prime factorization of $n!$? In other words, given $n, p \in \mathbb{N}^+$ with p prime, how can we calculate $ord_p(n!)$? Let's look at an example. Suppose that we want to calculate $ord_3(30!)$. Since $ord_p(ab) = ord_p(a) + ord_p(b)$ (see Problem 7 on Homework 12), we have

$$ord_3(30!) = \sum_{k=1}^{30} ord_3(k).$$

Now several terms in this sum equal 1. For example, we have $ord_3(3) = 1, ord_3(6) = 1, ord_3(12) = 1$, etc. We also have $ord_3(9) = 2$, $ord_3(18) = 2$, and $ord_3(27) = 3$. To evaluate sums like this more generally, we need to determine the number of summands of each value.

Proposition 7.3.6. *Let $n, p \in \mathbb{N}^+$ with p prime. We have*

$$\begin{aligned} ord_p(n!) &= \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \dots \\ &= \sum_{k=1}^{\left\lfloor \frac{\ln n}{\ln p} \right\rfloor} \left\lfloor \frac{n}{p^k} \right\rfloor. \end{aligned}$$

Proof. Since $ord_p(ab) = ord_p(a) + ord_p(b)$ for all $a, b \in \mathbb{N}^+$, we know that

$$ord_p(n!) = \sum_{m=1}^{n} ord_p(m).$$

Notice that exactly $\left\lfloor \frac{n}{p} \right\rfloor$ many numbers in $[n]$ are divisible by p, and hence exactly $\left\lfloor \frac{n}{p} \right\rfloor$ many terms in the sum are greater than or equal to 1. In general, for any $k \in \mathbb{N}^+$, exactly $\left\lfloor \frac{n}{p^k} \right\rfloor$ many numbers in $[n]$ are divisible by p^k, and hence exactly $\left\lfloor \frac{n}{p^k} \right\rfloor$ many terms in the sum are greater than or equal to k. Thus, for the each $k \in \mathbb{N}^+$, the number of summands exactly equal to k is $\left\lfloor \frac{n}{p^k} \right\rfloor - \left\lfloor \frac{n}{p^{k+1}} \right\rfloor$. It follows that

$$\begin{aligned}
ord_p(n!) &= \sum_{m=1}^{n} ord_p(m) \\
&= 1 \cdot \left(\left\lfloor \frac{n}{p} \right\rfloor - \left\lfloor \frac{n}{p^2} \right\rfloor \right) + 2 \cdot \left(\left\lfloor \frac{n}{p^2} \right\rfloor - \left\lfloor \frac{n}{p^3} \right\rfloor \right) + 3 \cdot \left(\left\lfloor \frac{n}{p^3} \right\rfloor - \left\lfloor \frac{n}{p^4} \right\rfloor \right) + \dots \\
&= 1 \cdot \left\lfloor \frac{n}{p} \right\rfloor + (2-1) \cdot \left\lfloor \frac{n}{p^2} \right\rfloor + (3-2) \cdot \left\lfloor \frac{n}{p^3} \right\rfloor + \dots \\
&= \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \dots .
\end{aligned}$$

Finally, given $k \in \mathbb{N}^+$, notice that we have $p^k \leq n$ if and only if $\ln(p^k) \leq \ln n$ (since $f(x) = \ln x$ is an increasing function), which is if and only if $k \cdot \ln p \leq \ln n$, which is if and only if $k \leq \frac{\ln n}{\ln p}$. Therefore, we can cut off the above sum at $\left\lfloor \frac{\ln n}{\ln p} \right\rfloor$. □

Using this result, we can compute the prime factorization of $n!$. Now in order to determine the prime factorization of

$$\binom{2n}{n} = \frac{(2n!)}{n! \cdot n!},$$

it appears that we need to square the powers of the primes in $n!$, and subtract the result from the powers of the primes in $(2n)!$. In order to determine the fallout of this process, we will need the following simple result.

Lemma 7.3.7. *For any $x \in \mathbb{R}$, we have $\lfloor 2x \rfloor - 2\lfloor x \rfloor \in \{0, 1\}$.*

Proof. Let $x \in \mathbb{R}$ be arbitrary. Let $m = \lfloor x \rfloor \in \mathbb{Z}$, and let $z = x - m$, so $0 \leq z < 1$. We then have $x = m + z$, and hence $2x = 2m + 2z$. We now consider two cases:

- *Case 1:* Suppose that $0 \leq z < \frac{1}{2}$. We then have $0 \leq 2z < 1$. Since $2x = 2m + 2z$ and $2m \in \mathbb{Z}$, it follows that $\lfloor 2x \rfloor = 2m$. Now $\lfloor x \rfloor = m$, so $\lfloor 2x \rfloor - 2\lfloor x \rfloor = 2m - 2m = 0$.
- *Case 2:* Suppose that $\frac{1}{2} \leq z < 1$. We then have $1 \leq 2z < 2$, so $0 \leq 2z - 1 < 1$. Since

$$\begin{aligned}
2x &= 2m + 2z \\
&= 2m + 1 + (2z - 1),
\end{aligned}$$

  and $2m + 1 \in \mathbb{Z}$, it follows that $\lfloor 2x \rfloor = 2m + 1$. Now $\lfloor x \rfloor = m$, so $\lfloor 2x \rfloor - 2\lfloor x \rfloor = 2m + 1 - 2m = 1$.

Therefore, we have $\lfloor 2x \rfloor - 2\lfloor x \rfloor \in \{0, 1\}$ in either case. □

We can now state our fundamental theorem about the prime factorization of $\binom{2n}{n}$.

Theorem 7.3.8. *Let $n \in \mathbb{N}^+$ and let $N = \binom{2n}{n}$.*

1. *For all primes p, we have $ord_p(N) \leq \frac{\ln(2n)}{\ln p}$, and hence $p^{ord_p(N)} \leq 2n$.*

2. *For all primes p with $p > \sqrt{2n}$, we have $ord_p(N) \leq 1$.*
3. *If $n \geq 3$ and $\frac{2}{3}n < p \leq n$, then $ord_p(N) = 0$.*
4. *If $n < p \leq 2n$, then $ord_p(N) = 1$.*

Before jumping into the proof, let's pause to analyze what the theorem is saying in the context of an example. We noted above that

$$\binom{50}{25} = 2^3 \cdot 3^2 \cdot 7^2 \cdot 13^1 \cdot 29^1 \cdot 31^1 \cdot 37^1 \cdot 41^1 \cdot 43^1 \cdot 47^1.$$

The first part of the above theorem is saying the following: If we take any prime, the value of the corresponding prime power in the factorization will be at most 50. For example, $2^3 = 8$, $7^2 = 49$, and $43^1 = 43$ are all less than or equal to 50. Moving on, the second statement says that only very small primes can occur to a power larger than 1. In the above example, we have $\sqrt{50} \approx 7.07$, so any prime larger than 7 must occur to the zeroth or first power only. Moreover, the fourth statement guarantees that all primes between n and $2n$ really do appear. Finally, the third statement (which is not needed until later) promises the existence of a nontrivial gap: any prime in this range will not appear at all in the factorization. In our example with $n = 25$, we have $\frac{2}{3}n = \frac{50}{3} \approx 16.67$, so any prime p with $17 \leq p \leq 25$ (which in this case are the primes 17, 19, and 23) do not appear in the above factorization. We now turn to the proof.

Proof. Since

$$N = \frac{(2n)!}{n! \cdot n!},$$

we have

$$n! \cdot n! \cdot N = (2n)!.$$

Therefore, for each prime $p \in \mathbb{N}^+$, we have

$$ord_p(n!) + ord_p(n!) + ord_p(N) = ord_p((2n)!),$$

and hence

$$ord_p(N) = ord_p((2n)!) - 2 \cdot ord_p(n!).$$

Using Proposition 7.3.6, it follows that

$$ord_p(N) = \sum_{k=1}^{\left\lfloor \frac{\ln(2n)}{\ln p} \right\rfloor} \left(\left\lfloor \frac{2n}{p^k} \right\rfloor - 2 \cdot \left\lfloor \frac{n}{p^k} \right\rfloor \right)$$

for each prime $p \in \mathbb{N}^+$. Using this expression, we now prove the four statements.

1. Let $p \in \mathbb{N}^+$ be an arbitrary prime. By the Lemma 7.3.7, each term in above sum is either 0 and 1. Therefore, $ord_p(N) \leq \frac{\ln(2n)}{\ln p}$. Multiplying both sides by $\ln p > 0$, we see that $(\ln p) \cdot ord_p(N) \leq \ln(2n)$, so since $f(x) = e^x$ is an increasing function, it follows that

   $$e^{(\ln p) \cdot ord_p(N)} \leq e^{\ln(2n)}.$$

   Therefore,

   $$(e^{\ln p})^{ord_p(N)} \leq e^{\ln(2n)},$$

   and hence

   $$p^{ord_p(N)} \leq 2n.$$

2. Let $p \in \mathbb{N}^+$ be prime with $p > \sqrt{2n}$. Squaring both sides, it follows that $p^2 > 2n$. Since $f(x) = \ln x$ is an increasing function, it follows that $\ln(p^2) > \ln(2n)$, so $2 \ln p > \ln(2n)$, and thus $\frac{\ln(2n)}{\ln p} < 2$. Using part 1 and the fact that $ord_p(N) \in \mathbb{N}$, it follows that $ord_p(N) \leq 1$. (Alternatively, since $p^2 > 2n$ and we know that $p^{ord_p(N)} \leq 2n$ by part 1, it follows that we must have $ord_p(N) \leq 1$.)

3. Suppose that $n \geq 3$ and let $p \in \mathbb{N}^+$ be prime with $\frac{2}{3}n < p \leq n$. Notice that since $p > \frac{2}{3}n$ and $n \geq 3$, we have $p > \frac{2}{3} \cdot 3 = 2$, so $p \geq 3$. Multiplying both sides of $\frac{2}{3}n < p$ by 3, we see that $2n < 3p$. Now since $p \geq 3$, it follows that $2n < 3p \leq p^2$. As in the proof of part 2, it then follows that $\frac{\ln(2n)}{\ln p} < 2$. Examining the above sum, we conclude that it only has one term, and so

$$ord_p(N) = \left\lfloor \frac{2n}{p} \right\rfloor - 2 \cdot \left\lfloor \frac{n}{p} \right\rfloor.$$

Since $\frac{2}{3}n < p$, we can multiply through by $\frac{3}{p}$ to conclude that $\frac{2n}{p} < 3$. Also, since $p \leq n$, we have $2p \leq 2n$, so $2 \leq \frac{2n}{p}$. Therefore, $2 \leq \frac{2n}{p} < 3$, and hence $\left\lfloor \frac{2n}{p} \right\rfloor = 2$. Multiplying through by $\frac{1}{2}$, we see that $1 \leq \frac{n}{p} < \frac{3}{2}$, so $\left\lfloor \frac{n}{p} \right\rfloor = 1$. Therefore, $ord_p(N) = 2 - 2 \cdot 1 = 0$.

4. Let $p \in \mathbb{N}^+$ be prime with $n < p \leq 2n$. Since $p \geq 2$, we have $2n < 2p \leq p^2$, so as in part 3 we have $\frac{\ln(2n)}{\ln p} < 2$ and hence
$$ord_p(N) = \left\lfloor \frac{2n}{p} \right\rfloor - 2 \cdot \left\lfloor \frac{n}{p} \right\rfloor.$$
Now multiplying $n < p \leq 2n < 2p$ through by $\frac{1}{p}$, we see that $\frac{n}{p} < 1 \leq \frac{2n}{p} < 2$. Thus, we have $\left\lfloor \frac{2n}{p} \right\rfloor = 0$ and $\left\lfloor \frac{n}{p} \right\rfloor = 1$, so $ord_p(N) = 1 - 2 \cdot 0 = 1$.

$\square$

Using information about the prime factorization of $\binom{2n}{n}$, we can now obtain the following number-theoretic bounds.

Corollary 7.3.9. *For all $n \in \mathbb{N}^+$, we have*

$$n^{\pi(2n) - \pi(n)} \leq \binom{2n}{n} \leq (2n)^{\pi(2n)}.$$

Proof. Let $N = \binom{2n}{n}$. Let $N = p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell}$ be the prime factorization of N into distinct primes with $k_i > 0$ for all i.

For each prime $p \in \mathbb{N}^+$, we know that $p^{ord_p(N)} \leq 2n$ by Theorem 7.3.8, so $p_i^{k_i} \leq 2n$ for all i. Furthermore, since $p_i \leq 2n$ for each i, we have that $\ell \leq \pi(2n)$. Therefore,

$$N = p_1^{k_1} p_2^{k_2} \cdots p_\ell^{k_\ell} \leq (2n)^{\pi(2n)}.$$

On the other hand, for each prime p with $n < p \leq 2n$, we know that $p \mid N$ by Theorem 7.3.8, so there exists an i with $p = p_i$. Now there are $\pi(2n) - \pi(n)$ many such primes, and since each of them is greater than n, it follows that

$$n^{\pi(2n) - \pi(n)} \leq N.$$

$\square$

Theorem 7.3.10. *For all $n \in \mathbb{N}$ with $n \geq 2$, we have*

$$\pi(n) \geq \frac{\ln 2}{2} \cdot \frac{n}{\ln n}.$$

Proof. We first prove this in the case where n is even. Suppose that $n = 2m$. We have

$$2^m \le \binom{2m}{m} \le (2m)^{\pi(2m)},$$

so

$$m \ln 2 \le \pi(2m) \ln(2m),$$

and hence

$$\pi(2m) \ge \ln 2 \cdot \frac{m}{\ln(2m)},$$

or

$$\pi(2m) \ge \frac{\ln 2}{2} \cdot \frac{2m}{\ln(2m)}.$$

Thus

$$\pi(n) \ge \frac{\ln 2}{2} \cdot \frac{n}{\ln n}$$

whenever n is even. Now consider an odd $n \ge 2$. Write $n = 2m-1$ where $m \ge 2$. Notice than $2m$ is an even number greater than or equal to 4, so $2m$ is not prime, and hence $\pi(2m-1) = \pi(2m)$. We therefore have

$$\begin{aligned}\pi(n) &= \pi(2m-1)\\ &= \pi(2m)\\ &\ge \frac{\ln 2}{2} \cdot \frac{2m}{\ln(2m)}.\end{aligned}$$

Now notice that the function $f(x) = \frac{x}{\ln x}$ is increasing for $x \ge 3$ because

$$f'(x) = \frac{\ln x - 1}{(\ln x)^2},$$

which is greater than 0 for $x > 3$. Since $2m-1 \ge 3$, it follows that

$$\frac{\ln 2}{2} \cdot \frac{2m}{\ln(2m)} \ge \frac{\ln 2}{2} \cdot \frac{2m-1}{\ln(2m-1)},$$

so

$$\pi(n) \ge \frac{\ln 2}{2} \cdot \frac{n}{\ln n}.$$

□

We can improve the bound a bit if we use slightly better lower bounds on $\binom{2n}{n}$. For example,

$$(3.99)^n \le \frac{4^n}{2n}$$

for all sufficiently large n. Using $\ln(3.99)$ in place of $\ln 2$ in the above argument, it is possible to conclude that

$$\pi(n) \ge \frac{\ln 3.99}{2} \cdot \frac{n}{\ln n}.$$

for all sufficiently large n.

In the other direction, we have

$$n^{\pi(2n)-\pi(n)} \le \binom{2n}{n} \le 4^n$$

so
$$(\pi(2n) - \pi(n)) \cdot \ln n \leq n \ln 4,$$
and hence
$$\pi(2n) - \pi(n) \leq \ln 4 \cdot \frac{n}{\ln n},$$
which is the same as saying that
$$\pi(2n) - \pi(n) \leq 2 \ln 2 \cdot \frac{n}{\ln n}.$$
In order to obtain an upper bound on $\pi(m)$, we need to use this result to repeatedly divide m by 2. The easiest case to use is when m is a power of 2. For example, suppose that $m = 32$. We have
$$\begin{aligned}
\pi(32) &= (\pi(32) - \pi(16)) + (\pi(16) - \pi(8)) + (\pi(8) - \pi(4)) + (\pi(4) - \pi(2)) + \pi(2) \\
&\leq 2 \ln 2 \cdot \frac{16}{\ln 16} + 2 \ln 2 \cdot \frac{8}{\ln 8} + 2 \ln 2 \cdot \frac{4}{\ln 4} + 2 \ln 2 \cdot \frac{2}{\ln 2} + 1 \\
&= 2 \ln 2 \cdot \frac{16}{4 \ln 2} + 2 \ln 2 \cdot \frac{8}{3 \ln 2} + 2 \ln 2 \cdot \frac{4}{2 \ln 2} + 2 \ln 2 \cdot \frac{2}{\ln 2} + 1 \\
&= 2 \cdot \left(\frac{16}{4} + \frac{8}{3} + \frac{4}{2} + \frac{2}{1} \right) + 1 \\
&= 2 \cdot \left(\frac{16}{4} + \frac{8}{3} + \frac{4}{2} + \frac{2}{1} + \frac{1}{2} \right)
\end{aligned}$$

Thus, in order to provide an upper bound on $\pi(2^m)$, we want to provide an upper bound on
$$\frac{2^1}{1} + \frac{2^2}{2} + \frac{2^3}{3} + \frac{2^4}{4} + \cdots + \frac{2^{m-1}}{m-1}$$
Since
$$1 + 2 + 2^2 + \cdots + 2^{m-1} = \frac{2^m - 1}{2 - 1} = 2^m - 1,$$
which is essentially the next term in the sum, it might be reasonable to hope that
$$\frac{2}{1} + \frac{2^2}{2} + \frac{2^3}{3} + \cdots + \frac{2^{m-1}}{m-1} \leq \frac{2^m}{m}.$$
Unfortunately, this is not true, as
$$\frac{2}{1} + \frac{2^2}{2} = 4 > \frac{2^3}{3},$$
and the problem continues. However, it turns out that inequality holds up to a multiplicative constant. In other words, there exists c such that
$$\frac{2}{1} + \frac{2^2}{2} + \frac{2^3}{3} + \cdots + \frac{2^{m-1}}{m-1} \leq c \cdot \frac{2^m}{m}$$
for all m. In fact, it's possible to show that $c = 2$ works (although you can bring that down with a bit of work). From this, it follows that
$$\pi(2^m) \leq 2c \cdot \frac{2^m}{m}$$
which is the same as saying that
$$\pi(2^m) \leq 2c \ln 2 \cdot \frac{2^m}{\ln(2^m)}$$

Now let n be arbitrary. Fix m with $2^m \leq n < 2^{m+1}$. Notice that $\ln n < (m+1)\ln 2$, and hence

$$\frac{1}{m+1} < \frac{\ln 2}{\ln n}.$$

It follows that

$$\begin{aligned}
\pi(n) &\leq \pi(2^{m+1}) \\
&\leq 2c \cdot \frac{2^{m+1}}{m+1} \\
&\leq 2c \ln 2 \cdot \frac{2n}{\ln n} \\
&\leq 4c \ln 2 \cdot \frac{n}{\ln n}.
\end{aligned}$$

Theorem 7.3.11. *For all $n \in \mathbb{N}$ with $n \geq 2$, we have*

$$\frac{\ln 2}{2} \cdot \frac{n}{\ln n} \leq \pi(n) \leq 8 \ln 2 \cdot \frac{n}{\ln n}.$$