Chapter 2

Relations, Functions, Partial Functions

2.1 What is a Function?

Roughly speaking, a function, f, is a rule or mechanism, which takes input values in some *input domain*, say X, and produces output values in some *output domain*, say Y, in such a way that to each input $x \in X$ corresponds a *unique* output value $y \in Y$, denoted $f(x)$.

We usually write $y = f(x)$, or better, $x \mapsto f(x)$.

Often, functions are defined by some sort of closed expression (a formula), but not always.

For example, the formula

$$y = 2x$$

defines a function. Here, we can take both the input and output domain to be $\mathbb{R}$, the set of real numbers.

Instead, we could have taken $\mathbb{N}$, the set of natural numbers; this gives us a different function.

In the above example, $2x$ makes sense for all input x, whether the input domain is $\mathbb{N}$ or $\mathbb{R}$, so our formula yields a function defined for all of its input values.

Now, look at the function defined by the formula

$$y = \frac{x}{2}.$$

If the input and output domains are both $\mathbb{R}$, again this function is well-defined.

However, what if we assume that the input and output domains are both $\mathbb{N}$?

This time, we have a problem when x is odd. For example, $\frac{3}{2}$ is not an integer, so our function is not defined for all of its input values.

It is a *partial function*, a concept that subsumes the notion of a function but is more general.

Observe that this partial function is defined for the set of even natural numbers (sometimes denoted $2\mathbb{N}$) and this set is called the *domain* (of definition) of f.

If we enlarge the output domain to be $\mathbb{Q}$, the set of rational numbers, then our partial function is defined for all inputs.

Another example of a partial function is given by

$$y = \frac{x+1}{x^2 - 3x + 2},$$

assuming that both the input and output domains are $\mathbb{R}$.

Observe that for $x = 1$ and $x = 2$, the denominator vanishes, so we get the undefined fractions $\frac{2}{0}$ and $\frac{3}{0}$.

This partial function "blows up" for $x = 1$ and $x = 2$, its value is "infinity" $(= \infty)$, which is not an element of $\mathbb{R}$. So, the domain of f is $\mathbb{R} - \{1, 2\}$.

In summary, partial functions need not be defined for all of their input values and we need to pay close attention to both the input and the ouput domain of our partial functions.

The following example illustrates another difficulty: Consider the partial function given by

$$y = \sqrt{x}.$$

If we assume that the input domain is $\mathbb{R}$ and that the output domain is $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x \geq 0\}$, then this partial function is *not defined for negative values of* x.

To fix this problem, we can extend the output domain to be $\mathbb{C}$, the complex numbers. Then we can make sense of $\sqrt{x}$ when $x < 0$.

However, a new problem comes up: Every negative number, x, has two complex square roots, $-i\sqrt{-x}$ and $+i\sqrt{-x}$ (where i is "the" square root of -1). Which of the two should we pick?

In this case, we could systematically pick $+i\sqrt{-x}$ but what if we extend the input domain to be $\mathbb{C}$.

Then, it is not clear which of the two complex roots should be picked, as there is no obvious total order on $\mathbb{C}$.

We can treat f as a *multi-valued function*, that is, a function that may return several possible outputs for a given input value.

Experience shows that it is awkward to deal with multi-valued functions and that it is best to treat them as relations (or to change the output domain to be a power set, which is equivalent to view the function as a relation).

Let us give one more example showing that it is not always easy to make sure that a formula is a proper definition of a function.

Consider the function from $\mathbb{R}$ to $\mathbb{R}$ given by

$$f(x) = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!}.$$

Here, $n!$ is the function *factorial*, defined by

$$n! = n \cdot (n-1) \cdots 2 \cdot 1.$$

How do we make sense of this infinite expression?

Well, that's where analysis comes in, with the notion of limit of a series, etc. It turns out that $f(x)$ is the exponential function $f(x) = e^x$.

Actually, e^x is even defined when x is a complex number or even a square matrix (with real or complex entries)! Don't panic, we will not use such functions in this course.

Another issue comes up, that is, the notion of *computability*.

In all of our examples, and for most (partial) functions we will ever need to compute, it is clear that it is possible to give a mechanical procedure, i.e., a computer program which computes our functions (even if it hard to write such a program or if such a program takes a very long time to compute the output from the input).

Unfortunately, there are functions which, *although well-defined mathematically, are not computable*!

For an example, let us go back to first-order logic and the notion of provable proposition.

Given a finite (or countably infinite) alphabet of function, predicate, constant symbols, and a countable supply of variables, it is quite clear that the set $\mathcal{F}$ of all propositions built up from these symbols and variables can be enumerated systematically.

We can define the function, Prov, with input domain $\mathcal{F}$ and output domain $\{0, 1\}$, so that, for every proposition $P \in \mathcal{F}$,

$$\text{Prov}(P) = \begin{cases} 1 & \text{if } P \text{ is provable (classically)} \\ 0 & \text{if } P \text{ is not provable (classically).} \end{cases}$$

Mathematically, for every proposition, $P \in \mathcal{F}$, either P is provable or it is not, so this function makes sense.

However, by Church's Theorem (see Section **??**), we know that there is **no** computer program that will terminate for all input propositions and give an answer in a finite number of steps!

So, although the function Prov makes sense as an abstract function, it is not computable.

Is this a paradox? No, if we are careful when defining a function not to incorporate in the definition any notion of computability and instead to take a more abstract and, in some some sense, naive view of a function as some kind of input/output process given by pairs
⟨input value, output value⟩ (without worrying about the way the output is "computed" from the input).

A rigorous way to proceed is to use the notion of ordered pair and of graph of a function.

Before we do so, let us point out some facts about "functions" that were revealed by our examples:

1. In order to define a "function", in addition to defining its input/output behavior, it is also important to specify what is its *input domain* and its *output domain*.
2. Some "functions" may not be defined for all of their input values; a function can be a *partial function*.
3. The input/output behavior of a "function" can be defined by a set of ordered pairs. As we will see next, this is the *graph* of the function.

2.2 Ordered Pairs, Cartesian Products, Relations, Functions, Partial Functions

Given two sets, A and B, one of the basic constructions of set theory is the formation of an *ordered pair*, $\langle a, b\rangle$, where $a \in A$ and $b \in B$.

Sometimes, we also write (a, b) for an ordered pair.

The main property of ordered pairs is that if $\langle a_1, b_1\rangle$ and $\langle a_2, b_2\rangle$ are ordered pairs, where $a_1, a_2 \in A$ and $b_1, b_2 \in B$, then

$$\langle a_1, b_1\rangle = \langle a_2, b_2\rangle \quad \text{iff} \quad a_1 = a_2 \quad \text{and} \quad b_1 = b_2.$$

Observe that this property implies that,

$$\langle a, b\rangle \neq \langle b, a\rangle,$$

unless $a = b$.

Thus, the ordered pair, $\langle a, b\rangle$, is not a notational variant for the set $\{a, b\}$; implicit to the notion of ordered pair is the fact that there is an order (even though we have not yet defined this notion yet!) among the elements of the pair.

Indeed, in $\langle a, b\rangle$, the *element a comes first and b comes second*.

Accordingly, given an ordered pair, $p = \langle a, b\rangle$, we will denote a by $pr_1(p)$ and b by $pr_2(p)$ (*first and second projection* or *first and second coordinate*).

Remark: Readers who like set theory will be happy to hear that an ordered pair, $\langle a, b\rangle$, can be defined as the set

$$\{\{a\}, \{a, b\}\}.$$

This definition is due to Kuratowski, 1921. An earlier (more complicated) definition given by N. Wiener in 1914 is $\{\{\{a\}, \emptyset\}, \{\{b\}\}\}$.

Figure 2.1: Kazimierz Kuratowski, 1896-1980

Now, from set theory, it can be shown that given two sets, A and B, the set of all ordered pairs, $\langle a, b\rangle$, with $a \in A$ and $b \in B$, is a set denoted $A \times B$ and called the *Cartesian product of A and B* (in that order). The set $A \times B$ is also called the *cross-product* of A and B.

By convention, we agree that $\emptyset \times B = A \times \emptyset = \emptyset$.

To simplify the terminology, we often say *pair* for *ordered pair*, with the understanding that pairs are always ordered (otherwise, we should say set).

Of course, given three sets, A, B, C, we can form $(A \times B) \times C$ and we call its elements (ordered) *triples* (or *triplets*).

To simplify the notation, we write $\langle a, b, c\rangle$ instead of $\langle\langle a, b\rangle, c\rangle$ and $A \times B \times C$ instead of $(A \times B) \times C$.

More generally, given n sets $A_1, \ldots, A_n$ $(n \geq 2)$, we define the set of *n-tuples*, $A_1 \times A_2 \times \cdots \times A_n$, as $(\cdots((A_1 \times A_2) \times A_3) \times \cdots) \times A_n$.

An element of $A_1 \times A_2 \times \cdots \times A_n$ is denoted by $\langle a_1, \ldots, a_n\rangle$ (an n-tuple).

We agree that when $n = 1$, we just have A_1 and a 1-tuple is just an element of A_1.

We now have all we need to define relations.

Definition 2.2.1 Given two sets, A and B, a (binary) *relation between A and B* is any triple, $\langle A, R, B\rangle$, where $R \subseteq A \times B$ is any set of ordered pairs from $A \times B$. When $\langle a, b\rangle \in R$, we also write aRb and we say that *a and b are related by R*. The set

$$dom(R) = \{a \in A \mid \exists b \in B,\ \langle a, b\rangle \in R\}$$

is called the *domain of R* and the set

$$range(R) = \{b \in B \mid \exists a \in A,\ \langle a, b\rangle \in R\}$$

is called the *range of R*. Note that $dom(R) \subseteq A$ and $range(R) \subseteq B$. When $A = B$, we often say that *R is a (binary) relation over A*.

The term *correspondence between A and B* is also used instead of the term relation between A and B and the word *relation* is reserved for the case where $A = B$.

It is worth emphasizing that two relations, $\langle A, R, B\rangle$ and $\langle A', R', B'\rangle$, are equal iff $A = A'$, $B = B'$ and $R = R'$.

In particular, if $R = R'$ but either $A \neq A'$ or $B \neq B'$, then the relations $\langle A, R, B\rangle$ and $\langle A', R', B'\rangle$ *are considered to be different*.

For simplicity, we usually refer to a relation, $\langle A, R, B\rangle$, as a relation, $R \subseteq A \times B$.

Among all relations between A and B, we mention three relations that play a special role:

1. $R = \emptyset$, the *empty relation*. Note that $dom(\emptyset) = range(\emptyset) = \emptyset$. This is not a very exciting relation!
2. When $A = B$, we have the *identity relation*,

   $$\mathrm{id}_A = \{\langle a, a\rangle \mid a \in A\}.$$

   The identity relation relates every element to itself, and that's it! Note that $dom(\mathrm{id}_A) = range(\mathrm{id}_A) = A$.
3. The relation $A \times B$ itself. This relation relates every element of A to every element of B. Note that $dom(A \times B) = A$ and $range(A \times B) = B$.

Relations can be represented graphically by pictures often called *graphs*. (Beware, the term "graph" is very much overloaded. Later on, we will define what a graph is.)

We depict the elements of both sets A and B as points (perhaps with different colors) and we indicate that $a \in A$ and $b \in B$ are related (i.e., $\langle a, b\rangle \in R$) by drawing an *oriented edge* (an arrow) starting from a (its *source*) and ending in b (its *target*). Here is an example:

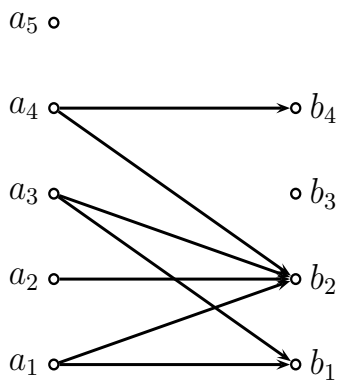


Figure 2.2: A binary relation, R

In Figure 2.2, $A = \{a_1, a_2, a_3, a_4, a_5\}$ and $B = \{b_1, b_2, b_3, b_4\}$.

Observe that a_5 is not related to any element of B, b_3 is not related to any element of A and some elements of A, namely, a_1, a_3, a_4, are related to several elements of B.

Now, given a relation, $R \subseteq A \times B$, some element $a \in A$ may be related to several distinct elements $b \in B$.

If so, R does not correspond to our notion of a function, because we want our functions to be single-valued.

So, we impose a natural condition on relations to get relations that correspond to functions.

Definition 2.2.2 We say that a relation, R, between two sets A and B is *functional* if for every $a \in A$, there is *at most one* $b \in B$ so that $\langle a, b\rangle \in R$. Equivalently, R is functional if for all $a \in A$ and all $b_1, b_2 \in B$, if $\langle a, b_1\rangle \in R$ and $\langle a, b_2\rangle \in R$, then $b_1 = b_2$.

The picture in Figure 2.3 shows an example of a functional relation.

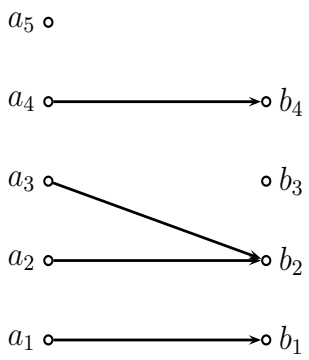


Figure 2.3: A functional relation G

Using Definition 2.2.2, we can give a rigorous definition of a function (partial or not).

Definition 2.2.3 A *partial function*, f, is a triple, $f = \langle A, G, B\rangle$, where A is a set called the *input domain of* f, B is a set called the *output domain of* f (sometimes *codomain of* f) and $G \subseteq A \times B$ is a functional relation called the *graph of* f (see Figure 2.4); we let $graph(f) = G$.

We write $f\colon A \to B$ to indicate that A is the input domain of f and that B is the codomain of f and we let $dom(f) = dom(G)$ and $range(f) = range(G)$.

For every $a \in dom(f)$, the unique element, $b \in B$, so that $\langle a, b\rangle \in graph(f)$ is denoted by $f(a)$ (so, $b = f(a)$). Often, we say that $b = f(a)$ is the *image of* a *by* f.

The range of f is also called the *image of* f and is denoted $\mathrm{Im}\,(f)$. If $dom(f) = A$, we say that f is a *total function*, for short, a *function with domain* A.

As in the case of relations, it is worth emphasizing that two functions (partial or total), $f = \langle A, G, B\rangle$ and $f' = \langle A', G', B'\rangle$, are equal iff $A = A'$, $B = B'$ and $G = G'$.

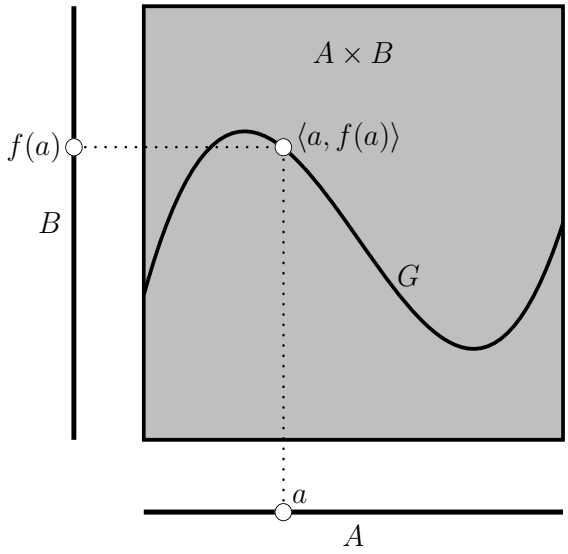


Figure 2.4: A (partial) function $\langle A, G, B\rangle$

In particular, if $G = G'$ but either $A \neq A'$ or $B \neq B'$, then the functions (partial or total) f and f' *are considered to be different*.

Remarks:

1. If $f = \langle A, G, B\rangle$ is a partial function and $b = f(a)$ for some $a \in dom(f)$, we say that *f maps a to b*; we may write $f : a \mapsto b$. For any $b \in B$, the set

$$\{a \in A \mid f(a) = b\}$$

is denoted $f^{-1}(b)$ and called the *inverse image* or *preimage of b by f*. (It is also called the *fibre of f above b*. We will explain this peculiar language later on.)

Note that $f^{-1}(b) \neq \emptyset$ iff b is in the image (range) of f. Often, a function, partial or not, is called a *map*.

2. Note that Definition 2.2.3 allows $A = \emptyset$. In this case, we must have $G = \emptyset$ and, technically, $\langle \emptyset, \emptyset, B \rangle$ is total function! It is the *empty function from $\emptyset$ to B*.

3. When a partial function is a total function, we don't call it a "partial total function", but simply a "function".

   The usual pratice is that the term "function" refers to a total function. However, sometimes, we say "total function" to stress that a function is indeed defined on all of its input domain.

4. Note that if a partial function $f = \langle A, G, B \rangle$ is not a total function, then $dom(f) \neq A$ and for all $a \in A - dom(f)$, there is **no** $b \in B$ so that $\langle a, b \rangle \in graph(f)$.

This corresponds to the intuitive fact that f does not produce any output for any value not in its domain of definition. We can imagine that f "blows up" for this input (as in the situation where the denominator of a fraction is 0) or that the program computing f loops indefinitely for that input.

5. If $f = \langle A, G, B\rangle$ is a total function and $A \neq \emptyset$, then $B \neq \emptyset$.

6. For any set, A, the identity relation, id_A, is actually a function $\mathrm{id}_A \colon A \to A$.

7. Given any two sets, A and B, the rules $\langle a, b\rangle \mapsto a = pr_1(\langle a, b\rangle)$ and $\langle a, b\rangle \mapsto b = pr_2(\langle a, b\rangle)$ make pr_1 and pr_2 into functions $pr_1 \colon A \times B \to A$ and $pr_2 \colon A \times B \to B$ called the *first and second projections*.

8. A function, $f : A \to B$, is sometimes denoted $A \xrightarrow{f} B$. Some authors use a different kind of arrow to indicate that f is partial, for example, a dotted or dashed arrow. We will not go that far!

9. The set of all functions, $f : A \to B$, is denoted by B^A. If A and B are finite, A has m elements and B has n elements, it is easy to prove that B^A has n^m elements.

The reader might wonder why, in the definition of a (total) function, $f : A \to B$, we do not require $B = \operatorname{Im} f$, since we require that $\operatorname{dom}(f) = A$.

The reason has to do with experience and convenience.

It turns out that in most cases, we know what the domain of a function is, but it may be very hard to determine exactly what its image is.

Thus, it is more convenient to be flexible about the codomain. As long as we know that f maps into B, we are satisfied.

For example, consider functions, $f\colon \mathbb{R} \to \mathbb{R}^2$, from the real line into the plane. The image of such a function is a *curve* in the plane $\mathbb{R}^2$.

Actually, to really get "decent" curves we need to impose some reasonable conditions on f, for example, to be differentiable. Even continuity may yield very strange curves (see Section 2.10).

But even for a very well behaved function, f, it may be very hard to figure out what the image of f is.

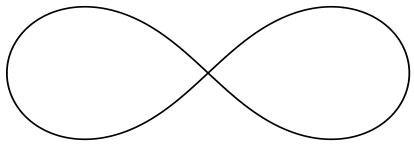


Figure 2.5: Lemniscate of Bernoulli

Consider the function, $t \mapsto (x(t), y(t))$, given by

$$\begin{aligned} x(t) &= \frac{t(1+t^2)}{1+t^4} \\ y(t) &= \frac{t(1-t^2)}{1+t^4}. \end{aligned}$$

The curve which is the image of this function, shown in Figure 2.5, is called the "*lemniscate of Bernoulli*".

Observe that this curve has a self-intersection at the origin, which is not so obvious at first glance.

2.3 Induction Principles on $\mathbb{N}$

Now that we have the notion of function, we can restate the induction principle (Version 2) stated at the end of Section 1.10 to make it more flexible.

We define a *property of the natural numbers* as any function, $P\colon \mathbb{N} \to \{\mathbf{true}, \mathbf{false}\}$.

The idea is that $P(n)$ holds iff $P(n) = \mathbf{true}$, else $P(n) = \mathbf{false}$. Then, we have the following principle:

Principle of Induction for $\mathbb{N}$ (Version 3).

Let P be any property of the natural numbers. In order to prove that $P(n)$ holds for all $n \in \mathbb{N}$, it is enough to prove that

(1) $P(0)$ holds and

(2) For every $n \in \mathbb{N}$, the implication $P(n) \Rightarrow P(n+1)$ holds.

As a formula, (1) and (2) can be written

$$[P(0) \wedge (\forall n \in \mathbb{N})(P(n) \Rightarrow P(n+1))] \Rightarrow (\forall n \in \mathbb{N})P(n).$$

Step (1) is usually called the *basis* or *base step* of the induction and step (2) is called the *induction step*.

In step (2), $P(n)$ is called the *induction hypothesis*.

That the above induction principle is valid is given by the

Proposition 2.3.1 *The Principle of Induction stated above is valid.*

Induction is a very valuable tool for proving properties of the natural numbers and we will make extensive use of it.

We will also see other more powerful induction principles. Let us give some examples illustrating how it is used.

We begin by finding a formula for the sum

$$1 + 2 + 3 + \cdots + n,$$

where $n \in \mathbb{N}$.

If we compute this sum for small values of n, say $n = 0, 1, 2, 3, 4, 5, 6$ we get

$$\begin{aligned} 0 &= 0 \\ 1 &= 1 \\ 1+2 &= 3 \\ 1+2+3 &= 6 \\ 1+2+3+4 &= 10 \\ 1+2+3+4+5 &= 15 \\ 1+2+3+4+5+6 &= 21. \end{aligned}$$

What is the pattern?

After a moment of reflection, we see that

$$\begin{aligned} 0 &= (0 \times 1)/2 \\ 1 &= (1 \times 2)/2 \\ 3 &= (2 \times 3)/2 \\ 6 &= (3 \times 4)/2 \\ 10 &= (4 \times 5)/2 \\ 15 &= (5 \times 6)/2 \\ 21 &= (6 \times 7)/2, \end{aligned}$$

so we conjecture

Claim 1:

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2},$$

where $n \in \mathbb{N}$.

For the *basis of the induction*, where $n = 0$, we get $0 = 0$, so the base step holds.

For the *induction step*, for any $n \in \mathbb{N}$, assume that

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

Consider $1 + 2 + 3 + \cdots + n + (n+1)$. Then, using the induction hypothesis, we have

$$\begin{aligned} 1 + 2 + 3 + \cdots + n + (n+1) &= \frac{n(n+1)}{2} + n + 1 \\ &= \frac{n(n+1) + 2(n+1)}{2} \\ &= \frac{(n+1)(n+2)}{2}, \end{aligned}$$

establishing the induction hypothesis and therefore, proving our formula. □

Next, let us find a formula for the sum of the first $n+1$ odd numbers:

$$1+3+5+\cdots+2n+1,$$

where $n \in \mathbb{N}$.

If we compute this sum for small values of n, say $n = 0, 1, 2, 3, 4, 5, 6$ we get

$$\begin{aligned} 1 &= 1 \\ 1+3 &= 4 \\ 1+3+5 &= 9 \\ 1+3+5+7 &= 16 \\ 1+3+5+7+9 &= 25 \\ 1+3+5+7+9+11 &= 36 \\ 1+3+5+7+9+11+13 &= 49. \end{aligned}$$

This time, it is clear what the pattern is: we get perfect squares.

Thus, we conjecture

Claim 2:

$$1 + 3 + 5 + \cdots + 2n + 1 = (n + 1)^2,$$

where $n \in \mathbb{N}$.

For the *basis of the induction*, where $n = 0$, we get $1 = 1^2$, so the base step holds.

For the *induction step*, for any $n \in \mathbb{N}$, assume that

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

Consider $1 + 3 + 5 + \cdots + 2n + 1 + 2(n + 1) + 1 = 1 + 3 + 5 + \cdots + 2n + 1 + 2n + 3$.

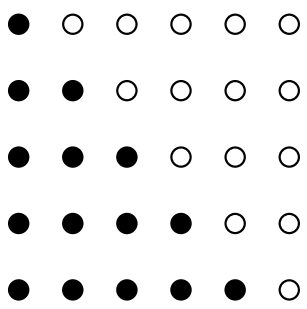
Then, using the induction hypothesis, we have

$$\begin{aligned} 1 + 3 + 5 + \cdots + 2n + 1 + 2n + 3 &= (n + 1)^2 + 2n + 3 \\ &= n^2 + 2n + 1 + 2n + 3 \\ &= n^2 + 4n + 4 = (n + 2)^2. \end{aligned}$$

Therefore, the induction step holds and this completes the proof by induction. □

The two formulae that we just discussed are subject to a nice *geometric interpetation* that suggests a closed form expression for each sum and this is often the case for sums of special kinds of numbers.

For the first formula, if we represent n as a sequence of n "bullets", then we can form a rectangular array with n rows and $n+1$ columns showing that the desired sum is half of the number of bullets in the array, which is indeed $\frac{n(n+1)}{2}$, as shown below for $n = 5$:

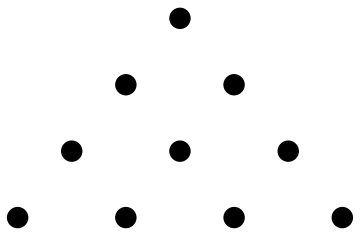


Thus, we see that the numbers,

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

have a simple geometric interpretation in terms of triangles of bullets.

For example, $\Delta_4 = 10$ is represented by the triangle



For this reason, the numbers, Δ_n, are often called *triangular numbers*. A natural question then arises: What is the sum

$$\Delta_1 + \Delta_2 + \Delta_3 + \cdots + \Delta_n?$$

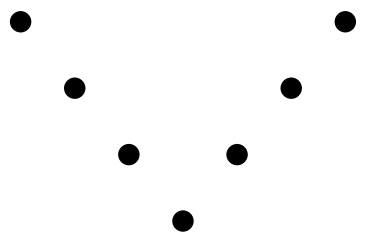
The reader should compute these sums for small values of n and try to guess a formula that should then be proved correct by induction. It is not too hard to find a nice formula for these sums.

The reader may also want to find a geometric interpretation for the above sums (stacks of cannon balls!).

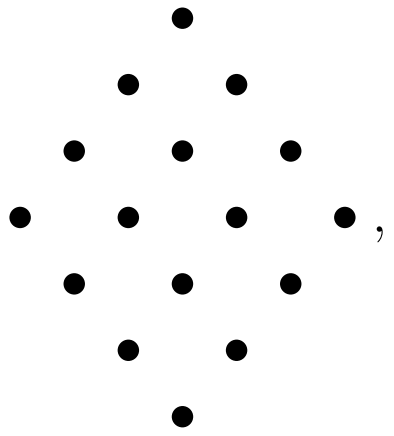
In order to get a geometric interpretation for the sum

$$1 + 3 + 5 + \cdots + 2n + 1,$$

we represent $2n + 1$ using $2n + 1$ bullets displayed in a V-shape; for example, $7 = 2 \times 3 + 1$ is represented by



Then, the sum $1 + 3 + 5 + \cdots + 2n + 1$ corresponds to the square



which clearly reveals that

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

A natural question is then: What is the sum

$$1^2 + 2^2 + 3^2 + \cdots + n^2?$$

Again, the reader should compute these sums for small values of n, then guess a formula and check its correctness by induction. It is not too difficult to find such a formula.

For a fascinating discussion of all sorts of numbers and their geometric interpretations (including the numbers we just introduced), the reader is urged to read Chapter 2 of Conway and Guy [4].

Sometimes, it is necessary to prove a property, $P(n)$, *for all natural numbers $n \geq m$, where $m > 0$*.

Our induction principle does not seem to apply since the base case is not $n = 0$.

However, we can define the property, $Q(n)$, given by

$$Q(n) = P(m+n), \quad n \in \mathbb{N},$$

and since *$Q(n)$ holds for all $n \in \mathbb{N}$ iff $P(k)$ holds for all $k \geq m$*, we can apply our induction principle to prove $Q(n)$ for all $n \in \mathbb{N}$ and thus, $P(k)$, for all $k \geq m$ (note, $k = m + n$).

Of course, *this amounts to considering that the base case is $n = m$* and this is what we always do without any further justification.

Here is an example. Let us prove that

$$(3n)^2 \leq 2^n, \quad \text{for all } \ n \geq 10.$$

The *base case* is $n = 10$.

For $n = 10$, we get

$$(3 \times 10)^2 = 30^2 = 900 \leq 1024 = 2^{10},$$

which is indeed true.

Let us now prove the *induction step*. Assuming that $(3n)^2 \leq 2^n$ holds for all $n \geq 10$, we want to prove that $(3(n+1))^2 \leq 2^{n+1}$.

Since

$$(3(n+1))^2 = (3n+3)^2 = (3n)^2 + 18n + 9,$$

if we can prove that $18n + 9 \leq (3n)^2$ when $n \geq 10$, using the induction hypothesis, $(3n)^2 \leq 2^n$, we will have

$$(3(n+1))^2 = (3n)^2 + 18n + 9 \leq$$
$$(3n)^2 + (3n)^2 \leq 2^n + 2^n = 2^{n+1},$$

establishing the induction step.

However,

$$(3n)^2 - (18n + 9) = (3n-3)^2 - 18$$

and $(3n-3)^2 \geq 18$ as soon as $n \geq 3$, so $18n + 9 \leq (3n)^2$ when $n \geq 10$, as required.

Observe that the formula $(3n)^2 \leq 2^n$ fails for $n = 9$, since $(3 \times 9)^2 = 27^2 = 729$ and $2^9 = 512$, but $729 > 512$. Thus, the base has to be $n = 10$.

There is another induction principle which is often more flexible that our original induction principle.

This principle, called *complete induction* (or sometimes *strong induction*), is stated below.

Complete Induction Principle for $\mathbb{N}$.

In order to prove that a predicate, $P(n)$, holds for all $n \in \mathbb{N}$ it is enough to prove that

(1) $P(0)$ holds (the base case) and

(2) for every $m \in \mathbb{N}$, if $(\forall k \in \mathbb{N})(k \leq m \Rightarrow P(k))$ then $P(m+1)$.

The difference between ordinary induction and complete induction is that in complete induction, the induction hypothesis, $(\forall k \in \mathbb{N})(k \leq m \Rightarrow P(k))$, *assumes that $P(k)$ holds for all $k \leq m$ and not just for m* (as in ordinary induction), in order to deduce $P(m+1)$.

This gives us more proving power as we have more knowledge in order to prove $P(m+1)$.

Complete induction will be discussed more extensively in Section 5.3 and its validity will be proved as a consequence of the fact that every nonempty subset of $\mathbb{N}$ has a smallest element but we can also justify its validity as follows:

Define $Q(m)$ by

$$Q(m) = (\forall k \in \mathbb{N})(k \leq m \Rightarrow P(k)).$$

Then, it is an easy exercise to show that if we apply our (ordinary) induction principle to $Q(m)$ (Induction Principle, Version 3), then we get the principle of complete induction.

Figure 2.6: Leonardo P. Fibonacci, 1170-1250

Here is an example of a proof using complete induction.

Define the sequence of natural numbers, F_n, (*Fibonacci sequence*) by

$$F_0 = 1,\ F_1 = 1,\ F_{n+2} = F_{n+1} + F_n,\ n \geq 0.$$

We claim that

$$F_n \geq \frac{3^{n-2}}{2^{n-3}}, \quad n \geq 3.$$

The *base case* corresponds to $n = 3$, where

$$F_3 = 3 \geq \frac{3^1}{2^0} = 3,$$

which is true.

Note that *we also need to consider the case $n = 4$ by itself* before we do the induction step because even though $F_4 = F_3 + F_2$, the induction hypothesis only applies to F_3 ($n \geq 3$ in the inequality above).

We have

$$F_4 = 5 \geq \frac{3^2}{2^1} = \frac{9}{2},$$

which is true since $10 > 9$.

Now for the *induction step* where $n \geq 3$, we have

$$\begin{aligned} F_{n+2} &= F_{n+1} + F_n \\ &\geq \frac{3^{n-1}}{2^{n-2}} + \frac{3^{n-2}}{2^{n-3}} \\ &\geq \frac{3^{n-2}}{2^{n-3}}\left(1 + \frac{3}{2}\right) = \frac{3^{n-2}}{2^{n-3}}\frac{5}{2} \geq \frac{3^{n-2}}{2^{n-3}}\frac{9}{4} = \frac{3^n}{2^{n-1}}, \end{aligned}$$

since $\frac{5}{2} > \frac{9}{4}$, which concludes the proof of the induction step.

Observe that we used the induction hypothesis for both F_{n+1} and F_n in order to deduce that it holds for F_{n+2}. This is where we needed the extra power of complete induction.

Remark: The Fibonacci sequence, F_n, is really a function from $\mathbb{N}$ to $\mathbb{N}$ defined recursively but we haven't proved yet that recursive definitions are legitimate methods for defining functions!

In fact, certain restrictions are needed on the kind of recursion used to define functions. This topic will be explored further in Section 2.5. Using results from Section 2.5, it can be shown that the Fibonacci sequence is a well-defined function (but this does not follow immediately from Theorem 2.5.1).

Induction proofs can be subtle and it might be instructive to see some examples of *faulty* induction proofs.

Assertion 1: For every natural numbers, $n \geq 1$, the number $n^2 - n + 11$ is an odd prime (recall that a prime number is a natural number, $p \geq 2$, which is only divisible by 1 and itself).

Proof. We use induction on $n \geq 1$. For the *base case*, $n = 1$, we have $1^2 - 1 + 11 = 11$, which is an odd prime, so the induction step holds.

For the *induction step*, assume that $n^2 - n + 11$ is prime. Then, as

$$(n+1)^2 - (n+1) + 11 = n^2 + n + 11,$$

we see that

$$(n+1)^2 - (n+1) + 11 = n^2 - n + 11 + 2n.$$

By the induction hypothesis, $n^2 - n + 11$ is an odd prime, p, and since $2n$ is even, $p + 2n$ is odd and therefore prime, establishing the induction hypothesis. □

If we compute $n^2 - n + 11$ for $n = 1, 2, \ldots, 10$, we find that these numbers are indeed all prime, but for $n = 11$, we get

$$121 = 11^2 - 11 + 11 = 11 \times 11,$$

which is not prime!

Where is the mistake?

What is wrong is the induction step: the fact that $n^2 - n + 11$ is prime does not imply that $(n+1)^2 - (n+1) + 11 = n^2 + n + 11$ is prime, as illustrated by $n = 10$. Our "proof" of the induction step is nonsense!

The lesson is: The fact that a statement holds for many values of $n \in \mathbb{N}$ *does not* imply that it holds for all $n \in \mathbb{N}$ (or all $n \geq k$, for some fixed $k \in \mathbb{N}$).

Interestingly, the prime numbers, k, so that $n^2 - n + k$ is prime for $n = 1, 2, \ldots, k-1$, are all known (there are only six of them!).

It can be shown that these are the prime numbers, k, such that $1 - 4k$ is a *Heegner number*, where the Heegner numbers are the nine integers:

$$-1,\ -2,\ -3,\ -7,\ -11,\ -19,\ -43,\ -67,\ -163.$$

The above results are hard to prove and require some deep theorems of number theory. What can also be shown (and you should prove it!) is that no nonconstant polynomial takes prime numbers as values for all natural numbers.

Assertion 2: Every Fibonacci number, F_n, is even.

Proof. For the *base case*, $F_2 = 2$, which is even, so the base case holds.

For the *induction step*, assume inductively that F_n is even for all $n \geq 2$. Then, as

$$F_{n+2} = F_{n+1} + F_n$$

and as both F_n and F_{n+1} are even by the induction hypothesis, we conclude that F_{n+2} is even. □

However, *Assertion 2 is clearly false*, since the Fibonacci sequence begins with

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34, \ldots.$$

This time, the mistake is that *we did not check the two base cases*, $F_0 = 1$ and $F_1 = 1$.

Our experience is that if an induction proof is wrong, then, in many cases, the base step is faulty. So, pay attention to the base step(s)!

A useful way to produce new relations or functions is to compose them.

2.4 Composition of Relations and Functions

We begin with the definition of the composition of relations.

Definition 2.4.1 Given two relations, $R \subseteq A \times B$ and $S \subseteq B \times C$, the *composition of R and S*, denoted $R \circ S$, is the relation between A and C defined by

$$R \circ S = \{\langle a, c\rangle \in A \times C \mid \exists b \in B,\ \langle a, b\rangle \in R \quad \text{and} \quad \langle b, c\rangle \in S\}.$$

One should check that for any relation $R \subseteq A \times B$, we have $\mathrm{id}_A \circ R = R$ and $R \circ \mathrm{id}_B = R$.

If R and S are the graphs of functions, possibly partial, is $R \circ S$ the graph of some function? The answer is yes, as shown in the following

Proposition 2.4.2 *Let $R \subseteq A \times B$ and $S \subseteq B \times C$ be two relations.*

(a) If R and S are both functional relations, then $R \circ S$ is also a functional relation. Consequently, $R \circ S$ is the graph of some partial function.

(b) If $dom(R) = A$ and $dom(S) = B$, then $dom(R \circ S) = A$.

(c) If R is the graph of a (total) function from A to B and S is the graph of a (total) function from B to C, then $R \circ S$ is the graph of a (total) function from A to C.

Proposition 2.4.2 shows that it is legitimate to define the composition of functions, possibly partial. Thus, we make the following

Definition 2.4.3 Given two functions, $f\colon A \to B$ and $g\colon B \to C$, possibly partial, the *composition of f and g*, denoted $g \circ f$, is the function (possibly partial)

$$g \circ f = \langle A, \mathit{graph}(f) \circ \mathit{graph}(g), C\rangle.$$

The reader must have noticed that the composition of two functions $f\colon A \to B$ and $g\colon B \to C$ is denoted $g \circ f$, whereas the graph of $g \circ f$ is denoted $\mathit{graph}(f) \circ \mathit{graph}(g)$.

This "reversal" of the order in which function composition and relation composition are written is unfortunate and somewhat confusing.

Once again, we are victim of tradition. The main reason for writing function composition as $g \circ f$ is that traditionally, the result of applying a function, f, to an argument, x, is written $f(x)$.

Then, $(g \circ f)(x) = g(f(x))$, because $z = (g \circ f)(x)$ iff there is some y so that $y = f(x)$ and $z = g(y)$, that is, $z = g(f(x))$.

Some people, in particular algebraists, write function composition as $f \circ g$, but then, they write the result of applying a function f to an argument x as xf. With this convention, $x(f \circ g) = (xf)g$, which also makes sense.

We prefer to stick to the convention where we write $f(x)$ for the result of applying a function f to an argument x and, consequently, we use the notation $g \circ f$ for the composition of f with g, even though it is the opposite of the convention for writing the composition of relations.

Given any three relations, $R \subseteq A \times B$, $S \subseteq B \times C$ and $T \subseteq C \times D$, the reader should verify that

$$(R \circ S) \circ T = R \circ (S \circ T).$$

We say that composition is *associative*.

Similarly, for any three functions (possibly partial), $f\colon A \to B$, $g\colon B \to C$ and $h\colon C \to D$, we have (associativity of function composition)

$$(h \circ g) \circ f = h \circ (g \circ f).$$

2.5 Recursion on $\mathbb{N}$

The following situation often occurs: We have some set, A, some fixed element, $a \in A$, some function, $g\colon A \to A$, and we wish to define a new function, $h\colon \mathbb{N} \to A$, so that

$$\begin{aligned} h(0) &= a, \\ h(n+1) &= g(h(n)) \qquad \text{for all} \quad n \in \mathbb{N}. \end{aligned}$$

This way of defining h is called a *recursive definition* (or a definition by *primitive recursion*).

I would be surprised if any computer scientist had any trouble with this "definition" of h but how can we justify rigorously that such a function exists and is unique?

Indeed, the existence (and uniqueness) of h requires proof.

The proof, although not really hard, is surprisingly involved and, in fact quite subtle. The reader will find a complete proof in Enderton [5] (Chapter 4).

Theorem 2.5.1 *(Recursion Theorem on $\mathbb{N}$) Given any set, A, any fixed element, $a \in A$, and any function, $g\colon A \to A$, there is a unique function, $h\colon \mathbb{N} \to A$, so that*

$$\begin{aligned} h(0) &= a, \\ h(n+1) &= g(h(n)) \qquad \textit{for all} \quad n \in \mathbb{N}. \end{aligned}$$

Theorem 2.5.1 is very important. Indeed, experience shows that it is used almost as much as induction!

As an example, we show how to define addition on $\mathbb{N}$. Indeed, at the moment, we know what the natural numbers are but we don't know what are the arithmetic operations such as $+$ or $*$! (at least, not in our axiomatic treatment; of course, nobody needs an axiomatic treatment to know how to add or multiply).

How do we define $m + n$, where $m, n \in \mathbb{N}$?

If we try to use Theorem 2.5.1 directly, we seem to have a problem, because addition is a function of two arguments, but h and g in the theorem only take one argument.

We can overcome this problem in two ways:

(1) We prove a generalization of Theorem 2.5.1 involving functions of several arguments, but with recursion only in a *single* argument. This can be done quite easily but we have to be a little careful.

(2) For any fixed m, we define $add_m(n)$ as $add_m(n) = m + n$, that is, we define addition of a *fixed* m to any n. Then, we let $m + n = add_m(n)$.

Since solution (2) involves much less work, we follow it. Let S denote the successor function on $\mathbb{N}$, that is, the function given by

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

Then, using Theorem 2.5.1 with $a = m$ and $g = S$, we get a function, add_m, such that

$$\begin{aligned} add_m(0) &= m, \\ add_m(n+1) &= S(add_m(n)) = add_m(n) + 1, \end{aligned}$$

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

Finally, for all $m, n \in \mathbb{N}$, we define $m + n$ by

$$m + n = add_m(n).$$

Now, we have our addition function on $\mathbb{N}$. But this is not the end of the story because we don't know yet that the above definition yields a function having the usual properties of addition, such as

$$\begin{aligned} m + 0 &= m \\ m + n &= n + m \\ (m + n) + p &= m + (n + p). \end{aligned}$$

To prove these properties, of course, we use induction!

We can also define multiplication. Mimicking what we did for addition, define $mult_m(n)$ by recursion as follows;

$$\begin{aligned} mult_m(0) &= 0, \\ mult_m(n+1) &= mult_m(n) + m \qquad \text{for all} \quad n \in \mathbb{N}. \end{aligned}$$

Then, we set

$$m \cdot n = mult_m(n).$$

Note how the recursive definition of $mult_m$ uses the addition function, $+$, previously defined.

Again, to prove the usual properties of multiplication as well as the distributivity of $\cdot$ over $+$, we use induction.

Using recursion, we can define many more arithmetic functions. For example, the reader should try defining exponentiation, m^n.

2.6 Inverses of Functions and Relations

Given a function, $f\colon A \to B$ (possibly partial), with $A \neq \emptyset$, suppose there is some function, $g\colon B \to A$ (possibly partial), called a *left inverse of f*, such that

$$g \circ f = \mathrm{id}_A.$$

If such a g exists, we see that f must be total but more is true.

Indeed, assume that $f(a) = f(b)$. Then, by applying g, we get

$$(g \circ f)(a) = g(f(a)) = g(f(b)) = (g \circ f)(b).$$

However, since $g \circ f = \mathrm{id}_A$, we have $(g \circ f)(a) = \mathrm{id}_A(a) = a$ and $(g \circ f)(b) = \mathrm{id}_A(b) = b$, so we deduce that

$$a = b.$$

Therefore, we showed that if a function, f, with nonempty domain, has a left inverse, then f is total and has the property that for all $a, b \in A$, $f(a) = f(b)$ implies that $a = b$, or equivalently $a \neq b$ implies that $f(a) \neq f(b)$.

We say that f is *injective*. As we will see later, injectivity is a very desirable property of functions.

Remark: If $A = \emptyset$, then f is still considered to be injective. In this case, g is the empty partial function (and when $B = \emptyset$, both f and g are the empty function from $\emptyset$ to itself).

Now, suppose there is some function, $h\colon B \to A$ (possibly partial), with $B \neq \emptyset$, called a *right inverse of f*, but this time, we have

$$f \circ h = \mathrm{id}_B.$$

If such an h exists, we see that it must be total but more is true.

Indeed, for any $b \in B$, as $f \circ h = \mathrm{id}_B$, we have

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

Therefore, we showed that if a function, f, with nonempty codomain has a right inverse, h, then h is total and f has the property that for all $b \in B$, there is some $a \in A$, namely, $a = h(b)$, so that $f(a) = b$.

In other words, $\mathrm{Im}\,(f) = B$ or equivalently, every element in B is the image by f of some element of A.

We say that f is *surjective*. Again, surjectivity is a very desirable property of functions.

Remark: If $B = \emptyset$, then f is still considered to be surjective but h is not total unless $A = \emptyset$, in which case f is the empty function from $\emptyset$ to itself.

If a function has a left inverse (respectively a right inverse), then it may have more than one left inverse (respectively right inverse).

If a function (possibly partial), $f\colon A \to B$, with $A, B \neq \emptyset$, happens to have *both a left inverse, $g\colon B \to A$, and a right inverse, $h\colon B \to A$*, then we know that f and h are total.

We claim that $g = h$, so that g is total and moreover g is uniquely determined by f.

Lemma 2.6.1 *Let $f\colon A \to B$ be any function and suppose that f has a left inverse, $g\colon B \to A$, and a right inverse, $h\colon B \to A$. Then, $g = h$ and moreover, g is unique, which means that if $g'\colon B \to A$ is any function which is both a left and a right inverse of f, then $g' = g$.*

This leads to the following definition.

Definition 2.6.2 A function, $f\colon A \to B$, is said to be *invertible* iff there is a function, $g\colon B \to A$, which is both a left inverse and a right inverse, that is,

$$g \circ f = \mathrm{id}_A \quad \text{and} \quad f \circ g = \mathrm{id}_B.$$

In this case, we know that g is unique and it is denoted f^{-1}.

From the above discussion, if a function is invertible, then it is both injective and surjective.

This shows that a function *generally does not have an inverse*.

In order to have an inverse a function needs to be injective and surjective, but this fails to be true for many functions.

It turns out that if a function is injective and surjective then it has an inverse. We will prove this in the next section.

The notion of inverse can also be defined for relations, but it is a somewhat weaker notion.

Definition 2.6.3 Given any relation, $R \subseteq A \times B$, the *converse* or *inverse* of R is the relation, $R^{-1} \subseteq B \times A$, defined by

$$R^{-1} = \{\langle b, a\rangle \in B \times A \mid \langle a, b\rangle \in R\}.$$

In other words, R^{-1} is obtained by swapping A and B and reversing the orientation of the arrows.

Figure 2.7 below shows the inverse of the relation of Figure 2.2:

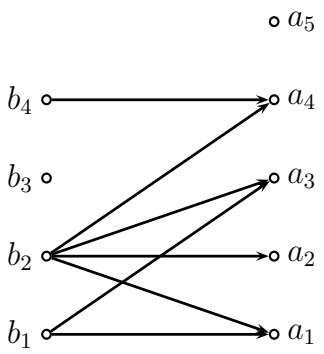


Figure 2.7: The inverse of the relation, R, from Figure 2.2

Now, if R is the graph of a (partial) function, f, beware that R^{-1} is generally *not* the graph of a function at all, because R^{-1} may not be functional.

For example, the inverse of the graph G in Figure 2.3 is *not* functional, see below:

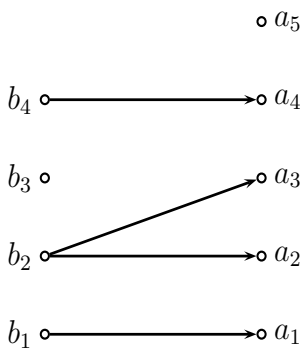


Figure 2.8: The inverse, G^{-1}, of the graph of Figure 2.3

The above example shows that one has to be careful not to view a function as a relation in order to take its inverse.

In general, this process does not produce a function. This only works if the function is invertible.

Given any two relations, $R \subseteq A \times B$ and $S \subseteq B \times C$, the reader should prove that

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}.$$

(Note the switch in the order of composition on the right hand side.)

Similarly, if $f\colon A \to B$ and $g\colon B \to C$ are any two invertible functions, then $g \circ f$ is invertible and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

2.7 Injections, Surjections, Bijections, Permutations

We encountered injectivity and surjectivity in Section 2.6. For the record, let us give

Definition 2.7.1 Given any function, $f\colon A \to B$, we say that *f is injective* (or *one-to-one*) iff for all $a, b \in A$, if $f(a) = f(b)$, then $a = b$, or equivalently, if $a \neq b$, then $f(a) \neq f(b)$.

We say that *f is surjective* (or *onto*) iff for every $b \in B$, there is some $a \in A$ so that $b = f(a)$, or equivalently if $\mathrm{Im}\,(f) = B$.

The function f is *bijective* iff it is both injective and surjective. When $A = B$, a bijection $f\colon A \to A$ is called a *permutation of A*.

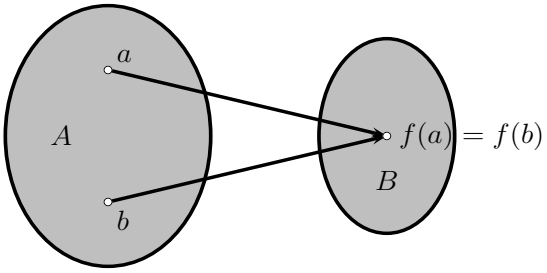


Figure 2.9: A non-injective function

Remarks:

1. If $A = \emptyset$, then any function, $f : \emptyset \to B$ is (trivially) injective.
2. If $B = \emptyset$, then f is the empty function from $\emptyset$ to itself and it is (trivially) surjective.
3. A function, $f : A \to B$, is **not injective** iff **there exist** $a, b \in A$ with $a \neq b$ and **yet** $f(a) = f(b)$, see Figure 2.9.

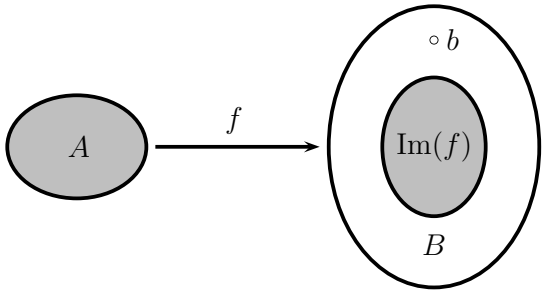


Figure 2.10: A non-surjective function

4. A function, $f\colon A \to B$, is **not surjective** iff **for some** $b \in B$, **there is no** $a \in A$ with $b = f(a)$, see Figure 2.10.

5. Since $\operatorname{Im} f = \{b \in B \mid (\exists a \in A)(b = f(a))\}$, a function $f\colon A \to B$ is always surjective onto its image.

6. The notation $f\colon A \hookrightarrow B$ is often used to indicate that a function, $f\colon A \to B$, is an injection.

7. If $A \neq \emptyset$, a function, $f\colon A \to B$, is injective iff for every $b \in B$, there *at most one* $a \in A$ such that $b = f(a)$.

8. If $A \neq \emptyset$, a function, $f\colon A \to B$, is surjective iff for every $b \in B$, there *at least one* $a \in A$ such that $b = f(a)$ iff $f^{-1}(b) \neq \emptyset$ for all $b \in B$.

9. If $A \neq \emptyset$, a function, $f\colon A \to B$, is bijective iff for every $b \in B$, there is *a unique* $a \in A$ such that $b = f(a)$.

10. When A is the finite set $A = \{1, \ldots, n\}$, also denoted $[n]$, it is not hard to show that there are $n!$ permutations of $[n]$.

The function, $f_1 \colon \mathbb{Z} \to \mathbb{Z}$, given by $f_1(x) = x + 1$ is injective and surjective.

However, the function, $f_2 \colon \mathbb{Z} \to \mathbb{Z}$, given by $f_2(x) = x^2$ is neither injective nor surjective (why?).

The function, $f_3 \colon \mathbb{Z} \to \mathbb{Z}$, given by $f_3(x) = 2x$ is injective but not surjective.

The function, $f_4 \colon \mathbb{Z} \to \mathbb{Z}$, given by

$$f_4(x) = \begin{cases} k & \text{if } x = 2k \\ k & \text{if } x = 2k + 1 \end{cases}$$

is surjective but not injective.

Remark: The reader should prove that if A and B are finite sets, A has m elements and B has n elements $(m \leq n)$ then the set of injections from A to B has

$$\frac{n!}{(n-m)!}$$

elements.

The following Theorem relates the notions of injectivity and surjectivity to the existence of left and right inverses.

Theorem 2.7.2 *Let $f\colon A \to B$ be any function and assume $A \neq \emptyset$.*

(a) The function f is injective iff it has a left inverse, g (i.e., a function $g\colon B \to A$ so that $g \circ f = \mathrm{id}_A$).

(b) The function f is surjective iff it has a right inverse, h (i.e., a function $h\colon B \to A$ so that $f \circ h = \mathrm{id}_B$).

(c) The function f is invertible iff it is injective and surjective.

The alert reader may have noticed a "fast turn" in the proof of the converse in (b). Indeed, we constructed the function h by choosing, for each $b \in B$, some element in $f^{-1}(b)$. How do we justify this procedure from the axioms of set theory?

Well, we can't! For this, we need another (historically somewhat controversial) axiom, the *axiom of choice*.

This axiom has many equivalent forms. We state the following form which is intuitively quite plausible:

Axiom of Choice (Graph Version).

For every relation, $R \subseteq A \times B$, there is a partial function, $f : A \to B$, with $graph(f) \subseteq R$ and $dom(f) = dom(R)$.

We see immediately that the axiom of choice justifies the existence of the function h in part (b) of Theorem 2.7.2.

Remarks:

1. Let $f\colon A \to B$ and $g\colon B \to A$ be any two functions and assume that

$$g \circ f = \mathrm{id}_A.$$

   Thus, f is a right inverse of g and g is a left inverse of f. So, by Theorem 2.7.2 (a) and (b), we deduce that f is injective and g is surjective. In particular, this shows that any left inverse of an injection is a surjection and that any right inverse of a surjection is an injection.

2. Any right inverse, h, of a surjection, $f\colon A \to B$, is called a *section* of f (which is an abbreviation for *cross-section*).

   This terminology can be better understood as follows: Since f is surjective, the preimage, $f^{-1}(b) = \{a \in A \mid f(a) = b\}$ of any element $b \in B$ is nonempty.

   Moreover, $f^{-1}(b_1) \cap f^{-1}(b_2) = \emptyset$ whenever $b_1 \neq b_2$.

Therefore, the pairwise disjoint and nonempty subsets, $f^{-1}(b)$, where $b \in B$, partition A.

We can think of A as a big "blob" consisting of the union of the sets $f^{-1}(b)$ (called fibres) and lying over B.

The function f maps each fibre, $f^{-1}(b)$ onto the element, $b \in B$.

Then, any right inverse, $h\colon B \to A$, of f picks out some element in each fibre, $f^{-1}(b)$, forming a sort of horizontal section of A shown as a curve in Figure 2.11.

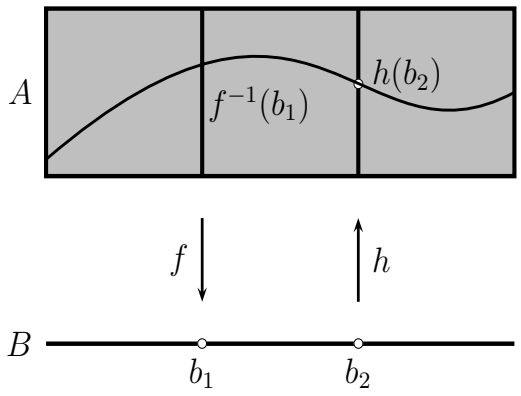


Figure 2.11: A section, h, of a surjective function, f.

3. Any left inverse, g, of an injection, $f\colon A \to B$, is called a *retraction* of f.

   The terminology reflects the fact that intuitively, as f is injective (thus, g is surjective), B is bigger than A and since $g \circ f = \mathrm{id}_A$, the function g "squeezes" B onto A in such a way that each point $b = f(a)$ in $\mathrm{Im}\, f$ is mapped back to its ancestor $a \in A$. So, B is "retracted" onto A by g.

Before discussing direct and inverse images, we define the notion of restriction and extension of functions.

Definition 2.7.3 Given two functions, $f\colon A \to C$ and $g\colon B \to C$, with $A \subseteq B$, we say that *f is the restriction of g to A* if $graph(f) \subseteq graph(g)$; we write $f = g \restriction A$. In this case, we also say that *g is an extension of f to B*.

2.8 Direct Image and Inverse Image

A function, $f : X \to Y$, induces a function from 2^X to 2^Y also denoted f and a function from 2^Y to 2^X, as shown in the following definition:

Definition 2.8.1 Given any function, $f : X \to Y$, we define the function $f : 2^X \to 2^Y$ so that, for every subset A of X,

$$f(A) = \{y \in Y \mid \exists x \in A,\ y = f(x)\}.$$

The subset, $f(A)$, of Y is called the *direct image of A under f*, for short, the *image of A under f*. We also define the function $f^{-1} : 2^Y \to 2^X$ so that, for every subset B of Y,

$$f^{-1}(B) = \{x \in X \mid \exists y \in B,\ y = f(x)\}.$$

The subset, $f^{-1}(B)$, of X is called the *inverse image of B under f* or the *preimage of B under f*.

Remarks:

1. The overloading of notation where f is used both for denoting the original function $f\colon X \to Y$ and the new function $f\colon 2^X \to 2^Y$ may be slightly confusing.

   If we observe that $f(\{x\}) = \{f(x)\}$, for all $x \in X$, we see that the new f is a natural extension of the old f to the subsets of X and so, using the same symbol f for both functions is quite natural after all.

   To avoid any confusion, some authors (including Enderton) use a different notation for $f(A)$, for example, $f[\![A]\!]$.

   We prefer not to introduce more notation and we hope that the context will make it clear which f we are dealing with.

2. The use of the notation f^{-1} for the function $f^{-1}\colon 2^Y \to 2^X$ may even be more confusing, because we know that f^{-1} is generally not a function from Y to X.

   However, it *is* a function from 2^Y to 2^X. Again, some authors use a different notation for $f^{-1}(B)$, for example, $f^{-1}[\![A]\!]$. We will stick to $f^{-1}(B)$.

3. The set $f(A)$ is sometimes called the *push-forward of A along f* and $f^{-1}(B)$ is sometimes called the *pullback of B along f*.

4. Observe that $f^{-1}(y) = f^{-1}(\{y\})$, where $f^{-1}(y)$ is the preimage defined just after Definition 2.2.3.

5. Although this may seem counter-intuitive, the function f^{-1} has a better behavior than f with respect to union, intersection and complementation.

Proposition 2.8.2 *Given any function, $f\colon X \to Y$, the following properties hold:*

(1) For any $B \subseteq Y$, we have

$$f(f^{-1}(B)) \subseteq B.$$

(2) If $f\colon X \to Y$ is surjective, then

$$f(f^{-1}(B)) = B.$$

(3) For any $A \subseteq X$, we have

$$A \subseteq f^{-1}(f(A)).$$

(4) If $f\colon X \to Y$ is injective, then

$$A = f^{-1}(f(A)).$$

The next proposition deals with the behavior of $f\colon 2^X \to 2^Y$ and $f^{-1}\colon 2^Y \to 2^X$ with respect to union, intersection and complementation.

Proposition 2.8.3 *Given any function, $f\colon X \to Y$, the following properties hold:*

(1) For all $A, B \subseteq X$, we have

$$f(A \cup B) = f(A) \cup f(B).$$

(2)

$$f(A \cap B) \subseteq f(A) \cap f(B).$$

Equality holds if $f\colon X \to Y$ is injective.

(3)

$$f(A) - f(B) \subseteq f(A - B).$$

Equality holds if $f\colon X \to Y$ is injective.

(4) For all $C, D \subseteq Y$, we have

$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D).$$

(5)

$$f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D).$$

(6)

$$f^{-1}(C - D) = f^{-1}(C) - f^{-1}(D).$$

As we can see from Proposition 2.8.3, the function $f^{-1}\colon 2^Y \to 2^X$ has a better behavior than $f\colon 2^X \to 2^Y$ with respect to union, intersection and complementation.

2.9 Equinumerosity; The Pigeonhole Principle and the Schröder–Bernstein Theorem

The notion of *size of a set* is fairly intuitive for finite sets but what does it mean for infinite sets?

How do we give a precise meaning to the questions:

(a) Do X and Y have the same size?

(b) Does X have more elements than Y?

For finite sets, we can rely on the natural numbers. We count the elements in the two sets and compare the resulting numbers.

If one of the two sets is finite and the other is infinite, it seems fair to say that the infinite set has more elements than the finite one.

But what if both sets are infinite?

Remark: A critical reader should object that we have not yet defined what a finite set is (or what an infinite set is).

Indeed, we have not!

This can be done in terms of the natural numbers but, for the time being, we will rely on intuition.

We should also point out that when it comes to infinite sets, experience shows that *our intuition fails us miserably*. So, we should be very careful.

Let us return to the case where we have two infinite sets.

For example, consider $\mathbb{N}$ and the set of even natural numbers, $2\mathbb{N} = \{0, 2, 4, 6, \ldots\}$. Clearly, the second set is properly contained in the first.

Does that make $\mathbb{N}$ bigger?

On the other hand, the function $n \mapsto 2n$ is a *bijection* between the two sets, which seems to indicate that they have the same number of elements.

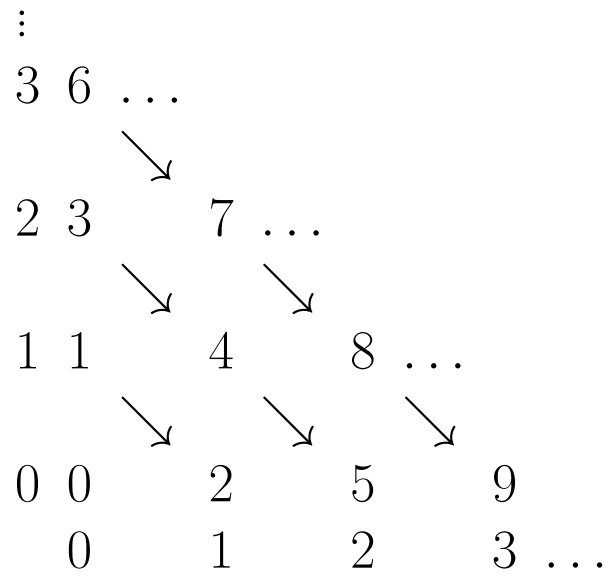
Similarly, the set of squares of natural numbers, Squares $= \{0, 1, 4, 9, 16, 25, \ldots\}$ is properly contained in $\mathbb{N}$ and many natural numbers are missing from Squares.

But, the map $n \mapsto n^2$ is a bijection between $\mathbb{N}$ and Squares, which seems to indicate that they have the same number of elements.

A more extreme example is provided by $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$.

Intuitively, $\mathbb{N} \times \mathbb{N}$ is two-dimensional and $\mathbb{N}$ is one-dimensional, so $\mathbb{N}$ seems much smaller than $\mathbb{N} \times \mathbb{N}$.

However, it is possible to construct bijections between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$ (try to find one!). In fact, such a function, J, has the graph partially showed below:



The function J corresponds to a certain way of enumerating pairs of integers.

Note that the value of $m + n$ is constant along each diagonal, and consequently, we have

$$\begin{aligned} J(m,n) &= 1 + 2 + \cdots + (m+n) + m, \\ &= ((m+n)(m+n+1) + 2m)/2, \\ &= ((m+n)^2 + 3m + n)/2. \end{aligned}$$

For example,

$J(2,1) = ((2+1)^2+3\cdot 2+1)/2 = (9+6+1)/2 = 16/2 = 8.$

The function

$$J(m,n) = \frac{1}{2}((m+n)^2 + 3m + n)$$

is a bijection but that's not so easy to prove!

Perhaps even more surprising, there are bijections between $\mathbb{N}$ and $\mathbb{Q}$. What about between $\mathbb{R} \times \mathbb{R}$ and $\mathbb{R}$?

Again, the answer is yes, but that's harder to prove.

These examples suggest that the notion of bijection can be used to define rigorously when two sets have the same size.

This leads to the concept of equinumerosity.

Definition 2.9.1 A set A is *equinumerous* to a set B, written $A \approx B$, iff there is a bijection $f : A \to B$.

We say that A is *dominated* by B, written $A \preceq B$, iff there is an injection from A to B.

Finally, we say that A is *strictly dominated* by B, written $A \prec B$, iff $A \preceq B$ and $A \not\approx B$.

Using the above concepts, we can give a precise definition of finiteness.

Firstly, recall that for any $n \in \mathbb{N}$, we defined $[n]$ as the set $[n] = \{1, 2, \ldots, n\}$, with $[0] = \emptyset$.

Definition 2.9.2 A set, A, is *finite* if it is equinumerous to a set of the form $[n]$, for some $n \in \mathbb{N}$. A set, A, is *infinite* iff it is not finite. We say that A is *countable* (or *denumerable*) iff A is dominated by $\mathbb{N}$.

Two pretty results due to Cantor (1873) are given in the next Theorem.

These are among the earliest results of set theory.

We assume that the reader is familiar with the fact that every number, $x \in \mathbb{R}$, can be expressed in decimal expansion (possibly infinite).

For example,

$$\pi = 3.14159265358979\cdots$$

Theorem 2.9.3 *(Cantor's Theorem) (a) The set, $\mathbb{N}$, is not equinumerous to the set, $\mathbb{R}$, of real numbers.*

(b) For every set, A, there is no surjection from A onto 2^A. Consequently, no set, A, is equinumerous to its power set, 2^A.

The proof of (a) uses a famous proof method due to Cantor and known as a *diagonal argument*.

As there is an obvious injection of $\mathbb{N}$ into $\mathbb{R}$, Theorem 2.9.3 shows that $\mathbb{N}$ is strictly dominated by $\mathbb{R}$.

Also, as we have the injection $a \mapsto \{a\}$ from A into 2^A, we see that every set is strictly dominated by its power set.

So, we can form sets as big as we want by repeatedly using the power set operation.

Remark: In fact, $\mathbb{R}$ is equinumerous to $2^{\mathbb{N}}$, but we will not prove this here.

The following proposition shows an interesting connection between the notion of power set and certain sets of functions.

To state this proposition, we need the concept of characteristic function of a subset.

Given any set, X, for any subset, A, of X, define the *characteristic function of A*, denoted χ_A, as the function, $\chi_A \colon X \to \{0, 1\}$, given by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

In other words, χ_A tests membership in A: For any $x \in X$, $\chi_A(x) = 1$ iff $x \in A$.

Observe that we obtain a function, $\chi \colon 2^X \to \{0, 1\}^X$, from the power set of X to the set of characteristic functions from X to $\{0, 1\}$, given by

$$\chi(A) = \chi_A.$$

We also have the function, $\mathcal{S} \colon \{0, 1\}^X \to 2^X$, mapping any characteristic function to the set that it defines and given by

$$\mathcal{S}(f) = \{x \in X \mid f(x) = 1\},$$

for every characteristic function, $f \in \{0, 1\}^X$.

Proposition 2.9.4 *For any set, X, the function, $\chi\colon 2^X \to \{0,1\}^X$, from the power set of X to the set of characteristic functions on X is a bijection whose inverse is $\mathcal{S}\colon \{0,1\}^X \to 2^X$.*

In view of Proposition 2.9.4, there is a bijection between the power set 2^X and the set of functions in $\{0,1\}^X$.

If we write $2 = \{0,1\}$, then we see that the two sets looks the same!

This is the reason why the notation 2^X is often used for the power set (but others prefer $\mathcal{P}(X)$).

There are many other interesting results about equinumerosity. We only mention four more, all very important.

Theorem 2.9.5 *(Pigeonhole Principle) No set of the form $[n]$ is equinumerous to a proper subset of itself, where $n \in \mathbb{N}$,*

Although the Pigeonhole Principle seems obvious, the proof is not. In fact, the proof requires induction.

Corollary 2.9.6 *(Pigeonhole Principle for finite sets) No finite set is equinumerous to a proper subset of itself.*

The pigeonhole principle is often used in the following way:

If we have m distinct slots and $n > m$ distinct objects (the pigeons), then when we put all n objects into the m slots, two objects must end up in the same slot.

Figure 2.12: Johan Peter Gutav Lejeune Dirichlet, 1805-1859

This fact was apparently first stated explicitly by Dirichlet in 1834. As such, it is also known as *Dirichlet's box principle*.

Let A be a finite set. Then, by definition, there is a bijection, $f : A \to [n]$, for some $n \in \mathbb{N}$.

We claim that such an n is unique.

If A is a finite set, the unique natural number, $n \in \mathbb{N}$, such that $A \approx [n]$ is called the *cardinality of A* and we write $|A| = n$ (or sometimes, $\text{card}(A) = n$).

Remark: The notion of cardinality also makes sense for infinite sets.

What happens is that every set is equinumerous to a special kind of set (an initial ordinal) called a *cardinal number* but this topic is beyond the scope of this course.

Let us simply mention that the cardinal number of $\mathbb{N}$ is denoted $\aleph_0$ (say "aleph" 0).

Corollary 2.9.7 *(a) Any set equinumerous to a proper subset of itself is infinite.*

(b) The set $\mathbb{N}$ *is infinite.*

The image of a finite set by a function is also a finite set. In order to prove this important property we need the following two propositions:

Proposition 2.9.8 *Let* n *be any positive natural number, let* A *be any nonempty set and pick any element,* $a_0 \in A$*. Then there exists a bijection,* $f\colon A \to [n+1]$*, iff there exists a bijection,* $g\colon (A - \{a_0\}) \to [n]$*.*

Proposition 2.9.9 *For any function, $f\colon A \to B$, if f is surjective and if A is a finite nonempty set, then B is also a finite set and there is an injection, $h\colon B \to A$, such that $f \circ h = \mathrm{id}_B$. Moreover, $|B| \leq |A|$.*

Instead of using Theorem 2.7.2 (b), which relies on the Axiom of Choice, the proof of Proposition 2.9.9 proceeds by induction on the cardinality of A.

Corollary 2.9.10 *For any function, $f\colon A \to B$, if A is a finite set, then the image, $f(A)$, of f is also finite and $|f(A)| \leq |A|$.*

Corollary 2.9.11 *For any two sets, A and B, if B is a finite set of cardinality n and is A is a proper subset of B, then A is also finite and A has cardinality $m < n$.*

If A is an infinite set, then the image, $f(A)$, is not finite in general but we still have the following fact:

Proposition 2.9.12 *For any function, $f\colon A \to B$, we have $f(A) \preceq A$, that is, there is an injection from the image of f to A.*

Here are two more important facts that follow from the Pigeonhole Principle for finite sets and Proposition 2.9.9.

Proposition 2.9.13 *Let A be any finite set. For any function, $f\colon A \to A$, the following properties hold:*

(a) If f is injective, then f is a bijection.

(b) If f is surjective, then f is a bijection.

The proof of Proposition 2.9.13 is left as an exercise (use Corollary 2.9.6 and Proposition 2.9.9).

Proposition 2.9.13 *only holds for finite sets*.

Indeed, just after the remarks following Definition 2.7.1 we gave examples of functions defined on an infinite set for which Proposition 2.9.13 fails.

A convenient characterization of countable sets is stated below:

Proposition 2.9.14 *A nonempty set, A, is countable iff there is a surjection, $g\colon \mathbb{N} \to A$, from $\mathbb{N}$ onto A.*

The following fact about infinite sets is also useful to know:

Theorem 2.9.15 *For every infinite set, A, there is an injection from $\mathbb{N}$ into A.*

The proof of Theorem 2.9.15 is actually quite tricky.

It requires a version of the axiom of choice and a subtle use of the Recursion Theorem (Theorem 2.5.1).

The intuitive content of Theorem 2.9.15 is that $\mathbb{N}$ *is the "smallest" infinite set*.

An immediate consequence of Theorem 2.9.15 is that every infinite subset of $\mathbb{N}$ is equinumerous to $\mathbb{N}$.

Here is a characterization of infinite sets originally proposed by Dedekind in 1888.

Proposition 2.9.16 *A set, A, is infinite iff it is equinumerous to a proper subset of itself.*

Let us give another application of the pigeonhole principle involving sequences of integers.

Given a finite sequence, S, of integers, $a_1, \ldots, a_n$, a *subsequence of S* is a sequence, $b_1, \ldots, b_m$, obtained by deleting elements from the original sequence and keeping the remaining elements in the same order as they originally appeared.

More precisely, $b_1, \ldots, b_m$ is a subsequence of $a_1, \ldots, a_n$ if there is an injection, $g\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, such that $b_i = a_{g(i)}$ for all $i \in \{1, \ldots, m\}$ and $i \leq j$ implies $g(i) \leq g(j)$ for all $i, j \in \{1, \ldots, m\}$.

For example, the sequence

$$1 \quad \mathbf{9} \quad 10 \quad \mathbf{8} \quad 3 \quad 7 \quad 5 \quad 2 \quad \mathbf{6} \quad \mathbf{4}$$

contains the subsequence

$$9 \quad 8 \quad 6 \quad 4.$$

An *increasing subsequence* is a subsequence whose elements are in strictly increasing order and a *decreasing subsequence* is a subsequence whose elements are in strictly decreasing order.

For example, 9 8 6 4 is a decreasing subsequence of our original sequence.

We now prove the following beautiful result due to Erdös and Szekeres:

Theorem 2.9.17 *(Erdös and Szekeres) Let n be any nonzero natural number. Every sequence of $n^2 + 1$ pairwise distinct natural numbers must contain either an increasing subsequence or a decreasing subsequence of length $n + 1$.*

Remark: The proof is not constructive in the sense that it does not produce the desired subsequence; it merely asserts that such a sequence exists.

Our next theorem is the historically famous Schröder-Bernstein Theorem, sometimes called the "Cantor-Bernstein Theorem."

Cantor proved the theorem in 1897 but his proof used a principle equivalent to the axiom of choice.

Schröder announced the theorem in an 1896 abstract. His proof, published in 1898, had problems and he published a correction in 1911.

The first fully satisfactory proof was given by Felix Bernstein and was published in 1898 in a book by Emile Borel.

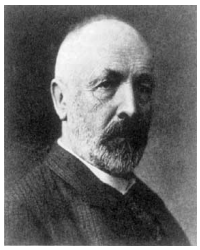



Figure 2.13: Georg Cantor, 1845-1918 (left), Ernst Schröder, 1841-1902 (middle left), Felix Bernstein, 1878-1956 (middle right) and Emile Borel, 1871-1956 (right)

A shorter proof was given later by Tarski (1955) as a consequence of his fixed point theorem. We postpone giving this proof until the section on lattices (see Section 5.2).

Theorem 2.9.18 *(Schröder-Bernstein Theorem) Given any two sets, A and B, if there is an injection from A to B and an injection from B to A, then there is a bijection between A and B. Equivalently, if $A \preceq B$ and $B \preceq A$, then $A \approx B$.*

The Schröder-Bernstein Theorem is quite a remarkable result and it is a main tool to develop cardinal arithmetic, a subject beyond the scope of this course.

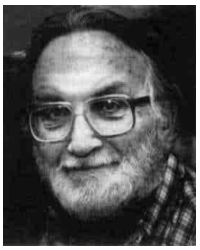


Figure 2.14: Max August Zorn, 1906-1993

Our third theorem is perhaps the one that is the more surprising from an intuitive point of view. If nothing else, it shows that our intuition about infinity is rather poor.

Theorem 2.9.19 *If A is any infinite set, then $A \times A$ is equinumerous to A.*

The proof is more involved than any of the proofs given so far and it makes use of the axiom of choice in the form known as *Zorn's Lemma* (see Theorem 5.1.3).

In particular, Theorem 2.9.19 implies that $\mathbb{R} \times \mathbb{R}$ is in bijection with $\mathbb{R}$.

But, geometrically, $\mathbb{R} \times \mathbb{R}$ is a plane and $\mathbb{R}$ is a line and, intuitively, it is surprising that a plane and a line would have "the same number of points."

Nevertheless, that's what mathematics tells us!

Our fourth theorem also plays an important role in the theory of cardinal numbers.

Theorem 2.9.20 *(Cardinal comparability) Given any two sets, A and B, either there is an injection from A to B or there is an injection from B to A (that is, either $A \preceq B$ or $B \preceq A$).*

The proof requires the axiom of choice in a form known as the *Well-Ordering Theorem*, which is also equivalent to Zorn's lemma. For details, see Enderton [5] (Chapters 6 and 7). □

Theorem 2.9.19 implies that there is a bijection between the closed line segment

$$[0,1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

and the closed unit square

$$[0,1] \times [0,1] = \{(x,y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq 1\}$$

As an interlude, in the next section, we describe a famous space-filling function due to Hilbert.

Such a function is obtained as the limit of a sequence of curves that can be defined recursively.

2.10 An Amazing Surjection: Hilbert's Space Filling Curve

In the years 1890-1891, Giuseppe Peano and David Hilbert discovered examples of *space filling functions* (also called *space filling curves*).

These are surjective functions from the line segment, $[0, 1]$ onto the unit square and thus, their image is the whole unit square!

Such functions defy intuition since they seem to contradict our intuition about the notion of dimension, a line segment is one-dimensional, yet the unit square is two-dimensional.

They also seem to contradict our intuitive notion of area.


Figure 2.15: David Hilbert 1862-1943 and Waclaw Sierpinski, 1882-1969

Nevertheless, such functions do exist, even continuous ones, although to justify their existence rigouroulsy requires some tools from mathematical analysis.

Similar curves were found by others, among which we mention Sierpinski, Moore and Gosper.

We will describe Hilbert's scheme for constructing such a square-filling curve.

We define a sequence, (h_n), of polygonal lines, $h_n\colon [0,1] \to [0,1]\times[0,1]$, starting from the simple pattern h_0 (a "square cap" $\sqcap$) shown on the left in Figure 2.16.

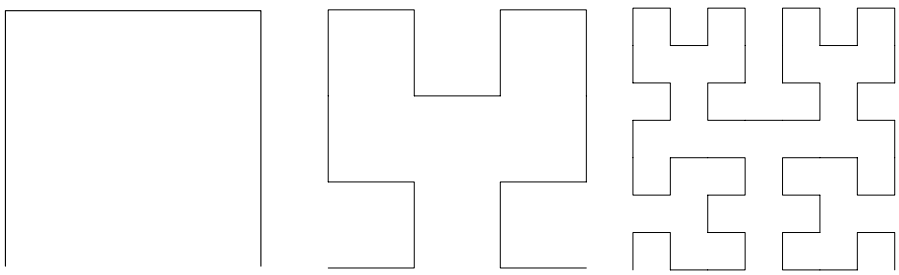


Figure 2.16: A sequence of Hilbert curves h_0, h_1, h_2

The curve h_{n+1} is obtained by scaling down h_n by a factor of $\frac{1}{2}$, and connecting the four copies of this scaled–down version of h_n obtained by rotating by $\pi/2$ (left lower part), rotating by $-\pi/2$ and translating right (right lower part), translating up (left upper part), and translating diagonally (right upper part), as illustrated in Figure 2.16.

It can be shown that the sequence (h_n) converges (uniformly) to a continuous curve $h\colon [0,1] \to [0,1] \times [0,1]$ whose trace is the entire square $[0,1] \times [0,1]$.

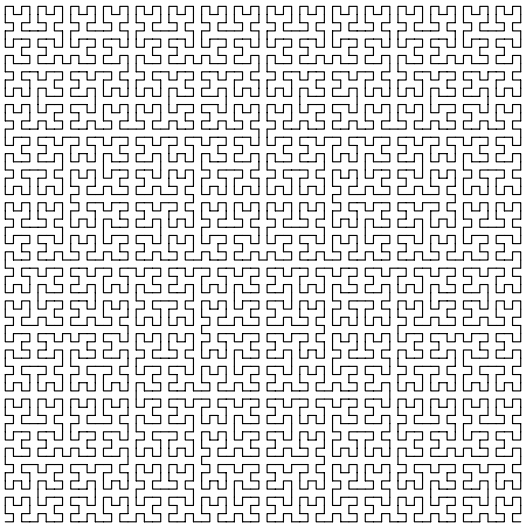


Figure 2.17: The Hilbert curve h_5

The Hilbert curve h is surjective, continuous, and nowhere differentiable. It also has infinite length!

The curve h_5 is shown in Figure 2.17.

You should try writing a computer program to plot these curves!

By the way, it can be shown that no continuous square-filling function can be injective.

It is also possible to define cube-filling curves and even higher-dimensional cube-filling curves! (see some of the web page links in the home page for CIS260)

2.11 Strings, Multisets, Indexed Families

Strings play an important role in computer science and linguistics because they are the basic tokens that languages are made of.

In fact, formal language theory takes the (somewhat crude) view that a language is a set of strings (you will study some formal language theory in CIS262).

A string is a finite sequence of letters, for example "Jean", "Val", "Mia", "math", "gaga", "abab".

Usually, we have some alphabet in mind and we form strings using letters from this alphabet.

Strings are not sets, the order of the letters matters: "abab" and "baba" are different strings.

What matters is the position of every letter. In the string "aba", the leftmost "a" is in position 1, "b" is in position 2 and the rightmost "b" is in position 3.

All this suggests defining strings as certain kinds of functions whose domains are the sets $[n] = \{1, 2, \ldots, n\}$ (with $[0] = \emptyset$) encountered earlier. Here is the very beginning of the theory of formal languages.

Definition 2.11.1 An *alphabet*, Σ, is any **finite** set.

We often write $\Sigma = \{a_1, \ldots, a_k\}$. The a_i are called the *symbols* of the alphabet.

Examples:

$\Sigma = \{a\}$

$\Sigma = \{a, b, c\}$

$\Sigma = \{0, 1\}$

A string is a finite sequence of symbols. Technically, it is convenient to define strings as functions.

Definition 2.11.2 Given an alphabet, Σ, a *string over Σ (or simply a string) of length n* is any function

$$u\colon [n] \to \Sigma.$$

The integer n is the *length* of the string, u, and it is denoted by $|u|$. When $n = 0$, the special string, $u\colon [0] \to \Sigma$, of length 0 is called the *empty string, or null string*, and is denoted by ϵ.

Given a string, $u\colon [n] \to \Sigma$, of length $n \geq 1$, $u(i)$ is the i-th letter in the string u.

For simplicity of notation, we denote the string u as

$$u = u_1 u_2 \dots u_n,$$

with each $u_i \in \Sigma$.

For example, if $\Sigma = \{a, b\}$ and $u\colon [3] \to \Sigma$ is defined such that $u(1) = a$, $u(2) = b$, and $u(3) = a$, we write

$$u = aba.$$

Strings of length 1 are functions $u\colon [1] \to \Sigma$ simply picking some element $u(1) = a_i$ in Σ.

Thus, we will identify every symbol $a_i \in \Sigma$ with the corresponding string of length 1.

The set of all strings over an alphabet Σ, including the empty string, is denoted as Σ^*.

Observe that when $\Sigma = \emptyset$, then

$$\emptyset^* = \{\epsilon\}.$$

When $\Sigma \neq \emptyset$, the set Σ^* is countably infinite. Later on, we will see ways of ordering and enumerating strings.

Strings can be juxtaposed, or concatenated.

Definition 2.11.3 Given an alphabet, Σ, given two strings, $u\colon [m] \to \Sigma$ and $v\colon [n] \to \Sigma$, the *concatenation, $u \cdot v$, (also written uv) of u and v* is the string, $uv\colon [m+n] \to \Sigma$, defined such that

$$uv(i) = \begin{cases} u(i) & \text{if } 1 \leq i \leq m, \\ v(i-m) & \text{if } m+1 \leq i \leq m+n. \end{cases}$$

In particular, $u\epsilon = \epsilon u = u$.

It is immediately verified that

$$u(vw) = (uv)w.$$

Thus, concatenation is a binary operation on Σ^* which is associative and has ϵ as an identity.

Note that generally, $uv \neq vu$, for example for $u = a$ and $v = b$.

Definition 2.11.4 Given an alphabet Σ, given any two strings $u, v \in \Sigma^*$ we define the following notions as follows:

u is a prefix of v iff there is some $y \in \Sigma^*$ such that

$$v = uy.$$

u is a suffix of v iff there is some $x \in \Sigma^*$ such that

$$v = xu.$$

u is a substring of v iff there are some $x, y \in \Sigma^*$ such that

$$v = xuy.$$

We say that *u is a proper prefix (suffix, substring) of v* iff u is a prefix (suffix, substring) of v and $u \neq v$.

For example, ga is a prefix of $gallier$, the string $lier$ is a suffix of $gallier$ and all is a substring of $gallier$

Finally, languages are defined as follows.

Definition 2.11.5 Given an alphabet Σ, a *language over Σ (or simply a language)* is any subset, L, of Σ^*.

The next step would be to introduce various formalisms to define languages, such as automata or grammars but you'll have to take CIS262 to learn about these things!

We now consider multisets. We already encountered multisets in Section 1.2 when we defined the axioms of propositional logic.

As for sets, in a multiset, *the order of elements does not matter*, but as in strings, multiple occurrences of elements matter.

For example,

$$\{a, a, b, c, c, c\}$$

is a multiset with two occurrences of a, one occurrence of b and three occurrences of c.

This suggests defining a multiset as a function with range $\mathbb{N}$, to specify the multiplicity of each element.

Definition 2.11.6 Given any set, S, a *multiset, M, over S* is any function, $M : S \to \mathbb{N}$. A *finite multiset, M, over S* is any function, $M : S \to \mathbb{N}$, such that $M(a) \neq 0$ only for finitely many $a \in S$. If $M(a) = k > 0$, we say that a *appears with mutiplicity k in M*.

For example, if $S = \{a, b, c\}$, we may use the notation $\{a, a, a, b, c, c\}$ for the multiset where a has multiplicity 3, b has multiplicity 1, and c has multiplicity 2.

The empty multiset is the function having the constant value 0.

The *cardinality* $|M|$ of a (finite) multiset is the number

$$|M| = \sum_{a \in S} M(a).$$

Note that this is well-defined since $M(a) = 0$ for all but finitely many $a \in S$. For example

$$|\{a, a, a, b, c, c\}| = 6.$$

We can define the *union* of multisets as follows: If M_1 and M_2 are two multisets, then $M_1 \cup M_2$ is the multiset given by

$$(M_1 \cup M_2)(a) = M_1(a) + M_2(a), \quad \text{for all} \quad a \in S.$$

A multiset, M_1, is a *submultiset* of a multiset, M_2, if $M_1(a) \leq M_2(a)$, for all $a \in S$.

The *difference of M_1 and M_2* is the multiset, $M_1 - M_2$, given by

$$(M_1 - M_2)(a) = \begin{cases} M_1(a) - M_2(a) & \text{if } M_1(a) \geq M_2(a) \\ 0 & \text{if } M_1(a) < M_2(a). \end{cases}$$

Intersection of multisets can also be defined but we will leave this as an exercise.

Let us now discuss indexed families.

The Cartesian product construct, $A_1 \times A_2 \times \cdots \times A_n$, allows us to form finite indexed sequences, $\langle a_1, \ldots, a_n \rangle$, but there are situations where we need to have infinite indexed sequences.

Typically, we want to be able to consider families of elements indexed by some index set of our choice, say I.

We can do this as follows:

Definition 2.11.7 Given any, X, and any other set, I, called the *index set*, the set of *I-indexed families (or sequences) of elements from X* is the set of all functions, $A \colon I \to X$; such functions are usually denoted $A = (A_i)_{i \in I}$.

When X is a set of sets, each A_i is some set in X and we call $(A_i)_{i \in I}$ a *family of sets (indexed by I)*.

Observe that if $I = [n] = \{1, \ldots, n\}$, then an I-indexed family is just a string over X.

When $I = \mathbb{N}$, an $\mathbb{N}$-indexed family is called an *infinite sequence* or often just a *sequence*.

In this case, we usually write (x_n) for such a sequence ($(x_n)_{n\in\mathbb{N}}$, if we want to be more precise).

Also, note that although the notion of indexed family may seem less general than the notion of arbitrary collection of sets, this is an illusion.

Indeed, given any collection of sets, X, we may choose the set index set I to be X itself, in wich case X appears as the range of the identity function, $\mathrm{id}\colon X \to X$.

The point of indexed families is that the operations of union and intersection can be generalized in an interesting way.

We can also form infinite Cartesian products, which are very useful in algebra and geometry.

Given any indexed family of sets, $(A_i)_{i\in I}$, the *union of the family* $(A_i)_{i\in I}$, denoted $\bigcup_{i\in I} A_i$, is simply the union of the range of A, that is,

$$\bigcup_{i\in I} A_i = \bigcup range(A) = \{a \mid (\exists i \in I),\ a \in A_i\}.$$

Observe that when $I = \emptyset$, the union of the family is the empty set.

When $I \neq \emptyset$, we say that we have a *nonempty family* (even though some of the A_i may be empty).

Similarly, if $I \neq \emptyset$, then the *intersection of the family,* $(A_i)_{i\in I}$, denoted $\bigcap_{i\in I} A_i$, is simply the intersection of the range of A, that is,

$$\bigcap_{i\in I} A_i = \bigcap range(A) = \{a \mid (\forall i \in I),\ a \in A_i\}.$$

Unlike the situation for union, when $I = \emptyset$, the intersection of the family does not exist. It would be the set of all sets, which does not exist.

It is easy to see that the laws for union, intersection and complementation generalize to families but we will leave this to the exercises.

An important construct generalizing the notion of finite Cartesian product is the product of families.

Definition 2.11.8 Given any family of sets, $(A_i)_{i \in I}$, the *product of the family* $(A_i)_{i \in I}$, denoted $\prod_{i \in I} A_i$, is the set

$$\prod_{i \in I} A_i = \{a \colon I \to \bigcup_{i \in I} A_i \mid (\forall i \in I),\ a(i) \in A_i\}.$$

Definition 2.11.8 says that the elements of the product $\prod_{i \in I} A_i$ are the functions, $a \colon I \to \bigcup_{i \in I} A_i$, such that $a(i) \in A_i$ for every $i \in I$.

We denote the members of $\prod_{i\in I} A_i$ by $(a_i)_{i\in I}$ and we usually call them *I-tuples*.

When $I = \{1, \ldots, n\} = [n]$, the members of $\prod_{i\in[n]} A_i$ are the functions whose graph consists of the sets of pairs

$$\{\langle 1, a_1\rangle, \langle 2, a_2\rangle, \ldots, \langle n, a_n\rangle\}, \quad a_i \in A_i,\ 1 \leq i \leq n,$$

and we see that the function

$$\{\langle 1, a_1\rangle, \langle 2, a_2\rangle, \ldots, \langle n, a_n\rangle\} \mapsto \langle a_1, \ldots, a_n\rangle$$

yields a bijection between $\prod_{i\in[n]} A_i$ and the Cartesian product $A_1 \times \cdots \times A_n$.

Thus, if each A_i is nonempty, the product $\prod_{i\in[n]} A_i$ is nonempty. But what if I is infinite?

If I is infinite, we smell choice functions. That is, an element of $\prod_{i\in I} A_i$ is obtained by choosing for every $i \in I$ some $a_i \in A_i$.

Indeed, the axiom of choice is needed to ensure that $\prod_{i\in I} A_i \neq \emptyset$ if $A_i \neq \emptyset$ for all $i \in I$! For the record, we state this version (among many!) of the axiom of choice:

Axiom of Choice (Product Version)

For any family of sets, $(A_i)_{i\in I}$, if $I \neq \emptyset$ and $A_i \neq \emptyset$ for all $i \in I$, then $\prod_{i\in I} A_i \neq \emptyset$.

Given the product of a family of sets, $\prod_{i\in I} A_i$, for each $i \in I$, we have the function $pr_i\colon \prod_{i\in I} A_i \to A_i$, called the *ith projection function*, defined by

$$pr_i((a_i)_{i\in I}) = a_i.$$