2 What do mathematicians study, and why?[3]

The short answer is: mathematicians study the *mathematical objects* that inhabit the *mathematical universe*, with the goal of discovering and confirming the universal principles of this universe: the patterns, symmetry and laws that are obeyed by different types of objects. In this section we'll take our first look at this abstract universe, and at the universal principles it obeys.

2.1 Introducing the mathematical universe

The mathematical universe is populated by countless *mathematical objects*, which are classified according to *object types*. Here are some examples of object types:

Natural numbers

Real numbers

Sets

Sets of real numbers (that is, sets whose members are real numbers)

Matrices with real number entries.

Functions mapping the set of real numbers to the set of real numbers

Matrices whose entries are functions

Operators (functions that map functions to functions)

Groups

Vector Spaces

Topological spaces

Rings

Modules

Sheaves

In mathematics, some simple-sounding names, such as groups, rings and sheaves, represent objects with very precise and intricate structure.

Objects belonging to a given object type can be manipulated and transformed in ways that are particular to that type. Natural numbers can be added together and multiplied together. Natural numbers can be compared according to which is greater. We can add two numbers,

[3] Version 2/2/2015. ©2015 Michael E. Saks

or add two matrices of the same shape, but we can't add two matrices of different shapes. We can take the derivative of a (differentiable) function, but we can't take the derivative of a set of real numbers. We can compute the determinant of a square matrix, but not of a non-square matrix, or of a function mapping the real numbers to the real numbers. We can take the union of two sets, but we can't take the union of two real numbers.

In the mathematical universe, complicated types of objects are built up from simple types of objects. In fact, mathematicians have shown that starting only from the concept of a set, one can build up all other types of objects (including, numbers, functions, and all of the other objects listed above). That all of mathematics can be boiled down to the theory of sets is a fascinating achievement of mathematical logic, but is not so important for what we do here. Our goal is to learn how to think like a professional mathematician, and a mathematician does not think of numbers as built up from sets. He thinks of numbers as . . . well . . . numbers. So we start with a larger (but still fairly small) collection of object types that serve as out building blocks for the mathematical universe. These object types are:

Numbers: Real Numbers and **Integers**

Collections: Sets, **Lists**, and **Indexed Collections**

Correspondences between sets: Functions

Most students in this course will have worked extensively with numbers, and some with sets, lists and functions and perhaps indexed families. Even if you are familiar with these concepts, it's important that we review these from the point of view of a professional mathematician.

Important common features of mathematical objects

When encountering a new type of mathematical object, there are three fundamental characteristics of the object to pay special attention to.

Equality. One of the most important concepts in the mathematical universe is *equality.* When we say that two mathematical objects are equal we mean that they are the *same object.* When we write 5+3=8 we mean that 5+3 represents a mathematical object (in this case, a number), that 8 represents a mathematical object, and that both objects are, in fact, the same.

Whenever we introduce a new mathematical type, one of the most important things to know about this type is: For two objects of this type, what are ways to establish that two objects are equal?

Operations and transformations Every type of object has certain basic ways that the object can be maninpulated. Some of these transform one object of that type into another object of the same type, for example you can take the negative of a number to produce another number, and you can take the reciprocal of a (non-zero) number to get produce another number You can take the complement of a set, which produces another set.

More commonly, there are ways of combining two objects of a type to get another object of the same type; this is called a *binary operation* on the type. (The word binary means "two" and we call the operation binary because it combines two objects into one.) Two numbers can be combined to get another number by adding them, or multiplying them. Two sets can be combined by taking their union or intersection. (Later we'll see that a binary operation such as addition can itself be viewed as a more mathematical object.)

Comparison of objects Objects of the same type can also be compared in various ways: we compare numbers according to which is less than the other, we compare sets according to which is a subset of the other.

Real numbers and Integers

Any student reading this book presumably knows that the set of *integers* consists of the number 0, the positive integers 1,2,3,... and the negative integers -1,-2,-3,.... The set of integers is a subset of the set of *real numbers*, which can be viewed as corresponding to the set of points on the number line.

Real numbers can be combined by adding them together, or multiplying them together, or subtracting one from another, or dividing one by another (provided the second number of not zero). Each of these operations gives a way of combining two real numbers to get one real number. The real numbers are ordered by $<$, and numbers that are > 0 are classified as *positive* and those that are < 0 are classified as *negative.* The operations of $+, \times, -, \div$ and the ordering $<$ satisfies a number of familiar properties: for example, addition is commutative and associative, multiplication distributes over addition, whenever $a < b$ and $c < d$ we also have $a + c < b + d$, etc.) We'll discuss these properties systematically later on.

The collection of integers satisfy many of the same properties as the collection of real numbers, but there are some crucial differences. The sum, difference, or product of two integers is an integer, but in general the quotient of two integers is not necessarily integer. Also, between any two real numbers a and b with $a < b$ there is a real number c with $a < c < b$ (so that c is *strictly between* a and b, but it is not true that between any two integers a and b with $a < b$ there is an integer c with $a < c < b$. (Why not?)

We'll look more closely at the basic properties of numbers in Section **??** (to be filled int).

Sets, Lists and Indexed Collections

In astronomy, there are many basic types of objects such as stars, planets, moons, asteroids, and comets. Then there are also more complex objects, such as *solar systems.* A solar system consists of a star, together with all of the objects (planets, moons, comets, etc.) that orbit around that star. A solar system is a *single astronomical object* that consists of a collection of other astronomical objects.

In the mathematical universe, we can also collect together objects to create a single new object There are two main types of collection objects: *sets* and *lists.* We'll also see that the notion of list can be generalized to the notion of an *indexed collection.*

Sets

A *set* is collection of objects where we ignore the ordering of the objects and we ignore duplicate copies of any object. For a set, the objects belonging to the set are called the *members* of the set, and the only thing that's important about the set is which objects are members, and which aren't. For a set A, we define the notation:

$x \in A$ means that x is a member of A

$x \notin A$ means that x is not a member of A.

Two sets A and B are considered to be equal if they have the same members. More precisely we have the following definitions:

A is a subset of B, written $A \subseteq B$ means that every element of A belongs to B.

A is a superset of B, written $A \supseteq B$ means that B is a subset of A.

$A = B$ means that $A \subseteq B$ and $B \subseteq A$.

Here are some examples of sets:

$S = \{2, 4, 6, 8, 10\}$

$T = \{10, 8, 6, 4, 2\}$

$U = \{2, 2, 4, 4, 6, 6, 8, 8, 10, 10\}$

$V = \{2, 6, 10\}$.

$W = \{2\}$

$X = \{\}$

In thie above notation, we list we use one of the standard notations for sets, where we list the members, separated by commas and surround the set by braces $\{\}$.

The above sets are all *finite*. We'll define more precisely what we mean by finite later, but for now we'll use the intuitive idea that a finite set is one where you can count all the elements to obtain a definite number. The set of all integers and the set of all real numbers are examples of infinite sets.

Notice that sets S, T and U are actually all the same set, so $S = T = U$. That is because our definition of equality of sets ignores the order in which members are listed, and ignores repeated appearances of a member.

Set W has size 1, and set X has size 0 and is called the *empty set*. The empty set is also denoted $\emptyset$.

Notice that V is a subset of each of the sets listed above it, W is a subset of each of the subsets listed above it, and X is a subset of each of the subsets listed above it. This last fact follows from the general principle that the empty set is a subset of every set.

If a set A is finite, we define the *size* or *cardinality* of A, $|A|$ to be the number of members of the set.

If a and b are integers, the set of integers greater than or equal to a and less than or equal to b is abbreviated by $\{a, \ldots, b\}$. Notice that if $a = b$ this is just the singleton set $\{a\}$ and if $a > b$ then this is the empty set.

Sets (unlike lists) may be infinite. For example the set of all real numbers is an infinite set, as is the set of all integers. These sets are denoted by special symbols:

$\mathbb{R}$ is the set of real numbers, and $\mathbb{R}_{>0}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{<0}$ and $\mathbb{R}_{\leq 0}$ denote, respectively, the sets of positive real numbers, nonnegative real numbers, negative real numbers and nonpositive real numbers.

$\mathbb{Z}$ is the set of integers, and $\mathbb{Z}_{>0}$, $\mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{<0}$ and $\mathbb{Z}_{\leq 0}$ denote, respectively, the sets of positive integers, nonnegative integers, negative integers and nonpositive integers.

These sets are infinite, but easy to describe. Most infinite sets are hard to describe. Obviously, with an infinite set we can't describe it by listing all of its members. Later we'll see some additional ways of describing sets.

For any set S, the *power set of* S, denoted $\mathcal{P}(S)$, is the set of all subsets of S. For example $\mathcal{P}(\{1,2,3\})$ is the set $\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. We also write $\mathcal{P}_{\mathbf{fin}}(S)$ for the set of all finite subsets of S. Thus $\mathcal{P}_{\mathbf{fin}}(\mathbb{Z}\}$ is a set which has a member the set of all integers from 1 and 100, but does not have as a member the set of even integers.

The main operations for combining sets are *union* (written $\cup$), *intersection* (written $\cap$) and *difference* (written $\backslash$). For sets A and B we define:

$A \cup B$ is the set obtained by merging A and B. An object x is a member of $A \cup B$ provided that $x \in A$ or $x \in B$.

$A \cup B$ is the set of elements that belong to both A and B. Thus an object x is a member of $A \cap B$ if $x \in A$ and $x \in B$.

$A \backslash B$ is the set of elements that belong to A but not B. Thus an object x

When working with sets, it is often the case that we are looking at the subsets of one specific set, which we refer to as the *universe*. For example, we may be considering only subsets of real numbers. The universe is sometimes clear from context, and other times we need to specify it. If A is a set then the *complement of* A denoted A^c is the set $U \backslash A$ where U is the chosen universe.

Lists

A *list* is a different kind of collection of objects, in which the objects are presented in a specific order and we pay attention to repeated elements.. Each separate appearance of an object in a list is called an *entry* in the list. Here are some examples of lists:

$a = (2, 4, 6, 8, 10)$

$b = (10, 8, 6, 4, 2)$

$c = (3, 3, 3, 3)$

$d = (1, 3, 1, 3, 1, 3)$

$e = (17).$

$f = ().$

These lists are represented in the standard notation for lists, in which the entries are separated by commas and the list is surrounded by parentheses (). Lists a and b each have 5 entries. Notice that entries of these two lists are the same, but in different order, and for this reason $a \neq b$.

List c has 4 entries, all of which are the same. List e has a single entry, and list f has no entries, and is called the *empty list*.

All of the above lists have entries that are numbers, but as we'll see, the entries of a list can be any mathematical object. You've probably already encountered lists of numbers in prevous courses, where they were called *vectors*.

Unlike sets, lists are restricted to have a finite number of entries. For a list v, the number of entries is called its *length* and is denoted $\mathbf{length}(v)$. Each entry of v appears in a specific *position* that is a number from 1 to $\mathbf{length}(v)$. Thus in the list a, 8 appears in positon 4, and in list b, 8 appears in position 2. We often write L_j for the entry of L appearing in position j. Thus for list d we have $d_1 = d_3 = d_5 = 1$ and $d_2 = d_4 = d_6 = 3$.

Two lists v and w are considered to be equal if they have the same length, and for each j between 1 and $\mathbf{length}(v)$ we have $v_j = w_j$ (so that the entries in matching positions are the same).

There is a natural operation for combining two lists into one list called *concatenation*; the concatenation of list a w“nith list b, denoted $a \circ b$, is obtained by appending the items of b to the items of a, for example $(1, 6, 1) \circ (2, 3)$ is the list $(1, 6, 1, 2, 3)$.

We also introduce a non-standard notation. If A and B are sets of lists then $A \circledcirc B$ is the set of all lists of the form $a \circ b$ where $a \in A$ and $b \in B$.

If A is any set, we write A^* for the set of all lists with entries coming form A. If k is an integer, we write A^k for the set of all possible lists of length k whose entries are from the set A For example, as you probably learned in previous courses, $\mathbb{R}^3$ is the set of lists (vectors) of length 3 with real number entries, and $\mathbb{R}^*$ is the set of lists of real numbers of all possible (finite) lengths.

Remark 2.1. Notice that the meaning of A^k *depends on the type of object A represents* If A is a number, then A^k means the number obtained by multiplying k A's together. If A is a set then A^k is a set of lists of length k This is an example of a common feature of mathematical terminology: the meaning of notation depends on the types of objects involved This is one of many reasons why in a mathematical discussion, it is crucial to be aware of the types of each object under consideration. There is an interesting connection between the notation A^k when A is a set and A^k when A is a number If A is a finite set, then A^k is also a finite set (whose members are lists of length k) It turns out that the size of the set A^k, denoted $|A^k|$, is equal to $|A|^k$, which is the size of A raised to the k power.

Lists of size 2 are called *ordered pairs.* If A and B are sets, then $A \times B$, the *product of A and B* denotes the set of al ordered pairs (a, b) such that $a \in A$ and $b \in B$. Thus, $A \times A$ is the same as A^2.

Remark 2.2. The correct interpretation of the notation A^k and the notaiton $A \times B$ for sets is rather subtle. As we now discuss, mathematicians routinely misuse this notation. This misuse of of the kind that rarely if ever causes confusion or difficulty, but it's something for novices to avoid.

1. The sets A and A^1 are different; A is a set of objects, and A^1 is a set of lists of length 1. For example, if $A = \{1, 2, 3\}$ then A^1 is the set $\{(1), (2), (3)\}$. The members of A correspond to the members of A^1 in an obvious way, but they are not the same sets. Still, mathematicians often treat A and A^1 as the same set because for most purposes they are, but strictly speaking this is incorrect, and it could conceivably lead to confusion or mistakes.

2. Unlike multiplication of numbers, multiplication of sets is neither commutative or associative. If A,B,C are sets then $(A \times B) \times C$ is not the same set as $A \times (B \times C)$. Each member of $(A \times B) \times C$ are ordered pairs, whose first entry is an ordered pair from $A \times B$ and whose second entry is a member of C. Each member of $A \times (B \times C)$ is also ordered pairs, whose first entry is a member of A and whose second entry is an ordered pair from $B \times C$. For example if $1 \in A$, $2 \in B$ and $3 \in C$ then $((1, 2), 3) \in (A \times B) \times C$ while $(1, (2, 3)) \in A \times (B \times C)$. It is easy to see that every member of $(A \times B) \times C$ corresponds to a member of $A \times (B \times C)$, and for most purposes one can think of corresponding members as the same. Mathematicians do this routinely, but we won't do that here to emphasize the importance of paying attention to the types of objects.

3. Similarly, $A \times A^2$, $A^2 \times A$ and A^3 are all different sets. The members of $A \times A^2$ are ordered pairs whose second entry is an ordered pair, the members of $A^2 \times A$ are ordered pairs whose first entry is an ordered pair, and A^3 is a set of lists of length 3. Note, however that $A^1 \circledcirc A^2 = A^2 \circledcirc A^1 = A^3$ since the members of all three sets are the lists of length three with entries in A. As mentioned, the operation $\circledcirc$ is not a standard mathematical operation, and most mathematicians view $A \times A^2$ to means the same as $A \circledcirc A^2$, even though they are not really the same.

Nested types

The above examples of sets and lists were all collections of numbers. We can have more sets or lists whose members or entries are themselves sets or lists. For example:

- $a = \{\{1,2\},\{2,3\},\{3,5\}\}$
- $b = \{\{3,2\},\{1,2\},\{5,3\}\}$
- $c = (\{1\},\{1,2\},\{1,2,3\},\{1,2,3,4\})$
- $d = (\{1\},\{2,1\},\{3,2,1\},\{4,3,2,1\})$
- $e = (\{1,2,3,4\},\{1,2,3\},\{1,2\},\{1\})$
- $f = \{\emptyset\}$
- $g = \{\}$
- $h = (\emptyset)$
- $j = \{(\{1,2,3\},\{1,3\}),(\{2,4\},\{\}),(\{1,3,5,6\},\{2,4,6,8\})$

The object a is a set with three members. Each member of a is itself a set having two members. The object b is also a set with three members, with each member being a set having two members. In fact we have $a = b$! This is because each member of a is a member b and vice versa.

Objects c,d and e are lists, each having four entries, where each entry is a set. We have $c = d$ because they have the same length, and $c_j = d_j$ for j between 1 and 4. However, $c \neq e$ since, for example, $c_1 \neq e_1$.

Object f and g are sets. g is the empty set and has size 0. Note that f is not equal to g and is not the empty set. Instead f is a set having exactly one member, and that member is the empty set. Object h is the empty list and is not the same object as f because h is a list and f is a set. Object j is a set, whose members are lists of size 2 (order pairs) where the entries of each list are sets of integers.

Indexed collections

We now consider a generalization of lists called *indexed collections*. If d is a list of length 5 whose entries belong to set S then the entries d_1, d_2, d_3,d_4, d_5 are each members of S, The subscripts 1,2,3,4, and 5 serve as reference labels. These reference labels are called *indices*, and one can think of each of index as pointing to the appropriate item on the list. The set of indices, $\{1,2,3,4,5\}$ is called the *index set* of the list.

In a list the index set is always of the form $\{1,\ldots,m\}$ where m is the length of the list. When we think of a list in this way, it is natural to consider the possibility of using a different index set for the collection. If we call our index set J, then for each $j \in J$ we have an object

d_j. If all of the objects in the collection are of some type T, we call this collection an *indexed collection* of objects from T with index set J. We denote such a collection using the notation $d = (d_j : j \in J)$, which means that d has one entry for each index j belonging to J.

Example 2.1. Suppose we take an indexed collection of numbers m with the index set the set $J = \{1, 2, 3\} \times \{1, 2, 3\}$ which (as discussed earlier) is the set of ordered pairs (i, j) with both entries in $\{1, 2, 3\}$. For example, our indexed collection might be given by $m_{(1,1)} = 7, m_{(1,2)} = 4, m_{(1,3)} = 1, m_{(2,1)} = 8, m_{(2,2)} = 5, m_{(2,3)} = 2, m_{(3,1)} = 9, m_{(3,2)} = 6, m_{3,3)} = 3$. Because the index set is ordered pairs, it is natural to visualize this collection by arranging the entries in a 3 by 3 table:

i/j	1	2	2
1	7	4	1
2	8	5	2
3	9	6	3

Thus an indexed collection with index set $\{1, 2, 3\} \times \{1, 2, 3\}$ is naturally thought of as a *3 by 3 matrix*. In general an indexed collection with index set $\{1, \ldots, m\} \times \{1, \ldots, n\}$ is called an m by n matrix.

We can take any other index set. Another index set that arises commonly is the index set $\mathbb{Z}_{>0}$ of positive integers. An indexed collection with this index set has the form $a = (a_j : j \in \mathbb{Z}_{>0})$ and is called a *sequence*, which can be thought of as a list that doesn't end.

Functions

The last group of basic objects are mathematical objects that represent a correspondence between two sets of objects. These are called *functions* and are at the heart of modern mathematics. Most students start this course with some familiarity with functions, and view functions as formulas such as $f(x) = x^2 - 2x + 1$ or as curves drawn in the x-y plane. These examples and pictures are useful, but may give a misleading picture of what kind of object a function is.

The idea of a function is of a small "computer" that can receive certain mathematical objects as input, and for each input object it might receive, produces an output that depends on the input. Each function f has a specified set of input objects that it will accept, called the *domain* of f and written as $\mathbf{Dom}(f)$. When the function f is given an object x belonging to the domain, the function responds with an output. Each time you give f the same input x you get the same output. The outputs that f gives for two different inputs x and z may be the same or different.

Figure 1 illustrates the idea of a function.

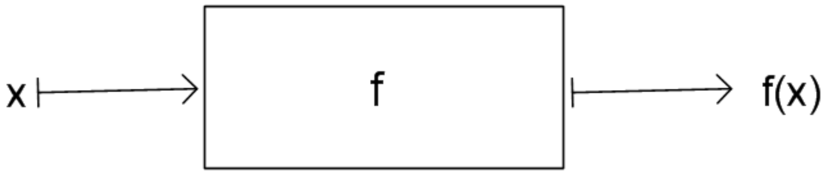

Figure 1: A pictorial view of a function

Let's express this idea a bit more carefully. A function is a mathematical object g with the following characteristics:

1. Associated to g is a set, called the *domain* of g and denoted $\mathbf{Dom}(g)$. This set is the set of "permissible inputs" to g.

2. To each permissible input $x \in \mathbf{Dom}(g)$, g associates a mathematical object denoted $g(x)$ (which is read "g of x"). $g(x)$ is called the *image of g on x* and represents the output of the function g when x is input. The set consisting of $g(x)$ for all domain members x is called the *range* of g and is denoted $\mathbf{Rng}(g)$.

A *target set* for the function g is a set that contains $\mathbf{Rng}(g)$. The notation $g : S \longrightarrow T$ means that g is a function with domain S and target set T. Since the target set can be any set that contains $\mathbf{Rng}(g)$, we generally choose T to be an easy to describe set that contains $\mathbf{Rng}(g)$.

For most of the functions you've encountered in the past, the domain and target sets are both $\mathbb{R}$, or subsets of $\mathbb{R}$. In fact, a function is allowed to have any set as its domain, and any (nonempty) set as its target. Below we'll see various examples of functions with different domains.

Specifying a function. To fully describe a function g, you need to specify the set $\mathbf{Dom}(g)$, and for each $x \in \mathbf{Dom}(g)$ you must specify what $g(x)$ is. If the domain of g is a finite set we can simply describe the domain by listing its elements, and listing a function table.

Example 2.2. Let h be the function with $\mathbf{Dom}(h) = \{1, 2, 3\}$ given by the table:

x	1	2	3	4	5
$h(x)$	3	2	2	1	3

There are various ways to visualize a function. In figure 2, we present a *domain-target diagram*. Here the elements of the domain are pictured on the left, and the output elements are pictured on the right. There is an arrow from each domain member to its corresponding output member.

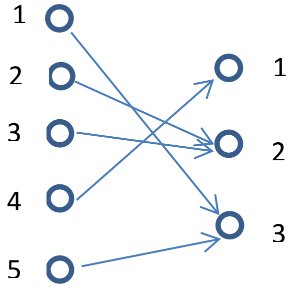

Figure 2: Domain-Target diagram of function h from Example 2.2

When the domain and target are subsets of the real numbers, we can illustrate the function using the familiar *function graph in the x-y plane* as in Figure 3. For each point plotted in the graph, the x-coordinate is a domain value, and the y coordinate is the function value.

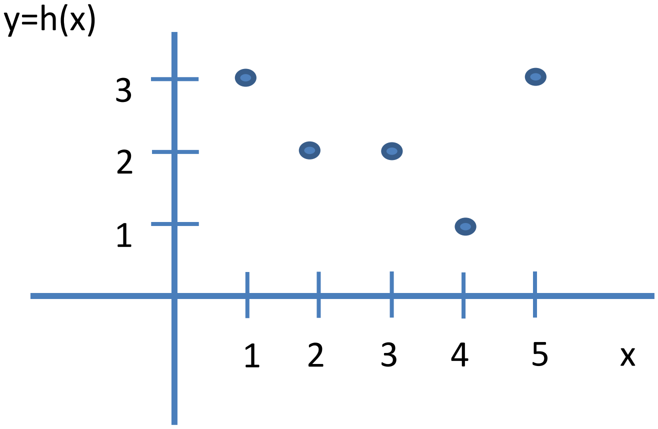

Figure 3: A graph of function h from Example 2.2

Image and inverse image Suppose $f : X \longrightarrow Y$. For a subset X' of X, the *image of* X' *under* f, denoted $f[X']$ is the set of all output values $f(x)$ for $x \in X'$. For a subset Y' of Y, the *inverse image of* Y' *under* f, denoted $f^{-1}[Y']$ is the set of domain values x such that $f(x) \in Y'$. For example, for the function h in Example 2.2 $f[\{1, 2, 3\}] = \{2, 3\}$ and $f^{-1}[\{2, 3\}] = \{1, 2, 3, 5\}$. Both the Domain-Target diagram and the graph of the function are useful for computing the image and inverse inputs of sets. We note that the image of the entire domain X under f is $\mathbf{Rng}(f)$ and that for any set Y' containing $\mathbf{Rng}(f)$, $f^{-1}[Y'] = X$.

If the domain is infinite then we can't describe the function by a table, so we need other ways to specify a function. The most common way of specifying a function is to give the domain, together with a *rule* that precisely describes how the output object is built from the input object.

Example 2.3. Let f be the function with domain $\mathbb{R}$ given by the rule $f(x) = x^2$ for all $x \in \mathbb{R}$.

Here's a simple but important function.

Example 2.4. Let D be an arbitrary set. The *identity function* on D is the function $\mathbf{id}_D$ with domain D given by the rule $\mathbf{id}_D(x) = x$ for all $x \in D$.

The rules for describing a function may be complicated.

Example 2.5. Let h be the function defined on $\mathbb{Z}$ by the rule:

$$h(x) = \begin{cases} x/2 & \text{if } x \text{ is even} \\ 3x + 1 & \text{if } x \text{ is odd} \end{cases}$$

A rule of this form is expressed by dividing the domain into pieces. Here we divide the set of integers into the set of even integers and the set of odd integers.

Next we consider some examples of functions whose domain and/or target are not sets of numbers. Here's an example of a function whose domain is $\mathbb{Z}$, that maps each domain element to a set of integers (rather than a single integer).

Example 2.6. Here's a function that maps a number to a set of numbers. Let $\mathbf{Div}$ be the function with domain $\mathbb{Z}_{>0}$ given by the rule $\mathbf{Div}(n)$ is the set of positive integer divisors of n. Thus $\mathbf{Div}(12) = \{1, 2, 3, 4, 6, 12\}$ and $\mathbf{Div}(0) = \mathbb{Z}_{>0}$.

We can take our target set to be $\mathcal{P}_{\mathbf{fin}}(\mathbb{Z})$ and write $\mathbf{Div} : \mathbb{Z} \longrightarrow \mathcal{P}_{\mathbf{fin}}(\mathbb{Z})$, since every output value is a finite set of integers.

Example 2.7. Consider the function m whose domain is the set of finite subsets of real numbers given by the rule that for any set S, $m(S)$ is the least member of S. For example, $m(\{17, -\pi, 1, 100.25\}) = -\pi$. This function has domain $\mathcal{P}_{\mathbf{fin}}(\mathbb{R})$ and target set $\mathbb{R}$.

When are two functions equal? The criterion for two functions f and g to be equal is that they have the same domain, and that for each member x of the domain $f(x) = g(x)$.

Well-defined rules for functions. A proper description of f by a rule should clearly state the domain of f, and provide clear instructions that given x allows one to determine $f(x)$. A rule that does this is said to be *well-defined.* A function rule may fail to be well-defined for two main reasons: (1) The function is undefined for some element of the domain, which means that rule supplied does not make sense for that element, or (2) The rule is ambiguous for some element of the domain, so that for that domain element the rule either produces more than one output value.

Example 2.8. (An invalid function definition.) Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be defined by $f(x) = x/(x+1)$.

This function definition is invalid because the rule does not make sense for all values of the domain, namely if $x = -1$ then the rule does not produce a real number.

Example 2.9. (An ambiguous function definition.) Let $r : \mathbb{R}_{\geq 0} \longrightarrow \mathtt{R}$ be defined by the rule $r(x)$ is the real number such that $r(x)^2 = x$.

This rule is ambiguous because for each input x, there are two choices of $r(x)$ such that $r(x)^2 = x$.

In some cases we may wish to repair a faulty function rule so that the result is a function. There are a few ways this can be done, and the choice depends on the particular situation. The easiest way to repair a function is to remove all of the "troublesome" inputs from the domain. So for Example 2.8, we could redefine the domain to be $\mathbb{R} - \{-1\}$. For Example 2.9, the rule is ambiguous for every member of the domain, so removing all of these inputs from the domain will leave a very uninteresting function. In the case of an ambiguous rule, we can try to restrict the target so as to eliminate the ambiguity. In Example 2.9, if we reduce the target set to

$\mathbb{R}_{\geq 0}$ then the rule becomes unambiguous, since now for each input x there is one and only one choice of $r(x) \in \mathbb{R}_{\geq 0}$ so that $r(x)^2 = x$. In the case of a rule that is undefined for some values, we can simply specify a value for the function for those values. For example, the function in Example 2.8 could be modified to:

$$f(x) = \begin{cases} x/(x+1) & \text{if } x \neq -1 \\ 0 & \text{if } x = -1 \end{cases}$$

In this case, our choice to set $f(-1)$ to 0 is arbitrary, which may or may not be okay depending on the context. Later we'll see that in some cases we'll see that there is a natural choice for filling in undefined values.

Example 2.10. (Another invalid function definition.) Suppose we define the function $h : \mathbb{Z} \longrightarrow \mathbb{Z}$ be the function given by the rule $h(n) = n/2$. If we take an odd input, then the output is not an integer and therefore it is outside the specified target set. If our given situation allows for h to output numbers that are not integers, one easy repair is to change the target set to $\mathbb{R}$. If we need the function to output integers, then either we have to restrict the domain to the set of even integers, or we have to modify the function rule for odd integers.

Composition of functions The primary way to combine two functions into one is by *composition*. The idea of function composition is represented in Figure 4, namely the composition of two functions is the function built by chaining the functions together so that the output of one becomes the input of the other.

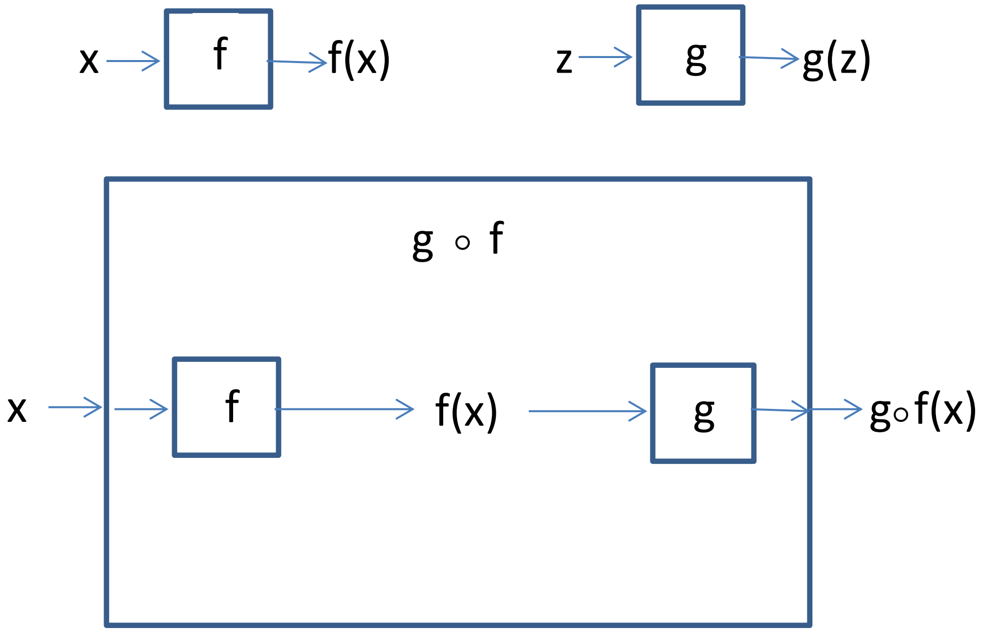

Figure 4: The function $f \circ g$ obtained by composing functions f and g.

The normal situation for composing functions is when we have two functions: $f : S \longrightarrow T$ and $g : U \longrightarrow V$, where the target T of f is a subset of the domain U of g. We define $g \circ f : S \longrightarrow V$ to be the function given by the rule $g \circ f(s) = g(f(s))$ for all $s \in S$. For this rule to make sense we need that $f(s)$ belong to $\mathbf{Dom}(g)$. this rule makes sense for all $s \in S$,

Functions versus Indexed collections The observant reader may notice a close similarity between functions and indexed collections. A function f associates to each member of the domain an element in the target. An indexed collection C of objects in T with index set J consists of a collection of objects $(C_j : j \in J)$ where each $C_j \in T$.

In fact functions and indexed collections are really just two different ways of representing the same thing. Given a function $f : X \longrightarrow Y$ we can build an indexed collection $(C_x : x \in X)$ of objects in Y with index set x where $C_x = f(x)$, so we can think of the function as an indexed collection. Similarly, if we start with an indexed collection $(C_j : j \in J)$ of objects in T, we can build from it a function $g : J \longrightarrow T$ defined by $g(j) = C_j$. In particular, a list b of length m can be associated to a function whose domain is $\{1, \ldots, m\}$.

Given that functions and indexed collections are essentially the same thing, why do we bother defining both of them? We do this because both concepts appear frequently in the literature (with functions being much more common.) There is a subtle difference in how they are used. When we study indexed collections, we are often (though not always) mostly interested in the objects in the collection and not in which indices correspond to which objects. The indices are simply a convenient label to reference all of the objects. With functions, we generally care very much about the correspondence between domain elements and target elements.

2.2 Introducing universal principles in mathematics

Having introduced some basic objects of study in mathematics, let's turn to the question of what the goal of that study is. On a very basic level, we want to discover the "facts" of mathematics. But there are many facts, we're particularly interested in the *important* ones. Let's start with an example of two mathematical facts:

Fact A**.** $(7-3)(7+3) = 7^2 - 3^2$

Fact B**.** For any two numbers a and b, $a^2 - b^2 = (a+b)(a-b)$.

The first is a fact from arithmetic about the two numbers 3 and 7.

The second fact is related to the first, but is a much more interesting and powerful fact than the first. It asserts a *universal principle* that *works for any two numbers.*

This universal principle implies the first fact, and an endless number of others, such as:

Fact C**.** $(7101036 - 3354251)(7101036 + 3354251) = 7101036^2 - 3354251^2$

Propositions A and C are said to be *instances* of the universal proposition B. They are obtained by substituting specific values for a and b. Since, as we know from high school algebra, Proposition B is a true proposition, it tells us that Fact C must be true also. Notice that Fact C is not something that would be easy to verify directly (even with a calculator).

Universal principles are the most important type of facts. Universal principles allow us to summarize *vast amounts of knowledge* in a *single sentence.*

Every intellectual endeavor, such as astronomy, microbiology, economics, or history, is the search for *universal principles* which reflect a truth that applies in many situations: "galaxies rotate","living organisms use oxygen to convert biological fuel into usable energy","As the supply of a good goes up, the price goes down", or "flu shots protect people against the flu". Universal principles in the sciences and social sciences (such as the ones stated above) are often only approximations to the truth. The principle "living organisms use oxygen to convert biological fuel into usable energy" does not apply to all living organisms, since some do this conversion without oxygen. The principle "the supply of a good goes up, the price goes down" is not always true since there are many other factors that determine the price of a good, and this principle is understood to mean that increases in supply of a good *tend to be* accompanied with a reduction in price. These principles capture some general knowledge that apply broadly but may be incorrect in some situations..

The universal principles of mathematics concern relationships and patterns that are present among mathematical objects. Unlike universal principles in other fields, universal principles of mathematics are required to be *completely clear and unambiguous*, clearly stating the situations to which it applies, clearly excluding situations to which it does not apply, and clearly stating the conclusion that must hold in the situations to which they do apply.

How do we know that a proposed universal principle is indeed true? In the physical, biological and social sciences, universal principles are conjectured based on a combination of observation (experiment) and reasoning based on broader knowledge of the field. Once formulated, a general principle in the sciences is proved or disproved by experiment. A general principle in these fields is always open to reconsideration if new evidence comes up that contradicts it.

In mathematics, possible universal principles are also formulated based on a combination of reasoning and observation (of examples). However, the confirmation of universal principles in mathematics is not done by experiment. Instead it is done by *deductive proof.* Deductive proof is a central part of mathematics, and will become a central focus of this course. For now, we will look at some examples of the statements (and not the proofs) of some universal principles of mathematics.

Earlier we mentioned the following famous examples of a universal principle:

Universal Principle 2.1. (Pythagorean Theorem) For any right triangle, the square of the length of its hypotenuse is equal to the sums of the squares of its two legs.

For example, if T is a right triangle whose legs have length 3 and 7, then the hypotenuse h must satisfy $h^2 = 7^2 + 3^2 = 58$ and so the length of the hypotenuse must be $\sqrt{58}$.

Remark 2.3. In stating this general principle, I've assumed that the reader is already familiar with the following terminology:

- *right triangle*,
- *hypotenuse* of a right triangle, and
- *leg* of a right triangle.

Obviously, a reader who doesn't know what a right triangle is, or doesn't know what a hypotenuse is, won't be able to understand this principle.

When reading any general principle, or any mathematical sentence, the first thing to ask yourself is "Do I know the precise meaning of each piece of mathematical terminology used in this sentence?" If the answer is no, STOP! Find out what every term means. It's not enough to have a vague idea what the terminology means. If you don't know the precise meaning of every term means, it is unlikely that you'll be able to work correctly with the principle.

Universal Principle 2.2. For any real number a and any real number b, the average of the squares of a and b is at least the square of the average of a and b, that is: $\frac{a^2+b^2}{2} \geq \left(\frac{a+b}{2}\right)^2$.

For example, if $a = 12$ and $b = 17$ then the average of the squares is $(144 + 289)/2 = 216.5$ while the square of their average is $((12 + 17)/2)^2 = 210.25$. This is a principle in the area of *real number inequalities*: here the expressions $\frac{a^2+b^2}{2}$ and $\left(\frac{a+b}{2}\right)^2$ are determined by the choice of a and b and the principle says that for any choice of a and b, the first expression is always less than or equal to the second.

When given any universal principle in mathematics, one question a mathematician asks herself is: Can this principle be extended to cover more general situations? For example, the previous principle applies to any pair of numbers. It turns out that we can generalize this principle so that it applies to any finite list of numbers.

Universal Principle 2.3. For any (finite) list of real numbers, the average of the squares of the numbers is at least the square of the average of the numbers.

For example, for the list of numbers $4, 6, 3, 4$, the average is 4.25, and the square of the average is $18\frac{1}{16}$, while the average of the squares is $\frac{1}{4}(16 + 36 + 9 + 16) = 19\frac{1}{4}$.

The next general principle expresses a basic fact of algebra that you probably learned in high school. Recall that a *monomial* in the variable x is an expression of the form cx^n where n is a nonnegative integer called the *degree* of the monomial, and c is a real number called the *coefficient* of the monomial. A monomial is said to be *nonzero* if its coefficient is not zero. A polynomial in variable x is a sum of a finite number of monomials, such as $5x^4 + \pi x^5 + (-5/4)x^2$. A *root* of the polynomial p is a number which when substituted for x makes the polynomial evaluate to 0. For example the roots of $x^3 + (-9)x^2 + 15x + (-7)$ are 1 and 7, since if you substitute 1 for x the polynomial evaluates to 0, and if you substitute 7 for x the polynomial evaluates to 0, but if you substitute any other number, the polynomial evaluates to a nonzero number. A polynomial is said to be nonzero if there is at least one number such that substituting that number for x results in a nonzero value for the polynomial.

Universal Principle 2.4. For any polynomial p, if p is nonzero then the number of roots of p is at most the largest degree of any of the monomials in the sum.

In the example given just before the statement, the largest degree of any of the monomials in the sum is 3, and the polynomial has exactly 2 roots.

Here's a general principle in the field of *number theory*, which is the study of the integers and their properties:

Universal Principle 2.5. For any positive integer a and positive integer b, if b is a prime number then $a^b - a$ is a multiple of b.

For example, if $a = 3$ and $b = 7$ then $a^b - a = 3^7 - 3$ is equal to 2184, which is indeed a multiple of 7, since it is equal to 7 times 314.

This is a somewhat surprising principle, and a mathematician reading this principle for the first time, would ask herself: Is this really always true? If so why?

Here's another strange and surprising principle from number theory:

Universal Principle 2.6. For any positive integer n such that n is prime, and $n-1$ is divisible by 4, there are positive integes a and b such that $n = a^2 + b^2$.

For example, the number 29 is a prime number such that $29 - 1$ is divisible by 4 and the conclusion is true by choosing $a = 5$ and $b = 2$ since $29 = 5^2 + 2^2$. The number 73 is a prime number such that $73 - 1$ is divisible by 4, and the conclusion is true by choosing $a = 8$ and $b = 3$, since $73 = 8^2 + 3^2$.

The next principle applies to elementary set theory:

Universal Principle 2.7. For any two sets A,B and C, if $A \neq B$ then $A \cup C \neq B \cup C$ or $A \cap C \neq B \cap C$.

Notice that in the conclusion there are two possible conditions, and the principle guarantees that at least one of them is true, but doesn't say which. For example, if A is the set $\{1, 2, 3, 4\}$ and B is the set $\{2, 3, 4, 5\}$ and C is the set $\{1, 3, 5\}$ then $A \cup C = B \cup C = \{1, 2, 3, 4, 5\}$ and $A \cap C = \{1, 3\}$ while $B \cap C = \{3, 5\}$ and so $A \cap C \neq B \cap C$, as asserted by the principle.

We've now seen various examples of some universal principles. One of the main things we'll be studying in this course is: given a possible universal principle, how do mathematicians know that the principle is true? This question will lead us to the idea of *mathematical proof*. Before we can start discussing mathematical proof, we'll need to address a more basic question: how do mathematicians communicate about mathematics?