UMASS AMHERST MATH 300: NOTES FOR FALL 05

FARSHID HAJIR

Contents

Dear Reader,

Hi, I am the slightly-edited and rearranged notes Farshid typed up for his students in Math 300 (*Introduction to Fundamental Concepts of Mathematics*) starting with the Spring term of 2005. Since you've read this far, it appears you're taking Math 300 this semester, so you might be spending a fair bit of time with me. Allow me, therefore, to tell you a bit about myself.

I am rough, as in unpolished: there are typos sprinkled throughout, and even errors, some of them are actually intentional (see the note at the end of Part 1), not to mention unconventional words and even made-up words like "Shazzam" and "anyhoo." I am also uneven: at times you'll find I'm quite conversational and quirky (there are little one-act "plays" in most chapters), and at other times quite formal and stodgy, like a regular mathematics book for $\geq$ 300-level classes. I am, after all, supposed to help prepare you for the rather formal structures you will be studying as you continue your mathematically oriented learning experiences. I'm also a bit wordy and long-winded, explaining the same thing over and over again in only slightly different formulations, saying it this way, then that way, putting it in different words, finding alternate expressions for it, in short repeating myself, repeating myself. I guarantee Farshid will admonish you to be "cogent, correct, and concise!" But just between us, finding it rather difficult to be the latter himself, he likes to pretend that "repetition is a key hallmark of any educational process."

Having just listed some of my obvious faults, allow me a certain measure of complacecy in hoping that you will find these faults to be counterbalanced by my being a little less boring than other texts that cover the same material, such as the textbook used for Math 300 this term (Gilbert and Vanstone's). Not to say that the latter text is boring, it's quite good apparently, but I'm willing to bet it doesn't have any one-act plays starring actual students from the class! Nor will you find Mr. Noodle mentioned anywhere within its pages, and key ideas in the proof will not be highlighted by *SHAZZAM!*. As it is, I think Farshid will be making you pay through the nose to buy the book because he thinks it's valuable to have a real "published," "well-organized," "example-laden" textbook (alongside yours truly) that explains the same material in a different way: "The contrasts, as well as the similarities," he says, will "enhance the students' ability to learn." You be the judge; I think you can guess where I stand on this issue.

As the semester progresses, I suspect Farshid will attempt to polish me up and make cross-references, add examples, diagrams, all that jazzy stuff, which is alright I suppose, but I sure hope he doesn't try to change my informal character; I don't want to be prim and proper and pompous, like most mathematics texts. That would be a pity. Although I have to admit, grudgingly, that I wouldn't mind having a real cover one day. Anyhoo, don't bother printing out all my gazillion pages yet, as I'm bound to undergo quite a few changes.

Maybe every week you can print out the "part" that you happen to be studying then. Also watch Farshid's web site for update notes.

Well, I hope you enjoy Math 300. If you have suggestions for improving me, send them to Farshid: unfortunately, the tyrant *auteur* has complete final cut on my contents.

Your humble servant and guide,

Math 300 Notes, September 8th 2005 Edition.

Part 1. Problem Solving, Inductive vs. Deductive Reasoning, An introduction to Proofs

1. Introductory Notes

1.1. **Propositions.** Much of mathematics is about solving problems. The process of doing mathematics is, however, far more complex than "simply" solving problems. For one thing, you have to create the problems in the first place! One of the things I would like to show you in this course is how mathematicians create mathematical objects and tools in order to solve specific problems. In the process of doing so, the objects and tools created bring forth new questions and new problems, from which new objects and tools blossom.

But, for now, to whet your appetite in this first week of class, let us start with the more familiar territory of problems: Do keep in mind that there will be **plenty** of "theory-building" coming later. Think of the problems below as puzzles, for that is what they are. Have fun with them. Examine their different components and how they fit together, make observations, jot down notes, locate and investigate patterns, try easier versions of the problem, relate the given problem to others you may have worked on, act the problem out, "see" the problem as a movie in your head, draw, doodle, sketch pictures, make graphs and tables and charts.

You will probably find it helpful to keep in mind the 4 steps George Polya advocates in approaching a problem:

i) Read and Understand the Problem
ii) Dream up a plan to solve the problem
iii) Carry out the plan
iv) Look back, returning to i) if necessary.

In some of the problems below (and throughout the semester, heck, throughout your life), you will be asked to provide a "proof" of some "Proposition." Let us define how we will use these terms.

A *Proposition* is a common name for a mathematical statement in which all concepts that appear have been given a valid definition. A Proposition may be true or false. A *Proof* of a Proposition is an argument which establishes the truth of the Proposition. Here is an alternative description: a proof of P is a convincing argument which removes any doubt concerning the truth of P. If a Proposition is true, we say that it holds. If a Proposition is false, we say that it does not hold.

Here are some examples of Propositions.

$P1$: The sum of two even integers is even.

[What is meant here is shorthand for the following more precise version: Let us define our terms first: the integers are the elements of the familiar set

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

on which we are given the usual operations of addition and multiplication. An integer a is even if there exists an integer m such that $a = 2m$. Now here, finally, is the meaning of our statement: If a is an arbitrary even integer, and b is an arbitrary even integer, then $a + b$ is an even integer. Is $P1$ true? Can you construct a proof of $P1$ for yourself?]

$P2$: Every triangle is isosceles. [Although this is a very good example of a *patently false* proposition, assuming the notion of distance between two points is the usual one, there is a very important class of distance-measurements (called p-adic distance, where p is a prime) for which the statement is true! If you campaign long and hard, say one second, then I will be glad to tell you more about p-adic metrics.]

$P3$: If a and b are integers and $b \neq 0$, then there exists exactly one ordered pair of integers (q, r) such that $a = bq + r$ and $0 \leq r \leq |b| - 1$. [See if you can figure out what commonly grasped mathematical process can convince you of the truth of this abstract-sounding statement!]

$P4$: Every dog will have his day.

The status of the statement $P4$ as a Proposition is dubious, unless we give precise definitions to all the terms involved. In any case, the establishment of such a Proposition is outside the scope of mathematics.

1.2. **Deductive Reasoning.** Suppose you somehow establish the following facts.

- A. Every cat named Tom is grey.
- B. Garret has a cat named Tom.

I am reasonably certain you have already established in your mind a third fact that follows from these two, namely that Garret has a grey cat. We say that the general statement A and the particular statement B together imply

- C. Garret has a grey cat.

This an example of *Deductive Reasoning*. That is to say, **if** you believe Statement A **and if** you believe Statement B, **then** you must necessarily believe Statement C. Note that whether or not you actually believe Statement A or Statement B is irrelevant to this discussion. The key idea being illustrated here is the validity of "implication arrow" $A + B \Rightarrow C$, not the validity of either A or B or C on its own. What we are establishing is that you cannot logically believe statements A and B to be true without also believing C to be true. The notation $P \Rightarrow Q$ is read "P implies Q" or "Q follows (logically) from P," or "If P, then Q" (meaning "If P holds, then Q holds.") or "Q is an implication of P." Okay that's enough.

In Mathematics, we begin with *axioms,* statements which we accept as true. We also give precise definitions for a variety of mathematical objects which, through much experience, we have decided recur often and deserve a name of their own and worthy of study. Our job is then to seek out true statements which follow logically from (i.e. are implied by) the axioms and the definitions we have given. How to choose these axioms and definitions is dictated in a natural way by the historical evolution of mathematics itself. And whether a given true statement about these objects is "interesting" or "useful" is also historically judged. You can be sure that the objects, axioms, concepts and theorems you encounter as a student of mathematics have "paid their dues" through hundreds and thousands of years of selective pressure; if they are still around, it means they are interesting and important.

Let me repeat: Our jobs as mathematicians is to seek out interesting statements which follow logically from (i.e. are implied by) the axioms and the definitions we have given. The statements that we establish to be logical consequences of the axioms are usually called "Propositions," the more important ones get dubbed "Theorems." When a Theorem has interesting consequences which follow without much difficulty from the theorem, we call the resulting propositions a "Corollary" of the theorem. When we need an intermediary or auxiliary result on the way toward proving a Proposition, we call that a "Lemma." Certain Lemmas become so indispensable, they become the mathematical equivalent of a screwdriver (the tool, not the drink, where is your mind?). It's great to prove a theorem that becomes famous and gets your name attached to it, but many mathematicians secretly pine for having a Lemma of theirs that becomes famous. You can decide for yourself: would you rather be known as the inventor of the screwdriver (the tool not the drink!) or the inventor of the HydroMagneticDestabilizingMultiChannelTransmogrifier? On an even more basic level, when we are plodding along trying to solve a problem having to do with one set of objects, over many years and much trial and error, a collection of properties and ideas may coallesce together and inspire us to define a new mathematical object. Sometimes these objects then are so fascinating on their own, that they become the thing many people want to study. How satisfying do you think that might be?

1.3. **Inductive Reasoning.** Before I say anything about Inductive Reasoning itself, let me say that later in the course we will discuss a very important Deductive Reasoning Tool known as "Mathematical Induction," and that Mathematical Induction and Inductive Reasoning are NOT to be confused with each other. Inductive Reasoning produces a statement that may or may or not be true. Mathematical Induction is a useful method for proving certain kinds of statements.

The term *inductive reasoning* refers to generalization based on observed patterns. It represents an "educated guess." It is used in all the sciences, as well as in everyday life.

Here are some examples:

$1 = 1^2$ is a perfect square; $1 + 3 = 4 = 2^2$ is a square; $1 + 3 + 5 = 3^2$ is a square, $1 + 3 + 5 + 7 = 4^2$ is a square; $1 + 3 + 5 + 7 + 9 = 5^2$ is a perfect square! Wow, (here comes the observation): when we take a small bunch of odd numbers in order (starting with 1) and add them up, we get a perfect square. Dude, maybe, (here comes the generalization) no matter how many odd numbers you take (starting with 1 and going in order) and add them up, you always get a perfect square! This is inductive reasoning. The next step in the inductive reasoning process is to test a few more cases to see if the pattern continues to hold: $1 + 3 + 5 + 7 + 9 + 11$ is just $(1 + 3 + 5 + 7 + 9) + 11$, i.e. the previous number plus 11, i.e. $25 + 11 = 36$. If we add the next odd prime we get $36 + 13 = 49$ which is the next square! If we add $49 + 15 = 64$ it's the next square. Dude, this seems like no accident. There has got to be a "reason" behind this. If you feel that way too, you are thinking like a mathematician already. Later in the course, we will discuss how Mathematical Induction can be used to prove the following more precise version of our observation:

Proposition. For any integer $n \geq 1$, the sum of the first n odd numbers is n^2.

NOTE ON HOMEWORK PROBLEMS AND BEING ON YOUR TOES IN CLASS: There number of problems given will be typically smaller than you expect from a mathematics class; but the problems will have several parts or will require quite a bit of thinking. So, get started on the homework right away, because it will take you a long time to come up

with the solutions and write them up carefully. One of the goals of this course is to help you express your arguments cogently, concisely, and correctly (**"the three co's"**). This is much harder than it sounds. **A certain number of points for the assignment go toward each of the co's.** Expect to suck at the three co's for a while, but if you keep working at it and study the style of the arguments you encounter throughout this course, you will improve steadily. By the way, as I am lecturing, feel free to rate my proofs on the three co's. You might have already guessed that I suck at "concisely." Let me be totally up-front about this: sometimes in class, I will present proofs that suck on purpose, even in correctness. Your job is to catch those instances. Why do I do this? Sometimes it helps to see someone else fall into a ditch in order to learn how to avoid one yourself.

Part 2. Logic and Sets

2. Elementary Logic

In mathematics, we are concerned with statements which are either True (T) or False (F). This is called the "truth value" or "validity" of the statement. If a statement is true, we say it holds; if false, we say that it does not hold. Most the time, the statements we encounter will depend on specific values taken by "variables" contained within them. For example, the statement $S_1 : x + 1 = 3$ may be true or false depending on the situation. The same goes for $S_2 : x - 1 = 1$. The same goes for $S_3 : y = 18$. But whereas for S_1 and S_3, there are four possible "truth value combinations", namely (S_1, S_3) is (T, T) or (T, F) or (F, T) or (F, F), (since we have been given no relationship between x and y), it is easy to see that S_1 can be derived from S_2 and vice versa. In other words, the only truth value combinations for (S_1, S_2) are (T, T) and (F, F). If you want, you can imagine that there is some kind of "logical glue" binding S_1 and S_2 together. Determining whether two given statements are held together by a logical glue or not is, in some sense, a large part of doing mathematics.

The basic relationship that can exist between two statements is that of *implication*, meaning if the validity of one statement P always guarantees the validity of another one Q, then we write $P \Rightarrow Q$ (to be read P implies Q) or Q follows from P. In $P \Rightarrow Q$, we call P the *premise* or *hypothesis* and Q is the *conclusion.* Let us review this: if P and Q are statements, then we can make a new statement $R : P \Rightarrow Q$ out of them. The *individual* validity of P or Q by itself does not affect the validity of $R : P \Rightarrow Q$. Rather, for R to be true, all one needs is that WHENEVER P HAPPENS TO BE TRUE, Q HAPPENS TO BE TRUE ALSO. So, R is false only we can establish some instance of P being TRUE and simultaneously Q being FALSE. In other words, $P \Rightarrow Q$ is TRUE if and only if P is true and Q is false is impossible.

Let us introduce some more notation. Here are some more ways to create new statements from given ones. First, there is the negation $-P$ of a statement P. This is a statement which is TRUE whenever P is FALSE and is FALSE whenever P is TRUE. In other words, the negation of P has the opposite truth value of P. Note that $-(-P) = P$. If P, Q are statements, then $P \wedge Q$, read P AND Q, is true if and only if BOTH P and Q are true. On the other hand, $P \vee Q$ read P OR Q is true if at least one of the two statements P,Q is true. Note the symmetries $P \wedge Q = Q \wedge P$ and $P \vee Q = Q \vee P$.

We have already seen that given P and Q, we can creat the statement $P \Rightarrow Q$. The converse of $P \Rightarrow Q$ is $Q \Rightarrow P$. A statement which always has truth value T is called a *tautology*, and one which always has truth value F is called a *contradiction.* The *contrapositive* of $P \Rightarrow Q$ is $-Q \Rightarrow P$. Two statements are called equivalent if the truth value of one is always equal to the truth value of the other. (They have coinciding truth tables).

Consider $P : x = 2$ and $Q : x^2 \leq 4$. We have $-P : x \neq 2$ and $-Q : x^2 > 4$. The implication $P \Rightarrow Q$ is true, but its converse $Q \Rightarrow P$ is false, because there are many values of x other than 2 such that $x^2 \leq 4$ (such as $x = -2$). The implication $-Q \Rightarrow -P$, however, i.e. the contrapositive of $P \Rightarrow Q$ is true, for if $x^2 > 4$, then $|x| > 2$ so $x \neq 2$. Let us show that the contrapositive of $P \Rightarrow Q$ is always equivalent to it. Well, $P \Rightarrow Q$ is true if and only if $P \wedge -Q$ is a contradiction. On the other hand, $-Q \Rightarrow -P$ is true if and only if $-Q \wedge -(-P)$ is a contradiction. But clearly $-Q \wedge -(-P) = P \wedge -Q$, so we are done. In the homework, you will recheck the equivalence of an implication and its contrapositive by examining their truth tables and showing that they coincide.

Since we know that an implication and its contrapositive are equivalent, sometimes instead of proving a statement directly, we can prove its contrapositive instead and be done more quickly. Here is an example.

PROBLEM: Prove that if n is an integer and n^2 is odd, then n is odd.

SOLUTION: We will prove the contrapositive, which will suffice for establising the theorem. The contrapositive is:

If n is an integer and n is not odd, then n^2 is not odd.

Recall that an integer not being odd means it is even, so we must prove: If n is an even integer, then n^2 is an even integer. This is easy: If n is an even integer, then $n = 2m$ for some $m \in \mathbb{Z}$ by definition. Then $n^2 = 4m^2 = 2(2m^2)$, so n^2 is even. This completes the proof.

3. Sets

At some early point in your childhood, you learned about binary phenomena. Binary phenomena are those that involve **two** possible "outcomes" or "states." (Yes—No, On—Off, 0—1,...). Have you ever seen a child delight in repeatedly switching lights on and off in a room? (Anyway, even if not, chances are good, you *were* such a child at some point). The source of this delight is the control they exert on the state of this binary phenomenon. I think it is also one of the reasons why very young children like Books of Opposites.

A very important concept in mathematics is that of a *set*, and at the heart of the concept of set is a binary phenomenon, namely that of "belongs—does not belong." This deserves some explanation.

What is a set? Basically, a set is a collection of objects. What a set does is to divide the world into two parts: the part that belongs to the set and the part that doesn't. If you introduce any object to the set, it will admit that object to be a member of the set or not to be a member of the set according to some definite, immovable, solid criterion. The binary phenomenon at the heart of the definition of a set is belong — not belong. To be a little more precise,

Definition 3.1. A *set* X is a well-defined rule for determining whether any given object belongs or does not belong to X. Those objects that belong to X are called *elements* of X or the *members* of X. Here, "well-defined" means that the rule for belonging is very precisely described, that no ambiguity enters into checking whether an object belongs or does not belong to X; it does not depend, for instance, on who the person checking the condition is, or at what time of day the checking takes place. Below I will give some examples of "poorly-defined" rules of belonging to give you a better idea of "well-defined" means.

One simple way of describing the rule which distinguishes between the objects that belong to the set or not is simply to give a list of the elements of the set. To list the elements, we put them inside curly brackets ($\{...\}$) and separate them by commas. For instance, here are some sets defined in this way:

$$X = \{a, b, c\}, Y = \{-2, 2, 4, 0, -4\}, , Z = \{\text{Aaron, Anna, Laura, Garret}\}.$$

These sets are all *finite* meaning they have a finite number of elements. We will talk much more about sizes of sets in a moment. Let us see if we can give some alternative ways of defining the sets we defined above.

Farshid: Can you describe the set X verbally without listing its elements explicitly?

Jaclyn: I got one: X is the set consisting of the first three letters of the alphabet.

Emily: That's pretty good, but I have a minor objection. I don't think this is precise enough yet. When you say "first three," somebody might list the letters of the alphabet in some other crazy way (say backwards) and then for them the "first three" would be z, x, y not a, b, c like we want.

Jaclyn: Hmm, that's easy to fix. I revise my definition thus: X is the set consisting of the first three (in forward alphabetic order) letters of the alphabet.

Alby: I thought Emily was going to say that it's not precise enough because we should specify they are all lowercase!

Jaclyn: Point well-taken, just add "lowercase" to my definition.

Nicolai: Dudes, you are forgetting there are more alphabets in the world "than are yet dreamt of in your philosophies!"[1] For instance, if an Iranian sees this definition, she might list the elements of X as "aleph, beh, peh!"[2]

Jaclyn: Touchée! I should add "English" to my defintion also, so now I have X is the set consisting of the first three (in forward alphabetic order) lowercase letters in the English Alphabet.

Farshid: I think it's reasonable to say that this is a precise definition of the set X. It has three elements, namely a, b, c. Good work.

I hope this little "play" helps describe what I mean by "well-defined." Here are some other examples of not well-enough-defined sets. You're probably too old to be familiar with "Elmo's World" but who knows, you're awfully young. Perhaps you know *Mr. Noodle*?! In case you don't know him, it suffices to know that at the start of each Elmo's World episode, Elmo asks Mr. Noodle to do someting related to "what Elmo is thinking about." Mr. Noodle then proceeds to make a real hash of it (which toddlers simply LOVE because they love feeling superior to Mr. Noodle and they all shout out advice to Mr. Noodle). Anyway, you get the idea.

Let's ask Mr. Noodle to give us a set.

Mr. Noodle: Let B be the set of beautiful people in the world.

John: Objection. Dude, that is, like SO not well-defined.

Mr. Noodle: Oh yeah. Let C be the set of Farshid's favorite numbers.

Steve: I can't even begin to register the depth of my objections. Would that be the set of Farshid's favorite numbers when he was 3 years old, or at this very moment right now, or .... ? Besides, how are **we** supposed to know which numbers are his favorite? Try to rely on a more universally-understood concept, Mr. Noodle!

Mr. Noodle: Okay, you're right. Let D be the set of positive numbers.

Jennifer: That's better, Mr. Noodle, but what kind of numbers do you mean (whole numbers, fractions, real numbers....)?

Mr. Noodle: Oh, sorry, you're right. This business of elements of sets is so hard!! Let E be the set with no elments!

Whoa! You might think Mr. Noodle is off his rocker, but actually he has just managed to give us a well-defined and, I might add, very important, set, namely *the empty set.* The empty set is **the set with no elements.** There are two standard notations for the empty set, namely either $\{\}$ or $\emptyset$. I think you will agree that the rule "nothing belongs to the empty

[1]Apparently Nicolai is taking a course on Shakespeare

[2]He is also taking a course on Elementary Farsi

set" is a very solid, impeccably well-defined rule! There is absolutely no ambiguity about that rule: the bouncer guarding the gate of the empty set has a very simple job, namely don't let anyone into the club, because nobody belongs. You might be uncomfortable with it at first: shouldn't all sets have at least one element? The answer is no, it's very convenient and useful to accept the empty set as a well-defined set; you'll see in a moment why.

Here are some more standard examples of sets (of numbers) you will encounter yet and again in this course as well as during the rest of your mathematical career.

The bold letter $\mathbb{N}$ is reserved for the set of **natural numbers**. The elements of $\mathbb{N}$ are the counting numbers, namely 1, 2, 3, 4, 5, 6, 7, ...

Remark. In some textbooks, 0 is also considered a natural number. Because of this ambiguity, I usually avoid using the notation $\mathbb{N}$ and specify more directly what I mean; for instance, I will write $\mathbb{Z}_{>0}$ for the set of positive integers, $\mathbb{Z}_{\geq 0}$ for the set of non-negative integers, etc.

The bold letter $\mathbb{Z}$ is reserved for the set of **integers**, positive, negative and 0. Namely $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. Leopold Kronecker, a very influential and important 19th century number theorist once said something to the effect that $\mathbb{Z}$ is at the heart of all mathematics. Being a number theorist myself, I agree with him.

Another set of great importance is the set $\mathbb{R}$ of all real numbers. We will discuss the definition of this set in a bit more detail later on, but for now, think of a number line extending in two infinite directions, with the integers marked on it at regular intervals. The real numbers are exactly the distances from 0 along this line to any point on the line; it is positive if the point is to the right of 0 and negative if the point is to the left of 0. A real number r can be specified by giving a "decimal expansion" for it, i.e. by an integer followed by a decimal point followed by an infinite sequence of digits (0-9). Later, we will also discuss the complex numbers $\mathbb{C}$; these are all the numbers of the form $a + b\sqrt{-1}$ where a, b are real numbers.

Now, more notation. We write $x \in X$ (read x is in X, or x belongs to X) to mean that x is an element of X and $x \notin X$ to mean that x is not an element of X. If X and Y are sets, we write $Y \subseteq X$ or sometimes $Y \subset X$ if every element of Y is an element of X. In this case, we say that Y is a subset of X. Another, sometimes useful, way of saying this is that no element of Y fails to be in X. Let's think about this a little. What does a set Y have to do in order for it NOT to be a subset of X? It would have to have an element $y \in Y$ which is not in X. In other words, it would have to have an element $y \in Y$ which fails to be in X. Therefore, if Y has no elements that fail to be in X, then Y is a subset of X.

If you think back to the definition of a set, you will see that two sets X and Y are equal (written $X = Y$) if and only if $Y \subseteq X$ and $X \subseteq Y$. In other words, two sets are the same exactly when their membership lists coincide. By the way, recall that any given object x either belongs to a set X or not, there is no concept of half-beloning or not, or belonging twice. For instance, if $X = \{a, b, c\}$ and $Y = \{a, b, c, c, c, c, c\}$, then $X = Y$. (It's as if the guy "c" keeps losing his membership card, so we had to issue multiple club membership cards for him, which does not change in any way the fundamental fact that he is a member of Club X.

Recall from our class discussion that in mathematics, the truth or falsity of any given statement is not as important as the *implications* that truth or falsity has on other statements; in other words, a mathematician is an investigator of the logical links between statements. In the same way, what is interesting to us is not so much sets of mathematical interest as **the**

relationships between sets. These "relationships" are mediated by *functions between sets.* I think it would be hard to refute that the single most important concept in all of mathematics is that of a function between sets.

Definition 3.2. A *map* or a *function* f from a set X to a set Y is a precise rule which assigns, to each element $x \in X$, a well-determined element $f(x) \in Y$. We write $f : X \to Y$ to denote a function with source X and target Y.

You have encountered many functions (usually from $\mathbb{R}$ to $\mathbb{R}$) in Calculus, so they are somewhat familiar. You probably expect a function to be given by an explicit "formula" but there are many acceptable ways of defining a function that do not involve a formula. What is needed is precision, so that there is no ambiguity in how to assign a value $f(x) \in Y$ to a given $x \in X$. For example, if we define $f : \mathbb{R} \to \mathbb{R}$ by $f(x)$ is the nearest integer to x, then it is easy to that $f(17.4) = 17$ for example, and $f(17) = f(16.9) = 17$, but what is $f(17.5)$?! This number is equally close to 17 or 18. As defined, f is not a function. If we define $f(x) = [x]$ is the greatest integer not exceeding x, then f is well-defined and $[17.5] = 17$. Here is another example. Let α be the set whose elements are the 26 lowercase English letters, and β the set of 26 uppercase English letters. The map sending a lowercase letter to its corresponding uppercase letter is a well-defined map from α to β.

Definition 3.3. Let X, Y be sets. Then $X \cap Y = \{a | a \in X \wedge a \in Y\}$ is called the intersection of X and Y, and $X \cup Y = \{a | a \in X \vee a \in Y\}$ is the union of X and Y. We say that X and Y are disjoint if $X \cap Y = \{\}$ is empty. We define the set exclusion $X \setminus Y$, sometimes denoted $X - Y$ as follows:

$$X \setminus Y = X - Y = \{x \in X | x \notin Y\}.$$

In other words, $X \setminus Y$ consists of all those members of X which are not also members of Y.

Little Warmup Exercises: A. The President of Costless meets the President of Price-Chopped and brags: "If C denotes the set of members of Costless and P denotes the set of members of *PriceChopped*, then $P \setminus C$ is empty." The President of PriceChopped retorts: You don't say. Consider this: I am not just the President of PriceChopped, I am also one of its members; however, I am *not*, nor have I ever been, a member of Costless. The President of Costless then leaves in a huff. Explain why.

B. Suppose X, Y, Z are sets and let $A = X \setminus Y$, $B = Y \setminus X$. First show that A and B are disjoint. Suppose $\Delta = \{A, B, Z\}$ is a partition of $X \cup Y$. (See below for what "partition" means). What can you say about $X \cap Y$? Explain. [3]

It is clearly desirable that intersection be an operation that takes as input two sets and gives as output another set. Since many pairs of sets have no element in common, whose intersection is then void, it is a matter of tremendous convenience to define the empty set to be the set with no elements.

Try this experiment: locate a kid aged at least 3 years – the optimal age for this experiment is probably 5 or 6; if the kid is a teenager, the experiment will probably fail miserably as it will be difficult as the reaction will probably consist of little more than "What*ever*; you're weird". Anyway, find a kid and then say "Hey Kid, listen to this: 1,2,3,4,6,5,7,8,9,10. Chances are the kid will go ballistic. Kid: "Nooooo! That's not how it goes!" You: "What's

[3] free advice: draw some blobs and call them X, Y etc. then shade in different regions for A, B etc.

wrong with it?" Kid: You just can't do that! It's against the rules." You: "But if you put ten pennies in front of me and ask me to count them, I'll get the right anser my way." Kid: "I don't care; you just can't say 6 before 5, don't you get it?" (I'm still waiting for the kid who would say: "Yeah, but if I put 5 or 6 pennies in front of you, you'll mess up!")

What this experiment demonstrates (I have tried it quite often myself) is that early in life, people become familiar with the notion that the set $\mathbb{Z}$ is endowed with an ordering, in other words, the integers come in a specific order. To a child, it's very hard, in fact, to separate the concept of "number" from that of "counting" and of course counting requires that the numbers be recited in a specific order. The notation that we use for ordering the integers is $\leq$. How is this ordering defined? As usual, there are multiple perspectives. If you list the integers on the number line in the usual way, with 0 in the middle and travelling to the right to 1, then 2, etc. and starting at 0 and going left to -1, then -2, etc., then for $x, y \in Z$, $x \leq y$ means that y is on top of or to the right of x. The same procedure serves to define an ordering of the real numbers. Another perspective is that $\mathbb{Z}$ is divided into two subsets: $\mathbb{Z}^+ := \{1, 2, 3, 4, \cdots\}$, $\mathbb{Z}^- := \{-1, -2, -3, -4, \ldots\}$, and $\{0\}$.[4]. Then $x \leq y$ means that $y - x \in \mathbb{Z}^+ \cup \{0\}$.

The notion that the integers are ordered according to their "size" brings up, at some early age, the question of the "largest" and "smallest" numbers. Surprisingly early, children can decide on their own that there is no largest integer. This is a fundamental and first-rate mathematical theorem. Once they become familiar with negative numbers, children also accept fairly quickly that there is no smallest integer.

However, if one restricts attention to positive integers $\mathbb{N}$ or non-negative integers $\mathbb{Z}_{\geq 0}$, then there is of course a smallest integer, namely 1 (respectively 0). A slightly more fancy version of this observation has a fancy name since it is often invoked as a useful and fundamental fact of arithmetic. But first, let's have a definition.

Definition 3.4. Suppose S is a subset of $\mathbb{R}$. We say that $l \in S$ is a least element of S if for all $s \in S$, we have $l \leq s$. We say that $g \in S$ is a greatest element of S if for all $s \in S$, we have $s \leq g$.

The Well-Ordering Principle Given any subset $S \subseteq \mathbb{Z}_{>0}$ of the set of positive integers, there exists a least element of S.

WARNING. I have purposely inserted an error in the above statement. This error is of a standard type called "forgetting to include one of the hypotheses." Can you find the missing hypothesis?! [If not, read on, you'll find it later.]

What is responsible for the Well-ordering Principle is the fact is that the integers are a "discrete set." Its elements don't get bunched up together, so to speak; they maintain a respectful distance from each other.

NON-EXAMPLE[5]: For instance, the open interval $I = (0, 1) = \{x \in \mathbb{R} | 0 < x < 1\}$ has no least element and no greatest element. Let us begin to give a proof of this fact. Let us prove that I has no largest element (the other half is left to the reader, that means YOU!)

[4] "There are two kinds of mathematicians: those who can count, and those who can't." – John Conway

[5] You no doubt know the important maxim that whenever you learn a new word, it's a good idea to construct a simple sentence containing that word. Well, in learning mathematics, we have something along these lines: Whenver you are presented with a new definition, you need to do two things; first, create for yourself an example of the object just defined, then create *several* examples of objects which "just miss" fitting the definition you have just encountered. Chances are you will learn as much, or perhaps more, from the *non*-examples as from the genuine bona-fide example.

Interlude on Proof by Contradiction

We must show that I does NOT have a greatest element. One way to do this is to use the **proof technique known as "Contradiction."** The idea behind this method is simple: you wish to prove that some premise P implies some conclusion Q. Recall that $P \Rightarrow Q$ simply means that $P \wedge -Q$ is impossible, i.e. is a contradiction, (i.e. its truth table has constant value FALSE.) So, the idea behind proving $P \Rightarrow Q$ via contradiction is to place yourself inside a universe where P is true and $-Q$ is true (i.e. P is true and Q is false). Our job is to show that this is impossible, so we look around this world we are imagining and look for cracks in its foundation: once we find a statement that is logically untenable and which must follow from the assumptions $P \wedge -Q$, then we will conclude that this world cannot exist, so $P \wedge -Q$ must be impossible, so whenver P is true, Q must also be true, i.e. $P \Rightarrow Q$.

Back to the statement we wish to prove, namely, $I = (0, 1)$ does not have a greatest element.

Proof By Contradiction. Let us suppose $I = (0, 1)$ does in fact admit a greatest element x. By definition, $x \in I$ and if $y \in I$, then $x \leq y$. Let $y = x + (1 - x)/2$. [This is the definitive step; we have looked around in this weird universe where x is a greatest element, and noticed that since there is quite a gap between x and 1, we can slip in the number y in $(0, 1)$ and check that $y \not\leq x$; this contradicts the fact that x is a greatest element of I]. In other words, on the one hand, $y = 1/2 + x - x/2 = (1 + x)/2 < (1 + 1)/2$ since $x < 1$ thanks to $x \in I$ so $y < 1$, and on the other hand, $y - x = (1 - x)/2 > 0$ i.e. $y > x$. But then $x < y < 1$ so $y \in I$, so x is NOT the greatest element of I. This is a contradiction. Thus the hypothesis that I has a greatest element leads to an absurdity; it must therefore be false. We conclude that I has no greatest element. I have made the argument quite repetitive here, because I want to hammer home all the different components. Now prove that I has no least element, and give your argument cogently, concisely and correctly.

Definition 3.5. A set is called *finite* if it has a finite number of distinct elements. To be more formal, a set X is finite if there exists a non-negative integer n such that for all sequences $x_0, x_1, x_2, \ldots, x_n$ of elements of X, there exist integers $0 \leq i < j \leq n$ such that $x_i = x_j$. If X is finite, then we can define the order of X, also called its cardinality and denoted by $|X|$ as follows: $|X|$ is the **least** non-negative integer n with the property that for all sequences $x_0, x_1, x_2, \ldots, x_n$ of elements of X, there exist integers $0 \leq i < j \leq n$ such that $x_i = x_j$. Why does such an integer n exist? By the assumption that X is finite, we know that there exists at least one such non-negative integer n; now by the Well-Ordering Principle [6] there must be a least such integer $n \geq 0$.

In less formal language, the maximal number of distinct elements in X is called the size of X or the cardinality of X and is denoted by $|X|$ or sometimes $\#X$.

A set which is not finite is called *infinite*. Equivalently, a set X is infinite if and only if for every positive integer n, there exists a sequence of n (pairwise) distinct elements of X. In this case, we write $|X| = \infty$. As we will see later, this statement is slightly misleading, as there are in fact multiple *gradations* of infinity.[7]

For example, let us look at the finite set $X = \{a, c, b, c, c, b\}$; for this set, we have the sequence a, c, b with no repetitions, of length 3, but in any sequence of length 4 or more

[6] NOTE that the missing hypothesis in the statement of the Well-Ordering Principle was that $S \neq \{\}$!!

[7] **Say what?!?**... stay tuned!

coming from this set, one of these letters at least must be repeated, so $|X| = 3$. On the other hand, the set $I = (0, 1)$ is infinite because, for instance, given any positive integer n, the sequence of length n $3/4, 1/2, 1/3, \ldots, 1/n$ consists of n pairwise distinct elements of I.

Of course, in real life, this is not how we count how many things are in a set. Again, imagine interviewing a child: Put a bunch of dots on a page, say two or three times the kid's age and ask her to count how many dots there are. The child will probably point to each dot and count in sequence, 1,2,3, etc. An attempt will be made to make sure of two things: a) each dot is included in the count, and b) no dot is counted more than once. In other words, she will try to "cover" each dot exactly once. What is the child's method? She attempts to set up a "one-to-one correspondence" between the dots on the page and a set of type $\{1, 2, 3, 4, \ldots, n-1, n\}$. Then the child will know that there are n dots. The important principle behind her method is that whenever two sets are in one-to-one correspondence, then they must have the same cardinality, or size.

Definition 3.6. A map $f : X \to Y$ is called *onto (or surjective)* if for all $y \in Y$, there exists $x \in X$ such that $f(x) = y$. A map $f : X \to Y$ is called one-to-one (or injective) if whenever $x_1, x_2 \in X$ and $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$. A map $f : X \to Y$ is called *bijective* (or a *set-equivalence*, or most commonly a *one-to-one correspondence*) if it is injective and surjective. We say that two sets X and Y are equivalent if there exists a bijective map $f : X \to Y$.

Example 3.7. The sets $X = \{a, b, c, d, e, f, g\}$ and $Y = \{A, B, C, D, E, F, G\}$ are equivalent. For instance, the map that sends a lower-case letter to the corresponding upper-case letter is clearly one-to-one and onto. The sets $A = \{1, 2, 3\}$ and $B = \{9, 0\}$ are not equivalent, because any map $f : A \to B$ is not injective. How to prove this?

Lemma 3.8. *Suppose $f : X \to Y$ is an injective map of a set X into some set Y. If X has n distinct elements, then Y has at least n distinct elements.*

Proof. Let $x_1, \ldots, x_n$ be n distinct elements of X. Since f is injective, and $x_i \neq x_j$ for $1 \leq i < j \leq n$, we have $f(x_i) \neq f(x_j)$ for the same i, j. Hence the elements $f(x_1), \ldots, f(x_n)$ are n distinct elements of Y. So, Y has at least n distinct elements. □

The moral is that you cannot shove a set of size 3 into a set of size 2 in a one-to-one fashion!

If X is a set, then *the power set* of X is the set $\mathcal{P}(X)$ consisting of all subsets of X. In other words, the *elements* of $\mathcal{P}(X)$ are *subsets* of X, and every subset of X is in fact a member of $\mathcal{P}(X)$.

Now, let's get one thing straight. If X is any set, any set at all, then the empty set $\{\}$ is a subset of X. Let's check why in two ways. First, the direct way is to check that every member of the empty set is in X. You might say that since the empty set has no elements, then we cannot do this ... but I say, since the empty set has not elements, what we have to check is an *empty condition*, so it holds by fiat, so to speak. There is nothing to check, so the fact is true. (If you wish, think that a statement is true unless proven false). Now here is the second way: perhaps you will find it more convincing. In order for the empty set to fail to be a subset of X, we have to produce an element of $\{\}$ which does not belong to X. Since we can find no such suspect, the empty set must be a subset of X!

Example 3.9. List all subsets of $X = \{a, b, c\}$. Well, there is only one subset of size 3, namely X itself. There are three ways to "leave out" an element, so there are three subsets of size 2, namely $\{a, b\}, \{a, c\}, \{b, c\}$. Similarly, there are three ways to pick one element to leave in, so there are three subsets of size one, $\{a\}, \{b\}, \{c\}$. Finally, there is one subset of size 0, namely $\{\}$. Altogether, there are $1+3+3+1 = 8$ subsets of X. Thus, $\mathcal{P}(X)$ has size 8.

Definition 3.10. A partition Δ of a set X is a subset $\Delta \subseteq \mathcal{P}(X)$ of the power set of X with the following two properties: i) if $Y_1, Y_2 \in \Delta$, then $Y_1 \cap Y_2 = \{\}$, and ii) the union of all elements of Δ is X. In other words, a partition of X a collection of pairwise disjoint subsets of X whose collective union is X.

In a future unit, we will explore the importance of partitions in more depth when we talk about *equivalence classes.*

Example 3.11. A. Let $X = \{1, 2, 3, 4\}$. Then $\Delta = \{\{1\}, \{2, 3\}, \{4\}\}$ is a partition of X. At the two extremes, we can partition X via the largest possible subdivision, namely $\Delta = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, or the smallest one $\Delta = \{X\}$.

B. If S is the set of students at an elementary school, then two natural partitions are the "class partition" and the "grade partition." Which one of these do you think is potentially "finer?"

Now, a caveat regarding our treatment of sets, for the sake of honesty. I have given a relatively informal definition of what it means for a rule defining a set to be "well-defined," but in truth, this concept is actually an extremely deep and subtle one. For practical purpose, we may and we will almost entirely avoid all the depth and all the weird subtleties (to begin exploring them, if you wish, consult some books on logic or "Gödel-Escher-Bach." by D. Hofstadter). For most mathematicians, the particular sets we use in everyday work are of a simple, standard type, which do not exhibit any pathological behavior.

So far, we have defined sets by listing all the elements. This is not always the most convenient method. Sets are often described in terms of the conditions satisfied by elements of a larger set. For instance $\mathbb{N} = \{x \in \mathbb{Z} | x \geq 1\}$.[8] This should be read as "the set of natural numbers consists of **all** x in the set of integers that satisfy $x \geq 1$. "

Some more examples: $\{x \in \mathbb{R} | x^2 = 1\}$ should be read as the set of all real numbers whose square is equal to 1. (Thus, the elements of this set are 1 and -1.

$\{x \in \mathbb{Z} | x > 0 \wedge x \text{ is odd } \}$ is the set of positive odd integers, whereas $\{x \in \mathbb{Z} | x > 0 \wedge x \text{ is odd } \}$ is the union $\{x \in \mathbb{Z} | x > 0\} \cup \{2k + 1 | k \in \mathbb{Z}\}$.

Finally, let us note the use of the "dummy" variable [9] namely $\{x \in \mathbb{R} | x^2 = 1\} = \{\zeta \in \mathbb{R} | \zeta^2 = 1\}$.

Definition 3.12. Suppose $X_1, \ldots, X_n$ are sets. The direct product of $X_1, \ldots, X_n$, denoted by $X_1 \times X_2 \times \cdots \times X_n$, is the set of all ordered n-tuples $(x_1, x_2, \cdots, x_n)$ where $x_i \in X_i$ for $i = 1, 2, \ldots, n$.

For examples, $\mathbb{R} \times \mathbb{R}$ is the set of all ordered pairs (x, y) where $x, y \in \mathbb{R}$. In other words, $\mathbb{R} \times \mathbb{R}$ can be thought of as the set of coordinates of points in the usual xy plane.

[8] The little vertical line | should be read as "such that"; sometimes instead, one uses the colon : instead of |.

[9] the poor variables get no respect.

Part 3. Sets and Maps

4. Partitions of Sets

If X is a set, then *the power set* of X is the set $\mathcal{P}(X)$ consisting of all subsets of X. In other words, the *elements* of $\mathcal{P}(X)$ are *subsets* of X, and every subset of X is in fact a member of $\mathcal{P}(X)$.

Recall that two sets Y and Z are called *disjoint* or sometimes *non-overlapping* if $Y \cap Z = \{\}$. Instead of saying "Y intersect Z is empty," sometimes I might say Y and Z *do not meet*, or Y and Z *have no one in common* or Y and Z *have no intersection.* You shouldn't take the latter phrase too literally, of course, because Y and Z *do* have an intersection even when they are disjoint, it's just that their intersection happens to be a set devoid of elements, i.e. is the empty set. If X, Y, Z are three sets, we say that X, Y, Z are disjoint if $X \cap Y \cap Z = \{\}$. Let's be very precise about what we mean by the intersection of a bunch of sets (so far we only defined the intersection of two sets).

Say we have a collection of sets and we want to take their intersection. First of all, how do we list this "collection" of sets? One way is to "index" them; in other words, we have some auxiliary set A which we will use to index our sets, meaning for each $\alpha \in A$, we have one set X_α in our collection. We can then say that we have a collection $\mathcal{C} = \{X_\alpha | \alpha \in A\}$ of sets. Their intersection is defined as

$$\cap_{\alpha \in A} X_\alpha = \{x | x \in X_\alpha \text{ for all } \alpha \in A\}.$$

If this intersection is the empty set, we say that the collection of sets is disjoint. There is a *much* stronger condition, namely that of the collection of sets being *pairwise disjoint.* This means that for *every* pair $\alpha, \beta \in A$, $X_\alpha \cap X_\beta = \{\}$. For instance, the collection $\{1,2\}, \{2,3\}, \{3,4\}$ is disjoint (nobody is in all three sets simultaneously) but not pairwise disjoint.

It happens often that we want to divide up a set into smaller pieces. For example, consider the set of students in Math 300 this semester. We may break them up (or "classify" them) in quite a few natural ways: 1. according to the student's major; 2. according to the student's registration status (fresh., soph., jr., sr., other) 3. according to the student's TA. Pick any one of these schemes for breaking up the set of Math 300 students into smaller subsets and check that it enjoys the following property: every student belongs to exactly one subset. This means that the subsets do not overlap, and that the union of all the subsets is in fact the whole class. We say that each of these schemes for breaking up the class is a *partition* of it.

Definition 4.1 (BOGUS ALERT[10])**.** A partition Δ of a set X is a subset $\Delta \subseteq \mathcal{P}(X)$ of the power set of X with the following two properties: i) Given $Y, Z \in \Delta$ with $Y \neq Z$, then $Y \cap Z = \{\}$, and ii) the union of all elements of Δ is X. In other words, a partition of X is a collection of pairwise disjoint subsets of X whose collective union is X.

[10]Recall that occasionally in class and in these notes, I will make bogus, slightly wrong, statements. Okay, let's be frank about it, I will just lie on purpose. The goal is to show you how doing mathematics is a progressive process, full of false starts and second attempts. We learn a lot everytime we crash the plane. Sometimes when I want you to be on the lookout for my lies, I will give you the BOGUS ALERT signal, and sometimes I won't. In this particular case, I want you to be on the lookout for a missing assumption which would make the definition more elegant.

In the near future, we will explore the importance of partitions in more depth when we talk about *equivalence classes.*

Example 4.2. A. Let $X = \{1, 2, 3, 4\}$. Then $\Delta = \{\{1\}, \{2, 3\}, \{4\}\}$ is a partition of X. At the two extremes, we can partition X via the largest possible subdivision, namely $\Delta = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, or the smallest one $\Delta = \{X\}$.

B. How many partitions does the empty set have?

C. If X is a non-empty set, then $\Delta = \{X\}$ is a partition, indeed it is the *coarsest* possible partition of X. At the other extreme, if Δ consists of all singleton subsets of X, i.e. $\Delta = \{\{x\}|x \in X\}$, then Δ is the *finest* partition of X.

D. If S is the set of students at an elementary school, then two natural partitions are the "class partition" and the "grade partition." Which one of these do you think is potentially "finer?"

E. Consider the partition $X = \{X, \{\}\}$. It *is* a partition, according to our definition, because $X \cap \{\} = \{\}$ and $X \cup \{\} = X$. But is it not a slightly annoying partition? Why do I find it annoying? I'll tell you why. Suppose X is a set and Δ is a partition of X where none of the chosen subsets is empty. So, Δ specifies a certain scheme for how to divide up Δ into "teams." Now suppose we add to Δ the empty set. Has the scheme for dividing up the teams changed at all? No. But the partition has! Putting the empty set in the partition doesn't really add or subtract anything from how the set is actually broken up; it doesn't serve any useful purpose. It just dirties the waters, so to speak. So, it would be more clean, more tight, more *elegant*, if we don't allow the empty set to be one of the subsets allowed in the partition. This is the source of the BOGUS ALERT I gave you earlier. Accordingly, below is BOGUS-ALERT-FREE definition of partitions.

Definition 4.3. If X is a set, and $\Delta \subseteq \mathcal{P}(X)$ is a collection of subsets of X, then Δ is a *partition* of X if: i) for all $Y \in \Delta$, $Y \neq \{\}$; ii) for all $Y, Z \in \Delta$ with $Y \neq Z$, $Y \cap Z = \{\}$, and iii) $\cup_{Y \in \Delta} Y = X$[11] . In other words, a partition of X is a collection of non-empty, pairwise disjoint (i.e. non-overlapping) subsets of X whose union is X.

Now, how many partitions does the empty set have?!

I think it is probably intuitively clear what I meant by "coarser" and "finer" partions in the above examples. Here is a new kind of exercise or "problem" you may not have tried before: try your hand at making up a formal definition by completing the following.

Definition 4.4. Let X be a set, and suppose Δ and Δ' are two distinct partitions of X. We say that Δ is *finer* than Δ' if ..... We say that Δ is *coarser* than Δ' if Δ' is finer than Δ.

5. Maps between sets

One of the maxims of modern mathematics is that "objects are not as important as the relationships between objects." The meaning of this maxim will perhaps become clearer as the course goes on, and I don't want to dwell on it here except to point out how, in a way, we have already encountered one instance of this when studying implications. Recall that in pondering an implication statement $P \Rightarrow Q$, the truth of P or that of Q in any given situation is not so important; what *is* important for $P \Rightarrow Q$ is whether there is some logical

[11]The notation $\cup_{Y \in \Delta} Y$ means "the union of all Y as Y runs over all elements of Δ."

glue which rules out Q being false whenever P happens to be true; whether that "logical glue" exists or not is what determines the truth or falsity of $P \Rightarrow Q$, so it's the *relationship between P and Q* that is of import when we ponder $P \Rightarrow Q$.

I have said that sets are the primal "words" in the language of mathematics. Combining this with the maxim stressing the importance of relationships between objects, we conclude that a central role in mathematics must be played by the gadgets that measure relationships between sets. The most common such gadgets, as we alreay mentioned, are called *functions* or *maps.* A more general, more flexible, but less frequently encountered, notion for expressing relationships between sets is that of a *relation*, which we will talk about a little later. For now, let us concentrate on functions and begin by recalling its definition.

Definition 5.1. A *map* or a *function* f from a set X to a set Y is a precise rule which assigns, to each element $x \in X$, a well-determined element $f(x) \in Y$. We write $f : X \to Y$ to denote a function with *source* X and *target* Y.

You have encountered many functions (usually from $\mathbb{R}$ to $\mathbb{R}$) in Calculus, so they are somewhat familiar. You probably expect a function to be given by an explicit "formula" but there are many acceptable ways of defining a function. What is needed for a definition to be valid is simply a high degree of precision, so that there is no ambiguity in how to assign **one single value** $f(x) \in Y$ to any given $x \in X$. For example, if we define $f : \mathbb{R} \to \mathbb{R}$ by $f(x)$ is the nearest integer to x, then it is easy to that $f(17.4) = 17$ for example, and $f(17) = f(16.9) = 17$, but what is $f(17.5)$?! This number is equally close to 17 or 18. As defined, f is not a function. If we define $f(x) = [x]$ is the greatest integer not exceeding x, then f is well-defined and $f(17) = [17.5] = 17$.

In calculus, one often specifies a function in shorthand by alluding to part of a formula or symbol defining it, viz. "the sine function", or "the exponential function" or "$1/x$" with utter disregard for the specification of its source and target. Therefore, you must retrain yourself in two ways: first, think of functions as having three components: I. a source set; II. a target set; III. a rule describing how to assign one single element in the target set to each element in the source. Second, be more flexible about what kinds of sets can be sources and targets for maps and also about how the rule for assigning elements of the target to the source elements is described.

Every map $f : X \to Y$ from a set X to a set Y establishes a connection, a relationship, between these two sets. Sometimes this connection is very "tight," setting up a correspondence between the two sets showing them to be similar, and sometimes not. There are two important ways to measure how "tightly" a function binds two sets. Consider for example, the case of P being the players on the current UMass Amherst Hockey Team, and N being the set of whole numbers between 0 and 99 inclusive. There is a map $f : P \to N$ which assigns to each player his jersey number. This relationship is tight in one way, namely no two players are assigned the same number (for obvious reasons, so that the map can be used to identify players uniquely based on their jersey number). But there is another way in which this map is **not** tight, namely there are gaps, so to speak, between the numbers that have been assigned to the players. Not every number in the range 0—99 actually corresponds to a player (for instance Zech Klann is no. 39, but no one wears the no. 38 jersey).

As another example, if X is the set of TAs for Math 300 this term, and $Y = \{\text{undergrad, grad}\}$, then I'm sure you'll agree there is a naturally defined map $f : X \to Y$ which assigns to each TA his/her degree program status. Now for this map, there are no "missed values" because,

for instance, Molly is a graduate TA, and Aaron is an undergraduate TA, so the map is tight as far as hitting every possible value is concerned. But now there are four undergraduate TAs, so you cannot necessarily determine the identity of a TA based soleley on his/her degree program status.

On the other hand, say Y is the set consisting of the 26 lowercase English letters and $X = \{n \in \mathbb{Z} | 1 \leq n \leq 26\}$. The map $f : X \to Y$ sending n to the nth letter of the alphabet in the standard order does not miss any letters and no two distinct n go to the same letter. In other words, each number corresponds to exactly one letter and vice versa. In this case, the map is as tight as possible: we call it a "set-equivalence" (or sometimes a "one-to-one correspondence"[12]).

Let us formalize these concepts.

Definition 5.2. A map $f : X \to Y$ is called *onto (or surjective)* if for all $y \in Y$, there exists $x \in X$ such that $f(x) = y$. A map $f : X \to Y$ is called one-to-one (or injective) if whenever $x_1, x_2 \in X$ and $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$. A map $f : X \to Y$ is called *bijective* (or a *set-equivalence*) if it is injective and surjective. We say that two sets X and Y are *equivalent* if there exists a bijective map $f : X \to Y$.

Example 5.3. The sets $X = \{a, b, c, d, e, f, g\}$ and $Y = \{A, B, C, D, E, F, G\}$ are equivalent. For instance, the map that sends a lower-case letter to the corresponding upper-case letter is clearly one-to-one and onto. The sets $A = \{1, 2, 3\}$ and $B = \{9, 0\}$ are not equivalent, because any map $f : A \to B$ is not injective. How to prove this?

Lemma 5.4. *Suppose $f : X \to Y$ is an injective map of a set X into some set Y. If X has n distinct elements, then Y has at least n distinct elements.*

Proof. Let $x_1, \ldots, x_n$ be n distinct elements of X. Since f is injective, and $x_i \neq x_j$ for $1 \leq i < j \leq n$, we have $f(x_i) \neq f(x_j)$ for the same i, j. Hence the elements $f(x_1), \ldots, f(x_n)$ are n distinct elements of Y. So, Y has at least n distinct elements. □

The moral is that you cannot shove a set of size 3 into a set of size 2 in a one-to-one fashion!

6. Composites and inverses of maps

Let's say you want to fly from Columbus Ohio to London England. Chances are there are no direct flights. So, what do you do? You fly from Columbus to Chicago say, then catch a flight from Chicago to London. You "concatenate" the two flights to create a path from Columbus to London. You can do a similar thing with maps of sets. Suppose $f : X \to Y$ is a map and $g : Y \to Z$ is a map. Then, can we define a map from X to Z by "concatenating" the two maps to get a map $h : X \to Z$? How do we specify a map? We take an arbitrary $x \in X$ and we have to describe how to attach an element of Z to it. Well, how about $h(x) = g(f(x))$? In other words, first we send x to $f(x) \in Y$, then, having safely arrived at $f(x)$, since $f(x)$ is in Y, we can send $f(x)$ to Z via g, giving us $g(f(x))$. The picture is

$$\begin{array}{ccc} X & \to & Y \to Z \\ x & \mapsto & f(x) \mapsto g(f(x)) \end{array}$$

[12]Make sure you note that there is a big differnce between a map being one-to-one and a map being a "one-to-one correspondence"; the former just means injective but the latter means injective AND surjective. This is one of the reasons I prefer the terms injective/surjective over one-to-one and onto.

Note that in the picture, the map f comes before g, but in the "algebra", g comes before (to the left of) f. This can be a little disorienting at first, but you'll get used to it with practice. We write that $h = g \circ f$. This means that the source of h is the source of f and the target of h is the target of g, and that for an arbitrary x in the source of f, $h(x) = g(f(x))$. We say that the map h is the composite of the maps f and g.

You have seen composition of functions before when you studied the chain rule in calculus for example. Thus, the function $h(x) = \sin(x^2)$ (a function from $\mathbb{R}$ to $\mathbb{R}$) for example is a composite of two maps, namely, if $f : \mathbb{R} \to \mathbb{R}$ is the map given by $f(x) = x^2$ and $g : \mathbb{R} \to \mathbb{R}$ is given by $g(y) = \sin(y)$, then $g \circ f = h$. (You should check this).

If X and Y are sets, we let $\mathrm{Maps}(X, Y)$ be the set consisting of all maps from X to Y.

Now let us study maps from a set X to itself. (These are sometimes called self-maps of X). Among all maps from X to itself, there is one very special one, called the *identity map* and usually denoted id_X; it is a very neutral map, as it is defined simply by $\mathrm{id}_X(x) = x$ for all $x \in X$.

Recall that a bijective map $f : X \to Y$ sets up a one-to-one correspondence between X and Y, so that if at any point we find ourselves at some point in Y, we know exactly one point in X to which it corresponds and if we are at a particular point in X, we know exactly which single point in Y it is linked with. For example, if there are four students in a section, say Emily Braley, Amy Jackson, Mike Malloy, and John McColgan, then we have a map $F \to L$ where $F = \{$Emily, Amy, Mike,John$\}$ and $L = \{$Braley, Jackson, Malloy, McColgan$\}$ which links each students' first name with his or her last name. Note that this link is injective and surjective so that we can "go back and forth" from one list to the other without any ambiguity.

More generally, if $f : X \to Y$ is a map, then we may ask: Under what conditions will there may be a map g in the reverse direction $g : Y \to X$ such that the maps f and g "undo" each other, i.e. $g(f(x)) = x$ for all $x \in X$ and $f(g(y)) = y$ for all $y \in Y$? We can rephrase this as follows: Given a map $f : X \to Y$, when is there a map $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$? If such a map exists, we call it an *inverse function* for f and denote it by f^{-1}.

IMPORTANT EXAMPLE: It is very significant that for an inverse function, we insist not only that g undoes f, but that in turn f undoes g. As an example to illustrate that this does make a difference , let $X = \{1\}$ and let $Y = \{-1, 1\}$. Consider the map $f : X \to Y$ defined by $f(1) = 1$. Since X has only one element, a moment's thought will convince you that there is only one map from Y to X, namely, $g : Y \to X$ which sends everyone to 1, i.e. $g(1) = g(-1) = 1$. Now it is clear that g undoes f, i.e. $g \circ f = \mathrm{id}_X$. But note that $f \circ g \neq \mathrm{id}_Y$!. Indeed, $(f \circ g)(1) = 1$ and $(f \circ g)(-1) = 1$, so f does not undo g. If you think about it, you will see that since X and Y do not have the same cardinality, no map from X to Y could ever undo a map from Y to X.

Definition 6.1. We say that a map $f : X \to Y$ admits an inverse if there is a map $g : Y \to X$ such that $g(f(x)) = x$ for all $x \in X$ and $f(g(y) = y$ for all $y \in Y$. In other words, g is an inverse of f if $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.

Now we can give a very satisfying answer to our question about which maps admit inverses. The following is a very important and useful theorem. You should study its statement and proof carefully.

Theorem 6.2. *A map $f : X \to Y$ admits an inverse if and only if f is bijective.*

Proof. Let us suppose f is bijective. As we saw in class, this means that for each $y \in Y$, there is a unique x such that $f(x) = y$. Now define $g : Y \to X$ as follows: for each $y \in Y$, since there is one and only one $x \in X$ such that $f(x) = y$, then by putting $g(y) = x$, g is well-defined. We also clearly have $g(f(x)) = g(y) = x$, so $(g \circ f)(x) = x$ for all $x \in X$, so $g \circ f = \mathrm{id}_X$; similarly, for $y \in Y$, $f(g(y)) = f(x) = y$. Thus, we have shown that if f is bijective, then it admits an inverse. Let us now prove the converse. Suppose g is an inverse for f. Let us show that f is injective. Suppose $x, x' \in X$, and $f(x) = f(x')$. Then $g(f(x)) = g(f(x'))$. But $g(f(x)) = x$ and $g(f(x')) = x'$ by the fact that g is an inverse to f, hence $x' = x$. We have shown that the only way for $x, x' \in X$ to have the same image $f(x) = f(x')$ under f is for x to be equal to x'. Thus, f maps distinct elements of X to distinct elements of Y, i.e. f is injective. Now we'll show that f is surjective. Given $y \in Y$, we must find $x \in X$ such that $f(x) = y$; that's easy: let $x = g(y)$. Then $f(x) = f(g(y))$ and $f(g(y)) = y$ by the definition of "inverse function." □

NOTE: Often, an easy way to show that a function is bijective is to find an inverse for it, as in the following example.

Example 6.3. Consider the map $f : [0, 1] \to [7, 11]$ given by $f(x) = 4x + 7$ for $x \in [0, 1]$. Show that f is bijective.

It suffices to note that if $y = 4x+7$, then we can uniquely solve for x, and get $x = (y-7)/4$. Thus, if we define the map $g : [7, 11] \to [0, 1]$ by $g(y) = (y - 7)/4$, then $g(f(x)) = x$ for all $x \in [0, 1]$ and $f(g(y)) = y$ for all $y \in [7, 11]$.

Proposition 6.4. *(a) If a map $f : X \to Y$ admits an inverse map g, then the inverse map g is unique, and we will denote it by f^{-1}.*

(b) In this case, the map $f^{-1} : Y \to X$ is invertible also, and its inverse is f, i.e $f^{-1^{-1}} = f$.

Proof. (a) Suppose $g, g' : Y \to X$ are both maps from Y to X that are inverses to f. Thus, $g(f(x)) = g'(f(x)) = x$ for all $x \in X$ and $f(g(y)) = f(g'(y)) = y$ for all $y \in Y$. Since f admits an inverse, by the above Theorem, f is bijective. What is our task? We must show that the functions $g : Y \to X$ and $g' : Y \to X$ are really the same function. That means we have to show for each $y \in Y$, $g(y) = g'(y)$. Well, given, $y \in Y$, since f is surjective, we can find $x \in X$ such that $f(x) = y$. Then, $g(y) = g(f(x)) = x = g'(f(x)) = g'(y)$. This completes the proof that $g = g'$.

(b) Now suppose $g = f^{-1}$. By the definition of inverse function, we have that $g(f(x)) = x$ for all $x \in X$ and $f(g(y)) = y$ for all $y \in Y$, hence f is an inverse for g. By the uniqueness of the inverse, we must have $g^{-1} = f$ or $f^{-1^{-1}}$, which is what we wished to prove. □

Finally, we recall here a definition from the previous homework.

Definition 6.5. Suppose $X_1, \ldots, X_n$ are sets. The *direct product* (or *Cartesian product*) of $X_1, \ldots, X_n$, denoted by $X_1 \times X_2 \times \cdots \times X_n$, is the set of all ordered n-tuples $(x_1, x_2, \cdots, x_n)$ where $x_i \in X_i$ for $i = 1, 2, \ldots, n$.

For examples, $\mathbb{R} \times \mathbb{R}$ is the set of all ordered pairs (x, y) where $x, y \in \mathbb{R}$. In other words, $\mathbb{R} \times \mathbb{R}$ can be thought of as the set of coordinates of points in the usual xy plane, and $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ can be thought of as the set of all points in three dimensional space.

Part 4. Counting Principles and Finite Sets

7. Three Counting Principles

7.1. **The Well-Ordering Principle.** The notion that the integers are ordered according to their "size" brings up, at some early age, the question of the "largest" and "smallest" numbers. Surprisingly early, children can decide on their own that there is no largest integer. This is a fundamental and first-rate mathematical theorem. Once they become familiar with negative numbers, children also accept fairly quickly that there is no smallest integer (by symmetry!).

However, if one restricts attention to positive integers $\mathbb{N}$ or non-negative integers $\mathbb{Z}_{\geq 0}$, then there is of course a smallest integer, namely 1 (respectively 0). The same is true of any subset of $\mathbb{N}$ or of $\mathbb{Z}_{\geq 0}$. This simple but extremely handy observation has a fancy name since it is often invoked as a useful and fundamental fact of arithmetic. But first, let's have a definition.

Definition 7.1. Suppose S is a subset of $\mathbb{R}$. We say that $l \in S$ is a *least element* of S (sometimes *minimal element*) if for all $s \in S$, we have $l \leq s$. We say that $g \in S$ is a *greatest element* of S (or *maximal*) if for all $s \in S$, we have $s \leq g$.

The Well-Ordering Principle Given any **non-empty** subset $S \subseteq \mathbb{Z}_{\geq 0}$ of the set of positive integers, there exists a least element (or *minimal element*) of S.

WARNING. In the above statement, if you forget to say that S is non-empty, then the statement becomes false! The empty set has no least element, because it has no elements whatsoever.

What is responsible for the Well-ordering Principle is the fact that the integers are a "discrete set." Its elements don't get bunched up together, so to speak; they maintain a respectful distance from each other. This is a "topological" property of the integers as a subset of the real numbers.

Non-Example 7.2. For instance, the open interval $I = (0, 1) = \{x \in \mathbb{R} | 0 < x < 1\}$ has no least element and no greatest element. A proof of this (using the the "proof by contradiction" method) was given earlier in these series of notes. Perhaps you dimly remember skipping that. Well, now go back and read it, or better yet, try to give a proof for yourself.

Example 7.3. Let S be the set of positive integers that when added to themselves are divisible by 12, i.e.

$$S = \{x \in \mathbb{Z} | x > 0 \text{ and } x + x = 12y \text{ for some } y \in \mathbb{Z}\}.$$

It's not too hard to see that the least element of S is 6.

7.2. **The Pigeon-Hole Principle.** Let's say you have a bunch of pigeons and a bunch of pigeon-holes. If all the pigeons must be placed in the pigeon-holes, and you have **more pigeons than pigeon-holes**, then at least two pigeons will have to share a pigeon-hole. That's the Pigeon-hole Principle. In Germany, they call this the Shoebox principle[13] (with shoes playing the role of pigeons and boxes ... you get the idea). Here is a game you should play: pick your own favorite imagery for objects: pigeons, shoes, ducks, pencils, ... as well

[13]or the much more amusing *Schubfachschluss-Prinzip; on this topic, you should read Act I of Oscar Wilde's* The Importance of Being Earnest

as places to put them: holes, boxes, slots, ... to get your own principle, such as The Pencil-box principle or the Ducks-in-row principle. Then translate your phrase into some abstruse language to get a really cool Principle of Your Very Own.

You might think this is a colossally silly thing to make into a "principle" but you would be wrong; it is very useful and important. Sometimes a slightly fancier version is handy.

The Pigeon-hole Principle Suppose $k \geq 1$ is a positive integer. If N pigeons are to be placed in H pigeon-holes and $N > kH$, then there will be at least one pigeon-hole housing more than k pigeons.

It might sound fancy, but it's very simple. Let's use the "proof by contradiction" way of looking at it. If there are no pigeon-holes housing more than k pigeons, then every pigeon-hole houses at most k pigeons, so then the number of pigeons is at most kH, but $kH < N$. So,

$$N = \text{ the number of pigeons } < N.$$

This is rather embarrassing, so there must be at least one pigeon-hole housing more than k pigeons.

7.3. **The multiplication counting principle.** Let's say your breakfast every weekday morning consists of one bagel and one cup of coffee at the Marcus Café. There are the following choices for the bagel: plain, poppy seed, sesame, onion, cinnamon raisin, salt, and everything, nine choices in all. Your coffee choices are small regular, small decaf, large regular, and large decaf. Now how many different possible breakfasts can you have? You could list them all in a systematic way, for example, by saying: there are four breakfasts where the bagel is plain (indexed by the four different coffee choices), there are four breakfasts where the bagel is poppy seed, etc. In fact, there are four breakfasts for each type of bagel, and there are 9 types of bagels, so the total number of breakfasts is $4+4+4+4+4+4+4+4+4 = 36$. Of course, you could also say it's $9 \cdot 4 = 36$. Or if you say there are 9 breakfasts where the cofee is regular, etc. you'd get $4 \cdot 9 = 36$ breakfasts. Now then you say, aha, but I can get the bagel toasted or not, so now how many breakfasts are there? Well, since toasted/not toasted is two choices, it just doubles the number of breakfasts, to 72. Oh, I forgot, I can also choose No topping, Cream Cheese topping, or butter that's three choices, tripling the number of breakfasts to 216. Oh I forgot that for the coffee I might choose sugar or not sugar, doubling the number of choices again to 432. Then I might opt for Regular Milk, Lowfat Milk, Skim Milk, Cream, Half/Half, or No-Gosh-Darn-Cold-Liquid-In-my-Gosh-Darn-Coffee-Thank-You-Very-Much, sextupling the number of choices to a mere 2592 possible breakfasts. No wonder I never vary my breakfast![14]

We have just seen an example of the "Multiplication Counting Principle:" The total number of possible outcomes of a compound event is the **product** of the number of possible outcomes for each of the constituent events.

Again, this is a simple and extremely nimble principle. Why is it true? Well, say you have two finite sets X_1 and X_2 and you take their Cartesian product $X_1 \times X_2$. Recall that this is the set of all ordered pairs of type (x_1, x_2) where $x_1 \in X_1$ and $x_2 \in X_2$. There are $|X_2|$ elements of the Cartesian product whose first coordinate is x_1, an equal number whose first coordinate is x_2, etc. giving us a total of $|X_2|$ added to itself $|X_1|$ times which is just

[14]Poppy-Seed Bagel with no topping, untoasted, unsugared coffee and No-Gosh-Darn-Cold-Liquid-In-my-Gosh-Darn-Coffee-Thank-You-Very-Much.

$|X_1| \cdot |X_2|$. In other words, $|X_1 \times X_2| = |X_1| \times |X_2|$. It takes little more effort to prove the more general rule that if $X_1, X_2, \ldots, X_n$ are $n \geq 2$ finite sets, then

$$|X_1 \times X_2 \times \cdots \times X_n| = |X_1| \times \cdots \times |X_n|.$$

For instance, in a short while, we'll give a proof of this using Mathematical Induction.

8. Finite sets

We have been studying sets and maps between sets. Suppose two sets X and Y meet at a party and want to find out about each other. What would be the first question one set would ask another, do you think? Perhaps the most fundamental question is "Are you the empty set?" There is a profound difference, after all, between the empty set and all other sets. If the set you have met is empty, then you know exactly who she is. But if the set you have met is non-empty, then what question would you ask next? I think a perfect sensible question is "How many elements do you have?"

In particular, to "size" up the set you have just met, what you are really curious about is whether she is infinite or finite. What does this mean? Oh Oh, I think another one-act play is about to start.

(In)finite sets Starring: Farshid as Professor Dude, and Introducing: Jeremiah, Ashley, Shaohan, Nicolai, Jeremy, Eric, Kwylan, Ben, Bert, Matthew, Amy, Emily and Alby.

Farshid: Class, what does it mean for a set X to be infinite?

Jeremiah: An infinite set is one that has "infinitely many elements."

Farshid: Sorry I don't understand, what does it mean to have infinitely many elements?

Ashley: Don't play stupid, you know what that means.

Farshid: No, really, I want a rigorous, solid, definition of what it means for a set to be infinite. So far, it has been put forward that a set is infinite if and only if it has infinitely many elements. That is possibly a good definition, but only if someone can then define rigorously what it means for a set to have infintely many elements. For instance, for the set $\{1, 2, 3\}$, here are infinitely many elements: 1,1,1,1,1,1,1,1,1,.... but $\{1, 2, 3\}$ is clearly not what you would call an infinite set. The point is: You have to give a rigorous definition of what it means to give infinitely many elements of the set.

I submit that when it comes to infinity, nothing is "obvious" or to be taken for granted. In this course, one thing you will hopefully learn is that Infinity is a big tough guy on the block; you don't mess around with it! So, I must insist, please define: either a) what it means for a set to be infinite, and/or b) what it means for a set to have infinitely many elements.

Shaohan: I cannot answer your question immediately, but I would like to say that infinite is the opposite of finite.

Nicolai: That's a good point. Dude, if you can define what it means for a set to be finite, then I'll give you a definition of what it means to be infinite, namely a set is called infinite if it is not finite!

Farshid: That is perfectly acceptable, in theory, assuming you can give a good definition of what it means for a set to be finite.

Julie: Well, a set is finite if you can count how many elements it has! That's easy.

Jeremy: I think Professor Hajir is looking for something more formal than that.

Julie: Who's Professor Hajir? Oh, you mean that Farshid guy? Okay, no problem. A set is finite if you can list all its elements in a finite amount of time.

Farshid: You are on to something, Julie. However, do you think that this is really a workable definition? Will you ever be able to prove that a set is NOT finite using that definition? Let me suggest that you should try to take "time" out of your definition. You said a minute ago that a set is finite if you can count how many elements it has. Imagine you are 4 years old and I give you a bag of apples. How would you count the number of apples in the bag?

Eric: Dude, like, I'll handle that one. You take one apple out and you go "1," you take out another apple and you go "2," and so on, and so forth, until you have taken out all the apples. If the last number that you said is, "n", then the number of apples is "n."

Bert: Wait, you know what you are doing when you're doing that? You are setting up a one-to-one correspondence between the bag of apples and the set $\{1, 2, ..., n-1, n\}$. Each apple gets labelled with exactly one integer between 1 and n, so the map you are creating is a bijection.

Eric: Well, I never thought of it *that* way, but I agree with what you say.

Amy: Okay, how about this: A set X is finite if there exists an integer $n \geq 0$ such that the set X is equivalent to the set $\{1, 2, 3, \ldots, n\}$, i.e. there is a bijection $f : X \to \{1, 2, \ldots, n\}$.

Matthew: Why $n \geq 0$? Why not $n \geq 1$?

Kwylan: Don't forget the empty set dudes!

Matthew: Gotcha. Hey, you know the "n" in Amy's definition is how many elements the set has, we should call it the size of the set.

Farshid: Good point! This is great, but I still want to know what it means to be an infinite set!

Emily: Um, like, haven't you been paying attention? We already defined what it means for a set to be finite, and now we say that a set is infinite it is NOT finite.

Farshid: You're right, I'm just trying to catch up here. Let me see if I can summarize for myself what conclusions you have reached. First of all, your strategy for checking whether a given set is finite or not (and if it is finite to know its size), is to compare it to a standard family of "counting sets." To facilitate the summary, let's give a name to these special counting sets: for each integer $n \geq 0$, let

$$Z_n = \{x \in \mathbb{Z} | 1 \leq x \leq n\}.$$

Bert: Hey, there's that $n = 0$ again... I still don't like that. Your definition doesn't make sense if $n = 0$ because you get

$$Z_0 = \{x \in \mathbb{Z} | 1 \leq x \leq 0\}!!!!$$

That's not well-defined because there are no such x!!!

Ben: I see what you don't like: $1 \leq x \leq 0$ is stupid.

Shaohan: It is not so much stupid as false. In other words, since $1 \leq x \leq 0$ simply does not hold for any x, it is a "contradiction" i.e. a false statement.

Farshid: Let's look at Z_0. You are abosolutely right, Bert, when you say that Z_0 consists of all integers x such that x is bigger than 1 and less than 0. Of course, there are no such integers! But that doesn't mean that we have to flush everything down the toilet and say it's not well-defined. For a set to be well-defined, the bouncer should be completely sure about who to let in and who to bounce out. For the set Z_0, the instructions to the bouncer are: there are no x that pass the test for getting into Z_0, so bounce the heck out of every single x that comes along. This is not an ambiguous rule, it's pefectly well-defined, it's perhaps the

most-well-defined rule for any bouncer! It's just that Z_0 is the set with no elments, i.e. Z_0 is the empty set, that's all.

Bert: Oh, okay, sorry man, that was bothering me. I see now.

Farshid: No problem. Now, $Z_0 = \{\}$ is the empty set. It has size 0; $Z_1 = \{1\}$, it has size 1; $Z_2 = \{1, 2\}$, it has size 2, and so on. For an integer $n \geq 1$, $Z_n = Z_{n-1} \cup \{n\} = \{1, 2, 3, \ldots, n\}$ and $|Z_n| = n$. Finally, here is the summary of our discussion.

GIVEN A SET X, IF THERE EXISTS A NON-NEGATIVE INTEGER n SUCH THAT X IS EQUIVALENT TO THE SET $\{1, 2, 3, \ldots, n\}$, THEN THE SET X IS *finite* AND WE SAY THAT ITS **size** IS n. IF THERE IS NO SUCH n, THEN WE SAY THAT X IS INFINITE.

Alby (sotto voce): Leave it to this guy to take something so simple as counting and making it sound complicated.

Farshid: What was that?

Alby: Oh nothing, I was just saying how much I like this class.

Farshid: Well, I'll be damned.

Definition 8.1. A set is called *finite* if it has a "limited" number of distinct elements. To be more formal, a set X is finite if there exists a non-negative integer n such that for all sequences $x_0, x_1, x_2, \ldots, x_n$ of elements of X, there exist integers $0 \leq i < j \leq n$ such that $x_i = x_j$. In other words, a set is finite if for some $n \geq 0$, every list of $n + 1$ elements of the set must have a repetition. If X is finite, then we can define the order of X, also called its cardinality and denoted by $|X|$ as follows: $|X|$ is the **least** non-negative integer n with the property that for all sequences $x_0, x_1, x_2, \ldots, x_n$ of elements of X, there exist integers $0 \leq i < j \leq n$ such that $x_i = x_j$. Why does such an integer n exist? By the assumption that X is finite, we know that there exists at least one such non-negative integer n; now by the Well-Ordering Principle there is a unique smallest such integer $n \geq 0$ and this n is the cardinality or size of the set.

In less formal language, the maximal number of distinct elements in X is called the size of X or the cardinality of X and is denoted by $|X|$ or sometimes $\#X$.

A set which is not finite is called *infinite*. Equivalently, a set X is infinite if and only if for every positive integer n, there exists a sequence of n (pairwise) distinct elements of X. In this case, we write $|X| = \infty$. As we will see later, this statement is slightly misleading, as there are in fact multiple *gradations* of infinity.[15]

For example, let us look at the finite set $X = \{a, c, b, c, c, b\}$; for this set, we have the sequence a, c, b with no repetitions, of length 3, but in any sequence of length 4 or more coming from this set, one of these letters at least must be repeated, so $|X| = 3$. On the other hand, the set $I = (0, 1)$ is infinite because, for instance, given any positive integer n, the sequence of length n $1/2, 2/3, 3/4 \ldots, (n-1)/n, n/(n+1)$ consists of n pairwise distinct elements of I. Thus, there exist arbitrarily long sequences of pairwise distinct elements in I so I must be infinite.

Of course, in real life, this is not how we count how many things are in a set. Again, imagine interviewing a child: Put a bunch of dots on a page, say two or three times the kid's age and ask her to count how many dots there are. The child will probably point to each dot and count in sequence, 1,2,3, etc. An attempt will be made to make sure of two things: a) each dot is included in the count, and b) no dot is counted more than once. In

[15]We will see that the set of real numbers, for example, is "a lot more infinite" than the set of integers!

other words, she will try to "cover" each dot exactly once. What is the child's method? She attempts to set up a "one-to-one correspondence" between the dots on the page and a set of type $\{1, 2, 3, 4, \ldots, n-1, n\}$. Then the child will know that there are n dots. The important principle behind her method is that whenever two sets are in one-to-one correspondence, then they must have the same cardinality, or size.

Lemma 8.2. *Suppose $f : X \to Y$ is an injective map of a set X into some set Y.*

(a) For any integer $n \geq 2$, if X has n pairwise distinct elements, then Y has at least n pairwise distinct elements.

(b) If Y is finite, then X is finite also and, moreover, $|X| \leq |Y|$.

Proof. First, we prove (a). Let $x_1, \ldots, x_n$ be n pairwise distinct elements of X. Since f is injective, and $x_i \neq x_j$ for $1 \leq i < j \leq n$, we have $f(x_i) \neq f(x_j)$ for the same i, j. Hence the elements $f(x_1), \ldots, f(x_n)$ are n pairwise distinct elements of Y. So, Y has at least n distinct elements.

Now to prove (b). We assume that Y is finite. Let $n = |Y|$. We claim that X cannot have $n + 1$ pairwise distinct elements. For if it did, then by part (a), Y would would have $n + 1$ pairwise distinct elements which would contradict $n = |Y|$. Since X cannot have $n + 1$ pairwise distinct elements, we must have $|X| \leq n$. In particular, X is finite and $|X| \leq |Y|$. □

Now we are ready to prove the theorem we have been after.

Given a set X, if there exists a non-negative integer n such that X is equivalent to the set $\{1, 2, 3, \ldots, n\}$, then the set X is *finite* and we say that its **size** is n. If there is no such n, then we say that X is infinite. Got it?

Theorem 8.3. *If X and Y are equivalent sets, and if one of them is finite, then the other is finite also and $|X| = |Y|$.*

Proof. Suppose Y is finite. Since X and Y are equivalent sets, there exists a bijective map $f : X \to Y$. Since f is injective, by Lemma 8.2, X is also finite and $|X| \leq |Y|$. Now we have the map $f^{-1} : Y \to X$ which is also bijective (why?) so again by the same Lemma, $|Y| \leq |X|$. Since $|X| \leq |Y|$ and $|Y| \leq |X|$, we have $|X| = |Y|$. If we start with the assumption that X is finite instead of Y, then we repeat the argument, starting with a bijection $g : Y \to X$. □

Part 5. (Equivalence) Relations and Partitions

9. Relations

Recall the concept of a function f from a source set X to a target set Y. It is a rule for mapping each element x of the source to a single, well-defined, element of the target, which we call $f(x)$. A function from X to Y gives a very neat relationship between these two sets. Not all relationships between two sets are so "neat," and we will now consider a more general notion, that of a *relation* between X and Y.

Definition 9.1. Suppose X and Y are sets and $R \subseteq X \times Y$ is an arbitrary subset of the Cartesian product of X and Y. We say that R determines a *relation* from X to Y in the following way: If $x \in X$ and $y \in Y$ we say that x is related to y (and write $x \sim_R y$) if $(x, y) \in R$, and we say that x is not related to y ($x \not\sim_R y$) if $(x, y) \notin R$. If R is a relation from X to Y, and $x \in X$, we say that the fiber above x is the set $R_{x,\bullet} = \{y \in Y \mid (x, y) \in R\}$. Similarly, for $y \in Y$ $R_{\bullet,y} = \{x \in X \mid (x, y) \in R\}$ is the fiber below y. The Graph of the relation R is simply the set R itself. If X and Y are subsets of $\mathbb{R}$ with X, Y lying along the x-axis, y-axis as usual, then $R_{x,\bullet}$ is simply the intersection of the vertical line passing through x with the graph of R and $R_{\bullet,y}$ is the intersection of the horizontal line through y with this graph. The relation R determines a function $X \to Y$ determines a function if and only if for each $x \in X$, the fiber above it, $R_{x,\bullet}$ is a singleton, i.e. contains a single element, which is then the value of the function at x. Thus, a relation can be thought of as a function when its graph passes the "vertical line test."

Here is an informal summary of the above formal definition. First, informally speaking, a relation R between X and Y is a rule which, given $x \in X$ and $y \in Y$ determines whether x is "related" to y or not. If x is related to y, we write $x \sim_R y$ and otherwise we write $x \not\sim_R y$. The graph of R is defined by $\{(x, y) \in X \times Y \mid x \sim_R y\}$. If for each $x \in X$, there is a unique $y \in Y$ such that $x \sim_R y$, then R is a special kind of relation called a function; in that case, we write y is the value of this function at x if $x \sim_R y$.

Let us consider some examples.

Example 9.2. Let $X = Y = \mathbb{R}$, and let

$$R = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

In other words, x is related to y if and only if (x, y) is a point on the unit circle. The graph of the relation is just the unit circle. For instance, $1 \sim_R 0$ because $1^2 + 0^2 = 1$ and $0.5 \not\sim_R 0.5$ because $0.5^2 + 0.5^2 = 0.5 \neq 1$. Now let's talk about fibers. We have $R_{0,\bullet} = \{1, -1\}$ because $0^2 + y^2 = 1$ has two solutions $y = \pm 1$. On the other hand, the fiber below 1 has only one element, because $R_{\bullet,1} = \{x \in X \mid x^2 + 1 = 1\} = \{0\}$. By looking at the picture, one can see immediately that the fiber above x has $0, 1$ or 2 elements depending on whether $|x|$ is greater than 1, equal to 1, or less than one, respectively. In particular, this relation is *not* a function.

Example 9.3. Let $X = [0, 2]$ and $Y = [0, 12]$. Let $R = \{(x, y) \in \mathbb{R}^2 \mid y - 3x^2 = 0\}$. The graph of this relation is a piece of the parabola $y = 3x^2$. Since there is exactly one y-value in Y for each x value in X, this relation actually determines a function $X \to Y$, namely $\mathcal{R}(x) = 3x^2$.

Example 9.4. A relation can also be determined by just listing which elements of X are related to which of Y. For example, let $X = \{a, b, c, d\}$ and $Y = \{0, 1\}$. Then, we define a relation R by defining the related pairs to be are $a \sim 0$, $b \sim 1$, $b \sim 0$, $c \sim 1$. This means that these are ALL the related pairs. In other words, the subset R of $X \times Y$ is simply $R = \{(a, 0), (b, 1), (b, 0), (c, 1)\}$. This relation is not a function: here are two reasons; the fiber above b is too big (it is {0,1}) and the fiber above d is too little (it's empty!).

Example 9.5. Let $X = Y = [-1, 1]$ and define $x \sim_R y$ if and only if $xy \geq 0$. Thus, $R = \{(x, y) \in [0, 1] \times [0, 1] \mid xy \geq 0\}$. What does the graph of this relation look like? Think about it on your own for a minute. No Peekie until one minute's worth of thinking has occurred.

What did you get? The union of the unit squares in the first and third quadrants (including the axes), I hope. Verify that $R_{x,\bullet} = [0, 1]$ if $x > 0$ and that $R_{x,\bullet} = [-1, 0]$ if $x < 0$. What is $R_{0,\bullet}$?

In some of the examples above, the sets X and Y were the same, so the relation from X to Y can be called a self-relation on X; we call this a "relation on X" for short. Certain relations on X (those which satisfy three very nice properties we'll describe in a moment) are alled *equivalence relations*; they play a very important role in mathematics! We now define them.

Definition 9.6. A relation $\sim$ from a set X to itself is called an *equivalence relation* if it is **reflexive** (i.e. $x \sim x$ for all $x \in X$), **symmetric** (i.e. for all $x, y \in X$, $x \sim y \Rightarrow y \sim x$), and **transitive** (i.e. for all $x, y, z \in X$, $(x \sim y) \wedge (y \sim z) \Rightarrow x \sim z$). These three conditions can be stated as follows: for each $x \in X$, x is always in the fiber above itself (reflexive), for each $x, y \in X$, if x is in the fiber above y, then y is in the fiber above x (symmetric), and finally for $x, y, z \in X$, if y is in the fiber above x and z is in the fiber above y, then z is in the fiber above x (transitive).

Given our philosophy of the importance of non-examples, let us start with two relations which are not equivalence relations.

Non-Example 9.7. Let us look at the first example of a relation we gave, namely $X = Y = \mathbb{R}$ and $x \sim y$ if and only if $x^2 + y^2 = 1$. This relation fails two of the three tests needed for being an equivalence relation. First, it is not reflexive: for $x \in X$, we do not always have $x \sim x$. For starters, for an equivalence relation, there are no empty fibers (why?) but this relation has plenty of empty fibers (just consider numbers of absolute value greater than 1). In fact, there are just 2 real numbers x such that $x \sim x$! What are they? This relation also fails transitivity. (Give an example). This relation is, however, symmetric.

Example 9.8. Let $X = Y = \mathbb{Z}$ with the divisibility relation, i.e. $x|y$ if and only if the equation $y = dx$ has a unique solution $d \in \mathbb{Z}$. Note that this is an asymmetric relation: $x|y$ does not imply $y|x$. But $x|y$ and $y|z$ implies $x|z$ and $x|x$ of course. We see that this relation is reflexive and transitive, but not symmetric. It plays a very important role in the study of the algebraic properties of the set $\mathbb{Z}$.

Non-Example 9.9. Now consider $X = Y = [-1, 1]$ with $x \sim y$ if and only if $xy \geq 0$. We have xRx because $x^2 \geq 0$ for all $x \in X$. If $xy \geq 0$, then clearly $yx \geq 0$ also. But this relation fails the last property of transitivity: for example, $1 \sim 0$ and $0 \sim -1$ are true but $1 \sim -1$ is false!

Example 9.10. We can tweak the previous relation just a little bit and make it into an equivalence relation. Namely, if we take $X_0 = [-1,0) \cup (0,1]$, i.e. $X_0 = [-1,1] \setminus \{0\}$, and define $x \sim y$ if and only if $xy \geq 0$, then $\sim$ defines an equivalence relation on X_0. Or, if we keep $X = [-1,1]$ as before but now define $x \sim y$ to mean that either $xy > 0$ or $x = y = 0$. You should check that in either of these cases, we get an equivalence relation.

Many relations from a set X to itself that occur "naturally" are equivalence relations. When that happens, it is a signal to us that the property defining the relation is a useful way of understanding the set. For example, if you consider the set of students at some elementary school, and consider the relation $\sim_1$ of "being in the same grade," then YOU CAN EASILY CONVINCE YOURSELF [**This is a shorthand way of saying : It's very important for you to do this exercise!**] that this is an equivalence relation. What that means is that breaking up the students according to their grade is a meaningful and interesting way to "organize" the set of elementary school students. Another relation $\sim_2$ of "being in the same class" is also an equivalence relation. They are not necessarily the same relation (because a school might have so many students that it has three different Kindergarten classes, for example). Another relation might be the classify the students according to age.

Now let us consider the fibers of an equivalence relation R on X. For $x \in X$, since $R_{x,\bullet} = R_{\bullet,x}$, we write simply R_x for either of these sets. This may appear an innocent definition, but the reader is hereby alreted to the fact that the fibers of an equivalence relation play a tremendously important role in all of mathematics! Often we annotate an equivalence relation with just a $\sim$ instead of giving a name (such as R) to its graph, so it's useful to have some other kind of notation for these fibers.[16] Consequently, we make the following very important definition.

Definition 9.11. If $\sim$ is an equivalence relation on a set X with graph $R = \{(x,y) \in X \times X \mid x \sim y\}$, then for each $x \in X$ the *$\sim$-equivalence class* of x, or simply the *equivalence class of* x, is the fiber above x, namely

$$R_x = \mathbf{cl}(x) = \widetilde{x} = [x] = \{y \in X \mid x \sim y\}.$$

It is the set of all elements of X equivalent to x. The set of all equivalence classes, i.e. the set $\widetilde{X} = \{\widetilde{x} \mid x \in X\}$ is called "X modulo $\sim$," and sometimes denoted by $X/\sim$ and called a "quotient of X." Note that $\widetilde{X} \subseteq \mathcal{P}(X)$ since its elements are subsets of X. The (surjective) map $X \twoheadrightarrow \widetilde{X}$ defined by $x \mapsto \mathbf{cl}(x) = \widetilde{x}$ is called the "quotient map," or the "natural map" from X to $\widetilde{X}$.

Example 9.12. Consider the set $\mathbb{Z}$ of all integers. If $a \in \mathbb{Z}$ we say that the *parity* of a is *even* if $a = 2b$ for some $b \in \mathbb{Z}$ and that its parity is *odd* otherwise. Let us write $a \sim b$ whenever a and b have the same parity. In other words, $a \sim b$ means that a and b are both odd or both even. Another way to say this is that $a - b$ is even. Let us check that this defines an equivalence relation on $\mathbb{Z}$: first, $a - a$ is always $0 = 2 \cdot 0$ hence always even. If $a - b$ is even, then $b - a = -(a - b)$ is also even. If $a - b$ and $b - c$ are even, then $a - c$ is even, because $a - c = (a - b) + (b - c)$ is the sum of two evens. Thus, parity defines an equivalence relation on $\mathbb{Z}$ with two equivalence classes (the odds O and the evens E). The quotient of $\mathbb{Z}$ by this equivalence relation is the set $\{O, E\}$.

[16] In general, whenever a mathematical concept is given multiple names or multiple notations, this should serve as a clear indication to the student that a concept of great import has just been encountered

Part of the Fundamental Theorem of Equivalence Relations says the following: Suppose X is a set and $\sim$ is an equivalence relation on X; then $\widetilde{X}$ is a partition of X, i.e. $\sim$ *partitions* the set into non-overlapping non-empty subsets that cover the whole set. You might say, it polarizes the set into non-overlapping "clans" of equivalent elements. It might help you to think of equivalent elements (i.e. those which are related to each other by the given equivalence relation) as "relatives" and the set of all relatives of a given element x is the *clan* of x, a less technical term for "equivalence class" of x. The fact that "clan" and "class" both start with "cla" is linguistically useful; is that an accident?

To justify the remarks of the previous paragraph, let us make some important observations about $\mathbf{cl}(x) = \widetilde{x}$ where x is an arbitrary element of X (relative to a fixed equivalence relation $\sim$ on X).

- $\mathbf{cl}(x) = \widetilde{x}$ is never empty (because $x \in \mathbf{cl}(x)$!). "Every clan has at least one member."
- For the same reason, every element of X belongs to some clan. (If $x \in X$, then $x \in \mathbf{cl}(x)$).
- On the other hand, if x and z and two elements of X, then their equivalence classes $\mathbf{cl}(x)$ and $\mathbf{cl}(z)$ are either identical or disjoint (recall that two sets are called disjoint if their intersection is the empty set $\emptyset$). In other words, if $x, y \in X$, then either $\widetilde{x} = \widetilde{y}$ or $\widetilde{x} \cap \widetilde{y} = \emptyset$. The proof of this is left as an important homework exercise.

Now let us recall the definition of a partition of a set.

Definition 9.13. Suppose X is a set, A is another set (an "auxiliary" or "indexing" set) and $\Delta = \{X_\alpha | \alpha \in A\}$ is a collection of subsets of X. We say that $\{X_\alpha\}$ is a *partition of X with indexing set A* if (1) for all $\alpha \in A$, $X_\alpha \neq \emptyset$; (2) for all $\alpha, \beta \in A$, $X_\alpha \cap X_\beta = \emptyset$; (2) $\cup_{\alpha \in A} X_\alpha = X$. In other words, a partition of X is a collection of non-empty pairwise disjoint (non-overlapping) subsets of X whose collective union is X. the subests X_α are non-overlapping and together entirely cover X.

Remark. As you will show in one of your homework problems, a collection $\{X_\alpha | \alpha \in A\}$ of non-empty subsets of X is a partition of X with indexing set A if and only if for every $x \in X$ there exists a unique $\alpha \in A$ such that $x \in X_\alpha$.

Remark. If $\Delta = \{X_\alpha \mid \alpha \in A\}$ is a partition of X with indexing set A, then the map $A \to \Delta$ defined by $\alpha \mapsto X_\alpha$ is a one-to-one correspondence from A to Δ. Thus, the point of the indexing set A is just to have a convenient way to refer to the elements of Δ.

Example 9.14. let $X = \mathbb{Z}$, let $A = \{0, 1\}$, let X_0 bet the set of even integers, and let X_1 be the set of odd integers. Then $\{X_0, X_1\}$ is a partition of $\mathbb{Z}$ because every integer is either odd or even (and no integer is both odd and even). You may note that this partition is mandated by the parity equivalence relation we discussed earlier. If we impose the parity relation on the integers and then order all the integers to band together into the corresponding clans, we will have exactly two clans, the evens and the odds, i.e. X_0 and X_1. The partition $\Delta = \{X_0, X_1\}$ is the set of clans under the parity equivalence relation.

Thus, if X is a set and $\sim$ is an equivalence relation on X, then X breaks up (is partitioned into) non-overlapping spanning equivalence classes. The set of these equivalence classes, i.e. $\widetilde{X}$ or $X/\sim$ called "X modulo $\sim$," is also called *the partition of X associated to $\sim$*. The elements of X are, on the one hand, subsets of X, on the other hand, they should be thought of as the "clans" which together make up X. Thus, if $x \in X$ is a "point" in X, then $\mathbf{cl}(x) = \widetilde{x}$ does double duty: as $\mathbf{cl}(x) \subset X$ it is a subset of X, and as $\widetilde{x}$, it is "a point" in the set $\widetilde{X}$.

For example, for $\mathbb{Z}$ under the parity equivalance, the set $\mathbb{Z}/\sim$ is the set $\{X_0, X_1\}$ consisting of two elements. Note that the elements of $\mathbb{Z}/\sim$ are themselves sets: X_0 is the set of even integers and X_1 is the set of odd integers, but as elements of the set $\mathbb{Z}/\sim$, we just think of them as "the even equivalence class" and the "the odd equivalence class." One psychological technique is to say that the equivalence relation gloms all odds together into one object and all the evens together into another object (it "forgets" or erases the distinguishing features of the integers and remembers only their parity). Any single member of an equivalence class is then a "representative" of that class, just as any member of a clan is a representative of his or her clan.

The main fact that one should understand about partitions also happens to be the main fact one should understand about equivalence relations, namely that **To specify a partition of X is "the same as" specifiying an equivalence relation on X and vice versa.** The process of going back and forth between the two concepts, the first half of which we have already outlined above, is as follows.

From an equivalence relation to a partition: If $(X, \sim)$ is a set together with an equivalence relation $\sim$ on it, then the set $\widetilde{X} = \{\mathbf{cl}(x) \mid x \in X\}$ of the clans of X, also called $X/\sim$ read "X modulo $\sim$", is a partition of X.

From a partition to an equivalence relation: On the other hand, if (X, Δ) is a set together with a partition of it, then this partition induces an equivalence relation on X via the rule $x \sim y$ if and only if there exists $S \in \Delta$ such that $x \in S$ and $y \in S$, i.e. if and only if x and y belong to the same piece of the partition.

Note that the equivalence relation we have attached to Δ is the unique equivalence relation under which the partition formed by the equivalence classes is just the partition we started with. Likewise, the reader should check that if you start with $(X, \sim)$ then pass to (X, Δ) and then attach an equivalence relation to Δ, then you get back the original $\sim$.

Theorem 9.15 (The Fundamental Theorem of Equivalence Relations)**.** *Suppose X is a set.*

(a) If $\sim$ is an equivalence relation on X, then the set of $\sim$-equivalence classes, $\widetilde{X} = X/\sim$, is a partition of X.

(b) If Δ is a partition of X, then Δ induces an equivalence relation $\sim_\Delta$ by the rule $x \sim y$ if and only if $x \in S$ and $y \in S$ for some $S \in \Delta$.

(c) If Δ is a partition of X, then $X/\sim_\Delta = \Delta$, and if $\Delta = \widetilde{X}$, then $\sim_\Delta = \sim$.

The point of the above theorem is that equivalence relations and partitions are two ways of looking at the same thing. Sometimes it is more convenient to use the partition language, and other times it is more useful to think in terms of relations. It is a frequent theme in a mathematician's experience that two objects that had been under separate study are revealed to offer different perspectives on the same underlying idea. Whenever this happens, the cohabitation, by the two seemingly different concepts, of the same "idea-landscape," serves to illuminate both concepts and to elevate the latter to a higher category of importance in the

consciousness of the mathematician. On this issue, consult the wonderful book[17] by Barry Mazur, one of the most distinguished, not to mention eloquent, mathematicians of our times.

10. The set of rational numbers

Earlier we introduced the set of rational numbers in a practical way by defining:

$$\mathbb{Q} = \{\frac{m}{n} \mid m \in \mathbb{Z}, n \in \mathbb{Z}^+, \gcd(m,n) = 1\}.$$

The problem, if you want to call it that, with this definition is that as written, $\frac{2}{-4}$ is not a rational number. We have to make the added stipulation that $\frac{2}{-4} = \frac{-1}{2}$ by putting the fraction in reduced form. The problem here is to be able to "forget" the fact that the fraction $\frac{2}{-4}$ determine by the ordered pair of numbers 2 and -4 looks different from the fraction $-1/2$. Our machinery of equivalence classes gives a neat solution to this little dilemma. Namely, to construct the set of rational numbers, let $\mathbb{Z}_0 = \mathbb{Z} \setminus \{0\}$ be the set of non-zero integers. We start with the set $X = \mathbb{Z} \times \mathbb{Z}_0$. On this set, we define the equivalence relation $(a,b) \equiv (c,d)$ if and only if $ad - bc = 0$.[18] [Don't take my word for it: you must actually check that this really is an equivalence relation!] Then we define the set $\mathbb{Q}$ of rational numbers to be $\mathbb{Q} = X/\sim$! In other words, a rational number is really an equivalence class of (infinitely) many ordered pairs of integers (the second of which is non-zero). For instance,

$$\{(n,-2n) \mid n \in \mathbb{Z}_0\} = \{\ldots,(3,-6),(2,-4),(1,-2),(-1,2),(-2,4),(-3,6),\ldots\}$$

is a rational number. You might object that this is a horrendously cumbersome way of "dealing" with what is after all a pretty basic object, and you would be right. But you cannot argue witht the fact that our construction is very rigorous and "correct" somehow. If we are ever in doubt about the truth of some subtle point about rational numbers, we can fall back on this very rigorous understanding. On some level, our mind keeps track of the fact that rational numbers can be represented in infinitely many different ways. For instance, in wishing to add $1/2$ to $1/3$, we prefer to think of these as $3/6$ and $2/6$ respectively. What allowed us to give a definition of $\mathbb{Q}$ without all this fuss was that we are able to give in each equivalence class in X a well-chosen representative, namely the one with positive second coordinate whose first coordinate is least in absolute value. This is a different way of saying (a,b) where $b > 0$ and $\gcd(a,b) = 1$. By the way, we have a natural injection $\mathbb{Z} \hookrightarrow \mathbb{Q}$ given by $a \mapsto \mathbf{cl}(a,1)$, which you should think of as $a \mapsto a/1$.

Now we will express our previous proof that $\mathbb{Q}$ is a countable set in a slightly different way. Recall that a set S is countable if its elements can be listed, with or without repetition, as $s_1, s_2, \ldots$. In other words, S is countable if and only if there exists a surjection $\mathbb{N} \twoheadrightarrow S$. Thus, if S is a countable set and $\sim$ is an equivalence relation on S, then $\widetilde{S}$ is also countable. (see the homework exercises). Since $\mathbb{N}$ and $\mathbb{Z}_0$ are countable, and since the direct product of two countable sets is countable, we have $\mathbb{N} \times \mathbb{Z}_0$ is countable, hence so is $\mathbb{Q} = \widetilde{\mathbb{N} \times \mathbb{Z}_0}$ where this is the quotient under the equivalence relation $(a,b) \sim (c,d) \Leftrightarrow ad - bc = 0$.

[17] Barry Mazur, Imagining Numbers (especially the square root of minus fifteen) FSG 2002.

[18] Where else have you seend that expression $ad - bc$ before?!

Part 6. Induction

11. Remembrances of Things Past

Let's take a moment to recall some material which will hopefully be familiar to you from your study of infinite series and matrices, etc.

A finite sum $a_1 + a_2 + \cdots + a_n$ can be expressed compactly as $\sum_{j=1}^{n} a_j$. Note that it could also be written as $\sum_{umass=1}^{n} a_{umass}$. If $a_1, a_2, \ldots$ is an infinite sequence of real numbers, then for the infinite sum $a_1 + a_2 + \cdots$ we write

$$\sum_{j=1}^{\infty} a_j = \lim_{n\to\infty} \sum_{j=1}^{n} a_j,$$

i.e. the sum of the series is defined to be the limit of the partial sums, should this limit exist. If it does not, we say that the series diverges.

A product of terms $a_1 a_2 \cdots a_n$ is written as $\prod_{j=1}^{n} a_j$.

You have probably encountered the factorial notation (when you studied the Taylor series of e^x for example): for an integer $n \geq 0$, we define $n! = \prod_{j=1}^{n} j$. Note that $0! = 1$ because by definition, 0! is an empty product: an empty product should be interpreted as 1 always, just as an empty sum should be interpreted as 0. A very useful fact is that $n! \approx e^{-n} n^{n-0.5} \sqrt{2\Pi}$; this rather good approximation (for large n) is known as *Stirling's formula*.

If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ are two matrices with real coefficients, then the product of the two matrices AB is defined to be the matrix

$$AB = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}.$$

Matrices represent linear transformations of a plane (equipped with a basis) and the matrix product defined above represents the composite of the two linear transformations corresponding to A and B, respectively under a fixed basis. The determinant of the matrix A is defined to be $ad - bc$; one can check directly that $\det(AB) = \det(A)\det(B)$, i.e. the determinant map is multiplicative. For a more conceptual explanation, you should have attended an Undergraduate Colloquium on March 24, 2005 [notes on my web page might pop up one day ...]

12. Mathematical Induction

One of the most powerful methods for proving certain kinds of propositions is mathematical induction. Keep in mind that inductive reasoning is quite distinct from mathematical induction (from now on I will just say "induction" for short). Induction is based on the well-ordering principle. There are several different "flavors" of induction; we will talk about three: ordinary induction, complete induction, and two-variable induction.

The basic idea is as follows. Imagine a typical guy. Let's call him Larry. Larry is a snacker. His favorite snack is Potato Chips. Larry has a peculiar habit. If he eats a chip from a bag, then he simply must eat another chip from that bag (as long as another chip exists in the bag). So, what can we conclude from this? That if you give Larry a bag of chips **and** he eats a first chip, then he will completely eat the whole bag no matter how many chips are in the bag. That's it, that's Mathematical Induction. Note the main "engine" of

the conclusion we reached about Larry is the "inductive step" that each consummed chip necessarily will entail the consumption of another chip. The other necessary hypothesis (for Larry to eat the whole bag) is of course is that he take that first plunge and actually eat the first chip (the "base case"). Here is another analogy. Imagine you set up a bunch of dominoes in your room and you space them so closely that you know for sure that should any domino ever topple forward, then its succeeding neighbor will also fall forward. Then, if any domino along the chain falls forward, **all** of the succeeding ones will as well.

Okay, now you are psychologically prepared for a more formal statement of the Principle of Mathematical Induction.

Theorem 12.1 (The (Ordinary) Principle of Mathematical Induction)**.** *Suppose for each integer $n \geq 1$, $P(n)$ is a statement. Assume moreover that*

- *$P(1)$ is true, and*
- *Whenever $P(k)$ happens to be true for some $k \geq 1$, then $P(k+1)$ is also true, i.e.*

$$P(k) \Rightarrow P(k+1) \text{ for all } k \geq 1.$$

Then $P(n)$ is true for all $n \geq 1$.

Proof. We will give a proof by contradiction. Suppose it is not the cae that $P(n)$ holds for all n, i.e. the set $S = \{n \geq 1 \mid P(n) \text{ is false } \}$ is non-empty. By the well-ordering principle, S has a least element, let us call it m. We know that $m > 1$ since $P(1)$ is true by assumption. Since m is the smallest element of S, we know that $m-1 \notin S$. By definition of S, therefore, $P(m-1)$ is true. Since $P(k) \Rightarrow P(k+1)$ for all $k \geq 1$, we conclude that $P(m)$ is true, but $m \in S$ so $P(m)$ is false. This is a contradiction. So S must be empty after all, i.e. $P(n)$ holds for all $n \geq 1$. □

Remark. Problem 5 on HW5 was just a different way of stating the above Principle.

We will now use mathematical induction to prove some formulas that we guessed using inductive reasoning but have not yet verified, as well as some other kinds of statements.

Example 12.2. For each integer $n \geq 1$, prove that $P(n) : 1 + 2 + \cdots + n = n(n+1)/2$. This is crying out for proof by induction.

We first check $P(1)$: $1 = 1(1+1)/2$ CHECK.

Now we check the "induction step" i.e. $P(k) \Rightarrow P(k+1)$ for $k \geq 1$. So, suppose $P(k)$ holds for some $k \geq 1$. Then $1 + 2 + \cdots + k = k(k+1)/2$. We have to show that this implies $P(k+1)$. Okay, let's try to calculate $1+2+\cdots+k+(k+1)$ using what we have been allowed to assume, namely that the sum of the first k integers has a nifty closed-form formula. Okay

$$\begin{aligned} 1 + 2 + \cdots + k + (k+1) &= [1 + 2 + \cdots + k] + (k+1) \\ &= k(k+1)/2 + (k+1) \\ &= (k+1)(1 + k/2) \\ &= (k+1)(k+2)/2. \end{aligned}$$

Hey, we just showed that, for $k \geq 1$, $P(k+1)$ follows if $P(k)$ is true. That completes the "induction step." So, by the principle of mathematical induction, $P(n)$ is true for all $n \geq 1$.

Example 12.3. Show that if we order the odd positive integers in increasing size (as usual), for $n \geq 1$, the nth one is $2n-1$. In other words, $P(n)$: if O_n is the nth odd integer, then $O_n = 2n-1$. Let's check the base case: $n = 1$: $O_1 = 1$ and $2(1)-1 = 1$ so it checks out okay.

Now let's do the inductive step. If, for some $k \geq 1$, $O_k = 2k-1$, then, since $O_{k+1} = O_k + 2$, we have $O_{k+1} = (2k-1)+2 = (2k+1)$. We might scratch our heads here a little and say "Um, like, are we done now?" or "What the heck are we trying to do here?!" The answers are "Not quite dude" and "Like you have to show $P(k) \Rightarrow P(k+1)$ dude," respectively. Let's just work backwards just a tiny bit: what is $P(k+1)$? It says that $O_{k+1} = 2(k+1)-1$. So far, we have $O_{k+1} = 2k+1$. Oh but wait, $2k+1 = 2k+2-1 = 2(k+1)-1$ so we got it.

Just to be totally clear let's now re-give the induction step. We suppose that $k \geq 1$ is some integer and that $P(n)$ happens to hold for $n = k$, then try to derive that $P(n)$ would follow for $n = k+1$. So, we assume $O_k = 2k-1$, then compute $O_{k+1} = 2k-1+2 = 2k+1 = 2k+2-1 = 2(k+1)-1$, so $P(k) \Rightarrow P(k+1)$.

Example 12.4. Prove that for $n \geq 1$, the sum of the first n odd positive integers is n^2. So, let us define T_n to be the sum of the first n odd integers, i.e. $T_n = 1+3+5+\cdots+(2n-1)$, since we figured out that $2n-1$ is the nth odd number. Let's check the base case. Always do that first. In fact, it pays to do just a few more cases after the base case. So, for $n = 1$, $T_n = T_1 = 1$ and $n^2 = 1^2 = 1$ so they agree. We also check $T_2 = 1+3 = 2^2$, $T_3 = 1+3+5 = 9$ and that's good too. Okay, let's move to the heart of the matter, then inductive step. We must show that $P(k)$ implies $P(k+1)$ for all $k \geq 1$. Thus, if $P(k)$ holds, then $T_k = k^2$ which implies that

$$T_{k+1} = T_k + (2k+1) = k^2 + 2k + 1 = (k+1)^2,$$

so $P(k+1)$ holds, just as easy as that. The equality $T_n = n^2$ has now been proved for all $n \geq 1$ by the principle of mathematical induction.

Example 12.5. Show that for $n \geq 5$, $2^n > n^2$. So we have $P(n) : 2^n > n^2$. You mean out "statement" is not an equality, it's an *in*equality? And we don't start with $n = 1$? Dude, can we still apply induction?! Of course we can, Dude! Base Case and Inductive Step that's what it's all about! Base Case $n = 5$: $2^5 = 32 > 25 = 5^2$ no sweat. Inductive Step: must show, for $k \geq 5$, that $2^k > k^2$ implies $2^{k+1} > (k+1)^2$. So suppose for some integer $k \geq 5$, $2^k > k^2$. Then $2^{k+1} = 2 \cdot 2^k > 2k^2$. Looking ahead (or working backwards) it sure would be nice if $2k^2$ would be gracious enough to dominate $(k+1)^2$ for $k \geq 5$. So, we boldly claim $2k^2 > (k+1)^2$ for $k \geq 5$ which solves the problem assuming we can prove our claim. Now, we write down $2k^2 \stackrel{?}{>} (k+1)^2$ on a Blue Wall napkin and expand and bring stuff from one side to the other until we see the following argument work itself out in reverse order of how we will now present it:[19] Since $k \geq 5$, we have $k > k-2 > 1$ so $k(k-2) > 1$ so $k^2 - 2k > 1$ so $k^2 > 2k+1$ so $2k^2 = k^2 + k^2 > k^2 + 2k + 1$ so $2k^2 > (k+1)^2$.

Example 12.6. Prove by induction that $n^3+(n+1)^3+(n+2)^3$ is a multiple of 9 for all $n \geq 1$. The base case: $n = 1$, $1+8+27 = 36 = 4*9$ CHECK. Now we must show $P(k) \Rightarrow P(k+1)$ for all $k \geq 1$. So we assume the Induction Hypothesis: $k^3 + (k+1)^3 + (k+2)^3 = 9t$ for some integer t. We must SHOW: $(k+1)^3 + (k+2)^3 + (k+3)^3 = 9s$ for some integer s. We recognize the first two terms of this sum as the last two terms of the sum in the induction

[19] One problem with reading mathematics is that the little scratchwork is never kept, even off in the margins, the proofs are all presented neat and crisp. I'm not saying this should change, but if that stuff isn't there in the text you are reading, then you need to be supplying it as you read!

hypothesis, so we calculate

$$\begin{aligned}(k+1)^3+(k+2)^3+(k+3)^3 &= 9t+(k+3)^3-k^3\\ &= 9t+k^3+9k^2+27k+27-k^3\\ &= 9t+9k^2+27k+27\\ &= 9s \qquad s=(t+k^2+3k+3)\in\mathbb{Z}.\end{aligned}$$

By the Principle of Mathematical Induction, we have proved that the sum of three consecutive integer cubes is divisible by 9.

Part 7. Number Theory

13. A little number theory

The set $\mathbb{Z}$ is really much more marvelous than you think. We have already discussed its first marvelous quality: its infinitude. What makes it even more marvelous are the two binary operations $+$ and $\times$. What is a binary operation, you ask? Good question.

Definition 13.1. Let X be a set. A binary operation on a set is a map $\mu : X \times X \to X$. Thus, an operation is a rule, which, given an ordered pair (x, x') of elements of x, produces an element $x'' = \mu(x, x')$ of X in a well-determined way. An alternative notation is often more convenient, namely if $\bullet$ stands for some kind of "operational" symbol, then instead of writing $\mu(x, x')$, we write more compactly $x \bullet x'$. An operation $\bullet$ on X is called *commutative* if $a \bullet b = b \bullet a$ for all $a, b \in X$. It is called *associative* if for all $a, b, c \in X$, $(a \bullet b) \bullet c = a \bullet (b \bullet c)$.

Non-Example 13.2. If $\mathbb{N}$ is the set of natural numbers, ordinary addition $(+)$ defines a commutative operation on $\mathbb{N}$. However, subtraction $(-)$ does not define an operation on $\mathbb{N}$ because for $a, b \in \mathbb{N}$, it is not always the case that $a - b \in \mathbb{N}$.

Example 13.3. The operations $+, \times, -$ on $\mathbb{Z}$ are familiar to you; addition and multiplication are associative and commutative, but subtraction is neither. Why did I leave out $\div$? Well, $\div$ does not actually define an operation on $\mathbb{Z}$ because given $a, b \in \mathbb{Z}$, it is not always that case that $a \div b$ is in $\mathbb{Z}$. Let us define an operation on $\mathbb{Z}$ as follows: given $a, b \in \mathbb{Z}$, we put $a \bullet b = |a^2 - b^2|$. Then $\bullet$ is a well-defined operation on $\mathbb{Z}$. It is clearly commutative. Is it associative?

Going back to what I started with, the set $\mathbb{Z}$ is really marvelous because it has two *compatible* operations $+, \times$ defined on it. What this means is that the two operations "respect" each other: namely, if $a, b, c \in \mathbb{Z}$, then $a \times (b + c) = (a \times b) + (a \times c)$. We say that $\times$ distributes over $+$. Moreover, these operations satisfy a whole host of other properties.[20]

Of the two basic operations $(+, \times)$ on $\mathbb{Z}$, the more subtle of the two is multiplication. How numbers are put together additively is not too mysterious: each integer n decomposes additively into a sum of n 1's: $n = 1 + 1 + \cdots + 1$. As we traverse the number line, this decomposition grows in a regular fashion, picking up one more "1" as it goes. But how numbers decompose *multiplicatively* is much less predictable as we traverse the number line.[21]This comment hopefully serves to explain a little the claim that multiplication is more subtle than addition.

The most subtle and interesting concept then, for the algebraic structure of $\mathbb{Z}$, is that of *divisibility*. Divisibility is a relation on $\mathbb{Z}$ which is transitive and reflexive but not symmetric; thus it is not an equivalence relation. Its importance is reflected in the multiplicity (excuse the pun) of names for this concept.

Definition 13.4. If n and d are integers, we write $d|n$ if and only if the equation $dx = n$ has a unique solution $x \in \mathbb{Z}$. The following phrases are all equivalent:

- $d|n$,
- d divides n,

[20]As your study of algebra continues these properties will collectively come to be known as "ring properties"; by the way, "algebra" is the study of sets equipped with certain kinds of operations.

[21]For instance, 17 is indecomposable, $18 = 2 \cdot 3 \cdot 3$, 19 is indecomposable,$20 = 2 \cdot 2 \cdot 5$, $21 = 3 \cdot 7$ etc.

- n is a multiple of d,
- n is is divisible by d,
- d is a factor of n,
- d is a divisor of n,
- there exists a unique $x \in \mathbb{Z}$ such that $n = dx$ (in shorthand, we write this as $n/d \in \mathbb{Z}$).

Example 13.5. For any integer $d \in \mathbb{Z} \setminus \{0\}$, we have $d|0$, as the equation $dx = 0$ has a unique solution $x = 0$. On the other hand, the statement $0|n$ is false for every $n \in \mathbb{Z}$, because the equation $0x = n$ has no solutions if $n \neq 0$ and infinitely many solutions if $n = 0$! In summary, 0 **doesn't divide anybody but** 0 **is divisible by everybody other than itself.**

Definition 13.6. For any integer $n \neq 0$, let $\mathrm{Div}(n) = \{d \in \mathbb{Z} \mid d|n\}$ be the set of divisors of n. We let $\mathrm{Div}^+(n) = \{d \in \mathbb{Z} \mid d > 0, d|n\}$ be the set of positive divisors of n and put $\sigma_0(n) = |\mathrm{Div}^+(n)|$.

Since $d \in \mathrm{Div}(n) \Rightarrow |d| \leq |n|$, $\mathrm{Div}(n)$ is a finite set, and in fact, we have the very crude bounds $|\mathrm{Div}(n)| \leq 2n$ and $|\mathrm{Div}^+(n)| \leq n$.

The set $\mathbb{Z} \setminus \{0\}$ is equipped with the involution "multiplication by -1". This involution reduces many issues having to do with multiplicative properties of integers to essentially the same question on the set $\mathbb{N}$ of positive integers. In other words, for every positive divisor of n there is exacly one negative divisor of n, so it suffices to work with $\mathrm{Div}^+(n)$ and this is often more convenient.

Now, suppose $n \in \mathbb{N}$. Every $d \in \mathrm{Div}^+(n)$, can be graphically represented by a $d \times e$ grid of $n = de$ dots arranged in d rows and e columns. We note that $|\mathrm{Div}^+(n)|$ is never empty since $1|n$ and $n|n$, corresponding to the $1 \times n$ and $n \times 1$ arrangements of n points. Now, for certain integers $n \geq 2$, no other rectangular arrangement is possible; these n are called *primes.*

Definition 13.7. A positive integer n is *prime* if $|\mathrm{Div}^+(n)| = 2$. In other words, n is prime if and only if it has exactly two positive divisors, namely 1 and n. An integer n is called *composite* if $|\mathrm{Div}^+(n)| \geq 3$. Thus, every integer > 1 is either prime or composite.

Non-Example 13.8. Note that 1 has but a single positive divisor hence 1 **is not a prime** according to our definition. It is not a composite either! It is clear that it plays a very special role in multiplication, since 1 divides every integer. A number which divides every element of $\mathbb{Z}$ is called *unit*. The only units in $\mathbb{Z}$ are ± 1. The number 1 is further distinguished by its role as the *identity* for multiplication, namely $1 \cdot a = a$ for all $a \in \mathbb{Z}$.

Example 13.9. The primes less than 50 are $2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47$.

The importance of primes for arithmetic is that every integer can be decomposed into a product of primes, and that, up to reordering of the factors, this "prime decomposition" is unique (this is a very important fact, known as The Fundamental Theorem of Arithmetic). Here is an analogy. In this sense, the primes are to arithmetic what "the elements" are to chemistry: we understand molecules in terms of the elements which constitute them. For now, let us show that every integer is a product of primes.

Theorem 13.10. *If $n > 1$ is an integer, then n is a product of primes.*

Proof. We will give a proof by complete mathematical induction. So, for $n \geq 2$, we define the statement $P(n)$ as follows.

$$P(n) : n \text{ is a "product of primes"},$$

which is shorthand for the following more precise statement: there exist primes $p_1, \cdots, p_r$ and positive integerrs $a_1, \ldots, a_r$ such that $n = p_1^{a_1} \ldots p_r^{a_r}$. Note that n is a "product of primes" thus encompasses the possibility that n is a prime itself. Let us check the validity of the base case, i.e. $P(2)$; 2 is a prime, so 2 is a product of primes. We will now proceed to the "induction step" of complete induction, so we have to establish

(*) Given an arbitrary $k \geq 2$, if $P(j)$ holds for $2 \leq j \leq k$, then $P(k+1)$ holds.

The statement (*) can be restated as $P(2) \wedge P(3) \wedge \cdots \wedge P(k) \Rightarrow P(k+1)$. So, we assume for $2 \leq m \leq k$, m is a product of primes and seek to show that $k+1$ is a product of primes. If $k+1$ is a prime itself, then it is a product of primes and we would be done. The other possibility is that $k+1$ is not a product of primes, i.e. $k+1$ is composite (since $k+1 > 1$). Thus, there exist integers $1 < d \leq e < k+1$ such that $de = k+1$. But by the induction hypothesis, since d, e are integers in the range $[2, k]$, each of them is a product of primes, hence so is $k+1 = de$. This establish (*).

We have thus established $P(n)$ for all $n \geq 2$ by complete induction on n. □

Second Proof of Theorem 13.10. Let us now give a proof by contradiction which relies on the Well-ordering principle. The strategy is to use the following lemma.

Lemma 13.11. *If an integer $m \geq 2$ is not a product of primes, then there exists an integer $1 < k < m$ such that k is not a product of primes.*

Proof. Since n is not a product of primes, n itself is not a prime. Since $n \geq 2$, $|\mathrm{Div}^+(n)| \geq 2$, and since n is not prime $|\mathrm{Div}^+(n)| \neq 2$, hence $|\mathrm{Div}^+(n)| > 2$. Therefore, there exists $d \in \mathrm{Div}^+(n) \backslash$ with $1 < d < n$, which implies that $n = de$ with $1 < e < n$ also. Since n is not a product of primes, at least one of d, e must not be a product of primes; since both d and e are in the range $[2, n-1]$, this proves the existence of an integer k, $1 < k < n$ such that k is not a product of primes. □

It should be clear how to prove the theorem using the Lemma. Suppose the theorem is false. Thus, there exists an integer $n \geq 2$ such that n is not a product of primes. Applying the lemma to this n, we get an integer $1 < k_1 < n$. Applying the lemma to k_1, now we get an integer $1 < k_2 < k_1 < n$. It is clear that if we repeat this procedure, we obtain infinitely many integers in the range $[2, n-1]$ which is a contradiction. To be even more precise, repeating this procedure $n-1$ times, we obtain $1 < k_{n-1} < k_{n-2} < \cdots < k_2 < k_1 < n$, i.e. we have $n-1$ distinct integers in the range $[2, n-1]$ which is impossible. This contradiction proves that every integer is a product of primes.

We can rephrase the endgame of this proof a little more efficiently by using the well-ordering principle. If we assume the theorem is false, i.e. that the set

$$\{n \geq 2 \mid n \text{ is not a product of primes}\}$$

is non-empty, then by the well-ordering principle, there exists a least element $m \geq 2$ which is not the product of primes. By the Lemma, there exists $1 < k < m$ such that k is not a product of primes, contradicting the minimality of m. □

The following theorem and proof, going back at least to Euclid, is a classic.

Theorem 13.12. *There are infinitely many primes.*

Proof. How do we show that a set X is infinite? One way to do so would be to show that if $F \subseteq X$ is any non-empty **finite** subset of X, then $X \setminus F$ is non-empty. So, let $\mathbf{P}$ be the set of all primes, and let F be a finite set of primes. Since F is finite, it has a maximal element, say ℓ. Now, define $N = 1 + \prod_{p \leq \ell} p$, i.e. N is one more than the product of all the primes up to and including ℓ. Since $N > 1$, it is a product of primes by the previous theorem, so there exists at least one prime q which divides N. We claim that $q \notin F$. This is because $q|N$ so N/q is an integer, but for $p \in F$, $N/p = x + 1/p$ for some integer x. In other words, for any $p \in F$, when N is divided by p, the remainder is 1, but when N is divided by q, the remainder is 0. Hence, $q \notin F$. We have shown that for every finite subset F of $\mathbf{P}$, $\mathbf{P} \setminus F \neq \emptyset$, hence $\mathbf{P}$ is infinite. □

Recall that for an integer $n \neq 0$, $\mathrm{Div}^+(n)$ is a finite set. If m is another non-zero integer, then $\mathrm{Div}^+(n) \cap \mathrm{Div}^+(m)$ is clearly a finite set and is non-empty since it contains 1, hence it has a unique largest element, which we denote by $\gcd(m, n) = \gcd(n, m)$ and of course call the greatest common divisor of m and n. Apparently the "greatest common factor" is much more à la mode in schools these days. In the reverse direction, for an integer n, we let $n\mathbb{Z} = \{nm \mid m \in \mathbb{Z}\} = \{k \mid k \text{ is divisible by } n\}$ be the set of integer multiples of n. If m and n are non-zero integers, then It is clear that $n\mathbb{Z} \cap m\mathbb{Z} \cap \mathbb{N}$ is not empty since it contains $|mn|$. Thus, this set must have a least element (by the Well-ordering principle) which we call $\mathrm{lcm}(m, n) = \mathrm{lcm}(n, m)$, the least common multiple of m and n.

Calculating the greatest common divisor and least common multiple of pairs of integers is an important computational task in many situtations, so fortunately there is a very efficient procedure for calculating them. It is based on the Division algorithm, which is simply what we usually call Long Division.

Theorem 13.13 (The Division, or Euclidean, Algorithm)**.** *Given integers $n, k \in \mathbb{N}$, there exists a unique pair (q, r) where $q \in \mathbb{N}$ and $r \in \{0, 1, 2, 3, \ldots, k-1\}$ such that $n = qk + r$. We call q the quotient and $r = \mathrm{Rem}[n \div k]$ the remainder of $n \div k$.*

Proof. First let us prove uniqueness of the pair (q, r). Suppose for pairs (q, r) and (q', r') where $q, q' \in \mathbb{N}$ and $r, r' \in \{0, 1, 2, 3, \ldots, k-1\}$, we have $n = qk + r = q'k + r'$. By switching the pairs if necessary, we may assume that $q \geq q'$. Then $(q - q')k = r' - r$. We claim that $q = q'$ and prove this by contradiction. If not, then $q - q' > 0$ and $k \geq 1$ together imply that $r' - r = (q - q')k \geq k$ which in turn implies that $r' \geq k$ since $r \geq 0$, but $r' < k$ by assumption, giving the desired contradiction. Thus, $q = q'$, and since $r' - r = (q - q')k$, we get $r' = r$ also. This proves uniqueness of the specified pair (q, r). Now let us establish the existence of this pair.

Since $k \neq 0$ by assumption, the set $S = \{x \in \mathbb{N} \mid xk \leq n\}$ is finite. It therefore has a largest element; we put $q = \max S$ for this maximal element, and let $r = n - qk$. It remains to show that $0 \leq r \leq k - 1$. Since $q \in S$, $qk \leq n$ so $r = n - qk \geq 0$. To show that $r \leq k - 1$, let us use proof by contradiction. If $r \geq k$, then

$$n = qk + r = qk + k + r - k = (q + 1)k + (r - k),$$

which, since $r - k \geq 0$ would show that $q + 1 \in S$ contradicting the fact that $q = \max S$. Thus $0 \leq r \leq k - 1$. □

The division algorithm can be used to calculate the greatest common divisor of two given positive integers efficiently. Let us examine the key idea. Given $n_1, n_2 \geq 1$, we rearrange them if necessary to have $n_1 \geq n_2$. If $n_1 = n_2$, then we rejoice because then $\gcd(n_1, n_2) = n_2$ without any further ado. The strategy is to replace the pair (n_1, n_2) by a **smaller** pair (n_2, n_3) with the **same** gcd! So where do we get n_3 from? Easy, we take n_3 to be the remainder when n_1 is divided by n_2! Thus, we need to prove a little lemma.

Lemma 13.14. *If $n \geq k \geq 1$ and $n = qk + r$ with $q \in \mathbb{N}$, then $\gcd(n, k) = \gcd(k, r)$.*

Proof. Recall that $\gcd(n, k)$ is by definition the largest element of $\mathrm{Div}^+(n) \cap \mathrm{Div}^+(k)$, thus it suffices to prove that

$$\mathrm{Div}^+(n) \cap \mathrm{Div}^+(k) = \mathrm{Div}^+(k) \cap \mathrm{Div}^+(r).$$

To prove the above equality of sets, we well show that each set is contained in the other. So, suppose $d|n$ and $d|k$. Then $d|qk$ so $d|(n - qk)$ i.e. $d|r$. So now $d|k$ and $d|r$ showing that $\mathrm{Div}^+(n) \cap \mathrm{Div}^+(k) \subseteq \mathrm{Div}^+(k) \cap \mathrm{Div}^+(r)$. On the other hand, if $d|r$ and $d|k$, then $d|qk$ so $d|(qk+r)$ i.e. $d|n$, showing the reverse inclusion. This completes the proof of the lemma. □

The strategy for computing $\gcd(n_1, n_2)$ should now be clear. Let $n_3 = \mathrm{Rem}[n_1 \div n_2]$, and indeed for each $i \geq 3$, successively define $n_i = \mathrm{Rem}[n_{i-2} \div n_{i-1}]$. Then, $n_2 > n_3 > \cdots$ gives a strictly decreasing sequence of remainders (which are automatically non-negative!), i.e. $n_2 > n_3 > \cdots \geq 0$, thus this sequence must eventually hit 0. Let $s \geq 2$ be the least integer such that $n_s = 0$. Thus, we have

$$\gcd(n_1, n_2) = \gcd(n_2, n_3) = \cdots = \gcd(n_{s-2}, n_{s-1}) = \gcd(n_{s-1}, 0), \qquad n_{s-1} \neq 0.$$

Since $n_{s-1} \neq 0$, $\gcd(n_{s-1}, 0) = n_{s-1}$. Another perspective is that since $n_s = \mathrm{Rem}[n_{s-2} \div n_{s-1}] = 0$, we have $n_{s-1}|n_{s-2}$ and hence $\gcd(n_{s-2}, n_{s-1}) = n_{s-1}$. Either way, we find $\gcd(n_1, n_2) = n_{s-1}$ is the penultimate remainder (just before getting remainder 0).

Example 13.15. Let us use the above algorithm to compute $\gcd(432, 60)$. So, $n_1 = 432$, $n_2 = 60$. We get $432 = 7 \cdot 60 + 12$ so $n_3 = 12$, and $60 = 5 \cdot 12$ so $n_4 = 0$. Thus, $\gcd(432, 60) = n_3 = 12$. Let's do one more. What is $\gcd(89, 55)$? Letting $n_1 = 89$, $n_2 = 55$, we have $n_3 = 34$, $n_4 = 21$, $n_5 = 13$, $n_6 = 8$, $n_7 = 5$, $n_6 = 3$, $n_7 = 2$, $n_8 = 1$, $n_9 = 0$. Phew, $\gcd(89, 55) = n_8 = 1$.

Definition 13.16. If m, n are integers, we say that m and n are coprime or relatively prime to each other if $\gcd(m, n) = 1$.

Theorem 13.17 (Bezout's Theorem)**.** *Suppose a, b are integers and $d = \gcd(a, b)$. Then there exist integers x, y such that $ax + by = d$. In particular, if a and b are relatively prime, then some integer linear combination of a and b is 1. Indeed, for $m \in \mathbb{Z}$, the equation $aX + bY = m$ is solvable with $X, Y \in \mathbb{Z}$ if and only if $d|m$.*

Sketch of Proof. The integers x, y can in fact be found via the repeated application of the Euclidean algorithm we described for computing $\gcd(a, b)$. Recall that we put $n_1 = \max(a, b)$, $n_2 = \min(a, b)$ and define recursively n_{j+1} to be the remainder of n_{j-1} divided by n_j for $j \geq 2$, viz. $n_{j-1} = q_j n_j + n_{j+1}$. Then $\gcd(a, b) = n_{s-1}$ where $n_s = 0$ (with s minimal for this property). From $n_{s-3} = q_{s-2} n_{s-2} + n_{s-1}$, we climb one level higher to $n_{s-1} = n_{s-3} - q_{s-2} n_{s-2} = n_{s-3} - q_{s-2}(n_{s-4} - n_{s-3} q_{s-3})$, and so on until we obtain an expression $n_{s-1} = x n_1 + y n_2$ for some integers x, y. Once we have $x, y \in \mathbb{Z}$ with $ax + by = d$ where

$d = \gcd(a, b)$, then for any multiple m of d, say $m = kd$, we have $aX + bY = m$ for $X = kx$ and $Y = ky$. On the other hand, suppose $aX + bY = m$ with integers X, Y. We want to show that then m is a multiple of d, so let us divide and see: we have $m = qd + r$ for integers q, r where $0 \leq r < d$. By multiplying $ax + by = d$ by q, we find $axq + byq = m - r$. We subtract this from $aX + bY = m$ to find $a(X - xq) + b(Y - yq) = r$. Since $d|a$ and $d|b$, we conclude that $d|r$, which, when combined with the inequality $0 \leq d < r$ gives $r = 0$, i.e. $d|m$ as desired. □

Remark. We should note the following important interpretation of the theorem. The set of $\mathbb{Z}$-linear combinations of a and b is exactly the set of $\mathbb{Z}$-multiples of their greatest common divisor, i.e. we have an equality of sets

$$\{aX + bY \mid X, Y \in \mathbb{Z}\} = \{kd \mid k \in \mathbb{Z}\}.$$

Even more briefly, one can write $a\mathbb{Z} + b\mathbb{Z} = d\mathbb{Z}$ where $d = \gcd(a, b)$. Later when you study rings (in Math 412), you will come to interpret this statement as "The ideal generated by a and b is principal, generated by $\gcd(a, b)$."

Example 13.18. Determine $\gcd(200, 126)$ and express it as a linear combination of these integers. We write

$$\begin{aligned}
200 &= 126 + 74 \qquad n_3 = 74 \\
126 &= 74 + 52 \qquad n_4 = 52 \\
74 &= 52 + 22 \qquad n_5 = 22 \\
52 &= 2 \cdot 22 + 8 \qquad n_6 = 8 \\
22 &= 2 \cdot 8 + 6 \qquad n_7 = 6 \\
8 &= 6 + 2 \qquad n_8 = 2 \\
6 &= = 3 \cdot 2 \qquad n_9 = 0.
\end{aligned}$$

Thus, $n_8 = 2 = \gcd(200, 126)$. Reversing the steps, we have

$$\begin{aligned}
2 &= 8 - 6 \\
&= 8 - (22 - 2 \cdot 8) \\
&= -22 + 3 \cdot 8 \\
&= -22 + 3(52 - 2 \cdot 22) \\
&= 3 \cdot 52 - 7 \cdot 22 \\
&= 3 \cdot 52 - 7(74 - 52) \\
&= -7 \cdot 74 + 10 \cdot 52 \\
&= -7 \cdot 74 + 10(126 - 74) \\
&= 10 \cdot 126 - 17 \cdot 74 \\
&= 10 \cdot 126 - 17(200 - 126) \\
&= 27 \cdot 126 - 17 \cdot 200.
\end{aligned}$$

There is a nice method, advocated by W.A. Blankinship (*Amer. Math. Monthly, 1963*), for keeping track of the straightforward but somewhat messy book-keeping of the above algorithm. It produces $\gcd(n_1, n_2)$ and the "Bezout numbers" x, y such that $xn_1 + yn_2 = \gcd(n_1, n_2)$ all in one shot. Namely, to find $\gcd(n_1, n_2)$, we write them in a column next to

the 2×2 identity matrix, then we do the usual operations for finding $n_3, n_4, \ldots$ but apply each operation to the whole row. We stop when we reach a row that begins with 0. The penultimate row will then be d, x, y where $d = \gcd(n_1, n_2)$ and $d = xn_1 + yn_2$! Instead of giving a formal algorithm (and proving that it does what we say), we will be satisfied with reworking the above example with Blankinship as our guide.

Example 13.19. To find $\gcd(200, 126)$, we follow the same steps as before, but carry the algebra to the entire row each time:

$$\begin{array}{rrr} 200 & 1 & 0 \\ 126 & 0 & 1 \\ 74 & 1 & -1 \\ 52 & -1 & 2 \\ 22 & 2 & -3 \\ 8 & -5 & 8 \\ 6 & 12 & -19 \\ 2 & -17 & 27 \\ 0 & 63 & -100. \end{array}$$

We read off that $-17 \cdot 200 + 27 \cdot 126 = 2$. We also can read off $6 = 12 \cdot 200 - 19 \cdot 126$ etc. in case we wanted to. Note that $0 = 63 \cdot 200 - 100 \cdot 126$, in other words, the sixty-third multiple of 200 is also the hundredth multiple of 126, and so this number, $63 \cdot 200 = 100 \cdot 126 = 12600$ is a common multiple of 200 and 126. Are you thinking what I'm thinking? This must be the *least* common multiple of 200 and 126! Yes, that is true.

Remark. If you are familiar with row operations on matrices, you will note that the sequence of moves in the Blankinship algorithm is nothing more than that. I leave it as a challenge to the interested reader to investigate (and prove if true) whether the last row of the Blankinship algorithm will always display $0\ r\ s$ where $|rn_1| = |sn_2| = \operatorname{lcm}(n_1, n_2)$.

The Bezout theorem has a bunch of important and useful consequences.

Theorem 13.20. *If p is a prime and $p|ab$ where $a, b \in \mathbb{Z}$, then $p|a$ or $p|b$.*

Proof. Let us write $ab = pk$ for some $k \in \mathbb{Z}$. If a and b are both divisible by p, then we are done. So, let us assume one of them is not divisible by p, say b. Then by Bezout's theorem, there exist $x, y \in \mathbb{Z}$ such that $bx + py = 1$. Multiplying this last equation by b, we find $abx + apy = a$, or $p(k + ay) = a$, so $p|a$. □

Corollary 13.21. *If $n \geq 1$ and $m_1, \cdots, m_n \in \mathbb{Z}$ are n integers whose product is divisibe by p, then at least one of these integers is divisible by p, i.e. $p|m_1 \cdots m_n$ implies that then there exists $1 \leq j \leq n$ such that $p|m_j$.*

Proof. The proof is by induction on n, and is left as an exercise. □

Corollary 13.22. *For $a, b, c \in \mathbb{Z}$, if $a|bc$, and $\gcd(a, b) = 1$, then $a|c$.*

Proof. We use Bezout to write $ax + by = 1$ with $x, y \in \mathbb{Z}$. We multiply this by c to get $axc + bcy = c$, then note that $a|axc$ and $a|bcy$, so $a|axc + bcy = c$. □

Another consequence of the Bezout theorem is the following. Let's give it a fanciful name in the hope that you will remember its statement. It will be extremely useful to you when you study group theory.

Theorem 13.23 (The Supremacy of gcd and lcm)**.** *Suppose $a, b \in \mathbb{Z}$. Every common multiple of a and b is a multiple of their least common multiple* $\operatorname{lcm}(a, b)$ *and every common divisor of a and b is a divisor of their greatest common divisor* $\gcd(a, b)$*. In other words,*

$$\begin{aligned} a|c, b|c &\implies \operatorname{lcm}(a, b)|c \\ d|a, d|b &\implies d|\gcd(a, b). \end{aligned}$$

Proof. Let $l = \operatorname{lcm}(a, b)$ and $g = \gcd(a, b)$. First, let's show that $a|c, b|c \Rightarrow l|c$. We may write $c = as$ and $c = bt$ for integers s, t. We want to show that c divided by l gives remainder 0, so let's divide and see! We have $c = lq + r$ for some integer q and some r satisfying $0 \leq r < l$. We have $c = at = lq + r$ so $r = at - lq$. Since l is a multiple of a, we then have $a|r$. Similarly, $c = bu = lq + r$ so $r = bu - lq$ is a multiple of b. Thus, r is a common multiple of a and b. But $0 \leq r < l$ and l is the least (positive!) common multiple of a and b so r cannot be positive. Thus $r = 0$, i.e. l divides c.

Now let's show that $d|a, d|b \Rightarrow d|g$. We may write $a = de$ and $b = df$ with $e, f \in \mathbb{Z}$ (by assumption), and $g = ax + by$ for $x, y \in \mathbb{Z}$ (by Bezout). Assembling all of this together, we get $g = dex + dfy = d(ex + fy)$, hence $d|g$. □

Now let us state and prove the Fundamental Theorem of Arithmetic. It says that, except for the way the prime factors are ordered, how a number breaks up into prime factors is unique.

Theorem 13.24 (The Fundamental Theorem of Arithmetic)**.** *If $n \in \mathbb{N}$, then there is a unique function $e_n : \mathbf{P} \to \mathbb{Z}_{\geq 0}$ from the set of all primes $\mathbf{P}$ to the set of non-negative integers such that*

$$n = \prod_{p \in \mathbf{P}} p^{e_n(p)}.$$

The function e_n vanishes on all but finitely many primes.

Proof. We have already shown in Theorem 13.10 that every integer > 1 is a product of primes (and 1 is an "empty" product of primes, i.e. the function e_0 is just the function that takes the value 0 at every prime). To show uniqueness, let us proceed by contradiction (hoping to use the well-ordering principle once again). So, we suppose that there exist positive integers $n > 1$ that admit at least two distinct factorizations. By the well-ordering principle, there exists a least such integer, let us call it m. Thus, there exist two factorizations, $m = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$, where the p_i, q_j are all primes, not necessarily distinct, ordered so that $p_1 \leq p_2 \leq \ldots \leq p_r$ and $q_1 \leq q_2 \leq \ldots \leq q_s$. By assumption, the lists $p_1, \ldots, p_r$ $q_1, \ldots, q_s$ are not identical. We can assume without loss of generality[22] that $p_1 \leq q_1$. By Corollary 13.21, $p_1|q_i$ for some $1 \leq i \leq s$. Since q_i and p_1 are both primes, we then have $p_1 = q_i$. So, $p_1 \leq q_1 \leq q_i = p_1$, so $p_1 = q_1$. Letting $m' = m/p_1$, we have, $m' = p_2 \cdots p_r = q_2 \cdots q_s$. Now these two factorizations must be distinct, since the two distinct factorizations of m are gotten by including the equal prime factors p_1 and q_1 at the beginning of each one. Thus, $0 < m' < m$ and m' has two distinct factorizations, contradicting the fact that m is the

[22]This oft-quoted phrase warns the reader that the author is about to make an assumption, but that this assumption is not central to the validity of the proof. Without the assumption, a simple and obvious modification or repetition of the argument can be made to account for all possible cases. For instance, in this case, if it happens that $q_1 \leq p_1$, then we simply repeat the argument, replacing all the q's by p's and vice versa.

least positive integer admitting two distinct factorizations. This contradiction completes the proof. □

To compute $\mathrm{lcm}(m,n)$, one can compute $\gcd(m,n)$ and then use part (c) of the following fact.

Theorem 13.25. *Suppose $m,n \geq 1$ and*

$$m = \prod_{p \in \mathbf{P}} p^{e_m(p)}, \qquad n = \prod_{p \in \mathbf{P}} p^{e_n(p)}.$$

Then

(a)

$$\gcd(m,n) = \prod_{p \in \mathbf{P}} p^{\min(e_m(p), e_n(p))}.$$

(b)

$$\mathrm{lcm}(m,n) = \prod_{p \in \mathbf{P}} p^{\max(e_m(p), e_n(p))}.$$

(c) If $m,n \geq 1$, then $\mathrm{lcm}(m,n) \cdot \gcd(m,n) = mn$.

Proof. We leave the proof to the interested reader. □

14. Some more number theory

One of the biggest mysteries in number theory is the following problem:

Major Problem. Explain the distribution of prime numbers on the number line.

If we list the primes in order, then it becomes apparent fairly quickly that they start to "thin out." In other words, if you take an interval of length N for a large but fixed N, then look at N consecutive positive integers, starting with $a+1$, then the chances that this interval $[a+1, a+N]$ contains a prime goes to zero as a goes to infinity. Here is a "movie" of this phenomenon: If you take a "window" of fixed width and shift it to the right, the chances that you catch a prime for any given frame goes to zero as you shift to the right.

In a sense, you should expect that the primes are "outmuscled" by the composites, because everytime you have a bunch of primes, you can combine them in many ways in order to make composites, but there is only one way to make a prime. In particular, in one of the homework problems, you will show that no matter how large N is, as you shift to the right, you are bound to hit a frame with no primes in it. Here is another way in which composites "outmuscle" the primes.

Example 14.1. A sequence $x_0, x_1, x_2, \cdots$ in $\mathbb{Z}$ is called *arithmetic* if there exists an integer a (called the *addend*) such that $x_{n+1} - x_n = a$ for all $n \geq 1$. Equivalently, $x_n = x_0 + na$. Show that any arithmetic sequence in $\mathbb{Z}$ with non-zero addend contains infinitely many composites.

Here is a proof. Suppose $(x_n)_{n \geq 0}$ is an arithmetic sequence with addend a. If all the x_n are composite, we are certainly done! If not, let $p = x_0 + ma$ be prime for some $m \geq 0$. We claim that if $n = m + kp$, where $k \geq 1$, then x_n is composite. Once we prove the claim, we are done, of course. To prove the claim, note first that x_n is divisible by p because

$$x_n = x_0 + an = x_0 + a(m+kp) = x_0 + am + akp = p + akp = p(1+ak).$$

Note that $x_n = p(1 + ak) > p$ for $k \geq 1$, hence x_n is divisible by p and greater than p hence it is composite, proving the claim.

On the other hand, primes are "persistent" in some ways. For instance, as we proved, there are infinitely many of them! A much more subtle and powerful theorem, first formulated by Lagrange, and finally proved by Peter Gustav Lejeune Dirichlet in 1837, says that in any arithmetic progression that has the potential of having infinitely many primes does have infinitely many primes.

Theorem 14.2 (Dirichlet's Theorem on Primes in Arithmetic Progressions)**.** *If $x_0, a \in \mathbb{N}$ and $\gcd(a, x_0) = 1$, then the arithmetic sequence $x_0, x_0 + a, x_0 + 2a, \ldots, x_0 + an, \ldots$ contains infinitely many primes.*

Note that the assumption $\gcd(x_0, a) = 1$ is needed, because otherwise all the elements in the sequence are divisible by $d > 1$ where $d = \gcd(x_0, a)$. Dirichlet's ideas for proving this theorem consitute the foundations of an entire branch of modern mathematics known as "analytic number theory."

Another "prime persistence" theorem, due to Chebyshev, is known as "Bertrand's Postulate." It says that for $n \geq 1$, the interval $(n, 2n]$ contains at least one prime. Here is an unsolved problem.

Question 14.3. Is it true that for all large enough n, (say $n \geq 117$), the interval $[n, n + \sqrt{n}]$ contains a prime? (It is believed that the answer is "yes" but no proof or counterexample is known at present).

A spectacular and recent "prime persistence" theorem is the following.

Theorem 14.4 (Peter Green and Terence Tao, 2004)**.** *Given $N \geq 1$, there exists integers $x_0, a \in N$ such that $x_0 + a, x_0 + 2a, \ldots, x_0 + Na$ are all primes. In other words, there are arbitrarily long arithmetic progressions of primes.*

See `http://arxiv.org/abs/math.NT/0404188` for their paper. You may not understand much, but you'll get a glimpse of what a mathematical "preprint" (an article in pre-published form) looks like. One of the most exciting aspects of their proof is that it uses techniques of "ergodic theory," a branch of "analysis" (calculus).

The fact that the study of smooth functions $\mathbb{R} \to \mathbb{R}$ should say anything about arithmetic properties of whole numbers might be surprising at first, but this tradition actually goes way back to Leonhard Euler at least who gave a proof of the infinitude of primes based on the fact that the harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

diverges! Euler's observation led Georg Bernhard Riemann to the study of the function

$$\zeta(x) = 1 + \frac{1}{2^x} + \frac{1}{3^x} + \cdots + \frac{1}{n^x} + \cdots$$

which Euler had introduced, but which we now call the **Riemann Zeta Function**. Riemann observed that the *analytic* properties of this function reveal some deep *arithmetic* facts about the distribution of primes on the number line! The connection between them is sealed by the Euler Product Formula:

$$\zeta(x) = \prod_{p \in \mathbf{P}} \frac{1}{1 - \frac{1}{p^x}}$$

which in turn holds because of the Fundamental Theorem of Arithmetic. In 1859, Riemann outlined a program, completed by de la Vallée-Poussin and Hadamard independently in 1896, for proving a conjecture of Gauss which we now call the Prime Number Theorem. To state it, let us define the Prime Counting Function $\pi(x) = |\{p \in \mathbf{P} \mid p \leq x\}|$ which counts the number of primes in the interval $[1, x]$. Up close, this function is quite choppy, as it jumps by 1 everytime it encounters a prime. But if you look at its graph on a very large interval, it looks remarkably smooth. So the question is: Is there a nice simple continuous function whose graph approaches the graph of $\pi(x)$ as x tends to infinity? The answer is "Yes," and one function which fits the bill is $x/\ln(x)$.

Theorem 14.5 (The Prime Number Theorem)**.** *We have*

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\ln(x)} = 1.$$

The theorem says that $\pi(x)$ and $x/\ln(x)$ are about the same size. For example, in 1959, Derrick Lehmer (my mathematical grandfather) calculated on his super-duper computer that $\pi(10^{10}) = 455052511$,i.e. is about 455 million. Let us compare that to $10^{10}/\ln(10^{10}) \approx 434294482$ or about 434 million. I wonder whether you are impressed by this or not. On the one hand, we are off by about 21 million primes! On the other hand, calculating $10^{10}/\ln(10^{10})$ takes just a second whereas counting how many primes there are up to 10^{10} is serious business. In retrospect, 21 million primes out of 455 million is only about a 4.5% error. Not too bad at all! Nonetheless, one would like to understand how much the error $\pi(x) - x/\ln(x)$ is, or at least put a cap on this error. Riemann's method does give us a bound on this error, but the bound is MUCH bigger than the actual errors we observe. Riemann has an explanation for that too: He thinks it is highly likely that the roots of his function (not just for x in $\mathbb{R}$ for complex numbers x) all lie on a certain line. To find out whether this is true or not is one of the hottest problems in Mathematics. It is known as The Riemann Hypothesis, the subject of various recent popular books.

Part 8. Counting and Uncountability

15. The classification of sets according to size

Let us begin by reviewing some facts about finite sets. Two main facts are as follows.

Lemma 15.1 (Main Lemma of Finite Sets)**.** *Suppose $f : X \hookrightarrow Y$ is an injective map of sets. Then,*

(a) If $x_1, \ldots, x_n$ is a non-repeating sequence of length $n \geq 1$ in X, then $f(x_1), \ldots, f(x_n)$ is a non-repeating sequence of length n in Y.

(b) If Y is finite, then so is X and $|X| \leq |Y|$.

Theorem 15.2 (Main Theorem of Finite Sets)**.** *Suppose X and Y are finite sets. Then, $X \sim Y$ if and only if $|X| = |Y|$.*

Proof. If $X \sim Y$, then there exists a bijective map $f : X \to Y$. Since f is injective, $|X| \leq |Y|$ by the Main Lemma of Finite Sets. Since f is bijective, we also have $f^{-1} : Y \to X$ which is bijective hence injective, so we also get $|Y| \leq |X|$. Since $|X| \leq |Y| \leq |X|$, we must have $|X| = |Y|$. The other direction is also easy, it is left as a homework problem. □

A set X is infinite if and only if there exists an injective map $f : \mathbb{N} \hookrightarrow X$. For, if such a map exists, then $f(1), f(2), f(3), \ldots$ is a non-repeating infinite sequence in X, and in the other direction, if $x_1, x_2, \ldots$ is an infinite non-repeating sequence in X, then by putting $f(n) = x_n$ for $n \geq 1$, we see immediately that f gives an injection $\mathbb{N} \hookrightarrow X$.

How should we measure the "size" of infinite sets? Our first thought might be (as it was for a period lasting at least 4000 years!) that all infinite sets are of the same size. It was only in the 19th century that Georg Cantor cured us of this naïveté singlehandedly. Here is an analogy. Suppose in all your life, you have never seen or heard of a car. The fastest method of travel you know about is riding a horse. Then one day you discover cars. In the beginning of your discovery, you will not be able to tell the difference between a Yugo and a Ferrari. To you, these new "machine horses" are so extraordinarily fast and superior to your prior experiences that the relative differences between them are irrelevant. But after driving in a Yugo for a while, then moving on to a Chevrolet and a Mercedes and a Ferrari, you come to distinguish and understand the differences between them. Then one day you see a Cessna plane cruising down a runway and you think well, here is another machine horse, and then you see it take off and fly, and you realize "Whoa! Now that thing belongs to a whole higher class of machine horses by itself!!"

So it was for our mathematical understanding of infinity. At first, you are familiar with finite sets. You realize early on that some finite sets are "alike" (they have the same size) and are in that sense "equivalent." Then one fine day, you realize "Hey, there is NO largest counting number!" i.e. "Whoa! the set of counting numbers (which 'classify' finite sets) is not a finite set itself!" In other words, you have just discovered the existence of infinite sets. This is such a monumental discovery, infinity is such a remote concept, that to you just assume that all infinite sets are of the same size. This relative homogeneity of infinite sets is expressed in your language: you have a name for sets of size 1, size 2, size 3, etc. but you have only one name for sets which are not finite (infinite) reinforcing the notion that all infinites sets are of the same size.

Let us look more closely to see whether all infinite sets really are of the same size. Well, what does it mean for two sets to be of the same size? For finite sets, it means that there

is a one-to-one correspondence between the two sets: we can line up the elements of one set against the elements of the other set in an exact match-up. Cantor had the important insight that this definition should be used for infinite sets as well. Recall that we write $X \sim Y$ if there is a bijection $X \to Y$.

Definition 15.3. For any pair of arbitrary sets, X and Y, we say that X and Y have the same size and write $|X| = |Y|$ if and only if $X \sim Y$.

Having accepted this definition, we now must dare ask, with Cantor, a fundamental question.

Let's imagine a conversation Cantor could have had with a fictitious colleague (and a very old friend) who has just come for a visit from out of town, let us call him Professor Groeg Rotnac, as they stroll along on Ulmestrasse one summer evening in the early 1870s.

Professors Cantor and Rotnac Take a Walk

Prof Rotnac: What work is occupying you these days, esteemed colleague?

Prof Cantor: I have been fascinated and considerably puzzled by a fundamental question which on the surface appears quite simple but whose complexities, in my opinion, run very deep. This question has been troubling me for years and only recently have I been able to make any progress. Given the simplicity of the question, I hesitate to mention it, even, to such a dear and old friend as yourself.

Prof. Rotnac: Come now, we have known each other practically from infanthood! We have no need of reticence! Please, I am curious, I will treat your question with respect.

Prof. Cantor: Very well, I would expect no less from you. You would agree that two sets, be they finite or infinite, should be said to be of the same size if there is a one-to-one correspondence, i.e. a bijection, between them.

Prof. Rotnac: On that, we are in complete agreement. You might say, on that there is a correspondence between our thoughts. Ha Ha Ha.

Prof. Cantor: Yes, that is very witty of you. Very good. Now, here is my innocent-sounding question: If X, Y are two infinite sets, are they necessarily of the same size? In other words, if X, Y are infinite sets, then is there necessarily a bijection from X to Y?

Prof. Rotnac: Mit Verlaub, with all due respect, I believe you are teasing me.

Prof. Cantor: Need I remind you that you promised you would take my question at face value? Here is my question again, please treat it seriously: If X and Y are infinite sets, can one always find a bijection from X to Y?

Prof. Rotnac: Jawohl, Natürlich, of course one can, my dear Georg! Das ist ein Kinderspiel: It is child's play not worthy of a Herr Doctor Professor of your standing!

Prof. Cantor: Indulge me and present your kinder-proof, bitte.

Prof. Ratnoc: Very well, list the elments of X and Y as $x_1, x_2, x_3, \ldots$ and $y_1, y_2, y_3, \ldots$, respectively. Then map x_i to y_i for each $i \geq 1$.

Prof. Cantor: Your proof rests on the notion that if X is any infinite set, then one can simply list all its elements in a single infinite list, i.e. that there is a bijection from the

natural numbers $\mathbb{N}$[23] to X. Then, since there is a bijection from X to $\mathbb{N}$ and another from $\mathbb{N}$ to Y, composing them one gets a bijection from X to Y. Is that about it?

Prof. Ratnoc: You have expressed my thoughts perfectly, as you always do.

Prof. Cantor: Yes, but this brings us to the crux of the matter. I find your claim that one can always list all the elements of an arbitrary infinite set in a single infinite list LÄSTIG, troubling at best, indeed downright false as we shall hopefully see in a moment. Pray discuss how you would justify your claim that if X is any infinite set, then its elements can be listed in one infinite list.

Prof. Rotnac: Well, simply choose one element at random, call it x_1, then a different one, call it x_2, then an x_3 different form x_1 and x_2 and so on. Since the set is infinite, one will never run out of elements to choose, and one will eventually list every single element.

Prof. Cantor: I am afraid that will not work.

Prof. Rotnac: Of course it will.

Prof. Cantor: NEIN.

Prof. Rotnac: I see that you are quite sure of yourself. Very well, why not, pray?

Prof Cantor: The trouble with your procedure is your claim that that "one will eventually list every single element." Let me illustrate with a very simple example. Suppose we attempt to use your procedure to list all the elements of the set $\mathbb{N}$. You said I can choose the elements as I please. Suppose I happen to choose 2, then 4, then 6, then 8, and so on, always skipping the odd numbers. In this way, I will never stop and will never list all the elements since I miss every odd number.

Prof. Rotnac: Very well, the procedure does not work if you apply it in a deliberately obtuse manner, but surely there must be a way of listing the elements correctly so that none gets left out. I grudgingly admit that I have not yet given you a correct procedure that will always list all the elements of a given infinite list, but I do not give up on the claim that such a procedure exists. For instance, of course one *can* list all the elements of $\mathbb{N}$, simply ask a child to do it and she will do it in the right order!

Prof. Cantor: For the set $\mathbb{N}$, yes, of course, but recall that we are asking if the elements of an arbitrary infinite set X can always be given in one infinite list (let us call that simply "listing the elements" and a set for which we can list the elements in one infinite list I will call "listable"). So far, I was simply showing you that your procedure is not guaranteed to work. Let us continue our investigation for various infinite sets and attempt to find the means of listing the elements, which you so adamantly maintain must exist.

Prof. Rotnac: Very well, how about Herr Dr Professor Kronecker's favored set, $\mathbb{Z}$?

Prof. Cantor: (With barely disguised disgust) BITTE do not speak of that distinguished yet pompous gentleman.

Prof. Rotnac: Please forgive me; I did not know that mentioning his name would upset you so. But do let us take up the case of the set of all integers, $\mathbb{Z}$, positive, negative and 0. This is a good test case to start with, because it is usually thought of as a "two-sided" infinite list. and you would like a one-sided infinite list? Hmm, just a moment. I have it! We start at 0, then go to 1, then circle around to -1, then circle around to 2, then circle around to -2, and so on, getting: $0, 1, -1, 2, -2, 3, -3, 4, -4, \ldots$. Every element is now listed in one infinite list.

[23]For the sake of convenience, we allow several anachronisms, such as the use of words like "bijection" in this discourse. Another anachronism is the nuse of $\mathbb{N}$ to stand for the set of natural numbers. The notation $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and so on was not standardized until many years later when Bourbaki came along

Prof. Cantor: Excellent work, MEINE FREUND! You have discovered a method I call "interleaving." Here is what I mean: Given two sets X and Y which are listable (say as $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$), then $X \cup Y$ is also listable by interleaving the two sets, as follows: $x_1, y_1, x_2, y_2, x_3, y_3, \ldots$, just as you interleaved the positives and negatives. Moreover, suppose given a finite collection of sets $X_1, X_2, \ldots, X_n$, each of which is listable, then you can imagine ....

Prof. Rotnac: Tut tut, just a moment, I know what you are about to say: by listing the first element of the first, the first element of the second, etc. up the the first element of the last set, then going to the second element of the first set and so on, this shows that $X_1 \cup \cdots \cup X_n$ is listable!

Prof. Cantor: Sharp as the axe of the mightiest lumberjack of the Black Forest is your mind, Herr Dr Professor.

Prof. Rotnac: Very funny. Well, that takes care of quite a few infinite sets already!

Prof. Cantor: Yes, you can well imagine the next step: what if one has an infinite list of listable sets, is their union listable? Again, there is a simple way to show that this is so (I should say it's simple after one sees how to do it, but I can confess to my esteemed colleage because he is also an intimate friend that it took me many months to come up with this answer!). Namely, let us write $x_{11}, x_{12}, x_{13}, \cdots$ for the first list, then $x_{21}, x_{22}, x_{23}, \cdots$ for the second list and so on. So that x_{ij} is the jth element in the ith list. Do you see the lists in your mind's eye?

Prof. Rotnac: Yes, yes, you have put them there beautifully. I see them as a 2-dimensional grid with the first list at the bottom going off to the right, the second list just above it and so on.

Prof. Cantor: You flatter me, but it's true that you imagine them just as I do. Well, now, let us say that the element x_{ij} has "weight" $i + j$. The lightest element, so to speak, is x_{11} and it has weight 2, the next two lightest are x_{12} and x_{21} of weight 3, then next three lightest are of weight 4, they are x_{31}, x_{22}, x_{13}. The next four lightest are of weight 5, they are $x_{14}, x_{23}, x_{32}, x_{41}$. And then ...

Prof. Rotnac: ICH HABE DEIN SPIEL DURCHSCHAUT, I can see your game! You list the elements by increasing weight, and since there are only finitely elements of each weight, you simply order those finitely many elements however you please as you go along. In fact, I think you have been zigzagging them!

Prof. Cantor: Exactly! I have always admired your quick uptake of ideas.

Prof. Rotnac: Well, as I said before, it appears you are well on your way to proving that every infinite set is listable!

Prof. Cantor: That is what I thought as well. Notably, with the "filtering by weight" idea I have just described, one sees that the set of rational numbers $\mathbb{Q}$ is listable. But alas, my success stalled there for quite some time, for I have been utterly unsuccessful in listing all the elements of the set of real numbers $\mathbb{R}$. So much so, that I now believe that this set is not listable!

Prof. Rotnca: Now now, do not be hasty. I understand why you are discouraged, but do not give up hope. No one has yet found the holy grail, but that does not prove that it does not exist. Patience, I am sure your fertile mind will soon devise a method for listing all the real numbers as well. After all, each real number can be expressed as infinite list of finitely many digits. Will not some refinement of your zigzagging on diagonals provide the answer?

Prof. Cantor: Ah, I have indeed come to the conclusion that it does provide the answer, but the answer is NOT the one that you expect my dear colleague.

Prof. Rotnac: You mean to say that you have found a proof that the real numbers are not listable?!

Prof. Cantor: Precisely.

Prof. Rotnac: (Unbelieving, but excited all the same) Come now, be serious. How can such a monstrous thing be true? You have piqued my curiosity. I am sure you are mistaken, however. Come, let us have your so-called "proof," and if it is not overly elaborate, within minutes we will have found a shortcoming in your argument or my name is not Groeg Rotnac.

Prof. Cantor: Very well, I must admit I would be relieved to find an error in the argument, which is indeed quite short, but I am afraid there is none. In fact we will prove that the set of real numbers bounded by 0 and 1, the so-called unit interval, already has so many elements that they cannot all be given in one infinite list. We will proceed by Widerspruch, contradiction. So, suppose you offer me a list $x_1, x_2, x_3, \ldots$ of real numbers between 0 and 1 and claim that **all** real numbers between 0 and 1 figure among your list. I will now show you that, with all due respect, you cannot have done so. Namely, I will produce a real number between 0 and 1 which cannot be on your list. You see what I wish to do?

Prof. Rotnac: Yes, I see what you wish to do, but cannot see how you will possibly achieve it!

Prof. Cantor: Neither could I for a long time. You will agree, readily, I believe, that by definition, each real number larger than 0 but smaller than 1 can be represented in a unique way with an infinite decimal expansion $c_1c_2c_3\ldots$ where the digits c_i belong to the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Let us say we write the first number on your list as $x_1 = 0.a_{11}a_{12}a_{13}\ldots$ the second number on your list as $x_2 = 0.a_{21}a_{22}a_{23}, \ldots$ so that the jth digit of the ith number on the list is a_{ij}. Here of course the digits a_{ij} belong to the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Are you with me?

Prof. Rotnac: Yes, I am. Go on, please.

Prof. Cantor: Recall that my goal is to show you that some number in the unit interval must have been left out of the list. Here is how to do it. I will construct such a number $y = b_1b_2b_3\ldots$ by making sure that y differs from each number in the list for at least one digit. In fact, I will choose it one digit at a time so that for each $j \geq 1$, b_j is different from a_{jj}. For instance, if a_{11} is not 5, I will choose b_1 to be 5, whereas if a_{11} is 5 then I will choose b_1 to be some other digit, say 6 for the sake of definiteness. If a_{22} is not 5, I will choose b_2 to be 5 and otherwise I will choose it to be 6. And so on. Thus, y cannot be x_1 because they differ in the first digit, and y cannot be x_2 because they differ in the second digit and so on. For any $j \geq 1$, y cannot be x_j because their jth digits do not match. Thus, y (which will be some real number in the interval $[0.5555\ldots, 0.6666\ldots] = [5/9, 6/9]$) does not appear on the list $x_1, x_2, \ldots$ no matter how this list was constructed. This proves that the real numbers bounded by 0 and 1 are already too numerous to be put into one single list.

Prof. Rotnac: I am stunned! You have just shown that the set of real numbers is **more infinite** somehow than the set of natural numbers. You have opened my mind to a universe whose existence I did not even suspect! Infinite sets come in different sizes? Infinite sets come in different sizes.

Prof. Cantor: You believe the proof then?

Prof. Rotnac: Indeed I do. You are travelling down the diagonal modifying the digits as you go along. This "diagonal argument" of yours is simplicity itself. I congratulate you!

Have you found any other sets that are not listable? Have you found sets that are even bigger than the set of real numbers?

Prof. Cantor: Actually, yes to both questions. Let $\mathcal{P}(\mathbb{N})$ be the set of all subset of $\mathbb{N}$. If you begin to list its elements, you will see that they are quite numerous. Other than the empty set and the complete set $\mathbb{N}$ itself, there is one infinite list of subsets of size 1, a 2-dimensinonal infinite grid of subsets of size 2, a 3-dimensional infinite grid of subsets of size 3 and so on. I have found a rather roundabout way of showing that $\mathcal{P}(\mathbb{N})$ is not listable, on the one hand, and also of showing that $\mathcal{P}(\mathbb{N})$ is of the same size as the set of real numbers on the other. Let me first explain why $\mathcal{P}(\mathbb{N})$ is not listable, the term I prefer, incidentally is "countable" in which category I also include the finite sets. So, here is why $\mathcal{P}(\mathbb{N})$ is not countable. If if were, then there would be a bijection, which is in particular a surjection, form $\mathbb{N}$ to $\mathcal{P}(\mathbb{N})$. But I claim that for any set X, and any map $f : X \to \mathcal{P}(X)$, f is not surjective.

Prof. Rotnac: Let me see now, you are saying that if $f : X \to \mathcal{P}(X)$ is any map, then there is some subset Y of X such that $Y \neq f(x)$ for all $x \in X$?

Prof Cantor: Admirably put, yes. Here is a set Y which is always missed by f, so to speak. Consider the subset Y_f of X which consists of those elements $x \in X$ such that x does not belong to $f(x)$.

Prof. Rotnac: That is confusing. Let me see, to each x in X we have an associated subset $f(x)$ of X and you are saying that you admit x to the set Y_f if and only if x itself does **not** belong to the subset $f(x)$ associated to it by f. Yes, I see your definition of Y_f is rather bizarre but gives a pefectly valid subset of X.

Prof Cantor: Very good. Now I claim that Y_f is not in the image of f, i.e. is not of the form $f(y)$ with $y \in X$. Why, you ask? Recall that for any given $x \in X$, x belongs to Y_f if and only if x does not belong to $f(x)$. Recall that y belongs to Y_f if and only if y does not belong to $f(y)$. So if we had $f(y) = Y_f$, then this last statement would become "y belongs to Y_f if and only if y does not belong to Y_f." This is absurd! Thus, for all $y \in X$, $f(y) \neq Y_f$ and so f is not surjective.

Prof. Rotnac: Just when I think you have ceased surprising me, you astound me with an even more devious proof! That is brilliant. Moreover, with this new proof, not only do you have **two** infinite sets of different size, but you can make as many infinite sets of different size as you want.

Prof. Cantor: I can? How can I?

Prof. Rotnac: Well, say you start with $\mathbb{N}$, then $\mathcal{P}(\mathbb{N})$ is more numerous, then $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ is more numerous still, $\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))$ is more numerous still and so on!

Prof. Cantor: Yes, of course, that is truly marvelous! Thank you my friend for that wonderful observation.

Prof. Rotnac: You had already thought of that yourself, hadn't you?

Prof. Cantor: To be truthful, yes, but after much thinking. Whereas you made the leap immediately, and I wanted to share in the excitement of your discovery.

Prof. Rotnac: One thing puzzles me still. How does $\mathcal{P}(\mathbb{N})$ compare with the set of real numbers? Which one is bigger?

Prof. Cantor: Actually, they are both of the same size!

Prof. Rotnac: Is that so? Ah yes, you already said that they were. But I don't see why.

Prof. Cantor: Well, first of all, there is a natural one-to-one correspondence between $\mathcal{P}(\mathbb{N})$ and $\text{Maps}(\mathbb{N}, \{0,1\})$, that is to say the set of all maps from $\mathbb{N}$ to the binary set $\{0,1\}$. Here is how to ....

Prof. Rotnac: Just a moment, forgive me for interrupting but I beg you do not tell me! I want to figure this out on my own. Let's see now. You are saying that to give a subset of $\mathbb{N}$ is the same as giving a map from $\mathbb{N}$ to $\{0,1\}$, eh? Hmm... to give a subset X of $\mathbb{N}$ means to admit certain natural numbers to X and exclude others. And to give a map from $\mathbb{N}$ to $\{0,1\}$ means that certain natural numbers will be assigned the number 1 and others will be assigned the number 0. I have got it! Being assigned a "1" can mean admit to X and being assigned a "0" means you are to be excluded from X. To be perfectly formal, here is a bijective map $F : \mathcal{P}(\mathbb{N}) \to \text{Maps}(\mathbb{N}, \{0,1\})$. If $X \subseteq \mathbb{N}$, we have to describe a map $F(X) \in \text{Maps}(\mathbb{N}, \{0,1\})$. We define $F(X) : \mathbb{N} \to \{0,1\}$ by putting $F(X)(n) = 1$ if $n \in X$ and $F(X)(n) = 0$ if $n \notin X$. The map F is bijective because, for example, it has an inverse $G : \text{Maps}(\mathbb{N}, \{0,1\})$ defined in a simple way. Given a map $f : \mathbb{N} \to \{0,1\}$, we let $G(f) = \{n \in \mathbb{N} \mid f(n) = 1\}$. Then $F(G(f))$ is clearly the map f and $G(F(X))$ is just the set X.

Prof. Cantor: Masterfully done, I am sure. So, as I was saying, $\mathcal{P}(\mathbb{N})$ is in one-to-one correspondence with $\text{Maps}(\mathbb{N}, \{0,1\})$ and the set of maps from $\mathbb{N}$ to $\{0,1\}$ is in one-to-one correspondence with the interval $[0,1]$.

Prof. Rotnac: There you go confusing me again, just fresh from my triumph. Give me a moment to think...You have a string of 0's and 1's and you wish to assign to it a number between 0 and 1 in such a way that every real number is covered once. Hmm... If you had a string of digits between 0 and 9 then I would just say write out the decimal expansion and that would do the trick, but you are giving me only 0's and 1's.

Prof. Cantor: Pray remind me, what is the decimal expansion?

Prof. Rotnac: I assume you are playing dumb for my benefit. Very well, I will play along. If $a_1, a_2, a_3, \ldots$ is a sequence of digits from the set $\{0, 1, \ldots, 9\}$, then we get a corresponding decimal expansion $0.a_1a_2a_3\cdots$ which is defined to be the real number $a_1/10 + a_2/100 + a_3/1000 + \cdots$ and this converges to a number in the interval $[0,1]$. Every number in $[0,1]$ can be represented as a decimal expansion in exactly one way, with $0 = 0.000000\cdots$ and $1 = 0.99999\cdots$.

Prof. Cantor: Excellent. You are nearly there, meine Kollege. I am only giving you two symbols instead of ten, so instead of representing numbers in base ten (decimal expansion), you should represent them ...

Prof. Rotnac: In Base Two, natürlich! That's it. Allow me describe it in complete detail. There is a bijective map α from $\text{Maps}(\mathbb{N}, \{0,1\})$ to $[0,1]$ defined as follows: Given a map $f : \mathbb{N} \to \{0,1\}$, we associate a number $\alpha_f \in [0,1]$ to it by putting $\alpha_f = f(1)/2 + f(2)/4 + f(3)/8 + \cdots + f(n)/2^n + \cdots$. This series converges to a number in $[0,1]$. Note that the smallest possible α_f occurs for the function f_0 with constant value 0 giving $\alpha_{f_0} = 0$ and the greatest possible value occurs for the function f_1 with constant value 1 giving $\alpha_{f_1} = 1/2 + 1/4 + \cdots = 1$. To show that every real number in between 0 and 1 is in the image of α, i.e. has a binary expansion, we must construct a map $\beta : [0,1] \to \text{Maps}(\mathbb{N}, \{0,\})$ which just gives the binary (instead of the decimal) expansion of a real number. Namely, if $x \in [0,1]$, we define a map $\beta_x : \mathbb{N} \to \{0,1\}$ as follows. We know that $0 \leq x \leq 1$ so we try to determine which half of this range contains x. Namely, if $x < 1/2$, we put $x_1 = 0$ and if $x \geq 1/2$, we put $x_1 = 1$. Now we have $0 \leq x - x_1/2 \leq 1/2$, so we ask which half of this interval $x - x_1/2$ belongs to, i.e.

if $x - x_1/2 < 1/4$, we put $x_2 = 0$, otherwise we put $x_2 = 1$. Then if $x - x_1/2 - x_2/4 < 1/8$ we put $x_3 = 0$, otherwise $x_3 = 1$. In general, having found x_i for $1 \leq i \leq n$, we define x_{n+1} by putting $x_{n+1} = 0$ if $x - x_1/2 - x_2/4 - \ldots - x_n/2^n < 1/2^{n+1}$ and $x_{n+1} = 1$ otherwise. It is easy to see that $x - x_1/2 - x_2/4 - \cdots - x_n/2^n \leq 1/2^n$ and $1/2^n$ goes to 0 of course as n becomes large, hence the series $x_1/2 + x_2/4 + \cdots$ converges to x by construction. If we define $f_x : \mathbb{N} \to \{0, 1\}$ by $f_x(n) = x_n$, then $x = f(1)/2 + f(2)/4 + f(3)/8 + \cdots + f(n)/2^n + \cdots$, so that with $\beta(x) = f_x$, we have β is the inverse of α, and so α is bijective.

Prof. Cantor: Dotting the i's and crossing the t's perfectly as usual. So now we see that $\mathcal{P}(\mathbb{N})$ is in one-to-one correspondence with the closed interval $[0, 1]$ of real numbers. Now that we have seen infinite sets come in different sizes, I have begun to give names to sizes which infinite sets can take, I call these sizes "cardinal numbers" or "cardinalities." Sets which are in one-to-one correspondence with $\mathbb{N}$ I call of size $\aleph_0$. Those which are in one-to-one correspondence with $\mathcal{P}(\mathbb{N})$ I call of size $\aleph_1$. Those which are in one-to-one correspondence with $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ I call of size $\aleph_2$ and so on.

Prof. Rotnac: So the set of real numbers is of which cardinality?

Prof. Cantor: Of cardinality $\aleph_1$. For it is easy to see that the set of all real numbers is in one-to-one correspondence with those in the interval $(0, 1)$ and the latter is in one-to-one correspondence with $[0, 1]$. I have two different proofs that $[0, 1]$ is more numerous than $\mathbb{N}$, one via what you call the diagonal argument, and the other based on the fact that $[0, 1]$ is equivalent to $\mathcal{P}(\mathbb{N})$ and no surjective map $\mathbb{N} \to \mathcal{P}(\mathbb{N})$ can exist. The two proofs are in fact related in a simple way.

Prof. Rotnac: Which one is your favorite?

Prof. Cantor: You might imagine that I am more fond of the second, because it allows me to see that there is a whole infinite sequence of infinite cardinals, but the diagonal argument allowed me for the first time to show that infinite sets come in different sizes, so it is particularly dear to me.

Prof. Rotnac: My dear friend, you have uncovered eine grundsätzliche Wahrheit, a fundamental truth of nature, which has eluded our science for thousands of years! This is heady stuff, Georg! Have you written to our friend Richard about your findings?

Prof. Cantor: I am so pleased that you are in agreement with my results and methods. I have written to Herr Dedekind some of my discoveries, but now that I have your reassurance, I will write to him an expanded version of my results. Speaking of heady stuff, here we are in front of our favorite pub, Wirsthaus Mozart. Shall we go in for a celebratory pint of something cold?

Prof. Rotnac: We shall do more than that, meine freund! We will go on a Bierreise machen, a pub crawl to celebrate your results. And I will pay!

Definition 15.4. A set X is called *countable* if it is either finite or equivalent to $\mathbb{N}$. A set is called uncountable if it is not countable. An infinite countable set, i.e. one which is equivalent to $\mathbb{N}$, is also sometimes called countably infinite and is said to have size α_0 (read "aleph nought" or "aleph null.") A set which is equivalent to $\mathcal{P}(\mathbb{N})$ is said to have size $\aleph_1$.

Lemma 15.5. *A set X is countable if there exists a single infinite list (with or without repetition) which contains all the elements of X.*

Proof. First suppose X is countable. If it is finite, then certainly its elements can be listed. If it is infinite, then by definition, there is a bijection $f : \mathbb{N} \to X$ hence $f(1), f(2), \ldots$ is a single infinite list containing all the elements of X (with no repetition in fact). In the other direction, suppose $x_1, x_2, x_3, \ldots$ is an infinite list (possibly with repetition) with the property that for all $x \in X$, there exists $n \in \mathbb{N}$ such that $x = x_n$. If X is finite, then it is certainly countable. Now suppose X is infinite. From the list $x_1, x_2, \ldots$, we easily produce a list *without* repetition that contains all the elements of X. Namely, let $n_1 = 1$ and put $y_1 = x_1$. Let $y_2 = x_{n_2}$ where $n_2 = \min\{n > n_1 | x_n \notin \{y_1\}\}$. Let $y_3 = x_{n_3}$ where $n_3 = \min\{n > n_2 | x_n \notin \{y_1, y_2\}\}$. Similarly, having defined $n_1, n_2, \ldots, n_k$ and y_k as above, we define $n_{k+1} = \min\{n > n_k | x_n \notin \{y_1, y_2, \ldots, y_k\}\}$ and put $y_{k+1} = x_{n_{k+1}}$. Then the map $f : \mathbb{N} \to X$ defined by $f(m) = y_m$ is a bijection, so X is countable. □

Theorem 15.6. *(a) If $X_1, \ldots, X_N$ is a finite collection of countable sets, then $X_1 \cup \cdots \cup X_N$ is countable.*

(b) If $X_1, X_2, \ldots$ is a countably infinite collection of countable sets, then $\cup_{n=1}^{\infty} X_n$ is countable.

The two parts can be summarized by saying that "a countable union of countable sets is countable."

Proof. The proof of both assertions was given in the discussion between Cantor and his friend above. □

In practice, mathematicians encounter finite sets, countable sets, and sets of size $\aleph_1$ on an everyday basis. Sometimes when we want to compare two infinite sets to see if there could be some kind of relationship between them, it's a good first question to ask if they are both countable.

Recall that for infinite sets, we have given a meaning to the equality $|X| = |Y|$, namely this means that $X \sim Y$, i.e. there is a bijection from X to Y. We will also define a partial ordering on the size of sets, as follows.

Definition 15.7. If X and Y are sets, we write $|X| \leq |Y|$ if there exists an injective map $X \hookrightarrow Y$. We write $|X| < |Y|$ if there exists an injective map $X \hookrightarrow Y$ but no bijective map $X \xrightarrow{\sim} Y$ exists.

Note that the partial ordering we have defined on sizes of sets is transitive. Namely, suppose $|X| \leq |Y|$ and $|Y| \leq |Z|$. Then does it follows that $|X| \leq |Z|$? Yes, for suppose $f : X \hookrightarrow Y$ and $g : Y \hookrightarrow Z$ are injections. Then $g \circ f$ injects X into Z. But now suppose $|X| \leq |Y|$ and $|Y| \leq |X|$. We would expect in this case that $|X| = |Y|$. In other words, if X injects into Y and Y injects into X, then there is a bijection from X to Y. Cantor was not able to show this, but in 1898, Schroeder and Bernstein succeeded in demonstrating this property. Here is their theorem.

Theorem 15.8 (Schroeder-Bernstein)**.** *(a) Suppose $B \subseteq A$ and there is an injection $f : A \to B$. Then there is a bijection $h : A \xrightarrow{\sim} B$.*

(b) Suppose there is an injective map $f : X \hookrightarrow Y$ and an injective map $g : Y \hookrightarrow X$. Then there is a bijection $h : X \xrightarrow{\sim} Y$.

(c) Suppose $|X| \leq |Y|$ and $|Y| \leq |X|$. Then $|X| = |Y|$.

Proof. The proof that follows was adapted from `http://planetmath.org`

(a) Inductively define a sequence (C_n) of subsets of A by $C_0 = A \setminus B$ and $C_{n+1} = f(C_n)$. Then the C_n are pairwise disjoint. We will prove this by contradiction. Suppose the countable collection of sets $C_0, C_1, C_2, \ldots$ is not pairwise disjoint. Let

$$Z = \{j \geq 0 | C_j \cap C_k \neq \emptyset \text{ for some } k > j\}.$$

We have assumed that Z is not empty and are seeking a contradiction. By the Well-Ordering Principle, since Z is non-empty, it has a least element, call it m. Let $k > m$ be the least integer larger than m such that $C_m \cap C_k \neq \emptyset$; it exists because $m \in Z$. Since C_0 is disjoint with any following C_n, we have $0 < m$ and thus that $C_m = f(C_{m-1})$ and $C_k = f(C_{k-1})$. But this implies that $f(C_{m-1} \cap C_{m-1}) \neq \emptyset$ and so $C_{m-1} \cap C_{k-1}$ cannot be empty, hence $m-1 \in Z$, contradicting the minimality of m. This contradiction shows that the collection $C_0, C_1, \ldots$ is pairwise disjoint.

Now let $C = \bigcup_{k=0}^{\infty} C_k$, and define $h : A \to B$ by

$$h(z) = \begin{cases} f(z), & z \in C \\ z, & z \notin C \end{cases}.$$

If $z \in C$, then $h(z) = f(z) \in B$. But if $z \notin C$, then $z \in B$, and so $h(z) \in B$. Hence h is well-defined; h is injective by construction. Let $b \in B$. If $b \notin C$, then $h(b) = b$. Otherwise, $b \in C_k = f(C_{k-1})$ for some $k \geq 0$, and so there is some $a \in C_{k-1}$ such that $h(a) = f(a) = b$. Thus b is bijective; in particular, if $B = A$, then h is simply the identity map on A.

Now (b) is a simple consequence of (a). Suppose $f : X \to Y$ and $g : Y \to X$ are injective. Then the composition $g \circ f : X \to g(Y)$ is also injective. By the lemma, there is a bijection $h' : X \to g(Y)$. The injectivity of g implies that $g^{-1} : g(Y) \to Y$ exists and is bijective. Define $h : X \to Y$ by $h(z) = g^{-1} \circ h'(z)$; this map is a bijection.

For (c), we simply apply (b) to conclude that if X injects into Y and Y injects into X, then X and Y have the same cardinality. □

Note that if X is an infinite set, then $\mathbb{N}$ injects into X, hence for all infinite sets, $|\mathbb{N}| \leq |X|$. The picture that emerged from Cantor's research is the following schematic of sets classified according to their size.

$\{\}$	$\{1\}$	$\{1,2\}$	$\{1,2,3\}$	$\cdots$	$\mathbb{N}$	$\mathcal{P}(\mathbb{N})$	$\mathcal{P}(\mathcal{P}(\mathbb{N}))$	$\cdots\cdots$
0	1	2	3	$\cdots$	$\aleph_0$	$\aleph_1$	$\aleph_2$	$\cdots\cdots$

Cantor had established that $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are of size $\aleph_0$ and that $\mathcal{P}(\mathbb{N}), \mathbb{R}$ and $[0,1]$ are of size $\aleph_1$. Cantor believed that there were no sets of size strictly between $\aleph_0$ and $\aleph_1$, in other words, he believed that if X is an uncountable subset satisfying $|\mathbb{N}| < |X| \leq |\mathbb{R}|$, then X is equivalent to $\mathbb{R}$. Equivalently, he believed that every uncountable subset of $\mathbb{R}$ is equivalent to $\mathbb{R}$. This statement became known as the Continuum Hypothesis. Following fundamental work of Curt Gödel in 1940 and that of Paul Cohen in 1962, the Continuum Hypothesis was found to be in a most mysterious class of problems, those which are "undecidable." What this means in practice, is that believing the Continuum Hypothesis to be true does not lead to any contradictions within the totality of all logical consequences of the standard axioms of set theory, and the same can be said of believing the Continuum Hypothesis to be false! In other words, there are two perfectly consistent paradigms in which to do set theory: one in which we accept the standard axioms of set theory plus the Continuum Hypothesis and another in which we accept the standard axioms of set theory plus the negation of the

Continuum Hypothesis! In practice, most problems that mathematicians work on are not affected one way or another by which of these these two set-theoretical paradigms we work in, because the sets we usually deal with are, so to speak, too small. However, in the field of mathematical logic, one of the objects of study to consider ramifications of adopting one set of logic axioms versus another. Please note, with my apologies!, that I grossly misrepresented the status of knowledge concerning the Continuum Hypothesis in class, although I stated the Hypothesis itself correctly.

Part 9. Complex Numbers

16. In the beginning ...

In the beginning, there is the number 1. Then $1+1$ makes 2, $1+2$ makes 3, and the rest is history. We get all the positive whole numbers. After a while, we ponder the reverse of this "adding 1" process and discover the "take away 1" process, which gives us $3-1$ is 2, $2-1$ is 1, $1-1$ makes ... a new and wondrous number, namely 0. Moreover, $0-1$ makes -1, -1 minus 1 makes -2 and so on giving us the negative numbers. The new system of numbers,

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$$

has many wonderful qualities and internal relationships. For instance, -2 is defined to be $-1-1$ but it is also $(-1)+(-1)$. There are quite a few popular books which discuss the history of 0 and negative numbers. According to an ancient tradition of mathematics, which lasted through the fifteenth and sixteenth centuries in European mathematical circles (!), negative solutions of simple equations were not accepted as "proper" well-behaved solutions and were discarded at the end of the solving process. These days, of course, any accountant knows that negative numbers cannot be simply discarded ... or do they? [Can you say "Enron?" How about "Worldcom?"]

Once one chooses to accept the system $\mathbb{Z}$ of integers as a legitimate collection of numbers, it is difficult to avoid noting its charms. There they are, the integers, marching off to infinity at perfectly regular discrete intervals in two directions, with a neutral point 0 smack in the middle perfectly balancing out the positive and negative integers. As we discussed earlier, there are two basic *simply fantasmagoric, luscious* binary operations defined on $\mathbb{Z}$, namely $+$ and $\times$. What this implies is that for any fixed integer $k \in \mathbb{Z}$ we can attach two *actions* τ_k and δ_k corresponding to addition by k and multiplication by k. We use τ_k to stand for translation by k and δ_k for dilation by k in order to emphasize the geometric meaning of these operators: τ_k shifts all the integers k places (to the right if $\text{sign}(k) = +1$ and to the left if $\text{sign}(k) = -1$), and δ_k dilates everything (streches $\mathbb{Z}$) by a factor of k (if $\text{sign}(k) = -1$ the there is a stretch by the factor $|k|$ followed by a 180-degree rotation of the line). To give a precise algebraic definition, for each $k \in \mathbb{Z}$, we have two self-maps τ_k, δ_k of $\mathbb{Z}$, namely

$$\begin{aligned} \tau_k : \mathbb{Z} \to \mathbb{Z} \text{ defined by} \qquad & \tau_k(n) = k + n \text{ for all } n \in \mathbb{Z} \\ \delta_k : \mathbb{Z} \to \mathbb{Z} \text{ defined by} \qquad & \delta_k(n) = kn \text{ for all } n \in \mathbb{Z}. \end{aligned}$$

Note that τ_k is a bijection for all $k \in \mathbb{Z}$. But δ_k is a bijection for only *very few* integers k. Which ones?! Well, the map δ_k is injective as long as $k \neq 0$ (prove it!). But the image of δ_k is what we have been calling $k\mathbb{Z}$, namely the set of all multiples of k. The latter is all of $\mathbb{Z}$ if and only if $k = \pm 1$. Note that ± 1 are the only integers which have no prime divisors (they are "units"). They are also the only ones whose multiplicative reciprocal belongs to $\mathbb{Z}$.

Any algebraic manipulation with integers can be reinterpreted in terms of the "geometry" of the maps $\tau : \mathbb{Z} \to \text{Maps}(\mathbb{Z}, \mathbb{Z})$ and $\delta : \mathbb{Z} \to \text{Maps}(\mathbb{Z}, \mathbb{Z})$. For instance, why were we motivated to create the negative numbers in the first place? Because the translation map τ_1 (or rather its restriction to $\mathbb{N}$[24]) naturally wants to have an inverse map. Now $\tau_1|_{\mathbb{N}}$ doesn't

[24]Whenever $f : X \to Y$ is a map, and S is a subset of X, we get a map $S \to Y$ in a simple way, namely by restricting the "domain" to S. In other words, we define the *restriction of f to S*, denoted $f|_S$ by $f|_S : S \to Y$, $s \mapsto f(s)$ for all $s \in S$.

quite have an inverse because 1 is not in its image! Thus, we are led to creating a symbol 0 which serves as $\tau_1^{-1}(1)$. One then proves that τ_0 is the identity map, i.e. $0+n=0$ for all n. By the same process, $-1=\tau_1^{-1}(0)$, and now one proves that $\tau_{-1}=\tau_1^{-1}$!

Note that for any integer k, τ_k is the compositum of k copies of τ_1, resp. τ_{-1}, if k is positive, resp. negative.

Recalling how the lack of surjectivity of τ_k led to the creation of negative numbers, we recall that the maps $\delta_k:\mathbb{Z}\to\mathbb{Z}$ are injective for $k\neq 0$, but they turn out not to be surjective for $|k|>1$. So, for $|k|>1$, and $n\in\mathbb{Z}$, we want an element $\delta_k^{-1}(n)$, which is in general not in $\mathbb{Z}$. So, we "create" a new set $\mathbb{Q}$ which is the smallest set which fills in the "missing" numbers. Namely, we create a set $\mathbb{Q}$ by defining on the set $\mathbb{Z}\times\mathbb{Z}_0$[25] the equivalence relation $(a,b)\sim(c,d)$ if and only if $ad-bc=0$, then we put $\mathbb{Q}=(\mathbb{Z}\times\mathbb{Z}_0)/\sim$ for the set of equivalence classes. We think of the equivalence class $\widetilde{(a,b)}$ as the fraction a/b. We have a natural injection $\mathbb{Z}\hookrightarrow\mathbb{Q}$ given by $n\mapsto\widetilde{(n,1)}$. In this way, we "think of" $\mathbb{Z}$ as a subset of $\mathbb{Q}$. On $\mathbb{Q}$, we know how to define addition and multiplication by the "think-of-them-as-fractions" yoga, namely

$$\widetilde{(a,b)}+\widetilde{(c,d)}=\widetilde{(ad+bc,bd)},\qquad \widetilde{(a,b)}\times\widetilde{(c,d)}=\widetilde{(ac,bd)}.$$

Now the maps $\tau_k,\delta_k:\mathbb{Q}\to\mathbb{Q}$ is easy to define, just as before as addition and multiplication by k maps, with the advantage, now, that τ_k is a bijection of $\mathbb{Q}$ to itself for all k and δ_k is a bijection of $\mathbb{Q}$ to itself for all $k\neq 0$. The fact that we cannot make τ_0 into a bijection by extending its field of definition is due to its being so *horrifically* (is that a word?) non-injective!

We discussed earlier in the semester that the Greeks were quite happy with their system of numbers ($\mathbb{Q}$) until they discovered that the equation $x^2-2=0$ does not have a solution in this number system. The Greeks, knew, however, that the quantity $\sqrt{2}$ can be approximated as well as one wishes by a sequence of rational numbers, e.g. $1, 1.4, 1.41, 1.414, \ldots$. What is amiss is that this sequence does not have a limit in the set $\mathbb{Q}$ itself. Thus, what is needed is to have a number system in which all sequences of rational numbers which "should" tend to an acutal number *do* have a limit. Thus, another extension of their number system was required. The eventual solution was to write rational numbers in decimal expansion, and then note that there are lots of decimal expansions that do not express rational numbers, because the decimal expansion of a rational number always ends in a repeating finite pattern. Thus, the set of all real numbers, $\mathbb{R}$ is defined to be the set of numbers expressible as a decimal expansion. One then shows that any sequence of real numbers which "should" have a limit does indeed have a limit. (The concept of "should have a limit" which I am leaving vague here can be made precise of course and goes under the name of "Cauchy sequence" for the fabulous mathematician Auguste Louis Cauchy. You will learn more about this in Math 523).

Thus, the set of real numbers, $\mathbb{R}$ is "complete," it has all the properties that one would want. Translation by a real number is invertible. Multiplication by a real number other than 0 is invertible. The real numbers have no "wholes," meaning any sequence of reals that are getting closer and closer together actually converge to a real number. Moreover, Newton showed how to find real solutions of cubic, fifth degree, and higher odd degree polynomial equations by an "iteration" procedure. But the seeming "perfection" of the real number

[25]Recall that $\mathbb{Z}_0=\mathbb{Z}\setminus\{0\}$

system is tarnished by "only" one minor defect, namely, the simple equation $x^2 + 1 = 0$ does not have any roots! Thi story goes back thousands of years: from ancient times, we have known that there is a simple "quadratic formula" for solving all quadratic equations, namely, if

$$ax^2 + bx + c = 0, \qquad a, b, c \in \mathbb{R}, a \neq 0,$$

then we put

$$\Delta = b^2 - 4ac, x_1 = \frac{-b + \sqrt{\Delta}}{2a}, \qquad x_2 = \frac{-b - \sqrt{\Delta}}{2a},$$

and then $ax^2 + bx + c = a(x - x_1)(x - x_2)$, so that x_1, x_2 are the roots of the equation.

Great. The only trouble is how to interpret the symbol $\sqrt{\Delta}$ when Δ is a negative real number. In other words, how does one solve the equation $y^2 - \Delta = 0$ when $\Delta < 0$? If y is a real quantity, then $y^2 \geq 0$ so y^2 simply cannot equal $\Delta < 0$. The easy answer is to say, Dude, you just proved these equations don't have solutions, so just let it rest man! However, if you've followed the "arc" of the story so far, you will have noticed that almost every time that a certain "solution" did not seem to exist to a simple problem, there was a way to create a larger set within which a solution does exist and this new set is a bigger and better place to do mathematics. So the ambitious answer would be: Dude, you have shown there are no **real** numbers y that satisfy the equation $y^2 = \Delta$ when $\Delta < 0$, so are there some **"non-real"** numbers that do satisfy the equation?!

Historically what happened is that people came to realize that if they sometimes allowed the use of *symbols* such as $\sqrt{-1}$ in their calculations, as long as these symbols were handled with care, then they could be manipulated in the usual way and (this next bit was very important historically) they could use these manipulations to find *real* solutions of other equations! This was a great advantage to those who braved the new world of "imaginary" numbers as they came to be called. There is a fascinating story here [for which I highly recommend the book "Imagining Numbers" by Professor Barry Mazur as summer reading for all of you], but let's move on to describing the set $\mathbb{C}$ of complex numbers, a place where all polynomial equations have solutions!

Where to start? We want to solve $y^2 = \Delta$ where Δ is negative, so how about we try $y^2 = -1$ for starters. We know that y cannot be a real quantity, so we just invent a symbol (traditionally the symbol used is i and we will stick with that) and stipulate that this symbol designates a fixed object with the property that $i^2 = -1$. We will then want this i to interact with real numbers in reasonable ways, so for example, we should be able to add 5 to i to get an object $i + 5$. Also we should be able to multiply by real numbers, say $5i$, $-i$ should be admitted to our system of numbers. We will want some standard commutative/associative rules, so that, for example, $i + i = 2 \cdot i = i \cdot 2$, $2 \cdot (3i) = 6i$ etc. In short, we want to define a new system of numbers to be all those that are "generated" by the real numbers $\mathbb{R}$ and this one new symbol i but keeping the usual nice rules of addition and multiplication. So, we put

$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\},$$

and call this the set of complex numbers. Note that we can inject $\mathbb{R}$ into $\mathbb{C}$ via $a \mapsto a + 0i$. In this way, we think of $\mathbb{R}$ as a subset of $\mathbb{C}$. So, a complex number z is nothing other than a pair (a, b) of real numbers. What's the big deal about that? Well, as a set, maybe, that's what $\mathbb{C}$ is but on this set we now define some cool operations. Namely, given $z, w \in \mathbb{C}$, we write $z = a + bi$ and $w = c + di$, then we define $z + w = (a + c) + (b + d)i$, i.e. we add two complex numbers coordinate-wise, that's easy. Multiplication is a much niftier operation, so

let's be careful here. How should we define $zw = (a+bi)(c+di)$? Recalling that $i^2 = -1$ is its defining quality, its *raison d'être*, its *mantra*, its ... enough already, and that we want to maintain the usual algebraic rules, we compute

$$(a+bi)(c+di) = ac + adi + bic + bidi = ac + (ad+bc)i - bd = (ac-bd) + (ad+bc)i.$$

So that's what we do: on the set $\mathbb{C}$ we define

$$(a+bi) + (c+di) = (a+c) + (b+d)i, \qquad (a+bi)(c+di) = (ac-bd) + (ad+bc)i.$$

One then checks that all the usual algebraic rules do in fact hold!

Now, in this new wonderful set $\mathbb{C}$, what can we then say about the equation $y^2 = -1$? Well, naturally i is a solution of it. But even more than that, $-i$ is also a solution! Just plug it in and see:

$$(-i)^2 = (-i)(-i) = (-1)(-1)(i)(i) = i^2 = 1.$$

So we have $(y-i)(y+i) = y^2 - i^2 = y^2 + 1$, i.e. y^2+1 now factors completely and has exacly two roots in $\mathbb{C}$, namely $i, -i$. How do we know? Because if $zw = 0$ (where $z, w \in \mathbb{C}$), then either $z = 0$ or $w = 0$. We say that $\mathbb{C}$ has no *zero-divisors*.

The observant reader might now be thinking "Dude, like we had to do, like, *all this* just to solve, um, $y^2 = -1$?! And now we have to go through all these hoops again to solve, like, $y^2 = -2$, and $y^2 = -3$ etc.?" No, Dude, we don't. Watch: if $i^2 = -1$, and $\Delta = -r$ for some real number $r > 0$, then we know we have a positive real number $\sqrt{r}$, so we observe that $\pm i\sqrt{r}$ are two solutions of $y^2 = \Delta$. (As before, these are the only solutions in $\mathbb{C}$).

Consequently, we have the following theorem.

Theorem 16.1. *If $a, b, c \in \mathbb{R}$ with $\Delta = b^2 - 4ac = -r < 0$, then the equation $az^2 + bz + c = 0$ has exaclty two solutions in $\mathbb{C}$, namely*

$$z_\pm = \frac{-b \pm i\sqrt{r}}{2a}.$$

Much more is in fact true. It turns that if we consider any polynomial equation with coefficients of any degree $n \geq 1$ in $\mathbb{C}$, then it will always factor completely into linear factors over $\mathbb{C}$. This is known as the fundamental theorem of algebra (although it is an analytic fact!), and the first proof of it was given by Gauss in 1799. Keep in mind Gauss was born in 1777. Moroever, the field $\mathbb{C}$ does not have any "holes" meaning it is complete, meaning any sequence in $\mathbb{C}$ which ought to converge does in fact do so. Thus, the set of complex numbers is some kind of heavenly place to work with numbers.

Theorem 16.2 (Fundamental Theorem of Algebra)**.** *Suppose $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 x$ where $a_k \in \mathbb{C}$ for $k = 1, \ldots, n$, and $a_n \neq 0$. Then there exist integers $r, s \geq 0$ with $r + 2s = n$, real numbers $\alpha_1, \ldots, \alpha_r$ and non-real complex numbers $\beta_1, \ldots, \beta_s$ such that*

$$f(x) = a_n(x-\alpha_1)\cdots(x-\alpha_r)\cdot(x-\beta_1)\cdots(x-\beta_s)\cdot(x-\overline{\beta_1})\cdots(x-\overline{\beta_s}).$$

The numbers r, s as well as the unordered sequences $\alpha_1, \ldots, \alpha_r$ and $\beta_1, \cdots, \beta_s$ are uniquely determined by f.

You will probably study a proof of this wonderful theorem in Math 421 using the concept of differentiablity for a function $f : \mathbb{C} \to \mathbb{C}$. Just to give a very simple example, suppose $w \in \mathbb{C}$ and $w \neq 0$. Then the equation $z^2 - w$ has two roots. Namely, if we write $w = re^{i\theta}$, with $r > 0$, then with $z_\pm = \pm\sqrt{r}e^{i\theta/2}$, we have $z^2 - w = (z - z_+)(z - z_-)$.

17. A constructive existence theorem

Some of you might feel lulled into a sense of complacency by the previous section, culminating in Theorem 16.1. Others might be feeling somewhat unsatisfied or unconvinced on the point of the "reality" of the solution there proposed. Oh Oh, here comes another one of those `$&%(*^&$%` plays!

==========

The reality of imaginary numbers

Javier: Hey, professor Dude. So you're saying you can solve an equation like $z^2 = -1$ by "**inventing**" a symbol i and calling that the solution. What kind of a cop-out is that? I thought this course was supposed to be about proving/justifying everthing that we do logically. How can you just say "Oh happy me, let me just *invent* a solution, la di da!" and expect us to accept that?! What would you do if I just "invented" answers and gave them names to whatever equation you gave me on an exam?

Kristen: Way harsh.

Farshid: No, no, Javier has a valid point here. The point is: fine, maybe wonderful things will happen if I just invent a solution to some equation, but until I *prove* that a solution does exist, it is a valid question to ask how I know that the equation $z^2 = -1$ has a solution. This issue was actually the subject of bitter controversy in the 17th and 18th centuries. Gauss, for instance, was very critical of how his predecessors just posited (philosophers love that word) the existence of solutions and went merrily along playing with these posited (invented out of thin air) solutions. Gauss was the first person to put the complex numbers on firm ground and prove what we now call The Fundamental Theorem of Algebra (more on that later).

Matt T.: So you're saying that complex numbers don't really exist?

Farshid: Oh not at all, I'm just saying I haven't yet proved that they exist. I will do so now.

Alby: No way, dude, there is no way you can do it.

Farshid: Let me try and the class, including you, can be the judge of that. Before I start, let's agree on what it is I'm setting out to do. We all agree that the equation $z^2 = -1$ has no solution with $z \in \mathbb{R}$. What I want to construct is a set, let me call it $\mathbf{C}$ for now, which contains $\mathbb{R}$ but also contains a special element i such that $i^2 = -1$. Actually what I will do is slightly different. I will construct a set $\mathbf{C}$ which does not contain $\mathbb{R}$ strictly speaking, but an exact copy $\mathbf{R}$ of the set of real numbers written in a slightly unusual way. To explain what I mean, recall when we constructed the rational numbers as $\mathbb{Z} \times \mathbb{Z}_0/\sim$ i.e. as equivalence classes $\widetilde{(a,b)}$ of pairs (a,b) of integers with $b \neq 0$ with $(a,b) \sim (c,d)$ if and only if $ad = bc$? Well the set $\mathbb{Z}$ is not strictly speaking a subset of $\mathbb{Q} = \mathbb{Z} \times \mathbb{Z}_0/\sim$ now, is it? But we have a natural injection $\mathbb{Z} \hookrightarrow \mathbb{Q}$ by sending $a \mapsto \widetilde{(a,1)}$. In the same way, we we will have a very natural injective map $\mathbb{R} \hookrightarrow \mathbf{C}$ whose image $\mathbf{R}$ we will think of as a copy of the real numbers.

Brent: I think, with that explanation, you've confused the heck out of half the class and put the other half to sleep.

Farshid: You're right, sorry, let me just spit it out. Alright, here we go. You remember matrices, right? Let's look at the set $\mathbf{M}$ of all 2×2 matrices with real entries:

$$\mathbf{M} = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mid a, b, c, d \in \mathbb{R}\}.$$

I want to look at those matrices that satisfy $a = d$ and $b = -c$.

Matt L.: Why? That seems like a whacko idea.

Farshid: You'll see. So let me define $\mathbf{C}$ to be the matrices that have repeating diagonal entries and whose "anti-diagonal" entries are negatives of each other, i.e.

$$\begin{aligned} \mathbf{C} &= \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbf{M} \mid c = -b, d = a\} \\ &= \{\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right) \mid a, b \in \mathbb{R}\}. \end{aligned}$$

If we define a map $m : \mathbb{R}^2 \to \mathbf{M}$ by $(a, b) \mapsto m_{a,b}$ where

$$m_{a,b} = \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right),$$

then $\mathbf{C}$ is simply the image of the map m. You will admit, I hope, that the existence of the set $\mathbf{C}$ is not in question.

Class: (Grudingly) Granted. Get to the point, Dude.

Farshid: Thank you. Next, I want to show you that the real numbers live very happily inside $\mathbf{C}$, can you guess how, Kate?

Kate: I guess the real numbers correspond to matrices with zeros in the "b" and "$-b$" position.

Farshid: Brilliant, yes! How did you come up with that guess?

Kate: Well, I went with the coincidence-of-notation theory. In other words, I ...

Farshid: (interrupting) Wait, what was that? What's the "coincidence- of-notation theory?"

Kate: What I mean is that the set $\mathbf{C}$ you have defined is supposed to be like the complex numbers, right?

Farshid: Yeah, so?

Kate: And you said the elements of $\mathbb{C}$ (what you talked about before) are $a + bi$ with real numbers a, b. But now the elements of $\mathbf{C}$ are matrices $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$. Notice how the same letters are used? Coincidence? I personally don't think so. I think in some twisted way $a + bi$ the complex number is associated with $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ the matrix. So, to answer your question about which of these matrices corresponds to a real number, I went back to ask the same about the $a + bi$ and it's pretty obvious which of those are real numbers, the ones with $b = 0$. So then going back the matrices, I guess the "real ones" are the ones with $b = 0$ i.e. with zeros on the anti-diagonals.

Farshid: Wow, that is great detective work and perfectly correct in its findings, although I suspect the "coincidence-of-notation" theory can lead you down the road of confusion and false turns too. But in this case, I gotta hand it to you, you nailed it. We do define $\mathbf{R}$ to be the set of matrices

$$\begin{aligned} \mathbf{R} &= \{\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right) \in \mathbf{C} \mid b = 0\} \\ &= \{\left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right) \mid a \in \mathbb{R}\}. \end{aligned}$$

In other words, $\mathbf{R}$ is the image of the (clearly injective) map $\mathbb{R} \hookrightarrow \mathbf{C}$ defined by $a \mapsto \left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)$.

Julie: Will this discussion ever come back to the equation $z^2 = -1$?

Farshid: Yes, very soon. Good point. So how do we express the usual number -1 in this weird matrixy world, Julie?

Julie: I guess as $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$?

Farshid: Yes, so we look for a matrix Z (of the right type, i.e. with negative antidiagonals and equal entries on the diagonal) such that when you multiply it (matrix multiplication) by itself you get the matrix Julie just said. Can anyone see such a matrix?

Class: Silence.

Farshid: Okay, if I gave you ten minutes, even five, even three, to work together, and you really went at it, I'm sure you'd find it by experimentation, but that would be too obvious and boring a way to get the answer. Any other ideas for how to proceed?

Jonah: Oh come on man. For Pete's sake, you know the matrix, just tell us what it is!

Farshid: Okay, you're right, it's $\left(\begin{smallmatrix} -2342+\sqrt{5} & 19497+\sqrt{5} \\ -19497-\sqrt{5} & -2342+\sqrt{5} \end{smallmatrix}\right)$. Would you mind just squaring that matrix and verifying that it gives $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$?

Jonah: (Annoyed, gets to work, nonetheless).

Farshid: While Jonah is working on that ...

Shaohan: I have an idea, which, in principle will work. To ease the notation, recall that the matrix $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ has the name $m_{a,b}$. Then we want to find a, b such that $m_{a,b}^2 = m_{-1,0}$.

Farshid: (Hiding his glee) Yes, go on.

Shaohan: The equation $m_{a,b}^2 = m_{-1,0}$ will give us 4 equations, one for each of the entries. We can try to solve those four equations to find the a and the b.

Farshid: Excellent, you and Ashley and Jing work on that and see what you get. Jonah, how's it going?

Jonah: Still working on it, thanks to you jabbering away.

Farshid: (Not sorry at all) Sorry, keep working.

Mike: Jen and I think we've figured out the answer.

Farshid: Did you just experiment, trial-and-errorifically? That would be too dull.

Mike: Well, not exactly. I really liked Kate's theory, so we went with that to try to guess the anwer; it's more like "steal-and-check" as opposed to "trial-and-error." Anyway, remember how Kate said she thinks $m_{a,b}$ is secretly the number $a + bi$?

Farshid: I believe she used the phrase "in a twisted way."

Mike: Yes, well, that's not the point is it? Anyhoo, if $m_{a,b}$ is $a + bi$, and we're looking for what matrix represents i right? Then we should

Jen C.: Look at the matrix $m_{0,1}$ because $i = a + bi$ with $a = 0$ and $b = 1$.

Ashley: Okay, we followed Shaohan's method and we got the matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, which is the matrix $m_{0,-1}$, as one solution.

Farshid: Jonah, any luck?

Jonah: Dude, I multiplied the whole darn thing out three times and I got the same answer three times and it doesn't give what you said!!

Farshid: (Feigning sorrow) Oh, sorry my man, my bad. But don't worry, we seem to be getting the right answer some other way. Let's note that $m_{0,1} = -m_{0,-1}$ so if one of them squared is $m_{-1,0}$ then so is the other one. Alright, let's check $m_{0,1}$ then:

$$m_{0,1}^2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right),$$

just as we wanted. Hurray. Notice that in the set $\mathbf{C}$, we have two solutions of $z^2 = -1$, namely $m_{0,1}$ and $m_{0,-1}$. So let us define $I = m_{0,1}$. This is very nice because then $m_{a,b} = a + bI$

with the usual operations of matrix addition and multiplying a matrix by a scalar. We have therefore settled the existence of a set $\mathbf{C}$ which contains a copy of the real numbers and in which the equation $z^2 = -1$ has a solution. Moreover, the usual algbraic operations on $\mathbb{R}$ extends to $\mathbf{C}$. In practice, we could continue to work with these matrices, but having now rigorously shown the existence of complex numbers (in a rather concrete way, in fact), we can now return with an easy conscience to the paradise of $\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}$; should we ever have any doubts, we can travel back to $\mathbf{C}$ to make sure our intuition is based on a firm foundation.

==========

18. The geometry of $\mathbb{C}$

We have seen that, geometrically speaking, the set $\mathbb{C}$ is just the plane $\mathbb{R}^2 = \{(a, b) \mid a, b \in \mathbb{R}\}$. Namely we have a very nice bijective map $\mathbb{R}^2 \to \mathbb{C}$ given by $(a, b) \mapsto a + bi$. Moreover on this set, there is a "vector space" rule for adding elements (geometrically it is the familiar "parallelogram rule") which is just gotten by adding coordinate-wise which is the usual way adding numbers in $\mathbb{R}^2$. On the set $\mathbb{C}$, however, we have defined a very cool *multiplication* rule which one probably would not have thought of if the point $(0, 1)$ had not been given the interpretation as the quantity $i = \sqrt{-1}$. This multiplication rule is commutative, associative and distributes over addition, i.e. $zw = wz$, $(zv)w = z(vw)$, and $z(v + w) = zv + zw$ for all $z, v, w \in \mathbb{C}$.

The map $\mathbb{R}^2 \to \mathbb{C}$ allows us to think of complex numbers as vectors in the plane. Continuing with the vector metaphor, If $z = a + bi$ with $a, b \in \mathbb{R}$, the real part of z is defined to be $\Re(z) = a$ and the imaginary part of z is defined to be $\Im(z) = b$. In other words, these are, respectively the $(1, 0)$ and $(0, 1)$ components of (a, b). We always have $z = \Re(z) + \Im(z)i$. We define the *modulus* (or *length*) of $z = a + bi$ to be

$$|z| = \sqrt{\Re(z)^2 + \Im(z)^2} = \sqrt{a^2 + b^2}.$$

By definition, the modulus of z is a non-negative real number, measuring its distance (in the usual sense) from the origin. The complex numbers with fixed real part lie on a vertical line, those with fixed imaginary part lie on a horizontal line, and those with fixed modulus lie on a circle centred at the origin. We have already encountered a symmetry of the complex numbers, the one where i and $-i$ are interchanged. We define, for $z = a + bi$, its *complex conjugate* to be $\overline{z} = a - bi$. We have

$$\Re(z) = \frac{z + \overline{z}}{2}, \qquad \Im(z) = \frac{z - \overline{z}}{2}.$$

Note that $|z|^2 = z\overline{z}$.

You might recall that an alternate method for locating points in the plane is called *polar coordinates*. It specifies a point z in the plane by giving a pair (r, θ) where r, θ are real numbers. Here, $|r|$ is the modulus of z (we allow r to be negative, in which case we move backwards along the ray indicated by θ) and θ is an angle, measured in radians, with $\theta = 0$ being the angle subtended by the x-axis and positive angles indicating a counterclockwise motion. In some text, when we want to indicate a complex number z whose polar coordinates are r and θ, one writes $z = r\mathrm{cis}\theta$. A little recollection of trigonometry suffices for noting that

$r\mathrm{cis}\theta = r(\cos\theta + i\sin\theta)$. If $z = |z|(\cos\theta + i\sin\theta)$, then the angle θ is called the *argument* of z, sometimes denoted $\arg z$. Note that if $z \neq 0$, $\arg z$ is determined only up to an integer multiple of 2π, and $\arg 0$ is not well-defined at all! For instance, Brian might say that "the" argument of i is $\pi/2$, and Rick might say that i has argument $-3\pi/2$. They would both be right.

Here is a wonderful and useful theorem.

Theorem 18.1. *For a complex number $z = r(\cos\theta + i\sin\theta)$ with $r, \theta \in \mathbb{R}$, we have $z = re^{i\theta}$.*

Euler's Proof. For this proof, I will assume that you are familiar with Taylor series from calculus. Recall the following lusciously and everywhere converging power series representations:

$$\begin{aligned} e^z &= \sum_{k=0}^{\infty} \frac{z^k}{k!} \\ \cos z &= \sum_{j=0}^{\infty} \frac{(-1)^j z^{2j}}{(2j)!} \\ \sin z &= \sum_{m=0}^{\infty} \frac{(-1)^k z^{2m+1}}{(2m+1)!}. \end{aligned}$$

It turns out that these power series converge and are valid even if the argument is a complex number. Before we plug in, let's take a peek ahead and notice that the expression i^k will appear in the formulas; this is a periodic function of period four, giving $1, i, -1, -i, 1, i, -1, -i, 1, \ldots$, so we will split our sum into the even k's giving alternating $1, -1$ and the odd k's giving alternating $i, -i$.

So, we plug in $z = i\theta$ into the first one and compute:

$$\begin{aligned} e^{i\theta} &= \sum_{k=0}^{\infty} \frac{i^k\theta^k}{k!} \\ &= \sum_{k\text{even}} \frac{i^k\theta^k}{k!} + \sum_{k\text{odd}} \frac{i^k\theta^k}{k!} \\ &= \sum_{j=0}^{\infty} \frac{i^{2j}\theta^{2j}}{(2j)!} + \sum_{m=0}^{\infty} \frac{i^{2m+1}\theta^{2m+1}}{(2m+1)!} \\ &= \sum_{j=0}^{\infty} \frac{(-1)^j\theta^{2j}}{(2j)!} + \sum_{m=0}^{\infty} \frac{(-1)^m i\theta^{2m+1}}{(2m+1)!} \\ &= \cos(\theta) + i\sin(\theta). \end{aligned}$$

Presto. □

Now let's pull a few rabbits out of our hat with this fabulous result. First of all, let's say z_1, z_2 are points on the unit circle, say $z = e^{i\theta_1}$ and $w = e^{i\theta_2}$. Remembering rules for how to multiply exponentials, we have $zw = e^{i(\theta_1+\theta_2)}$. In other words, multiplying two points on the unit circle keeps them on the unit circle but adds their arguments.

If $n \in \mathbb{Z}$, clearly $z^n = (e^{i\theta})^n = e^{in\theta}$, or $z^n = \cos n\theta + i\sin n\theta$, which seems to indicate that if you multiply z by itself n times, then that just makes it go n times further along the circle.

Note that if $t \in \mathbb{R}$, then $|e^{it}| = \sqrt{\cos^2 t + \sin^2 t} = 1$. Let us solve the equation $z^n = 1$ where $n \in \mathbb{N}$. Rewriting $z = re^{i\theta}$ with $r \geq 0$, we have $r^n e^{in\theta} = 1$ giving $r^n = 1$. Since $r \geq 0$, we conclude that $r = 1$. Moreover, $\arg(1) = \arg(e^{in\theta})$, so we have $n\theta = 0 + 2\pi k$ with $k \in \mathbb{Z}$. In other words, θ can take n values $\{2\pi k/n \mid k = 0, 1, 2, \ldots, n-1\}$ which are distinct modulo $2\pi\mathbb{Z}$. These give n equally-spaced points on the unit circle, 1 being one of them.

To see what multiplication of complex numbers means geometrically, for each $w \in \mathbb{C}$, let's define $\delta_w : \mathbb{C} \to \mathbb{C}$ by $z \mapsto wz$, i.e. $\delta_w(z) = wz$. This is just an extension of the maps δ_k we defined for integers k to all of $\mathbb{C}$. What does the map δ_5 do? Let's see, if $z = a + bi$, then $5z = 5a + 5bi$, so it multiplies the imaginary and real parts by 5. The point z will move along the line from 0 to z and travel to a point 5 times as far from 0 as it was. So δ_5 is a radial "expansion-by-5" map. You can see that for all $s \in \mathbb{R}_{>0}$, that δ_s is an expansion by s map. (If $0 < s < 1$, then it's more of a dilation, isn't it?) You should verify that δ_{-1} is rotation by 180 degrees around 0. More generally, if $s \in \mathbb{R}$ is negative, then the map δ_s is expansion by $|s|$ followed by a 180 degree rotation around 0. Of course, δ_0 is the killer map that sends everybody and her cousin to 0.

Now, what about δ_i? We calculate that $i(a + bi) = ai - b = -b + ai$ so multiplying by i sends the Euclidean point (a, b) to the point $(-b, a)$. This represents a vector which is perpendicular to the vector (a, b) and has moved counterclockwise. In other words, $a + bi$ has moved 90 degrees counterclockwise, with unchanged modulus. Aha, so δ_i has a nice geometric meaning, it is rotation by $\pi/2$ radians counterclockwise. In retrospect, we could have seen that coming, because $i = e^{i\pi/2}$ so if $z = re^{i\theta}$, then $iz = re^{i(\theta+\pi/2)}$. In other words, $|\delta_i z| = |z|$ and $\arg(\delta_i z) = \arg(z) + \pi/2$. If that's not rotation counterclockwise by $\pi/2$ then I'm Mr. Noodle.

What δ_w is geometrically for an arbitrary w should now be clear, for instance by thinking of it as a composite map.