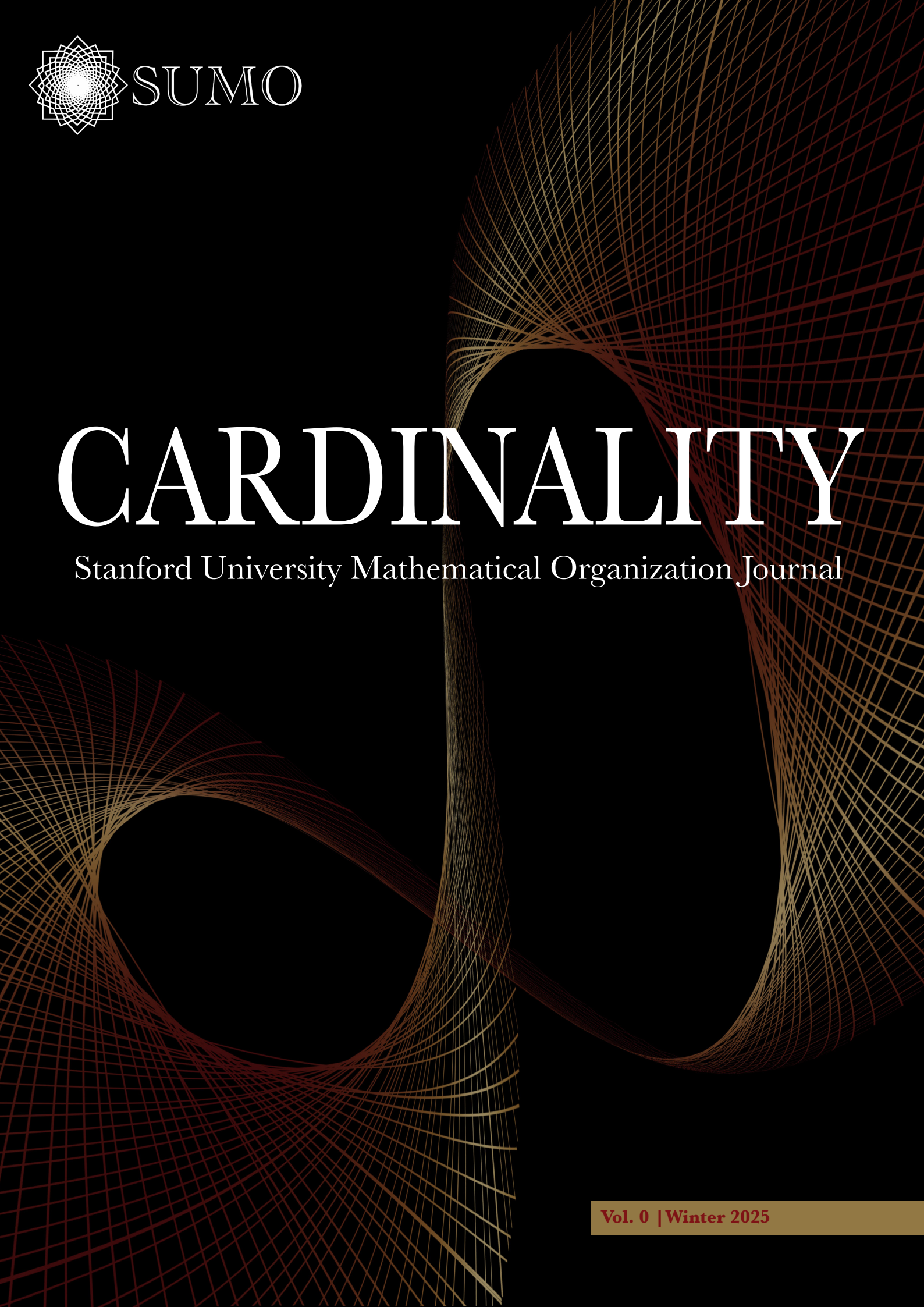

SUMO
CARDINALITY
Stanford University Mathematical Organization Journal
Vol. 0 | Winter 2025

Letter from the Editor

Dear reader,

Welcome to the 0^{th} issue of *Cardinality*: a celebration of the Stanford undergraduate mathematics community.

The idea for this magazine started over a year ago, when we asked: what if we collected people's WIM (writing-in-the-major) papers or senior honors theses to put them in one place? The idea soon extended to include a broader scope of mathematical work: expository papers, research papers, mathematical art, and more.

At its core, we wanted this to showcase the brilliance and diversity of the Stanford undergraduate mathematics community. So, when Aaryan suggested the name *Cardinality*, we ran with it.

This 0^{th} issue is a proof of concept, the beginning of an idea. In their article, *Cardinality and Infinity*, Aaryan Sukhadia '25 writes about infinite cardinalities and the unsolvability of the halting problem: "Turing's answer to the halting problem, rather than being a death knell, told us that there would always be more mathematics to do. In answering one question we raise several more; in proving one result we motivate several more." This ethos is shared in Eric Gao's '25 article *Try Trisecting Triangles*, on the constructibility of various real numbers by straightedge and compass. By recasting constructibility in terms of the algebraic machinery of field extensions, we can easily answer classical questions about trisecting angles or doubling cubes in the negative. Try all you want, but some angles can't be trisected.

In contrast, sometimes all it takes to understand a problem is finding a clever application of a well-known fact. Justin Wu '25 writes about clever applications of the pigeonhole principle to problems in information theory, number theory, and combinatorics, exemplifying the philosophical style of a branch of combinatorics called Ramsey theory.

Yet, there is more to mathematics than just building abstract theory and coming up with clever arguments. Mathematics is done by people, for people, with people—and nowhere is this humanistic spirit more clear than in art. As such, we're proud to include superb visual art, custom-made for each article, by our graphics editors Kae Heller '27 and Paul Gontard '27. In addition, Karen Ge's '23 poem *Instructions on Finishing Your Math Homework* eloquently captures the phenomenological character of doing mathematics: "Think about love, about the way it dissolves into a blank page and finds itself again in your mistakes."

This 0^{th} issue has been a labor of love—love that finds itself in the minutiae of equation spacing and punctuation, love that finds itself in the arresting beauty of pure mathematics, but above all, love that finds itself in the pride of being a part of this community.

My deepest gratitude to everyone who has contributed to this issue, big or small. Now let us celebrate.

Best wishes,

Andrew Lee '25

Editor-in-Chief

Masthead

Contents

Cardinality and Infinity

Aaryan Sukhadia

1 Precursor: Hilbert's *Entscheidungsproblem*

Before LED screens or mouse cursors existed, David Hilbert, one of the eminent mathematicians of the 21st century, dreamed of an algorithm that could be given any set of axioms, along with a statement, and could return if that statement was true given the axioms (putting every mathematician out of a job). Unfortunately for Hilbert (and fortunately for future job-seeking mathematicians), Alonzo Church came by in 1935 to show this was impossible. To add insult to injury, his PhD student, Alan Turing, gave a *different* proof that this was impossible only a year later in 1936. He achieved this by reducing the problem to a different one, and proving the following theorem:

Theorem 1.1. *There is no Turing machine that can decide whether or not a given Turing machine will run forever.*

Amazingly, Alan Turing was able to produce a robust theory of computer science before computers ever existed. His theoretical framework for describing what a computer could do was known as a **Turing machine**. A Turing machine has a specific way of reading inputs and computing outputs, but for our purposes we can think of any modern program as a Turing machine, as they are general enough to be equivalent. In fact, a stronger result is true: virtually *any* model of computation is equivalent to a Turing machine.

This is the **halting problem**: given a program P and an input for P, is there a program that can determine whether P will eventually terminate or continue running forever? Theorem 1.1 tells us the answer is no.

Remark. We will be making the assumption that these programs have infinite memory allocated to them. If memory was finite, we would only have a finite number of configurations to check, and we could easily deduce if a program would terminate or continue forever in some sort of loop.

Our journey through this problem will take us to the infinite, via the lens of cardinality, and lead us into some interesting questions as we do so.

2 Grasping Infinity

Definition 2.1. *Two sets A, B have the same **cardinality** if there exists a bijection $f : A \xrightarrow{\sim} B$. The cardinality of A is denoted $|A|$.*

Another way to say this is that A and B are **equinumerous**, which we denote by $A \sim B$. For finite sets, this concept is fairly intuitive, but it becomes a little more nuanced when working with infinite sets.

2.1 Examples

We can find a bijection $f : \mathbb{N} \xrightarrow{\sim} \mathbb{Z}$ as follows:

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ even,} \\ \frac{-n-1}{2} & \text{if } n \text{ odd.} \end{cases}$$

This is a classic example of a well-known quirk of infinite sets: they can be equinumerous to a *proper* subset of themselves. The German mathematician Richard Dedekind believed that this was a *defining* property of what it meant for a set to be infinite.

Definition 2.2. *A set that is equinumerous to a proper subset of itself is called **Dedekind infinite**.*

This definition coincides with the standard notion of infinite if and only if we assume the axiom of choice.

Question 1. How would one *rigorously* define the "standard" notion of what it means for a set to be infinite?

Perhaps even more surprisingly, $\mathbb{N}$ is not only equinumerous to $\mathbb{Z}$, but also to $\mathbb{Q}$, the rationals. We can find this bijection by listing out every rational number in a well-defined order. We can order then first by sum of

Artwork by Kae Heller

numerator and denominator in lowest form, then numerator, then denominator. For example, our ordering would start off like:

$$\frac{0}{1}, \frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \frac{1}{5}, \dots$$

Then we can assign a unique natural number to every rational number simply by their position in this order, giving us our bijection.

Question 2. The astute reader will have noticed that the above bijection entirely neglects the negative rationals! What small modification can be made to the argument to include them?

Considering that the rationals are a dense subset of the real line (i.e., you can always find infinitely many rationals in however small an interval you choose), the more you think about the fact that they are equinumerous to $\mathbb{N}$ the more absurd the notion becomes. Far more absurd, then, is the fact that $\overline{\mathbb{Q}}$, the set of algebraic numbers (solutions to polynomials with rational coefficients) is equinumerous to $\mathbb{N}$. This not only includes all of the rationals but also a lot of irrationals and complex numbers like $\sqrt{2}, \phi$, and $7i$. Try to see if you can come up with the bijection yourself (Hint: a solution to a $\mathbb{Q}$-polynomial is also a solution to a $\mathbb{Z}$-polynomial!)

Definition 2.3. *The cardinality of $\mathbb{N}, \mathbb{Z}$, $\mathbb{Q}$, and $\overline{\mathbb{Q}}$ is denoted $\aleph_0$ (pronounced 'aleph null').*

The notation above should seem suggestive. If we are suggesting that the size of these infinite sets deserves a "zeroeth subscript," what does that imply for the notion of infinite sizes? As we'll see, mathematicians weren't simply content with grasping and making rigorous the notion of infinity. They wanted to go beyond.

3 Infinitifying Infinity

Definition 3.1. *Given a set X, the **power set** of X, denoted $\wp(X)$, is the set of all subsets of X.*

Theorem 3.2 (Cantor's Theorem)**.** *A set X cannot have the same cardinality as its power set.*

Example 3.3. *The above result should feel somewhat intuitive for finite sets, and let's check this. Let's do the set with just one element in it, $\{a\}$. The subsets of this set are itself, and the empty set $\emptyset$. Thus $\wp(\{a\}) = \{\emptyset, \{a\}\}$, which we leave to the reader to check has greater cardinality than the original set.*

Question 3. Try generalizing: for a finite set with n elements, how many elements should the power set have?

Proof of 3.2. By method of "exercise for the reader," we have dealt with the (perhaps obvious) case of finite sets, but in the world of the infinite nothing can be assumed to be "obvious" when it comes to cardinality.

Assume for sake of contradiction that they *do* have the same cardinality, in which case there must exist a bijection $S : X \xrightarrow{\sim} \wp(X)$ (we use S as in the "set assigned to x"). Since $\wp(X)$ is the set of subsets of X, for any $x \in X$ there are two possibilities, either $x \in S(x)$ or $x \notin S(x)$. Consider the set

$$P \coloneqq \{x \in X : x \notin S(x)\}.$$

In other words, $P \in \wp(X)$ contains all the members of X whose corresponding subset $S(x) \in \wp(X)$ does not contain x. Now we consider the inverse $y = S^{-1}(P) \in X$, which must exist if S is a bijection. There are two possibilities:

1. $y \in P \implies y \notin S(y) = P$,
2. $y \notin P = S(y) \implies y \in P$.

Either way, we get a contradiction, so such a bijection cannot exist. □

What if we take X to be the set of natural numbers, $\mathbb{N}$? We have shown that $\wp(\mathbb{N})$ cannot have the same cardinality as $\mathbb{N}$, but it is still clearly infinite. We could repeat this to get another infinite set $\wp(\wp(\mathbb{N}))$ that is larger still! This gives us a striking result about infinities: there must be *hierarchies* of infinities that can grow *infinitely* large. We usually refer to $\aleph_0$ is as the "smallest transfinite cardinal."

Question 4. Why is $\aleph_0$ the "smallest infinity"? In other words, why isn't there a "smaller" infinite size?

Proposition 3.4. *$\wp(\mathbb{N})$ has the same cardinality as $\mathbb{R}$.*

Incorrect!!!!! Proof. A common proof of this fact listed in math articles and even in university lectures and textbooks is the following: first we find a bijection between $\mathbb{R}$ and the open interval $(0, 1)$, perhaps with something like

$$\begin{aligned} f : (0,1) &\xrightarrow{\sim} \mathbb{R} \\ x &\mapsto \tan\left(\pi x - \tfrac{\pi}{2}\right). \end{aligned}$$

It is not too difficult to verify the above function is a valid bijection. So now our task is reduced to finding a bijection between $\wp(\mathbb{N})$ and $(0, 1)$.

Let us imagine a list $S_1, S_2, \dots$ of subsets of $\mathbb{N}$. We arrange them in a table as shown in Figure 1. In row i, column j, we write a 1 if $j \in S_i$, and a 0 otherwise. This gives us an infinite binary string for each subset, with distinct binary strings corresponding to distinct subsets.

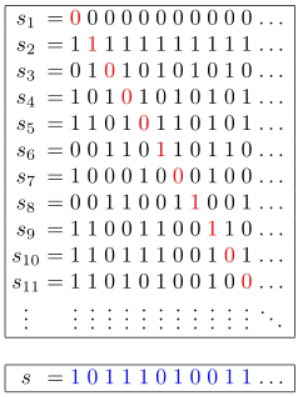


Figure 1: Cantor's Diagonalization

This also gives us another argument for why $\mathbb{N}$ and $\wp(\mathbb{N})$ cannot have the same cardinality. If a bijection $f : \mathbb{N} \xrightarrow{\sim} \wp(\mathbb{N})$ did exist, then we could list out every possible subset as in the figure above, each subset corresponding to S_i for a distinct natural number i. Let us construct a new binary string in the following manner: for each row i, we can switch the value of the ith "bit" to give us a completely new subset S that cannot be any of the listed S_i, as shown in Figure 1. However, this contradicts the assumption that the S_i's accounted for all possible subsets of $\mathbb{N}$.

Thus a bijection cannot exist. This is known as **Cantor's diagonalization argument**. As an exercise, show that this argument is equivalent to the argument we used to prove Cantor's theorem.

Thus, each element $S_i \in \wp(\mathbb{N})$ can be expressed as an infinite string of 1s and 0s. Placing this infinite string after a decimal point and viewing it as a binary number, we can view each subset S_i as a number between 0 and 1. For example, $\{0\}$ would map to $\frac{1}{2}$, $\{1\}$ would map to $\frac{1}{4}$ and $\{0, 1\}$ would map to $\frac{3}{4}$. Thus each subset of $\mathbb{N}$ gives us a distinct binary number between 0 and 1, and clearly writing any such number out in binary expansion we can associate it with a bit vector of a subset of $\mathbb{N}$. Thus we have our desired bijection. □

The above proof is wrong in not just one, but multiple ways. See if you can figure out why, and perhaps you won't give an incorrect proof to your students in the future. A correct proof might help in spotting the mistakes.

Correct! Proof. We must show that there is a bijection between $\wp(\mathbb{N})$ and the real numbers. To do this we use the **Schröder–Bernstein theorem**, which states that if we can find injections $f : A \hookrightarrow B$ and and $g : B \hookrightarrow A$, then there exists a bijection $h : A \xrightarrow{\sim} B$. This is the type of theorem that is initially seemingly obvious, but frustratingly hard to rigorously prove, but obvious again once you see the proof (which I encourage you to search up, after trying to derive it yourself, of course).

To find an injection $f : \wp(\mathbb{N}) \hookrightarrow (0, 1)$, using the notation of $S_i \subseteq \mathbb{N}$ as in the above incorrect proof, we have the function

$$f(S_i) = \sum_{n \in S_i} \frac{1}{10^n}.$$

Verification of injectivity is left as an exercise.

To do the converse, we use the fact that $\mathbb{N}$ is equinumerous to $\mathbb{Q}$, which tells us that $\wp(\mathbb{N})$ is equinumerous to $\wp(\mathbb{Q})$. Thus, an injection $g : \mathbb{R} \hookrightarrow \wp(\mathbb{Q})$ is equivalent to an injection to the power set of the naturals. We use the function:

$$f(r) = \{q \in \mathbb{Q} : q < r\}.$$

The reason why this function is injective has to do with the fact that, given any real numbers $r_1, r_2 \in \mathbb{R}$, we can find a rational between them. For example, we could write down their decimal expansion, and look at the first decimal point where they differ and take a rational between them.

Thus, by the Schröder–Bernstein theorem, $\wp(\mathbb{N})$ and $\mathbb{R}$ are equinumerous. □

4 Proof of the Halting Problem's Unsolvability

Let the set of Turing machines (programs) be $\mathbb{P}$. So how do power sets and cardinalities help us solve the halting problem? To answer this, we ask another question: *what exactly is a program*?

The answer: a block of binary code. A string of 1s and 0s. This means that each program is, at the end of the day, simply a natural number. This gives us an injection $\mathbb{P} \hookrightarrow \mathbb{N}$ (in other words, there's at most as many programs as natural numbers). However, clearly we can think of an infinite number of distinct programs (e.g. a program that runs a loop one time, another than runs a loop two times, etc.), so we have an injection $\mathbb{N} \hookrightarrow \mathbb{P}$. Thus, $\mathbb{P} \sim \mathbb{N}$. Likewise the input to each program is also a block of binary code so, at the end of the day, simply another natural number. Thus, we go back to the table in Figure 1.

Let the rows represent programs, and the column represent inputs. Note that this table includes all possible programs running on all possible inputs. If program i halts on input j, we put a 0 in box (i, j). If it runs forever we put a 1 in box (i, j).

Now, we use Cantor's diagonalization argument to construct a new program $\mathcal{P} \in \mathbb{P}$ whose output is *different* from every possible output of each row. How does this program work? Well, it takes the elements across the diagonal and flips it. In other words, if program S_i halts on input i, then $\mathcal{P}$ returns 1 and otherwise it returns a 0. If the halting problem is indeed solvable, and there exists some magical 'oracle' program that can tell whether any program will run on any input, then we can construct this program $\mathcal{P}$ as follows:

```
def oracle( program P, input I):
    if(P(I).halts()):
        return(1)
    else:
        return(0)
```

Listing 1: Our 'Oracle' code that can solve the halting problem

However, note that this program cannot appear anywhere in our enumeration of programs that we constructed, and thus it cannot exist! This contradicts the assumption of the existence of a subroutine capable of determining whether any given program halts. This proof rests on the fact that the set of all possible program-input combinations is in some sense 'bigger' than the set of all programs themselves, and so there cannot exist a program whose output matches the output of every program-input combination.

Another way to think about it is if we change the code slightly, we can derive a program that leads to a contradiction, by making programs take *themselves* as inputs:

```
def oracle( program P):
    if(P(P).halts()):
        while(True):
            run
    else:
        return(0)
```

Letting $\mathcal{O}$ represent the oracle program, if we consider what happens when we run $\mathcal{O}(\mathcal{O})$, we get a paradox. If the oracle determines this will run forever, it will halt, but if it detects it will halt, it will run forever. This results in a contradiction, and thus $\mathcal{O}$ cannot exist.

5 Postcursor

In going on this journey to prove the halting problem I realize I may have raised more questions than answered, but in a sense this is precisely the spirit of mathematics. Turing's answer to the halting problem, rather than being a death knell, told us that there would always be more mathematics to do. In answering one question we raise several more; in proving one result we motivate several more.

That is the ultimate pedagogical philosophy underpinning Cardinality. Mathematics is and always has been a field of shared knowledge, shared ideas and shared learning. We hope the writings and learnings of the Stanford community that this publication goes on to display inspires and galvanizes the never-ending, feverish pursuit of mathematical beauty and truth.

5.1 Further Reading

This article is written deliberately with multiple loose ends, to keep the reader questioning and wondering. If any are curious enough to explore further, here are some good places to do so:

- More Set Theory: If you want to find out more about infinite cardinals, their weird cousins the ordinals, or why exactly Dedekind infinite is, Keith Kearnes of the University of Colorado has a fantastic repository of course materials at his website.
- More Computing Theory: In my view, there is no better place to learn the foundations of computing theory than from the monumental paper from the man after which the Turing machines are named. Reading it should hopefully give you the sense that you yourself could have come up with these ideas and results.

Instructions for Finishing Your Math Homework

Karen Ge

Keep your eyes open. Remember your body
doesn't belong to anything else
but the chords of breath ringing in your bones.
He tells you the most beautiful
things, scratches them in white,
just as you slip into a wild sleep.

You must write the most difficult lines at 2
in the morning,
when your soul belongs to every corner of the world,
and starry flakes fall like ash
from a forest struck by lightning.
You must keep paying
attention.

Think about love, about the way it dissolves
into a blank page
and finds itself again in your mistakes.
Set your life on fire.
Tell yourself it is holy.
And trace the golden rays of sun
as they weave, like fractal droplets, into a melody

Artwork by Paul Gontard

Try Trisecting Triangles

Eric Gao

1 Introduction

Many geometric objects can be constructed using only a straightedge and a compass. For instance, it is relatively straightforward to construct an equilateral triangle given a line segment by doing the following:

1. Open the compass so that the endpoint of the first leg is on one end of the line segment and the endpoint of the second leg is on the other end of the line segment;
2. Keeping that width, draw one circle centered around one end of the line segment and draw a second circle centered around the other end of the line segment;
3. Pick one of the intersections between the two circles (there should be two such intersections);
4. Use a straight edge to connect the two endpoints of the original line segment and the intersection identified in the previous step.

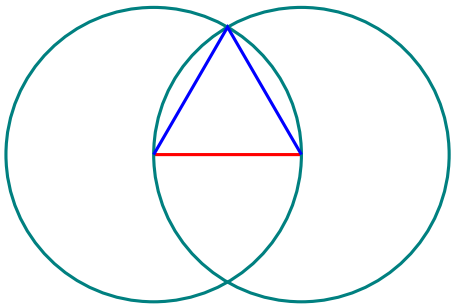


Figure 1: Constructing an equilateral triangle. Original line segment. Step two. Step four.

Each of the two lines drawn in step four connects the center of a circle with radius equal to the original line to a point on the circle, and thus has length equal to the original line. As such, all three line segments have the same length and form a triangle, and thus the constructed triangle is equilateral.

While equilateral triangles are interesting on their own, many more shapes are constructable as well. For instance, regular hexagons can be constructed by constructing six equilateral triangles, angles can be bisected, and even a regular heptadecagon (19 sides) can be constructed.[1]

On the other hand, what are the limits of such geometric constructions? It turns out that some seemingly simple objects elude the power of the straightedge and compass. For instance, triangles and angle bisectors are constructable,

[1] See this Wikipedia file for an animated construction.

Artwork by Kae Heller

series of actions yields the supposedly non-constructable object? Even the ancient Greeks could not come up with a fully satisfactory answer, but the remainder of this paper will do just that.

2 Constructable Numbers

What does it formally mean for something to be constructable? Before that notion can be formalized, what are a straightedge and compass, mathematically? Geometrically, constructions are created with physical objects (a straightedge and compass) on some surface (a sheet of paper) in the real world (not to be confused with the $\mathbb{R}$ world). Working backwards through these notions, start by identifying a two-dimensional surface with $\mathbb{R}^2$ so every point p that might be drawn has some associated coordinates $(p_x, p_y) \in \mathbb{R}^2$. With this in mind, straightedges and compasses can be formalized.

Definition 2.1 (Straightedge)**.** *A **straightedge** S is a function that takes in two distinct points $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$ and returns the line connecting them, namely the set*

$$S((x_1, y_1), (x_2, y_2)) = \left\{(x, y) \in \mathbb{R}^2 \colon y = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1\right\}$$

if $x_1 \neq x_2$ and

$$S((x_1, y_1), (x_2, y_2)) = \left\{(x, y) \in \mathbb{R}^2 \colon x = x_1\right\}$$

if $x_1 = x_2$.

Geometrically, this amounts to aligning a straightedge so both the two input points are on the straightedge, and then drawing a line along the edge of the straightedge.

Definition 2.2 (Compass)**.** *A **compass** C is a function that takes in three points $(x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbb{R}^2$ with $(x_2, y_2) \neq (x_3, y_3)$ and returns a circle centered around (x_1, y_1) with radius equal to the distance between (x_2, y_2) and (x_3, y_3). In particular, it returns the set*

$$C((x_1, y_1), (x_2, y_2), (x_3, y_3)) = \left\{(x, y) \in \mathbb{R}^2 \colon \sqrt{(x - x_1)^2 + (y - y_1)^2} = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}\right\}.$$

Geometrically, this amounts to drawing a circle centered at (x_1, y_1) with radius equal to the distance between (x_2, y_2) and (x_3, y_3). This can be executed by opening up a compass until the endpoints of the two legs are on (x_2, y_2) and (x_3, y_3), followed by putting the "pointy leg" on (x_1, y_1) and tracing out the "drawing leg."

Geometric constructions start at some point; call this point the origin or $(0, 0)$. Straightedges and compasses require two distinct points as inputs. As such, without loss of generality, let $(0, 1)$ be the second point. What points are constructable? Perhaps a randomly chosen point just happens to be (π, e) which makes those two values "constructable", but that is an unsatisfactory answer. Instead, constructable numbers must be *pinned down* by geometric constructions. Take $(0, 0)$ and $(0, 1)$ to be constructable. Then:

Definition 2.3 (Constructable Point)**.** *A point $p \in \mathbb{R}^2$ is **constructable** if starting from the points $\{(0, 0), (1, 0)\}$, the point p is in the output of a finite number of applications of Straightedge or Compass.*

The set of constructable numbers are then the set of possible coordinates of constructable points: $x \in \mathbb{R}$ is a constructable number if (x, y) or (y, x) is a constructable point for some $y \in \mathbb{R}$. Another natural way to define constructable numbers is to say that $x \in \mathbb{R}$ is constructable if it is possible to find constructable points $p_1, p_2 \in \mathbb{R}^2$ such that the (Euclidean) distance between p_1, p_2 is x. It turns out that these two notions perfectly coincide.

Proposition 2.4 (Equivalence of Length and Coordinate Definitions)**.** *A number $x \in \mathbb{R}$ is the distance between two constructable points if and only if there exists $y \in \mathbb{R}$ such that (x, y) or (y, x) is a constructable point.*

Proof. In the forward direction, suppose x is the distance between constructable points (x_1, y_1) and (x_2, y_2). Let $L = S((0, 0), (1, 0))$ be the line connecting $(0, 0)$ and $(0, 1)$. Geometrically, this is the x-axis. Next, let $C = C((0, 0), (x_1, y_1), (x_2, y_2))$. Geometrically, this is a circle centered around the origin with radius equal to $x = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

One intersection of L and C is $(x, 0)$. This point is in L as

$$0 = \frac{0 - 0}{1 - 0}(x - 0) + 0$$

while it is also in C as

$$\sqrt{(x - 0)^2 + (0 - 0)^2} = x = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

As such, $(x, 0)$ is constructable.

Conversely, suppose (x, y) is constructable (the other case where (y, x) is constructable follows similarly). Then, do the following construction:

1. Construct the point $(2x, 0)$ by taking $S((0,0),(1,0)) \cap C((x,y),(x,y),(0,0))$ (there will be two points in the intersection with the other being $(0,0)$);
2. Construct the point $(x, -y)$ by taking $C((0,0),(x,y),(0,0)) \cap C((2x,0),(x,y),(0,0))$ (there will be two points in the intersection with the other being (x,y));
3. Construct the point $(x, 0)$ by taking $S((x,y),(x,-y)) \cap S((0,0),(0,1))$.

Then, the distance between $(0,0)$ and $(x,0)$ is x, as desired. □

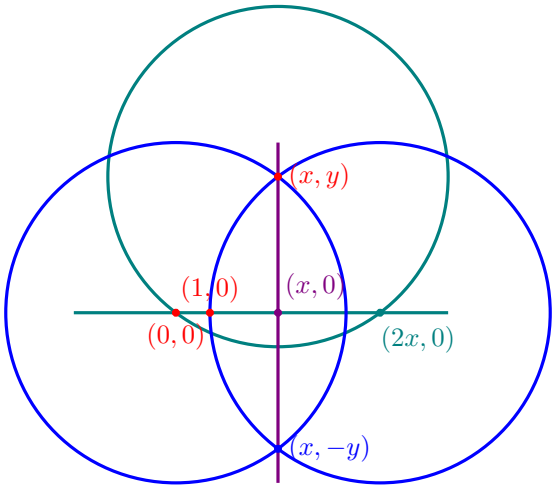


Figure 3: Proof of Proposition 2.4. Starting points. Step one. Step two. Step three.

Geometrically, the forward direction demonstrates how compasses can be used to translate distances while the converse direction demonstrates how it is possible to project one coordinate onto its respective "axis". As a result:

Corollary 2.4.1. *A number $x \in \mathbb{R}$ is constructable if and only if the point $(0, x)$ is constructable.*

It turns out that it is possible to give a complete algebraic characterization of constructable numbers.

Proposition 2.5 (Characterization of Constructable Points)**.** *A number $x \in \mathbb{R}$ is constructable if and only if there exists a sequence $\{k_i, \mathcal{C}_i\}_{i=0}^{n}$ such that:*

1. *$\mathcal{C}_0 = \mathbb{Q}$;*
2. *For all $i \geq 0$, $k_i \in \mathcal{C}_i$;*
3. *For all $i \geq 0$, $\mathcal{C}_{i+1} = \mathcal{C}_i[\sqrt{k_i}]$, where $\mathcal{C}_i[\sqrt{k_i}] = \{a + b\sqrt{k_i} : a, b \in \mathcal{C}_i\}$;*
4. *$x \in \mathcal{C}_n$.*

By construction, each $\mathcal{C}_i$ is a field[2] over addition and multiplication (since $\mathbb{Q}$ is a field and completing the square allows for the field to be closed over multiplicative inverses). The forward direction proceeds in three parts: showing that all integers are constructable, all quotients of constructable numbers are constructable (and hence all rationals are constructable), and that if a number is constructable, its square root is constructable.

Lemma 2.6 (Integers are Constructable)**.** *If $n \in \mathbb{Z}$ is an integer, then n is constructable.*

Proof. First, we show that all positive integers are constructable. In particular, we show that all points of the form $(n, 0)$ are constructable for natural numbers n by induction. Clearly, $(1, 0)$ is constructable. Next, suppose $(n-1, 0)$ is constructable. Then,

$$S((0,0),(n-1,0)) \cap C((n-1,0),(0,0),(1,0)) = \{(n-2,0),(n,0)\}$$

[2] A **field** is a set equipped with two operations, addition and multiplication, such that "math" works: Addition is associative with multiplication, addition and multiplication are commutative, additive and multiplicative identities and inverses exist, and addition distributes over multiplication.

as it is the intersection of the x-axis and a circle of radius one centered around $(n-1,0)$. Thus, $(n,0)$ is constructable and all natural numbers are constructable. By a similar argument, we can show that all negative integers are constructable, and we are done. □

A similar construction also gives that sums and differences of constructable numbers are constructable.

Lemma 2.7 (Quotients are Constructable)**.** *If a and b are constructable with $a \neq 0$, then so is $\frac{b}{a}$.*[3]

Proof. Suppose a and b are constructable numbers. Then, $(a,0)$ and $(b,0)$ are constructable points. By a similar construction as that in the proof of Proposition 1, perpendiculars are constructable, so the y-axis is a constructable line and $(0,1)$ is a constructable point. By constructing perpendiculars twice, it is possible to construct parallel lines through a point. Then, construct a line through $(b,0)$ parallel to the line connecting $(0,1)$ and $(a,0)$. The intersection of that line and $S((0,0),(0,1))$ will be the point $\left(0,\frac{b}{a}\right)$ by similar triangles. □

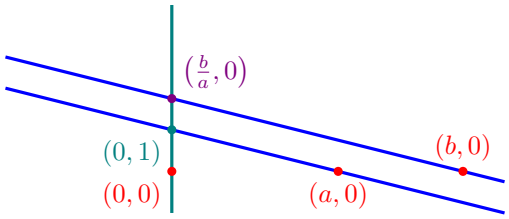


Figure 4: Proof of Lemma 2.7. Starting points. Step one. Step two. Desired point.

A similar construction creates the triangle $(0,0),(x,0),(0,b)$ similar to the triangle $(0,0),(a,0),(0,1)$. By similar triangles, $x = ab$ and hence products of constructable numbers are constructable as well.

Lemma 2.8 (Square Roots)**.** *Suppose a is constructable. Then, $\sqrt{a}$ is constructable.*

Proof. By Lemmas 1 and 2, $\frac{a+1}{2} = \frac{a}{2} + \frac{1}{2}$ and $\frac{a-1}{2} = \frac{a}{2} - \frac{1}{2}$ are constructable, so the points $\left(0,\frac{a}{2}-\frac{1}{2}\right),\left(0,\frac{a}{2}+\frac{1}{2}\right)$ are constructable. Then, by power of a point with respect to $(0,0)$, the x-coordinates of the points in

$$S((0,0),(1,0)) \cap C\left(\left(0,\tfrac{a}{2}-\tfrac{1}{2}\right),(0,0),\left(0,\tfrac{a}{2}+\tfrac{1}{2}\right)\right)$$

are $\pm\sqrt{a}$.

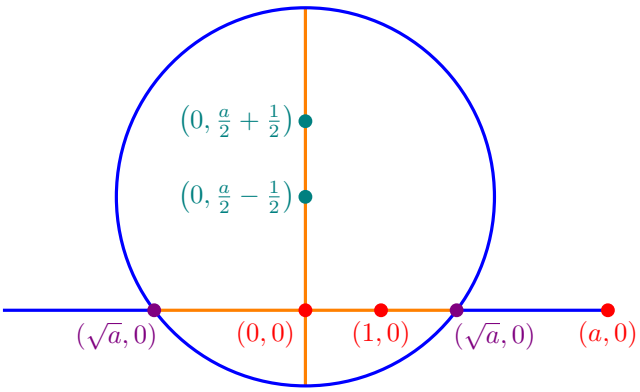


Figure 5: Proof of Lemma 2.8. Starting points. Step one. Step two. Desired points. Power of a point segments.

This is because the diameter of the circle is

$$2\sqrt{(0-0)^2+\left(0-\left(\tfrac{a}{2}+\tfrac{1}{2}\right)\right)^2} = 2\left(\tfrac{a}{2}+\tfrac{1}{2}\right) = a+1$$

and a point on the circle is $(0,-1)$ as

$$\sqrt{\left(-1-\left(\tfrac{a}{2}-\tfrac{1}{2}\right)\right)^2} = \sqrt{\left(0-\left(\tfrac{a}{2}+\tfrac{1}{2}\right)\right)^2}$$

[3] The assumption of $a \neq 0$ may be relaxed if your straightedge and compass are large enough.

so the point $(0, 0)$ splits a diameter along the y-axis into segments of lengths a and 1 while splitting the portion of the x-axis into two equal segments of length λ. It must be that $\lambda^2 = a \cdot 1$, so the intersections must be at $(-\sqrt{a}, 0)$ and $(\sqrt{a}, 0)$. □

Taken together, Lemmas 2.6, 2.7, and 2.8 give that if a real number can be made using some sequence of adjoining square roots, then it is constructable. The converse of Proposition 2.5 similarly proceeds in three steps, analyzing that in each of the three valid geometric constructions, nothing other than square roots can be created.

Lemma 2.9 (Intersection of Two Circles). *Suppose a, b, c, d, e, f are points in $\mathbb{R}^2$ with coordinates in some field $\mathcal{C}_i$. Let $C_1 = C(a, b, c)$ and $C_2 = C(d, e, f)$ with $C_1 \neq C_2$. If $(x, y) \in C_1 \cap C_2$, then $x \in \mathcal{C}_i[k_i]$ for some $k_i \in \mathcal{C}_i$ (and an analogous result holds for y).*

Proof. Suppose $(x, y) \in C_1 \cap C_2$. Let $a = (a_x, a_y)$ and so forth with b, c, d, e, f. Then, (x, y) satisfies

$$\sqrt{(x - a_x)^2 + (y - a_y)^2} = \sqrt{(c_x - b_x)^2 + (c_y - b_y)^2} \text{ and } \sqrt{(x - d_x)^2 + (y - d_y)^2} = \sqrt{(f_x - e_x)^2 + (f_y - e_y)^2}$$

Let

$$r_1 = \sqrt{(c_x - b_x)^2 + (c_y - b_y)^2} \quad \text{and} \quad r_2 = \sqrt{(f_x - e_x)^2 + (f_y - e_y)^2}$$

be constructable numbers. Squaring all expressions above gives that

$$(x - a_x)^2 + (y - a_y)^2 = x^2 - 2xa_x + a_x^2 + y^2 - 2ya_y + a_y^2 = r_1^2$$

and

$$(x - d_x)^2 + (y - d_y)^2 = x^2 - 2xd_x + d_x^2 + y^2 - 2yd_y + d_y^2 = r_2^2.$$

Subtracting the two equations from each other gives that

$$2(d_x - a_x)x + a_x^2 - d_x^2 + 2(d_y - a_y)y + a_y^2 - d_y^2 = r_1^2 - r_2^2.$$

If $a_y = d_y$ or $a_x = d_x$, then x, y already have closed-form solutions. Going forward, suppose $a_y \neq d_y$ and $a_x \neq d_x$.

As fields are closed under addition and multiplication, all terms (other than x, y) are in $\mathcal{C}_i$ so the above is a linear equation in x, y. Thus, solving for y in terms of x yields

$$y = a'x + b'$$

for some $a', b' \in \mathcal{C}_i$. Substituting this back gives that x is characterized by

$$x^2 - 2xd_x + d_x^2 + (a'x + b')^2 - 2(a'x + b')d_y + d_y^2 = r_2^2.$$

Once again using the fact that $\mathcal{C}_i$ is closed under addition and multiplication, algebraic manipulation gives that the above is equivalent to

$$a''x^2 + b''x + c'' = 0$$

for some $a'', b'', c'' \in \mathcal{C}_i$. The Quadratic Formula then gives that

$$x = \frac{-b'' \pm \sqrt{(b'')^2 - 4a''c''}}{2a''} = -\frac{b''}{2a''} \pm \frac{1}{2a''}\sqrt{(b'')^2 - 4a''c''}.$$

As $\mathcal{C}_i$ is closed under inverses, $-\frac{b''}{2a''}, \frac{1}{2a''} \in \mathcal{C}_i$. As $\mathcal{C}_i$ is closed under multiplication and addition, $(b'')^2 - 4a''c'' \in \mathcal{C}_i$ as well. Thus, taking $k_i = (b'')^2 - 4a''c''$ gives that

$$x = -\frac{b''}{2a''} \pm \frac{1}{2a''}\sqrt{(b'')^2 - 4a''c''} \in \{m + n\sqrt{(b'')^2 - 4a''c''} : m, n \in \mathcal{C}_i\} \subset \mathcal{C}_i[\sqrt{k_i}]$$

as desired. □

The proofs of the remaining two cases are largely similar, so some details are omitted.

Lemma 2.10 (Intersection of a Line and a Circle). *Suppose a, b, c, d, e are points with coordinates in some field $\mathcal{C}_i$. Let $L = S(a, b)$ and $C = C(c, d, e)$. If $(x, y) \in L \cap C$, then $x \in \mathcal{C}_i[\sqrt{k_i}]$ for some $k_i \in \mathcal{C}_i$.*

Proof. Let $(x, y) \in L \cap C$. If $a_x = b_x$ then L is vertical, so $x = a_x \in \mathcal{C}_i$. Otherwise, (x, y) satisfies

$$y = \frac{b_y - a_y}{b_x - a_x}(x - a_x) + a_y$$

and

$$\sqrt{(x - c_x)^2 + (y - c_y)^2} = \sqrt{(e_x - d_x)^2 + (e_y - d_y)^2}.$$

Letting $r = \sqrt{(e_x - d_x)^2 + (e_y - d_y)^2}$ so $r^2 \in \mathcal{C}_i$,

$$y = a'x + b'$$

for some $a', b' \in \mathcal{C}_i$ and

$$(x - c_x)^2 + (y - c_y)^2 = r^2.$$

Substituting gives that x is characterized by

$$(x - c_x)^2 + ((a'x + b') - c_y)^2 = r^2$$

which can be re-written as

$$a''x^2 + b''x + c'' = 0$$

for some $a'', b'', c'' \in \mathcal{C}_i$. Then, the quadratic formula gives that $x \in \mathcal{C}_i[\sqrt{k_i}]$ for some $k_i \in \mathcal{C}_i$. □

Lemma 2.11 (Intersection of Two Lines)**.** *Suppose a, b, c, d are points with coordinates in some field $\mathcal{C}_i$. Let $L_1 = S(a, b)$ and $L_2 = S(c, d)$ If $(x, y) \in L_1 \cap L_2$, then $x \in \mathcal{C}_i[\sqrt{k_i}]$ for some $k_i \in \mathcal{C}_i$.*

Proof. Let $(x, y) \in L_1 \cap L_2$. Then,

$$y = \frac{b_y - a_y}{b_x - a_x}(x - a_x) + a_y$$

and

$$y = \frac{d_y - c_y}{d_x - c_x}(x - c_x) + c_y.$$

Thus, x is characterized by

$$\frac{b_y - a_y}{b_x - a_x}(x - a_x) + a_y = \frac{d_y - c_y}{d_x - c_x}(x - c_x) + c_y$$

which is a linear equation with coefficients in $\mathcal{C}_i$. Thus, $x \in \mathcal{C}_i$ so taking k_i to be anything in $\mathcal{C}_i$ gives that $x \in \mathcal{C}_i[\sqrt{k_i}]$. □

3 Non-Constructable Numbers

The characterization of constructable numbers in the language of Proposition 2.5 forms the foundation for the proofs of why certain geometric objects are not constructable.

3.1 Trisecting an Angle

Is it possible to trisect every angle? To produce a negative result, it suffices to show that *some* angle is not trisectable: in particular, a 60 degree angle cannot be trisected. If a 60 degree angle could be trisected, then 20 degree angles could be constructed; a $20° - 70° - 90°$ triangle with hypotenuse of length one would have legs of length $\cos(20°)$ and $\sin(20°)$, making them constructable numbers. The following proposition shows that this can not be the case.

Proposition 3.1. *The number $\cos(20°)$ is not constructable.*

Proof. Observe that by the triple angle formula,

$$\frac{1}{2} = \cos(60°) = \cos(3 \cdot 20°) = 4\cos^3(20°) - 3\cos(20°)$$

so $\cos(20°)$ is a root of the polynomial

$$f(x) = 4x^3 - 3x - \frac{1}{2}$$

or equivalently, $\cos(20^\circ)$ is a root of the polynomial

$$g(x) = 8x^3 - 6x - 1.$$

As such, it is sufficient to show that no root of g is constructable.

The proof proceeds inductively. First, we will show no root of g is in $\mathbb{Q}$. Then, we will show if there are no roots of g in $\mathcal{C}_i$ for some field $\mathcal{C}_i$, then there are no roots of g in $\mathcal{C}_i[\sqrt{k_i}]$ for any $k_i \in \mathcal{C}_i$.

Towards a contradiction, suppose $g(x)$ has some rational root r. By the rational roots theorem,

$$r \in \left\{\pm 1, \pm\frac{1}{2}, \pm\frac{1}{4}, \pm\frac{1}{8}\right\}$$

. However, none of these work: to reduce the number of cases, note that

$$8r^3 - 6r - 1 = 0 \iff 8r^3 - 6r = 1 \iff r(8r^2 - 6) = 1 \iff |r|(8r^2 - 6) = 1$$

so the sign of r does not matter. Then:

- If $r = \pm 1, |r|(8r^2 - 6) = 2 \neq 1$;
- If $r = \pm\frac{1}{2}, |r|(8r^2 - 6) = -2 \neq 1$;
- If $r = \pm\frac{1}{4}, |r|(8r^2 - 6) = -\frac{11}{8} \neq 1$;
- If $r = \pm\frac{1}{8}, |r|(8r^2 - 6) = -\frac{47}{64} \neq 1$.

As such, $g(x)$ has no rational roots and $\cos(20^\circ)$ cannot be in $\mathbb{Q}$.

Next, suppose g has no roots some field $\mathcal{C}_i$. Towards a contradiction, suppose g has a root $r \in \mathcal{C}_i[\sqrt{k_i}]$ for some $k_i \in \mathcal{C}_i$. Then,

$$r = a + b\sqrt{k_i}$$

for some $a, b \in \mathcal{C}_i$. As r is a root of $g(x)$, it holds that $g(a + b\sqrt{k_i}) = 0$. Expanding gives that

$$(8a^3 + 24ab^2k_i - 4a + 1) + (24a^2b + 8b^3k_i - 4b)\sqrt{k_i} = 0.$$

It must be the case that $24a^2b + 8b^3k_i - 4b = 0$. If not, either $\sqrt{k_i} = 0$ or

$$\sqrt{k_i} = -\frac{8a^3 + 24ab^2k_i - 4a + 1}{24a^2b + 8b^3k_i - 4b}.$$

In either case, this would mean that $\sqrt{k_i} \in \mathcal{C}_i$, so $\cos(20^\circ) = a + b\sqrt{k_i} \in \mathcal{C}_i$, a contradiction.
Thus,

$$8a^3 + 24ab^2k_i - 4a + 1 = 24a^2b + 8b^3k_i - 4b = 0$$

Then, replacing b with $-b$ gives

$$(8a^3 + 24ab^2k_i - 4a + 1) + (-24a^2b - 8b^3k_i + 4b)\sqrt{k_i} = 0 - 0\sqrt{k_i} = 0$$

so $a - b\sqrt{k_i}$ is a root of g as well. Note the quadratic term vanishes, so the sum of roots of g is

$$-\frac{0}{8} = 0$$

so by Vieta's formula and the fundamental theorem of algebra, the final root of g is

$$0 - (a + b\sqrt{k_i}) - (a - b\sqrt{k_i}) = -2a.$$

However, fields are closed under multiplication so $-2a \in \mathcal{C}_i$ and g has a root in $\mathcal{C}_i$, which is a contradiction. This completes the proof. □

3.2 Doubling the Cube

Given a line segment ℓ, is it possible to construct a new line segment ℓ' such that a cube with edges ℓ' has twice the volume of a cube with edges ℓ? In general, it will not be possible: consider ℓ having length one, so a cube with edges ℓ has volume one. This problem then becomes whether or not it is possible to construct a line segment with length $\sqrt[3]{2}$. Unfortunately (or perhaps fortunately?), the answer to this question is no.

Proposition 3.2. *The number $\sqrt[3]{2}$ is not constructable.*

Proof. The proof will proceed similar to the proof of Proposition 3.1. Observe that

$$(\sqrt[3]{2})^3 = 2$$

so $\sqrt[3]{2}$ is a root of

$$f(x) = x^3 - 2.$$

Similar to before, if r is a root of x, then r is not rational and if there are no roots of f in the field $\mathcal{C}_i$, then there are no roots of f in the field $\mathcal{C}_i[\sqrt{k_i}]$ for any $k_i \in \mathcal{C}_i$.

First, there are no rational roots of f. By the rational roots theorem, if r is a rational root of f then

$$r \in \{\pm 1, \pm 2\}.$$

Checking these possibilities for whether or not $r^3 = 2$ yields

$$(1)^3 = 1, \quad (-1)^3 = -1, \quad (2)^3 = 8, \quad (-2)^3 = -8,$$

so f has no rational roots.

Next, suppose f has no roots in the field $\mathcal{C}_i$. Towards a contradiction, suppose f has a root r in $\mathcal{C}_i[\sqrt{k_i}]$ for some $k_i \in \mathcal{C}_i$. Then, $r = a + b\sqrt{k_i}$ for $a, b \in \mathcal{C}_i$. As $f(r) = 0$,

$$(a^3 + 3ab^2 - 2) + (3a^2b + b^3k_i)\sqrt{k_i} = 0.$$

Then, $3a^2b + b^3c = 0$ as, if not, either $\sqrt{k_i} = 0$ or $\sqrt{k_i} = -\frac{a^3+3ab^2-2}{3a^2b+b^3k_i}$ is in $\mathcal{C}_i$, contradicting $r \notin \mathcal{C}_i$. As such, $a - b\sqrt{k_i}$ is another root of f. By sum of roots, the final root of f is

$$0 - (a + b\sqrt{k_i}) - (a - b\sqrt{k_i}) = -2a$$

and we note $2a \in \mathcal{C}_i$, contradicting the original assumption that f has no roots in $\mathcal{C}_i$. □

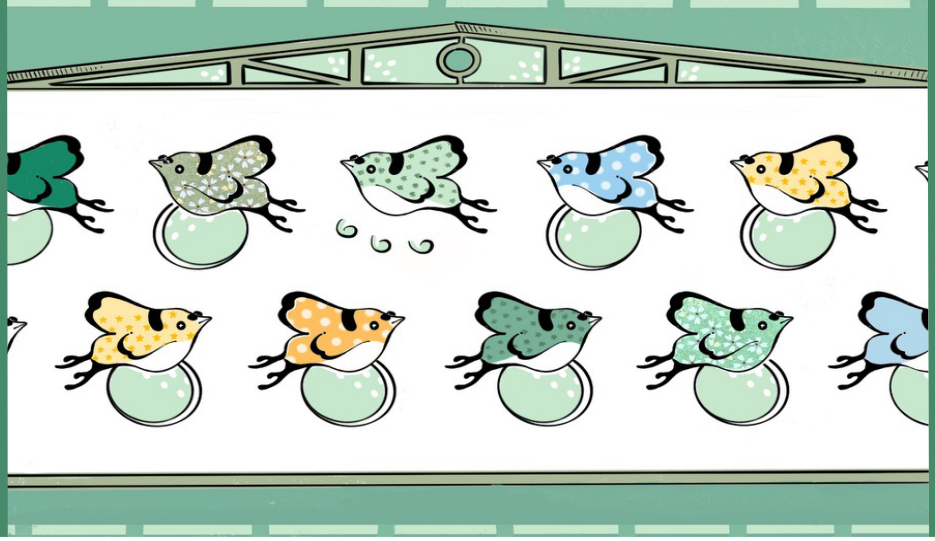
4 Conclusion

It was once famously said:[4]

> Since the dawn of time, human beings have asked some fundamental questions: who are we? why are we here? is there life after death? Unable to answer any of these, in this paper we will consider cohomology classes on a compact projective manifold that have a property analogous to the Hard-Lefschetz Theorem and Hodge-Riemann bilinear relations.

Unable to parse the last sentence, this paper grasps lower-hanging fruit and instead develops connections between algebraic structures and geometric constructions to ultimately answer why it is impossible to trisect every angle or double every cube, a question dating back to ancient Greece. Constructable numbers in the plane generally correspond to roots of polynomials and can be characterized by taking the rationals and sequentially adjoining square roots of numbers that are known to be constructable.

[4] "On Hodge-Riemann Cohomology Classes" by Julius Ross and Matei Toma, 2021.



PIGEONHOLES

Justin Wu

1 Preliminary Examples

The Pigeonhole Principle (which we take for granted) states: If n items are put into m containers with $n > m$, then at least one of the containers must have more than one item in it. Let's see some straightforward examples to get some flavor of its use.

Theorem 1.1. *There must be two people in New York City with the same number of hairs on their head.*

To illustrate, suppose (and this is an overestimate) the maximum number of hairs on a head is $300,000$. Since there are more than $300,000$ people in New York City who all have some natural number of hairs on their head, by the Pigeonhole Principle, at least two people will have the same number of hairs. Let's now move to a combinatorial example.

Theorem 1.2. *In every group of six people there will be at least three mutual acquaintances or three mutual strangers.*

This is one of the motivating examples in Ramsey Theory, a field in combinatorics where the general philosophy is that every very large structure (often graphs) contains a large well-organized substructure (properties or substructures). In particular, Ramsey Theory is concerned with finding order in chaos. To prove the theorem, we turn to graph theory. Consider the complete graph K_6 with six vertices (representing the six people), where all the edges are colored either red or blue, based on whether the people are acquaintances or strangers. We want to prove that there must exist a red triangle or a blue triangle.

Proof. Choose some vertex v in K_6. Since we're in K_6, there are 5 edges incident to v, and so by Pigeonhole, 3 of them must be the same color. Without loss of generality, let this color be red. Suppose these 3 edges connect the vertex v to the vertices a, b, c. Now, if any of the edges $(ab), (bc), (ac)$ are blue, then we have a blue triangle. If not, then all 3 of those edges must be red, and they thus form a red triangle. □

Let's end the preliminaries with a more applied example from computer science.

Theorem 1.3. *There is no lossless data compression algorithm that shortens every file.*

This theorem essentially says that it's impossible to create a magic compression algorithm that makes every single file smaller without losing any information. Imagine you're trying to pack all your clothes into a suitcase. There's a limit to how much you can compress everything before you either have to leave something out (lose data) or acknowledge that not everything can be made smaller. In the world of data, this means there are always going to be some files that, when compressed, either stay the same size or might even need to become larger to ensure they can be fully recovered (decompressed) without losing any bits of information.

Proof. Suppose that there does exists some lossless data compression algorithm A that can shorten every file. We consider all possible files of a certain length n bits. Then there are 2^n such files. If A shortens every file, then every file of length n must be mapped to some file of length less than n. But there are fewer files of length less than n than there are of length n. To see this, note that the total number of files of length less than n is

$$\sum_{i=0}^{n-1} 2^i = 2^n - 1,$$

which is less than 2^n, the total number of files of length n. So, by the Pigeonhole Principle, at least two files of length n need to be compressed to the same shorter file. This contradicts the lossless condition for A, since losslessness requires a unique decoding of each compressed file back to its original form. So we are done. □

In fact, it follows that any such compression algorithm necessarily *increases* the size of some files.

Artwork by Paul Gontard

2 Minkowski's Theorem

Minkowski's theorem is a pretty intuitive, but incredibly versatile theorem that lends itself in many areas of algebra and number theory. It's a powerful tool for developing much of the geometry of numbers, and sees itself in proofs of the finiteness of the class number and Dirichlet's unit theorem in particular.

Theorem 2.1. *Every symmetric (w.r.t the origin) convex set in $\mathbb{R}^n$ with volume $> 2^n$ contains a nonzero lattice point (that is, a point in $\mathbb{Z}^n \setminus \{0\}$).*

Let's prove the case for $n = 2$.

Proof for $n = 2$. Let S be convex and origin-symmetric with $\text{Area}(S) > 4$. We now consider taking each point in $S \pmod 2$. That is, we consider the map

$$(x, y) \mapsto (x \bmod 2, y \bmod 2).$$

Here, by mod 2, we mean the system where for $a, b \in \mathbb{R}$, $a \equiv b$ if and only if $a - b \in 2\mathbb{Z}$. Thus, this is a function of the form $f : S \to \mathbb{R}^2/2\mathbb{Z}^2$. Intuitively, we have cut the plane into 2 by 2 squares, and we're stacking each of these squares on top of each other. It is clear that $f(S)$ has area ≤ 4, since its codomain is a 2 by 2 square. Hence, after stacking, there must be some overlapping regions, since $\text{Area}(S) > 4$. We can view this as some geometric form of the Pigeonhole Principle! In particular, if we have some volume-nondecreasing function f compacting a set A into a set B with smaller volume, then f cannot be injective.

So there must be distinct points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ such that $f(x) = f(y)$. By our definition of f, it follows that

$$(x_1 \bmod 2, x_2 \bmod 2) = (y_1 \bmod 2, y_2 \bmod 2),$$

hence the points differ by some integer translation with factor 2: $x + (2i, 2j) = y$ for some $i, j \in \mathbb{Z}$ with at least one being nonzero. Since S is origin-symmetric and convex, $-x \in S$ and the line segment connecting y and $-x$ is fully contained in S. Let z be the midpoint of this line. Then

$$z = \frac{-x + y}{2} = \frac{-x + (x + (2i, 2j))}{2} = (i, j)$$

is a lattice point. So S contains some nonzero lattice point, which finishes the proof. □

The general theorem applies to any lattice Λ:

Theorem 2.2. *Let Λ be a lattice and S a bounded origin-symmetric convex subset in $\mathbb{R}^n$. If $\text{vol}(S) > 2^n \det(\Lambda)$, then S contains at least one point in $\Lambda \setminus \{0\}$.*

Proving this reduces to the case of Theorem 2.1 after applying some affine mapping. This helps us prove some classical results in elementary number theory, like the two-square and four-square theorems.

3 Dirichlet Approximation Theorem

The Pigeonhole Principle often appears when needed to obtain bounds with a finite number of objects. Consider for example, the following problem, which appeared on the 2006 Putnam Exam:

Show for every $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^n$, there exists nonempty $S \subseteq X$ and $m \in \mathbb{Z}$ such that

$$\left| m + \sum_{s \in S} s \right| \leq \frac{1}{n+1}.$$

Proof. The idea is to take $s_i = x_1 + \cdots + x_i$ for $i \in \{0, \ldots, n\}$ and consider the fractional part $\{s_i\} = s_i - \lfloor s_i \rfloor$ of each s_i. In particular, sort the $\{s_i\}$ in ascending order and denote this new list as $t_0, \ldots, t_n$. Now, since $t_0 = 0$, we have that the sum

$$(t_1 - t_0) + \cdots + (t_n - t_{n-1}) + (1 - t_n)$$

adds up to 1. So by the Pigeonhole Principle, one of the differences is $\leq \frac{1}{n+1}$. It is then a matter of casework (conditioning on whether the difference is $1 - t_n$) to obtain the set S. □

One of the powers of the Pigeonhole Principle is exactly this idea: to discretize a space so that we can find the existence of some suitable object. Now we state Dirichlet Approximation. The proof follows a similar flavor to that of the above problem.

Theorem 3.1 (Dirichlet Approximation)**.** *For every number $\alpha \in \mathbb{R}$ and every $n \in \mathbb{Z}^+$, there is some $q \in \{1, \ldots, n\}$ and $p \in \mathbb{Z}$ such that*

$$|q\alpha - p| < \frac{1}{n}$$

Proof. Consider the set of numbers $\{s\alpha\} \in [0, 1)$ for each $s \in \{0, 1, \ldots, n\}$, where again $\{s\alpha\} = s\alpha - \lfloor s\alpha \rfloor$ is the fractional part. Now note that

$$[0, 1) = \left[0, \frac{1}{n}\right) \cup \left[\frac{1}{n}, \frac{2}{n}\right) \cup \cdots \cup \left[\frac{n-1}{n}, 1\right).$$

Note that there are n disjoint intervals here. So there must be two such $\{s\alpha\}$ in the same interval. In particular, applying Pigeonhole here, there must be $s_1, s_2 \in \{0, 1, \ldots, n\}$ with $s_1 < s_2$ where

$$|\{s_2\alpha\} - \{s_1\alpha\}| < \frac{1}{n}.$$

Now let $q = s_2 - s_1$ and $p = \lfloor s_2\alpha \rfloor - \lfloor s_1\alpha \rfloor$. Then $|q\alpha - p| < \frac{1}{n}$. That $q \in \{1, \ldots, n\}$ is clear as $s_1 < s_2$ and $s_1, s_2 \in \{0, 1, \ldots, n\}$. This finishes the proof of Dirichlet's approximation theorem! □

What does this tell us? Well, it says that we can approximate any real number with a sequence of good rational approximations! A direct consequence of this is that the inequality

$$0 < \left| \alpha - \frac{p}{q} \right| < \frac{1}{q^2}$$

is satisfied by infinitely many $p, q \in \mathbb{Z}$. Indeed, we can view this result as some stronger form of the rationals being dense in $\mathbb{R}$. An interesting note: this approximation theorem can be proven using Minkowski's theorem as well. We leave the details to the reader. (Hint: construct a bounded set in the plane that describes $|\alpha x - y|$. This set should contain a nonzero lattice point!)

4 Solvability of Pell's Equation

Pell's equation is a certain Diophantine equation of the form $x^2 - Dy^2 = 1$ for nonsquare $D \in \mathbb{Z}^+$. It is a foundational equation in Algebraic Number Theory. In particular, notice that

$$x^2 - Dy^2 = \big(x + y\sqrt{D}\big)\big(x - y\sqrt{D}\big)$$

is the norm of $x + y\sqrt{D}$ in the ring $\mathbb{Z}[\sqrt{D}]$. Thus, solutions to Pell's equation give the norm 1 units in $\mathbb{Z}[\sqrt{D}]$. One proof that there exists some nontrivial solution (solutions other than $(\pm 1, 0)$) is the following elementary proof, applying the Pigeonhole Principle three times.

Theorem 4.1. *If $D \in \mathbb{Z}^+$ is nonsquare, then Pell's equation $x^2 - Dy^2 = 1$ has at least one nontrivial solution.*

The proof follows Dirichlet's argument, which is a simplification of a proof by Lagrange.

Proof. Fix $B_1 > 1$. Employ Dirichlet Approximation to get $a_1, b_1 \in \mathbb{Z}^+$ with $\big|a_1 - b_1\sqrt{D}\big| < \frac{1}{B_1} < \frac{1}{b_1}$. Now take $B_2 > b_1$ with $\frac{1}{B_2} < \big|a_1 - b_1\sqrt{D}\big|$, and obtain another pair (a_2, b_2) with $\big|a_2 - b_2\sqrt{D}\big| < \frac{1}{B_2} < \frac{1}{b_2}$. Repeating this, we get an infinite sequence $\frac{a_j}{b_j}$ of increasingly tight approximations to $\sqrt{D}$.

Now, we aim to show that there are infinitely many $a - b\sqrt{D} \in \mathbb{Z}[\sqrt{D}]$ with absolute norm at most $3\sqrt{D}$. This will form the basis of our next Pigeonhole. Let $a, b \in \mathbb{Z}$ such that $\big|a - b\sqrt{D}\big| < \frac{1}{b}$. Then

$$\big|a + b\sqrt{D}\big| \le \big|a - b\sqrt{D}\big| + \big|2b\sqrt{D}\big| \le \big|3b\sqrt{D}\big|$$

and

$$|a^2 - Db^2| = \big|a + b\sqrt{D}\big|\big|a - b\sqrt{D}\big| \le \frac{1}{b} \cdot \big|3b\sqrt{D}\big| = 3\sqrt{D}$$

so there are infinitely many $a - b\sqrt{D}$ with absolute norm at most $3\sqrt{D}$.

Now, by the Pigeonhole Principle, there must be some $n \in \mathbb{Z}$ with $n < 3\sqrt{D}$ such that $a^2 - Db^2 = n$ is true for infinitely many positive integer pairs (a, b). Now, consider reducing each pair modulo $|n|$ to get pairs $(a \bmod |n|, b \bmod |n|)$. Applying Pigeonhole again, these pairs must repeat, so in particular, there are distinct positive integer solutions $(a_1, b_1), (a_2, b_2)$ satisfying $a_1^2 - Db_1^2 = a_2^2 - Db_2^2 = n$, and $a_1 \equiv a_2 \bmod |n|$, $b_1 \equiv b_2 \bmod |n|$.

Write $a_1 = a_2 + nk_1$ and $b_1 = b_2 + nk_2$. Then we have

$$\begin{aligned} &a_1 + b_1\sqrt{D} \\ &\quad = a_2 + nk_1 + b_2\sqrt{D} + nk_2\sqrt{D} \\ &\quad = a_2 + b_2\sqrt{D} + n\big(k_1 + k_2\sqrt{D}\big) \\ &\quad = a_2 + b_2\sqrt{D} + \big(a_2^2 - b_2^2 D\big)\big(k_1 + k_2\sqrt{D}\big) \\ &\quad = \big(a_2 + b_2\sqrt{D}\big)\big(1 + \big(a_2 - b_2\sqrt{D}\big)\big(k_1 + k_2\sqrt{D}\big)\big). \end{aligned}$$

Similarly, we have

$$a_1 - b_1\sqrt{D} = \big(a_2 - b_2\sqrt{D}\big)\big(1 + \big(a_2 + b_2\sqrt{D}\big)\big(k_1 - k_2\sqrt{D}\big)\big).$$

Now, we can combine like terms to write these as

$$\begin{aligned} a_1 + b_1\sqrt{D} &= \big(a_2 + b_2\sqrt{D}\big)\big(a + b\sqrt{D}\big) \\ a_1 - b_1\sqrt{D} &= \big(a_2 - b_2\sqrt{D}\big)\big(a - b\sqrt{D}\big). \end{aligned}$$

Now multiply these to get

$$n = \big(x_1 + y_1\sqrt{D}\big)\big(x_1 - y_1\sqrt{D}\big) = n(a^2 - db^2)$$

So we recover some $a, b \in \mathbb{Z}$ with $a^2 - db^2 = 1$ as desired. It remains to show that this (a, b) is nontrivial, which is a routine check. In particular, if $(a, b) = (1, 0)$, we would find that this contradicts $(a_1, b_1) \neq (a_2, b_2)$, and if $(a, b) = (-1, 0)$, this contradicts that $a_1, a_2 > 0$. So we are done. □

The same proof (modified slightly) gives infinite solutions to the Pell's equation. Such a fact reveals that there are infinitely many units in real quadratic rings $\mathbb{Z}[\sqrt{D}]$ for nonsquare $D \in \mathbb{Z}^+$.

5 A Theorem of van der Waerden

We come back to Ramsey Theory. Recall the motivating philosophy: every very large structure contains a large well-organized substructure. Terrence Tao in his book *Additive Combinatorics* [TV06] said that you could view Ramsey Theory as "the set of generalizations and repeated applications of the pigeonhole principle." Indeed, the picture is clear: when our objects get too large, eventually, things repeat (and end up in the same pigeonhole).

Van der Waerden's theorem [Van27] echoes this philosophy.

Theorem 5.1. *Given $r, k \in \mathbb{Z}$, there is some $N \in \mathbb{Z}^+$ such that if the integers $\{1, 2, \ldots, N\}$ are colored with one of r different colors, then there are at least k integers in arithmetic progression (AP) whose elements are of the same color.*

This is quite profound: arbitrarily long AP substructure exists within the integers! The least N here is the van der Waerden number $W(r, k)$. Determining these values for most r and k is still open. We prove the special case for $W(2, 3)$:

Theorem 5.2. *There is some number $N \in \mathbb{Z}^+$ such that if the integers $\{1, 2, \ldots, N\}$ are colored red or blue, then there must be at least one monochromatic 3-AP. In particular, $W(2, 3) \le 325$.*

Proof. Let $c(n)$ be a coloring of $A = \{1, \ldots, 325\}$. We find three elements in AP. Divide A into 65 blocks of the form $\{5b+1, \ldots, 5b+5\}$ for $b \in \{0, 64\}$. Now there are 32 possible colorings for each block, so by the Pigeonhole principle, every 33 consecutive blocks must have two blocks colored exactly the same. In particular, there are $b_1, b_2 \in \{0, \ldots, 32\}$ with

$$c(5b_1 + j) = c(5b_2 + j)$$

for all $j \in \{1, \ldots, 5\}$. Now, among $5b_1+1, 5b_1+2, 5b_1+3$, there are at least two that are of the same color. Call these $5b_1 + a_1$ and $5b_1 + a_2$ with $a_1 < a_2$ and without loss of generality, let them be colored red.

Now take $a_3 = 2a_2 - a_1$. If $5b_1 + a_3$ is red, then we are done.

Now, suppose $5b_1+a_3$ is blue. Then $5b_2+a_3$ is also blue. Take $b_3 = 2b_2-b_1$. We know $b_3 \leq 64$ so $5b_3+a_3 \leq 325$. If $5b_3 + a_3$ is red, then $5b_1 + a_1, 5b_2 + a_2, 5b_3 + a_3$ is a red arithmetic progression. Otherwise, $5b_1 + a_3, 5b_2 + a_3, 5b_3 + a_3$ is a blue arithmetic progression. This finishes the proof. □

Over the years, mathematicians have extended and generalized van der Warden's theorem in many ways. Szemerédi's theorem generalizes this to subsets of integers with positive density, and of course, the celebrated Green–Tao theorem which extends Szemerédi's theorem, stating that the sequence of primes contains arbitrarily long arithmetic progressions.

6 Erdős–Szekeres Theorem

The Erdős–Szekeres theorem [ES35] is a result in Ramsey Theory, that offers an incredibly slick application of the Pigeonhole Principle.

Theorem 6.1. *Any sequence of distinct real numbers with length $> (r-1)(s-1)$ contains a monotonically increasing subsequence of length r or a monotonically decreasing subsequence of length s.*

Proof. Suppose we have some real sequence with length $(r-1)(s-1)+1$. Now, label each n_i in the sequence with (a_i, b_i), where a_i and b_i are the lengths of the longest monotonically increasing/decreasing subsequences ending with n_i. Now, notice that the labels for each of these numbers must be different. In particular, for any $i < j$, when $n_i \leq n_j$, then $a_i < a_j$, and when $n_i \geq n_j$, then $b_i < b_j$. Now, if $a_i \leq r-1$ and $b_i \leq s-1$, then there are $(r-1)(s-1)$ possible labels. Thus, by the Pigeonhole Principle, there must be some i for which $a_i \geq r$ or $b_i \geq s$. If the former, then n_i is part of a monotonically increasing sequence of length $\geq r$, and if the latter, then n_i is part of a increasing sequence of length $\geq s$. This finishes. □

Erdős–Szekeres generalizes to constant sequences too. Consider, for example, the following statement.

Theorem 6.2. *Any sequence of $(n-1)^3+1$ real numbers contains a constant, increasing, or decreasing subsequence of length n.*

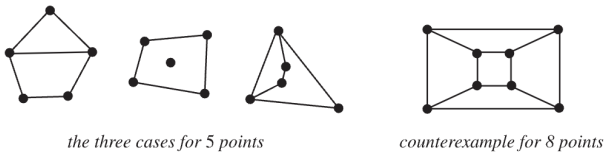
Proof. Suppose we have some sequence of real numbers of length $(n-1)^3+1$. Now, label each n_i in the sequence with (a_i, b_i, c_i), where a_i, b_i, c_i are the lengths of the longest monotonically increasing/decreasing, or constant subsequences ending with n_i, respectively. Now notice that the labels for each of these numbers must be different. In particular, for any $i < j$, when $n_i < n_j$, then $a_i < a_j$, when $n_i > n_j$, then $b_i < b_j$, and when $n_i = n_j$, then $c_i < c_j$. Now, if $a_i \leq n-1, b_i \leq n-1$, and $c_i \leq n-1$, then there are $(n-1)^3$ possible labels for the numbers until n_i. Thus, by the Pigeonhole Principle, there must be some $i < j$ for which $(a_i, b_i, c_i) = (a_j, b_j, c_j)$, but we've already seen why this is not possible. Thus, there is some i for which $a_i \geq n$, $b_i \geq n$, or $c_i \geq n$. This finishes, since if $a_i \geq n$, then n_i is part of an increasing subsequence of length n, if $b_i \geq n$, then n_i is part of a decreasing subsequence of length n, and if $c_i \geq n$, then n_i is part of a constant subsequence of length n. □

7 The Happy Ending Problem

We end with a discussion that is not so directly related to Pigeonhole as it is related to Erdős–Szekeres.

Theorem 7.1. *Among any five points in the plane, four are in convex position.*

To see this visually, consider the plot in Figure 6. No matter how we place our five points, we can always connect four into a convex quadrilateral. The proof is straightforward.



the three cases for 5 points *counterexample for 8 points*

Figure 6: Examples of the Happy Ending Problem [Pet13]

Proof. We'll show that either the five points form a convex pentagon, or some subset of four points among them forms a convex quadrilateral. If four or five of the points are vertices of the convex hull of these points, then we are trivially done. Otherwise, the convex hull will be in the form of a triangle with two points inside of it (otherwise, the convex hull would contain four or five points). In this case, we can choose the two inner points and one of the triangle sides to be in convex position. □

For a visual explanation of this proof, see [Pet13].

The problem extends: of interest is the least n points in the plane such that at least k are in convex position. Some known constants now are that nine points give five

in convex position and seventeen points give six in convex position. This problem is known as the Happy Ending Problem. Erdős worked on this problem with Szekeres, and he coined this the Happy Ending Problem, since the research eventually led to the marriage of George Szekeres and Esther Klein, a happy ending.

References

[ES35] Paul Erdös and George Szekeres. "A combinatorial problem in geometry". In: *Compositio mathematica* 2 (1935), pp. 463–470.

[Pet13] Ivars Peterson. "Planes of Budapest". In: *MAA Online* (2013).

[TV06] Terence Tao and Van H Vu. *Additive combinatorics*. Vol. 105. Cambridge University Press, 2006.

[Van27] Bartel Leendert Van der Waerden. "Beweis einer baudetschen vermutung". In: *Nieuw Arch. Wiskunde* 15 (1927), pp. 212–216.