In mathematics, the truth of propositions is established with a *proof*. If there is no proof then the proposition cannot be used. This lecture is about what a proof is and how you may go about finding one.

There are clear and stringent rules about what qualifies as a mathematical proof. Two economists may debate vigorously about economic truth: One could make a case that raising taxes would improve the economy, while the other one might argue that lowering them would have that effect. A prosecution lawyer might try to convince a jury that the accused broke the law, while a defence lawyer would argue that he didn't. In contrast, mathematicians overwhelmingly agree about what is true and what is false: Every claim must come with a proof. A mathematician with sufficient training in his or her specialty ought to be able to verify the correctness of the claimed proof, or to spot a mistake if the proof is incorrect.[1]

While *verifying* the correctness of a proof is a skill you can master with some effort and self-discipline, *creating* proofs is a different story. Mathematics is full of propositions that nobody knows how to prove. For some, like Goldbach's conjecture, the search for a proof has been going on for hundreds of years. In 1998 the Clay Mathematics Institute collected seven famous propositions and offered a 1 million US Dollar prize for each proof. So far only one has been proven.[2]

Coming up with proofs is not completely dark magic. There are general guidelines for what kind of strategy might help with what type of proposition. However, it is important to remember that — unlike, say, the recipe you learn in school for calculating square roots — these are not guaranteed to succeed.

1 What is a proof?

A *proof* of a proposition is a sequence of *logical deductions* from *axioms* and previously proved propositions that concludes with the proposition in question.

Instead of trying to explain, in general, what axioms and logical deductions are, let us see an example of a proof. For now we won't ask how someone came up with it.

First we need to state the proposition that we intend to prove. A proposition for which a (correct) proof is given is called a *theorem*. Before we state our theorem, we need to *define* a few concepts that will show up in it.

The theorem I have in mind is about friendships. Let's call two people *strangers* if they are not friends. A *group of friends* is a collection of people in which every two of them are friends, and a *group of strangers* is a collection of people in which every two are strangers.

Theorem 1. *Any group of 6 people includes a group of 3 friends or a group of 3 strangers.*

Proof. Let Alice be one of the six people. The proof is by case analysis. We consider two cases:

- **Case 1:** Alice is friends with at least 3 other people.
- **Case 2:** Alice is a stranger to at least 3 other people.

One of these two cases must hold: There are 5 people besides Alice, and these are divided into friends of Alice and strangers to Alice. The bigger group has at least 3 people.

Now let's discuss Case 1. Let's give the group of people who are friends with Alice a name – call it F for "friends of Alice". We consider two subcases:

[1] An interesting recent case is Mochizuki's claimed proof of the ABC conjecture.
[2] The prize money was refused.

- **Subcase 1.1:** At least two people within F are friends. Let's call them Bob and Charlie. Then Alice, Bob, and Charlie form a group of 3 friends.
- **Subcase 1.2:** No two people within F are friends. Take any three people in F. They form a group of 3 strangers.

We conclude that the Theorem holds in Case 1.

We are left with Case 2. Let's give the group of people who are strangers to a a name – call it S for "strangers to Alice". We consider two subcases:

- **Subcase 2.1:** At least two people within S, call them Bob and Charlie, are strangers to one another. Then Alice, Bob, and Charlie form a group of 3 strangers.
- **Subcase 2.2:** No two people within S are strangers. Take any three people in S. They form a group of 3 friends.

The theorem also holds in Case 2, and so it holds in all the cases. □

Theorem 1 talks about *collections* of people and *friendships* among people. The axioms are facts that we view as self evident. For example, the following two axioms are implicitly used in our proof:

Axiom 1. *For any two people x and y, if x and y are friends, then y and x are also friends.*

Axiom 2. *If a group that has Alice in it has 6 people, then there are 5 people other than Alice.*

Let us now look at the proof. This is a proof by *case analysis*. Case analysis is a logical *deduction rule*. It says that we can prove a proposition P like this: Split all logical possiblities into two cases C_1 and C_2, prove that C_1 and C_2 cover all possibilities, prove that C_1 implies P, and prove that C_2 implies P.

$$\frac{C_1 \text{ OR } C_2 \qquad C_1 \longrightarrow P \qquad C_2 \longrightarrow P}{P}$$

It should be clear that this deduction rule is *sound* – it only proves true statements – but if in doubt you can always write out a truth table. Let's do it just this once. Here, $\star$ is shorthand for $(C_1 \text{ OR } C_2) \text{ AND } (C_1 \longrightarrow P) \text{ AND } (C_2 \longrightarrow P)$.

P	C_1	C_2	$\star$	$\star \longrightarrow P$
T	**T**	**T**	**T**	**T**
T	**T**	**F**	**T**	**T**
T	**F**	**T**	**T**	**T**
T	**F**	**F**	**F**	**T**
F	**T**	**T**	**F**	**T**
F	**T**	**F**	**F**	**T**
F	**F**	**T**	**F**	**T**
F	**F**	**F**	**F**	**T**

Next, the proof has to tell us what the two cases (C_1 and C_2) are. Here, C_1 is the proposition "Alice is friends with at least 3 people" and C_2 is the proposition "Alice is strangers to at least 3 people."

Now, we expect to be given proofs of the propositions C_1 OR C_2 (the cases cover all possibilities), $C_1 \longrightarrow P$ (the theorem holds in case 1) and $C_2 \longrightarrow P$ (the theorem holds in case 2). By the case analysis deduction rule, once we validate these proofs we'll be sure that Theorem 1 is true.

Let's start with C_1 OR C_2. This says "Alice is friends with 3 people, or Alice is a stranger to 3 people". The next sentence explains why this must be true: Among friends and strangers to Alice there are 5 people, so the bigger of the two groups must contain at least $5/2 = 2.5$ people. As 2.5 is not an integer, there must be at least 3 people in this group.

This appears like a sensible argument. It is fine to leave it at that. But how does it follow from our axioms? We will see so shortly. For now let us "package" this proposition C_1 OR C_2 as a *lemma* and give its proof later:

Lemma 2. *In every group of six people including Alice, Alice is friends with at least three or stranger to at least three.*

A lemma is just like a theorem – a proposition with a proof. Usually, the theorems are the ones we are really interested in, and lemmas are intermediate propositions that are used in the proofs of theorems.

Now comes the proof of the theorem in Case 1. For this part, we can *assume* C_1: Alice is friends of at least 3 people. You can think of it as another axiom, but just for this part of the proof. We divide C_1 into two subcases: Those 3 contain a pair of friends (C_{11}), or they are all strangers to one another (C_{12}). Clearly, C_{11} OR C_{12} always holds. Next, we see that C_{11} implies the theorem (analysis of Subcase 2.1) and C_{12} implies the theorem (anaysis of Subcase 2.2). So the theorem holds in all subcases of Case 1.

The last part of the proof is structurally similar: By the same type of reasoning, the theorem is shown to hold in all subcases of Case 2. A mathematics book may omit this part altogether and say "Case 2 is proved analogously to Case 1". Before you become practiced at proofs, it may be better to refrain from doing this and work out all the cases in detail.

Before we embark on the challenging task of discovering proofs, let us have one final word about axioms. What, exactly, are we allowed to assume as an axiom or as a previously proved proposition when we prove a theorem? For us, this will consist of the "common sense" facts you have learned in school, as well as propositions we have previously proved in class. For example, if you are asked to prove a theorem in your homework, it is okay to use Theorem 1 as a previously proved statement.

In the beginning of the 20th century logicians spent considerable effort trying to agree on a small collection of axioms that ought to be enough to prove all known mathematics. One of the proposals are the so-called ZFC axioms of set theory; you can read about them in the textbook. In principle, you can *define* any mathematical object as a set of some kind and then write any proof relying on just these nine axioms. In practice, deriving a proposition as simple as $\forall n\colon n+n=2\times n$ from the ZFC axioms may take many pages of proof and explanation, so we won't be doing that.

2 How to prove it

Let's start by proving a simple theorem:

Theorem 3. *The sum of two even integers is even.*

How do we go about proving such a theorem? First, let us unwind this statement in terms of quantifiers:

For all integers m and n, if m is even and n is even, then $m+n$ is even.

This is a universally quantified proposition about two integers, which we call m and n. We need to show that following implication:

$$(m \text{ is even}) \text{ AND } (n \text{ is even}) \longrightarrow (m+n \text{ is even}).$$

Let's assume that m is even and n is even. This means there exist integers a and b such that $m=2a$ and $n=2b$. But then $m+n=2a+2b=2(a+b)$, so $m+n$ is also twice an integer, and therefore even.

This is a common method for proving a statement of the form "If P then Q". We assume P, do a bit of reasoning, see what consequences we get, and eventually hope to end up with Q.

Once you figured out the reasoning, here is how you may *write* this proof:

Proof of Theorem 3. Let us call the two integers m and n. Assume m is even and n is even. Then there exist integers a and b such that $m=2a$ and $n=2b$. It follows that $m+n=2a+2b=2(a+b)=2c$, where $c=a+b$. Therefore m is also even. □

Let's do another one:

Theorem 4. *The product of two odd integers is odd.*

We follow the same pattern.

Proof. Call the integers m and n. Since m and n are both odd, we can write $m = 2a + 1$ and $n = 2b + 1$ for some integers a and b. Then

$$mn = (2a + 1)(2b + 1) = (2a)(2b) + 2a + 2b + 1 = 2(2ab + a + b) + 1 = 2c + 1$$

where $c = 2ab + a + b$. It follows that mn is also odd. □

In these examples, the path to the proof was clear; we just need to move along (and avoid making mistakes in the process). Other times we need to do some "scratch work," that is reasoning which won't make it into the proof but helps us figure things out. Here is one such example:

Theorem 5. *The square of an odd number is of the form $8k + 1$ for some integer k.*

Let's call our number n. Since n is odd, we can write $n = 2t + 1$ for some integer t. Then

$$n^2 = (2t + 1)^2 = 4t^2 + 4t + 1.$$

Why should this be of the form $8k + 1$? We want to show that given t, we can always find a k such that

$$4t^2 + 4t + 1 = 8k + 1$$

which we can simplify to $t^2 + t = 2k$. Namely, we are now left to show that $t^2 + t$ is always even. To make sure we are on the right track, we can try some examples: $1^2 + 1 = 2$, $2^2 + 2 = 4 + 2 = 6$, $3^2 + 3 = 9 + 3 = 12$, all even.

It seems there are two cases: t is even, in which case so is t^2 and also $t^2 + t$, or t is odd, in which case so is t^2, and so $t^2 + t$ is also even. This covers all possibilities. We now need to summarize them nicely into a proof.

Before we do so, let's revisit the last step and see if there is an easier way to explain why $t^2 + t$ is always even. If we factor this expression, we get $t^2 + t = t(t + 1)$. Now if t is even, so is $t(t + 1)$, and if t is odd, then $t + 1$ is even and so is $t(t + 1)$. This simplifies our case analysis a bit.

Proof of Theorem 5. Assume n is odd, so we can write $n = 2t + 1$ for some integer t. Then

$$n^2 = (2t + 1)^2 = 4t^2 + 4t + 1 = 4t(t + 1) + 1$$

We now prove the theorem by case analysis.

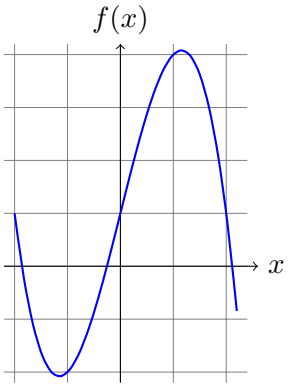
- **Case 1:** t is even. Then we can write $t = 2r$ for some r and $4t(t + 1) + 1 = 8r(t + 1) + 1 = 8k + 1$ for $k = r(t + 1)$.
- **Case 2:** t is odd. Then $t + 1 = 2r$ for some r and $4t(t + 1) + 1 = 8tr + 1 = 8k + 1$ for $k = tr$.

The two cases cover all possibilities and the claim holds in each case. □

Here is another one where some scratch work of a different sort is helpful:

Theorem 6. *If x is a real number with $0 \leq x \leq 2$, then $-x^3 + 4x + 1 > 0$.*

This is a universally quantified proposition and there are infinitely many x to consider, so we need to be a bit clever here. Fortunately, we live in an age of computers so we start by plotting the graph of $f(x) = -x^3 + 4x + 1$:



This picture is not a proof; we must derive the theorem by logical deduction. So where do we start?

From the picture we can see that in the range of interest $0 \le x \le 2$, $f(x)$ is not only greater than zero, but always exceeds 1, namely

$$\text{If } 0 \le x \le 2, \text{ then } -x^3 + 4x + 1 \ge 1.$$

The predicate $-x^3 + 4x + 1 \ge 1$ is the same as $-x^3 + 4x \ge 0$. But now we can *factor* the left hand side as

$$-x^3 + 4x = x(4 - x^2) = x(2 - x)(2 + x).$$

When x is between 0 and 2, all of the terms $x, 2 - x, 2 + x$ are nonnegative, and so must be their product. There!

We are not finished yet. We must now summarize our conclusions neatly into a proof with clear logical deductions.

Proof of Theorem 6. Assume x is a real number such that $0 \le x \le 2$. Then all of the numbers x, $2 - x$, and $2 + x$ must be nonnegative. It follows that $x(2 - x)(2 + x) \ge 0$. Multiplying out the left hand side, we obtain $-x^3 + 4x \ge 0$. Therefore $-x^3 + 4x + 1 \ge 1 > 0$, as claimed. □

3 Some proof patterns

The contrapositive

The *contrapositive* of a proposition of the form $P \longrightarrow Q$ is the proposition (NOT Q) $\longrightarrow$ (NOT P). The two are logically equivalent. You can draw your own truth table to verify this.

A number r is *rational* if we can write $r = n/d$ where both n and d are integers, e.g. 1/2, 3/2, 5/17, 8/16. A number is *irrational* if it is not rational.

Theorem 7. *Assume $r \ge 0$. If r is irrational, then $\sqrt{r}$ is irrational.*

Let us try to prove this theorem. We assume r is irrational. So r cannot be written as a fraction n/d for any integers n and d. Where do we go from here? An assumption like this doesn't tell us much about $\sqrt{r}$, so it is not clear how to reach any conclusion about it. Instead, let us try the contrapositive:

> Assume $r \ge 0$. If $\sqrt{r}$ is rational, then r is rational.

This is now much easier to prove.

Proof of Theorem 7. We prove the contrapositive. Assume $r \ge 0$ and $\sqrt{r}$ is rational. Then we can write $\sqrt{r} = n/d$ for some integers n and d. It follows that $r = n^2/d^2$, and so r is also rational. □

Sometimes the "if... then" structure of the proposition may not be completely apparent as in Lemma 2 using this method. Recall what the lemma says:

In every group of six people including Alice, Alice is friends with at least three or stranger to at least three.

Proof of Lemma 2. Suppose that Alice is friends with at most two others and stranger to at most two others. Then the group, which comprises Alice, her friends, and the strangers, consists of at most 1+2+2 = 5 people. Therefore the group does not have six people. □

Proving equivalences

A common way to prove a proposition of the form P IFF Q, that is, an equivalence, is to prove separately that P implies Q and that Q implies P:

$$\frac{P \longrightarrow Q \qquad Q \longrightarrow P}{P \text{ IFF } Q}$$

Here is an example.

Theorem 8. *For every integer n, n^2 is even if and only if n is even.*

Proof. First, we prove that if n is even then n^2 is even. If n is even, we can write $n = 2k$ for some integer k, so $n^2 = 4k^2 = 2(2k^2)$, which is also even.

Now, we prove that if n^2 is even then n is even. We prove the contrapositive: If n is odd, then n^2 must also be odd. In Theorem 5 we showed that if n is odd then n^2 is of the form $8k + 1 = 2(4k) + 1$, which is an odd number. □

Proof by contradiction

Say you want to prove a proposition P. In a proof by contradiction, you start by assuming P is *false*, and then you deduce that this assumption applies a falsehood. So P must have been true:

$$\frac{(\text{NOT } P) \longrightarrow \mathbf{F}}{P}$$

Here is a famous example:

Theorem 9. *$\sqrt{2}$ is irrational.*

This is a universally quantified statement: For all n and d, we cannot write $\sqrt{2}$ as n/d. You could try different choices of n and d and see for yourself that they don't work. Where to go from here?

Proof. Assume, for contradiction, that $\sqrt{2}$ is rational. Then we can write $\sqrt{2} = n/d$ where n and d are integers. Furthermore, let's take n and d so that they have no common factor greater than 1, so the fraction is written in lowest terms.

Squaring both sides, we obtain $2 = n^2/d^2$ and so $n^2 = 2d^2$. So n^2 is even. Then n must also be even (by Theorem 8), and so n^2 is a multiple of 4. Because $2d^2 = n^2$, d^2 must be even, so d is also even.

We conclude that both n and d are even. But we assumed that they have no common factor greater than 1. This contradicts our assumption that $\sqrt{2}$ is rational. □

Proofs by contradiction can be confusing because you begin by assuming a statement that is, in fact, false. So some of the claims you will be making inside the proof will also be false. You need to keep in mind at all times that you are operating under a false assumption, and intermediate claims, like "d is even", are only true within that context. Because of this confusion, I generally recommend proofs by contradiction only as a last resort, when all your other attempts at a proof have failed.

Here is another example. The object we will be interested in is a rectangular table populated with distinct numbers. A number in the table is a *saddle* if it is the largest in its column and the smallest in its row. For example `6` is a saddle in this table:

```
9 6 8
3 2 1
7 4 5
```

Moreover `6` is the only saddle; there is no other in this table. This is no surprise:

Theorem 10. *Every table with distinct numbers has at most one saddle.*

You can convince yourself that this is reasonable by trying out more tables. Some will have one saddle, some will have zero, but none will have more than one. How do we prove it?

Proof. Assume, for contradiction, that the table has two saddles x and y. We proceed by cases.

- x and y are in the same row. Then we have two smallest numbers in the same row, contradicting that all numbers are distinct.
- x and y are in the same column. Then we have two largest numbers in the same column, contradicting that all numbers are distinct.
- x and y are in different rows and columns. Let u be the unique number in the table in the same row as x and in the same column as y. Since x and y are saddles, $u > x$ and $u < y$. Now let v be the number in the same column as x and in the same row as y. Since x and y are saddles, $v < x$ and $v > y$. Putting these inequalities together we obtain

  $$x < u < y < v < x$$

  which is impossible. This contradicts our assumption that x and y are both saddles. □

4 More proofs

We continue with the same setup. Take a table and sort its rows. They sort its columns. Are the rows still sorted? Let's try it out:

```
9 6 7          6 7 9          1 3 4
1 8 5  sort rows→  1 5 8  sort cols→  2 5 8
4 3 2          2 3 4          6 7 9
```

The rows remain sorted. How do we explain this?

Conjecture 1. *In any table of distinct numbers, after sorting the rows and then the columns, the rows remain sorted.*

The numbers in the first table are in arbitrary order. After sorting the rows, we know that all consecutive pairs in a given row must be in increasing order:

```
6 < 7 < 9
1 < 5 < 8
2 < 3 < 4
```

We want to argue that, after sorting the columns, the rows are still in order. Namely, in each row the first number is smaller than the second one and the second number is smaller than the third one. Now, if we

want to argue that the first number in each row remains smaller than the second, the third column of the table is not relevant. We can focus on the sub-table spanned by the first two columns only:

$$\begin{array}{ccc} 6 & < & 7 \\ 1 & < & 5 \\ 2 & < & 3 \end{array} \xrightarrow{\text{sort cols}} \begin{array}{ccc} 1 & < & 3 \\ 2 & < & 5 \\ 6 & < & 7 \end{array}$$

Why do the rows remain in order after sorting the columns? Let's start by looking at the first row after sorting. This row will contain the smallest number in the first column (1) followed by the smallest number in the second column (3). Which of the two will be larger? Well the smallest number in the second column must be larger than *at least one* number in the first column, namely the number that was next to it (2). It must therefore be the larger of the two:

$$\begin{array}{ccc} 6 & < & 7 \\ 1 & < & 5 \\ 2 & < & 3 \end{array} \xrightarrow{\text{sort cols}} \begin{array}{ccc} 1 & < & 3 \\ 2 & & 5 \\ 6 & & 7 \end{array}$$

Now let's look at the second row in the sorted table. This row will comprise the second smallest numbers in both columns. Which is larger? We can reason in a similar manner: The two smallest numbers in column 2 are both larger than their neighbors in column 1 ($5 > 1$ and $3 > 2$). So the second smallest number in column 2 (5) must be larger than the second smallest number in column 1 (2). The second row will therefore remain sorted.

We now have all the ingredients for the proof of our theorem:

Theorem 11. *In any table of distinct numbers with sorted rows, after sorting the columns, the rows remain sorted.*

First we prove a lemma:

Lemma 12. *Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be two columns of numbers, all of them distinct. Assume $x_i < y_i$ for all i (between 1 and n). Then for every k (between 1 and n), the k-th smallest number in the x-column is smaller than the k-th smallest number in the y-column.*

Proof. Any k numbers in the y-column are larger than the k numbers in the x-column that are in the same rows. Therefore the k smallest numbers in the y-column are all larger than *some* k distinct numbers in the x-column. In particular, the k-th smallest number in the y column must be larger than at least k numbers in the x-column, so it must be larger than the k-th smallest number in the x-column. □

Proof of Theorem 11. Let x_{ij} and y_{ij} be the entry in row i and column j before and after sorting the columns, respectively. Let n be the number of rows and m be the number of columns. As the rows are initially sorted, $x_{ij} < x_{i(j+1)}$ for all i and j (where i ranges from 1 to n and j ranges from 1 to $m-1$). We apply Lemma 12 to the j-th and $(j+1)$-st columns $x_{1j}, \ldots, x_{nj}$ and $x_{1(j+1)}, \ldots, x_{n(j+1)}$. Lemma 12 says that the k-th smallest numbers in these two columns must be in increasing order. Therefore $y_{ij} < y_{i(j+1)}$ for all i (between 1 and n) and all j (between 1 and $m-1$). It follows that the rows of y are all sorted. □

Experiment and don't give up easily!

When you start out trying to prove a theorem, you rarely know what is the right method ahead of time. So play around, experiment, backtrack, and don't be afraid. The "correct" approach will often reveal itself after a few trials and errors.

Theorem 13. *There exist irrational numbers a and b such that a^b is rational.*

Where do we start? Let's try some examples. Well, the only number we know for sure is irrational is $\sqrt{2}$, so let's try setting $a = \sqrt{2}$ and $b = \sqrt{2}$. Is $\sqrt{2}^{\sqrt{2}}$ rational or irrational? It looks pretty irrational to me, so it doesn't seem that this should work out.[3]

Ah, but if $\sqrt{2}^{\sqrt{2}}$ is irrational, then we have one more irrational number to play with. So why don't we try $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}^{\sqrt{2}}$. Then

$$a^b = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}^{\sqrt{2}}} = \sqrt{2}^{\sqrt{2}\cdot(\sqrt{2})^{\sqrt{2}}} = \sqrt{2}^{\sqrt{2}^{\sqrt{2}+1}}$$

What a mess! Let's backtrack and try instead $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$. Then

$$a^b = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^{(\sqrt{2})^2} = \sqrt{2}^2 = 2$$

which is a rational number! Let's summarize this reasoning into a proof.

Proof. The proof is by case analysis.
Case 1: $\sqrt{2}^{\sqrt{2}}$ is rational. In this case, the theorem is true for $a = \sqrt{2}$ and $b = \sqrt{2}$.
Case 2: $\sqrt{2}^{\sqrt{2}}$ is irrational. In this case, the theorem is true for $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$ because $a^b = 2$. □

This type of proof is sometimes called a *win-win argument*. It doesn't matter if $\sqrt{2}^{\sqrt{2}}$ is rational or not. In either case you win. You may not always get this lucky, but it doesn't hurt to try.

5 How to write and present a proof

For this class, it is not enough that you know how to come up with proofs. You must also write and present them properly. Writing a proof is not easy. On the one hand the proof must be clear and precise. On the other hand, it should be easy to read and understand (by humans, not by machines). For general advice on how to write proofs, see Section 1.9 in your textbook.

Presenting a proof to others is also challenging. Your listeners may not be familiar with the notation. Steps in the proof that are obvious to you may take longer for others to grasp. So start from the beginning and go slowly; do not introduce too many new concepts at once; give examples along the way; and encourage questions from your audience.

6 Truth and proof*

Mathematical proofs are guaranteed to be *sound*: If you start with a set of axioms and rigorously follow the deduction rules, any proposition you derive must be true. In other words, everything that is provable is true. How about the converse: Do all true propositions have proofs?

This sounds like a trick question so let's try to unwrap its meaning. What it says is that for every proposition P, if P is true than P has a proof. The meaning of "P has a proof" should be clear: This means we can derive P from our axioms after some sequence of deductions. But what does "P is true" really mean?

[3] This part of the argument is not conclusive: "It looks pretty irrational" doesn't make a number irrational. Perhaps we'll come back to it later, but we might as well try something easier first.

Before we answer this question let's do an exercise. Suppose we want to figure out things about integers and we start with the following axioms:

$$\forall x\colon x + 0 = x \tag{A1}$$
$$\forall x \exists y\colon x + y = 0 \tag{A2}$$
$$\forall x, y\colon x + y = y + x \tag{A3}$$
$$\forall x, y, z\colon (x + y) + z = x + (y + z) \tag{A4}$$

Clearly these axioms hold true for the integers. Now I want to prove that

$$\forall x\colon x + x = 0 \longrightarrow x = 0. \tag{P}$$

Although proposition (P) is true, it can never be proved from axioms (A1-A4). The reason is that there is a world (logicians call this a *model*) in which axioms (A1-A4) are true, but proposition (P) is false. This is the world $\mathbb{Z}_2$ consisting of the elements 0 and 1 in which addition is specified by the formulas

$$0 + 0 = 0, \qquad 0 + 1 = 1, \qquad 1 + 0 = 1, \qquad \text{and} \qquad 1 + 1 = 0.$$

Any proposition we prove from axioms (A1-A4) must be true not only for the integers, but also for $\mathbb{Z}_2$, so in particular proposition (P) cannot be proved from axioms (A1-A4).

In this case, it is easy to explain what went wrong: Although axioms (A1-A4) are certainly true about the integers, they do not specify the integers completely because they also describe, for instance, $\mathbb{Z}_2$. The issue here is not the logic, but the axioms: We need more of them in order to "pin down" the integers.

More generally, suppose we start with some collection of axioms A. When can we hope to prove a given proposition P? At a minimum, we should ensure that P is true in *every world* in which the axioms A are all satisfied. Let's take, for example, the proposition

$$\forall x, z, w\colon z + x = w + x \longrightarrow z = w.$$

This one is true for both the integers and $\mathbb{Z}_2$, so (based on our experience so far) we may hope that it can be proved from axioms (A1-A4). Indeed, here is the proof: Starting with the assumption

$$z + x = w + x$$

we get that for every y

$$(z + x) + y = (w + x) + y$$

which, applying (A4) on both sides, can be rewritten as

$$z + (x + y) = w + (x + y).$$

Now choosing y as in axiom (A2) we get that $x + y = 0$ and so

$$z + 0 = w + 0$$

from where, after applying axiom (A1) on both sides we obtain the desired conclusion

$$z = w.$$

This example illustrates a general phenomenon called *completeness*: If a proposition P is true in every world in which the axioms A are also true, then P is provable from A.

To be precise, completeness says that P is provable from A together with a fixed collection of self-evident *logical axioms* by applying one of a few specific *deduction rules*. One collection of logical axioms

and deduction rules for which completeness holds is the Hilbert System. This system includes infinitely many logical axioms, but has only one deduction rule called *Modus Ponens*:

$$\frac{P \quad P \longrightarrow Q}{Q}. \tag{1}$$

Writing proofs in Hilbert System format is not particularly natural for humans, but can always be done in principle.

This sounds like very good news: As long as we start with a collection of axioms that accurately describe the world we have in mind, we can in principle prove everything from them (as long as it is true). For example, if we want to determine the truth of propositions about numbers (integers) that involve the symbols 0, 1, and +, then the axioms of Presburger arithmetic suffice.

Automated theorem proving To a student of mathematics like you, proving theorems is a creative, challenging, and (I hope) enjoyable activity. In principle, however, theorem proving can be done in a purely methodical way that requires no creativity whatsoever. Suppose you want to know if some proposition X about numbers is true or not. Take all pairs of axioms (the Presburger arithmetic axioms plus the Hilbert System logical axioms) of the form P and $P \longrightarrow Q$ and apply the deduction rule (1) to derive some theorems Q. Now take all pairs of axioms and theorems you have obtained so far and repeat the process. By completeness, every true proposition will eventually show up among your list of theorems. In particular, one of the propositions X (if it happens to be true) or NOT X (otherwise) will show up at some point.[4]

So if you want to prove theorems, all you have to do is write a computer program that takes as its input the axioms and the proposition under investigation and performs all of the above calculations. Why do we bother with all the wishy-washy proof strategies like the ones in these lecture notes and not simply prove all theorems in this automated manner? Not only would automation save us a huge amount of effort, but it would also eliminate the pesky mistakes that, every once in a while, make us come up with incorrect proofs.

You may guess that automated theorem proving of the type I described doesn't work so well in practice because it is way too slow. Suppose I wanted to prove a theorem (about numbers) like

$$\exists x, y\colon x = y + 1 \text{ AND } x + x = y + y + y + 1.$$

This one is pretty easy for a human to figure out. The computer, on the other hand, will keep spitting out theorems until, eventually, this particular one appears on the list. There is no way to know how long this is going to take, and for "random" theorems like this one you should be prepared to wait for a very very long time.

The theorem prover I described is particularly stupid in the sense that it doesn't try to mimic human reasoning at all, so theorems that may be of interest to humans will be lost in a sea of computer-generated junk. This has partly to do with the choice of axioms, the choice of deduction rules, and the order in which they are applied. The field of automated theorem proving is concerned with the design, implementation, and application of such systems. Some of these are completely automated, while others are interactive; when they get stuck, they ask the user to provide a hint.

Incompleteness There is another, much more surprising (although maybe less relevant in practice) obstacle to automated theorem proving that has nothing to do with the *efficiency* of such procedures. To explain the notion of incompleteness, we first need to broaden our horizons a bit.

Talking about addition of integers gets pretty boring pretty fast. Once multiplication enters the picture the propositions become much more exciting. Multiplication allows us to talk about things like prime numbers and formulate very difficult problems like Goldbach's conjecture.

[4]The method I described is not quite correct because the number of axioms is infinite, so the first round of theorem-proving will take forever. It can be modified to eliminate this "bug". Can you figure out how?

But what is the big deal about multiplication? In second grade you learned that multiplying m and n is the same as adding n to itself m times:

$$m \times n = \underbrace{n + n + \cdots + n}_{m \text{ times}}$$

What about a proposition like "A number is even only if its square is even"? Well, this says

$$\forall m\colon (\exists n\colon \underbrace{m + m + \cdots + m}_{m \text{ times}} = n + n) \longrightarrow (\exists k\colon m = k + k)$$

The $\cdots$ look a bit fishy, and indeed they are. It turns out that it is impossible to *define* multiplication using only the symbols 0, 1, and + and the notation of quantifier logic.

In order to prove theorems about numbers that involve addition *and* multiplication, we need more axioms. One collection of axioms that was proposed after some careful thought are the axioms of Peano arithmetic. You can rewrite proofs of propositions like "If n^2 is even then n is even" into Peano arithmetic without terrible effort.

In 1931 Kurt Gödel produced a proposition about numbers (with 0, 1, +, and $\times$) that is true, but cannot be proved from the axioms of Peano arithmetic. Knowing what we know, it seems reasonable to conclude that the problem should lie with the axioms, as they probably do not describe numbers sufficiently accurately, but maybe also some other unintended structure like $\mathbb{Z}_2$. This is not the case. Gödel actually proved something much more surprising that has nothing to do with the specific content of the Peano axioms:

Gödel's Incompleteness Theorem. *For every collection of reasonable*[5] *axioms A about numbers (with 0, 1, +, and $\times$) that are true there exists a proposition P about numbers that is true, but is not provable from A.*

So an automatic search for a proof of, say, Goldbach's conjecture may well be doomed from the start: We can never be sure that the Peano axioms, or any other "self-evident" set of axioms we start with, is sufficient to prove it. (Most working mathematicians believe that, in this particular case, the Peano axioms should be sufficient.)

Gödel's incompleteness theorem is one of the most surprising theorems in all of mathematics. What is even more surprising is that even though this theorem talks about propositions concerning integers, it is fundamentally related to computer programs. Let me explain how. Take a good look at the following java program:

```
public class X {public static void main(String[] args){int[][] t = new int[][]{{2
02,1026,1100,396,324,1080,192,609,555,888,72,432},{3,9,8,5},{2,2,5,9},{4,6,1,9,2,
11},{4,6,1,9,3,2,11,7,0,5,10},{2,1,5,9},{1,9,2,5},{0,2,10,5,1,6,3,11,8,4},{10,4,2
,6},{1,10,2,3,5,9,7,4,11,6},{7,0,3,6},{2,9,10,1},{7,1,10,6},{12,0,-0}};do{while(t
[13][1]+1<t[t[13][0]].length){t[13][2]=t[0][t[t[13][0]][t[13][1]]];t[0][t[t[13][0
]][t[13][1]]]=t[0][t[t[13][0]][++t[13][1]]];t[0][t[t[13][0]][t[13][1]++]]=t[13][2
];}}while(!(--t[13][0]<=(int)Math.sin(Math.PI))&&((t[13][1]=0)<1));while(t[4][2]<
=t[9][5]+3)System.out.print((char)(t[0][t[4][2]-1]/t[4][2]++));}}
```

What does program X do? It is very difficult to tell just by looking at the code. You could try to type it up in your machine and run it. You run it for 1 hour, 2 hours, 10 days, it does nothing... will it eventually output something and terminate or is it stuck in an infinite loop? So maybe you write some computer code that tries do some automated program analysis and determine if program X will eventually terminate. In the most famous paper in computer science ever written, Alan Turing showed that your analysis tool will, in general, not be of much help:

[5] "Reasonable" is a technical term that prevents cheating by say, taking *all* propositions that are true about numbers as axioms. It means there exists a computer program that prints a possibly infinite list of all the axioms.

Turing's Theorem. *There does not exist a computer program T such that (1) T terminates on every input and (2) when given the code of a computer program X as input, T outputs "yes" if X eventually terminates and "no" otherwise.*

Gödel's Incompleteness Theorem is, in fact, a *special case* of Turing's Theorem. How so? Well, for every computer program X, the proposition "X eventually terminates" is either true or false. So our automated theorem prover should eventually tell us which is the case. But which axioms should we feed to it to get its reasoning started? It looks like we need some axioms that describe the logic of computer programs. What are they?

This is a trick question. We *already saw* the axioms of computer programs. They are the same as the axioms of Peano arithmetic. But how can this be? The Peano axioms are about numbers, not about programs. It turns out that a computer program is a number in disguise. After all, a computer's memory is nothing more than a long string of bits, and what a computer program does is merely some fancy copy-paste operations on such strings. So any proposition about computer programs is essentially a proposition about strings with operations like bit lookup, copying, and pasting. In Question 6 of Homework 1 you saw how propositions about strings can be formulated as propositions about numbers.[6] In fact, *every proposition about computer programs can be expressed as a proposition about numbers.*

Now consider the following implementation of Turing's program T: On input X, translate the statement "X eventually terminates" into a proposition P_X about numbers and run the automatic theorem prover on this proposition, starting with your favorite axioms about numbers. If the prover finds a proof of P_X output "yes". If it finds a proof of NOT P_X output "no".

There are two possibilities. If a proof of P_X or a proof of NOT P_X exists for every program X, then the automatic theorem prover will eventually find this proof and T will correctly determine if X terminates or not. But this is not allowed by Turing's theorem. Therefore the second possibility must hold: There must exist a program X for which neither of the propositions P_X and NOT P_X has a proof. One of these two must be true but not provable from the axioms, confirming Gödel's Incompleteness Theorem.

References

This lecture is based on Chapter 1 of the text *Mathematics for Computer Science* by E. Lehman, T. Leighton, and A. Meyer. Material from slides by Prof. Lap Chi Lau were also used in the preparation. Section 6 is partially based on the book *A Mathematical Introduction to Logic* (2nd edition) by Herbert Enderton. The code snippet is from `https://gist.github.com/jorgeatorres/442094`.

[6] In that exercise, besides 0, 1, + and × we also used the exponentiation symbol E, but it turns out the last one is not really needed.