Lecture 13
The First Incompleteness Theorem

Michael Beeson

The First Incompleteness Theorem

There is a true sentence of $\mathbf{PA}$ that is not provable in $\mathbf{PA}$.

Let us see why this is called "incompleteness":

- A theory T is complete if for every sentence ϕ, either $T \vdash \phi$ or $T \vdash \neg\phi$.
- If ϕ is true, i.e. satisfied in the standard model, then it is not the case that $\mathbf{PA} \vdash \neg\phi$, since all theorems of $\mathbf{PA}$ hold in the standard model. (That's what it means to be a model.)
- So, if a true sentence ϕ is unprovable, then $\mathbf{PA}$ is incomplete.
- Conversely, if $\mathbf{PA}$ is incomplete, then there is some sentence ψ such that neither ψ nor $\neg\psi$ is provable. In that case, one of them is a true unprovable sentence.

Turing's Proof

(I am not claiming historical accuracy in the name of this proof.)

- Let φ_e be the function computed by Turing machine e.
- We will show that if $\mathbf{PA}$ is complete, we can solve the halting problem.
- The idea is that if $\mathbf{PA}$ is complete, then if $\varphi_e(e)$ doesn't halt, there is a proof that it doesn't halt, so we can find out that it doesn't halt by searching for a proof.
- So to find out if $\varphi_e(e)$ halts or not, we search both for a computation by machine e at input e, and also for a proof that $\varphi_e(e)$ doesn't halt.
- If $\mathbf{PA}$ is complete, we must find one or the other, thus solving the halting problem.
- On the next slides we write this up properly.

Two notational matters

- We write $\mathbb{N} \models \psi$ to abbreviate $\langle \mathbb{N}, +, \cdot, ', 0 \rangle \models \psi$.
- People often write $\mathbf{T}$ both for the $\mathbf{T}$-predicate and the formula that represents and defines it.
- But I will try to preserve the distinction, writing $\mathbf{T}$ for the predicate and $\mathbb{T}$ for the formula.

Turing's proof done properly

$$\varphi_e(e) \text{ halts iff } \mathbb{N} \models \exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$$

$$\varphi_e(e) \text{ does not halt iff } \mathbb{N} \models \neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$$

Suppose, for proof by contradiction, that $\mathbf{PA}$ is complete. Then

$$\varphi_e(e) \text{ does not halt iff } \vdash \neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$$

Now we solve the halting problem as follows:

$$g(e,n) \quad = \quad \begin{cases} 1 & \text{if } \mathbf{T}(e,e,n) \\ 0 & \text{if } \mathrm{Prf}(n, \ulcorner \neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k) \urcorner) \text{ and not } \mathbf{T}(e,e,n) \\ 2 & \text{otherwise} \end{cases}$$

Since Prf is primitive recursive, g is primitive recursive. Now define

$$f(e) = g(e, \mu k\, (g(e,k) < 2))$$

Then $f(e) = 1$ if $\varphi_e(e)$ halts, and 0 if it doesn't halt, but f is μ-recursive and hence Turing computable. Contradiction, QED.

More details; every step shown

We claim that f solves the halting problem. Suppose $\varphi_e(e)$ is defined. Then $\exists k\, \mathbf{T}(e, e, k)$. Fix such a k; then $\mathbf{T}(e, e, k)$, so $g(e, k) = 0$. We claim $f(e) = 1$. If not, then it is because there is a proof of $\neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$.

Since $\mathbf{PA}$ proves only true theorems, $\neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$ is true; hence $\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$ is false; hence $\varphi_e(e)$ is not defined, contradicting our assumption that $\varphi_e(e)$ is defined. This proves that $f(e) = 1$ when $\varphi_e(e)$ is defined.

Details continued

Now suppose that $\varphi_e(e)$ is not defined. Then for all k, it is not the case that $\mathbf{T}(e, e, k)$, so for all k, $g(e, k) \neq 1$. Hence $g(e, n) = 0$ if and only n is a proof of $\neg\exists k\, \mathbb{T}(\bar{e}, \bar{e}, k)$. By the assumption that $\varphi_e(e)$ is not defined, this is a true sentence. By the assumption that $\mathbf{PA}$ proves every true sentence, it has a proof in $\mathbf{PA}$. Let n be the Gödel number of such a proof. Then $g(e, n) = 0$; and since for all k, $g(e, k) \neq 1$, we have $f(e) = 0$.

Therefore $f(e) = 1$ if $\varphi_e(e)$ halts and $f(e) = 0$ if $\varphi_e(e)$ does not halt. Therefore f solves the halting problem. Since f is μ-recursive, it is Turing computable. Hence f cannot solve the halting problem. This contradiction completes Turing's proof of the incompleteness theorem.

Remarks on Turing's proof

- It does not produce a specific true unprovable sentence.
- It does, however, show that for some number e, the formula expressing that $\varphi_e(e)$ does not halt is true but unprovable. It just doesn't exhibit a particular e.
- It relies heavily on the machinery we developed, to know that Prf is primitive recursive, which we need in order to see that g on the previous slide is primitive recursive, and to know that primitive recursive functions are Turing computable.
- Turing's work came five years after Gödel's, so Gödel's proof did not involve Turing machines and the halting problem.

The self-reference lemma and Gödel's proof

- The theorem we are about to state and prove is traditionally known as the "self-reference lemma", though it certainly deserves to be called a theorem.
- It encapsulates Gödel's method of self-reference.
- Gödel certainly knew this theorem, because he gave several applications of it, but I do not think he ever stated it.
- Even more surprisingly, it does not occur in Kleene's textbook.
- Even more surprisingly, it does not occur in Shoenfield's famous 1967 textbook.
- I have it handwritten inside the back cover of my copy of Shoenfield. I learned it orally from my teachers.

Notation

When ψ is a formula with a free variable z, and ϕ is another formula, the result of substituting the numeral for $\ulcorner\phi\urcorner$ into ψ would be written

$$\psi[x := \overline{\ulcorner\phi\urcorner}]$$

Because this is typographically complicated, it is often shortened to $\psi[z := \ulcorner\phi\urcorner]$, or even $\psi(\ulcorner\phi\urcorner)$.

This is actually unambiguous abbreviation, since the only syntactically correct interpretation of the notation $\psi(\ulcorner\phi\urcorner)$ is the former (more complicated) expression, since $\ulcorner\phi\urcorner$ is a number, not a term, and you need a term (in this case a numeral) to substitute.

However, in the following proof we shall use the completely precise notation, for the benefit of readers encountering this material for the first time.

The Self-reference Lemma (statement, not yet proof)

Let $\psi(z, \mathbf{x})$ be a formula of $\mathbf{PA}$ with free variables z and $\mathbf{x}$. Then there exists a formula ϕ with free variables $\mathbf{x}$ such that

$$\mathbf{PA} \vdash \phi \equiv \psi[z := \overline{\ulcorner \phi \urcorner}]$$

Somewhat less precisely,

$$\mathbf{PA} \vdash \phi \equiv \psi(\ulcorner \phi \urcorner)$$

Remark. ϕ says, "I have the property ψ."

Representation of $Subst$

We could add function symbols for $Subst$ and Num (conservatively over $\mathbf{PA}$). I choose not to do so; but then we have to represent $Subst$ and Num by formulas to discuss them in $\mathbf{PA}$.

Let $\mathbb{S}(a, b, n, z)$ be a formula representing $Subst$. Then

$$\vdash \mathbb{S}(\overline{\ulcorner t \urcorner}, \overline{\ulcorner x \urcorner}, \overline{\ulcorner A \urcorner}, \bar{z}) \quad \text{if} \quad z = \ulcorner A[x := t] \urcorner$$

$$\vdash \neg\, \mathbb{S}(\overline{\ulcorner t \urcorner}, \overline{\ulcorner x \urcorner}, \overline{\ulcorner A \urcorner}, \bar{z}) \quad \text{if} \quad z \neq \ulcorner A[x := t] \urcorner$$

$$\vdash \exists! z\, (\mathbb{S}(\overline{\ulcorner t \urcorner}, \overline{\ulcorner x \urcorner}, \overline{\ulcorner A \urcorner}, z)$$

and similarly for Num and its representing formula $\mathbb{N}um$.

Enumerating the formulas with one free variable

We define A_n as follows:

- If n is the Gödel number of a formula with (at least) the (specific) variable x free, then A_n is that formula.
- Otherwise A_n is the formula $x = x$.

Hence for every n, A_n is a formula with exactly the free variable x.

Yet more fun with Gödel numbers

- The Gödel number of $A_n[x := \bar{n}]$, or for short $A_n(\bar{n})$, is given by
$$\ulcorner A_n[x := \bar{n}] \urcorner = Subst(Num(n), \ulcorner x \urcorner, n)$$
- That function is represented by the formula
$$\exists w \mathbb{S}(w, \overline{\ulcorner x \urcorner}, n, z) \wedge \mathbb{N}um(n, w))$$
and it would be possible to bound the quantifier on w if required.
- Let ψ be a fixed formula with one free variable x.
- We claim that there is a formula with one free variable x, which can be informally described as
$$\psi(\ulcorner A_x(\bar{x}) \urcorner)$$
- More precisely that has to be
$$\exists z, w \, (\mathbb{S}(w, \overline{\ulcorner x \urcorner}, x, z) \wedge \mathbb{N}um(x, w) \wedge \psi[x := z])$$

$\psi(\ulcorner A_x(\bar{x}) \urcorner)$ continued

We said that has to be, precisely,

$$\chi(x) := \exists z, w\, (\mathbb{S}(w, \overline{\ulcorner x \urcorner}, x, z) \wedge \mathbb{N}um(x, w) \wedge \psi[x := z])$$

Note that x is the only free variable, if ψ has exactly one free variable. But if ψ has more free variables then these variables occur also in χ (and z should be chosen different from all free variables of ψ).

Proof of the self-reference lemma

Recall $\chi(x) := \exists z, w\,(\mathbb{S}(w, \overline{\ulcorner x \urcorner}, n, z) \wedge \mathbb{N}um(x, w) \wedge \psi[x := z])$, which we abbreviate as $\psi(\overline{\ulcorner A_x(\bar{x}) \urcorner})$. Define

$$\phi := \chi[x : \overline{\ulcorner \chi \urcorner}]$$

or informally, ϕ is $\chi(\overline{\ulcorner \chi \urcorner})$.

We claim $\vdash \phi \equiv \psi[x := \overline{\ulcorner \phi \urcorner}]$. Let $n = \ulcorner \chi \urcorner$. We have (provably)

$$\begin{array}{rcll} \phi & \equiv & \chi(\overline{\ulcorner \chi \urcorner}) \equiv \chi(\bar{n}) & \\ & \equiv & \psi(\overline{\ulcorner A_x(\bar{x}) \urcorner})[x := \bar{n}] & \\ & \equiv & \psi(\overline{\chi(\overline{\ulcorner \chi \urcorner})}) & \text{because } A_n \text{ is } \chi \\ & \equiv & \psi(\overline{\ulcorner \phi \urcorner}) & \end{array}$$

Self-reference and fixed points

We note the similarity between the proof of the self-reference lemma and the fixed-point theorem of λ-calculus. There we took $\omega = F(\lambda x\,(xx))$ and then showed $\omega\omega = F(\omega\omega)$.

Here we took $\chi = \psi(A_x(x))$ and showed $\chi(\chi) \equiv \psi(\chi(\chi))$, ignoring Gödel numbers and numerals to show the parallel structure of the proofs.

If there was imitation, chronology tells us Church imitated Gödel, not the other way around. But remember, Gödel didn't state the self-reference lemma explicitly.

Gödel's proof of the first incompleteness theorem

Take $\psi(n) := \neg\exists x \, \mathrm{Prf}(x, n)$. By the self-reference lemma choose

$$\phi \equiv \psi[n := \overline{\ulcorner \phi \urcorner}]$$

Then ϕ says "I am not provable."

Suppose, for proof by contradiction, that ϕ is false. Then it is provable, and hence true, since all theorems of $\mathbf{PA}$ are true. Contradiction. Hence ϕ is not false. Hence it is true. Hence it is not provable. QED.

Remark on Gödel's presentation

Gödel's original paper does not separate the self-reference lemma out as a theorem in its own right, but gives the diagonal argument in the specific case when ψ in the self-reference lemma is the formula $\neg\exists k\,\mathrm{Prf}(z,k)$.

Other applications of self-reference

Church proved that the truth set of arithmetic is not recursive. Now we can prove it is not arithmetical, i.e. not definable by any formula of $\mathbf{PA}$. Of course by countability, there are many undefinable sets of integers, but this is the first one to be explicitly exhibited (both historically, and in these lectures).

First let us precisely define the truth predicate: $True(n)$ is true if and only if n is the Gödel number of a sentence ϕ of $\mathbf{PA}$ such that $\mathbb{N} \models \phi$. Recall this is short for $\langle N, +, \cdot, ', 0\rangle \models \phi$.

The definition can be made more explicit this way. For each fixed n, we can define the truth predicate $True_n$ for formulas of no more than n symbols by clauses like this one: $True_{n+1}(\ulcorner \forall x A(x) \urcorner)$ if and only if for all m, $True_n(\ulcorner A[x := \bar{m}] \urcorner)$. Thus we need about n quantifiers to define truth for formulas of complexity up to n. This makes it not too surprising that truth for all formulas of $\mathbf{PA}$ is not arithmetical.

Tarski 1948: Undefinability of truth

The predicate $True(n)$ is not arithmetical.

Proof. Suppose that $True$ were definable by a formula $\psi(x)$ of $\mathbf{PA}$. By the self-reference lemma, we could then choose ϕ to say "I am false". More precisely,

$$\mathbf{PA} \vdash \phi \equiv \neg\psi(\overline{\ulcorner\phi\urcorner})$$

If ϕ is true, then $\psi(\overline{\ulcorner\phi\urcorner})$ is false, so ϕ is false, since ψ supposedly defines $True$. Hence ϕ is false. But then similarly, ϕ is true, contradiction. That completes the proof.

Thus the "liar's paradox" becomes not a paradox, but a theorem on the undefinability of truth. Tarski generalized this to show that no language stronger than some minimal strength can define its own truth.