

## The Diagonalization Lemma, Rosser and Tarski

Peter Smith

January 15, 2010

- The Diagonalization Lemma
- Incompleteness from the Diagonalization Lemma
- Rosser’s Theorem
- Tarski’s Theorem
- The Master Argument

We’ve now proved our key version of the First Theorem, Theorem 42. If  $T$  is the right kind of  $\omega$ -consistent theory including enough arithmetic, then there will be an arithmetic sentence  $G_T$  such that  $T \not\vdash G_T$  and  $T \not\vdash \neg G_T$ . Moreover,  $G_T$  is constructed so that it is true if and only if unprovable-in  $T$  (so it is true).

Now recall that, for a p.r. axiomatized theory  $T$ ,  $Prf_T(m, n)$  is the relation which holds just if  $m$  is the super g.n. of a sequence of wffs that is a  $T$  proof of a sentence with g.n.  $n$ . This relation is p.r. decidable (see §25.4). Assuming  $T$  extends  $Q$ , it can capture any p.r. decidable relation, including  $Prf_T$  (§22). So we can define

**Defn. 50.**  $Prf_T(x, y)$  stands in for a  $T$ -wff that canonically captures  $Prf_T$ .

NB, for some of what follows, *any* wff  $Prf_T$  that captures  $Prf_T$  will do (it doesn’t have to be ‘canonical’ in the sense of Defn. 34). But it’s convenient to fix on some canonical way, hence  $\Sigma_1$  way, of capturing  $Prf_T$ . Next,

**Defn. 51.** Put  $Prov_T(y) =_{\text{def}} \exists v Prf_T(v, y)$ : we’ll call such an expression a provability predicate for  $T$ .

$Prov_T(\bar{n})$  is true, of course, on the standard arithmetic interpretation of  $T$  just if  $n$  numbers a  $T$ -theorem, i.e. a wff for which some number numbers a proof of it. Which means that  $Prov_T(\ulcorner \varphi \urcorner)$  is true just when  $\varphi$  is a theorem. Hence the aptness of the label ‘provability predicate’ for  $Prov_T$ .<sup>1</sup> Note,  $Prov_T$  is also  $\Sigma_1$ .

So our observation  $G_T$  is true if and only if unprovable-in  $T$  can in fact be *expressed* inside  $T$  itself, by the wff  $G_T \leftrightarrow \neg Prov_T(\ulcorner G_T \urcorner)$ . Moreover,  $T$  can quite easily *prove* this, i.e.  $T \vdash G_T \leftrightarrow \neg Prov_T(\ulcorner G_T \urcorner)$  (see §34.2).

Now, when we look at how  $T$  proves that last result, we notice an interesting generalization. Take *any* open sentence  $\varphi(x)$ . Then, we can *always* find a sentence  $\delta$  such that  $T \vdash \delta \leftrightarrow \varphi(\ulcorner \delta \urcorner)$ . This generalization is called The Diagonalization Lemma (and was first explicitly isolated by Carnap as a principle that could be seen as underlying Gödel’s incompleteness proof).

What we are going to do in this episode is first prove the Lemma *directly*, i.e. without proving  $T \vdash G_T \leftrightarrow \neg Prov_T(\ulcorner G_T \urcorner)$  first. The direct proof is stunningly easy. In the next following section, we see how to get from the Lemma back to the incompleteness theorem. Then we note how to use the Lemma to improve Gödel’s original theorem and drop the  $\omega$ -consistency requirement (giving us Rosser’s Theorem). In the final formal section of this episode we see how the same Lemma also yield Tarski’s theorem about the undefinability of arithmetic truth in arithmetic.

We finish with a brief reflection on the fundamental roots of the incompleteness phenomenon.

<sup>1</sup>Recall: We assume we’ve fixed on some acceptable scheme for coding up wffs of a given theory  $T$ ’s language by using Gödel numbers. If  $\varphi$  is an expression, then we’ll denote its Gödel number in our logician’s English by ‘ $\ulcorner \varphi \urcorner$ ’. We use ‘ $\overline{\varphi}$ ’ as an abbreviation inside  $T$ ’s language for the standard numeral for ‘ $\ulcorner \varphi \urcorner$ ’. (§26)

### 33 The Diagonalization Lemma

First a reminder of something familiar:

**Defn. 15.** *The theory  $T$  captures the one-place function  $f$  by the open wff  $\varphi(x, y)$  iff, for any  $m, n$ ,*

- i. if  $f(m) = n$ , then  $T \vdash \varphi(\bar{m}, \bar{n})$ ,*
- ii. if  $f(m) \neq n$ , then  $T \vdash \neg\varphi(\bar{m}, \bar{n})$ .*

Now let's introduce what is obviously a *very* closely related idea:

**Defn. 52.** *The theory  $T$  captures\* the one-place function  $f$  by the open wff  $\varphi^*(x, y)$  iff, for any  $m, n$ , if  $f(m) = n$ , then  $T \vdash \forall y(\varphi^*(\bar{m}, y) \leftrightarrow y = \bar{n})$ .*

It's trivial that a wff that captures\*  $f$  will also capture  $f$  (why?). And there's nearly a converse:

**Theorem 44.** *If  $T$  extends  $\mathbf{Q}$ , then if  $f$  is captured by some wff  $\varphi$ , then there's also a wff  $\varphi^*$  which captures\*  $f$ .*

There's a little trick for defining a capturing\*  $\varphi^*$  from a capturing  $\varphi$ , a trick which is explained in §12.2 of the book. But I'm not going to explain that here – it would be seriously boring to delay over the details. (And when I do a second edition of the book, I'm tempted to define capturing\* from the off, and then we wouldn't need the boring details!) So take the mini-theorem on trust, and work out how to prove it from the book if you must!

Now some more reminders:

**Defn. 42.** *The diagonalization of  $\varphi$  is  $\exists y(y = \overline{\varphi} \wedge \varphi)$ .*

**Theorem 31.** *There is a p.r. function  $diag(n)$  which, when applied to a number  $n$  which is the g.n. of some wff, yields the g.n. of that wff's diagonalization.*

And following on from those, here's a new definition: if  $T$  is a theory that can contains  $\mathbf{Q}$ , it can capture (and hence capture\*) all p.r. functions, then it in particular can capture\* the function  $diag$ . So we put

**Defn. 53.**  $Diag_T(x, y)$  *is a  $T$ -wff which captures\*  $diag$ .*

And now we can state and prove

**Theorem 45 (Diagonalization Lemma).** *If  $T$  extends  $\mathbf{Q}$ , and  $\varphi$  is a one-place open sentence of  $T$ 's language, then there is sentence  $\delta$  such that  $T \vdash \delta \leftrightarrow \varphi(\overline{\delta})$ .*

To avoid unsightly rashes of subscripts, let's henceforth drop subscript  $T$ 's. Then we can argue like this:

*Proof.* Put  $\alpha =_{\text{def}} \forall z(Diag(y, z) \rightarrow \varphi(z))$ , and let  $\delta$  be the diagonalization of  $\alpha$ . Since diagonalizing  $\alpha$  yields  $\delta$ , we have  $diag(\overline{\alpha}) = \overline{\delta}$ . Hence  $T \vdash \forall z(Diag(\overline{\alpha}, z) \leftrightarrow z = \overline{\delta})$  since by hypothesis  $Diag$  captures\*  $diag$  in  $T$ . But just from the definition of  $\delta$ ,  $T \vdash \delta \leftrightarrow \forall z(Diag(\overline{\alpha}, z) \rightarrow \varphi(z))$ . Hence, substituting the provable equivalents, we have  $T \vdash \delta \leftrightarrow \forall z(z = \overline{\delta} \rightarrow \varphi(z))$ , which trivially gives  $T \vdash \delta \leftrightarrow \varphi(\overline{\delta})$ .  $\square$

I promised that it was going to be easy!

Finally, a bit of jargon before proceeding. By a certain abuse of mathematical terminology, we say

**Defn. 54.** *If  $\delta$  is such that  $T \vdash \delta \leftrightarrow \varphi(\overline{\delta})$ , then it is said to be a fixed point for  $\varphi$ .*

So the Diagonalization Lemma is often called the Fixed Point Theorem – every one-place open sentence has a fixed point.

## 34 Incompleteness from the Diagonalization Lemma

### 34.1 Recovering the First Theorem

First we have following general observation about provability predicates (as a reality check, ask yourself where subscript  $T$ 's really belong in this statement and its proof):

**Theorem 46.** *Suppose  $T$  is p.r. axiomatized, contains  $\mathbf{Q}$ , and some sentence or other  $\gamma$  is a fixed point for  $\neg\text{Prov}$  – i.e.,  $T \vdash \gamma \leftrightarrow \neg\text{Prov}(\ulcorner\gamma\urcorner)$ . Then (i) if  $T$  is consistent,  $T \not\vdash \gamma$ . And (ii) if  $T$  is  $\omega$ -consistent,  $T \not\vdash \neg\gamma$ .*

*Proof.* (i) Suppose  $T \vdash \gamma$ . Then  $T \vdash \neg\text{Prov}(\ulcorner\gamma\urcorner)$ . But if there is a proof of  $\gamma$ , then for some  $m$ ,  $\text{Prf}(m, \ulcorner\gamma\urcorner)$ , so  $T \vdash \text{Prf}(\bar{m}, \ulcorner\gamma\urcorner)$ , since  $T$  captures  $\text{Prf}$  by  $\text{Prf}$ . Hence  $T \vdash \exists x \text{Prf}(x, \ulcorner\gamma\urcorner)$ , i.e. we also have  $T \vdash \text{Prov}(\ulcorner\gamma\urcorner)$ , making  $T$  inconsistent. So if  $T$  is consistent,  $T \not\vdash \gamma$ .

(ii) Suppose  $T \vdash \neg\gamma$ . Then  $T \vdash \text{Prov}(\ulcorner\gamma\urcorner)$ , i.e.  $T \vdash \exists x \text{Prf}(x, \ulcorner\gamma\urcorner)$ . But given  $T$  is consistent, there is no proof of  $\gamma$ , i.e. for every  $m$ , not- $\text{Prf}(m, \ulcorner\gamma\urcorner)$ , whence for every  $m$ ,  $T \vdash \neg\text{Prf}(\bar{m}, \ulcorner\gamma\urcorner)$ . So we have a  $\varphi$  such that  $T$  proves  $\exists x \varphi(x)$  while it refutes each instance  $\varphi(\bar{m})$ , which makes  $T$   $\omega$ -inconsistent. So if  $T$  is  $\omega$ -consistent,  $T \not\vdash \neg\gamma$ .  $\square$

But the Diagonalization Lemma asserts the existence a sentence  $\gamma$  such that  $T \vdash \gamma \leftrightarrow \neg\text{Prov}(\ulcorner\gamma\urcorner)$ . Moreover, since  $\text{Prov}$  is  $\Sigma_1$ ,  $\neg\text{Prov}$  is  $\Pi_1$ , and the diagonalization construction produces a  $\Pi_1$  sentence. So we immediately recover Theorem 42.

### 34.2 Relating old and new

Briefly: how does the specific Gödel sentence  $G$  as we *originally* constructed it via the definitions in §28 stand to the generic Gödel sentences  $\gamma$ s we've just been talking about?

Well, how does our proof of the Diagonalization Lemma tell us to construct a  $\gamma$  such that  $T \vdash \gamma \leftrightarrow \neg\text{Prov}(\ulcorner\gamma\urcorner)$ ? It says: first form a wff  $\alpha = \forall z(\text{Diag}(y, z) \rightarrow \neg\text{Prov}(z))$ , and then diagonalize  $\alpha$  to get  $\gamma$ . So think more about  $\alpha$ . Unpacking a bit,  $\alpha$  is  $\forall z(\text{Diag}(y, z) \rightarrow \neg\exists x \text{Prf}(x, z))$ , which is equivalent to  $\forall x \forall z \neg(\text{Diag}(y, z) \wedge \text{Prf}(x, z))$ , i.e. to  $\forall x \neg \exists z(\text{Diag}(y, z) \wedge \text{Prf}(x, z))$ .

But now note that  $\exists z(\text{Diag}(y, z) \wedge \text{Prf}(x, z))$  captures the  $Gdl$  relation. So this makes  $\alpha$  like  $\forall x \neg Gdl(x, y)$  – and hence  $\gamma$  (the diagonalization of  $\alpha$ ) like  $G$  (the diagonalization of  $\forall x \neg Gdl(x, y)$ ).

Ok, that's motivational. But now let's check that we can give a direct proof of

**Theorem 47.** *If  $T$  is p.r. axiomatized and contains  $\mathbf{Q}$ ,  $T \vdash G_T \leftrightarrow \neg\text{Prov}_T(\ulcorner G_T \urcorner)$ .*

But don't get bogged down in this proof – it's just for the record.

*Proof.* Recall, dropping subscripts,

$$G =_{\text{def}} \exists y(y = \ulcorner U \urcorner \wedge U),$$

where ' $\ulcorner U \urcorner$ ' stands in for the numeral for  $U$ 's g.n.: further recall

$$U =_{\text{def}} \forall x \neg Gdl(x, y)$$

where  $Gdl(x, y)$  captures our old friend, the relation  $Gdl$ , where  $Gdl(m, n)$  holds when  $m$  codes for a proof of the diagonalization of the wff with number  $n$  (a proof in  $T$ , of course!).

Now, by definition then,

$$Gdl(m, n) =_{\text{def}} \text{Prf}(m, \text{diag}(n)).$$

But the one-place p.r. function  $\text{diag}$  is captured\* by an open wff  $\text{Diag}(x, y)$ . We can therefore now retrospectively fix on the following definition:

$$Gdl(x, y) =_{\text{def}} \exists z(\text{Prf}(x, z) \wedge \text{Diag}(y, z)).$$

And now let's do some elementary manipulations:

$$\begin{aligned}
G &\leftrightarrow \forall x \neg \text{Gdl}(x, \overline{\ulcorner U \urcorner}) \\
&\leftrightarrow \forall x \neg \exists z (\text{Prf}(x, z) \wedge \text{Diag}(\overline{\ulcorner U \urcorner}, z)) && \text{(definition of Gdl)} \\
&\leftrightarrow \forall x \forall z \neg (\text{Prf}(x, z) \wedge \text{Diag}(\overline{\ulcorner U \urcorner}, z)) && \text{(pushing in the negation)} \\
&\leftrightarrow \forall z \forall x \neg (\text{Prf}(x, z) \wedge \text{Diag}(\overline{\ulcorner U \urcorner}, z)) && \text{(swapping quantifiers)} \\
&\leftrightarrow \forall z (\text{Diag}(\overline{\ulcorner U \urcorner}, z) \rightarrow \neg \exists x \text{Prf}(x, z)) && \text{(rearranging after ‘}\forall z\text{’)} \\
&\leftrightarrow \forall z (\text{Diag}(\overline{\ulcorner U \urcorner}, z) \rightarrow \neg \exists v \text{Prf}(v, z)) && \text{(changing variables)} \\
&=_{\text{def}} \forall z (\text{Diag}(\overline{\ulcorner U \urcorner}, z) \rightarrow \neg \text{Prov}(z)) && \text{(definition of Prov)}
\end{aligned}$$

Since this is proved by simple logical manipulations, that means we can prove the equivalence inside the formal first-order logic built into  $Q$  and hence in  $T$ . So

$$T \vdash G \leftrightarrow \forall z (\text{Diag}(\ulcorner U \urcorner, z) \rightarrow \neg \text{Prov}(z)).$$

Now, diagonalizing  $U$  yields  $G$ . Hence, just by the definition of *diag*, we have  $\text{diag}(\ulcorner U \urcorner) = \ulcorner G \urcorner$ . Since by hypothesis  $\text{Diag}$  captures\* *diag* as a function, it follows by definition that

$$T \vdash \forall z (\text{Diag}(\ulcorner U \urcorner, z) \leftrightarrow z = \ulcorner G \urcorner).$$

Putting those two results together, we immediately get

$$T \vdash G \leftrightarrow \forall z (z = \ulcorner G \urcorner \rightarrow \neg \text{Prov}(z)).$$

But the right-hand side of that biconditional is trivially equivalent to  $\neg \text{Prov}(\ulcorner G \urcorner)$ . So we’ve proved the desired result.  $\square$

## 35 Rosser’s Theorem

### 35.1 Rosser’s basic trick

One half of the First Theorem requires the assumption that we are dealing with a theory  $T$  which is not only consistent but is  $\omega$ -consistent. But we can improve on this in two different ways:

1. We can keep the *same* undecidable sentence  $G_T$  while invoking the weaker assumption of ‘1-consistency’ in showing that  $T \not\vdash \neg G_T$ .
2. Following Barkley Rosser, we can construct a *different* and more complex sentence  $R_T$  such that we only need to assume  $T$  is plain consistent in order to show that  $R_T$  is formally undecidable.

Since Rosser’s clever construction yields the better result, that’s the result we’ll talk about here (I say something about 1-consistency in the book).

So how does Rosser construct an undecidable sentence  $R_T$  for  $T$ ? Well, essentially, where Gödel constructs a sentence  $G_T$  that indirectly says ‘I am unprovable in  $T$ ’, he constructs a ‘Rosser sentence’  $R_T$  which indirectly says ‘if I am provable in  $T$ , then my negation is already provable’ (i.e. it says that if there is a proof of  $R_T$  with super g.n.  $n$ , then there is a proof of  $\neg R_T$  with a smaller code number).

### 35.2 Implementing the trick

Consider the relation  $\overline{\text{Prf}}_T(m, n)$  which holds when  $m$  numbers a  $T$ -proof of the *negation* of the wff with number  $n$ . This relation is obviously p.r. given that  $\text{Prf}_T$  is; so assuming  $T$  is has the usual properties it will be captured by a wff  $\overline{\text{Prf}}_T(x, y)$ . So let’s consider *the Rosser provability predicate* defined as follows:

**Defn. 55.**  $\text{RProv}_T(x) =_{\text{def}} \exists v (\text{Prf}_T(v, x) \wedge (\forall w \leq v) \neg \overline{\text{Prf}}_T(w, x))$ .

Then a sentence is Rosser-provable in  $T$  – its g.n. satisfies the Rosser provability predicate – if it has a proof (in the ordinary sense) and there’s no ‘smaller’ proof of its negation.

Now we apply the Diagonalization Lemma, not to the negation of a regular provability predicate (which is what we just did to get Gödel’s First Theorem again), but to the negation of the Rosser provability predicate. The Lemma then tells us,

**Theorem 48.** *Given that  $T$  is p.r. axiomatized and contains  $Q$ , then there is a sentence  $R_T$  such that  $T \vdash R_T \leftrightarrow \neg \text{RProv}_T(\ulcorner R_T \urcorner)$ .*

We call such a sentence  $R_T$  a Rosser sentence for  $T$ .

Another semantic incompleteness result is immediate:

**Theorem 49.** *If  $T$  is a sound p.r. axiomatized theory including  $Q$  (and because sound therefore consistent),  $T \not\vdash R_T$  and  $T \not\vdash \neg R_T$ , where  $R_T$  is a Rosser sentence*

*Proof.* Assume  $T$ 's soundness, then its theorems are true, and  $R_T$  is true if and only if it is not Rosser-provable. Suppose  $R_T$  were a theorem. Then it would be true since all theorems are true. So it is not Rosser-provable, which means that 'if  $R_T$  is provable,  $\neg R_T$  is already provable' would be true, and also this conditional would have a true antecedent. We can infer that  $\neg R_T$  is provable. Which makes  $T$  inconsistent, contrary to hypothesis. Therefore  $R_T$  is unprovable. Which shows that the material conditional 'if  $R_T$  is provable,  $\neg R_T$  is already provable' has a false antecedent, and hence is true. In other words,  $R_T$  is true. Hence its negation  $\neg R_T$  is false, and is therefore unprovable since only truths are provable in a sound theory.  $\square$

As we said, however, in order to show that neither  $R_T$  nor  $\neg R_T$  is provable we do not need the semantic assumption that  $T$  is sound. The syntactic assumption of  $T$ 's consistency is enough.

### 35.3 Rosser's Theorem

We can now show that

**Theorem 50.** *Let  $T$  be consistent p.r. axiomatized theory including  $Q$  and let  $\rho$  be any fixed point for  $\neg \text{RProv}_T(x)$ . Then  $T \not\vdash \rho$  and  $T \not\vdash \neg \rho$ .*

And since the Diagonalization Lemma tells us that there is a fixed point, it follows that  $T$  has an undecidable sentence. Sadly, however – and there's no getting away from it – the proof of Theorem 50 is messy and very unpretty. Masochists can check out the proof of Theorem 21.2 in the book (on p. 178). We then have to do more work to beef up that proof idea to show that in fact (as with Gödel's original proof) we can find a  $\Pi_1$  sentence which is undecidable so long as  $T$  is consistent (that work is done on p. 179). You do not need to know these proofs! – just that they exist.

## 36 Tarski's Theorem

Next, we visit twin peaks which can also be reached via the Diagonalization Lemma. The path is very straightforward, but it leads to a pair of rather spectacular results that are usually packaged together as *Tarski's Theorem*.

### 36.1 Truth-predicates and truth-definitions

Recall a familiar thought: 'snow is white' is true iff snow *is* white. Likewise for all other sensible replacements for 'snow is white'. In sum, every instance of ' $\varphi$  is true iff  $\varphi$ ' is true. And that's because of the meaning of the informal truth-predicate 'true'.

Now suppose we have fixed on some scheme for Gödel numbering wffs of the interpreted arithmetical language  $L$ . Then we can define a corresponding numerical property *True* as follows:

$\text{True}(n)$  is true iff  $n$  is the g.n. of a true sentence of  $L$ .

Suppose that the open wff  $T(x)$  – which belongs to an arithmetical language  $L'$  which includes  $L$  – expresses this numerical property *True*. Then, for any  $L$ -sentence  $\varphi$ ,

$\varphi$  is true iff  $\text{True}(\ulcorner \varphi \urcorner)$  iff  $T(\ulcorner \varphi \urcorner)$  is true.

Hence, for any  $L$ -sentence  $\varphi$ , every corresponding  $L'$ -sentence

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

is true. Which motivates our first main definition:

**Defn. 56.**

An open  $L'$ -wff  $\mathsf{T}(x)$  is a formal truth-predicate for  $L$  iff for every  $L$ -sentence  $\varphi$ ,  $\mathsf{T}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$  is true.

Here's a companion definition:

**Defn. 57.**

A theory  $\Theta$  (with language  $L'$  which includes  $L$ ) is a truth-theory for  $L$  iff for some  $L'$ -wff  $\mathsf{T}(x)$ ,  $\Theta \vdash \mathsf{T}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$  for every  $L$ -sentence  $\varphi$ .

Equally often, a truth-theory for  $L$  is called a 'definition of truth for  $L$ '.

### 36.2 The undefinability of truth

Suppose  $T$  is a nice arithmetical theory with language  $L$ . An obvious question arises: could  $T$  be competent to 'define' truth for its own language (i.e., can  $T$  include a truth-theory for  $L$ )? And the answer is immediate:

**Theorem 51.** *No consistent p.r. axiomatized theory which includes  $\mathsf{Q}$  can define truth for its own language.*

*Proof.* Assume  $T$  defines truth for  $L$ , i.e. there is an  $L$ -predicate  $\mathsf{T}(x)$  such that  $T \vdash \mathsf{T}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$  for every  $L$ -sentence  $\varphi$ . Since  $T$  has the right properties, the Diagonalization Lemma applies, so applying the Lemma to the negation of  $\mathsf{T}(x)$ , we know that there must be some sentence  $\mathsf{L}$  – a Liar sentence! – such that

1.  $T \vdash \mathsf{L} \leftrightarrow \neg \mathsf{T}(\ulcorner \mathsf{L} \urcorner)$ .

But, by our initial assumption, we also have

2.  $T \vdash \mathsf{T}(\ulcorner \mathsf{L} \urcorner) \leftrightarrow \mathsf{L}$ .

It is immediate that  $T$  is inconsistent, contrary to hypothesis. So our assumption must be wrong:  $T$  can't define truth for its own language.  $\square$

### 36.3 The inexpressibility of truth

That first theorem puts limits on what a nice theory can *prove* about truth. But with very modest extra assumptions, we can put limits on what a theory's language can even *express* about truth.

Consider our old friend  $L_A$  for the moment, and suppose that there is an  $L_A$  truth-predicate  $\mathsf{T}_A$  that expresses the corresponding truth property  $\mathit{True}_A$ . Since  $\mathsf{Q}$  is nice, the Diagonalization Lemma applies, in particular to the negation of  $\mathsf{T}_A(x)$ . So we know that for some  $L_A$  sentence  $\mathsf{L}$ ,

1.  $\mathsf{Q} \vdash \mathsf{L} \leftrightarrow \neg \mathsf{T}_A(\ulcorner \mathsf{L} \urcorner)$ .

But (and here comes the extra assumption we said we were going to invoke!) everything  $\mathsf{Q}$  proves is true, since  $\mathsf{Q}$ 's axioms are of course true and its logic is truth preserving. So

2.  $\mathsf{L} \leftrightarrow \neg \mathsf{T}_A(\ulcorner \mathsf{L} \urcorner)$

will also be a true  $L_A$  wff. But, by the assumption that  $\mathsf{T}_A$  is a truth-predicate for  $L_A$ ,

3.  $\mathsf{T}_A(\ulcorner \mathsf{L} \urcorner) \leftrightarrow \mathsf{L}$

must be true too. (2) and (3) immediately lead to contradiction again. Therefore our supposition that  $\mathsf{T}_A$  is a truth-predicate has to be rejected. Hence no predicate of  $L_A$  can even express the numerical property  $\mathit{True}_A$ .

The argument evidently generalizes. Take any language  $L$  rich enough for us to be able to formulate in  $L$  something equivalent to the very elementary arithmetical theory  $\mathsf{Q}$  (that's so we can prove the Diagonalization Lemma again). Call that an arithmetically adequate language. Then by the same argument, assuming  $\mathsf{Q}$  is a correct theory,

**Theorem 52.** *No predicate of an arithmetically adequate language  $L$  can express the numerical property  $\text{True}_L$  (i.e. the property of numbering a truth of  $L$ ).*

This tells us that while you can express *syntactic* properties of a sufficiently rich formal theory of arithmetic (like provability) inside the theory itself via Gödel numbering, you can't express some key *semantic* properties (like arithmetical truth) inside the theory.

### 36.4 A moral

Suppose  $T$  is a nice theory. Then (1) there are some numerical properties that  $T$  can capture (the p.r. ones for a start); (2) there are some properties that  $T$  can express but not capture (for example the property of Gödel-numbering a  $T$ -theorem – see the book, §21.4); and (3) there are some properties that  $T$ 's language  $L$  cannot even express (for example  $\text{True}_L$ , the numerical property of numbering-a-true- $L$ -wff).

It is not, we should hasten to add, that the property  $\text{True}_L$  is mysteriously ineffable, and escapes all formal treatment. A richer theory  $T'$  with a richer language  $L'$  may perfectly well be able to capture  $\text{True}_L$ . But the point remains that, however rich a given theory of arithmetic is, there will be limitations, not only on what numerical properties it can capture but even on which numerical properties that particular theory's language can express.

## 37 The Master Argument

Our results about the non-expressibility of truth of course point to a particularly illuminating take on the argument for incompleteness.

For example: truth in  $L_A$  isn't provability in PA, because while PA-provability *is* expressible in  $L_A$ , truth-in- $L_A$  *isn't*. So assuming that PA is sound and everything provable in it is true, this means that there must be truths of  $L_A$  which it can't prove. Similarly, of course, for other nice theories.

And in a way, we might well take this to be *the* Master Argument for incompleteness, revealing the roots of the phenomenon. Gödel himself wrote (in response to a query)

I think the theorem of mine that von Neumann refers to is . . . that a complete epistemological description of a language A cannot be given in the same language A, because the concept of truth of sentences in A cannot be defined in A. *It is this theorem which is the true reason for the existence of undecidable propositions in the formal systems containing arithmetic.* I did not, however, formulate it explicitly in my paper of 1931 but only in my Princeton lectures of 1934. The same theorem was proved by Tarski in his paper on the concept of truth.

In sum, as we emphasized before, arithmetical truth and provability in this or that formal system must peel apart.

And that's enough – at least in these notes – about the First Incompleteness Theorem. There's quite a bit more in the book, in Chs 19–23, which I'm not going to be covering in lectures. Enthusiasts will want to devour the lot! – but let me especially highlight the sections which amplify this episode, and then the sections you ought to know about for general logical purposes anyway:

1. More on the topics in the episode: §§19.1–19.3, Ch. 20, §21.1–21.6: browse through Ch. 23.
2. For second-order arithmetics: §§22.1–22.6.