13 The Diagonalization Lemma

In this chapter, we establish the so-called Diagonalization Lemma, and use it to re-prove the syntactic version of the First Theorem. In Chapter 15 we use the Lemma again to prove Rosser's Theorem (a technical improvement on the syntactic First Theorem which allows us to drop the assumption of ω-consistency); and in Chapter 16 the Lemma is put to work to prove Tarski's Theorem (a much deeper result, about the 'undefinability of truth').

The Lemma is important, then; and – as you will see – its derivation is not difficult to follow once you grasp the key construction. However, it is perhaps worth pausing at the end of the chapter to describe a nice idea due to Kripke which gets us a tweaked version of the Diagonalization Lemma more or less for free, and we show how this version can also be used to prove incompleteness.

A general observation. To make a smooth fit with our earlier discussions, we will continue to focus on p.r. axiomatized theories. However, the Interlude has given us an inkling of why the same results will carry over to effectively axiomatized theories more generally. But we won't need to make use of this point, except in Chapter 14.

13.1 Two quick reminders

Before getting down to business, it's useful to recall the definition of capturing a function:

Defn. 37. *The theory T captures the one-place function f by the open wff $\psi(\mathsf{x}, \mathsf{y})$ iff, for any m, n,*

i. if $f(m) = n$, then $T \vdash \psi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$,

ii. if $f(m) \neq n$, then $T \vdash \neg\psi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.

iii. $T \vdash \exists!\mathsf{y}\psi(\overline{\mathsf{m}}, \mathsf{y})$.

And this time, let's note explicitly that (i) and (iii) together imply

iv. if $f(m) = n$, then $T \vdash \forall\mathsf{x}(\psi(\overline{\mathsf{m}}, \mathsf{x}) \leftrightarrow \mathsf{x} = \overline{\mathsf{n}})$.

Next, let's restate Theorem 32 about the *diag* function, and now apply Theorems 27 and 28 to add an important clause about expressing and capturing it (we assume as usual that we have a normal Gödel-numbering scheme in place for the relevant theory T):

Theorem 46. *If T is a p.r. axiomatized theory which contains Q, there is a p.r. function $diag_T(n)$ which, when applied to a number n which is the g.n. of some T-wff with one free variable, yields the g.n. of that wff's diagonalization, and yields 0 otherwise. And, as with any p.r. function, T can express and capture this function by a Σ_1 wff $\mathsf{Diag}_T(\mathsf{x}, \mathsf{y})$.*

13.2 The Diagonalization Lemma

We saw that the Gödel sentence G_T is true if and only if it is unprovable in T. We constructed this wff using a diagonalization trick. We now note that the basic trick can be applied more generally. Roughly: take any arithmetizable property F; then we can use diagonalization to construct a T-sentence which is true if and only it has property F (and T can prove that equivalence).

More carefully, we have the two-part *Diagonalization Lemma* (Rudolf Carnap noted a version of (i): he is often, but I think wrongly, attributed a version of (ii) as well). As usual, assume a normal Gödel-numbering scheme is in play: then,

Theorem 47. *If T is a p.r. axiomatized theory which contains Q, and φ is a one-place open sentence of T's language, then there is a sentence δ such that (i) $\delta \leftrightarrow \varphi(\overline{\ulcorner\delta\urcorner})$ is true, and moreover (ii) $T \vdash \delta \leftrightarrow \varphi(\overline{\ulcorner\delta\urcorner})$.*

Dropping the subscript 'T', let's put $\alpha =_{\text{def}} \forall\mathsf{z}(\mathsf{Diag}(\mathsf{x}, \mathsf{z}) \to \varphi(\mathsf{z}))$.[1] Now let δ be the diagonalization of α. By definition, then, δ is $\forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z}) \to \varphi(\mathsf{z}))$.

We now show that (i) and (ii) are true for this δ. The proofs *look* a bit complicated at first glance; but after our initial construction of δ, we are really just applying various definitions and inferring easy consequences.

Proof for (i). Because diagonalizing α yields δ, by definition $diag(\ulcorner\alpha\urcorner) = \ulcorner\delta\urcorner$. Since Diag expresses $diag$, $\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \overline{\ulcorner\delta\urcorner})$ will be true; indeed, $\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z})$ is *only* satisfied by $\ulcorner\delta\urcorner$.

So $\forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z}) \to \varphi(\mathsf{z}))$ is true if and only if $\ulcorner\delta\urcorner$ satisfies $\varphi(\mathsf{z})$. In other words, δ is true if and only if $\varphi(\overline{\ulcorner\delta\urcorner})$ is true. ☒

Proof for (ii). Since Diag captures $diag$ in T, (iv) from Defn. 37 above tells us that *if* $diag(\ulcorner\alpha\urcorner) = \ulcorner\delta\urcorner$, *then* $T \vdash \forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z}) \leftrightarrow \mathsf{z} = \overline{\ulcorner\delta\urcorner})$.

But, as we noted before, $diag(\ulcorner\alpha\urcorner) = \ulcorner\delta\urcorner$. Hence we can conclude that, indeed, $T \vdash \forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z}) \leftrightarrow \mathsf{z} = \overline{\ulcorner\delta\urcorner})$.

So T proves the equivalence of $\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z})$ and $\mathsf{z} = \overline{\ulcorner\delta\urcorner}$. Therefore T can also prove the equivalence of δ, i.e. $\forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\alpha\urcorner}, \mathsf{z}) \to \varphi(\mathsf{z}))$, with $\forall\mathsf{z}(\mathsf{z} = \overline{\ulcorner\delta\urcorner} \to \varphi(\mathsf{z}))$. But the latter is trivially equivalent to $\varphi(\overline{\ulcorner\delta\urcorner})$.

Hence $T \vdash \delta \leftrightarrow \varphi(\overline{\ulcorner\delta\urcorner})$. ☒

A bit of jargon: by a mild abuse of mathematical terminology, we say

[1] Think of it this way: since we are representing the *diag* function with a two-place relational expression Diag rather than a one-place function expression diag, α is our substitute for the perhaps more natural but unavailable $\varphi(\mathsf{diag}(\mathsf{x}))$.

Defn. 52. *If δ is such that $T \vdash \delta \leftrightarrow \varphi(\overline{\ulcorner\delta\urcorner})$, then it is said to be a* fixed point *for φ.*

So the Diagonalization Lemma – or rather, part (ii) of it – is often called the Fixed Point Theorem: for appropriate theories T, every one-place open sentence has a fixed point.

13.3 Incompleteness from the Diagonalization Lemma

Suppose as usual that $Prf_T(m, n)$ is the relation which holds just if m numbers a T proof of a sentence with g.n. n (we continue to assume that we have a normal Gödel-numbering scheme in place). If T is p.r. axiomatized and contains Q, this relation is p.r. decidable and can be expressed and captured in T by a wff $\mathsf{Prf}_T(\mathsf{x}, \mathsf{y})$. And now we pick up again an idea we first met in §3.6:

Defn. 16. *Put* $\mathsf{Prov}_T(\mathsf{x}) =_{\mathrm{def}} \exists\mathsf{z}\mathsf{Prf}_T(\mathsf{z}, \mathsf{x})$ *(where the quantifier runs over all the numbers in the domain). Then* $\mathsf{Prov}_T(\overline{\mathsf{n}})$*, i.e.* $\exists\mathsf{z}\mathsf{Prf}_T(\mathsf{z}, \overline{\mathsf{n}})$*, is true iff some number Gödel-numbers a T-proof of the wff with Gödel-number n, i.e. is true just if the wff with code number n is a T-theorem. So* $\mathsf{Prov}_T(\mathsf{x})$ *is naturally called a* provability predicate*.*

$\mathsf{Prov}_T(\overline{\ulcorner\varphi\urcorner})$ is true, therefore, just when φ is a theorem.

And now we have a general result about fixed points for the *negation* of such a provability predicate:

Theorem 48. *Suppose T is p.r. axiomatized, contains Q, and some sentence γ is a fixed point for $\neg\mathsf{Prov}_T$; in other words, suppose $T \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\gamma\urcorner})$. Then (i) if T is consistent, $T \nvdash \gamma$. And (ii) if T is ω-consistent, $T \nvdash \neg\gamma$.*

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

Proof for (i). Suppose $T \vdash \gamma$. Then, since $T \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\gamma\urcorner})$, we have $T \vdash \neg\mathsf{Prov}(\overline{\ulcorner\gamma\urcorner})$. But if there *is* a proof of γ, then for some m, $Prf(m, \ulcorner\gamma\urcorner)$, so $T \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \overline{\ulcorner\gamma\urcorner})$, since T captures Prf by Prf. Hence $T \vdash \exists\mathsf{x}\,\mathsf{Prf}(\mathsf{x}, \overline{\ulcorner\gamma\urcorner})$, i.e. we also have $T \vdash \mathsf{Prov}(\overline{\ulcorner\gamma\urcorner})$, making T inconsistent. So if T is consistent, $T \nvdash \gamma$. ⊠

Proof for (ii). Suppose $T \vdash \neg\gamma$. Then, since $T \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\gamma\urcorner})$, we have $T \vdash \mathsf{Prov}(\overline{\ulcorner\gamma\urcorner})$, i.e. $T \vdash \exists\mathsf{x}\,\mathsf{Prf}(\mathsf{x}, \overline{\ulcorner\gamma\urcorner})$. Given T is consistent and proves $\neg\gamma$, there is no proof of γ, i.e. for every m, not-$Prf(m, \ulcorner\gamma\urcorner)$, whence for every m, $T \vdash \neg\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\ulcorner\gamma\urcorner})$. So we have a $\psi(\mathsf{x})$ such that T proves $\exists\mathsf{x}\psi(\mathsf{x})$ while it refutes each instance $\psi(\overline{\mathsf{m}})$.

Hence, if T is consistent and $T \vdash \neg\gamma$, T is ω-inconsistent. So if T is ω-consistent (and hence consistent), $T \nvdash \neg\gamma$. ⊠

But part (ii) of the Diagonalization Lemma tells us in particular that

Theorem 49. *If T is p.r. axiomatized, contains* Q*, then there exists a sentence γ such that* $T \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\gamma\urcorner})$.

So putting our two theorems together we get incompleteness again. And note that all that is required so far is that we are working with a wff Prf_T that captures Prf_T. Any fixed point γ for the negation of any provability predicate Prov_T will give us a formally undecidable wff.[2]

Suppose though that we start from a wff Prf_T that canonically captures Prf_T. Prf_T will then be Σ_1. Since Prov_T is built by existentially quantifying a Σ_1 wff, it too is Σ_1. So $\neg\mathsf{Prov}_T$ is Π_1. And we can check that the diagonalization construction then produces a Π_1 fixed point γ.[3]

So putting everything together, we recover again the syntactic First Theorem, Theorem 44.

13.4 Proving our old G_T is a fixed point for $\neg\mathsf{Prov}_T$

How does our new proof of the syntactic incompleteness theorem relate to the old one?

Our Gödel sentence G_T was the diagonalization of $\mathsf{U}(\mathsf{y}) =_{\text{def}} \forall\mathsf{x}\neg\mathsf{Prfd}(\mathsf{x},\mathsf{y})$. G_T is true if and only if unprovable-in-T. This fact can now be *expressed* inside T itself, by the wff $\mathsf{G}_T \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\mathsf{G}_T\urcorner})$ (and compare our informal Theorem 5 in §3.6). But T doesn't just express this fact but can *prove* it too. We have

Theorem 50. *If T is p.r. axiomatized and contains* Q*,* $T \vdash \mathsf{G}_T \leftrightarrow \neg\mathsf{Prov}_T(\overline{\ulcorner\mathsf{G}_T\urcorner})$.

In other words, our G_T is one of those fixed points for $\neg\mathsf{Prov}_T$. The proof is a bit fiddly but elementary:

Proof. Dropping subscripts again, $Prfd(m,n)$ holds iff $Prf(m, diag(n))$. We can therefore fix on the following canonical definition:

$$\mathsf{Prfd}(\mathsf{x},\mathsf{y}) =_{\text{def}} \exists\mathsf{z}(\mathsf{Prf}(\mathsf{x},\mathsf{z}) \land \mathsf{Diag}(\mathsf{y},\mathsf{z})).$$

And now let's do some elementary manipulations:

$$\begin{aligned}
\mathsf{G} =_{\text{def}}\ & \forall\mathsf{x}\neg\mathsf{Prfd}(\mathsf{x},\overline{\ulcorner\mathsf{U}\urcorner}) && \text{(Defn. 48)}\\
\leftrightarrow\ \ & \forall\mathsf{x}\neg\exists\mathsf{z}(\mathsf{Prf}(\mathsf{x},\mathsf{z}) \land \mathsf{Diag}(\overline{\ulcorner\mathsf{U}\urcorner},\mathsf{z})) && \text{(definition of } \mathsf{Prfd}\text{)}\\
\leftrightarrow\ \ & \forall\mathsf{x}\forall\mathsf{z}\neg(\mathsf{Prf}(\mathsf{x},\mathsf{z}) \land \mathsf{Diag}(\overline{\ulcorner\mathsf{U}\urcorner},\mathsf{z})) && \text{(pushing in the negation)}\\
\leftrightarrow\ \ & \forall\mathsf{z}\forall\mathsf{x}\neg(\mathsf{Prf}(\mathsf{x},\mathsf{z}) \land \mathsf{Diag}(\overline{\ulcorner\mathsf{U}\urcorner},\mathsf{z})) && \text{(swapping quantifiers)}\\
\leftrightarrow\ \ & \forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\mathsf{U}\urcorner},\mathsf{z}) \to \neg\exists\mathsf{x}\,\mathsf{Prf}(\mathsf{x},\mathsf{z})) && \text{(rearranging after `}\forall\mathsf{z}\text{')}\\
=_{\text{def}}\ & \forall\mathsf{z}(\mathsf{Diag}(\overline{\ulcorner\mathsf{U}\urcorner},\mathsf{z}) \to \neg\mathsf{Prov}(\mathsf{z})) && \text{(Defn. 16)}
\end{aligned}$$

[2] Warning. Some authors call *any* such fixed point for a predicate $\neg\mathsf{Prov}_T$ built from *any* wff which captures Prf_T a Gödel sentence for T. That's fine, as long as you are alert to the fact that not everything that is true of canonical Gödel sentences in the narrow sense we introduced in the preceding chapters is true of Gödel sentences in this new wider sense.

[3] Why? Put $\alpha =_{\text{def}} \forall\mathsf{z}(\mathsf{Diag}(\mathsf{y},\mathsf{z}) \to \neg\mathsf{Prov}(\mathsf{z}))$, or equivalently $\forall\mathsf{z}(\neg\mathsf{Diag}(\mathsf{y},\mathsf{z}) \lor \neg\mathsf{Prov}(\mathsf{z}))$. What's inside the universal quantifier is a disjunction of Π_1 wffs which is easily seen to be itself Π_1. So α, the universal quantification of a Π wff, is also Π_1. Diagonalizing α to get the fixed point γ means replacing a variable with a numeral, and the result is still Π_1.

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. Therefore we have

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

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

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

Putting those two results together, we immediately get

$$T \vdash \mathsf{G} \leftrightarrow \forall \mathsf{z}(\mathsf{z} = \overline{\ulcorner \mathsf{G} \urcorner} \to \neg \mathsf{Prov}(\mathsf{z})).$$

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

13.5 Diagonalization the easy way?

An afterword for enthusiasts. Our proof of the standard form of the Diagonalization Lemma was not particularly difficult. But it was a bit messy, and it relied on a not-immediately-obvious initial construction. So it is interesting see that we can get a variant Lemma more or less for free. But this time, the application of the variant Lemma takes a little more care and attention.[4]

Suppose, then, that T is the usual kind of theory, p.r. axiomatized and containing Q (and so incorporating the language of basic arithmetic). We can effectively list its open wffs with the one free variable x in the usual way $\varphi_0(\mathsf{x}), \varphi_1(\mathsf{x}), \varphi_2(\mathsf{x}), \varphi_3(\mathsf{x}), \ldots$. Now add to T's language a corresponding infinite sequence of *new* constants $\mathsf{c}_0, \mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \ldots$, and expand our assumed normal Gödel-numbering scheme for T to give code numbers to each of the new constants. Then:

Defn. 53. *The* Kripke extension T^K *of a theory T whose open wffs with free variable x are $\varphi_i(\mathsf{x})$ is the result of adding all the new constants c_i to T, together with all the new axioms $\mathsf{c}_i = \overline{\ulcorner \varphi_i(\mathsf{c}_i) \urcorner}$.*

The idea, then, is to introduce names c_i governed by axioms which ensure that a wff $\varphi_i(\mathsf{c}_i)$ is – via the coding – more directly 'about' itself, without the usual diagonalization trickery.

We need to read into the record six quick and easy observations:

O1. T^K is still p.r. axiomatized. We can still effectively determine (without open-ended searches) whether a given wff is an axiom – why?

O2. Any wff without c-constants which is provable in T^K is already provable in T. If we add constants which are in effect just shorthand for certain standard numerals (different constants, different numerals), then the only new wffs we can prove will be wffs with those constants.

[4] The construction in this section, due to Kripke, has long been folklore but was only published in 2020, in 'Gödel's Theorem and direct self-reference' (*arXiv* 2010.11979).

O3. Any wff of T^K with some c-constant(s) is provably-in-T^K equivalent to a wff without c-constants. Just use the c-axioms and Leibniz's Law to replace a constant with a standard numeral.

O4. For any wff α of T's language, if $T^K \nvdash \alpha$, then $T \nvdash \alpha$. Trivially, if the theory with extra axioms can't prove α, then the original theory can't.

O5. If T is consistent, so is T^K. Again, adding distinct constants for distinct numerals can't make a consistent theory inconsistent.

O6. If T is ω-consistent, so is T^K. We prove the contrapositive. Suppose T^K is ω-inconsistent, so there is an open wff $\psi(\mathsf{x})$, such that $T^K \vdash \exists \mathsf{x}\psi(\mathsf{x})$ yet for each number n we have $T^K \vdash \neg\psi(\overline{\mathsf{n}})$. Apply (O3) to ψ if necessary to get a provably equivalent constant-free ψ', and then $T^K \vdash \exists \mathsf{x}\psi'(\mathsf{x})$ while for each number n we have $T^K \vdash \neg\psi'(\overline{\mathsf{n}})$. But since these wffs in ψ' are constant-free, we can apply (O2) to get $T \vdash \exists \mathsf{x}\psi'(\mathsf{x})$ while for each number n we have $T \vdash \neg\psi'(\overline{\mathsf{n}})$, which makes the original T ω-inconsistent.

We now have the following variant of (the second part) of the original Diagonal Lemma:

Theorem 51. *If T is a p.r. axiomatized theory which contains Q, and φ is a one-place open sentence of T's language, then there is a sentence δ of T's Kripke extension T^K such that $T^K \vdash \delta \leftrightarrow \varphi(\overline{\ulcorner\delta\urcorner})$.*

Proof. Trivially, φ will be some φ_i. By the relevant c-axiom, $T^K \vdash \varphi_i(\mathsf{c}_i) \leftrightarrow \varphi_i(\overline{\ulcorner\varphi_i(\mathsf{c}_i)\urcorner})$. So just put $\delta = \varphi_i(\mathsf{c}_i)$ and we are done. ⊠

That was easy! Now take the case where $\varphi(\mathsf{x})$ is $\neg\mathsf{Prov}_{T^K}(\mathsf{x})$, the negation of the canonical provability predicate for T^K. This is formed in the (constant-free) language of basic arithmetic, i.e. is a one-place predicate already available in T. So we can invoke our theorem to show that there is in particular a fixed point γ of the language of T^K such that $T^K \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_{T^K}(\overline{\ulcorner\gamma\urcorner})$.

And now we can apply Theorem 48. Since we have a fixed point γ for $\neg\mathsf{Prov}_{T^K}$, we know that (i) if T^K is consistent, $T^K \nvdash \gamma$. And (ii) if T^K is ω-consistent, $T^K \nvdash \neg\gamma$.

Fine. But this doesn't quite get us back to the desired incompleteness theorem for our original theory T. For a start, the fixed point γ (on our current construction) involves one of those new constants, so γ won't belong to the language of T. So we now need to use our observations above.

By (O3), we know that there is a constant-free γ' which is provably equivalent to γ, so (i′) if T^K is consistent, $T^K \nvdash \gamma'$; and (ii′) if T^K is ω-consistent, $T^K \nvdash \neg\gamma'$. So then applying (O5) and (O4), we get (i″) if T is consistent, so is T^K, so $T^K \nvdash \gamma'$, and hence $T \nvdash \gamma'$. And applying (O6) and (O4) we get (ii″) if T is ω-consistent so is T^K, so $T^K \nvdash \neg\gamma'$, hence $T \nvdash \neg\gamma'$. Which gives us the First Incompleteness Theorem for T again.

In summary, then. To get incompleteness from a Diagonalization Lemma, you rely on the key Theorem 48 either way. But there are two routes. Either you prove the Lemma in the original way (with its slight air of trickery) and get to apply Theorem 48 'neat'. Or you use Kripke's construction to arrive at an easier variant Lemma (adding constants stipulated to do the necessary work), but this time have to take a bit of care to massage away the constants after you've applied Theorem 48. So you pays your money and you takes your choice!