16 Proving the Second Incompleteness Theorem

In his original 1931 paper, Gödel states a version of the Second Theorem. But he doesn't spell out a full proof. He simply claims that the reasoning for the First Theorem is so elementary that a strong enough theory must be able to replicate the reasoning and prove the Formalized First Theorem; and then he notes that this implies the Second Theorem.[1]

The hard work of taking a strong enough T and checking that it really does prove the Formalized First Theorem was first done for a particular case by David Hilbert and Paul Bernays in their *Grundlagen der Mathematik* of 1939. The details of their proof are – the story goes – due to Bernays, who had discussed it with Gödel during a transatlantic voyage.

Now, Hilbert and Bernays helpfully isolated what we can call *derivability conditions* on the predicate Prov_T, conditions whose satisfaction is indeed enough for a theory T to prove the Formalized First Theorem. Later, Martin H. Löb gave a rather neater version of these conditions, giving us the so-called *HBL conditions* have become standard and which we invoke in this chapter.

16.1 Sharpening the Second Theorem

Recall: $\mathsf{Prov}_T(\mathsf{y})$ abbreviates $\exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v}, \mathsf{y})$; and Prf_T is a Σ_1 expression. So arguing about provability inside T will involve establishing some general claims involving Σ_1 expressions. And how do we prove general arithmetical claims? Using induction is the default method.

It therefore looks a good bet that we will get $T \vdash \mathsf{Con}_T \to \neg\mathsf{Prov}_T(\overline{\ulcorner \mathsf{G}_T \urcorner})$ if T at least has Σ_1-induction – meaning that T's axioms include (the universal closures of) all instances of the first-order Induction Schema where the induction predicate φ is no more complex than Σ_1.

In §7.5 we introduced $\mathsf{I}\Sigma_1$ as the standard label for the result of adding that amount of induction to Q. So it is a plausible conjecture that we can in fact show the following, a fleshed-out version of our skeletal Theorem 57:

[1] I am told that Gödel's shorthand notebooks at the time suggest that he in fact hadn't then worked out a detailed proof of the first step.

Theorem 60. *If T is p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, then T proves the Formalized First Theorem, i.e. $T \vdash \mathsf{Con}_T \to \neg\mathsf{Prov}_T(\overline{\ulcorner \mathsf{G}_T \urcorner})$*

And if that is right, by the argument of §15.3 we get an improved version of Theorem 58.

Theorem 61. *If T is consistent, p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, then $T \nvdash \mathsf{Con}_T$.*

So how can we prove Theorem 60?

16.2 The box notation

To improve readability, we introduce some standard notation. We will henceforth abbreviate $\mathsf{Prov}_T(\overline{\ulcorner \varphi \urcorner})$ simply by $\Box_T\varphi$. This is perhaps a bit naughty, as the new notation hides away some Gödel-numbering. But the notation is standard and is safe enough if we keep our wits about us.[2]

So in particular, $\neg\mathsf{Prov}_T(\overline{\ulcorner \mathsf{G}_T \urcorner})$ can be abbreviated $\neg\Box_T\mathsf{G}_T$. Thus in our new notation, the Formalized First Theorem is $\mathsf{Con}_T \to \neg\Box_T\mathsf{G}_T$. Moreover, Con_T can now alternatively be abbreviated as $\neg\Box_T\bot$.

To reduce clutter, we will also very often drop the explicit subscript T from the box symbol and elsewhere, and let context supply it.

16.3 Proving the Formalized First Theorem

First a standard definition: we will say (dropping subscripts)

Defn. 58. *The derivability conditions hold in T if and only if, for any T-sentences φ, ψ,*

1. *If $T \vdash \varphi$, then $T \vdash \Box\varphi$;*
2. *$T \vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$;*
3. *$T \vdash \Box\varphi \to \Box\Box\varphi$.*

Here, (1) tells us that if T can prove φ, then (via the relevant Gödel-numbering which enables T to code up claims about provability-in-T), T 'knows' it can prove φ. (2) tells us that T is aware of modus ponens; if T knows it can prove $(\varphi \to \psi)$, and knows it can prove φ, then T knows it can prove ψ too. And (3) tells us that if T knows it can prove φ, it can code up that proof so it can prove that φ is provable. We might reasonably expect that these conditions will hold for a strong enough theory T.

And now we can indeed show that

[2] If you are familiar with modal logic, then you will immediately recognize the conventional symbol for the necessity operator. And the parallels and differences between '"$1 + 1 = 2$" is provable (in T)' and 'It is necessarily true that $1 + 1 = 2$' are highly suggestive. These parallels and differences are the topic of 'provability logic', the subject of a contemporary classic, Boolos's *The Logic of Provability.*

Theorem 62. *If T is p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, then the derivability conditions hold for T.*

However, demonstrating this is a seriously tedious task, and I certainly *don't* propose to do hack through the annoying technical details. In these notes we will just take this technical theorem as given.

We also can show, considerably more easily, that

Theorem 63. *If T is p.r. axiomatized, contains Q, and the derivability conditions hold for T, then T proves the Formalized First Theorem.*

Proof. This is just a mildly fun exercise in box-pushing, with the target of showing $T \vdash \mathsf{Con}_T \to \neg\Box_T \mathsf{G}_T$.

First, since T is p.r. axiomatized and contains Q, Theorem 50 holds. So, in our new symbolism, we have $T \vdash \mathsf{G} \leftrightarrow \neg\Box\mathsf{G}$.

Second, since T contains Q and standard logic, we have

$$T \vdash \neg\varphi \to (\varphi \to \bot).$$

(whichever way we read the absurdity constant – see Defn. 56). Given this and the derivability condition (C1), we get

$$T \vdash \Box(\neg\varphi \to (\varphi \to \bot)).$$

So given the derivability condition (C2) and using modus ponens, it follows that for any φ

A. $T \vdash \Box\neg\varphi \to \Box(\varphi \to \bot)$.

We now argue as follows, dropping the subscript Ts:

1.	$T \vdash \mathsf{G} \to \neg\Box\mathsf{G}$	Half of Theorem 50
2.	$T \vdash \Box(\mathsf{G} \to \neg\Box\mathsf{G})$	From 1, given C1
3.	$T \vdash \Box\mathsf{G} \to \Box\neg\Box\mathsf{G}$	From 2, using C2
4.	$T \vdash \Box\neg\Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	Instance of A
5.	$T \vdash \Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	From 3 and 4
6.	$T \vdash \Box\mathsf{G} \to (\Box\Box\mathsf{G} \to \Box\bot)$	From 5, using C2 and logic
7.	$T \vdash \Box\mathsf{G} \to \Box\Box\mathsf{G}$	Instance of C3
8.	$T \vdash \Box\mathsf{G} \to \Box\bot$	From 6 and 7
9.	$T \vdash \neg\Box\bot \to \neg\Box\mathsf{G}$	Contraposing
10.	$T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$	Definition of Con ⊠

Now we put the last two theorems together. If T is p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, the derivability conditions will hold (by Theorem 62) and it contains Q. Hence (by Theorem 63) if T is p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, it proves the Formalized First Theorem.

Which gives us our target Theorem 60.

16.4 The equivalence of Con_T with G_T

We'll assume throughout this section that T is a p.r. axiomatized theory such that the derivability conditions hold. And first, we will show that in that case, G_T and Con_T are provably equivalent in T.

We have just shown in the previous Theorem that $T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$ (suppressing subscripts!). We can now prove the converse, $T \vdash \neg\Box\mathsf{G} \to \mathsf{Con}$. In fact, we have a more general result:

Theorem 64. *For any sentence φ, $T \vdash \neg\Box\varphi \to \mathsf{Con}$.*

Proof. We argue as follows:

1.	$T \vdash \bot \to \varphi$	Logic!
2.	$T \vdash \Box(\bot \to \varphi)$	From 1, given C1
3.	$T \vdash \Box\bot \to \Box\varphi$	From 2, given C2
4.	$T \vdash \neg\Box\varphi \to \neg\Box\bot$	Contraposing
5.	$T \vdash \neg\Box\varphi \to \mathsf{Con}$	Definition of Con ☒

This little theorem has a rather remarkable corollary. Since T can't prove Con, the theorem tells us that T doesn't entail $\neg\Box\varphi$ for any φ at all. Hence T doesn't ever 'know' that it can't prove φ, even when it can't! In sum, suppose that T satisfies the now familiar conditions: by (C1), T knows all about what it *can* prove; but we have just shown that it knows nothing about what it *can't* prove.

Now, as a particular instance of our last theorem, we have $T \vdash \neg\Box\mathsf{G} \to \mathsf{Con}$. So putting that together with Theorem 63, we have $T \vdash \mathsf{Con} \leftrightarrow \neg\Box\mathsf{G}$. And now combine *that* with Theorem 50 which tells us that $T \vdash \mathsf{G} \leftrightarrow \neg\Box\mathsf{G}$, and lo and behold we've shown

Theorem 65. *If T is p.r. axiomatized and contains $\mathsf{I}\Sigma_1$, then $T \vdash \mathsf{Con} \leftrightarrow \mathsf{G}$.*

This means that, not only do we have $T \nvdash \mathsf{Con}$, we also have (assuming in addition that T is ω-consistent) $T \nvdash \neg\mathsf{Con}$. In other words, Con is formally undecidable by T.

But Con is *not* self-referential. That observation should scotch any lingering suspicion that undecidable sentences provided by Gödelian arguments are all tainted by potentially paradoxical self reference.

We next prove

Theorem 66. $T \vdash \mathsf{Con} \leftrightarrow \neg\Box\mathsf{Con}$.

Proof. The direction from right to left is an instance of Theorem 64. For the other direction, note that

1.	$T \vdash \mathsf{Con} \to \mathsf{G}$	Already proved
2.	$T \vdash \Box(\mathsf{Con} \to \mathsf{G})$	From 1, given C1
3.	$T \vdash \Box\mathsf{Con} \to \Box\mathsf{G}$	From 2, given C2

4. $T \vdash \neg\Box\mathsf{G} \rightarrow \neg\Box\mathsf{Con}$ Contraposing
5. $T \vdash \mathsf{Con} \rightarrow \neg\Box\mathsf{Con}$ Using Thm. 63 ⊠

So: this shows that Con (like G) is also a fixed point of the negated provability predicate (see again Defn. 52 and §13.3).

As we have noted before, some authors refer to any fixed point of the negated provability predicate as a Gödel sentence. Fine. That's one way of using the jargon. But if you adopt the broad usage, you must be careful with your informal commentary. For example, not all Gödel sentences in the broad sense indirectly 'say' *I am unprovable*: Con is a case in point.

16.5 Consistent theories that 'prove' their own inconsistency

An amusing and instructive afterthought to finish this chapter.

An *ω-consistent* (and so consistent) T can't prove $\neg\mathsf{Con}_T$, as we've just noted. By contrast, a consistent but *ω-inconsistent* T can have $\neg\mathsf{Con}_T$ as a theorem. (Being ω-inconsistent, T won't be sound; that's how it can have a false theorem!)

The proof is actually straightforward, once we note a simple lemma. Suppose S and R are two p.r. axiomatized theories, which share a deductive logic; and suppose every axiom of the simpler theory S is also an axiom of the richer theory R. It is then a trivial logical truth that, if the richer theory R is consistent, then the simpler theory S must be consistent too. And the arithmetical claim that encodes this fact can also be formally proved:

Theorem 67. *Under the given conditions, with theory R extending theory S, $R \vdash \mathsf{Con}_R \rightarrow \mathsf{Con}_S$.*

So now take our simpler theory S to be PA. Take the richer theory R to be PA augmented by the extra axiom $\neg\mathsf{G}_{\mathsf{PA}}$. We have met this richer theory briefly before: by theorem 43, R is consistent but ω-inconsistent. Since R trivially proves $\neg\mathsf{G}_{\mathsf{PA}}$, and PA (and hence R) proves G_{PA} is equivalent to $\mathsf{Con}_{\mathsf{PA}}$, R proves $\neg\mathsf{Con}_{\mathsf{PA}}$. So – using our last theorem and modus tollens! – R proves $\neg\mathsf{Con}_R$.

Summing that up,

Theorem 68. *Assuming* PA *is consistent, the theory* $\mathsf{PA} + \neg\mathsf{G}_{\mathsf{PA}}$ *is a consistent theory which 'proves' itself inconsistent (i.e. it entails the* negation *of its own consistency sentence).*

And since R proves $\neg\mathsf{Con}_R$,

Theorem 69. *There can be a* consistent *theory R is such that the theory $R + \mathsf{Con}_R$ is* inconsistent.

What are we to make of these apparent absurdities? Well, giving the language of R, i.e. $\mathsf{PA} + \neg\mathsf{G}_{\mathsf{PA}}$, its standard arithmetical interpretation, the theory is unsound (since the Gödel sentence G_{PA} is true). So we shouldn't trust what R says about anything, especially about its own inconsistency when we derive

$\neg\mathsf{Con}_R$ from it. R doesn't really *prove* (in the ordinary sense of establish-as-true) its own inconsistency, since we don't accept the theory as correct on the standard interpretation! That's why we used scare quotes in stating Theorem 68.