

inp.1 Löb's Theorem

inc:inp:lob:
sec The Gödel sentence for a theory \mathbf{T} is a fixed point of $\neg\text{Prov}_T(y)$, i.e., a sentence γ such that

$$\mathbf{T} \vdash \neg\text{Prov}_T(\ulcorner \gamma \urcorner) \leftrightarrow \gamma.$$

It is not derivable, because if $\mathbf{T} \vdash \gamma$, (a) by derivability condition (1), $\mathbf{T} \vdash \text{Prov}_T(\ulcorner \gamma \urcorner)$, and (b) $\mathbf{T} \vdash \gamma$ together with $\mathbf{T} \vdash \neg\text{Prov}_T(\ulcorner \gamma \urcorner) \leftrightarrow \gamma$ gives $\mathbf{T} \vdash \neg\text{Prov}_T(\ulcorner \gamma \urcorner)$, and so \mathbf{T} would be inconsistent. Now it is natural to ask about the status of a fixed point of $\text{Prov}_T(y)$, i.e., a sentence δ such that

$$\mathbf{T} \vdash \text{Prov}_T(\ulcorner \delta \urcorner) \leftrightarrow \delta.$$

If it were derivable, $\mathbf{T} \vdash \text{Prov}_T(\ulcorner \delta \urcorner)$ by condition (1), but the same conclusion follows if we apply modus ponens to the equivalence above. Hence, we don't get that \mathbf{T} is inconsistent, at least not by the same argument as in the case of the Gödel sentence. This of course does not show that \mathbf{T} does derive δ .

We can make headway on this question if we generalize it a bit. The left-to-right direction of the fixed point equivalence, $\text{Prov}_T(\ulcorner \delta \urcorner) \rightarrow \delta$, is an instance of a general schema called a *reflection principle*: $\text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$. It is called that because it expresses, in a sense, that \mathbf{T} can “reflect” about what it can derive; basically it says, “If \mathbf{T} can derive φ , then φ is true,” for any φ . This is true for sound theories only, of course, and this suggests that theories will in general not derive every instance of it. So which instances can a theory (strong enough, and satisfying the derivability conditions) derive? Certainly all those where φ itself is derivable. And that's it, as the next result shows.

inc:inp:lob:
thm:lob **Theorem inp.1.** *Let \mathbf{T} be an axiomatizable theory extending \mathbf{Q} , and suppose $\text{Prov}_T(y)$ is a formula satisfying conditions P1–P3 from ???. If \mathbf{T} derives $\text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$, then in fact \mathbf{T} derives φ .*

Put differently, if $\mathbf{T} \not\vdash \varphi$, then $\mathbf{T} \not\vdash \text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$. This result is known as Löb's theorem.

The heuristic for the proof of Löb's theorem is a clever proof that Santa Claus exists. (If you don't like that conclusion, you are free to substitute any other conclusion you would like.) Here it is: explanation

1. Let X be the sentence, “If X is true, then Santa Claus exists.”
2. Suppose X is true.
3. Then what it says holds; i.e., we have: if X is true, then Santa Claus exists.
4. Since we are assuming X is true, we can conclude that Santa Claus exists, by modus ponens from (2) and (3).
5. We have succeeded in deriving (4), “Santa Claus exists,” from the assumption (2), “ X is true.” By conditional proof, we have shown: “If X is true, then Santa Claus exists.”

6. But this is just the sentence X . So we have shown that X is true.

7. But then, by the argument (2)–(4) above, Santa Claus exists.

A formalization of this idea, replacing “is true” with “is **derivable**,” and “Santa Claus exists” with φ , yields the proof of Löb’s theorem. The trick is to apply the fixed-point lemma to the **formula** $\text{Prov}_T(y) \rightarrow \varphi$. The fixed point of that corresponds to the sentence X in the preceding sketch.

*Proof of **Theorem inp.1**.* Suppose φ is a **sentence** such that **T derives** $\text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$. Let $\psi(y)$ be the **formula** $\text{Prov}_T(y) \rightarrow \varphi$, and use the fixed-point lemma to find a **sentence** θ such that **T derives** $\theta \leftrightarrow \psi(\ulcorner \theta \urcorner)$. Then each of the following is **derivable** in **T**:

- (1) $\theta \leftrightarrow (\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi)$ inc:inp:lob:
L-1
- θ is a fixed point of $\psi(y)$
- (2) $\theta \rightarrow (\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi)$ inc:inp:lob:
L-2
- from eq. (1)
- (3) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow (\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi)$ inc:inp:lob:
L-3
- from eq. (2) by condition P1
- (4) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \text{Prov}_T(\ulcorner \text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi \urcorner)$ inc:inp:lob:
L-4
- from eq. (3) using condition P2
- (5) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow (\text{Prov}_T(\ulcorner \text{Prov}_T(\ulcorner \theta \urcorner) \urcorner) \rightarrow \text{Prov}_T(\ulcorner \varphi \urcorner))$ inc:inp:lob:
L-5
- from eq. (4) using P2 again
- (6) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \text{Prov}_T(\ulcorner \text{Prov}_T(\ulcorner \theta \urcorner) \urcorner)$ inc:inp:lob:
L-6
- by **derivability** condition P3
- (7) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \text{Prov}_T(\ulcorner \varphi \urcorner)$ inc:inp:lob:
L-7
- from eq. (5) and eq. (6)
- (8) $\text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$ inc:inp:lob:
L-8
- by assumption of the theorem
- (9) $\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi$ inc:inp:lob:
L-9
- from eq. (7) and eq. (8)
- (10) $(\text{Prov}_T(\ulcorner \theta \urcorner) \rightarrow \varphi) \rightarrow \theta$ inc:inp:lob:
L-10
- from eq. (1)
- (11) θ inc:inp:lob:
L-11
- from eq. (9) and eq. (10)
- (12) $\text{Prov}_T(\ulcorner \theta \urcorner)$ inc:inp:lob:
L-12
- from eq. (11) by condition P1
- φ from eq. (8) and eq. (12) □

With Löb’s theorem in hand, there is a short proof of the second incompleteness theorem (for theories having a **derivability** predicate satisfying conditions

P1–P3): if $\mathbf{T} \vdash \text{Prov}_T(\ulcorner \perp \urcorner) \rightarrow \perp$, then $\mathbf{T} \vdash \perp$. If \mathbf{T} is consistent, $\mathbf{T} \not\vdash \perp$. So, $\mathbf{T} \not\vdash \text{Prov}_T(\ulcorner \perp \urcorner) \rightarrow \perp$, i.e., $\mathbf{T} \not\vdash \text{Con}_{\mathbf{T}}$. We can also apply it to show that δ , the fixed point of $\text{Prov}_T(x)$, is **derivable**. For since

$$\mathbf{T} \vdash \text{Prov}_T(\ulcorner \delta \urcorner) \leftrightarrow \delta$$

in particular

$$\mathbf{T} \vdash \text{Prov}_T(\ulcorner \delta \urcorner) \rightarrow \delta$$

and so by Löb's theorem, $\mathbf{T} \vdash \delta$.

Problem inp.1. Let \mathbf{T} be a computably axiomatized theory, and let Prov_T be a **derivability** predicate for \mathbf{T} . Consider the following four statements:

1. If $T \vdash \varphi$, then $T \vdash \text{Prov}_T(\ulcorner \varphi \urcorner)$.
2. $T \vdash \varphi \rightarrow \text{Prov}_T(\ulcorner \varphi \urcorner)$.
3. If $T \vdash \text{Prov}_T(\ulcorner \varphi \urcorner)$, then $T \vdash \varphi$.
4. $T \vdash \text{Prov}_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$

Under what conditions are each of these statements true?

Photo Credits

Bibliography