Never trust an unsound theory

Rasmus Blanck

Joint work with Christian Bennet

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

Logic Seminar, 6 May 2022

Introduction

- What is a Gödel sentence?
- Can we talk about *the* Gödel sentence of a theory?
- Intuition: The sentence δ constructed by the diagonal lemma to satisfy $\mathrm{PA} \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$ is a Gödel sentence for T.
- Is any sentence satisfying $\mathrm{PA} \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$ a Gödel sentence for T?
- The sentence δ constructed by the diagonal lemma to satisfy $T \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$?
- Any sentence satisfying $T \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$?

Preliminaries

- T, S, U are some r.e., consistent extensions of PA.
- $\mathrm{Pr}_T(x)$ is a standard Σ_1 provability predicate based on some fixed p.r. binumeration of T in PA.
- A sentence ϕ is true iff $\mathbb{N} \models \phi$.
- T is sound if everything provable in T is true.
- We use U to emphasise that the thory in question may well be unsound.
- T is Σ_1-complete if every true Σ_1 sentence is provable in T.
- T is ω-consistent if, for every formula $\phi(x)$, if T proves $\neg\phi(0), \neg\phi(1), \ldots$ then $T \nvdash \exists x \phi x$.
- PA is sound, Σ_1-complete, and ω-consistent.
- We do not distinguish between formulas and (the numerals for) their Gödel numbers.
- δ is a fixed point of $\phi(x)$ over T iff $T \vdash \delta \leftrightarrow \phi(\delta)$.
  (δ is a T-fixed point of $\phi(x)$).
- δ is a *Gödelian* sentence of T iff $T \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$. So a Gödelian sentence of T is a T-fixed point of $\neg\mathrm{Pr}_T(x)$.

This talk is based on

Bennet & Blanck: Never trust an unsound theory.
Accepted for publication in *Theoria*. (Henceforth B&B)

which is written in response to

Lajevardi & Salehi: There may be many arithmetical Gödel sentences.
Philosophia Mathematica 29(2):278–287, 2021. (Henceforth L&S)

Summary of Lajevardi & Salehi

Two pertinent observations:

- The first incompleteness theorem applies to unsound theories too. (Depending on what we mean by "the first incompleteness theorem".)
- There are unsound theories that are ω-consistent.

Two theorems and one corollary:

1. For all sentences ϕ: $T \nvdash \phi$ iff there is an $S \vdash T$ s.t. $S \vdash \phi \leftrightarrow \neg\mathrm{Pr}_S(\phi)$.
2. For all T-fixed points ϕ of $\neg\mathrm{Pr}_T(x)$: ϕ is true iff T is sound.
3. Unsound theories have both true and false Gödelian sentences.

And one inconclusive argument:

- There are *Gödelian* sentences with different truth values, therefore we must not talk about the *Gödel* sentence.

Four versions of Gödel's first for unsound theories (1/2)

Theorem

Let U be any r.e., consistent extension of PA. If δ is any sentence satisfying $U \vdash \delta \leftrightarrow \neg\mathrm{Pr}_U(\delta)$, then $U \nvdash \delta$.

Proof.

Let δ be any sentence satisfying the equivalence. Suppose $U \vdash \delta$. Then $\mathrm{Pr}_U(\delta)$ is true. By Σ_1-completeness of PA we get $\mathrm{PA} \vdash \mathrm{Pr}_U(\delta)$, so $U \vdash \mathrm{Pr}_U(\delta)$, and $U \vdash \neg\delta$. Then U is inconsistent, so $U \nvdash \delta$. □

Theorem

Let U be any r.e., ω-consistent extension of PA. If δ is any sentence satisfying $U \vdash \delta \leftrightarrow \neg\mathrm{Pr}_U(\delta)$, then $U \nvdash \delta, \neg\delta$.

Proof.

Suppose $U \vdash \neg\delta$. Since U is consistent, $U \nvdash \delta$. So $\neg\mathrm{Prf}_U(\delta, k)$ is true for each $k \in \omega$. By Σ_1-completeness of PA, $U \vdash \neg\mathrm{Prf}_U(\delta, k)$ for each $k \in \omega$. But since $U \vdash \neg\delta$, $U \vdash \mathrm{Pr}_U(\delta)$, and $U \vdash \exists x \mathrm{Prf}_U(\delta, x)$. So U is ω-inconsistent. □

Four versions of Gödel's first (2/2)

Theorem

Let U be any consistent, r.e. extension of PA. If γ is any sentence satisfying $\mathrm{PA} \vdash \gamma \leftrightarrow \neg\mathrm{Pr}_U(\gamma)$, then $U \nvdash \gamma$ and γ is true.

Proof.

Suppose $U \vdash \gamma$. Then $\mathrm{PA} \vdash \mathrm{Pr}_U(\gamma)$, so $\mathrm{PA} \vdash \neg\gamma$. Then U, extending PA is inconsistent. Hence $U \nvdash \gamma$. So $\neg\mathrm{Pr}_U(\gamma)$ is true. By *soundness* of PA, $\gamma \leftrightarrow \neg\mathrm{Pr}_U(\gamma)$ is true, so γ is true. □

Theorem

Let U be any ω-consistent, r.e. extension of PA. If γ is any sentence satisfying $\mathrm{PA} \vdash \gamma \leftrightarrow \neg\mathrm{Pr}_U(\gamma)$, then $U \nvdash \gamma, \neg\gamma$ and γ is true.

Proof.

By combining the earlier proofs. □

Löb's theorem and Gödel's 2nd

Theorem (Löb's theorem)

If $T \vdash \mathrm{Pr}_T(\phi) \to \phi$*, then* $T \vdash \phi$*.*

Proved using Löb's derivability conditions:

- L1 If $T \vdash \phi$, then $\mathrm{PA} \vdash \mathrm{Pr}_T(\phi)$
- L2 $\mathrm{PA} \vdash \mathrm{Pr}_T(\phi \to \psi) \to (\mathrm{Pr}_T(\phi) \to \mathrm{Pr}_T(\psi))$
- L3 $\mathrm{PA} \vdash \mathrm{Pr}_T(\phi) \to \mathrm{Pr}_T(\mathrm{Pr}_T(\phi))$

Theorem (Gödel's 2nd)

If δ *is any sentence satisfying* $U \vdash \delta \leftrightarrow \neg\mathrm{Pr}_U(\delta)$*, then* $U \vdash \delta \leftrightarrow \mathrm{Con}_U$*.*

Proof.

By construction together with Löb's conditions. □

Theorem 1 (L&S)

For every sentence ϕ, the following are equivalent:

1. $T \nvdash \phi$
2. *there is a consistent theory S extending T such that $S \vdash \phi \leftrightarrow \neg \mathrm{Pr}_S(\phi)$.*

Theorem A (B&B)

For every formula $\theta(x)$, and every sentence ϕ, the following are equivalent:

1. $T \nvdash \neg(\phi \leftrightarrow \theta(\phi))$
2. *there is a consistent theory S extending T such that $S \vdash \phi \leftrightarrow \theta(\phi)$.*

Proof.

Trivial: Observe that $T \nvdash \neg(\phi \leftrightarrow \theta(\phi))$ iff $S = T + \phi \leftrightarrow \theta(\phi)$ is consistent.
It is sometimes useful to choose S more carefully:

1. If $T \nvdash \phi \rightarrow \neg\theta(\phi)$, take $S = T + \phi + \theta(\phi)$.
2. If $T \nvdash \neg\theta(\phi) \rightarrow \phi$, take $S = T + \neg\theta(\phi) + \neg\phi$. □

Proof of Theorem 1 from Theorem A.

- $2 \Rightarrow 1$: Let S be a consistent extension of T and ϕ a sentence such that $S \vdash \phi \leftrightarrow \neg\mathrm{Pr}_S(\phi)$. If $S \vdash \phi$, then $\mathrm{PA} \vdash \mathrm{Pr}_S(\phi)$, so S, extending PA, is inconsistent. Hence $S \nvdash \phi$, and therefore $T \nvdash \phi$.

- $1 \Rightarrow 2$: Suppose that $T \nvdash \phi$. By Löb's theorem, $T \nvdash \mathrm{Pr}_T(\phi) \to \phi$. This is case 2 of the proof of Theorem A (taking $\theta(x) \coloneqq \neg\mathrm{Pr}_T(x)$), so let $S = T + \mathrm{Pr}_T(\phi) + \neg\phi$. Since S extends T, we have $T \vdash \mathrm{Pr}_T(\phi) \to \mathrm{Pr}_S(\phi)$. Then $S \vdash \mathrm{Pr}_S(\phi) \wedge \neg\phi$, so S and ϕ are as desired. □

Theorem 2 (L&S, rephrased)

The following are equivalent:

1. *T is unsound.*
2. *$\neg\mathrm{Pr}_T(x)$ has a false fixed point over T.*

Theorem B (B&B)

The following are equivalent:

1. *T is unsound.*
2. *Every formula has a false fixed point over T.*

Proof.

$1 \Rightarrow 2$: Suppose that T is unsound, and let ψ be a false but T-provable sentence. Let $\theta(x)$ be any formula and let ϕ be such that $\mathrm{PA} \vdash \phi \leftrightarrow \theta(\phi) \wedge \psi$. Since $T \vdash \psi$ and $T \vdash \mathrm{PA}$, $T \vdash \phi \leftrightarrow \theta(\phi)$. Since $\mathrm{PA} \vdash \phi \rightarrow \psi$ and ψ is false, ϕ is also false.

$2 \Rightarrow 1$: Suppose that every formula has a false fixed point over T. The formula $x = x$ has a false fixed point ψ over T. But $T \vdash \psi = \psi$, so $T \vdash \psi$ and T is unsound. □

True and false Gödelian sentences

Corollary 3 (L&S)

Any unsound theory U has both true and false Gödelian sentences: There are sentences δ, γ such that

- $U \vdash \delta \leftrightarrow \neg \mathrm{Pr}_U(\delta)$,
- $U \vdash \gamma \leftrightarrow \neg \mathrm{Pr}_U(\gamma)$, *and*
- $U \vdash \delta \leftrightarrow \gamma$, *but*
- δ *is false, and* γ *is true.*

Proof.

- We get γ by constructing a fixed point of $\neg \mathrm{Pr}_U(x)$ over PA.
- Theorem B guarantees the existence of a false fixed point of $\neg \mathrm{Pr}_U(x)$ over U.
- The U-provable equivalence of δ and γ follows from Gödel's 2nd, since both sentences are U-provably equivalent to Con_U.

□

Diagnosis

- If δ is false and γ is true, of course $\delta \leftrightarrow \gamma$ is false.
- By Gödel's 1st, $U \nvdash \delta$. So $\neg\mathrm{Pr}_U(\delta)$ is true. It follows that $\delta \leftrightarrow \mathrm{Pr}_U(\delta)$ is true. This means that $\delta \leftrightarrow \neg\mathrm{Pr}_U(\delta)$ is false, even though it is provable in U.
- Similarly, $U \nvdash \gamma$ and $\neg\mathrm{Pr}_U(\delta)$ is true, but, by contrast, γ is true, so $\gamma \leftrightarrow \neg\mathrm{Pr}_U(\gamma)$ is true.
- Con_U is true, but δ is false.
- So U is wrong about many things: it proves $\delta \leftrightarrow \neg\mathrm{Pr}_U(\delta)$, $\delta \leftrightarrow \mathrm{Con}_U$, and $\delta \leftrightarrow \gamma$, even though all of these equivalences are false.
- PA, on the other hand, is sound. It does not prove any of these equivalences.
- Hence the fixed points of $\neg\mathrm{Pr}_U(x)$ over U are not the same as the ones over PA.
- This seems to be an instance of a more general phenomenon.

The importance of separating the two coordinates

Observation: A formula may have very different collections of fixed points over different theories.

Theorem (Löb)

The set of T-fixed points of $\mathrm{Pr}_T(x)$ is equal to $\mathrm{Th}(T)$.

Theorem C (B&B)

If S is a proper sub- or supertheory of T, then there is no formula $\theta(x)$ such that the set of S-fixed points of $\theta(x)$ is equal to $\mathrm{Th}(T)$.

Proof.

Suppose $\mathrm{Th}(T) \subsetneq \mathrm{Th}(S)$, and that $\theta(x)$ is a formula whose set of S-fixed points equals $\mathrm{Th}(T)$. Let $\psi \in \mathrm{Th}(S) \setminus \mathrm{Th}(T)$, and let χ be such that $\mathrm{PA} \vdash \chi \leftrightarrow \theta(\psi \wedge \chi)$. Since $S \vdash \psi$, it follows that $\psi \wedge \chi$ is a fixed point of $\theta(x)$ over S. By the assumption, $T \vdash \psi \wedge \chi$. Then $T \vdash \psi$, a contradiction.

The other case is similar: Let $\psi \in \mathrm{Th}(T) \setminus \mathrm{Th}(S)$, and let χ be such that $\mathrm{PA} \vdash \chi \leftrightarrow \neg\theta(\psi \vee \chi)$. □

Separating the coordinates again

Theorem 2 (L&S)

The following are equivalent:

1. *T is sound.*
2. *For all ϕ: if $T \vdash \phi \leftrightarrow \neg\mathrm{Pr}_T(\phi)$, then ϕ is true.*

Theorem D (Cf. Lajevardi & Salehi, 2019)

A. *The following are equivalent:*
   - A1 *T is sound*
   - A2 *For all ϕ: if $\phi \leftrightarrow \neg\mathrm{Pr}_T(\phi)$ is true, then ϕ is true.*

B. *The following are equivalent:*
   - B1 *S is sound*
   - B2 *For all ϕ: if $S \vdash \phi \leftrightarrow \neg\mathrm{Pr}_T(\phi)$, then ϕ is true.*

Proof of Theorem D

Proof.

- A1 ⇒ A2: Suppose that T is sound, and that $\phi \leftrightarrow \neg \mathrm{Pr}_T(\phi)$ is true. If $T \vdash \phi$, then ϕ is true, so $\neg \mathrm{Pr}_T(\phi)$ is true. Hence $T \nvdash \phi$. Then $\neg \mathrm{Pr}_T(\phi)$ is true, and so is ϕ.
- A2 ⇒ A1: Argue for the contrapositive. Suppose that T is unsound. Let ϕ be any T-provable but false sentence. Then $\mathrm{Pr}_T(\phi)$ is true and ϕ is false, so $\phi \leftrightarrow \neg \mathrm{Pr}_T(\phi)$ is true.
- B1 ⇒ B2: Suppose that S is sound, and $S \vdash \phi \leftrightarrow \neg \mathrm{Pr}_T(\phi)$. If $T \vdash \phi$, then $\mathrm{PA} \vdash \mathrm{Pr}_T(\phi)$, so $T \vdash \neg\phi$. Hence $T \nvdash \phi$, so $\neg \mathrm{Pr}_T(\phi)$ is true, and ϕ is true by the soundness of S.
- B2 ⇒ B1: Argue for the contrapositive. Suppose that S is unsound. Theorem B guarantees the existence of a false sentence ϕ such that $S \vdash \phi \leftrightarrow \neg \mathrm{Pr}_T(\phi)$. □

The Gödel sentence

- Claim: *the* sentence constructed using the fixed point lemma to satisfy $\mathrm{PA} \vdash \gamma \leftrightarrow \neg\mathrm{Pr}_T(\gamma)$ is a Gödel sentence for T.
- The choice of Gödel numbering, axiomatisation of PA, binumeration of T, and the details of the fixed point lemma all affect which particular syntactic object we end up with.
- A particular syntactic object might be a fixed point of $\neg\mathrm{Pr}_T(x)$ under some of these choices but not under others.
- So it really only makes sense to speak of a Gödel sentence of T relative to these technicalities.
- But: Given them, γ is surely a Gödel sentence.
- What warrants the *the* talk is that any sentence ϕ satisfying $\mathrm{PA} \vdash \phi \leftrightarrow \neg\mathrm{Pr}_T(\phi)$ also satisfies $\mathrm{PA} \vdash \phi \leftrightarrow \mathrm{Con}_T$, and that both of these equivalences are true. And so is ϕ.
- So, may we not "divide out" the insignificant properties of ϕ by closing under provable equivalence in PA, and speak of *the* Gödel sentence of T *over* PA?
- This makes the notion of the Gödel sentence dependent also on the choice of base theory.

Conclusion

- Separate the two coordinates in expressions like $S \vdash \delta \leftrightarrow \neg\mathrm{Pr}_T(\delta)$.
- It is sometimes important over which theory something is a fixed point.
- Construct your fixed points over a sound base theory.
- The notions of Gödelian sentences and Gödel sentences should not be equated: not every U-fixed point of $\mathrm{Pr}_U(x)$ is a Gödel sentence.
- ...since Gödel sentences are true?

Thank you!

References

Bennet, C. and Blanck, R. (2022?). Never trust an unsound theory. *Theoria*. Accepted for publication.

Lajevardi, K. and Salehi, S. (2019). On the arithmetical truth of self-referential sentences. *Theoria*, 85(1):8–17.

Lajevardi, K. and Salehi, S. (2021). There May Be Many Arithmetical Gödel Sentences. *Philosophia Mathematica*, 29(2):278–287.