Diagonal-Free Proofs of the Diagonal Lemma

Saeed Salehi

University of Tabriz & IPM

WORMSHOP 2017, Moscow

What is the Diagonal Lemma?

For any formula $\Psi(x)$ there exists some sentence η such that

$$\mathbb{N} \vDash \Psi(\bar{\eta}) \leftrightarrow \eta.$$

(The Semantic Diagonal Lemma)

This is usually provable in a Σ_1-complete theory:

$$\mathrm{T} \vdash \Psi(\bar{\eta}) \leftrightarrow \eta.$$

(The Syntactic Diagonal Lemma)

What is the Diagonal Lemma good for?

(E.G.) For Proving the Following Theorems:

- Gödel's (1st & 2nd) Incompleteness Theorem(s);
- Gödel–Rosser's Incompleteness Theorem;
- Tarski's Undefinability (of Truth) Theorem;
- Löb's Theorem; $\mathsf{T} \vdash \Box_{\mathsf{T}}(\Box_{\mathsf{T}}\mathcal{A} \to \mathcal{A}) \longrightarrow \Box_{\mathsf{T}}\mathcal{A}$.

What is wrong with the Diagonal Lemma?

Does anybody remember its proof? What about the sketch?
Even after (so) many years of teaching the lemma?

- Samuel Buss, *Handbook of Proof Theory* (Elsevier 1998, p. 119): "[Its] proof [is] quite simple but rather tricky and difficult to conceptualize."
- György Serény, The Diagonal Lemma as the Formalized Grelling Paradox, in: *Gödel Centenary* 2006 (Eds.: M. Baaz & N. Preining), Collegium Logicum vol. 9, Kurt Gödel Society, Vienna, 2006, pp. 63–66. *https://arxiv.org/pdf/math/0606425.pdf*
  http://math.bme.hu/∼sereny/poster.pdf
- Wayne Urban Wasserman, It ***Is*** "Pulling a Rabbit Out of the Hat": Typical Diagonal Lemma "Proofs" Beg the Question, (*Social Science Research Network*) SSRN (2008).
  http://dx.doi.org/10.2139/ssrn.1129038

What is really wrong with the (proof of the) Diag. Lem.?

Vann McGee (2002) *http://web.mit.edu/24.242/www/1stincompleteness.pdf*

"The following result is a cornerstone of modern logic:

> **Self-referential Lemma.** For any formula $\Psi(x)$, there is a sentence ϕ such that $\phi \leftrightarrow \Psi[\ulcorner \phi \urcorner]$ is a consequence of Q.

Proof: You would hope that such a deep theorem would have an insightful proof. No such luck. I am going to write down a sentence ϕ and verify that it works. What I won't do is give you a satisfactory explanation for why I write down the particular formula I do. I write down the formula because Gödel wrote down the formula, and Gödel wrote down the formula because, when he played the logic game he was able to see seven or eight moves ahead, whereas you and I are only able to see one or two moves ahead. I don't know anyone who thinks he has a fully satisfying understanding of why the Self-referential Lemma works. It has a rabbit-out-of-a-hat quality for everyone."

The Problem of Eliminating the Diagonal Lemma!?

- Henryk Kotlarski, The Incompleteness Theorems After 70 Years, *APAL* 126:1-3 (2004) 125–138.

The Diagonal Lemma, "being very intuitive in the natural language, is highly unintuitive in formal theories like Peano arithmetic. In fact, the usual proof of the diagonal lemma ... is short, but tricky and difficult to conceptualize. The problem was to eliminate this lemma from proofs of Gödel's result. This was achieved only in the 1990s".

Chaitin (1971) — Boolos (1989) — ···

Diagonal–Free Proofs ...

Some "Diagonal–Free" Proof of Tarski's Theorem:

1. A. Robinson, On Languages Which Are Based On Nonstandard Arithmetic, Nagoya Mathematical Journal (1963).

2. H. Kotlarski, Other Proofs of Old Results, MLQ (1998).

3. G. Serény, Boolos-Style Proofs of Limitative Theorems, MLQ (2004).
   - Xavier Caicedo, Lecturas Matemáticas (1993) (seminar 1987).

4. R. Kossak, Undefinability of Truth and Nonstandard Models, APAL (2004).

Toward a Big Surprise

Tarski's Theorem (on the Undefinability of Truth) in $\mathbb{N}$

$$\neg\exists\Phi\ \forall\eta\ \mathbb{N}\vDash\Phi(\bar{\eta})\leftrightarrow\eta$$

is equivalent with

$$\forall\Phi\ \exists\eta\ \mathbb{N}\vDash\neg\big[\Phi(\bar{\eta})\leftrightarrow\eta\big]$$

or, by the propositional equivalence,

$$\neg(p\leftrightarrow q)\equiv(\neg p\leftrightarrow q)$$

with the Semantic Diagonal Lemma

$$\forall\Psi(=\neg\Phi)\ \exists\theta\ \mathbb{N}\vDash\Psi(\bar{\theta})\leftrightarrow\theta.$$

A Big Surprise

So, any diagonal–free proof of Tarski's Undefinability Theorem

$$\neg\exists\Phi\ \forall\eta\ \mathbb{N} \vDash \Phi(\bar{\eta}) \leftrightarrow \eta$$

gives us a diagonal–free proof of the Semantic Diagonal Lemma

$$\forall\Psi\ \exists\boldsymbol{\theta}\ \mathbb{N} \vDash \Psi(\bar{\boldsymbol{\theta}}) \leftrightarrow \boldsymbol{\theta}$$

by which one can prove (diagonal–freely)
the semantic version of Gödel's 1st Incompleteness Theorem

$$\forall\, \mathrm{T}\ \exists\boldsymbol{\gamma}\ \big(\mathbb{N} \models \mathrm{T} \in \mathsf{RE} \Longrightarrow \mathrm{T} \nvdash \boldsymbol{\gamma}, \neg\boldsymbol{\gamma}\big).$$

More Surprises

H. Kotlarski (*APAL 2004, MLQ 1998*) proves (diagonal–freely) that

> Let T be any theory in $\mathcal{L}_{\mathrm{PA}}$ containing PA. Assume that there exists a formula Φ such that for every sentence η, $\mathrm{T} \vdash \eta \equiv \Phi(\ulcorner \eta \urcorner)$. Then T is inconsistent.

That is to say that for any consistent $\mathrm{T} \supseteq \mathrm{PA}$,

$$\neg\exists\Phi\ \forall\eta\ \mathrm{T} \vdash \Phi(\bar{\eta}) \leftrightarrow \eta$$

$$\forall\Phi\ \exists\eta\ \mathrm{T} \nvdash \Phi(\bar{\eta}) \leftrightarrow \eta$$

$$\Psi = \neg\Phi : \ [\mathrm{T} \nvdash \Phi(\bar{\eta}) \leftrightarrow \eta] \iff \mathrm{T} + [\Psi(\bar{\eta}) \leftrightarrow \eta] \text{ is consistent.}$$

$$\forall\Psi\ \exists\theta \text{ s.t. } \mathrm{T} + [\Psi(\bar{\theta}) \leftrightarrow \theta] \text{ is consistent.}$$

The Weak Diagonal Lemma

Any Diagonal–Free Proof of Tarski's Theorem for a theory T gives such a proof for the following **Weak Diagonal Lemma**.

> For any consistent $\mathrm{T} \supseteq \mathrm{PA}$ and any formula $\Psi(x)$ there exists a sentence θ such that $\mathrm{T} + [\Psi(\bar{\theta}) \leftrightarrow \theta]$ is consistent.

This is weak since cannot prove Gödel's 1st Incompleteness Theorem (by the way of Gödel's own proof):

Even though, for any consistent $\mathrm{T} + [\neg \mathtt{Pr}_{\mathrm{T}}(\bar{\theta}) \leftrightarrow \theta]$ we have $\mathrm{T} \nvdash \theta$, we may not have $\mathrm{T} \nvdash \neg\theta$:

For $\theta = \bot$ we have the consistency of $[\neg \mathtt{Pr}_{\mathrm{T}}(\bot) \leftrightarrow \bot] \equiv \neg \mathrm{Con}(\mathrm{T})$ with T (by Gödel's 2nd) but $\mathrm{T} \vdash \neg\bot$ even when T is ω−consistent!

However, the Weak Diagonal Lemma can prove Rosser's Theorem:

The Weak Diagonal Lemma $\Longrightarrow$ Gödel–Rosser's Theorem

The following theory is consistent for some $\boldsymbol{\rho}$:

$$\mathrm{T} + [\forall x\big(\, \mathtt{Proof}_{\mathrm{T}}(x, \bar{\boldsymbol{\rho}}) \to \exists y < x\, \mathtt{Proof}_{\mathrm{T}}(y, \overline{\neg \boldsymbol{\rho}})\,\big) \longleftrightarrow \boldsymbol{\rho}].$$

Call it T'.

- If $\mathrm{T} \vdash \boldsymbol{\rho}$ then $\mathtt{Proof}_{\mathrm{T}}(k, \bar{\boldsymbol{\rho}})$ for some $k \in \mathbb{N}$ and so $\mathrm{T}' \vdash \exists y < \bar{k}\, \mathtt{Proof}_{\mathrm{T}}(y, \overline{\neg \boldsymbol{\rho}})$ which contradicts $\mathrm{T}' \vdash \neg \mathtt{Proof}_{\mathrm{T}}(\ell, \overline{\neg \boldsymbol{\rho}})$ for all $\ell \in \mathbb{N}$ (by $\mathrm{T} \nvdash \neg \boldsymbol{\rho}$).
- If $\mathrm{T} \vdash \neg \boldsymbol{\rho}$ then $\mathtt{Proof}_{\mathrm{T}}(k, \overline{\neg \boldsymbol{\rho}})$ for some $k \in \mathbb{N}$. Also, $\mathrm{T}' \vdash \exists a$ such that $\mathtt{Proof}_{\mathrm{T}}(a, \bar{\boldsymbol{\rho}})$ and $\forall y < a\, \neg \mathtt{Proof}_{\mathrm{T}}(y, \overline{\neg \boldsymbol{\rho}})$. Thus, $k < a$ is impossible, so $a \leqslant k$ whence $a \in \mathbb{N}$. This contradicts $\mathrm{T}' \vdash \neg \mathtt{Proof}_{\mathrm{T}}(\ell, \bar{\boldsymbol{\rho}})$ for all $\ell \in \mathbb{N}$ (by $\mathrm{T} \nvdash \boldsymbol{\rho}$).

QED

The Weak Diagonal Lemma ¿⟹? Löb's Theorem ?

Does the Weak Diagonal Lemma imply Löb's Theorem?

$$\mathrm{T} \vdash \mathrm{Pr}_{\mathrm{T}}\big(\mathrm{Pr}_{\mathrm{T}}(\varphi) \to \varphi\big) \longrightarrow \mathrm{Pr}_{\mathrm{T}}(\varphi)$$

Only One Proof!?

Is There Any Diagonal–Free Proof For Löb's Theorem?

Is There Any Other Proof For Löb's Theorem?

Löb's Theorem $\Longrightarrow$ Gödel's 2nd Theorem

Löb's Theorem $\Longleftrightarrow$ Formalized Gödel's 2nd Theorem

$$\mathrm{Pr}_{\mathrm{T}}\big(\mathrm{Pr}_{\mathrm{T}}(\varphi) \to \varphi\big) \longrightarrow \mathrm{Pr}_{\mathrm{T}}(\varphi)$$
$$\neg\mathrm{Pr}_{\mathrm{T}}(\varphi) \longrightarrow \neg\mathrm{Pr}_{\mathrm{T}}\big(\neg\varphi \to \neg\mathrm{Pr}_{\mathrm{T}}(\varphi)\big)$$
$$\mathrm{Con}(\mathrm{T}+\neg\varphi) \longrightarrow \neg\mathrm{Pr}_{\mathrm{T}+\neg\varphi}\big(\mathrm{Con}(\mathrm{T}+\neg\varphi)\big)$$

for $\xi = \neg\varphi$

$$\mathrm{Con}(\mathrm{T}+\xi) \longrightarrow \neg\mathrm{Pr}_{\mathrm{T}+\xi}\big[\mathrm{Con}(\mathrm{T}+\xi)\big]$$

Thus Far ...

The Equivalences and The Implications:

Semantic Diagonal Lemma $\iff$ Semantic Tarski's Theorem
$\implies$ Semantic Gödel's 1st Theorem

Weak Diagonal Lemma $\iff$ Syntactic Tarski's Theorem
$\implies$ Gödel–Rosser's Theorem
$\implies$ 1st Incompleteness Theorem

Löb's Theorem $\iff$ Formalized Gödel's 2nd Theorem
$\implies$ 2nd Incompleteness Theorem

Diagonal–Free Proofs for Gödel's 2nd Theorem

1. T. Jech, On Gödel's Second Incompleteness Theorem, Proc. AMS (1994).
2. H. Kotlarski, On the Incompleteness Theorems, JSL (1994).
3. M. Kikuchi, A Note on Boolos' Proof of the Incompleteness Theorem, MLQ (1994).
4. M. Kikuchi, Kolmogorov Complexity and the Second Incompleteness Theorem, Arch. Math. Logic (1997).
5. H. Kotlarski, Other Proofs of Old Results, MLQ (1998).
6. Z. Adamowicz & T. Bigorajska, Existentially Closed Structures and Gödel's Second Incompleteness Theorem, JSL (2001).
7. G. Serény, Boolos-Style Proofs of Limitative Theorems, MLQ (2004).
8. H. Kotlarski, The Incompleteness Theorems After 70 Years, APAL (2004).

Diagonal–Free Proofs for Tarski's Theorem. I

A. Robinsion (1963):

If $\mathbf{\Phi}$ defined truth [in $\mathbb{N}$](in T $\supseteq$ PA) then let $\mathfrak{M}$ be a non-standard model [$\equiv \mathbb{N}$](of T) with $\mathbb{N} < a \in \mathfrak{M}$. Put

$$\mathfrak{M}' = \{t_i(a) \mid t_i \text{ is an } \mathfrak{M}\text{–Skolem term}, i \in \mathbb{N}\} \quad (\preccurlyeq \mathfrak{M}).$$

For any $n \in \mathbb{N}$ we have $\mathfrak{M}' \models \exists x \bigwedge_{i<n} x \neq t_i(a)$. So,

$$\mathfrak{M}' \models \exists x \forall y < n \; \mathbf{\Phi}(\ulcorner x \neq t_y(a) \urcorner).$$

By overspill there is some $b > \mathbb{N}$ in $\mathfrak{M}'$ such that

$$\mathfrak{M}' \models \exists x \forall y < b \; \mathbf{\Phi}(\ulcorner x \neq t_y(a) \urcorner).$$

Thus for some $c \in \mathfrak{M}'$ we have $\bigwedge_{j \in \mathbb{N}} : \mathfrak{M}' \models c \neq t_j(a)$;
a contradiction! QED

Diagonal–Free Proofs for Tarski's Theorem. II

H. Kotlarski (1998):

If $\mathbf{\Phi}$ defined truth [in $\mathbb{N}$](in T $\supseteq$ PA) then let

$$F(x) = \min y : \forall \bar{\varphi}, u \leqslant x [\exists v\, \mathbf{\Phi}(\ulcorner \varphi(u,v) \urcorner) \rightarrow \exists v < y\, \mathbf{\Phi}(\ulcorner \varphi(u,v) \urcorner)].$$

The unary function F is [$\mathbb{N}-$](T$-$)Definable,
but `Dominates` all the unary definable functions:
If f is definable by $\varphi(u,v)$ then for any $z > \ulcorner \varphi \urcorner$ we have $F(z) > f(z)$;
a contradiction! QED

Diagonal–Free Proofs for Tarski's Theorem. III

G. Serény (2004):

If $\mathbf{\Phi}$ defined truth [in $\mathbb{N}$](in T $\supseteq$ PA) then let

$$\mathtt{Def}^{<z}(y) = \exists\varphi\big[\ \|\varphi\| < z \wedge \mathbf{\Phi}(\ulcorner\forall\zeta[\varphi(\zeta) \leftrightarrow \zeta = y]\urcorner)\ \big].$$

where $\|\varphi\|$ is a measure of φ such that $\forall n \in \mathbb{N}$ there are finitely many ϕ with $\|\phi\| < n$

$$\mathtt{Berry}^{<v}(u) = \neg\mathtt{Def}^{<v}(u) \ \wedge\ \forall w < u\, \mathtt{Def}^{<v}(w).$$

$$\mathtt{Boolos}(x) = \mathtt{Berry}^{<5\ell}(x), \text{ where } \ell = \|\mathtt{Berry}^{<y}(x)\|.$$

$$\mathfrak{b} = \min z \neg\mathtt{Def}^{<5\ell}(z).$$

Now we have $\|\mathtt{Boolos}(x)\| < 5\ell$ and also

[$\mathbb{N} \models$](T $\vdash$) $\quad \mathtt{Boolos}(\zeta) \leftrightarrow \zeta = \mathfrak{b}$;

a contradiction! QED

Xavier Caicedo, "La Paradoja de Berry, o la Indefinibilidad de la Definibilidad y las Limitaciones de los Formalismos", *Lecturas Matemáticas*, (1993) 14:37–48. Presented in "el Seminario de Lógica de la Universidad Nacional de Colombia" (1987). Revised in 2004 at *http://goo.gl/yYnstW*.

More Equivalences ...

Tarski's Semantic Theorem:

$$\{\ulcorner\varrho\urcorner \in \mathbb{N} \mid \mathbb{N} \models \varrho\} = \mathrm{Th}(\mathbb{N}) \notin \mathtt{Def}(\mathbb{N}) = \{X \subseteq \mathbb{N} \mid \exists \psi\colon\ n \in X \leftrightarrow \psi(n)\}$$

$$\forall T \subseteq \mathrm{Th}(\mathbb{N}) \big[T \in \mathtt{Def}(\mathbb{N}) \Longrightarrow T \neq \mathrm{Th}(\mathbb{N}) \big]$$

$$\forall T \big[\mathbb{N} \models T \in \mathtt{Def}(\mathbb{N}) \Longrightarrow T \text{ is incomplete.} \big]$$

No Sound and Definable Theory is Complete!

Gödel–Smullyan Incompleteness Theorem

$$\vdots$$

Thank You!

Thanks to

The Participants . . . . . For Listening · · ·

and

The Organizers — For Taking Care of Everything · · ·