Löb's Theorem in Coq: Taking stock of where we are

Janis Bailitis
Bachelor Project, Saarland University
Advisors: Dr. Yannick Forster and Dr. Dominik Kirst
Supervisor: Prof. Dr. Gert Smolka

22nd April 2024

This is a brief memo providing an overview of the current state of my Bachelor project, focussing on the goals achieved so far, and how the results relate to the work previously done at the Programming Systems Lab. Further, we aim to point out how our approach fits into the broader landscape of (provability) logic.

1 Preliminaries

At first, let us fix a notational convention: If we have a formula φ and write $\varphi(x_1, \ldots, x_n)$, we indicate that φ has n free variables, unless stated otherwise.

We work in constructive type theory and build upon some key results obtained from formalising logic in type theory. We rely on results established by Kirst et al. [HK23],[KH23],[KP23]. Most importantly, we build upon a variant of Church's thesis for Robinson arithmetic, stating that any function $f : \mathbb{N} \to \mathbb{N}$ can be represented by a Σ_1 formula in Q. More formally, we assume:

Axiom 1.1 (CT_{Q}, cf. [HK23]) For all $f : \mathbb{N} \to \mathbb{N}$ there exists a Σ_1-formula $\varphi(x, y)$ such that for all $x : \mathbb{N}$, $\mathrm{Q} \vdash \dot{\forall} y.\varphi(\overline{x}, y) \dot{\leftrightarrow} y \equiv \overline{fx}$.

Later, this axiom was refined to account for partial functions $\mathbb{N} \rightharpoonup \mathbb{N}$. The initial variant of CT_{Q} can be derived from this statement.

Axiom 1.2 ((partial) CT_{Q}, cf. [KP23]) For all $f : \mathbb{N} \rightharpoonup \mathbb{N}$ there exists a Σ_1-formula $\varphi(x, y)$ such that for all $x, y : \mathbb{N}$, $fx \downarrow y \leftrightarrow \mathrm{Q} \vdash \dot{\forall} z.\varphi(\overline{x}, z) \dot{\leftrightarrow} z \equiv \overline{y}$.

Kirst and Peters [KP23, section 5] use the term CT_{Q} to refer to Axiom 1.2, while we will refer to Axiom 1.1 by CT_{Q} as we do not use partial CT_{Q} directly. Using CT_{Q}, we can derive the **Weak Representability Theorem** playing a pivotal role in our project:

Fact 1.3 (Weak Representability Theorem) Suppose $P : \mathbb{N} \to \mathbb{P}$ is an enumerable predicate. We can find a Σ_1-formula $\varphi(x)$ such that for all $x : \mathbb{N}$, $Pn \leftrightarrow \mathrm{Q} \vdash \varphi(\overline{x})$.

Another result following from partial CT_{Q} we will use is the **Strong Separation Theorem**:

Fact 1.4 (Strong Separation Theorem) Let $P, R : \mathbb{N} \to \mathbb{P}$ be disjoint, semi-decidable predicates. There exists a Σ_1-formula $\varphi(x)$ satisfying

$$\forall n.Pn \to \mathrm{Q} \vdash \varphi(\overline{n}) \wedge Rn \to \mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n})$$

This gives rise to the following fact for decidable predicates:

Corollary 1.5 Let $P : \mathbb{N} \to \mathbb{P}$ be a decidable predicate. There exists a Σ_1-formula $\varphi(x)$ such that

$$\forall n.Pn \leftrightarrow \mathrm{Q} \vdash \varphi(\overline{n}) \wedge \neg Pn \leftrightarrow \mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n}).$$

Proof Since P is decidable, both P and $\lambda n.\neg Pn$ are semi-decidable. Further, both are clearly disjoint. By the previous fact, we obtain a Σ_1-formula $\varphi(x)$ such that

$$\forall n.Pn \to \mathrm{Q} \vdash \varphi(\overline{n}) \wedge \neg Pn \to \mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n}).$$

For each of the the converses $\mathrm{Q} \vdash \varphi(\overline{n}) \to Pn$ and $\mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n}) \to \neg Pn$, we decide Pn. If the decision matches the conclusion of the implications, we are done. Otherwise, for the first converse, we know that $\neg Pn$ as well as $\mathrm{Q} \vdash \varphi(\overline{n})$. Exploiting $\neg Pn \to \mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n})$ gives us $\mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n})$. Contradictory as Q is consistent using the standard model. Similarly, one handles the case Pn when proving $\mathrm{Q} \vdash \dot{\neg}\varphi(\overline{n}) \to \neg Pn$. ∎

Equipped with these notions, we started our exploration of Löb's theorem.

2 Löb's Theorem

The initial goal was to show Löb's Theorem, a result of provability logic stating that when we prove a formula in a sufficiently strong formal system, we can assume provability of this formula without additional cost. More formally, there is a modality, the box-modality (written $\Box$), expressing provability of a formula. Löb's theorem can then be stated as the following rule, where $\mathcal{T}$ is an arbitrary theory:

Theorem 2.1 (Löb) Suppose φ is a sentence. In order to prove $\mathcal{T} \vdash \varphi$, it suffices to show $\mathcal{T} \vdash \Box\varphi \dot{\to} \varphi$.

Three questions arose: Firstly, how can $\Box$ be defined? Secondly – given a sensible definition of $\Box$ – how can Löb's Theorem be shown? Lastly, which theory should we choose? We assumed that Peano arithmetic suffices, so we raised the question whether Robinson arithmetic is still enough.

The second question admits a fortunate answer. As long as $\Box$ obeys four axioms (pointed out below), Löb's theorem can be derived. In particular, we can abstract away concrete implementation details of the modality's definition. The required axioms are as follows:

1. Necessitation: For all $\varphi : \mathbb{F}$, $\mathcal{T} \vdash \varphi$ implies $\mathcal{T} \vdash \Box\varphi$.
2. Modal fixed points: Whenever φ is a sentence, we can construct another sentence ψ such that $\mathrm{Q} \vdash \psi \dot{\leftrightarrow} (\Box\psi \dot{\to} \varphi)$.
3. Internal necessitation: For all $\varphi : \mathbb{F}$, we can derive $\mathcal{T} \vdash \Box\varphi \dot{\to} \Box\Box\varphi$.
4. Box distributivity: For all $\varphi, \psi : \mathbb{F}$, we have $\mathcal{T} \vdash \Box(\varphi \dot{\to} \psi) \dot{\to} \Box\varphi \dot{\to} \Box\psi$.

For any operation $\Box$ on formulas and theories $\mathcal{T}$ obeying these rules, Löb's theorem can be derived; such an argument has been done by Boolos et al. [BBJ07, theorem 18.4], with a slightly different handling of modal fixed points. A first version of this proof is formalised in Coq in the file `coq/Loeb_formalisation_fol.v`. As of now, the formalisation needs some minor adjustments to account for the current formulation of our axioms, but it is a purely modal-logical argument not using any internals of formulas and deduction systems (apart from the inference rules).

3 Defining the Box modality

In search of a definition for $\Box$, we wanted to obtain a formula $\mathsf{Prov}(x)$ characterising provability in our deduction system. At first, we worked in Robinson arithmetic. Intuitively, we asked for the property $\forall \tau : \mathbb{F}.\mathsf{Q} \vdash \tau \leftrightarrow \mathsf{Q} \vdash \mathsf{Prov}(\tau)$. By defining $\Box\tau := \mathsf{Prov}(\tau)$, we would have a candidate for the Box modality. Unfortunately, $\mathsf{Prov}(\tau)$ is not syntactically correct, as we may only substitute in terms, but not entire formulas. Gödel numbering is the road to success in resolving this issue:

Fact 3.1 (Gödel numberings) There exist functions $\mathsf{göd} : \mathbb{F} \to \mathbb{N}$ and $\mathsf{göd}^{-1} : \mathbb{N} \to \mathbb{F}$ inverting each other, i.e. $\forall n.\mathsf{göd}(\mathsf{göd}^{-1}(n)) = n$ and $\forall \varphi.\mathsf{göd}^{-1}(\mathsf{göd}(\varphi)) = \varphi$. We call $\mathsf{göd}(\varphi)$ the **Gödel number** of φ.

Remark 3.2 In the Coq mechanisation, we consistently use `g` for göd and `f` for $\mathsf{göd}^{-1}$. In the future, we will work with coercions to allow for a more readable Coq development.

Definition 3.3 For a formula φ, we define $\ulcorner \varphi \urcorner := \overline{\mathsf{göd}(\varphi)}$. This is the **Gödel numeral** or **quine quote** of φ.

The term 'quine quote' was used by Norrish [Nor18].

Having this machinery set up, we can derive the Box modality.

Lemma 3.4 There exists a Σ_1-formula $\mathsf{Prov}(x)$ satisfying the following equivalence: $\forall \tau.\mathsf{Q} \vdash \tau \leftrightarrow \mathsf{Q} \vdash \mathsf{Prov}(\ulcorner \tau \urcorner)$.

Proof We apply the weak representability theorem. The predicate $\lambda \varphi.\mathsf{Q} \vdash \varphi$ is enumerable by standard techniques, cf. [FKS19]. A routine argument concludes enumerability of $\lambda n.\mathsf{Q} \vdash \mathsf{göd}^{-1}(n)$. Applying this theorem, we obtain a Σ_1-formula $\mathsf{Prov}(x)$ such that for all $m : \mathbb{N}$

$$\mathsf{Q} \vdash \mathsf{göd}^{-1}(m) = (\lambda n.\mathsf{Q} \vdash \mathsf{göd}^{-1}(n))m \leftrightarrow \mathsf{Q} \vdash \mathsf{Prov}(\overline{m}).$$

If τ is any formula, we can plug in $\mathsf{göd}(\tau)$ for m and arrive at

$$\mathsf{Q} \vdash \tau = \mathsf{Q} \vdash \mathsf{göd}^{-1}(\mathsf{göd}(\tau)) \leftrightarrow \mathsf{Q} \vdash \varphi(\overline{\mathsf{göd}(\tau)}) = \mathsf{Q} \vdash \varphi(\ulcorner \tau \urcorner)$$

and are done. ∎

Finally, we can define the Box modality.

Definition 3.5 (Box modality) For any formula τ, we set $\Box\tau := \mathsf{Prov}(\ulcorner \tau \urcorner)$.

Immediately, we obtain the necessitation rule.

Fact 3.6 (Necessitation for Box) For all formulas τ, we have that $\mathsf{Q} \vdash \tau$ implies $\mathsf{Q} \vdash \Box\tau$.

Proof Unfold the definition of $\Box$. The remaining goal is a direct consequence of Lemma 3.4. ∎

Next, we were interested in whether this definition is strong enough for the remaining three properties our modality should have. Kirst noted that we may need to switch to Peano arithmetic for Box distributivity and internal necessitation, while the existence of modal fixed points may be derivable from Q alone, which was our subsequent aim. Kirst suspected that, along the way, we need Gödel's diagonal lemma, hence we inspected this result in depth.

4 The diagonal lemma

In computability theory, we are often interested in a program's behaviour on its own Gödel number. In a similar fashion, we can analyse a formula's truth or provability when its own Gödel numeral is substituted in for the free variables. More specifically, we are interested in substituting in a formula's Gödel numeral for one of the free variables.

Definition 4.1 (Diagonalisation, cf. [Smi22]) Let $\varphi(x)$ be any formula (with potentially more than one free variable). We say that $\varphi(\ulcorner\varphi\urcorner)$ is the **diagonalisation** of φ. This gives rise to a function $\mathrm{diag} : \mathbb{F} \to \mathbb{F}, \varphi \mapsto \varphi(\ulcorner\varphi\urcorner)$.

We can define a similar function on the level of Gödel numbers:

Definition 4.2 $\mathrm{diag}_{\mathbb{N}} : \mathbb{N} \to \mathbb{N}, n \mapsto \mathrm{göd}(\mathrm{diag}(\mathrm{göd}^{-1}(n)))$ sends a formula's Gödel number to the Gödel number of its diagonalisation.

Note that CT_{Q} applies to $\mathrm{diag}_{\mathbb{N}}$; this is why we introduced this function at all.

We have a straightforward interchangeability result.

Lemma 4.3 Let φ be a formula. Then $\mathrm{diag}_{\mathbb{N}}(\mathrm{göd}(\varphi)) = \mathrm{göd}(\mathrm{diag}(\varphi))$.

Proof Immediate from the definitions and the identity $\mathrm{göd}^{-1}(\mathrm{göd}(\varphi)) = \varphi$. ∎

Ultimately, our aim is to show the following result, the well known **diagonal lemma**.

Theorem 4.4 (Diagonal lemma for Q) Let $B(x)$ be a formula. There exists a sentence G satisfying $\mathsf{Q} \vdash G \dot{\leftrightarrow} B(\ulcorner G\urcorner)$.

Interestingly, diagonalisation is nowhere mentioned in the theorem statement. However, diagonal thinking plays a pivotal role in its proof. Before heading straight to showing the claim, let us discuss a preliminary result making our argument easier to digest.

Lemma 4.5 There exists a Σ_1-formula $\mathrm{dg}(x, y)$ such that for all $\psi : \mathbb{F}$, $\mathsf{Q} \vdash \dot{\forall} y.\mathrm{dg}(\ulcorner\psi\urcorner, y) \dot{\leftrightarrow} y \equiv \ulcorner\mathrm{diag}(\psi)\urcorner$.

Proof With CT_{Q}, we find a Σ_1-formula $\mathrm{dg}(x, y)$ satisfying, for all $x : \mathbb{N}$,

$$\mathsf{Q} \vdash \dot{\forall} y.\mathrm{dg}(\overline{x}, y) \dot{\leftrightarrow} y \equiv \overline{\mathrm{diag}_{\mathbb{N}}(x)}.$$

Suppose $\psi : \mathbb{F}$. We obtain, by setting $x = \mathrm{göd}(\psi)$,

$$\mathsf{Q} \vdash \dot{\forall} y.\mathrm{dg}(\ulcorner\psi\urcorner, y) \dot{\leftrightarrow} y \equiv \overline{\mathrm{diag}_{\mathbb{N}}(\mathrm{göd}(\psi))}.$$

We finish by rewriting with Lemma 4.3. ∎

Having all this in mind, we can discuss the proof of Theorem 4.4. Apart from getting the substitutions right, over which we abstract here, the argument itself is not very involved provided that we instantiate the existential quantifier with a sensible candidate. Indeed, finding this formula is a bit delicate and hence the true art in showing this claim. The proof idea is from [BBJ07].

Proof (of Theorem 4.4) First, we define G. We inspect the formula $F(x) := \dot{\exists} y.\mathrm{dg}(x, y) \dot{\wedge} B(y)$. We set $G := \mathrm{diag}(F) = F(\ulcorner F \urcorner)$. Clearly, G is closed.

We are left to show $\mathsf{Q} \vdash G \dot{\leftrightarrow} B(\ulcorner G \urcorner)$. After unfolding the definitions of G and F, we arrive at

$$\mathsf{Q} \vdash (\dot{\exists} y.\mathrm{dg}(\ulcorner F \urcorner, y) \dot{\wedge} B(y)) \dot{\leftrightarrow} B(\ulcorner G \urcorner).$$

The direction from left to right is straightforward: We introduce our assumption, and destruct the existential as well as the conjunction. That is, we have, for some term y,

$$\mathsf{Q}, \mathrm{dg}(\ulcorner F \urcorner, y), B(y) \vdash B(\ulcorner G \urcorner).$$

We apply Lemma 4.5 to our assumption $\mathrm{dg}(\ulcorner F \urcorner, y)$ and obtain the proof state

$$\mathsf{Q}, \mathrm{dg}(\ulcorner F \urcorner, y), B(y), y \equiv \ulcorner \mathrm{diag}(F) \urcorner \vdash B(\ulcorner G \urcorner).$$

Since $G = \mathrm{diag}(F)$, we finish by rewriting on the meta level followed by a rewrite with $y \equiv \ulcorner G \urcorner$ in the deduction system.

For the converse, we have to show

$$\mathsf{Q}, B(\ulcorner G \urcorner) \vdash \dot{\exists} y.\mathrm{dg}(\ulcorner F \urcorner, y) \dot{\wedge} B(y).$$

Instantiating the existential with $\ulcorner G \urcorner$ leaves us to prove

$$\mathsf{Q}, B(\ulcorner G \urcorner) \vdash \mathrm{dg}(\ulcorner F \urcorner, \ulcorner G \urcorner) \dot{\wedge} B(\ulcorner G \urcorner).$$

By the assumption and conjunction rules as well as weakening, we are left to show

$$\mathsf{Q} \vdash \mathrm{dg}(\ulcorner F \urcorner, \ulcorner G \urcorner)$$

following from Lemma 4.5 as $G = \mathrm{diag}(F)$. ∎

5 Applications of the diagonal lemma

Having established the diagonal lemma, we can draw our attention to some important corollaries, including the existence of modal fixed points, contributing to the proof of Löb's theorem.

5.1 Modal fixed points

We will now clarify how the diagonal lemma contributes to the fixed point theorem. The high-level idea is, given a sentence φ, to apply the diagonal lemma to the formula $\mathrm{Prov}(x) \dot{\rightarrow} \varphi$. The result will then satisfy our requirements. The following proof accounts for all technical details.

Theorem 5.1 (Modal fixed points) For all sentences φ, there exists a sentence ψ such that $\mathsf{Q} \vdash \psi \dot{\leftrightarrow} (\Box\psi \dot{\rightarrow} \varphi)$

Proof We use the diagonal lemma to obtain a sentence ψ satisfying $\mathsf{Q} \vdash \psi \dot{\leftrightarrow} (\mathrm{Prov}(\ulcorner \psi \urcorner) \dot{\rightarrow} \varphi)$. Applying the definition of Box gives us our goal $\mathsf{Q} \vdash \psi \dot{\leftrightarrow} (\Box\psi \dot{\rightarrow} \varphi)$. ∎

5.2 Tarski's indefinability result

In this and the subsequent section, we explore certain limits of our deductive system following from the diagonal lemma. From CT_{Q}, we concluded capability of Q to represent type-theoretic functions. In a similar fashion, we can ask whether there is a notion to 'represent' predicates $\mathbb{N} \to \mathbb{P}$ inside the deduction system. Indeed, there is, and one talks of a predicate being **definable**. The definition is taken from Boolos et al. [BBJ07].

Definition 5.2 (Definable) Let $P : \mathbb{N} \to \mathbb{P}$ be a predicate and $\mathcal{T}$ a theory. The formula $\varphi(x)$ **defines** P in $\mathcal{T}$ if, for all $n : \mathbb{N}$, we have both $Pn \to \mathcal{T} \vdash \varphi(\overline{n})$ and $\neg Pn \to \mathcal{T} \vdash \dot{\neg}\varphi(\overline{n})$.

Definition 5.3 (Definablility) A predicate $P : \mathbb{N} \to \mathbb{P}$ is said to be **definable** in a theory $\mathcal{T}$ if there exists a formula defining P in $\mathcal{T}$.

Remark 5.4 Every decidable predicate is definable in any extension of Q by Corollary 1.5 and weakening.

Next, we focussed on the predicate $\lambda n.\mathcal{T} \vdash \mathrm{göd}^{-1}(n)$ for any consistent extension $\mathcal{T}$ of Q. It turned out that it is not definable.

Lemma 5.5 (Indefinability lemma) Let $\mathcal{T}$ be a consistent theory extending Q. $\lambda n.\mathcal{T} \vdash \mathrm{göd}^{-1}(n)$ is not definable in $\mathcal{T}$.

Before proving this claim, we remind ourselves that we can use classical reasoning when proving distinguished claims, the so called **stable** claims (cf. [Smo24]):

Definition 5.6 Suppose $P : \mathbb{P}$. We say that P is **stable** if $\neg\neg P \to P$.

Fact 5.7 (Classical reasoning for stable claims) Suppose $P, Q : \mathbb{P}$.
We have $\mathrm{stable}(P) \to (Q \vee \neg Q \to P) \to P$.

We are now ready to show the indefinability lemma. The proof idea is due to Boolos et al. [BBJ07].

Proof (of Lemma 5.5) Suppose $P := \lambda n.\mathcal{T} \vdash \mathrm{göd}^{-1}(n)$ was definable witnessed by $\varphi(x)$. Using the diagonal lemma, we find a sentence G such that $\mathrm{Q} \vdash G \dot{\leftrightarrow} \dot{\neg}\varphi(\ulcorner G \urcorner)$.

We have to show $\bot$, which is stable by a straightforward argument. Using Fact 5.7, we assume that $\mathcal{T} \vdash G \vee \mathcal{T} \nvdash G$. Case analysis.

If $\mathcal{T} \vdash G$, there exists a list $A \subseteq \mathcal{T}$ witnessing this, i.e. $A \vdash G$. As $\varphi(x)$ defines P, and $P(\mathrm{göd}(G))$ holds, we conclude $\mathcal{T} \vdash \varphi(\ulcorner G \urcorner)$, i.e. $B \vdash \varphi(\ulcorner G \urcorner)$ for some list $B \subseteq \mathcal{T}$. By weakening, we obtain $\mathrm{Q}, A, B \vdash G \dot{\leftrightarrow} \dot{\neg}\varphi(\ulcorner G \urcorner)$, $\mathrm{Q}, A, B \vdash G$ as well as $\mathrm{Q}, A, B \vdash \varphi(\ulcorner G \urcorner)$. We apply $\mathcal{T}$'s consistency to our goal and have to show $\mathcal{T} \vdash \bot$. Straightforward from the assumptions, as Q, A, B is contained in $\mathcal{T}$.

If we obtain $\mathcal{T} \nvdash G$, we apply this assumption to our goal and have to show $\mathcal{T} \vdash G$. Now, since $\neg P(\mathrm{göd}(G))$ holds, we conclude $\mathcal{T} \vdash \dot{\neg}\varphi(\ulcorner G \urcorner)$, i.e. $A \vdash \dot{\neg}\varphi(\ulcorner G \urcorner)$ for some list $A \subseteq \mathcal{T}$. Again, by weakening, have $\mathrm{Q}, A \vdash G \dot{\leftrightarrow} \dot{\neg}\varphi(\ulcorner G \urcorner)$ as well as $\mathrm{Q}, A \vdash \dot{\neg}\varphi(\ulcorner G \urcorner)$. As our goal is $\mathcal{T} \vdash G$, we are done since Q, A is contained in $\mathcal{T}$. ∎

While this result is somewhat abstract, it has interesting consequences. One of which is **Tarski's Theorem** stating that the theory of sentences correct for $\mathbb{N}$ is not definable.[1]

[1] As of now, I am not sure how this 'correctness' is modelled in our set-up. In Boolos' et al. book, a formula is 'correct' if it is true in the his notion of standard interpretation, which most likely coincides with our [KH23, Definition 4] definition of $\vDash$ with $\mathbb{N}$ as domain type, picking the standard meta-level predicates for the predicate symbols. I am working on resolving this issue.

Corollary 5.8 (Tarski's Theorem.) The theory $\lambda\varphi.\mathbb{N} \vDash \varphi$ is a not definable.

Proof The theory in question is a consistent extension of Q and hence not definable by Lemma 5.5. ∎

Finally, we can conclude this section with an argument that for any consistent extension $\mathcal{T}$ of Q, it is not decidable whether a formula is derivable from $\mathcal{T}$ or not. Most prominently, we cannot decide whether a formula is a theorem of Q.

Corollary 5.9 Let $\mathcal{T}$ be a consistent extension of Q. The predicate $\lambda\varphi.\mathcal{T} \vdash \varphi$ is not decidable.

Proof Suppose, for contradiction, that $P := \lambda\varphi.\mathcal{T} \vdash \varphi$ was decidable. A routine argument derives decidability of $R := \lambda n.\mathcal{T} \vdash \text{göd}^{-1}(n)$. By Corollary 1.5, R is definable, whereas Lemma 5.5 asserts indefinability of R. Absurd. ∎

In essence, the results derived in this section indicate certain boundaries our deduction system has when it comes to interaction with the meta level – even if we allow for arbitrary consistent extensions of Q. This why these (and some more) theorems are also known as 'limitative theorems', for instance in book by Boolos et al. [BBJ07].

5.3 Gödel's first incompleteness theorem

Next, we will focus on another consequence of the diagonal lemma: The existence of a sentence that can neither be proven nor refuted in Q. That is, we will prove a variant of Gödel's first incompleteness result.

Theorem 5.10 (Gödel's first incompleteness theorem for Q) There exists a sentence G such that neither $Q \vdash G$ nor $Q \vdash \dot{\neg}G$.

Before heading to the proof, we need to draw our attention to intuitionistic and classical reasoning. As we want our results to be as general as possible, we do not stick to intuitionistic ($\vdash_i$) or classical ($\vdash_c$) reasoning in our theorem statements. Unfortunately, the current proof needs classical reasoning within the deduction system once, so we could only show that $Q \nvdash_c \dot{\neg}G$. While this immediately refutes any intuitionistic derivation due to classical reasoning being strictly stronger, the proof-mode had trouble destructing assumptions when in classical mode, so we needed to find a way circumventing this. We found such a way, and the details can be found in our proof. We would like to emphasise our reliance on the following result:

Fact 5.11 (Σ_1-conservativity) Let φ be a Σ_1-sentence such that $Q \vdash_c \varphi$. Then $Q \vdash \varphi$.

As opposed to the library, this is a slightly different yet derivable formulation.

For the proof of Gödel's result, we use Boolos' et al. version of Σ_1 stating that we can write $\text{Prov}(x) = \dot{\exists} y.\text{Prf}(x, y)$ for some Δ_1-formula $\text{Prf}(x, y)$. We elaborate on this step in the discussion.

Definition 5.12 (ω-(in)consistency) Let $\mathcal{T}$ be a theory. We say that $\mathcal{T}$ is **ω-inconsistent** if it proves $\mathcal{T} \vdash \dot{\exists} x.\tau(x)$ for formula some $\tau(x)$, but also admits a derivation of $\mathcal{T} \vdash \dot{\neg}\tau(\overline{n})$ for all $n : \mathbb{N}$.

A theory not being ω-inconsistent is said to be **ω-consistent**.

Axiom 5.13 Q is ω-consistent.

Proof (of Theorem 5.10) By the diagonal lemma, we can find a sentence G such that $\mathsf{Q} \vdash G \dot{\leftrightarrow} \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner)$. We show that $\mathsf{Q} \nvdash G$ as well as $\mathsf{Q} \nvdash \dot{\neg} G$.

Suppose $\mathsf{Q} \vdash G$. By Lemma 3.4, conclude $\mathsf{Q} \vdash \mathsf{Prov}(\ulcorner G \urcorner)$. Furthermore, we obtain $\mathsf{Q} \vdash \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner)$ from $\mathsf{Q} \vdash G \dot{\leftrightarrow} \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner)$. This is contradictory since Q is consistent using the standard model.

Now, assume $\mathsf{Q} \vdash \dot{\neg} G$. We show that Q is ω-inconsistent, contradicting Axiom 5.13. We pick $\mathsf{Prf}(\ulcorner G \urcorner, y)$ as witness for ω-inconsistency.

At first, we show that $\mathsf{Q} \vdash \dot{\exists} x.\mathsf{Prf}(\ulcorner G \urcorner, x)$. It suffices to verify $\mathsf{Q} \vdash \mathsf{Prov}(\ulcorner G \urcorner)$ by definition of Prf. By Fact 5.11, it is enough to show $\mathsf{Q} \vdash_c \mathsf{Prov}(\ulcorner G \urcorner)$. This fact is applicable since $\mathsf{Prov}(\ulcorner G \urcorner)$ is a Σ_1-sentence. By the contradiction rule, we are left to prove

$$\mathsf{Q}, \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner) \vdash_c \dot{\bot},$$

for which

$$\mathsf{Q}, \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner) \vdash \dot{\bot}$$

suffices. We apply our assumption $\mathsf{Q} \vdash \dot{\neg} G$ and end up in the proof state

$$\mathsf{Q}, \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner) \vdash G.$$

Due to $\mathsf{Q} \vdash G \dot{\leftrightarrow} \dot{\neg}\mathsf{Prov}(\ulcorner G \urcorner)$ obtained from the diagonal lemma, we are done.

Next, we verify that $\mathsf{Q} \vdash \dot{\neg}\mathsf{Prf}(\ulcorner G \urcorner, \overline{n})$ for all $n : \mathbb{N}$. Now that $\mathsf{Prf}(x, y)$ is Δ_1 and $\ulcorner G \urcorner$ as well as $\overline{n}$ are closed as numerals, we get $\mathsf{Q} \vdash \mathsf{Prf}(\ulcorner G \urcorner, \overline{n})$ or $\mathsf{Q} \vdash \dot{\neg}\mathsf{Prf}(\ulcorner G \urcorner, \overline{n})$. The latter case proves our goal. For the former case, we employ falsity elimination. With Q being consistent, we conclude $\mathsf{Q} \nvdash G$ from $\mathsf{Q} \vdash \dot{\neg} G$. From $\mathsf{Q} \vdash \mathsf{Prf}(\ulcorner G \urcorner, \overline{n})$, we derive $\mathsf{Q} \vdash \dot{\exists} x.\mathsf{Prf}(\ulcorner G \urcorner, x)$, which is equivalent to $\mathsf{Q} \vdash \mathsf{Prov}(\ulcorner G \urcorner)$. By Lemma 3.4, we acquire a derivation of $\mathsf{Q} \vdash G$, clashing with our assumption $\mathsf{Q} \nvdash G$. ∎

Remark 5.14 Let $\mathcal{T}$ be a consistent and ω-consistent extension of Q admitting a formula $\varphi(x)$ satisfying $\forall \tau : \mathbb{F}.\mathcal{T} \vdash \tau \leftrightarrow \mathcal{T} \vdash \varphi(\ulcorner \tau \urcorner)$. The same proof shows Gödel's incompleteness theorem for $\mathcal{T}$. We underline this claim's lack of verification.

If $\lambda\varphi.\mathcal{T} \vdash \varphi$ is enumerable, I postulate the existence of such a formula with a slightly generalised version of the weak representability theorem, but it is not fully certain whether this approach works.

Remark 5.15 In the proof above, we have only used the ω-consistency for the Σ_1-formula $\dot{\exists} x.\mathsf{Prf}(\ulcorner G \urcorner, x)$. Using the lemma `Sigma1_witness` from the library, we can show ω-consistency of Q for such a special kind of formulas. In consequence, we can show Gödel's theorem for Q without relying on more than CT_{Q} being true.

In the proof, we used many assumptions requiring sophisticated arguments. Most prominently, we used that $\mathsf{Prov}(x)$ is of the form $\dot{\exists} y.\mathsf{Prf}(x, y)$ and we relied on ω-consistency of Q on Σ_1-formulas. Also, we used Σ_1-conservativity, although this may be obsolete. Still, this proof is important since many textbooks, such as [Smi22] and [BBJ07] use a similar argumentation showing the claim.

We can, however, employ Fact 1.4 to make the proof significantly easier, which is what we will do next. At first, we need an adapted provability formula.

Lemma 5.16 There exists a Σ_1-formula $\mathsf{SProv}(x)$ satisfying

$$\forall \varphi : \mathbb{F}.\mathsf{Q} \vdash \varphi \to \mathsf{Q} \vdash \mathsf{SProv}(\ulcorner \varphi \urcorner) \wedge \mathsf{Q} \vdash \dot{\neg}\varphi \to \mathsf{Q} \vdash \dot{\neg}\mathsf{SProv}(\ulcorner \varphi \urcorner).$$

Proof We apply Fact 1.4. Since $\lambda\varphi.\mathsf{Q} \vdash \varphi$ is enumerable by standard techniques (cf. [FKS19]), $\lambda\varphi.\mathsf{Q} \vdash \dot\neg\varphi$ can be shown enumerable, too. We conclude that both $P := \lambda n.\mathsf{Q} \vdash \ddot{\mathrm{g}}\mathrm{od}^{-1}(n)$ and $R := \lambda n.\mathsf{Q} \vdash \dot\neg\ddot{\mathrm{g}}\mathrm{od}^{-1}(n)$ are also enumerable. As $\mathbb{N}$ is discrete, we obtain that P and R are semi-decidable. By virtue of Q's consistency, P and R are disjoint. We obtain a Σ_1-formula $\mathsf{SProv}(x)$, such that

$$\forall n : \mathbb{N}. Pn \to \mathsf{Q} \vdash \mathsf{SProv}(\overline{n}) \land Rn \to \mathsf{Q} \vdash \dot\neg\mathsf{SProv}(\overline{n}).$$

Now, suppose $\varphi : \mathbb{F}$. Setting $n = \ddot{\mathrm{g}}\mathrm{od}(\varphi)$ in the result from above yields

$$P(\ddot{\mathrm{g}}\mathrm{od}(\varphi)) \to \mathsf{Q} \vdash \mathsf{SProv}(\ulcorner\varphi\urcorner) \land R(\ddot{\mathrm{g}}\mathrm{od}(\varphi)) \to \mathsf{Q} \vdash \dot\neg\mathsf{SProv}(\ulcorner\varphi\urcorner).$$

We are done since $P(\ddot{\mathrm{g}}\mathrm{od}(\varphi)) = \mathsf{Q} \vdash \varphi$ and $R(\ddot{\mathrm{g}}\mathrm{od}(\varphi)) = \mathsf{Q} \vdash \dot\neg\varphi$. ∎

Remark 5.17 While we have $\mathsf{Q} \vdash \mathsf{Prov}(\ulcorner\varphi\urcorner) \to \mathsf{Q} \vdash \varphi$ for any formula φ, the implication $\mathsf{Q} \vdash \mathsf{SProv}(\ulcorner\varphi\urcorner) \to \mathsf{Q} \vdash \varphi$ has not been inspected yet.

We are now ready to present a much shorter proof of Gödel's first incompleteness theorem.

Proof (of Theorem 5.10, alternative) We apply the diagonal lemma to $\dot\neg\mathsf{SProv}(x)$ and obtain a sentence G such that $\mathsf{Q} \vdash G \dot\leftrightarrow \dot\neg\mathsf{SProv}(\ulcorner G\urcorner)$. We will show: $\mathsf{Q} \nvdash G$ and $\mathsf{Q} \nvdash \dot\neg G$.

Suppose $\mathsf{Q} \vdash G$. Our reasoning is the same as in our first proof. By Lemma 5.16, we obtain $\mathsf{Q} \vdash \mathsf{SProv}(\ulcorner G\urcorner)$. From $\mathsf{Q} \vdash G \dot\leftrightarrow \dot\neg\mathsf{SProv}(\ulcorner G\urcorner)$ and $\mathsf{Q} \vdash G$, we learn $\mathsf{Q} \vdash \dot\neg\mathsf{SProv}(\ulcorner G\urcorner)$. Contradiction to Q's consistency.

Next, let a derivation of $\mathsf{Q} \vdash \dot\neg G$ be given. From Lemma 5.16, we get $\mathsf{Q} \vdash \dot\neg\mathsf{SProv}(\ulcorner G\urcorner)$. Using $\mathsf{Q} \vdash G \dot\leftrightarrow \dot\neg\mathsf{SProv}(\ulcorner G\urcorner)$, we obtain $\mathsf{Q} \vdash G$. Again, this clashes with consistency of Q. ∎

6 Discussion

In some books, the diagonalisation of a formula is defined differently, for instance in Boolos', Burgess' and Jeffrey's book [BBJ07]. Instead of picking $\varphi(\ulcorner\varphi\urcorner)$ directly, the formula $\dot\exists x.x \equiv \ulcorner\varphi\urcorner \dot\land \varphi(x)$ is chosen. Clearly, both are logically equivalent, but the latter imposes technical overhead since we need additional rewrites when using it in proofs. Still, the second formulation does not use any substitutions. However, we opted for the first formulation; this also used in the literature, for instance by Smith [Smi22].

Both authors also use a property similar to our CT_Q-assumption. Since they are not working in constructive type theory, they impose computability conditions on the underlying function f: Both authors arrive at a formulation stating that $\varphi(x, y)$ captures f in Q if $\mathsf{Q} \vdash \dot\forall y.\varphi(\overline{x}, y) \dot\leftrightarrow y \equiv \overline{z}$, provided that $f(x) = z$. While Smith requires f to be primitive recursive and therefore total, we can always replace z with $f(x)$ and obtain our notion of CT_Q. Boolos et al., on the contrary, allow f to be an arbitrary recursive function, so $f(x)$ does not need to be defined for all x. Note that this notion of representability asserts $f(x) = z \to \mathsf{Q} \vdash \dot\forall y.\varphi(\overline{x}, y) \dot\leftrightarrow y \equiv \overline{z}$ but does not enforce the converse $\mathsf{Q} \vdash \dot\forall y.\varphi(\overline{x}, y) \dot\leftrightarrow y \equiv \overline{z} \to f(x) = z$ to hold true. Our definition of partial CT_Q requires both directions. Still, this does not make any difference in our setting as we only work with total functions.

Lastly, CT_Q states that the representing formula is Σ_1. This coincides with Smith's result, up to minor differences in the definition of Σ_1: Instead of being a Δ_1 (cf. [KP23, definition 23]; this property is semantic) formula preceded by existential quantifiers, Smith's Σ_1-formulas are those that

do not use quantification[2] (he calls them Δ_0 formulas; this property is purely syntactic) preceded by existentials. Boolos et al. are even stronger and claim that the representing formula is of the form $\dot{\exists} x.\tau(x)$ with $\tau(x)$ being Δ_0. Hermes and Kirst already showed a somewhat related theorem: Σ_1-compression, cf. [HK23].

Any Δ_0-sentence (that is, a Δ_0 formula with closed terms substituted in for the free variables) can be decided within Q (cf. [Smi22, theorem 23]), so any Δ_0-formula is a Δ_1-formula in our sense.

References

[BBJ07] George S. Boolos, John P. Burgess and Richard C. Jeffrey. **Computability and Logic**. 5. edition. Cambridge University Press, 2007.

[FKS19] Yannick Forster, Dominik Kirst and Gert Smolka. 'On synthetic undecidability in coq, with an application to the entscheidungsproblem'. In: **Proceedings of the 8th ACM SIGPLAN International Conference on Certified Programs and Proofs, CPP 2019, Cascais, Portugal, January 14-15, 2019**. Ed. by Assia Mahboubi and Magnus O. Myreen. ACM, 2019, pp. 38–51. DOI: 10.1145/3293880.3294091. URL: https://doi.org/10.1145/3293880.3294091.

[HK23] Marc Hermes and Dominik Kirst. 'An Analysis of Tennenbaum's Theorem in Constructive Type Theory'. In: **CoRR** abs/2302.14699 (2023). DOI: 10.48550/ARXIV.2302.14699. arXiv: 2302.14699. URL: https://doi.org/10.48550/arXiv.2302.14699.

[KH23] Dominik Kirst and Marc Hermes. 'Synthetic Undecidability and Incompleteness of First-Order Axiom Systems in Coq'. In: **J. Autom. Reason.** 67.1 (2023), p. 13. DOI: 10.1007/S10817-022-09647-X. URL: https://doi.org/10.1007/s10817-022-09647-x.

[KP23] Dominik Kirst and Benjamin Peters. 'Gödel's Theorem Without Tears - Essential Incompleteness in Synthetic Computability'. In: **31st EACSL Annual Conference on Computer Science Logic (CSL 2023)**. Ed. by Bartek Klin and Elaine Pimentel. Vol. 252. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, 30:1–30:18. ISBN: 978-3-95977-264-8. DOI: 10.4230/LIPIcs.CSL.2023.30. URL: https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.CSL.2023.30.

[Nor18] Michael Norrish. **LSS 2018: Computability and Incompleteness. Gödel's First Incompleteness Theorem, Tarski's Indefinability of Truth**. 2018. URL: https://comp.anu.edu.au/lss/lectures/2023/lss-computability-4.pdf.

[Smi22] Peter Smith. **Gödel Without (Too Many) Tears**. 2. edition. Logic Matters, Cambridge, 2022. URL: https://www.logicmatters.net/resources/pdfs/GWT2edn.pdf.

[Smo24] Gert Smolka. **Modeling and Proving in Computational Type Theory Using the Coq Proof Assistant**. Textbook under construction. 2024. URL: https://www.ps.uni-saarland.de/~smolka/drafts/mpctt.pdf.

[2] Strictly speaking, bounded quantification is allowed, but as this can always be recast as sequence of conjunctions or disjunctions, we can omit quantifiers at all.