Noname manuscript No.
(will be inserted by the editor)

Distilling the Requirements of Gödel's Incompleteness Theorems with a Proof Assistant

Andrei Popescu · **Dmitriy Traytel**

Received: date / Accepted: date

Abstract We present an abstract development of Gödel's incompleteness theorems, performed with the help of the Isabelle/HOL proof assistant. We analyze sufficient conditions for the applicability of our theorems to a partially specified logic. In addition to the usual benefits of generality, our abstract perspective enables a comparison between alternative approaches from the literature. These include Rosser's variation of the first theorem, Jeroslow's variation of the second theorem, and the Świerczkowski–Paulson semantics-based approach. As part of the validation of our framework, we upgrade Paulson's Isabelle proof to produce a mechanization of the second theorem that does not assume soundness in the standard model, and in fact does not rely on any notion of model or semantic interpretation.

Contents

Department of Computer Science, University of Sheffield, UK ·
Department of Computer Science, University of Copenhagen, Denmark

1 Introduction

Gödel's incompleteness theorems [14, 17] are landmark results in mathematical logic. Both theorems refer to consistent logical theories that satisfy some assumptions, notably that of "containing enough arithmetic." The first incompleteness theorem ($\mathcal{IT}_1$) says that there are sentences that the theory cannot decide, i.e., neither prove nor disprove; the second theorem ($\mathcal{IT}_2$) says that the theory cannot prove an internal formulation of its own consistency. It is generally accepted that $\mathcal{IT}_1$ and $\mathcal{IT}_2$ have a wide scope (and wider for $\mathcal{IT}_1$ than for $\mathcal{IT}_2$), covering many logics and logical theories. However, when it comes to rigorous presentation, typically these results are proved for particular, albeit paradigmatic cases, such as theories of arithmetic or hereditarily finite (HF) sets, within classical first-order logic (FOL); and even in these cases the constructions and proofs tend to be significantly sketchy and incomplete. Hence, the theorems' scope remains largely unexplored on a rigorous/formal basis.

The emergence of powerful proof assistants (also known as interactive theorem provers) has been slowly changing the rules of the game and, we argue, the expectation. Using proof assistants, we can reliably keep track of all the constructions and their properties. Proof automation (sometimes achieved through the cooperation between proof assistants and automatic theorem provers [25, 41]), makes complete, entirely rigorous proofs feasible. And indeed, researchers have successfully met the challenge of mechanizing $\mathcal{IT}_1$ [21, 36, 40, 53] and recently $\mathcal{IT}_2$ [40]. Besides reassurance, these verification *tours de force* have brought superior technical insight into the theorems. But they have taken place within the same solitary confinement of scope as the informal proofs.

This article takes steps towards a fully formal exploration of the incompleteness theorems and their wide scope, by a detailed analysis of their assumptions. We use Isabelle/HOL [34, 35] to establish general conditions under which the theorems apply to a partially specified logic. Our formalization is publicly available in the Archive of Formal Proofs [44–48], but is not necessary for following this article, which is self-contained and does not employ Isabelle jargon (except for the dedicated Appendix A).

After discussing related work (Section 2) and guiding principles (Section 3), we describe our formal development. The abstract part of this development starts by setting the stage of a partially specified logical system, some partially specified arithmetic components, and representability (Section 4), proving the diagonalization lemmas (Section 5), and proving several flavors of the end results in this setting: $\mathcal{IT}_1$ (Section 6) and $\mathcal{IT}_2$ (Section 7). Some of the abstract results (summarized in Section 8) are instantiated to concrete first-order logic theories (Section 9). We also discuss proof-engineering aspects (Appendix A) and include an index of our abstract assumptions (Appendix B).

We start with a notion of logic whose terms and formulas are kept abstract (Section 4.1). In particular, substitution and free variables are not defined, but axiomatized by some general properties. Provability is also axiomatized (Section 4.2). We distinguish between a *basic provability* relation, capturing minimal theories that are sufficiently expressive for represent-

ing concepts via Gödel encodings (e.g., Peano arithmetic or weaker theories), and a *(plain) provability* relation, capturing consistent or ω-consistent extensions of the minimal theories. Thus, basic provability is subsumed by provability. Yet, provability will be represented internally and reasoned about within basic provability.

On top of this logic substratum, we consider an arithmetic substratum, consisting of a set of closed terms called *numerals* and an order-like relation (Section 4.3). Our framework also incorporates encodings of formulas and proofs into numerals, the representability of various functions and relations as formulas (Section 4.4), the Hilbert–Bernays–Löb derivability conditions (Section 4.5), and standard models (Section 4.6).

Overall, our assumptions capture the notion of "containing enough arithmetics" in a general and flexible way. It is general because only few assumptions are made about the exact nature of formulas and numerals. It is flexible because different versions of the incompleteness theorems consider their own "amount of arithmetics" that makes it "enough," as proper subsets of these assumptions. Indeed, our formalization of the results (the diagonalization lemma in Section 5, $\mathcal{IT}_1$ in Section 6, and $\mathcal{IT}_2$ in Section 7) proceeds in an austere-buffet style: Every result picks just enough infrastructure needed for it to hold—ranging from diagonalization which requires very little, to Rosser's version of $\mathcal{IT}_1$ which is quite demanding. This approach caters for a sharp comparison between different formulations of the theorems, highlighting their tradeoffs: Gödel's original formulation of $\mathcal{IT}_1$ versus Rosser's improvement (Section 6.3), proof-theoretic versus semantic versions of $\mathcal{IT}_1$ (Section 6.4), and Gödel's original formulation of the $\mathcal{IT}_2$ versus Jeroslow's improvement (Section 7.3).

Abstractness is our development's main strength, but also a potential weakness: Are our hypotheses reasonable? Are they consistent? These questions particularly concern our axiomatization of free variables and substitution—a notoriously error-prone area. As a (partial) remedy, we instantiate part of our framework to Paulson's semantics-based $\mathcal{IT}_1$ and $\mathcal{IT}_2$ for hereditarily finite (HF) set theory [40], also upgrading Paulson's $\mathcal{IT}_2$ to a more general and standard formulation: for consistent (not necessarily sound) theories (Section 9).

This article extends our CADE 2019 conference paper [43] with a significantly more fine-grained and self-contained presentation of the results, which includes lemmas and detailed proof sketches. Given the existence of formal proofs in Isabelle, one may question the usefulness of paper proof sketches; however, we believe these are important for reaching out to a wider audience—perhaps interested in following the reasoning behind our fine-grained results discovered with the help of Isabelle, but not willing to read and understand Isabelle scripts. Compared to the conference paper, the results are also established in a more general setting, where we distinguish between basic provability and provability (as explained above). This generalization had been left as future work in the conference paper.

2 Related Work

There is a vast amount of literature on the incompleteness theorems and their extensions and ramifications. We only discuss works that are strongly related to the ideas and techniques we tackle in this article. Gödel initially gave a proof of $\mathcal{IT}_1$ and the rough proof idea of $\mathcal{IT}_2$ [17]. Hilbert and Bernays gave a first detailed proof of $\mathcal{IT}_2$ [22]. Subsequently, a large amount of work was dedicated to the (re)formulation, proof, and analysis of these results [5,50,56,57]. The now canonical line of reasoning goes through the derivability conditions devised by Hilbert and Bernays [22] and simplified by Löb [32]. These conditions have inspired a new branch of modal logic called provability logic [5, 58]. Jeroslow has proved that, contrary to prior belief, one of these conditions is redundant when proving $\mathcal{IT}_2$ [24].

Smullyan [59], Kreisel [28] and Jeroslow [24] were among the first to study abstract conditions on logics under which the incompleteness theorems apply. Feferman [13] gives an essential incompleteness account of $\mathcal{IT}_2$ applicable to extensions of Peano arithmetic in classical FOL. Buldt [7] surveys the state of the art on $\mathcal{IT}_1$ up to 2014 with a focus on the theorem's scope, also sketching the applicability to non-standard logics. Our abstract approach, based on generic syntax, provability and truth predicates, resembles the style of institution-independent model theory [12, 18] and our previous work on abstract completeness [4] and completeness of ordered resolution [51]. On distinguishing between two notions of provability, one stronger than the other, we take inspiration from Smorynski's account [57]. Dimensions of generality that our formalized work does not (yet) explore include quantifier-free logics [24] and arithmetical hierarchy refinements [27]. Our syntax axiomatization is inspired by algebraic theories of the λ-calculi syntax [15, 16, 42].

In the realm of mechanical proofs, the earliest substantial development was due to Sieg [55], who used a prover based on TEM (Theory of Elementary Meta-Mathematics) to formalize parts of the proofs of both $\mathcal{IT}_1$ and $\mathcal{IT}_2$. Full mechanical proofs of $\mathcal{IT}_1$ were subsequently achieved by Shankar [52, 53] in the Boyer–Moore prover, O'Connor in Coq [36], and Harrison in HOL Light [21]. Harrison also proved abstract versions of $\mathcal{IT}_1$ and $\mathcal{IT}_2$ in a simple LCF-style prover implemented in OCaml [19]. $\mathcal{IT}_2$ has only been fully proved recently—by Paulson in Isabelle/HOL [39, 40] (who also proved $\mathcal{IT}_1$).

All these mechanizations target theories over a fixed language in classical FOL: variants of the language of arithmetic (Harrison and O'Connor) and variants of the language of set theory (Sieg, Shankar, and Paulson, with Shankar also allowing the convenience of new symbols to be defined from the existing ones). The targeted theories are usually (finite) extensions of given standard FOL theories—so the results state the (finitary) *essential incompleteness* of these theories. Sieg considers the theory Z^*, Shankar finite extensions of the theory Z2, and Paulson finite extensions of HF set theory. Each of Z^*, Z2 and HF set theory are variations of Zermelo–Frankel set theory without the axiom of infinity, and have the same expressive power as Peano arithmetic [10, 61]. O'Connor targets self-representable extensions of the theory NN [23, §7.1], a modification of Robinson arithmetic obtained by replacing the dichotomy axiom (stating that any element is either 0 or a successor) with three axioms regulating the behavior of an additional binary relation symbol for strict order. Harrison targets Robinson arithmetic, and additionally proves a variant of $\mathcal{IT}_1$ for an abstract class of theories in the FOL language of Robinson arithmetic. On their way to $\mathcal{IT}_1$, Shankar and O'Connor prove representability of all partial, respectively primitive recursive functions—important standalone results. We will revisit some of these mechanized concrete results in Section 9, with the hindsight of our abstract framework.

Outside the realm of holistic interactive proof development, there have been efforts to fully automate parts of the proofs of Gödel's and related theorems [8, 49, 54].

3 Formal Design Principles

Our long-term goal is a framework that makes it easy to instantiate the incompleteness theorems and related results to different logics. This is a daunting task, especially for $\mathcal{IT}_2$, where a lot of seemingly logic-specific technicalities are required to even formulate the theorem. The challenge is to push as much as possible of the technical constructions and lemmas to a largely logic-independent layer.

To this end, we strive to make minimal assumptions with regard to structure and properties when inferring the results—we will call this the *Economy* principle. For example, we do not define, but axiomatize syntax in terms of a minimalistic infrastructure. We assume

a generic single-point substitution, then define simultaneous substitution and infer its properties. This is laborious, but worthwhile: Any logic that provides a single-point substitution satisfying our assumptions gets the simultaneous substitution for free.

As another instance of Economy, when faced with two different ways of formulating a theorem's conclusion we prefer the one that is *stronger under fewer assumptions*. (And dually, we prefer weakness for a theorem's assumptions.) For example, we discuss two variants of consistency: (1) "does not prove false" or (2) "there exists no formula such that itself and its negation are provable" (Section 7.3). While the statements are equivalent at the meta-level, their representations as object-logic formulas are not necessarily equivalent; in fact, (1) implies (2) under mild assumptions but not *vice versa*. So in our abstract theorems we prefer (1). Indeed, even if (2) implies (1) in all reasonable instances, why postpone for the instantiation time any fact that we can show abstractly?

Applying the Economy principle not only stocks up generality for instantiations, but also accurately outlines tradeoffs: How much does it cost (in terms of other added assumptions) to improve the conclusion, or to weaken an assumption of a theorem? For example, an Economy-based proof of Rosser's variant of $\mathcal{IT}_1$ reveals how much arithmetic we must factor in for weakening the ω-consistency assumption into consistency.

4 Abstract Assumptions

Roughly, the incompleteness theorems are considered to hold for logical theories that (1) contain enough arithmetic and (2) can themselves be arithmetized. Our goal is to give a general formulation of these favorable conditions. To this end, we identify some logic and arithmetic substrata consisting of structure and axioms that express the containment of (various degrees of) arithmetic more abstractly and flexibly than relative interpretations [63]. We also identify abstract notions of encodings and representability that have just what it takes for a working arithmetization.

4.1 The logical substratum: syntax

We start with some unspecified sets of variables (Var, ranged over by x, y, z), numerals (Num, ranged over by m, n), terms (Term, ranged over by s, t) and formulas (Fmla, ranged over by φ, ψ, χ). We assume that variables and numerals are particular terms, i.e., $\mathsf{Var} \subseteq \mathsf{Term}$ and $\mathsf{Num} \subseteq \mathsf{Term}$, and that Var is infinite. Free-variables and substitution operators, FVars and $\_\,[\_/\_]$, are assumed for both terms and formulas. We think of $\mathsf{FVars}(t)$ as the (finite) set of free variables of the term t, and similarly for formulas. We think of $s\,[t/x]$ as the term obtained from s by the (capture-avoiding) substitution of t for the free occurrences of variable x; and similarly for $\varphi\,[t/x]$, where φ is a formula.

In FOL, terms introduce no bindings, so any occurring variable is free. FOL terms fall under our framework, and so do terms with bindings as in λ-calculi and higher-order logic (HOL). To achieve this degree of inclusiveness while also being able to prove interesting results, we work under some well-behavedness assumptions about FVars and $\_\,[\_/\_]$:

(1) Free-variables and substitution act on variable terms as expected:
- $\mathsf{FVars}(x) = \{x\}$;
- $x\,[s/x] = s$, and $y\,[s/x] = y$ if $x \neq y$.

(2) Substitution on terms is vacuous outside the free variables:
$x \notin \mathsf{FVars}(t)$ implies $t\,[s/x] = t$; and similarly for substitution on formulas.

In addition, for the operators on formulas we assume the following:

(3) Free-variables distribute over substitution:
$\mathsf{FVars}\,(\varphi\,[s/x]) = \mathsf{FVars}(\varphi) - \{x\} \cup \mathsf{FVars}(s)$ if $x \in \mathsf{FVars}(\varphi)$.
(4) Substitution of a variable for itself is vacuous: $\varphi\,[x/x] = \varphi$.
(5) Substitution is compositional (under some freshness assumptions):
- $\varphi\,[s_1/x]\,[s_2/x] = \varphi\,[(s_1[s_2/x])\,/\,x]$;
- $\varphi\,[s_1/x_1]\,[s_2/x_2] = \varphi\,[(s_1[s_2/x_2])\,/\,x_1]$ if $x_2 \notin \mathsf{FVars}(\varphi)$;
- $\varphi\,[s_1/x_1]\,[s_2/x_2] = \varphi\,[s_2/x_2]\,[(s_1[s_2/x_2])\,/\,x_1]$ if $x_1 \neq x_2$ and $x_1 \notin \mathsf{FVars}(s_2)$.

Of the above assumptions, (1) only applies to, and only makes sense for, substitution on terms. By contrast, we assume (2) for both terms and formulas. The last group, (3)–(5) would makes sense for terms too, but is only assumed for formulas; this is in line with our Economy principle, since our results will not need these assumptions for terms. In these assumptions, just like in the rest of this paper, "$=$" denotes the usual equality of two mathematical entities (formally represented by the Isabelle/HOL equality), and not some more abstract equality. This means that our assumptions do not hold for "raw" formulas, but for formulas quotiented to alpha-equivalence, i.e., equivalence classes modulo alpha (of the kind provided, e.g., by using de Bruijn indices or the Nominal Isabelle package [64]); likewise, if the terms have bindings, they would need to be quotiented to alpha-equivalence to satisfy our assumptions.

The incompleteness theorems rely heavily on simultaneous substitution, whose properties are tricky to formalize—for example, Paulson's formalization article dedicates them ample space [40, 6.2]. To address this problem once and for all generically, we define simultaneous substitution, written $\varphi\,[t_1/x_1, \ldots, t_n/x_n]$, from the single-point substitution, $\varphi\,[t/x]$. Accordingly, we derive the properties of simultaneous substitution from the single-point substitution axioms. For example, we prove that $\mathsf{FVars}\,(\varphi\,[s_1/x_1, \ldots, s_n/x_n]) = \mathsf{FVars}(\varphi) \cup \bigcup \{\mathsf{FVars}(s_i) - \{x_i\} \mid i \in \{1, \ldots, n\} \text{ and } x_i \in \mathsf{FVars}(\varphi)\}$. The technicalities are delicate: To avoid undesired variable replacements, $\varphi\,[s_1/x_1, \ldots, s_n/x_n]$ must be defined as $\varphi\,[y_1/x_1] \ldots [y_n/x_n]\,[s_1/y_1] \ldots [s_n/y_n]$ for some fresh $y_1, \ldots, y_n$, the choice of which we must show to be immaterial. This definition's complexity is reflected in the proofs of its properties. But again, this one-time effort benefits any "customer" logic: In exchange for a well-behaved single-point substitution, it gets back a well-behaved simultaneous substitution.

We call a term with no free variables *closed* and a formula with no free variables a *sentence*. Sen denotes the set of sentences. We let $v_1, v_2, \ldots$ be fixed mutually distinct variables. Fmla_k denotes the set of formulas whose set of free variables is exactly $\{v_1, \ldots, v_k\}$. In particular, $\mathsf{Fmla}_0 = \mathsf{Sen}$. Given $\varphi \in \mathsf{Fmla}_k$, we write $\varphi\,(t_1, \ldots, t_n)$ instead of $\varphi\,[t_1/v_1, \ldots, t_n/v_n]$.

In addition to free variables and substitution, our theorems will require formulas to be equipped with some of the following: term equality ($\equiv$), Boolean connectives ($\bot$, $\top$, $\to$, $\neg$, $\wedge$, $\vee$), universal and existential quantifiers ($\forall$, $\exists$). When we need negation, we define it taking $\neg\,\varphi$ to be $\varphi \to \bot$. On the other hand, even in the presence of negation, we do not assume that $\vee$ and $\exists$ are definable from $\wedge$ and $\forall$ or vice versa. This is because, in line with the Economy principle, we will not assume classical logic except in results that need it. For the rest, we will only assume intuitionistic logic, where these operators are not inter-definable.

The above are not assumed to be constructors (syntax builders), but unspecified operators on terms and formulas, e.g., $\equiv\, : \mathsf{Term} \to \mathsf{Term} \to \mathsf{Fmla}$, $\bot \in \mathsf{Fmla}$, $\forall\, : \mathsf{Var} \times \mathsf{Fmla} \to \mathsf{Fmla}$. This caters for logics that do not have them as primitives. For example, HOL defines all connectives and quantifiers from λ-abstraction and either equality or implication.

Free variables and substitution are assumed to be well-behaved w.r.t. these operators, e.g., $\mathsf{FVars}\,(\forall x.\,\varphi) = \mathsf{FVars}\,(\varphi) - \{x\}$. Finally, numerals are assumed to be closed terms. Thanks to our substitution axioms, this implies that substitution on numerals is vacuous.

4.2 Logical substratum: provability

We fix two unary relations on formulas, $\vdash \subseteq \mathsf{Fmla}$ and $\vdash^{\mathsf{b}} \subseteq \mathsf{Fmla}$, called *provability* and *basic provability*, respectively. We write $\vdash \varphi$ instead of $\varphi \in \vdash$, and say the formula φ is *provable*; similarly, we write $\vdash^{\mathsf{b}} \varphi$ instead of $\varphi \in \vdash^{\mathsf{b}}$, and say the formula φ is *basic-provable*. Henceforth, we will assume that on sentences basic provability is included in provability: For all $\varphi \in \mathsf{Sen}$, $\vdash^{\mathsf{b}} \varphi$ implies $\vdash \varphi$. Typical instances of these relations will be as follows:

- for $\vdash^{\mathsf{b}}$, provability in some minimal theory, e.g., Robinson arithmetic or HF set theory
- for $\vdash$, provability in some recursive extension of such a minimal theory

As we will see, $\vdash^{\mathsf{b}}$ will be assumed to be sufficiently expressive to reason about $\vdash$, and sometimes also sound w.r.t. the standard model. Whenever certain formula connectives or quantifiers are needed in our results, we will assume that $\vdash^{\mathsf{b}}$ and $\vdash$ are closed under the usual (Hilbert-style) intuitionistic FOL axioms and rules with respect to these connectives and quantifiers. Stronger systems, such as those of classical logic, also satisfy these assumptions.

Consistency of $\vdash$, denoted $\mathsf{Con}_{\vdash}$, is defined as the impossibility to prove false, namely $\nvdash \bot$. Another central concept is ω-consistency—we carefully choose a formulation that works intuitionistically, with conclusion reminiscent of Gödel's negative translation [11]:

$\mathsf{OCon}_{\vdash}$: For all $\varphi \in \mathsf{Fmla}_1$, if $\vdash \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$ then $\nvdash \neg\neg (\exists x.\, \varphi(x))$.

Assuming classical deduction in $\vdash$, this is equivalent to the standard formulation: For all $\varphi \in \mathsf{Fmla}_1$, it is not the case that $\vdash \varphi(n)$ for all $n \in \mathsf{Num}$ and $\vdash \neg\, (\forall x.\, \varphi(x))$.

Occasionally, we will consider not only provability but also explicit proofs. We fix a set Proof of (entities we call) *proofs*, ranged over by p, q, and a binary relation between proofs p and sentences φ, written $p \Vdash \varphi$ and read "p is a proof of φ." We assume $\vdash$ and $\Vdash$ to be related as expected, in that provability is the same as the existence of a proof:

$\mathsf{Rel}^{\Vdash}_{\vdash}$: For all $\varphi \in \mathsf{Sen}$, $\vdash \varphi$ iff there exists $p \in \mathsf{Proof}$ such that $p \Vdash \varphi$.

Convention 1. In all shown results we will implicitly assume: (1) the generic syntax (free variable and substitution) axioms, (2) at least $\to$ and $\bot$ plus whatever connectives and quantifiers appear in the statement, (3) the inclusion of $\vdash^{\mathsf{b}}$ into $\vdash$ and (4) the closedness of $\vdash^{\mathsf{b}}$ and $\vdash$ under intuitionistic deduction rules for the assumed connectives and quantifiers. Other assumptions (e.g., order-like relation axioms, classical logic deduction, standard models, etc.) will be indicated explicitly. The appendix contains an index with the explicit assumptions.

In our proof sketches, arguing "by logic" will mean invoking closedness of $\vdash^{\mathsf{b}}$ or $\vdash$ under *intuitionistic* deduction rules; "by classical logic" will explicitly indicate a step that assumes closedness under classical deduction rules. We will label local facts in proofs for later reference by parenthesized Arabic or Roman numbers, such as (1), (2), (i), (ii). The first occurrence of a parenthesized number will label a fact by *preceding* it, as in "we obtain (ii) $\vdash \varphi \wedge \psi$", while later occurrences will mean we refer to it, as in "from (ii) we obtain …".

4.3 Arithmetic substratum

On one occasion, we will assume an order-like binary relation modeled by a formula $\prec \in \mathsf{Fmla}_2$. We write $t_1 \prec t_2$ instead of $\prec (t_1, t_2)$ and $\forall x \prec n.\, \varphi$ instead of $\forall x.\, x \prec n \to \varphi$. It turns out that at our level of abstraction it does not matter whether $\prec$ is a strict or a non-strict order. Indeed, we only require the following two properties, where $x \in M$ denotes $\bigvee_{m \in M} x \equiv m$ and $\bigvee$ expresses the disjunction of a finite set of formulas:

Ord_1: For all $\varphi \in \mathsf{Fmla}_1$ and $n \in \mathsf{Num}$, if $\vdash^{\mathsf{b}} \varphi(m)$ for all $m \in \mathsf{Num}$, then $\vdash \forall x \prec n.\, \varphi(x)$.
Ord_2: For all $n \in \mathsf{Num}$, there exists a finite set $M \subseteq \mathsf{Num}$ such that $\vdash \forall x.\, x \in M \vee n \prec x$.

Ord_1 states that if a property φ is basic-provable for all numerals, then its universal quantification bounded by any given numeral n is provable. Having in mind the arithmetic interpretation of numerals, it would also make sense to assume a stronger version of Ord_1, replacing "if $\vdash^{\mathsf{b}} \varphi(m)$ for all $m \in \mathsf{Num}$" by the weaker hypothesis "if $\vdash^{\mathsf{b}} \varphi(m)$ for all $m \in \mathsf{Num}$ such that $\vdash m \prec n$". But this stronger version will not be needed. Also, note that we formulate Ord_1 in the weakest possible way w.r.t. the choice of provability relations: with a hypothesis about $\vdash^{\mathsf{b}}$ and a conclusion about $\vdash$. Ord_2 states that, for any numeral n, any element x in the domain of discourse is provably either greater than n or equal to one of a finite set M of numerals.

If we instantiate our syntax to that of first-order arithmetic and take $\vdash^{\mathsf{b}}$ to be intuitionistic Robinson arithmetic (and $\vdash$ any larger relation), then both Ord_1 and Ord_2 hold when taking $\prec$ as either $<$ or $\leq$. In the presence of a numeral-restricted form of anti-symmetry of the relation (which would include $<$ but exclude $\leq$), the second condition is stronger:

Lemma 2. Assume $\vdash \forall x.\ x \prec n \to \neg\, n \prec x$. Then Ord_2 implies Ord_1.

Proof. Let $\varphi \in \mathsf{Fmla}_1$ and $n \in \mathsf{Num}$. Assume $\vdash^{\mathsf{b}} \varphi(m)$, in particular, $\vdash \varphi(m)$, for all $m \in \mathsf{Num}$ and let M be as in Ord_2. We must prove $\vdash \forall x \prec n.\ \varphi(x)$. Working inside the formal proof system $\vdash$, we fix x such that $x \prec n$. Thanks to the antisymmetry assumption, we obtain $\neg\, n \prec x$, which implies by Ord_2 that x equals some $m \in M$; this means that $\varphi(x)$ holds, as desired. □

4.4 Encodings and representability

Central in the incompleteness theorems are functions that encode formulas and proofs as numerals, $\langle \_ \rangle : \mathsf{Fmla} \to \mathsf{Num}$ and $\langle \_ \rangle : \mathsf{Proof} \to \mathsf{Num}$. For our abstract results, the encodings are not required to be injective or surjective. Various concepts will be assumed to be *representable* (via these encodings) inside our object logic, via the basic provability relation $\vdash^{\mathsf{b}}$. We will consistently employ $\vdash^{\mathsf{b}}$, and not $\vdash$, to represent concepts. On the other hand, $\vdash$ and its associated proof-of relation $\Vdash$ will be among the concepts we will want to represent.

Let $A_1, \ldots, A_m$ be sets, and let, for each of them, $\langle \_ \rangle : A_i \to \mathsf{Num}$ be an "encoding" function to numerals. Then, an m-ary relation $R \subseteq A_1 \times \ldots \times A_m$ is said to be *represented* by a formula $\textcircled{R} \in \mathsf{Fmla}_m$ if the following hold for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$:

- $(a_1, \ldots, a_m) \in R$ implies $\vdash^{\mathsf{b}} \textcircled{R}(\langle a_1 \rangle, \ldots, \langle a_m \rangle)$
- $(a_1, \ldots, a_m) \notin R$ implies $\vdash^{\mathsf{b}} \neg \textcircled{R}(\langle a_1 \rangle, \ldots, \langle a_m \rangle)$

R is said to be *weakly represented* by $\textcircled{R}$ if, for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$, it holds that $(a_1, \ldots, a_m) \in R$ if and only if $\vdash^{\mathsf{b}} \textcircled{R}(\langle a_1 \rangle, \ldots, \langle a_m \rangle)$. Occasionally, we will use the alternative formulation "$\textcircled{R}$ *(weakly) represents* R."

Let A be another set with $\langle \_ \rangle : A \to \mathsf{Num}$. An m-ary function $f : A_1 \times \ldots A_m \to A$ is said to be *represented* by a formula $\textcircled{f} \in \mathsf{Fmla}_{m+1}$ if for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$:

- $\vdash^{\mathsf{b}} \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, \langle f\ (a_1, \ldots, a_m) \rangle)$
- $\vdash^{\mathsf{b}} \forall x, y.\ \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, x) \wedge \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, y) \to x \equiv y$

A function f as above is *term-represented* by an operator $\textcircled{f} : \mathsf{Term}^m \to \mathsf{Term}$ if $\vdash^{\mathsf{b}} \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle) \equiv \langle f\ (a_1, \ldots, a_m) \rangle$ for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$.

When the formula by which a relation/function P is represented or term-represented is irrelevant, we call P *representable* or *term-representable*.

The terms "representability" and "weak representability" are fairly standard [50]. We refer to Raatikainen [50, §2.2] and Smith [56, §5.6] for an account of different terminologies used in the literature for (variations of) these concepts. In contrast, "term-representability" is a notion that we have introduced ourselves (and so is "cleanness", defined below).

It is immediate that, assuming $\vdash^{\mathsf{b}}$ consistent, if a relation R is weakly represented by a formula then it is also represented by that formula. Moreover, if we assume deductive injectivity of the encoding, i.e., $\vdash^{\mathsf{b}} \langle a_1 \rangle \equiv \langle a_2 \rangle$ implies $a_1 = a_2$ for all $a_1, a_2 \in A$, then the following holds: If a function f is represented by a formula, then its graph $\mathsf{Gr}(f)$ is represented (as a relation) by the same formula, in particular, representability of f implies representability of $\mathsf{Gr}(f)$. The converse, i.e., representability of $\mathsf{Gr}(f)$ implying representability of f (this time by a modified formula), also holds under some assumptions—essentially saying that there is an order-like relation on A that is represented by a formula $\prec$ as in Section 4.3. We do not elaborate on these aspects since they are not used in our end results. Smith works them out in detail in his monograph [56, §16]; he does it for the particular case of Robinson arithmetic, but in such a way that the more general assumptions under which the results hold can be depicted from his proofs. (Smith uses the following terminology: A relation or a function being "captured" means it is represented, and a function being "weakly captured" means its graph is represented as a relation.)

We will also need an enhancement of relation representability: Given $i < m$, we call the representation of an m-ary relation R by $\textcircled{R}$ *i-clean* if $\vdash^{\mathsf{b}} \neg \textcircled{R}(n_1, \ldots, n_m)$ for all numbers $n_1, \ldots, n_m$ such that n_i (the i'th number among them) is outside the image of $\langle \_ \rangle$ (i.e., there is no $a \in A_i$ with $n_i = \langle a \rangle$). Cleanness would be trivially satisfied if the encodings were surjective. However, surjectivity is not a reasonable assumption. For example, most of the numeric encodings used in the literature are injective but not surjective.

The key property of cleanness is that it makes a representation behave well with respect to universal quantification of negative statements. We illustrate this for the binary case:

Lemma 3. Assume $R \subseteq A \times B$ is represented by $\textcircled{R}$, and this representation is 1-clean. Then the following are equivalent:

(1) $(a, b) \notin R$ for all $a \in A$ (2) $\vdash^{\mathsf{b}} \neg \textcircled{R}(n, \langle b \rangle)$ for all $n \in \mathsf{Num}$

Proof. By representability, (1) is equivalent to $\vdash^{\mathsf{b}} \neg \textcircled{R}(\langle a \rangle, \langle b \rangle)$ for all $a \in A$. This, in turn, is equivalent to (2) by 1-cleanness, which lets us exclude numerals outside $\langle \_ \rangle$'s image. □

We let $\mathsf{S} : \mathsf{Fmla}_1 \to \mathsf{Sen}$ be the *self-substitution* function, which sends any $\varphi \in \mathsf{Fmla}_1$ to $\varphi(\langle \varphi \rangle)$, i.e., to the sentence obtained from φ by substituting the encoding of φ for the unique variable of φ. An alternative to the above "hard" version of S is the following "soft" version, which sends any $\varphi \in \mathsf{Fmla}_1$ to $\exists v_1.\, v_1 \equiv \langle \varphi \rangle \wedge \varphi$, where v_1 is the single free variable of φ. The soft version yields provably equivalent formulas and has the advantage that it is easier to represent inside the logic, since it does not require formalizing the complexities of capture-avoiding substitution. All our results involving S have been proved for both versions.

We will consider the properties $\mathsf{Repr}_{\neg}$, $\mathsf{Repr}_{\mathsf{S}}$, and $\mathsf{Repr}_{\Vdash}$, stating the representability of the functions $\neg$ and S, and of the relation $\Vdash$. In addition, $\mathsf{Clean}_{\Vdash}$ will state that the considered representation of $\Vdash$ is 1-clean, i.e., it is clean on the proof component. For the representing formulas of the above relations and functions we will use their circled names, $\textcircled{\neg}$, $\textcircled{\Vdash}$, etc.; for example, $\mathsf{Repr}_{\Vdash}$ means that (1) $p \Vdash \varphi$ implies $\vdash^{\mathsf{b}} \textcircled{\Vdash}(\langle p \rangle, \langle \varphi \rangle)$ and (2) $p \nVdash \varphi$ implies $\vdash^{\mathsf{b}} \neg \textcircled{\Vdash}(\langle p \rangle, \langle \varphi \rangle)$ for all $p \in \mathsf{Proof}$ and $\varphi \in \mathsf{Sen}$.

4.5 Derivability conditions

For several relations R, we will assume representability by formulas $\textcircled{R}$. However, the case of the provability relation $\vdash$ is special. It will have an associated formula $\textcircled{\vdash} \in \mathsf{Fmla}_1$, but we will assume for it conditions weaker than representability, and also additional conditions. The following are known as the Hilbert–Bernays–Löb derivability conditions [22, 32]:

HBL_1: $\vdash \varphi$ implies $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
HBL_2: $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\varphi \to \psi\rangle \to \textcircled{\vdash}\langle\psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
HBL_3: $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle \to \textcircled{\vdash}\langle\textcircled{\vdash}\langle\varphi\rangle\rangle$ for all $\varphi \in \mathsf{Sen}$.

Above and elsewhere, we omit parentheses when instantiating one-variable formulas with encodings of formulas to lighten notation—e.g., writing $\textcircled{\vdash}\langle\varphi\rangle$ instead of $\textcircled{\vdash}(\langle\varphi\rangle)$.

HBL_1 states that, if a sentence is provable, then its encoding is basic-provable inside the representation. We would obtain a weaker version of HBL_1 if we replaced $\vdash^{\mathsf{b}}$ with $\vdash$ in the conclusion, namely asking that $\vdash \varphi$ implies $\vdash \textcircled{\vdash}\langle\varphi\rangle$. HBL_3 is, roughly speaking, a formulation of this weaker version of HBL_1 "one level up," inside the proof system $\vdash^{\mathsf{b}}$. Finally, note that the provability relation is closed under *modus ponens*, in that $\vdash \varphi$ and $\vdash \varphi \to \psi$ implies $\vdash \psi$ for all $\varphi, \psi \in \mathsf{Sen}$. Thus, HBL_2 roughly states the same property inside the proof system. In short, the derivability conditions state that the representation of provability acts partly similarly to the provability relation. (The above internalizations are "rough" in that they use meta-level quantification instead of object-level quantification—we will come back to this in Section 7.2, in the context of $\mathcal{IT}_2$ where these conditions are being used.)

We will also be interested in the converse of HBL_1:

$\mathsf{HBL}_1^{\Leftarrow}$: $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.

The weak representability of $\vdash$ (as defined in Section 4.4) is the conjunction of HBL_1 and $\mathsf{HBL}_1^{\Leftarrow}$. Moreover, $\Vdash$'s representability implies HBL_1 for $\textcircled{\vdash}(x)$ defined to be $\exists y.\, \textcircled{\Vdash}(y, x)$:

Lemma 4. $\mathsf{Rel}_{\vdash}^{\Vdash}$ and $\mathsf{Repr}_{\Vdash}$ imply HBL_1.

Proof. Assume $\vdash \varphi$. Then there exists $p \in \mathsf{Proof}$ such that $p \Vdash \varphi$. By $\mathsf{Repr}_{\Vdash}$, we have $\vdash^{\mathsf{b}} \textcircled{\Vdash}(\langle p\rangle, \langle\varphi\rangle)$, hence $\vdash^{\mathsf{b}} \exists y.\, \textcircled{\Vdash}(y, \langle\varphi\rangle)$, as desired. (Note that we did not need the whole $\mathsf{Repr}_{\Vdash}$; one implication in the representability condition of $\Vdash$ would have sufficed.) □

Convention 5. Whenever we assume explicit proofs and representability of proof-of, the formula $\textcircled{\vdash}$ will be defined from $\textcircled{\Vdash}$ as shown above.

4.6 Standard models

We fix a unary relation $\models \subseteq \mathsf{Sen}$, representing *truth of a sentence in the standard model*. We write $\models \varphi$ instead of $\varphi \in \models$, and read it as "φ is true." We consider the assumptions:

$\mathsf{LCQ}_{\models}$: Logical connectives and quantifiers handle truth as expected:
(1) $\not\models \bot$; (2) for all $\varphi, \psi \in \mathsf{Sen}$, $\models \varphi$ and $\models \varphi \to \psi$ imply $\models \psi$;
(3) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \varphi(n)$ for all $n \in \mathsf{Num}$ then $\models \forall x.\, \varphi(x)$;
(4) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \exists x.\, \varphi(x)$ then $\models \varphi(n)$ for some $n \in \mathsf{Num}$;
(5) for all $\varphi \in \mathsf{Sen}$, $\models \varphi$ or $\models \neg\varphi$.
$\mathsf{Sound}_{\models}^{\vdash^{\mathsf{b}}}$ (basic provability is sound w.r.t. truth): $\vdash^{\mathsf{b}} \varphi$ implies $\models \varphi$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{TIP}_{\models}^{\vdash}$ (truth implies provability): $\models \textcircled{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.

Note that $\mathsf{Sound}_{\models}^{\vdash^{\mathsf{b}}}$ refers to $\vdash^{\mathsf{b}}$, not $\vdash$. Not having to assume that $\vdash$ is sound will allow us to capture, for example, consistent or ω-consistent extensions of Robinson arithmetic that are not sound in the standard natural numbers model.

$\mathsf{LCQ}_{\models}$(1–4) form a partial description of the connectives' and quantifiers' behavior w.r.t. truth: corresponding to elimination rules for $\bot$, $\to$ and $\exists$ and introduction rule for $\forall$. This partial description suffices for our results. Note that $\mathsf{LCQ}_{\models}$(4) is a strong form of existential elimination, saying that (the interpretations of) numerals are a complete set of witnesses for existential formulas valid in the standard model; in particular, this holds for the case

when the standard model is built of numerals only. $\mathsf{LCQ}_{\models}(5)$ states that the standard model decides every sentence. $\mathsf{TIP}^{\vdash}_{\models}$ is a form of completeness: It states that $\vdash$ can prove whatever the standard model "agrees" that can be proved by $\vdash$.

The above axiomatization of standard models will be used to obtain semantic versions of $\mathcal{IT}_1$. At the heart of these results there will be the connection between the representability of $\Vdash$ and $\mathsf{HBL}_1^{\Leftarrow}$ in the presence of standard models. Recall that, by Convention 5, whenever we assume $\Vdash$ representable, we also assume that $\vdash$'s representation $\textcircled{\vdash}$ is naturally defined from $\Vdash$'s representation $\textcircled{\Vdash}$ (matching the definition of $\vdash$ from $\Vdash$). This is crucial for $\mathcal{IT}_2$, where the internal definitions must faithfully capture the external ones [1], but not for $\mathcal{IT}_1$, where we only care about producing, no matter how, an undecided (and true) sentence. In fact, for recursively enumerable extensions of the Robinson arithmetic and related FOL theories, it is possible to produce an artificial provability formula $\textcircled{\vdash}$ that enjoys better properties than the above natural choice: While the latter satisfies HBL_1 but not necessarily $\mathsf{HBL}_1^{\Leftarrow}$, the former is guaranteed to satisfy both HBL_1 and $\mathsf{HBL}_1^{\Leftarrow}$ (i.e., to weakly represent provability). This is why, for example, in his abstract account, Buldt takes the liberty to assume not only HBL_1 but also $\mathsf{HBL}_1^{\Leftarrow}$ in his most general formulation of $\mathcal{IT}_1$ [7, Theorem 3.1]. We will not attempt to model such "artificial" versions of $\textcircled{\vdash}$ in our framework, but will focus on the "natural" one, which works for both $\mathcal{IT}_1$ and $\mathcal{IT}_2$.

On his way to formalizing $\mathcal{IT}_2$ for extensions of HF set theory (and thus having in mind the "natural" $\textcircled{\vdash}$), after proving HBL_1 Paulson notes [40, p.21]: "The reverse implication [namely $\mathsf{HBL}_1^{\Leftarrow}$], despite its usefulness, is not always proved." However, for the "natural" $\textcircled{\vdash}$, $\mathsf{HBL}_1^{\Leftarrow}$ does not come cheaply: It seems to require the soundness of $\vdash^{\mathsf{b}}$ w.r.t. truth in the standard model (which Paulson assumes), or at least the ω-consistency of $\vdash$. We can depict the situation abstractly, without knowing what standard models look like:

Lemma 6. (1) Assume $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and $\mathsf{OCon}_{\vdash}$. Then $\mathsf{HBL}_1^{\Leftarrow}$ holds.
(2) Assume $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\mathsf{TIP}^{\vdash}_{\models}$. Then $\mathsf{HBL}_1^{\Leftarrow}$ holds.
(3) Assume $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\mathsf{LCQ}_{\models}(1,2,4)$. Then $\mathsf{TIP}^{\vdash}_{\models}$ holds. (In particular, $\mathsf{HBL}_1^{\Leftarrow}$ holds.)

Proof. (1): Assume $\vdash \textcircled{\vdash}\langle\varphi\rangle$.
- Hence $\vdash \exists x.\, \textcircled{\Vdash}(x, \langle\varphi\rangle)$. (Recall Convention 5.)
- By logic, we obtain $\vdash \neg\neg\,(\exists x.\, \textcircled{\Vdash}(x, \langle\varphi\rangle))$.
- With $\mathsf{OCon}_{\vdash}$, we obtain $n \in \mathsf{Num}$ such that $\nvdash \neg\, \textcircled{\Vdash}(n, \langle\varphi\rangle)$, in particular $\nvdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(n, \langle\varphi\rangle)$.
- With $\mathsf{Clean}_{\Vdash}$, we obtain $p \in \mathsf{Proof}$ such that $n = \langle p\rangle$. Hence $\nvdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(\langle p\rangle, \langle\varphi\rangle)$.
- Since, by $\mathsf{Repr}_{\Vdash}$, we have that $p \nVdash \varphi$ implies $\vdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(\langle p\rangle, \langle\varphi\rangle)$, we obtain $p \Vdash \varphi$.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $\vdash \varphi$, as desired.
(2): To prove $\mathsf{HBL}_1^{\Leftarrow}$, let $\varphi \in \mathsf{Sen}$ and assume $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models \textcircled{\vdash}\langle\varphi\rangle$.
- With $\mathsf{TIP}^{\vdash}_{\models}$, we obtain $\vdash \varphi$, as desired.
(3): To prove $\mathsf{TIP}^{\vdash}_{\models}$, assume $\models \textcircled{\vdash}\langle\varphi\rangle$.
- Then $\models \exists x.\, \textcircled{\Vdash}(x, \langle\varphi\rangle)$.
- With $\mathsf{LCQ}_{\models}(4)$, we obtain $n \in \mathsf{Num}$ such that (i) $\models \textcircled{\Vdash}(n, \langle\varphi\rangle)$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\mathsf{LCQ}_{\models}(1,2)$, we obtain $\nvdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(n, \langle\varphi\rangle)$.
- Now the proof of $\vdash \varphi$ proceeds just like at point (1): using $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$. □

Lemma 6 shows that, in the presence of standard models with reasonable properties and the soundness of $\vdash^{\mathsf{b}}$, clean representability of the proof-of relation implies $\mathsf{HBL}_1^{\Leftarrow}$; and recall

from Lemma 4 that it also implies HBL_1. Interestingly, a converse of these implications also holds. To state it, we initially assume there is no "outer" notion of proof (i.e., no set Proof and no relation $\Vdash$), but only an "inner" one, given by a formula $\mathsf{Pf} \in \mathsf{Fmla}_2$ such that:

$\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$: $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle \leftrightarrow (\exists x.\, \mathsf{Pf}(x, \langle\varphi\rangle))$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{Compl}_{\mathsf{Pf}}$: $\models \mathsf{Pf}(n, \langle\varphi\rangle)$ implies $\vdash^{\mathsf{b}} \mathsf{Pf}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.
$\mathsf{Compl}_{\neg\mathsf{Pf}}$: $\models \neg\, \mathsf{Pf}(n, \langle\varphi\rangle)$ implies $\vdash^{\mathsf{b}} \neg\, \mathsf{Pf}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.

$\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$ is the inner version of $\mathsf{Rel}^{\Vdash}_{\vdash}$: It expresses that, *inside the representation*, proofs and provability are connected as expected. $\mathsf{Compl}_{\mathsf{Pf}}$ and $\mathsf{Compl}_{\neg\mathsf{Pf}}$ state that provability is complete on Pf statements about formula encodings, as well as on their negations; in traditional settings, this is true thanks to Pf being a Δ_1-formula. Now the converse result states that, thanks to standard models, HBL_1 and $\mathsf{HBL}^{\Leftarrow}_1$, we can define an outer notion of proof that is represented by the inner notion Pf:

Lemma 7. Assume $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{Compl}_{\neg\mathsf{Pf}}$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, $\mathsf{LCQ}_{\models}(4,5)$, HBL_1, $\mathsf{HBL}^{\Leftarrow}_1$. Take $\mathsf{Proof} = \mathsf{Num}$ and define $\Vdash$ by $n \Vdash \varphi$ iff $\vdash^{\mathsf{b}} \mathsf{Pf}\ (n, \langle\varphi\rangle)$. Then $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$ hold, with $\Vdash$ being represented by Pf (i.e., $\textcircled{\Vdash}$ being Pf).

Proof. To show $\mathsf{Rel}^{\Vdash}_{\vdash}$ in this context (that is, for this particular definitions of Proof and relation $\Vdash$), we must show the equivalence between (i) $\vdash \varphi$ and (ii) the existence of $n \in \mathsf{Num}$ such that $\vdash^{\mathsf{b}} \mathsf{Pf}\ (n, \langle\varphi\rangle)$.
First assume (i).
- With HBL_1, we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$.
- With $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, we obtain $\vdash^{\mathsf{b}} \exists x.\, \mathsf{Pf}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models \exists x.\, \mathsf{Pf}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{LCQ}_{\models}(4)$, we obtain $n \in \mathsf{Num}$ such that $\models \mathsf{Pf}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\mathsf{Pf}}$, we obtain (ii), as desired.
Now assume (ii).
- By logic, we obtain $\vdash^{\mathsf{b}} \exists x.\, \mathsf{Pf}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$.
- With $\mathsf{HBL}^{\Leftarrow}_1$, we obtain (i), as desired.
Showing half of $\mathsf{Repr}_{\Vdash}$ in this context is trivial, as it amounts to showing that $\vdash^{\mathsf{b}} \mathsf{Pf}\ (n, \langle\varphi\rangle)$ implies $\vdash^{\mathsf{b}} \mathsf{Pf}\ (n, \langle\varphi\rangle)$. For the other half, assume $\nvdash^{\mathsf{b}} \mathsf{Pf}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\mathsf{Pf}}$, we obtain $\not\models \mathsf{Pf}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{LCQ}_{\models}(5)$, we obtain $\models \neg\, \mathsf{Pf}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\neg\mathsf{Pf}}$, we obtain $\vdash^{\mathsf{b}} \neg\, \mathsf{Pf}\ (n, \langle\varphi\rangle)$, as desired.
Finally $\mathsf{Clean}_{\Vdash}$ is trivial in this context, since the encoding of proofs is the identity. □

The property $\mathsf{TIP}^{\vdash}_{\models}$ will be pivotal in the proofs of our semantic versions of $\mathcal{IT}_1$. As Lemma 6(3) shows, $\mathsf{TIP}^{\vdash}_{\models}$ follows from the soundness of $\vdash^{\mathsf{b}}$, reasonable properties of $\models$ (namely $\mathsf{LCQ}_{\models}(1,2,4)$), and the $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ trio; and the last trio follows by Lemma 7 from the other assumptions if we assume an additional reasonable property of $\models$ (namely $\mathsf{LCQ}_{\models}(5)$), together with $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{Compl}_{\neg\mathsf{Pf}}$, and the weak representability of $\vdash$ (i.e., HBL_1 and $\mathsf{HBL}^{\Leftarrow}_1$). One disadvantage of this indirect route for obtaining $\mathsf{TIP}^{\vdash}_{\models}$ is the need to have both $\mathsf{Compl}_{\mathsf{Pf}}$ and $\mathsf{Compl}_{\neg\mathsf{Pf}}$—which are very tedious to prove for concrete logics, especially $\mathsf{Compl}_{\neg\mathsf{Pf}}$. However, it turns out that we can directly prove $\mathsf{TIP}^{\vdash}_{\models}$ from a subset of the above assumptions, not including $\mathsf{Compl}_{\neg\mathsf{Pf}}$:

Lemma 8. Assume $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, and $\mathsf{LCQ}_{\models}(2,4)$. Define Proof and $\Vdash$ as in Lemma 7.
(1) Then $\models \textcircled{\vdash}\langle\varphi\rangle$ implies $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
(2) If in addition we assume $\mathsf{HBL}^{\Leftarrow}_1$, then $\mathsf{TIP}^{\vdash}_{\models}$ holds.

Proof. (1): Assume (i) $\models \textcircled{\vdash}\langle\varphi\rangle$.
- From $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$ and $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models \textcircled{\vdash}\langle\varphi\rangle \to (\exists x.\, \mathsf{Pf}(x, \langle\varphi\rangle))$.
- With (i) and $\mathsf{LCQ}_{\models}(2)$, we obtain $\models \exists x.\, \mathsf{Pf}(x, \langle\varphi\rangle)$.
- With $\mathsf{LCQ}_{\models}(4)$, we obtain $n \in \mathsf{Num}$ such that $\models \mathsf{Pf}(n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\mathsf{Pf}}$, we obtain $\vdash^{\mathsf{b}} \mathsf{Pf}(n, \langle\varphi\rangle)$.
- By logic, from this we obtain $\vdash^{\mathsf{b}} \exists x.\, \mathsf{Pf}(x, \langle\varphi\rangle)$.
- With $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, by logic we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$, as desired.

(2): Follows immediately from point (1), given that $\mathsf{HBL}^{\Leftarrow}_1$ and $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\varphi\rangle$ imply $\vdash \varphi$. □

Note that point (1) of the above lemma states that basic provability is complete for sentences of the form $\textcircled{\vdash}\langle\varphi\rangle$. For Robinson arithmetic and related theories, this follows from the completeness of provability for Σ_1-sentences (Σ_1-completeness).

5 Diagonalization

The formula diagonalization technique (due to Gödel and Carnap [9]) yields "self-referential" sentences. All we need for it to work is (logic plus) the representability of substitution.

Prop 9. Assuming $\mathsf{Repr}_{\mathsf{S}}$, for all $\psi \in \mathsf{Fmla}_1$ there exists $\varphi \in \mathsf{Sen}$ with $\vdash^{\mathsf{b}} \varphi \leftrightarrow \psi\langle\varphi\rangle$.

Proof. Assume $\mathsf{Repr}_{\mathsf{S}}$, where S is the "hard" self-substitution function. Let $\chi \in \mathsf{Fmla}_1$ be $\exists y.\, \textcircled{\mathsf{S}}(x, y) \wedge \psi(y)$. We take φ to be $\chi\langle\chi\rangle$ and must prove (1) $\vdash^{\mathsf{b}} \varphi \leftrightarrow \psi\langle\varphi\rangle$.
- From the fact that S is represented by $\textcircled{\mathsf{S}}$ we obtain (provably in the formal system $\vdash^{\mathsf{b}}$) that $\langle\varphi\rangle$ is the unique y for which $\textcircled{\mathsf{S}}(\langle\chi\rangle, y)$ holds.
- By logic, this implies $\vdash^{\mathsf{b}} (\exists y.\, \textcircled{\mathsf{S}}(\langle\chi\rangle, y) \wedge \psi(y)) \leftrightarrow \psi\langle\varphi\rangle$.
- By the definition of χ, the above means exactly (1).

A similar argument works for soft self-substitution. □

A sentence $\varphi \in \mathsf{Sen}$ is called:
- a *Gödel sentence* if $\vdash \varphi \leftrightarrow \neg\, \textcircled{\vdash}\langle\varphi\rangle$,
- a *basic Gödel sentence* $\vdash^{\mathsf{b}} \varphi \leftrightarrow \neg\, \textcircled{\vdash}\langle\varphi\rangle$,
- a *Rosser sentence* if $\vdash \varphi \leftrightarrow \neg\, (\exists x.\, \textcircled{\Vdash}(x, \langle\varphi\rangle) \wedge \mathsf{RosserTwist}(x, \langle\varphi\rangle))$,
- a *basic Rosser sentence* if $\vdash^{\mathsf{b}} \varphi \leftrightarrow \neg\, (\exists x.\, \textcircled{\Vdash}(x, \langle\varphi\rangle) \wedge \mathsf{RosserTwist}(x, \langle\varphi\rangle))$.

Above, the formula $\mathsf{RosserTwist}(x, y)$ is $\forall x'.\, x' \prec x \to \forall y'.\, \textcircled{\neg}(y, y') \to \neg\, \textcircled{\Vdash}(x', y')$. Here, y' represents the negation of y. If negation were represented not by a formula but by a unary function symbol $\textcircled{\neg}$, $\mathsf{RosserTwist}(x, y)$ would be written $\forall x'.\, x' \prec x \to \neg\, \textcircled{\Vdash}(x', \textcircled{\neg}(y))$.

Since $\vdash^{\mathsf{b}}$ is included in $\vdash$, any basic Gödel or Rosser sentence is in particular a Gödel or Rosser sentence, respectively. It will turn out that basic Gödel sentences will be needed for the model-theoretic versions of $\mathcal{IT}_1$, whereas (not necessarily basic) Gödel or Rosser sentences will suffice for the proof-theoretic versions.

Prop 10. Assuming $\mathsf{Repr}_{\mathsf{S}}$, there exist basic Gödel and basic Rosser sentences.

Proof. Follows immediately from Prop. 9, taking $\psi(x)$ to be $\neg\, \textcircled{\vdash}(x)$ and $\neg\, (\exists y.\textcircled{\Vdash}(y, x) \wedge \mathsf{RosserTwist}(y, x))$, respectively. □

Thus, any (basic) Gödel sentence is (basic-)provably equivalent to the negation of its own provability; in Gödel's words, it "says about itself that it is not provable" [17]. A Rosser sentence φ asserts its own unprovability in a weaker fashion: Rather than saying "I am not provable" (i.e., "it is not the case that there exists a proof p of me"), it says "it is not the case that there exists a proof p of me such that all smaller q are not proofs of $\neg\,\varphi$." Here, "smaller" refers to the order that the encoding of proofs as numerals imposes.

6 First Incompleteness Theorem

After last sections' preparations, we are now ready to discuss different versions of the incompleteness theorems, based on alternative assumptions. This section deals with $\mathcal{IT}_1$, and the next one with $\mathcal{IT}_2$.

For a consistent or ω-consistent theory that is sufficiently expressive (in particular able to express concepts about itself, such as formulas and provability), $\mathcal{IT}_1$ identifies sentences that are neither provable nor disprovable, and are also true in the standard model—these are usually the Gödel and Rosser sentences discussed in the previous section.

6.1 Informal account and roadmap

Before embarking on the formal analysis of $\mathcal{IT}_1$, it is worth recalling informally the line of reasoning behind some of its variants. (More details can be found, e.g., in Boolos's [5] and Smith's [56] monographs.) Gödel's original formulation referred to a system called P, a form of simple type theory enriched with the Dedekind–Peano axioms for natural numbers. However, it was soon recognized that the argument works for much weaker systems, notably Robinson arithmetic and *a fortiori* Peano arithmetic, as well as for any (ω-)consistent recursively axiomatizable FOL theories that extend these.

When reading the informal (but quite detailed) recollection that follows, the reader should feel free to think of any of the above systems as target systems—so the term "provable" will refer to provability in one of these systems. To simplify the discussion, we will assume the availability of classical logic reasoning, but the later formal analysis will refine this by singling out the results that only need intuitionistic logic. Moreover, here we will not distinguish between provability and basic provability, but leave this too for our later formal analysis. Enclosing a statement in double quotes will mean that we refer to its internalization as a sentence in the language of the considered system; for example, the provability of "n is not a proof of R" can be written using our formal notations as $\vdash \neg\, \textcircled{\vdash}(n, \langle \mathsf{R} \rangle)$.

(1) Let us first consider a purely proof-theoretic $\mathcal{IT}_1$, which ignores the notion of truth and focuses on undecidability.

(1.1) Gödel's original argument goes as follows, for a Gödel sentence G.

(1.1.1) That G is unprovable is argued straightforwardly: The provability of G on the one hand, by HBL_1, would imply that its provability is provable, and on the other hand, by virtue of G being a Gödel sentence, would imply that its unprovability is provable, thus contradicting consistency.

(1.1.2) That $\neg\,\mathsf{G}$ is unprovable needs a more subtle argument, which delves into actual proofs and their representation: The provability of $\neg\,\mathsf{G}$ would imply, by consistency, the unprovability of G, i.e., the nonexistence of any proof of G, i.e., by proof representability, the provability of "n is not a proof of G" for all n, i.e., by ω-consistency, the unprovability of "there exists a proof of G", i.e., unprovability of "G is provable", i.e., by virtue of G being a Gödel sentence, the unprovability of $\neg\,\mathsf{G}$.

(1.2) Rosser's variant removes the need for ω-consistency in Gödel's argument for $\neg\,\mathsf{G}$. This is done by using Rosser sentences R instead of Gödel sentences G. (Recall form Section 5 that Rosser sentences assert about themselves something *weaker* than their unprovability, namely the nonexistence of any proof of them such that Rosser's twist holds, i.e., there is no smaller proof of their negation.)

(1.2.1) Arguing that $\neg\,\mathsf{R}$ is unprovable goes the same as in Gödel's case until the point of establishing the provability of "n is not a proof of R" for all n, while additionally recording a proof p of $\neg\,\mathsf{R}$ (from the assumption that $\neg\,\mathsf{R}$ is provable), which by proof representability brings the provability of "$\langle p\rangle$ is a proof of $\neg\,\mathsf{R}$". So, taking $m = \langle p\rangle$, we have the provability of "m is a proof of $\neg\,\mathsf{R}$" for a fixed m, and also of "n is not a proof of R" for all n. Using a bit of Robinson arithmetic, this gives us the provability of "there exists no x such that x is a proof of R and Rosser's twist holds for x." Hence, by virtue of R being a Rosser sentence, we obtain the provability of R—which, given our initial assumption that $\neg\,\mathsf{R}$ is provable, contradicts consistency.

(1.2.2) On the other hand, due to the aforementioned weaker "self-assertion" in Rosser sentences, Rosser's argument for the unprovability of R is not as immediate as in Gödel's case, but itself needs to delve into proofs. First, proceeding in the same way as for $\neg\,\mathsf{R}$, we obtain a dual of the situation from there: the provability of "m is a proof of R" for a fixed m, and also of "n is not a proof of $\neg\mathsf{R}$" for all n. Again using a bit of Robinson arithmetic (a different bit than before!), we obtain the provability of "Rosser's twist holds for m", hence the provability of "there exists x such that x is a proof of R and Rosser's twist holds for x", hence, by virtue of R being a Rosser sentence, the provability of $\neg\,\mathsf{R}$—which, given our initial assumption that R is provable, contradicts consistency.

(2) Now we move to the argument for why the given undecided sentence is also true in the standard model. In what follows, truth and falsity will implicitly refer to the standard model.

(2.1) For a Gödel sentence G, we know that G is not provable, hence there is no proof of G, hence, by proof representability, it is provable that "n is not a proof of G" for all n. In particular, since deduction is sound w.r.t. truth, it is true that "n is not a proof of G" for all n, i.e., that "for all x, x is not a proof of G", i.e., that "G is not provable". Hence, by virtue of G being a Gödel sentence and deduction being sound, we obtain that G is true.

(2.2) The truth of a Rosser sentence R follows by the same argument as above, noting that we only used that a Gödel sentence is *implied* by the statement of its own unprovability, which is also true for Rosser sentences.

(3) As we will show later during the formal discussion, if stated carefully the above arguments do not need the full power of classical logic, but intuitionistic logic suffices. On the other hand, if we assume classical logic (i.e., double negation) and additional properties mentioned below, more direct arguments can be given for some of $\mathcal{IT}_1$'s components—more precisely, the arguments for the unprovability of $\neg\,\mathsf{G}$ and the truth of G no longer need to delve into proofs, but can stay at the level of provability. Below we only discuss the case of Gödel sentences; Rosser sentences can be treated in exactly the same way.

(3.1) To argue that $\neg\,\mathsf{G}$ is unprovable, we assume that it is provable. Hence, by virtue of G being a Gödel sentence and making essential use of classical logic, we obtain the provability of "G is provable". At this point, we invoke the converse of HBL_1 (i.e., the provability of any φ follows from the provability of φ's provability) to obtain the provability of G, which together with our assumption contradicts consistency.

(3.2) To argue that G is true, we assume otherwise and try to reach a contradiction (thus making essential use of classical negation). Since G is false, $\neg\,\mathsf{G}$ must be true, hence

by soundness and by virtue of G being a Gödel sentence, "not not G is provable" must be true, hence "G is provable" must be true. At this point, we invoke that the truth of provability implies provability, a property that we called $\mathsf{TIP}^{\vdash}_{\models}$ in Section 4.6, to reach the desired conclusion, namely that G is provable. In turn, $\mathsf{TIP}^{\vdash}_{\models}$ can be inferred from Σ_1-completeness (which states that, for all Σ_1 sentences, in particular for those asserting provability, their truth implies their provability) and the converse of HBL_1.

This concludes our informal recollection, which offers a roadmap for our formal and more abstract development that follows: Subsections 6.2, 6.3, 6.4 and 6.5 tackle the above points (1.1), (1.2), (2) and (3), respectively. We distill the exact assumptions needed in these arguments. This forms a basis for generalizing them to a large variety of logical systems, and also reveals some interesting properties required from the logic and arithmetic infrastructures and from the encodings that are not clearly visible in the concrete setting. In particular, we identify the purely intuitionistic line of reasoning that suffices for (1) and (2), the amount of arithmetic needed in (1.2), the tradeoffs between (1.1) and (1.2), and, in Subsection 6.6, the limits in combining provability with basic provability to widen these arguments' scope.

6.2 Gödel's proof-theoretic version

We start with an analysis of Gödel's original argument for the undecidability of Gödel sentences, which requires consistency for one half and ω-consistency for the other half.

Prop 11. Assume $\mathsf{Con}_{\vdash}$ and HBL_1. Then $\nvdash \mathsf{G}$ for all Gödel sentences G.

Proof. Let G be a Gödel sentence. To prove $\nvdash \mathsf{G}$, we assume (1) $\vdash \mathsf{G}$ and aim to reach a contradiction.
- From (1) and G being a Gödel sentence, we obtain $\vdash \neg \textcircled{\scriptsize$\vdash$}\langle \mathsf{G} \rangle$.
- From (1) and HBL_1, we obtain $\vdash^{\mathsf{b}} \textcircled{\scriptsize$\vdash$}\langle \mathsf{G} \rangle$, hence $\vdash \textcircled{\scriptsize$\vdash$}\langle \mathsf{G} \rangle$.
- The last two facts contradict $\mathsf{Con}_{\vdash}$. □

For showing that the Gödel sentences are not disprovable, a standard route is to assume explicit proofs, strengthen the consistency assumption to ω-consistency, and strengthen HBL_1 to representability of the proof-of relation.

Prop 12. Assume $\mathsf{OCon}_{\vdash}$, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$. Then $\nvdash \neg \mathsf{G}$ for all Gödel sentences G.

Proof. Let G be a Gödel sentence. To prove $\nvdash \neg \mathsf{G}$, we assume (1) $\vdash \neg \mathsf{G}$ and aim to reach a contradiction.
- From $\mathsf{OCon}_{\vdash}$, we obtain $\mathsf{Con}_{\vdash}$.
- With (1), we obtain $\nvdash \mathsf{G}$.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $p \nVdash \mathsf{G}$ for all $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$, by Lemma 3 we obtain $\vdash^{\mathsf{b}} \neg \textcircled{\scriptsize$\Vdash$}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- Since $\vdash^{\mathsf{b}}$ is included in $\vdash$, we obtain $\vdash \neg \textcircled{\scriptsize$\Vdash$}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{OCon}_{\vdash}$, we obtain $\nvdash \neg \neg \exists x.\, \textcircled{\scriptsize$\Vdash$}(x, \langle \mathsf{G} \rangle)$, i.e., $\nvdash \neg \neg \textcircled{\scriptsize$\vdash$}\langle \mathsf{G} \rangle$.
- With G being a Gödel sentence, we obtain $\nvdash \neg \mathsf{G}$, which contradicts (1). □

While the line of reasoning in the above proof is mostly well-known, it contains two subtle points about which the literature is not explicit (due to the usual focus on classical first-order arithmetic and particular choices of encodings).

First, we must assume the representation of the proof-of relation $\Vdash$ to be 1-*clean*, i.e., clean with respect to the proof component. Indeed, the argument crucially relies on converting the statement "$p \nVdash \mathsf{G}$ for all $p \in \mathsf{Proof}$" into "$\vdash^{\mathsf{b}} \neg \textcircled{\scriptsize$\Vdash$}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$," which

is only possible for 1-clean encodings. This assumption is needed in many of our results. By contrast, cleanness is never required with respect to the sentence component of proof-of or for the provability relation (which only involves sentence encodings). In short, cleanness is only needed for proofs, not for sentences.

Second, to reach the desired contradiction for our intuitionistic proof system $\vdash$, from "$\vdash \neg \textcircled{\Vdash}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$" it is not sufficient to employ standard ω-consistency, which would only give us $\nvdash \exists x.\ \textcircled{\Vdash}(x, \langle \mathsf{G} \rangle)$, i.e., $\nvdash \textcircled{\vdash}\langle \mathsf{G} \rangle$; the last together with $\vdash \mathsf{G} \leftrightarrow \neg\, \textcircled{\vdash}\langle \mathsf{G} \rangle$ would be insufficient for obtaining $\nvdash \neg\, \mathsf{G}$. However, our stronger version of ω-consistency, $\mathsf{OCon}_{\vdash}$, does the job. $\mathcal{IT}_1$ now follows by putting together Props. 10–12:

Theorem 13. ($\mathcal{IT}_1$) Assume $\mathsf{OCon}_{\vdash}$, $\mathsf{Rel}_{\vdash}^{\Vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$, $\mathsf{Repr}_{\mathsf{S}}$. Then the following hold:

(1) There exists a basic Gödel sentence. (2) $\nvdash \mathsf{G}$ and $\nvdash \neg\, \mathsf{G}$ for all Gödel sentences G.

Proof. (1): Immediate from Prop. 10.
(2): $\nvdash \neg\, \mathsf{G}$ follows by applying Prop. 12 to the assumptions, so it remains to show $\nvdash \mathsf{G}$.
- From $\mathsf{OCon}_{\vdash}$, we obtain $\mathsf{Con}_{\vdash}$.
- Applying Lemma 4 to $\mathsf{Rel}_{\vdash}^{\Vdash}$ and $\mathsf{Repr}_{\Vdash}$, we obtain HBL_1.
- Applying Prop. 11 to the last two facts, we obtain $\nvdash \mathsf{G}$, as desired. □

6.3 Rosser's version

Rosser's contribution to $\mathcal{IT}_1$ was an ingenious trick for weakening the ω-consistency assumption into plain consistency—as such, it is usually seen as a *strict improvement* over Gödel's version. While this is true for the concrete case of FOL theories extending arithmetic, from an abstract perspective the situation is more nuanced: The improvement is achieved at the cost of asking more from the logic. Our framework makes this tradeoff clearly visible. The idea is to use Rosser sentences instead of Gödel sentences to "repair" the ω-consistency assumption of Theorem 13 (inherited from Prop. 12).

Prop 14. Assume $\mathsf{Con}_{\vdash}$, Ord_2, $\mathsf{Rel}_{\vdash}^{\Vdash}$, $\mathsf{Repr}_{\Vdash}$ $\mathsf{Repr}_{\neg}$ and $\mathsf{Clean}_{\Vdash}$. Then $\nvdash \neg\, \mathsf{R}$ for all Rosser sentences R.

Proof. To prove $\nvdash \neg\, \mathsf{R}$, we assume (1) $\vdash \neg\, \mathsf{R}$ and aim to reach a contradiction.
- With $\mathsf{Rel}_{\vdash}^{\Vdash}$, we obtain $p \Vdash \neg\, \mathsf{R}$ for some $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, we obtain $\vdash^{\mathsf{b}} \textcircled{\Vdash}(\langle p \rangle, \langle \neg\, \mathsf{R} \rangle)$, hence (2) $\vdash \textcircled{\Vdash}(\langle p \rangle, \langle \neg\, \mathsf{R} \rangle)$.
- From (1) and $\mathsf{Con}_{\vdash}$, we obtain $\nvdash \mathsf{R}$.
- With $\mathsf{Rel}_{\vdash}^{\Vdash}$, we obtain $q \nVdash \mathsf{R}$ for all $q \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and Lemma 3, we obtain, for all $n \in \mathsf{Num}$, $\vdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(n, \langle \mathsf{R} \rangle)$, hence (3) $\vdash \neg\, \textcircled{\Vdash}(n, \langle \mathsf{R} \rangle)$.
- By Ord_2, we obtain a finite $M \subseteq \mathsf{Num}$ such that (4) $\vdash \forall x.\ (\bigvee_{m \in M} x \equiv m) \vee \langle p \rangle \prec x$
- We prove (5) $\vdash \forall x.\ \neg\, (\textcircled{\Vdash}(x, \langle \mathsf{R} \rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle))$. The proof is performed in the intuitionistic proof system of $\vdash$, but we describe it informally: We fix x, assume $\textcircled{\Vdash}(x, \langle \mathsf{R} \rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle)$, and aim to reach a contradiction. We perform a case distinction according to (4):
  - If x equals some $m \in M$, then $\textcircled{\Vdash}(m, \langle \mathsf{R} \rangle)$, which together with (3) leads to a contradiction.
  - If $\langle p \rangle \prec x$, then from $\mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle)$ and $\textcircled{\neg}(\langle \mathsf{R} \rangle, \langle \neg\, \mathsf{R} \rangle)$ (which holds thanks to $\mathsf{Repr}_{\neg}$ and $\vdash^{\mathsf{b}}$ being included in $\vdash$), we obtain $\neg\, \textcircled{\Vdash}(\langle p \rangle, \langle \neg\, \mathsf{R} \rangle)$, which together with (2) leads to a contradiction.
  - This concludes (our informal description of) the $\vdash$-formal proof of (5).

- From (5), by (intuitionistic) logic we obtain $\vdash \neg\,(\exists x.\, \textcircled{\vdash}(x, \langle \mathsf{R}\rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R}\rangle))$.
- Thanks to R being a Rosser formula, we obtain $\vdash \mathsf{R}$.
- Together with (1), this contradicts $\mathsf{Con}_{\vdash}$. □

Thus, ω-consistency (assumption $\mathsf{OCon}_{\vdash}$) has been weakened to consistency (assumption $\mathsf{Con}_{\vdash}$), but in exchange we needed to additionally assume a special formula $\prec$ satisfying Ord_2. This represents a quite strong commitment to the arithmetical ordering.

Even worse, this fix on the assumptions needed to show the unprovability of the negated formula ($\neg$ R) complicates the proof of the unprovability of the *direct* formula (R), which was trivial in Gödel's version (Prop. 11). Now we again need a cleanly representable proof-of relation, representable negation, and well-behavedness of the order-like relation $\prec$:

Prop 15. Assume $\mathsf{Con}_{\vdash}$, Ord_1, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ $\mathsf{Repr}_{\neg}$ and $\mathsf{Clean}_{\Vdash}$. Then $\nVdash \mathsf{R}$ for all Rosser sentences R.

Proof. To prove $\nVdash \mathsf{R}$, we assume (1) $\vdash \mathsf{R}$ and aim to reach a contradiction.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $p \Vdash \mathsf{R}$ for some $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, we obtain $\vdash^{\mathsf{b}} \textcircled{\Vdash}(\langle p\rangle, \langle \mathsf{R}\rangle)$, hence (2) $\vdash \textcircled{\Vdash}(\langle p\rangle, \langle \mathsf{R}\rangle)$.
- From (1) and $\mathsf{Con}_{\vdash}$, we obtain $\nVdash \neg\, \mathsf{R}$.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $q \not\Vdash \neg\, \mathsf{R}$ for all $q \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and Lemma 3, we obtain $\vdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(n, \langle \neg\, \mathsf{R}\rangle)$ for all $n \in \mathsf{Num}$.
- With Ord_1, we obtain (3) $\vdash \forall y \prec \langle p\rangle.\, \neg\, \textcircled{\Vdash}(y, \langle \neg\, \mathsf{R}\rangle)$.
- The following reasoning is performed in the (intuitionistic) proof system of $\vdash$, but we describe it informally.
  - By $\mathsf{Repr}_{\neg}$ and the fact that $\vdash^{\mathsf{b}}$ is included in $\vdash$, the only z such that $\textcircled{\neg}(z, \langle \neg\, \mathsf{R}\rangle)$ is $\langle \neg\, \mathsf{R}\rangle$.
  - With (3), we obtain $\vdash \mathsf{RosserTwist}(\langle p\rangle, \langle \mathsf{R}\rangle)$.
- From $\vdash \mathsf{RosserTwist}(\langle p\rangle, \langle \mathsf{R}\rangle)$ and (2), we obtain $\vdash \exists x.\, \textcircled{\vdash}(x, \langle \mathsf{R}\rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R}\rangle)$.
- Since R is a Rosser sentence, from (1) we obtain $\vdash \neg\, (\exists x.\, \textcircled{\vdash}(x, \langle \mathsf{R}\rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R}\rangle))$.
- The last two facts contradict consistency. □

Theorem 16. ($\mathcal{IT}_1$ *à la* Rosser) Assume $\mathsf{Con}_{\vdash}$, Ord_1, Ord_2, $\mathsf{Repr}_{\neg}$, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$, $\mathsf{Repr}_{\mathsf{S}}$. Then the following hold:

(1) There exists a basic Rosser sentence. (2) $\nVdash \mathsf{R}$ and $\nVdash \neg\, \mathsf{R}$ for all Rosser sentences R.

Proof. (1): Immediate from Prop. 10.
(2): $\nVdash \mathsf{R}$ follows by applying Prop. 15 to the assumptions, and $\nVdash \neg\, \mathsf{R}$ follows by applying Prop. 14 to the assumptions. □

Highlighted in the statements of Theorems 16 and 13 is the assumption tradeoff between the two versions of $\mathcal{IT}_1$: Rosser's weakening of ω-consistency into consistency is paid by additionally assuming representability of negation and an order-like relation satisfying Ord_1 and Ord_2. Certainly, negation representability is not a big price, since for concrete logics this tends to be a lemma that is anyway needed when proving HBL_1. On the other hand, the ordering assumptions seem to be a significant generality gap in favor of Gödel's version.

6.4 Semantic versions

A semantic version of $\mathcal{IT}_1$ is one that establishes not only the unprovability of Gödel or Rosser sentences and of their negations, but also the truth of these sentences. To capture this abstractly, we leverage our concept of truth from Section 4.6, denoted $\models$.

Theorem 17. (Semantic $\mathcal{IT}_1$) If we enrich the assumptions of Theorem 13 with $\mathsf{LCQ}_{\models}(2,3)$ and $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, then its conclusions can be enriched with the following:

(3) $\models \mathsf{G}$ for all basic Gödel sentences G.

Proof. We know from Theorem 13 that $\not\vdash \mathsf{G}$, and $\not\vdash \neg\,\mathsf{G}$. It remains to show $\models \mathsf{G}$.
- From $\not\vdash \mathsf{G}$ and $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain that $p \not\Vdash \mathsf{G}$ for all $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$, by Lemma 3 we obtain $\vdash^{\mathsf{b}} \neg\, \textcircled{\Vdash}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models \neg\, \textcircled{\Vdash}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{LCQ}_{\models}(3)$, we obtain (i) $\models \forall x. \neg\, \textcircled{\Vdash}(x, \langle \mathsf{G} \rangle)$.
- By logic we obtain $\vdash^{\mathsf{b}} (\forall x. \neg\, \textcircled{\Vdash}(x, \langle \mathsf{G} \rangle)) \to \neg\, \textcircled{\vdash}\langle \mathsf{G} \rangle$. (Recall Convention 5.)
- With the definition of basic Gödel sentence, by logic we obtain $\vdash^{\mathsf{b}} (\forall x. \neg\, \textcircled{\Vdash}(x, \langle \mathsf{G} \rangle)) \to \mathsf{G}$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models (\forall x. \neg\, \textcircled{\Vdash}(x, \langle \mathsf{G} \rangle)) \to \mathsf{G}$.
- With $\mathsf{LCQ}_{\models}(2)$ and (i), we obtain $\models \mathsf{G}$, as desired. □

The next variant of the semantic $\mathcal{IT}_1$ does not directly assume the existence of proofs and their representations, but "recovers" them using $\mathsf{HBL}^{\Leftarrow}_1$ as prescribed in Lemma 7:

Theorem 18. (Semantic $\mathcal{IT}_1$, second variant) The conclusions of Theorem 17 remain true if we replace its assumptions $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ with the assumptions $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{Compl}_{\neg\mathsf{Pf}}$, $\mathsf{HBL}^{\Leftarrow}_1$, $\mathsf{LCQ}_{\models}(4,5)$.

Proof. Immediate by Lemma 7 and Theorem 17, noting that, by Lemma 4, HBL_1 (which is needed by Lemma 7) is implied by $\mathsf{Rel}^{\Vdash}_{\vdash}$ and $\mathsf{Repr}_{\Vdash}$. □

Similar semantic theorems can be obtained for Rosser-style $\mathcal{IT}_1$:

Theorem 19. (Semantic $\mathcal{IT}_1$ *à la* Rosser) If we enrich the assumptions of Theorem 16 with $\mathsf{LCQ}_{\models}(2,3)$ and $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, then its conclusions can be enriched with the following:

(3) $\models \mathsf{R}$ for all basic Rosser sentences R.

Proof. Exactly the same as the proof of Theorem 17, but using Rosser sentences and applying Theorem 16 (rather than using Gödel sentences and applying Theorem 13). Note that the last part of the proof of $\models \mathsf{G}$ also works for R, because $\vdash^{\mathsf{b}} (\neg\, \textcircled{\vdash}\langle \mathsf{R} \rangle) \to \mathsf{R}$ follows from the definition of Rosser sentence (by logic). □

Theorem 20. (Semantic $\mathcal{IT}_1$ *à la* Rosser, second variant) The conclusions of Theorem 19 remain true if we replace its assumptions $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ with the assumptions $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{Compl}_{\neg\mathsf{Pf}}$, $\mathsf{HBL}^{\Leftarrow}_1$, $\mathsf{LCQ}_{\models}(4,5)$.

Proof. The same as Theorem 18's proof, but using Theorem 19 rather than Theorem 17. □

The assumption tradeoff between Theorems 17 and 18 on the one hand and Theorems 19 and 20 on the other hand is the same as that between their proof-theoretic counterparts (discussed in Section 6.3): $\mathsf{OCon}_{\vdash}$ on the Gödel side versus $\mathsf{Con}_{\vdash}$, Ord_1, Ord_2 and $\mathsf{Repr}_{\neg}$ on the Rosser side. An interesting phenomenon arises when $\vdash^{\mathsf{b}}$ and $\vdash$ are the same relation. Then soundness implies ω-consistency under reasonable assumptions:

Lemma 21. Assume $\vdash^{\mathsf{b}} = \vdash$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\mathsf{LCQ}_{\models}(1,2,3)$. Then $\mathsf{OCon}_{\vdash}$ holds.

Proof. Assume $\vdash \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\vdash^{\mathsf{b}} = \vdash$, we obtain $\models \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{LCQ}_{\models}(3)$, we obtain $\models \forall x. \neg\, \varphi(x)$.

- By logic and $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models (\forall x.\, \neg\, \varphi(x)) \rightarrow \neg\, (\exists x.\, \varphi(x))$.
- From the last two facts and $\mathsf{LCQ}_{\models}(2)$, we obtain $\models \neg\, (\exists x.\, \varphi(x))$.
- With $\mathsf{LCQ}_{\models}(1,2)$, we obtain $\not\models \neg\, \neg\, (\exists x.\, \varphi(x))$.
- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\vdash^{\mathsf{b}} = \vdash$, we obtain $\nvdash \neg\, \neg\, (\exists x.\, \varphi(x))$, as desired. □

Thus, if $\vdash^{\mathsf{b}} = \vdash$ and some reasonable properties hold for $\models$, then ω-consistency comes for free. Hence, in this case Gödel's versions, Theorems 17 and 18, are *strictly more general* than Rosser's versions, Theorems 19 and 20 (if we ignore the difference in the way Gödel and Rosser sentences are actually defined). This further illustrates the idea that Rosser's trick is not always an improvement.

6.5 Classical logic versions

The results so far do not require going beyond intuitionistic logic. But if we commit to classical logic for $\vdash$ (i.e., assume $\vdash \neg\, \neg\, \varphi \rightarrow \varphi$) and also assume $\mathsf{HBL}^{\Leftarrow}_1$, there is a well-known more direct argument for showing that Gödel sentences are not disprovable, which immediately proves $\mathcal{IT}_1$. (This is documented, for example, as Theorem 3.1 in Buldt's monograph [7].) However, in our generalized setting with two provability relations, this argument does not go through unless we strengthen $\mathsf{HBL}^{\Leftarrow}_1$ (which currently refers to $\vdash^{\mathsf{b}}$) to refer to $\vdash$:

$\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$: $\vdash \textcircled{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.

Theorem 22. (Classical $\mathcal{IT}_1$) Assume classical logic for $\vdash$, $\mathsf{Con}_{\vdash}$, HBL_1, $\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$, $\mathsf{Repr}_{\mathsf{S}}$. Then the following hold:

(1) There exists a basic Gödel sentence. (2) $\nvdash \mathsf{G}$ and $\nvdash \neg\, \mathsf{G}$ for all Gödel sentences G.

Proof. (1): Immediate from $\mathsf{Repr}_{\mathsf{S}}$ by Prop. 10.
(2): Let G be a Gödel sentence.
- From $\mathsf{Con}_{\vdash}$ and HBL_1, by Prop. 11 we obtain $\nvdash \mathsf{G}$.
- So we are left to prove $\nvdash \neg\, \mathsf{G}$. To this end, we assume (i) $\vdash \neg\, \mathsf{G}$ and aim to reach a contradiction.
- Since G is a Gödel sentence, by logic we obtain $\vdash \neg\, \neg\, \textcircled{\vdash}\langle \mathsf{G}\rangle$.
- By classical logic, from this we obtain $\vdash \textcircled{\vdash}\langle \mathsf{G}\rangle$.
- With $\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$, we obtain $\vdash \mathsf{G}$.
- With (i), this contradicts $\mathsf{Con}_{\vdash}$. □

Point (2) of the above theorem refers to Gödel sentences (defined using $\vdash$). Note that weakening the statement to refer to basic Gödel sentences (defined using $\vdash^{\mathsf{b}}$) would not help with relaxing the assumption $\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$ to $\mathsf{HBL}^{\Leftarrow}_1$; the former would still be needed to finish the proof. Of course, $\mathsf{HBL}^{\Leftarrow}_1$ and $\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$ coincide in the important case when $\vdash^{\mathsf{b}} = \vdash$.

Two semantic versions are possible for classical $\mathcal{IT}_1$. The first one additionally assumes some reasonable properties of $\models$, soundness for $\vdash^{\mathsf{b}}$, and $\mathsf{TIP}^{\vdash}_{\models}$:

Theorem 23. (Classical Semantic $\mathcal{IT}_1$) If we enrich the assumptions of Theorem 22 with $\mathsf{LCQ}_{\models}(1,2,5)$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, $\mathsf{TIP}^{\vdash}_{\models}$, then its conclusions can be enriched with the following:

(3) $\models \mathsf{G}$ for all basic Gödel sentences G.

Proof. We know from Theorem 22 that (i) $\nvdash \mathsf{G}$, and $\nvdash \neg\, \mathsf{G}$. It remains to show $\models \mathsf{G}$. To this end, we assume (ii) $\not\models \mathsf{G}$ and try to reach a contradiction.
- From (ii), by $\mathsf{LCQ}_{\models}(5)$ we obtain (iii) $\models \neg\, \mathsf{G}$.
- From the basic Gödel sentence definition we obtain $\vdash^{\mathsf{b}} \neg\, \mathsf{G} \rightarrow \neg\, \neg\, \textcircled{\vdash}\langle \mathsf{G}\rangle$.

- With $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$, we obtain $\models \neg\, \mathsf{G} \rightarrow \neg\, \neg\, \textcircled{\vdash}\langle \mathsf{G}\rangle$.
- With (iii), by $\mathsf{LCQ}_{\models}(2)$ we obtain $\models \neg\, \neg\, \textcircled{\vdash}\langle \mathsf{G}\rangle$.
- With $\mathsf{LCQ}_{\models}(1,2)$, we obtain $\not\models \neg\, \textcircled{\vdash}\, \langle \mathsf{G}\rangle$.
- With $\mathsf{LCQ}_{\models}(5)$, we obtain $\models \textcircled{\vdash}\langle \mathsf{G}\rangle$.
- With $\mathsf{TIP}^{\vdash}_{\models}$, we obtain $\vdash \mathsf{G}$, which contradicts (i). □

The second one replaces $\mathsf{TIP}^{\vdash}_{\models}$ with some assumptions that, in the presence of the others, ensure $\mathsf{TIP}^{\vdash}_{\models}$—hence is strictly less general than the first one:

Theorem 24. (Classical Semantic $\mathcal{IT}_1$, second version) The conclusions of Theorem 23 still hold if we replace $\mathsf{TIP}^{\vdash}_{\models}$ with the assumptions $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$ and $\mathsf{LCQ}_{\models}(4)$.

Proof. It suffices to show that $\mathsf{TIP}^{\vdash}_{\models}$ follows from its replacements and the other assumptions. We do this using Lemma 8(2). To apply this lemma, we need:
- $\mathsf{Rel}^{\mathsf{Pf}}_{\textcircled{\vdash}}$, $\mathsf{Compl}_{\mathsf{Pf}}$, $\mathsf{LCQ}_{\models}(4)$, which are assumed above;
- $\mathsf{LCQ}_{\models}(2)$, $\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$ and $\mathsf{HBL}^{\Leftarrow}_{1}$, which are assumptions of Theorem 23.

So from the lemma we infer $\mathsf{TIP}^{\vdash}_{\models}$, as desired. □

We used Gödel, not Rosser sentences in our classical semantic versions of $\mathcal{IT}_1$. Unlike for the (intuitionistic) semantic versions in Section 6.4, here a Rosser-style improvement would serve no purpose, since we already assume $\vdash$ to be consistent, not ω-consistent.

6.6 Benefits of the two-relation take on provability

Our framework distinguishes between basic provability ($\vdash^{\mathsf{b}}$) and provability ($\vdash$). This seems to be a rational design choice when aiming high in terms of generality for the incompleteness theorems. For example, this choice has been made explicitly by Smorynski [57] and more implicitly by Feferman [13] in their general accounts. Let us analyze what are the choice's benefits to $\mathcal{IT}_1$ in the context of our development. The main questions are of course whether the scope of these theorems has to gain from the two-relation approach, as opposed to working with only one relation; and, if so, by how much.

In some cases, the gain is undeniable: Our Section 6.4's semantic Theorems 17–20 gain significant generality by assuming soundness for $\vdash^{\mathsf{b}}$ only, and merely consistency or ω-consistency for $\vdash$. This covers the case of Gödel or Rosser sentences being true for unsound theories as well. And of course the above theorems are based on the proof-theoretic theorems in Sections 6.2 and 6.3, which means that the latter's two-relation formulations are also needed.

At the other extreme, in one case, namely the classical-logic-based Theorem 22, there is no gain. Indeed, say we ignore $\vdash^{\mathsf{b}}$ and modify all this theorem's assumptions to replace $\vdash$ for all occurrences of $\vdash^{\mathsf{b}}$—which is the same as assuming $\vdash^{\mathsf{b}} = \vdash$. Then we would lose no generality, because the modified assumptions would be *the same or weaker* than the original assumptions. In conclusion, Theorem 22 stays equally general if we identify $\vdash^{\mathsf{b}}$ and $\vdash$.

The other cases, namely the classical-semantic Theorems 23 and 24, are somewhere in between these two extremes: Their two-relation formulation is more general than a one-relation formulation, but the gain from this is doubtful. Like in Theorems 17–20, they allow an unsound $\vdash$ as an extension of a sound $\vdash^{\mathsf{b}}$. On the other hand, their assumptions HBL_1 and $\mathsf{HBL}^{\Leftarrow}_{1,\vdash}$ (inherited from Theorem 22) force $\vdash$ to coincide with $\vdash^{\mathsf{b}}$ on all sentences of the form $\textcircled{\vdash}\langle\varphi\rangle$; and it is not clear if one can find interesting classes of unsound relations $\vdash$ that satisfy this constraint (for standard choices of $\vdash^{\mathsf{b}}$).

7 Second Incompleteness Theorem

For a consistent theory that is sufficiently expressive, $\mathcal{IT}_2$ states that this theory cannot prove (the internal formalization of) its own consistency, which in our notations will be written as $\nvdash \neg \textcircled{\vdash}\langle\bot\rangle$. Here, "sufficient expressiveness" refers to something similar to the case of $\mathcal{IT}_1$, namely the theory's ability to express concepts about itself such as formulas and provability, but is a stronger requirement than for $\mathcal{IT}_1$: For $\mathcal{IT}_2$, the theory needs to be expressive enough to *formalize and prove part of* $\mathcal{IT}_1$. This includes Peano arithmetic and stronger theories but excludes Robinson arithmetic.

$\mathcal{IT}_2$ is of course a perfectly mathematical theorem, just like $\mathcal{IT}_1$. However, the informal paraphrasing of $\mathcal{IT}_2$'s conclusion, taking $\nvdash \neg \textcircled{\vdash}\langle\bot\rangle$ to mean that the theory cannot prove *its* own consistency, relies on an extra-mathematical assumption of an intensional nature [1] [13, §1]: that $\textcircled{\vdash}$ *adequately expresses the provability relation* $\vdash$. The mathematical property of $\textcircled{\vdash}$ (weakly) representing $\vdash$ is only an extensional approximation of this assumption. By contrast, $\mathcal{IT}_1$ only needs $\textcircled{\vdash}$ as an auxiliary concept used in its proof; the adequate expression of $\vdash$ is irrelevant there, and it is only (weak) representability that matters. When discussing variants of $\mathcal{IT}_2$, we will always work under the adequate expression assumption.

7.1 Informal account and roadmap

Similarly to the case of $\mathcal{IT}_1$, we start with an informal account of the argument behind $\mathcal{IT}_2$, where again we use double quotes for sentences that internalize certain statements in the language of the considered system.

(1) Gödel realized that $\mathcal{IT}_2$ follows by internally formalizing the positive half of his (proof-theoretic) $\mathcal{IT}_1$, henceforth referred to as $\mathcal{IT}_{0.5}$. It states the unprovability of a Gödel sentence G, covered by Section 6.1's point (1.1) and Prop. 11. This leads to the provability of "the theory is consistent implies that G is not provable". Moreover, by virtue of G being a Gödel sentence, $\mathcal{IT}_{0.5}$ itself implies the unprovability of "G is not provable". From the above together with consistency, we obtain the unprovability of "the theory is consistent".

The three derivability conditions HBL_{1-3} recalled in Section 4.5 were perfected by Löb [32] based on previous work by Hilbert and Bernays [22] to make the above informal argument fully rigorous without referring to internal formalization details (although such details do need to be worked out to prove the conditions). The way these conditions work together to achieve this goal will be discussed in Subsection 7.2. For now, we should just note that the unqualified requirement of internally formalizing $\mathcal{IT}_{0.5}$ is in itself not sufficient. The internalized concepts must exhibit certain similarities to the original concepts from one level up; and this is what the derivability conditions express. For example, the above informal argument had a silent shift from the provability of "the theory is consistent implies that G is not provable" (with the whole statement inside quotes) to the provability of "the theory is consistent" implies "G is not provable" (where the implication operator is outside the quotes, i.e., is positioned one level up)—which is where HBL_2 comes to help.

(2) An alternative line of reasoning due to Jeroslow [24] is often cited [50, 56, 57] as a simplification of the canonical route to prove $\mathcal{IT}_2$: Whereas traditionally $\mathcal{IT}_2$ requires all three derivability conditions, Jeroslow's version does not make use of HBL_2.

Jeroslow's approach relies on pseudo-terms. These are formulas that satisfy existence and uniqueness properties on one of their free variables, say, x, meaning that x denotes a uniquely identified item depending on any items denoted by the other free variables; in

short, pseudo-terms can essentially be treated like terms. In the informal discussion that follows, the reader is free to think of actual terms instead of pseudo-terms.

Jeroslow proved an alternative diagonalization lemma, producing pseudo-term fixpoints instead of formula fixpoints. In particular, one obtains a pseudo-term τ that is provably equal to the encoding of the sentence "non-τ is provable". If we let φ denote the latter sentence, we obtain that φ is provably equivalent to "$\neg\,\varphi$ is provable". Let us call any sentence satisfying this fixpoint property a *Jeroslow sentence*. Such a sentence states about itself something stronger-sounding than a Gödel (or Rosser) sentence: not that it is merely not provable, but that even its negation is provable. (We write "stronger-sounding" rather than "stronger" because it would be actually stronger only assuming the *provability* of consistency.)

Now, the argument for $\mathcal{IT}_2$ goes as follows. Assume that "the theory is consistent" is provable. Because a Jeroslow sentence φ asserts the provability of something, (a slightly stronger form of) HBL_3 applies, so φ provably implies "φ is provable". On the other hand, by virtue of being a Jeroslow sentence, φ also provably implies "$\neg\,\varphi$ is provable". So φ provably implies "the theory is inconsistent", which together with our assumption gives the provability of $\neg\,\varphi$. With HBL_1, we obtain the provability of "$\neg\,\varphi$ is provable", i.e, by virtue of φ being a Jeroslow sentence, the provability of φ. So both φ and its negation are provable, which contradicts consistency.

The above argument invokes HBL_1 and HBL_3 but not HBL_2. It is specific to Jeroslow sentences and cannot be achieved with Gödel or Rosser sentences. The argument has several loose ends, which will be addressed in our formal discussion. In light of that, it will become clear that the $\neg$ in "$\neg\,\varphi$" and the "non" in "non-τ" are different, but related operators: The former is formula negation (applied to φ), while the latter is substitution (with τ) in a pseudo-term that represents the operator on numerals corresponding to $\neg$ via formula encoding.

This concludes our informal discussion. Next, we engage in formal accounts of the above arguments: point (1) in Subsection 7.2 and point (2) in Subsection 7.3.

7.2 Standard version

Let us slightly rephrase the statement and proof of $\mathcal{IT}_{0.5}$ (Prop. 11) in a way that will make it convenient to highlight its internal formalization within the proof of $\mathcal{IT}_2$:

Prop. 11 (rephrased). Assume HBL_1. Let G be a Gödel sentence. Then $\mathsf{Con}_{\vdash}$ implies $\nvdash \mathsf{G}$.

Proof. Step 1. Since G is a Gödel sentence, $\vdash \mathsf{G}$ implies $\vdash \neg\,\textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle$.
Step 2. By $\vdash^{\mathsf{b}} \subseteq \vdash$ and HBL_1, $\vdash \mathsf{G}$ implies $\vdash \textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle$.
Step 3. By *modus ponens* (since $\neg\mathsf{G}$ is $\mathsf{G} \to \bot$), $\vdash \neg\,\textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle$ and $\vdash \textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle$ implies $\vdash \bot$.
Step 4. From the last three facts, $\vdash \mathsf{G}$ implies $\vdash \bot$.
Step 5. Hence $\mathsf{Con}_{\vdash}$ (i.e., $\nvdash \bot$) implies $\nvdash \mathsf{G}$, as desired. □

The standard proof of $\mathcal{IT}_2$ uses all three derivability conditions in key places in order to internalize the above proof of $\mathcal{IT}_{0.5}$:

Theorem 25. ($\mathcal{IT}_2$) Assume $\mathsf{Con}_{\vdash}$, HBL_1, HBL_2, HBL_3 and $\mathsf{Repr}_{\mathsf{S}}$. Then $\nvdash \neg\,\textcircled{\scriptsize$\vdash$}\langle\bot\rangle$.

Proof. Let G be a Gödel sentence, which exists by Prop. 10 and $\mathsf{Repr}_{\mathsf{S}}$.
Internalizing the proof of $\mathcal{IT}_{0.5}$:
- Step 1. Since G is a Gödel sentence, we obtain $\vdash \mathsf{G} \to \neg\,\textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle$.
With HBL_1 and *modus ponens*, we obtain $\vdash^{\mathsf{b}} \textcircled{\scriptsize$\vdash$}\langle\mathsf{G} \to \neg\,\textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle\rangle$.
With HBL_2, we obtain $\vdash^{\mathsf{b}} \textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle \to \textcircled{\scriptsize$\vdash$}\langle\neg\,\textcircled{\scriptsize$\vdash$}\langle\mathsf{G}\rangle\rangle$.

- Step 2. From HBL_3, we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle \mathsf{G}\rangle \to \textcircled{\vdash}\langle\textcircled{\vdash}\langle \mathsf{G}\rangle\rangle$.
- Step 3. From HBL_2, we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle\neg\textcircled{\vdash}\langle \mathsf{G}\rangle\rangle \wedge \textcircled{\vdash}\langle\textcircled{\vdash}\langle \mathsf{G}\rangle\rangle \to \textcircled{\vdash}\langle\bot\rangle$.
- Step 4. From the last three facts, we obtain $\vdash^{\mathsf{b}} \textcircled{\vdash}\langle \mathsf{G}\rangle \to \textcircled{\vdash}\langle\bot\rangle$.
- Step 5. This implies $\vdash^{\mathsf{b}} \neg\textcircled{\vdash}\langle\bot\rangle \to \neg\textcircled{\vdash}\langle \mathsf{G}\rangle$, hence (1) $\vdash \neg\textcircled{\vdash}\langle\bot\rangle \to \neg\textcircled{\vdash}\langle \mathsf{G}\rangle$.

Invoking $\mathcal{IT}_{0.5}$:
- From $\mathsf{Con}_{\vdash}$ and HBL_1, by Prop. 11 we obtain $\nvdash \mathsf{G}$.
- With the Gödel sentence definition, we obtain (2) $\nvdash \neg\textcircled{\vdash}\langle \mathsf{G}\rangle$.

Putting the two together:
- From (1) and (2), we obtain $\nvdash \neg\textcircled{\vdash}\langle\bot\rangle$, as desired. □

The above proof of $\mathcal{IT}_2$ starts with an internalization of aspects of the $\mathcal{IT}_{0.5}$'s proof. It does not literally formalize the end-to-end proof, but instead proceeds by plugging in the derivability conditions, which can be thought of as pre-formalized reasoning patterns.

- Step 2 is internalized using HBL_3, which asserts the provability of some instances of HBL_1, replacing object-level quantification with meta-level quantification. To see this, note that a full formalization of HBL_1 would be a sentence of the form $\forall x.\ \textcircled{\mathsf{Sen}}(x) \wedge \textcircled{\vdash}(x) \to \textcircled{\vdash}(\textcircled{\mathsf{inst}}(\textcircled{\textcircled{\vdash}}, \textcircled{\langle\_\rangle}(x)))$, where $\textcircled{\mathsf{Sen}}$, $\textcircled{\langle\_\rangle}$, $\textcircled{\mathsf{inst}}$ and $\textcircled{\textcircled{\vdash}}$ formalize membership to the set of sentences Sen, the encoding operator $\langle\_\rangle$, the formula-instantiation (i.e., substitution of a term for the first variable, v_1) operator, and the inner representation of provability $\textcircled{\vdash}$ (one further level inside), respectively. By instantiating the $\forall$-quantified x with $\langle\varphi\rangle$ for any $\varphi \in \mathsf{Sen}$, we obtain sentences that can be equivalently written in a more palatable from, $\textcircled{\vdash}\langle\varphi\rangle \to \textcircled{\vdash}\langle\textcircled{\vdash}\langle\varphi\rangle\rangle$, which are exactly the sentences whose provability is asserted by HBL_3.
- Similarly, Step 3 is internalized using HBL_2, which asserts the provability of some instances of the *modus ponens* rule, again replacing object-level quantification with meta-level quantification—whereas a full formalization of HBL_1 would be a sentence of the form $\forall x, y.\ \textcircled{\mathsf{Sen}}(x) \wedge \textcircled{\mathsf{Sen}}(y) \wedge \textcircled{\vdash}\langle x\rangle \wedge \textcircled{\vdash}\langle x \to y\rangle \to \textcircled{\vdash}\langle y\rangle$.
- The internalization of Step 1 is more interesting: To formalize the fact that $\vdash \mathsf{G}$ implies $\vdash \neg\textcircled{\vdash}\langle \mathsf{G}\rangle$, one takes advantage of the availability of the stronger and "more formal" property $\vdash \mathsf{G} \to \neg\textcircled{\vdash}\langle \mathsf{G}\rangle$, which is pushed inside the proof system via HBL_1, and then its implication is lifted one level up using HBL_2.
- Steps 4 and 5 are internalized by mapping meta-implication and meta-negation to the implication and negation operators, $\to$ and $\neg$, using the latter's deductive properties.

In summary, a judicious use of the derivability conditions and other *ad hoc* procedures are used to prove an internalized version of $\mathcal{IT}_{0.5}$, while avoiding the need to fully formalize the proof inside the system. (On the other hand, proving the derivability conditions does require a substantial internal formalization effort in the first place.)

Theorem 25's proof is concluded according to the plan sketched in Section 7.1: by combining the formalized and the original $\mathcal{IT}_{0.5}$ to obtain the unprovability of consistency.

Finally, let us scrutinize $\mathcal{IT}_2$ with respect to the benefit of the two-relation take on provability (as was done for $\mathcal{IT}_1$ in Section 6.6). We see that for $\mathcal{IT}_2$ there is no benefit from using two relations. The same reason as the one discussed for Theorem 22 applies: Replacing $\vdash^{\mathsf{b}}$ with $\vdash$ does not decrease generality. Thus, when discussing $\mathcal{IT}_2$, we can assume $\vdash^{\mathsf{b}} = \vdash$ without loss of generality. Note also that, even if we used a formula $\textcircled{\vdash^{\mathsf{b}}}$ corresponding $\vdash^{\mathsf{b}}$, no meaningful two-relation strengthening of $\mathcal{IT}_2$ would be in sight; in particular, the consistency of the basic theory $\vdash^{\mathsf{b}}$ could well be provable in the extended theory $\vdash$.

Convention 26. For the rest of Section 7, we will assume $\vdash^{\mathsf{b}} = \vdash$ and no longer refer to $\vdash^{\mathsf{b}}$.

Next we study Jeroslow's approach to $\mathcal{IT}_2$ [24]. To analyze its features and pitfalls, we need to recall into some notions and notations employed by Jeroslow.

A *pseudo-term* is a formula $\varphi \in \mathsf{Fmla}_{m+1}$ expressing a provably functional relation via "exists unique": $\vdash \forall x_1, \ldots, x_m.\ \exists! y.\ \varphi(x_1, \ldots, x_m, y)$. Note that we have already seen examples of pseudo-terms: Section 4.4's formulas $\textcircled{f}$ representing functions f.

We let PTerm, ranged over by σ, τ, be the set of pseudo-terms. While pseudo-terms are particular formulas, they will be treated as an extension of the notion of term. Indeed, a term t having free variables $v_1, \ldots, v_m$ can be regarded as the pseudo-term $v_{m+1} \equiv t$.

Let $\sigma \in \mathsf{Fmla}_{m+1}$ be a pseudo-term. Whereas $\mathsf{FVars}(\sigma) = \{v_1, \ldots, v_{m+1}\}$, the free variables of σ *as a pseudo-term*, written $\mathsf{FVarsP}(\sigma)$, will be $\{v_1, \ldots, v_m\}$; in this case, we will also write $\sigma \in \mathsf{PTerm}_m$. A pseudo-term σ is *closed* if $\mathsf{FVarsP}(\sigma) = \emptyset$, i.e., $\sigma \in \mathsf{PTerm}_0$.

Pseudo-terms can be composed freely with terms and other pseudo-terms in a term-like fashion, and also substituted in formulas, as indicated in the following notation.

Notation 27. Given $\sigma \in \mathsf{PTerm}_1$, $\tau \in \mathsf{PTerm}_m$, $t \in \mathsf{Term}$, and $\varphi \in \mathsf{Fmla}_1$, we write:

(1) $\sigma(t)$ instead of $\sigma(t, v_2)$, assuming $v_2 \notin \mathsf{FVars}(t)$ (note that $\sigma(t)$ is closed if t is closed);
(2) $\tau \equiv t$ instead of $\tau(v_1, \ldots, v_m, t)$;
(3) $\varphi(\tau)$ instead of $\exists y.\ \tau(v_1, \ldots, v_m, y) \wedge \varphi(y)$, which thanks to the pseudo-term property is provably equivalent to $\forall y.\ \tau(v_1, \ldots, v_m, y) \to \varphi(y)$;
(4) $\sigma(\tau)$ instead of $\exists y.\ \tau(v_1, \ldots, v_m, y) \wedge \sigma(y, v_{m+1})$, which again is provably equivalent to $\forall y.\ \tau(v_1, \ldots, v_m, y) \to \sigma(y, v_{m+1})$; note that $\sigma(\tau) \in \mathsf{PTerm}_m$.

Above, y is chosen to be distinct from the other occurring variables. It is possible to introduce multi-input extensions of this notation, but we will not need them. The notation smoothly integrates pseudo-terms with terms, as shown in the following example properties:

Example 28. (1) If $\vdash \sigma \equiv t$ (employing point (1) of the notation) and $\vdash \varphi(\sigma)$ (employing point (3)) then $\vdash \varphi(t)$, where $\varphi(t)$ is the usual instance of φ with t.
(2) If $\varphi \in \mathsf{Fmla}_1$, $\sigma \in \mathsf{PTerm}_1$ and $\tau \in \mathsf{PTerm}_0$, then $\vdash \varphi(\sigma)(\tau) \leftrightarrow \varphi(\sigma(\tau))$. Indeed:
  – On the left of $\leftrightarrow$, we use point (3) for φ and σ, which expands $\varphi(\sigma)$ to $\exists y.\ \sigma(v_1, y)$. Then, we use point (3) for $\varphi(\sigma)$ and τ, which yields $\exists y.\ \tau(y) \wedge \varphi(\sigma)(y)$. Combining the two, we obtain that $\varphi(\sigma)(\tau)$ abbreviates $\exists y.\ \tau(y) \wedge (\exists y'.\ \sigma(y, y') \wedge \varphi(y'))$.
  – On the right, we use point (4) for σ and τ, which expands $\sigma(\tau)$ to $\exists y.\ \tau(y) \wedge \sigma(y, v_1)$. Then we use point (3) for φ and $\sigma(\tau)$, which yields $\exists y.\ \sigma(\tau)(y) \wedge \varphi(y)$. Combining the two, we obtain that $\varphi(\sigma(\tau))$ abbreviates $\exists y.\ (\exists y'.\ \tau(y') \wedge \sigma(y', y)) \wedge \varphi(y)$.

Jeroslow fixes an abstract class of "computable" m-ary functions, $\mathcal{F}_m \subseteq \mathsf{Num}^m \to \mathsf{Num}$, for all arities $m \in \mathbb{N}$, on which he considers the following assumptions:

$\mathsf{Repr}_{\mathcal{F}}$: Every $f \in \mathcal{F}_m$ is represented by some pseudo-term $\textcircled{f} \in \mathsf{PTerm}_m$ under the identity encoding $\mathsf{Num} \to \mathsf{Num}$.
CapN: Some $\mathsf{N} \in \mathcal{F}_1$ correctly captures negation: $\mathsf{N}\langle\varphi\rangle = \langle\neg\,\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
CapSS: Some $\mathsf{ssub} : \mathsf{Fmla}_1 \to \mathcal{F}_1$ correctly captures substituted self-substitution: $\mathsf{ssub}\,(\psi)\,\langle\textcircled{f}\rangle = \langle \psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$ for all $\psi \in \mathsf{Fmla}_1$ and $f \in \mathcal{F}_1$.[1]

Note that, in CapSS, we take advantage of the introduced notation for pseudo-terms: If we spell out Notation 27(2), the highlighted text denotes $\exists y.\ \textcircled{f}(\langle\textcircled{f}\rangle, y) \wedge \psi(y)$. Moreover, employing Notation 27(1), the statement of $\mathsf{Repr}_{\mathcal{F}}$ for some $f \in \mathcal{F}_1$ and $n \in \mathsf{Num}$ would

[1] For the proof of $\mathcal{IT}_2$, we will not need this to work for arbitrary formulas ψ in Fmla_1, but only for $\textcircled{\mathsf{N}}$. However, we will not delve into such micro-optimizations; see also footnote 2.

be written as $\vdash \textcircled{f}(n) \equiv f(n)$; and combining CapN with the instance of $\mathsf{Repr}_{\mathcal{F}}$ for N, we obtain a fact that, using the same notation, can be written as $\vdash \textcircled{N}\langle\varphi\rangle \equiv \langle\neg\varphi\rangle$.

When our logical theory is a recursive extension of Robinson arithmetic and $\mathsf{Num} = \mathbb{N}$, $\mathcal{F}_m$ could be any sufficiently rich the set of m-ary computable functions, ranging from the primitive recursive functions to all total μ-recursive functions. Then, every $f \in \mathcal{F}_m$ would indeed be represented by a formula $\textcircled{f}$. Moreover, assuming a computable and injective encoding of formulas, $\langle\_\rangle : \mathsf{Fmla}_1 \to \mathbb{N}$, we can take $\mathsf{N} : \mathbb{N} \to \mathbb{N}$ to be the following computable function: Given input n, it checks if n has the form $\langle\varphi\rangle$; if so, it returns $\langle\neg\varphi\rangle$; if not, it returns any value (e.g., 0). And $\mathsf{ssub}\,(\psi)$ can be defined similarly, obtaining the desired property for every $\varphi \in \mathsf{Fmla}_2$, not necessarily of the form $\textcircled{f}$. In short, Jeroslow's assumptions cover arithmetic (but also potentially many other systems).

The heart of Jeroslow's approach lies in his diagonalization lemma, which offers pseudo-term fixpoints, and from them formula fixpoints as well:

Lemma 29. Assume CapSS and $\mathsf{Repr}_{\mathcal{F}}$ and let $\psi \in \mathsf{Fmla}_1$. Then there exists a closed pseudo-term τ such that $\vdash \tau \equiv \langle\psi(\tau)\rangle$. Moreover, taking $\varphi = \psi(\tau)$, we have $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$.

Proof. Let $f = \mathsf{ssub}\,(\psi)$ and $\tau = \textcircled{f}\langle\textcircled{f}\rangle$. From CapSS, we obtain $f\langle\textcircled{f}\rangle = \langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$. With $\mathsf{Repr}_{\mathcal{F}}$, we obtain $\vdash \textcircled{f}\langle\textcircled{f}\rangle \equiv \langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$, i.e., $\vdash \tau \equiv \langle\psi(\tau)\rangle$. By logic, from this we obtain $\vdash \psi(\tau) \leftrightarrow \psi(\langle\psi(\tau)\rangle)$, i.e., $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$. □

Lemma 29 can be used to produce Gödel and Rosser sentences, which can be used like in Sections 6, leading to variants of $\mathcal{IT}_1$. However, as discussed in Section 7.1, Jeroslow's main innovation affects $\mathcal{IT}_2$: It removes from the assumptions the second derivability condition, HBL_2.

Theorem 30. ($\mathcal{IT}_2$ *à la* Jeroslow) Assume $\mathsf{Con}_{\vdash}$, HBL_1, SHBL_3, $\mathsf{Repr}_{\mathcal{F}}$, CapN, CapSS, SHBL_3 denotes the condition: $\vdash \textcircled{\vdash}(\tau) \to \textcircled{\vdash}\langle\textcircled{\vdash}(\tau)\rangle$ for all closed pseudo-terms τ.

Then $\nvdash \mathsf{jcon}$, where jcon denotes $\forall x.\, \neg\,(\textcircled{\vdash}(x) \wedge \textcircled{\vdash}(\textcircled{N}(x)))$.

As with Rosser's trick, we analyze this innovation's tradeoffs from an abstract perspective. A first tradeoff is in the employment of a stronger version of the third condition, SHBL_3, holding for all closed pseudo-terms and not only those that encode sentences.

Another tradeoff is in the way consistency is expressed in the logic. Jeroslow does not conclude $\nvdash \neg\,\textcircled{\vdash}\langle\bot\rangle$, but something more elaborate, namely $\nvdash \mathsf{jcon}$. While the formula $\neg\,\textcircled{\vdash}\langle\bot\rangle$ internalizes the statement $\nvdash \bot$, jcon internalizes the equivalent statement "for all φ, it is not the case that $\vdash \varphi$ and $\vdash \neg\varphi$." But are the internalizations themselves equivalent, i.e., is it the case that $\vdash \neg\,\textcircled{\vdash}\langle\bot\rangle$ iff $\vdash \mathsf{jcon}$? This surely holds for many concrete logics, but it is only one direction that we can infer logic-independently, under mild assumptions:

Prop 31. Assume HBL_1, $\mathsf{Repr}_{\mathcal{F}}$, CapN. Then $\vdash \mathsf{jcon}$ implies $\vdash \neg\,\textcircled{\vdash}\langle\bot\rangle$.

Proof. Assume $\vdash \mathsf{jcon}$.
- Instantiating jcon with $\langle\bot\rangle$, we obtain $\vdash \neg\,(\textcircled{\vdash}\langle\bot\rangle \wedge \textcircled{\vdash}(\textcircled{N}\langle\bot\rangle))$
- From $\mathsf{Repr}_{\mathcal{F}}$ and CapN, we obtain $\vdash \textcircled{N}\langle\bot\rangle \equiv \langle\neg\bot\rangle$.
- From the last two facts, by logic we obtain $\vdash \neg\,(\textcircled{\vdash}\langle\bot\rangle \wedge \textcircled{\vdash}\langle\neg\bot\rangle)$.
- From $\vdash \neg\bot$ and HBL_1, we obtain $\vdash \textcircled{\vdash}\langle\neg\bot\rangle$.
- From the last two facts, by logic we obtain $\vdash \neg\,(\textcircled{\vdash}\langle\bot\rangle)$, as desired. □

It seems impossible to infer the other direction without knowing what $\textcircled{\vdash}$ looks like more concretely. Therefore, $\nvdash \neg\,\textcircled{\vdash}\langle\bot\rangle$, the original $\mathcal{IT}_2$'s conclusion, is *abstractly stronger than*, hence *preferable to* $\nvdash \mathsf{jcon}$. In short, Jeroslow somewhat weakens the theorem's conclusion.

Let us now look at (a slight rephrasing of) Jeroslow's proof:

Proof of Theorem 30. We assume (1) $\vdash \mathsf{jcon}$ and aim to reach a contradiction.
- Applying Lemma 29 to the formula $\textcircled{\vdash}(\textcircled{\mathsf{N}})$, we obtain a closed pseudo-term τ such that $\vdash \tau \equiv \langle\varphi\rangle$ and (2) $\vdash \varphi \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$, where φ denotes $\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))$.
- By SHBL_3 applied to $\textcircled{\mathsf{N}}(\tau)$, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau)) \rightarrow \textcircled{\vdash}\langle\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))\rangle$, i.e.,
(3) $\vdash \varphi \rightarrow \textcircled{\vdash}\langle\varphi\rangle$.
- From (2) and (3), we obtain $\vdash \varphi \rightarrow \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$.
- On the other hand, (1) instantiated with $\langle\varphi\rangle$ gives us $\vdash \neg\,(\textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle))$.
- From the last two facts, we obtain (4) $\vdash \neg\,\varphi$.
- With HBL_1, we obtain $\vdash \textcircled{\vdash}\langle\neg\,\varphi\rangle$.
- From $\mathsf{Repr}_{\mathcal{F}}$ and CapN, we obtain $\vdash \textcircled{\mathsf{N}}\langle\varphi\rangle \equiv \langle\neg\,\varphi\rangle$.
- From the last two facts, by logic we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$.
- With (2), we obtain $\vdash \varphi$. Together with (4), this contradicts $\mathsf{Con}_{\vdash}$. □

The above proof has a subtle gap, which makes Theorem 30 *incorrect* under its stated assumptions. The problem lies in the highlighted description of the formula φ. Strictly speaking (i.e., rigorously employing our Notation 27), the correct form of fact (2) is not $\vdash \varphi \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$ but $\vdash \varphi \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}})\langle\varphi\rangle$, and the correct φ is not $\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))$ but $\textcircled{\vdash}(\textcircled{\mathsf{N}})(\tau)$. So let us write φ for the correct version, $\textcircled{\vdash}(\textcircled{\mathsf{N}})(\tau)$, and φ' for $\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))$. Notice the difference: φ is obtained by first instantiating $\textcircled{\vdash}$ with $\textcircled{\mathsf{N}}$ and then instantiating the remaining formula with τ, whereas φ' is obtained by first instantiating $\textcircled{\mathsf{N}}$ with τ and then instantiating $\textcircled{\vdash}$ with the result. Both sentences occur in the proof: φ comes from Lemma 29, while φ' comes from SHBL_3. For most purposes in logic, the difference is minor, since (as we note in Example 28(2)) φ and φ' are provably equivalent. However, as we discuss below, shifting between φ and φ' must be done with care, since the proof uses them under the encoding $\langle\_\rangle$.

A first attempt to fill this gap would be to require $\langle\varphi\rangle = \langle\varphi'\rangle$, or at least $\vdash \langle\varphi\rangle \equiv \langle\varphi'\rangle$. The latter would be true under the assumption that *the encodings of provably equivalent sentences are provably equal*. But assuming this is unreasonable: Usually sentence equivalence is undecidable, so no computable encoding can achieve that.[2,3] A more feasible solution comes from noting that the proof does not need $\vdash \langle\varphi\rangle \equiv \langle\varphi'\rangle$, but could work with the weaker property $\vdash \textcircled{\vdash}\langle\varphi\rangle \rightarrow \textcircled{\vdash}\langle\varphi'\rangle$. The latter would be true under the following assumption:

WHBL_2: $\vdash \varphi \leftrightarrow \psi$ implies $\vdash \textcircled{\vdash}\langle\varphi\rangle \rightarrow \textcircled{\vdash}\langle\psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.

Since the $\rightarrow$ in WHBL_2 can be replaced with $\leftrightarrow$ without changing the meaning, WHBL_2 can be read as: *encodings of provably equivalent sentences are provably equiprovable*. Also, WHBL_2 is a weakening of

$\vdash \varphi \rightarrow \psi$ implies $\vdash \textcircled{\vdash}\langle\varphi\rangle \rightarrow \textcircled{\vdash}\langle\psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$

which, in the presence of HBL_1, is seen to be a weak form of HBL_2.[4] This motivates the name "WHBL_2". We are led to the following solution:

Correction 1. Theorem 30 becomes correct if we add WHBL_2 as an assumption.

---

[2] One may argue that we don't need $\vdash \langle\varphi\rangle \equiv \langle\varphi'\rangle$ in general, but only for the particular sentences φ and φ' used in the proof. However, for concrete logics these are complex sentences, so "hacking" an encoding to achieve that equality *while preserving the other desirable properties* seems difficult, if not impossible. This kind of argument could also be made, e.g., for the derivability conditions, in that $\mathcal{IT}_1$ and $\mathcal{IT}_2$ only need them to hold for particular sentences; and the reason why such optimizations are not helpful is similar to the above.

[3] Incidentally, this problem is also the reason why we need SHBL_3 instead of HBL_3: In the application of SHBL_3 to obtain $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau)) \rightarrow \textcircled{\vdash}\langle\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))\rangle$, we cannot work with $\langle\neg\,\varphi\rangle$ instead of $\textcircled{\mathsf{N}}(\tau)$. Indeed, even though $\vdash \textcircled{\mathsf{N}}(\tau) \equiv \langle\neg\,\varphi\rangle$ and hence $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau)) \leftrightarrow \textcircled{\vdash}\langle\neg\,\varphi\rangle$, we cannot conclude $\vdash \langle\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))\rangle \equiv \langle\textcircled{\vdash}\langle\neg\,\varphi\rangle\rangle$.

[4] Incidentally, the latter property can replace HBL_2 when formalizing Step 1, but not when formalizing Step 3 in Gödel-style $\mathcal{IT}_2$ (Theorem 25).

Proof. Indeed, using WHBL_2, we can fill the gap in Theorem 30's proof as follows:
… (same as before)
- Applying Lemma 29 to the formula $\textcircled{\vdash}(\textcircled{\mathsf{N}})$, we obtain a closed pseudo-term τ such that $\vdash \tau \equiv \langle\varphi\rangle$ and $\vdash \varphi \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}})\langle\varphi\rangle$, where φ denotes $\textcircled{\vdash}(\textcircled{\mathsf{N}})(\tau)$.
- Since $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}})\langle\varphi\rangle \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$ by Example 28(2), with the above we obtain
(2) $\vdash \varphi \leftrightarrow \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle)$.
- By SHBL_3 applied to $\textcircled{\mathsf{N}}(\tau)$, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau)) \to \textcircled{\vdash}\langle\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))\rangle$, i.e.,
(3') $\vdash \varphi' \to \textcircled{\vdash}\langle\varphi'\rangle$, where φ' denotes $\textcircled{\vdash}(\textcircled{\mathsf{N}}(\tau))$.
- Since $\vdash \varphi \leftrightarrow \varphi'$, by WHBL_2 we obtain $\vdash \textcircled{\vdash}\langle\varphi'\rangle \to \textcircled{\vdash}\langle\varphi\rangle$.
- From this and (3'), we obtain (3) $\vdash \varphi \to \textcircled{\vdash}\langle\varphi\rangle$.
… (same as before) □

In summary, one solution to filling the gap in Jeroslow's approach, which aimed at removing HBL_2, was to (re)introduce a weaker version of HBL_2, namely WHBL_2.

An alternative solution is to replace representation by pseudo-terms with actual term-representation (defined in Section 4.4). To this end, we amend SHBL_3 to quantify over all closed terms t instead of all closed pseudo-terms τ; moreover, also factoring in the observation that Jeroslow's proof does not need $\mathcal{F}_n$ for all n but $\mathcal{F}_1$ suffices, we change $\mathsf{Repr}_{\mathcal{F}}$ into:

$\mathsf{Repr}_{\mathcal{F}}$: Every $f \in \mathcal{F}_1$ is term-represented, under the identity encoding $\mathsf{Num} \to \mathsf{Num}$, by some $\textcircled{f}$ taken from a set $\mathsf{Ops} \subseteq (\mathsf{Term} \to \mathsf{Term})$ for which an encoding as numerals $\langle\_\rangle : \mathsf{Ops} \to \mathsf{Num}$ is given, and such that $\mathsf{FVars}(g(t)) = \mathsf{FVars}(t)$ and $(g(t))[s/x] = g(t[s/x])$ for all $g \in \mathsf{Ops}$, $s, t \in \mathsf{Term}$ and $x \in \mathsf{Var}$.

(In concrete logics, the elements of Ops can be constructors or derived operators on terms.)

Correction 2. Theorem 30 becomes correct if we work with terms rather than pseudo-terms and amend SHBL_3 and $\mathsf{Repr}_{\mathcal{F}}$ as indicated above.

Proof. Indeed, all the proofs of CapSS, Lemma 29 and Theorem 30 work if we switch from pseudo-terms to terms. □

In summary, our second solution requires the following amendment to Jeroslow's approach: For representing computable functions, we must have available not just pseudo-terms, but actual terms. This usually means that the logic has built-in Skolem symbols and axioms.

Finally, let us see what it takes to alleviate the second tradeoff: from $\nvdash$ jcon to the more desirable $\nvdash \neg \textcircled{\vdash}\langle\bot\rangle$. We consider the following condition:

HBL_4: $\vdash \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\psi\rangle \to \textcircled{\vdash}\langle\varphi \wedge \psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.

HBL_4 has a similar flavor as HBL_2, but refers to conjunction rather than implication: It states that conjunction introduction holds inside the proof system.

Theorem 32. If we modify Theorem 30 by applying Correction 1 (i.e., adding assumption WHBL_2) and adding assumption HBL_4, then its conclusion can be upgraded to $\nvdash \neg \textcircled{\vdash}\langle\bot\rangle$.

Proof. The only time when $\vdash$ jcon is used in the proof is via its specific instance $\vdash \neg (\textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}(\textcircled{\mathsf{N}}\langle\varphi\rangle))$, which by $\mathsf{Repr}_{\mathcal{F}}$ and CapN would follow from (1) $\vdash \neg (\textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\neg\varphi\rangle)$. So it suffices to show that the last follows from $\vdash \neg \textcircled{\vdash}\langle\bot\rangle$, WHBL_2 and HBL_4:
- From HBL_4, we obtain $\vdash \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\neg\varphi\rangle \to \textcircled{\vdash}\langle\varphi \wedge \neg\varphi\rangle$.
- From WHBL_2 and $\vdash \varphi \wedge \neg\varphi \leftrightarrow \bot$, we obtain $\vdash \textcircled{\vdash}\langle\varphi \wedge \neg\varphi\rangle \to \textcircled{\vdash}\langle\bot\rangle$.
- From the last two facts, we obtain $\vdash \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\neg\varphi\rangle \to \textcircled{\vdash}\langle\bot\rangle$.
- Hence $\vdash \neg \textcircled{\vdash}\langle\bot\rangle \to \neg (\textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}\langle\neg\varphi\rangle)$.
- With $\vdash \neg \textcircled{\vdash}\langle\bot\rangle$, we obtain (1), as desired. □

Note that a version of Theorem 32 relying on Correction 2 rather than Correction 1 would be weaker than Theorem 32, since WHBL_2 is necessary in the proof even if we work with terms instead of pseudo-terms.

In summary, Theorem 32 highlights the following assumption tradeoff in Jeroslow's approach, provided the same strong conclusion as in the standard $\mathcal{IT}_2$ is desired: the removal of HBL_2 against the addition of WHBL_2 and HBL_4 (and the slight strengthening of HBL_3 into SHBL_3). Whether this is a good tradeoff will of course depend on the logic's specificity, in particular, on its primitive rules of inference.

Jeroslow presented his approach for an abstract logical theory over a FOL language, which is not necessarily a FOL theory—so it found a natural fit in our generic framework. Jeroslow's account is extremely sketchy and notationally ambiguous. In spite of this account having become part of the $\mathcal{IT}_2$ folklore, very few subsequent authors present it rigorously, and none at its original level of generality. Smith's monograph gives a rigorous account for arithmetic [56, §33], silently performing Correction 2,[5] but failing to detect the need for SHBL_3 instead of HBL_3 (which Jeroslow had noticed). A mechanical proof assistant is of invaluable help with detecting such nuances and pitfalls.

We conclude with an anecdote involving our Isabelle formalization and Jeroslow's notations. Given the relative simplicity of Lemma 29, we were not too surprised that Isabelle's Sledgehammer [41] was able to prove it automatically. But Sledgehammer went further. It reported to have used the equality-reflexivity rule for $\vdash$ in the proof. And it had found a term (not a pseudo-term) t for which it had proved not just $\vdash t \equiv \langle\psi(t)\rangle$, but actual equality, $t = \langle\psi(t)\rangle$; in particular, the term was a numeral. All this was too good to be true. It took us some time to realize why that happened: Due to one of Jeroslow's notations, who wrote f instead of $\textcircled{f}$ (thus identifying a function with its representing pseudo-term), we had at first misstated CapSS, writing $\langle\psi(f\langle\textcircled{f}\rangle)\rangle$ instead of $\langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$; the former is still a valid expression, since f is a function between numerals which are particular terms. Embarrassingly, it took us even longer to realize why this variation discovered by chance was not an improvement of Jeroslow's diagonalization lemma: because the assumption CapSS becomes unreasonable. Indeed, no concrete computable function would then be able to act like the intended ssub (ψ): Given an input n, (1) decode it into a unique formula φ such that $n = \langle\varphi\rangle$, (2) decode φ into a unique function f such that $\varphi = \textcircled{f}$ and (3) proceed to apply f as part of producing $\langle\psi(f\langle\textcircled{f}\rangle)\rangle$. The second step requires an injective and computable encoding of computable functions into formulas, which is impossible.

8 Summary of the Abstract Results

Using our generic infrastructure (Section 4), we have formally proved Gödel-style and Rosser-style diagonalization lemmas (Section 5) and several abstract incompleteness results. They include several versions of $\mathcal{IT}_1$:

- Gödel's original $\mathcal{IT}_1$ (Theorem 13) and an $\mathcal{IT}_1$ based on classical logic (Theorem 22) required the formalization of some well-known arguments without change.
- Rosser's $\mathcal{IT}_1$ (Theorem 16) involved the generalization of a well-known argument: distilling two abstract conditions, Ord_1 and Ord_2.
- Novel semantic variants of $\mathcal{IT}_1$ (Theorems 17–20, 23 and 24) arose from analyzing the interplay between standard models, HBL_1's "iff" version, and proof representability.

[5] In preparation for stating Jeroslow's variant of $\mathcal{IT}_2$, Smith requests that the language has a built-in function symbol for each primitive recursive function (and the theory has corresponding axioms for its behavior), which ensures that one can use terms instead of pseudo-terms.

They also include two versions of $\mathcal{IT}_2$:

- The standard $\mathcal{IT}_2$ based on the three derivability conditions (Theorem 25) again only required formalizing a well-known argument.
- The alternative, Jeroslow-style $\mathcal{IT}_2$ (Theorem 30 with its two corrections, and Theorem 32) involved a detailed analysis and correction of an existing abstract result.

9 Concrete Instances

All the results presented so far operate abstractly, under certain assumptions—starting with a logic as generic as possible and adding structure and hypotheses as needed, while exploring conditions that enable different formulations of the results with various tradeoffs; concrete encodings and recursiveness are below the abstraction level of these results. By contrast, some of the previous mechanization projects, namely those by Shankar [52, 53], O'Connor [36], Harrison [21] and Paulson [40], focused on the impressive goal of "getting all the work done." They fully proved the incompleteness theorems in particular settings, which involved defining the concrete Gödel encodings. These two types of developments are complementary, and they both contribute to formally taming the complex ramifications of the incompleteness theorems.

This section will discuss concrete instances of the abstract results. We start by listing our mechanized instances (Subsection 9.1), and explain how they have been based on Paulson's prior Isabelle development (Subsection 9.2). When instantiating our abstract assumptions to Paulson's setting, not only did we recover his results, but were also able to upgrade them. This did require modifying some concrete proofs, but even when doing that we relied on top-down insight from the abstract results; in fact, as we are about to discuss, insight has traveled bottom-up as well. We also revisit major developments in other provers (Subsection 9.3), and finally briefly sketch a wider array of possible instances (Subsection 9.4).

9.1 Our mechanized instances

We first validate the assumptions about our abstract logic and arithmetic:

Prop 33. (1) Any FOL theory that extends Robinson arithmetic or HF set theory satisfies all the axioms in our logical and arithmetic substrata (in Sections 4.1, 4.2 and 4.3).
(2) If, in addition, the theory is sound, then, together with its corresponding standard model, it also satisfies all our model-theoretic axioms (in Section 4.6).

In particular, point (2) shows that our abstract framework for standard models applies equally well to $\mathbb{N}$ and the datatype of HF sets. In the latter case, Num becomes the entire set of closed terms, so that numerals can denote arbitrary HF sets. This illustrates the versatility of our abstract concept of numeral.

We instantiate two of our main theorems in three ways:

Theorem 34. Let T be a FOL theory that extends HF set theory with a finite set of axioms, and let $\vdash^{\mathsf{b}}$ and $\vdash$ be the same relation, namely provability from T.
(1) If T is sound in the standard HF set model, then the hypotheses of Theorems 24 and 25 are satisfied, i.e., $\mathcal{IT}_1$ (classical semantic version) and $\mathcal{IT}_2$ hold for T.
(2) If T is consistent, then the hypotheses of Theorem 25 are satisfied, i.e., $\mathcal{IT}_2$ holds for T.

The above instances are heavily based on the lemmas proved by Paulson in his Isabelle/HOL formalization of $\mathcal{IT}_1$ (covering both the proof-theoretic and the semantic aspect) and $\mathcal{IT}_2$ [39, 40]. Paulson formalized quite faithfully Świerczkowski's detailed account [61], but he also strengthened and slightly corrected it. Świerczkowski's work applies to HF set theory [61, 62], a classical FOL theory axiomatizing hereditarily finite sets by means of an induction principle stating that the universe is comprised of such sets only. Paulson extended Świerczkowski's incompleteness to essential incompleteness with respect to any finite sound extension of HF set theory within the same FOL language.

Our Theorem 34's point (1) is a restatement of Paulson's formalized results: theorems *Goedel_I* and *Goedel_II* in [40]. By contrast, point (2) is an upgrade of Paulson's *Goedel_II*, applicable to any finite consistent, though possibly unsound theory. This stronger version is a more standard form of $\mathcal{IT}_2$, free from any model-theoretic dependencies. Paulson proved both HBL_1 and $\mathsf{HBL}_1^{\Leftarrow}$ taking advantage of soundness, so to achieve the upgrade we had to discard $\mathsf{HBL}_1^{\Leftarrow}$ and re-prove HBL_1 by replacing any semantic arguments with proofs within the HF calculus. We also removed all invocations of the Σ_1-completeness lemma, which happened to depend on soundness due to Paulson's choice of Σ_1-sentence definition.

This instantiation process has offered us important feedback into the abstract results. A formal development such as ours is (largely) immune to reasoning errors, but not to missing out on useful pieces of generality. We experienced this firsthand with our assumptions about substitution. An *a priori* natural choice was to assume representability of the numeral substitution $\mathsf{Sb} : \mathsf{Fmla}_1 \times \mathsf{Num} \to \mathsf{Sen}$ (defined as $\mathsf{Sb}(\varphi, n) = \varphi(n)$), part of which means (1) $\vdash^{\mathsf{b}} \textcircled{\mathsf{Sb}}(\langle\varphi\rangle, n, \mathsf{Sb}(\varphi, n))$. Instead, Paulson had proved (2) $\vdash^{\mathsf{b}} \textcircled{\mathsf{Sb}}(\langle\varphi\rangle, \langle n\rangle, \mathsf{Sb}(\varphi, n))$. Unlike (1), Paulson's (2) applies the term encoding function $\langle\_\rangle : \mathsf{Term} \to \mathsf{Num}$ to numerals as well (which are particular terms); and since his $\langle\_\rangle$ function is injective, it is far from the case that $\langle n\rangle = n$ for all numerals n. Paulson's version makes more sense than ours when building the results bottom-up: Representability should not discriminate numerals, but filter them through the encodings like other terms. However, top-down our version also made sense: It yielded the incompleteness theorems under reasonable assumptions, which do hold, by the way, for HF set theory—even though in a bottom-up development one is unlikely to prove them. We resolved this discrepancy through a common denominator: the representability of self-substitution $\mathsf{S} : \mathsf{Fmla}_1 \to \mathsf{Sen}$ (Section 4.4), which made our results more general.

Paulson's formalization has also inspired our abstract treatment of standard models (Section 4.6). Since Paulson proved $\mathsf{HBL}_1^{\Leftarrow}$ and used classical logic, an obvious "port of entry" of his $\mathcal{IT}_2$ into our framework is Theorem 22, taking both $\vdash^{\mathsf{b}}$ and $\vdash$ to be Paulson's provability relation (which is classical provability in a finite extension of HF set theory). But this theorem tells us nothing about the Gödel sentences' truth. Delving deeper into Paulson's development, we noted that, following Świerczkowski, he (unconventionally) completely avoided $\mathsf{Repr}_{\Vdash}$, and did not even define $\Vdash$. This raised the question of whether $\mathsf{HBL}_1^{\Leftarrow}$ and $\mathsf{Repr}_{\Vdash}$ are somehow interchangeable in the presence of standard models (on which Paulson relies heavily); and we found that they indeed are, under mild assumptions about truth (as we discuss in Section 4.6). This analysis has led to variants of our semantic $\mathcal{IT}_1$, Theorems 18 and 20, which incidentally do not need classical logic. Although our Theorem 18 seemed like an excellent candidate to instantiate to Paulson's semantic $\mathcal{IT}_1$, its instantiation turned out to be difficult. All its assumptions were easy to fulfill based on what Paulson had already proved, except for $\mathsf{Compl}_{\neg\mathsf{Pf}}$. Indeed, whereas Paulson proved that his proof-of relation is a Σ_1-formula (which implies $\mathsf{Compl}_{\mathsf{Pf}}$ by Σ_1-completeness), he did not prove the

same for its negation (which would imply $\mathsf{Compl}_{\neg\mathsf{Pf}}$). Instead, we recovered Paulson's $\mathcal{IT}_1$ as an instance of our Theorem 24 (which requires classical logic).

There are two further improvements that we could perform to Paulson's formalization, leveraging our abstract results: (1) replacing the soundness assumption from Paulson's $\mathcal{IT}_1$ with consistency, and (2) removing all traces of classical reasoning in the object logic to port Paulson's $\mathcal{IT}_1$ and $\mathcal{IT}_2$ to intuitionistic logic. For the first improvement, we must prove the aforementioned missing link between Paulson's $\mathcal{IT}_1$ and our Theorem 18, namely showing that $\mathsf{Compl}_{\neg\mathsf{Pf}}$ holds in Paulson's setting; we are confident that this is true (any reasonable proof-of relation is a Δ_1-formula, implying that its negation is a Σ_1-formula), but the proof will be very laborious. The second improvement will have a large formal overlap with the first: To remove the uses of the unrestricted Excluded Middle axiom, we must prove that instances of this axiom hold intuitionistically for several formulas expressing decidable predicates, including many predicates that participate in the definition of Paulson's Pf, as well as Pf itself; and, in the presence of $\mathsf{Compl}_{\mathsf{Pf}}$, we have that $\mathsf{Compl}_{\neg\mathsf{Pf}}$ is equivalent to Excluded Middle holding for $\mathsf{Pf}(n, \langle\varphi\rangle)$.

9.3 Connection to results mechanized in other provers

Shankar's 1986 development. In pioneering work [52, 53], Shankar proved formally the proof-theoretic version of $\mathcal{IT}_1$ for any finite extension of the FOL theory Z2 [10], i.e., he proved Z2's finitary essential incompleteness. Z2 is a variation of HF set theory, the difference between the two being that the latter postulates an induction principle for all the HF sets, whereas the former singles out the natural numbers as those transitive HF sets that are totally ordered by membership and postulates induction for numbers only. The underlying object logic considered by Shankar was classical FOL enriched with definitions by the Skolemization of any proved "exists unique" sentences. He worked in Thm, an early version of the Boyer–Moore prover that eventually evolved into Nqthm [6] and then ACL2 [26]. This prover's logic, i.e., the meta-logic of Shankar's development, is a quantifier-free FOL enriched with induction and recursion principles for reasoning about total functions expressed in pure Lisp. This is significantly less expressive than HOL, and in fact close to primitive recursive arithmetic (PRA). Formally proving $\mathcal{IT}_1$ within the constraints of this minimalistic meta-logic was an impressive achievement even by today's standards.

Shankar's development follows a similar structure to Cohen's high-level informal presentation [10, §9] (which Shankar cites). He proved that all partial recursive functions are representable in Z2, a result we will refer to as $\mathcal{RR}$. Besides being a central result in itself, $\mathcal{RR}$ is a convenient tool for proving Gödel's theorems. Some proof developments for $\mathcal{IT}_1$, including the Świerczkowski-Paulson one, do not prove $\mathcal{RR}$ in its generality, but prove the representability of needed functions only. On the other hand, the $\mathcal{RR}$ route is usually the one preferred in textbooks due to its elegance and generality. As Shankar observed, textbook proofs of $\mathcal{IT}_1$ via $\mathcal{RR}$ often step from the meta-logic (where the usual informal mathematical discourse takes place) into a meta-meta-logic: The formula- and proof- manipulating functions needed for $\mathcal{IT}_1$ are defined (as usual) as meta-level functions, then a meta-meta-level argument is being made that they are recursive, in order to conclude that they are representable. In a mechanization, however, such an argument must stay in the meta-logic. Shankar achieves this by formalizing a pure Lisp interpreter that is able to evaluate any recursive function when taking its description as input. His formulation of $\mathcal{RR}$ refers to this interpreter, stating that the interpreter's partial-function behavior (in relational form) is representable in Z2. Each function needed in the proof of $\mathcal{IT}_1$ is proved to be representable by first showing it to be equivalent to its interpreted version. Special care is required to have

these definitions and proofs work in the meta-logic, where all functions must terminate—to that end, the interpreter takes an additional numeric argument representing the maximum allowed size of the computation.

Using notations close to the ones in this paper and bypassing the indirection through the interpreter, Shankar's proof of $\mathcal{IT}_1$ can be summarized as follows. He defined a partial function $\mathsf{THM} : \mathsf{Sen} \to \{0, 1\}$ that, upon an input φ, enumerates all the possible proofs and:

- terminates and returns 1 if a proof of φ is found;
- terminates and returns 0 if a proof of $\neg\,\varphi$ is found.

In particular, THM loops (i.e., is undefined) if neither φ nor $\neg\,\varphi$ is provable. Also, if both φ and $\neg\,\varphi$ are provable (meaning the considered extension of Z2 is inconsistent), then the output of THM depends on whose proof comes first in the enumeration. But regardless of that, it holds that $\mathsf{THM}\,(\varphi) = 1$ implies $\vdash \varphi$, and $\mathsf{THM}\,(\varphi) = 0$ implies $\vdash \neg\,\varphi$.

Let $\psi \in \mathsf{Fmla}_1$ be the formula that represents the unary relation $\{\varphi \in \mathsf{Sen} \mid \mathsf{THM}(\varphi\langle\varphi\rangle) = 1\}$; this is obtained by (i) invoking $\mathcal{RR}$ to produce a formula $\chi \in \mathsf{Fmla}_2$ that represents the graph of the partial function $\mathsf{THM} \circ \mathsf{S}$ (where S is the self-substitution operator), and (ii) substituting $\langle 1\rangle$ for χ's second variable. Let CS be the Cohen–Shankar sentence $\neg\,\psi\langle\neg\,\psi\rangle$.

Now, assume that $\vdash \mathsf{CS}$ or $\vdash \neg\,\mathsf{CS}$, meaning that $\mathsf{THM}\,(\mathsf{CS})$ terminates and returns 1 or 0. We have two cases, both of which contradict consistency:

- If $\mathsf{THM}\,(\mathsf{CS}) = 1$ (i.e., $(\mathsf{THM} \circ \mathsf{S})\,(\neg\,\psi) = 1$), then we have $\vdash \neg\,\psi\langle\neg\,\psi\rangle$ by THM's definition, and also $\vdash \psi\langle\neg\,\psi\rangle$ by $(\mathsf{THM} \circ \mathsf{S})$'s representability.
- If $\mathsf{THM}\,(\mathsf{CS}) = 0$ (hence $(\mathsf{THM} \circ \mathsf{S})\,(\neg\,\psi) \neq 1$), then we have $\vdash \neg\,\neg\,\psi\langle\neg\,\psi\rangle$ by THM's definition, and also $\vdash \neg\,\psi\langle\neg\,\psi\rangle$ by $(\mathsf{THM} \circ \mathsf{S})$'s representability.

The above proof, which is similar to Cohen's proof sketch,[6] does not make explicit reference to HBL_1, although this is of course a consequence of $\mathcal{RR}$ via the representability of the "proof of" relation. In fact, the proof makes use of the representability of $\mathsf{THM} \circ \mathsf{S}$, which is a variation of the representability of $\vdash$ (for particular sentences of the form $\varphi\langle\varphi\rangle$) featuring a positive version of the Rosser twist discussed in Section 5, but at the meta-level: The considered relation is not just provability, but provability by a proof p such that there is no proof q of the formula's negation occurring earlier in the enumeration.

The above argument is based on the diagonalization, though at the meta-level not at the object level as in Prop. 9. As Shankar remarked, the sentence CS says "my negation is provable by a proof that comes in the enumeration before any proof of me". This is true in the context of the above argument by contradiction, namely under the assumption that CS is decided (either provable or unprovable). Indeed, from the definitions of ψ and THM, we see that CS says "it is not the case that a proof of CS comes before a proof of $\neg\,\mathsf{CS}$", which, given the assumption, is equivalent to the above.

Let us refer to such sentence CS as Cohen-Shankar sentences (without claiming historical accuracy about the ideas behind them, which seem to go back at least as far as Smullyan [59]). They can alternatively be obtained by diagonalization in the object logic, namely using Prop. 9 and $\mathsf{Repr}_{\neg}$ to find CS such that $\vdash \mathsf{CS} \leftrightarrow (\exists x.\, ⓟ(x, \langle\neg\,\mathsf{CS}\rangle) \wedge \forall x'.\, (x' < x \to \neg\, ⓟ(x', \langle\mathsf{CS}\rangle)))$, where $<$ is the representation of the occurrence order in the enumeration of proofs used in THM's definition. While classically a Cohen–Shankar sentence is essentially the negation of a Rosser sentence, intuitionistically this is not the case. However, CS can replace the Rosser sentence in our proof-theoretic Rosser-style $\mathcal{IT}_1$, covered by Props. 14, 15 and Theorem 16. Indeed, a bit of mining reveals that the proofs of these results are suf-

[6] A difference is that Cohen employs "self-application" of recursive functions to their (incremented) encoding, whereas Shankar employs self-substitution $\varphi\langle\varphi\rangle$ to the same effect.

ficiently general to accommodate both types of sentences. Given any φ_1 and φ_2, let us define the one-variable formula $\mathsf{Twist}_{\varphi_1,\varphi_2}(x)$ to be $\textcircled{\vdash}(x, \langle\varphi_1\rangle) \wedge \forall x'.\ (x' < x \rightarrow \neg\, \textcircled{\vdash}(x', \langle\varphi_2\rangle))$. Note that, in the presence of $\mathsf{Repr}_\neg$, we have that (i) $\vdash \mathsf{R} \leftrightarrow \neg\, \mathsf{Twist}_{\mathsf{R},\neg\,\mathsf{R}}$ for all Rosser sentences R, and (ii) $\vdash \mathsf{CS} \leftrightarrow \mathsf{Twist}_{\neg\,\mathsf{CS},\mathsf{CS}}$ for all Cohen–Shankar sentences CS. Based on this observation, we can amend the proofs of Props. 14 and 15 by simply replacing $\neg\,\mathsf{R}$ with CS and R with $\neg\,\mathsf{CS}$, and using (ii) instead of (i). Let us illustrate this on the crucial point (5) in Prop. 14's proof, establishing that $\vdash \forall x.\neg\, \mathsf{Twist}_{\mathsf{R},\neg\,\mathsf{R}}(x)$, which is used to infer $\vdash \neg\,(\exists x.\, \mathsf{Twist}_{\mathsf{R},\neg\,\mathsf{R}}(x))$, hence (by (i)) $\vdash \mathsf{R}$, leading to a contradiction with the $\vdash \neg\,\mathsf{R}$ assumption. After the replacement, point (5) establishes that $\vdash \forall x.\neg\, \mathsf{Twist}_{\neg\,\mathsf{CS},\mathsf{CS}}(x)$, which is used to infer $\vdash \neg\,(\exists x.\, \mathsf{Twist}_{\neg\,\mathsf{CS},\mathsf{CS}}(x))$, hence (by (ii)) $\vdash \neg\,\mathsf{CS}$, leading to a contradiction with the $\vdash \mathsf{CS}$ assumption; and similarly for Prop. 15. Moreover, the proof of our semantic Rosser-style $\mathcal{IT}_1$ (Theorems 19 and 20) can be straightforwardly adapted to show, under our abstract assumptions, that CS is false (i.e., $\neg\,\mathsf{CS}$ is true) in the standard model—a fact also proved by Cohen in his concrete setting. In conclusion, our abstract results can be migrated from Rosser to Cohen–Shankar sentences without requiring classical reasoning in the object logic.

O'Connor's 2005 development. O'Connor proved formally the proof-theoretic version of $\mathcal{IT}_1$ [36, 37] for any self-representable extension of a classical FOL theory called NN [23, §7.1], i.e., he proved the essential incompleteness of NN with respect to self-representable extensions. Self-representability of a FOL theory means that its set of axioms is represented by a one-variable formula in that theory. NN is a modification of Robinson arithmetic obtained by replacing the dichotomy axiom (any element is either 0 or a successor) with three axioms regulating the behavior of an additional binary relation symbol for strict order, namely stating that (i) no element is smaller than 0, (ii) being smaller than the successor of an element implies being smaller than or equal to that element, and (iii) the order is total. NN has a similar (though not comparable) expressiveness to Robinson arithmetic. Like the latter, it is significantly less expressive than Peano arithmetic yet sufficient for $\mathcal{IT}_1$ (but not for $\mathcal{IT}_2$). O'Connor worked in the Coq prover [3], so his meta-logic is Coq's underlying Calculus of Inductive Constructions [38], an intuitionistic logic based on intensional type theory. Working out $\mathcal{IT}_1$'s theorem intuitionistically (though for a classical object logic)[7] was original, and revealed some interesting phenomena.

O'Connor's development followed the informal presentation from Hodel's textbook [23], with some notable modifications discussed below. Following Hodel, and similarly to Shankar, he combined a representability theorem for a class of computable functions with proofs that all functions needed for $\mathcal{IT}_1$ are in this class, hence are representable. However, unlike Hodel and Shankar, O'Connor did not prove representability for all recursive functions (a result we denoted by $\mathcal{RR}$), but stopped at the representability of *primitive* recursive functions—we will refer to this latter result as $\mathcal{PR}$. This restriction made the proofs that certain functions are in the considered class more difficult—notably, he reported on the difficulty of establishing that substitution is primitive recursive. On the other hand, O'Connor's formalized representability result is stronger than Shankar's on the theory expressiveness dimension, since it is proved for the minimalistic theory NN.

O'Connor proved a version of $\mathcal{IT}_1$ that would classically read as follows: For any consistent self-representable extension of NN, there exists a sentence φ such that neither φ nor $\neg\,\varphi$ is provable. Due to the intuitionistic meta-logic, O'Connor preferred the intuitionistically stronger (and classically equivalent) formulation: For any self-representable extension of NN, there exists a sentence φ such that, if φ or $\neg\,\varphi$ are provable, then that extension proves

[7] By contrast, most of our results have been established for intuitionistic object logics, while working in a classical meta-logic (the Isabelle/HOL variant of HOL [29]).

everything (i.e., is inconsistent). Another consequence of the intuitionistic meta-logic is the need for an additional assumption: that the given extension's set of axioms is decidable, i.e., its (meta-level) membership predicate satisfies Excluded Middle. The above universally quantified φ is witnessed by a Rosser sentence constructed via diagonalizaton, so the result essentially falls under Props. 9,10 and Theorem 16, where both $\vdash^{\mathsf{b}}$ and $\vdash$ are taken to be deduction in a self-representable extension of NN. (Note that all the FOL theories of interest for $\mathcal{IT}_1$ can already be represented in NN, not only in an extension of NN; and the corresponding (slightly weaker) version of O'Connor's result assuming NN-representability instead of self-representability is obtained by taking $\vdash^{\mathsf{b}}$ to be deduction in NN and $\vdash$ to be deduction in the considered extension.) Since here the FOL infrastructure is fixed, self-representability is equivalent to representability of the "proof of" relation (which O'Connor proved), hence it implies HBL_1 (which he did not mention explicitly but inlined in his proof). Incidentally, O'Connor's formalization improves on Hodel's account, who unnecessarily added an axiom to NN for coping with Rosser's trick [37, §6.4].

O'Connor's self-representability assumption in $\mathcal{IT}_1$ is more general than the standard recursive axiomatizability assumption. In informal accounts of essential incompleteness including Hodel's, this more general result is usually inlined in the proof and only the end result is stated, which assumes not self-representability but recursive axiomatizability; an exception is the account of Feferman, who assumes a generalized form of self-representability (namely representability in a sub-theory) in his statements of $\mathcal{IT}_{0.5}$ and $\mathcal{IT}_2$ (Theorems 5.3 and 5.6 in [13]). In a formal account, such more general results are valuable for easier reusability across different instances. O'Connor did not prove that all recursively axiomatizable extensions of NN are self-representable (which would have followed from $\mathcal{RR}$). However, he used his $\mathcal{PR}$ together with a proof that Peano arithmetic has its axioms primitively recursive to instantiate $\mathcal{IT}_1$ to Peano arithmetic. He also proved the consistency of this theory (by showing that the natural numbers form a model, via a semantic interpretation function wrapped up in a negative translation to ensure classical validity within the intuitionistic meta-logic). Thus, he obtained the theory's unconditional incompleteness.

Harrison's 2009-2010 development. Harrison [21] proved formally versions of $\mathcal{IT}_1$ for theories in the language of Robinson arithmetic with $\leq$ and $<$ included as primitive predicate symbols. In what follows, we will refer to this language as $\mathcal{LA}$, and by "Robinson arithmetic" we will mean the definitional extension of Robinson arithmetic as a theory in $\mathcal{LA}$ (with added axioms that define $\leq$ and $<$). Harrison worked in HOL Light [20], a proof assistant belonging to the HOL family together with Isabelle/HOL and HOL4.

In his development towards $\mathcal{IT}_1$, Harrison followed a semantic approach, based on ideas that go back to Gödel's introduction of his original paper [17]. The approach was promoted by Smullyan [60] for its simplicity and elegance, and Harrison himself further elaborated and improved on it in his textbook [19, §7]. The focus is no longer on the concept of a relation's representability (for a given theory), but on that of a relation's *definability* in the standard model (for a given language). In our notations, definability is obtained by replacing $\vdash^{\mathsf{b}}$ with $\models$ in either the representability or the weak representabilty condition.[8] (Harrison formalized an equivalent definition of definability using valuations in the model.) The advantage of definability over representability is that the former is typically much easier to prove for concrete relations, without having to work inside a formal proof system.

$\mathcal{LA}$ is sufficient to achieve the definability (in the standard model of natural numbers) of the relevant syntactic concepts. These include (soft) self-substitution, which gives a se-

[8] These give the same result under the reasonable assumption that truth satisfies Excluded Middle—our Section 4.6's $\mathsf{LCQ}_{\models}(5)$.

mantic version of diagonalization: Prop. 9 with $\vdash^{\mathsf{b}}$ replaced by $\models$. In turn, this leads to the semantic version of Tarski's theorem on the undefinability of truth, which concludes the non-existence of a one-variable formula T such that $\models \varphi \leftrightarrow \mathsf{T}\langle\varphi\rangle$ for all φ. And after showing that provability in Robinson arithmetic *is* definable, one obtains that provability is distinct from truth; in particular, for sound theories this implies the incompleteness of provability, a first version of the proof-theoretic $\mathcal{IT}_1$. In fact, Harrison proved something more general: If a theory T in $\mathcal{LA}$ is definable (in that its set of axioms is definable), then its set of provable sentences is definable, hence different from the set of true sentences. This leads to a form of essential incompleteness: Any sound definable theory in $\mathcal{LA}$, in particular, any extension of Robinson arithmetic with a sound definable set of axioms, is incomplete.

Harrison also pursued an alternative semantic route to $\mathcal{IT}_1$, which does not go through Tarski's theorem, but instead: (1) assumes (for starters) the soundness of the theory, (2) obtains a semantic version of Gödel sentences G using the semantic diagonal lemma, and (3) performs (what can be regarded as) a modification of the Gödel's original argument (the proofs of Props. 11 and 12), appealing to soundness whenever needed for shifting from provability to truth. The advantage of this last line of reasoning is that it can be sharpened: Noting that soundness is only needed for G, $\neg\,\mathsf{G}$ and $\bot$, and using the fact that G is a Π_1-sentence (making $\neg\,\mathsf{G}$ a Σ_1-sentence) if the theory is Σ_1-definable (i.e., definable by a Σ_1-formula), Harrison obtained the following stronger, symmetric version of proof-theoretic $\mathcal{IT}_1$: If a theory in $\mathcal{LA}$ is Σ_1-definable, then (i) if it also Π_1-sound then $\nvdash \mathsf{G}$ and (ii) if it also Σ_1-sound then $\nvdash \neg\,\mathsf{G}$ (where $\vdash$ denotes deduction from this theory, and X-soundness or X-completeness means soundness or completeness for all X-sentences). And from representability and the semantic Gödel-sentence property, under the assumptions of (i), it follows that $\models \mathsf{G}$. So he obtained both the proof-theoretic and the semantic component of $\mathcal{IT}_1$.

In the above statement of $\mathcal{IT}_1$, the Π_1-soundness assumption can be replaced by consistency plus Σ_1-completeness, since the latter two imply the former. Finally, using the Σ_1-completeness for Robinson arithmetic (and hence for any extension), Harrison formalized an essential incompleteness generalization and strengthening of the original Gödel-style $\mathcal{IT}_1$: For any consistent Σ_1-definable extension of Robinson arithmetic, we have $\nvdash \mathsf{G}$ and $\models \mathsf{G}$; and if the extension is also Σ_1-sound, then $\nvdash \neg\,\mathsf{G}$. In the presence of Σ_1-completeness, the Σ_1-soundness property (also called 1-consistency) is weaker than the ω-consistency property used originally by Gödel, which we assume in our Prop. 12 and Theorem 13.

Currently, refinements of $\mathcal{IT}_1$ based on arithmetical hierarchy considerations are below the level of abstraction of our general framework. On the other hand, the high-level aspects of the Smullyan–Harrison semantic line of reasoning could be incorporated in this framework, which has infrastructure for both provability and truth. Our Archive of Formal Proofs entry [44] already contains proof-theoretic and semantic versions of Tarski's theorem.

9.4 Other potential instances

Many other logics and logical theories satisfy our theorems' assumptions. We do *not* require the logic to be reducible to a single syntactic category of formulas, Fmla, a single pair of judgments, $\vdash^{\mathsf{b}}$ and $\vdash$, etc.; but only that such (well-behaved) formulas, provability relations, etc. are identifiable as part of that logic, e.g., localized to a given type and/or relativised by a given predicate. This allows our framework to capture most variants of higher-order logic and type theory (including the variant underlying Isabelle/HOL itself [29, 30]), and also, we believe, many of the logics surveyed by Buldt [7], including non-classical and fuzzy. But enabling "mass instantiation" that is both formal and painless requires more progress

on the agenda we started here: recognizing reusable construction and proof patterns and formalizing them as abstract results.

Acknowledgments. We thank Bernd Buldt for his patient explanations on material in his monograph, Russell O'Connor and Natarajan Shankar for their clarification of aspects of their work, and the CADE and Journal of Automated Reasoning reviewers for insightful comments and suggestions.

References

1. Auerbach, D.: Intensionality and the Gödel theorems. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **48**(3), 337–351 (1985)
2. Ballarin, C.: Locales: A module system for mathematical theories. J. Autom. Reason. **52**(2), 123–153 (2014)
3. Bertot, Y., Castéran, P.: Interactive Theorem Proving and Program Development – Coq'Art: The Calculus of Inductive Constructions. Texts in Theoretical Computer Science. An EATCS Series. Springer (2004)
4. Blanchette, J.C., Popescu, A., Traytel, D.: Unified classical logic completeness—A coinductive pearl. In: IJCAR 2014, pp. 46–60 (2014)
5. Boolos, G.: The Logic of Provability. Cambridge University Press (1993)
6. Boyer, R., Kaufmann, M., Moore, J.: The Boyer-Moore theorem prover and its interactive enhancement. Computers & Mathematics with Applications **29**(2), 27 – 62 (1995)
7. Buldt, B.: The scope of Gödel's first incompleteness theorem. Logica Universalis **8**(3), 499–552 (2014)
8. Bundy, A., Giunchiglia, F., Villafiorita, A., Walsh, T.: An incompleteness theorem via abstraction. Tech. rep., Istituto per la Ricerca Scientifica e Tecnologica, Trento (1996)
9. Carnap, R.: Logische syntax der sprache. Philosophical Review **44**(4), 394–397 (1935)
10. Cohen, P.J.: Set theory and the continuum hypothesis. New York: W.A. Benjamin (1966)
11. Davis, M.: The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems, and Computable Functions. Dover Publication (1965)
12. Diaconescu, R.: Institution-independent Model Theory, 1st edn. Birkhäuser (2008)
13. Feferman, S.: Arithmetization of metamathematics in a general setting. Journal of Symbolic Logic **31**(2), 269–270 (1966)
14. Feferman, S., Dawson Jr., J.W., Kleene, S.C., Moore, G., Solovay, R., van Heijenoort, J. (eds.): Kurt Gödel: Collected Works, Volume I: Publications 1929–1936. Oxford University Press (1986)
15. Fiore, M.P., Plotkin, G.D., Turi, D.: Abstract syntax and variable binding. In: Logic in Computer Science (LICS) 1999, pp. 193–202. IEEE Computer Society (1999)
16. Gabbay, M.J., Mathijssen, A.: Nominal (universal) algebra: Equational logic with names and binding. J. Log. Comput. **19**(6), 1455–1508 (2009)
17. Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik **38**(1), 173–198 (1931)
18. Goguen, J.A., Burstall, R.M.: Institutions: Abstract model theory for specification and programming. J. ACM **39**(1), 95–146 (1992)
19. Harrison, J.: Handbook of Practical Logic and Automated Reasoning. Cambridge University Press (2009)
20. Harrison, J.: HOL light: An overview. In: TPHOLs, pp. 60–66 (2009)
21. Harrison, J.: HOL Light proof of Gödel's first incompleteness theorem (2010). Located at `https://github.com/jrh13/hol-light/`, directory Arithmetic
22. Hilbert, D., Bernays, P.: Grundlagen der Mathematik, Vol. II. Springer-Verlag (1939)
23. Hodel, R.E.: An Introduction to Mathematical Logic, 2nd edn. Dover Publications (2013)
24. Jeroslow, R.G.: Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. J. Symb. Log. **38**(3), 359–367 (1973)
25. Kaliszyk, C., Urban, J.: HOL(y)Hammer: Online ATP service for HOL light. Mathematics in Computer Science **9**(1), 5–22 (2015)
26. Kaufmann, M., Manolios, P., Moore, J.S.: Computer-Aided Reasoning: An Approach. Kluwer Academic Publishers (2000)
27. Kikuchi, M., Kurahashi, T.: Generalizations of Gödel's incompleteness theorems for $\sum$ n -definable theories of arithmetic. Rew. Symb. Logic **10**(4), 603–616 (2017)
28. Kreisel, G.: Mathematical logic. In: T.L. Saaty (ed.) Lectures on modern mathematics, vol. 3. Wiley (1963)

29. Kunčar, O., Popescu, A.: A Consistent Foundation for Isabelle/HOL. In: ITP, pp. 234–252 (2015)
30. Kunčar, O., Popescu, A.: Comprehending Isabelle/HOL's consistency. In: ESOP, pp. 724–749 (2017)
31. Kunčar, O., Popescu, A.: From types to sets by local type definition in higher-order logic. J. Autom. Reason. **62**(2), 237–260 (2019)
32. Löb, M.: Solution of a Problem of Leon Henkin. The Journal of Symbolic Logic **20**(2), 115—-118 (1955)
33. Matichuk, D., Murray, T.C., Wenzel, M.: Eisbach: A proof method language for isabelle. J. Autom. Reason. **56**(3), 261–282 (2016)
34. Nipkow, T., Klein, G.: Concrete Semantics - With Isabelle/HOL. Springer (2014)
35. Nipkow, T., Paulson, L., Wenzel, M.: Isabelle/HOL — A Proof Assistant for Higher-Order Logic, *LNCS*, vol. 2283. Springer (2002)
36. O'Connor, R.: Essential incompleteness of arithmetic verified by Coq. In: TPHOLs, pp. 245–260 (2005)
37. O'Connor, R.: Incompleteness & Completeness: Formalizing Logic and Analysis in Type Theory. Ph.D. thesis, Radboud University Nijmegen, the Netherlands (2009)
38. Paulin-Mohring, C.: Introduction to the Calculus of Inductive Constructions. In: All about Proofs, Proofs for All (2015)
39. Paulson, L.C.: A machine-assisted proof of Gödel's incompleteness theorems for the theory of hereditarily finite sets. Rew. Symb. Logic **7**(3), 484–498 (2014)
40. Paulson, L.C.: A mechanised proof of Gödel's incompleteness theorems using Nominal Isabelle. J. Autom. Reasoning **55**(1), 1–37 (2015)
41. Paulson, L.C., Blanchette, J.C.: Three years of experience with Sledgehammer, a practical link between automatic and interactive theorem provers. In: IWIL 2010, pp. 1–11 (2010)
42. Popescu, A., Roşu, G.: Term-generic logic. Theor. Comput. Sci. **577**, 1–24 (2015)
43. Popescu, A., Traytel, D.: A formally verified abstract account of Gödel's incompleteness theorems. In: Automated Deduction - CADE 27, pp. 442–461 (2019)
44. Popescu, A., Traytel, D.: An abstract formalization of Gödel's incompleteness theorems. Archive of Formal Proofs (2020). URL `https://www.isa-afp.org/entries/Goedel_Incompleteness.html`
45. Popescu, A., Traytel, D.: From abstract to concrete Gödel's incompleteness theorems—part I. Archive of Formal Proofs (2020). URL `https://www.isa-afp.org/entries/Goedel_HFSet_Semantic.html`
46. Popescu, A., Traytel, D.: From abstract to concrete Gödel's incompleteness theorems—part II. Archive of Formal Proofs (2020). URL `https://www.isa-afp.org/entries/Goedel_HFSet_Semanticless.html`
47. Popescu, A., Traytel, D.: Robinson arithmetic. Archive of Formal Proofs (2020). URL `https://www.isa-afp.org/entries/Robinson_Arithmetic.html`
48. Popescu, A., Traytel, D.: Syntax-independent logic infrastructure. Archive of Formal Proofs (2020). URL `https://www.isa-afp.org/entries/Syntax_Independent_Logic.html`
49. Quaife, A.: Automated proofs of Löb's theorem and Gödel's two incompleteness theorems. J. Autom. Reasoning **4**(2), 219–231 (1988)
50. Raatikainen, P.: Gödel's incompleteness theorems. In: The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2018)
51. Schlichtkrull, A., Blanchette, J.C., Traytel, D., Waldmann, U.: Formalizing Bachmair and Ganzinger's ordered resolution prover. In: IJCAR 2018, pp. 89–107 (2018)
52. Shankar, N.: Proof-checking metamathematics. Ph.D. thesis, University of Texas (1986)
53. Shankar, N.: Metamathematics, Machines, and Gödel's Proof. Cambridge University Press (1994)
54. Sieg, W., Field, C.: Automated search for Gödel's proofs. Ann. Pure Appl. Logic **133**(1-3), 319–338 (2005)
55. Sieg, W., Lindstrom, I., Lindstrom, S.: Gödel's incompleteness theorems – a computer-based course in elementary proof theory. In: University-Level Computer-Assisted Instruction at Stanford, pp. 183–193 (1981)
56. Smith, P.: An introduction to Gödel's incompleteness theorems. Cambridge University Press (2007)
57. Smorynski, C.: The incompleteness theorems. In: J. Barwise (ed.) Handbook of Mathematical Logic, pp. 821–865. North-Holland (1977)
58. Smorynski, C.: Self-reference and modal logic. Springer, Berlin (1985)
59. Smullyan, R.M.: Theory of Formal Systems. (AM-47). Princeton University Press (1961)
60. Smullyan, R.M.: Gödel's Incompleteness Theorems. Oxford University Press (1992)
61. Świerczkowski, S.: Finite sets and Gödel's incompleteness theorems. Dissertationes Mathematicae **422**, 1–58 (2003)
62. Tarski, A., Givant, S.: A Formalization of Set Theory without Variables. Soc. Colloq. Publ. 41, Amer. Math. Soc., Providence, RI (1987)
63. Tarski, A., Mostowski, A., Robinson, R.: Undecidable Theories. Studies in Logic and the Foundations of Mathematics. North-Holland (1953). 3rd edition, 1971
64. Urban, C.: Nominal Techniques in Isabelle/HOL. J. Autom. Reason. **40**(4) (2008)

A More Details on the Isabelle Formalization

While in this article the focus is on the abstract formulation of Gödel's incompleteness theorems, our formalization's total 28 000 lines of Isabelle definitions and proofs split almost evenly between the abstract and the concrete results. More specifically, we organize our development in five entries in the Archive of Formal Proofs (AFP) [44–48]—a centralized repository for Isabelle proofs, which ensures the proofs' longevity despite the continuous, often not-backwards-compatible development of Isabelle.

Two of the five AFP entries cover the formalization's abstract parts. Our development of syntax-independent logic infrastructure as described in this article's Sections 4.1–4.3 spans 9 200 lines [48]. The abstract definitions more specifically related to Gödel's incompleteness theorems described in Sections 4.4–7 is at 4 000 lines more concise. We observe that line numbers are not necessarily a good measure for a proof's ingenuity. Logic infrastructure proofs revolve around setting up parallel substitution and a deduction calculus for further use—a tedious but routine task. In contrast, the abstract Gödel development involves subtle diagonalization arguments and requires careful symbolic manipulation of represented objects.

The remaining three AFP entries formalize the concrete instances of the abstract results detailed in Section 9. Prop. 33 instantiates our syntax-independent logic infrastructure with Robinson arithmetic. Our 1 600 lines long formalization of Robinson arithmetic [47] follows the style and structure of Paulson's formalization of HF set theory. In particular, we also use Nominal Isabelle [64] to represent binders. Our formalization of Theorem 34(1) [45] reproduces Paulson's formulation of Gödel incompleteness theorems for finite, sound extensions of HF set theory. This merely required us to discharge the abstract assumptions of Theorems 24 and 25 by instantiating them with results from Paulson's formalization—a simple exercise spanning 400 lines (not counting the 12 300 lines of Paulson's formalization). Formalizing the strengthened Theorem 34(2) [46] was significantly more difficult, because we could not simply reuse Paulson's formalization. Instead, we had to replace all of Paulson's semantic arguments with proofs within the HF calculus. In terms of proof-engineering, we started by copying Paulson's formalization (12 300 lines) and by removing from it every argument and definition that referred to standard models, which saved about 5 000 lines. After that, we reintroduced the arguments needed for Gödel's second incompleteness theorem and proved them within the HF calculus. The new proofs span about as much as we had removed, such that overall we obtain the stronger result in 12 800 lines.

Our formalization relies heavily on locales [2], Isabelle's mechanism for maintaining contexts with parameters and assumptions. The two abstract AFP entries [44, 48] declare 65 interdependent locales. These locales allow us to flexibly select just the needed assumptions for each theorem's variant. On the downside, complex locale hierarchies like ours tend to cause the formalizers to write seemingly redundant boilerplate code. In particular, every locale which extends another locale has to repeat the parameters (but fortunately not the assumptions) of the extended locale to ensure that correct type variables are used in the new locale.

In our locales, we fix explicit sets as universes of variables, numerals, terms, and formulas. Thus, any quantification over these entities must be expressed as bounded quantification over the fixed sets. This complicates the reasoning inside of the locales, because every step that uses a theorem with bounded quantification must discharge these additional universe-membership assumptions. We have even developed an *ad hoc* collection of specialized Eisbach proof methods [33] to deal with such assumptions. A natural alternative that would avoid these complications is to use types as universes. We opted for the set-based formulation instead of the type-based one, because set-based result can be instantiated more flexibly. For example, numerals are a subset of terms in Paulson's HF set theory and we instantiate our locales' universe of numerals with this subset. A type-based formulation would require introducing a separate type for numerals and lifting all results involving numerals to this type. Another alternative, the types-to-sets approach [31], combines the strengths of type-based and set-based theorems, at the expense of extending the logic, which we wanted to avoid.

The abstract parts of our formalization use declarative Isar proofs. This makes the proofs readable and ensures that they closely resemble the pen-and-paper arguments presented in this paper. In fact, the information flow for this algorithm went in the opposite direction: the pen-and-paper arguments constitute a (sometimes compressed) transcript of the formal Isar proofs. The concrete parts use a mixture of declarative and procedural (apply-style) proofs. Especially proofs in the HF calculus tend to follow the procedural style.

All our concrete theorems use Nominal Isabelle [64] to represent formulas with binders. This, however, is attributed to the fact that in all cases we took Paulson's formalization, which uses Nominal, as a blueprint. Our abstract development does not prescribe the usage of Nominal—it can similarly well accommodate de Bruijn indices, locally nameless terms, or other representations that equate alpha-equivalent terms.

B Main Property Index

$\mathsf{Con}_{\vdash}$: $\nvdash \bot$.
$\mathsf{OCon}_{\vdash}$: For all $\varphi \in \mathsf{Fmla}_1$, if $\vdash \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$ then $\nvdash \neg\, \neg\, (\exists x.\, \varphi(x))$.
$\mathsf{Rel}^{\Vdash}_{\vdash}$: For all $\varphi \in \mathsf{Sen}$, $\vdash \varphi$ iff there exists $p \in \mathsf{Proof}$ such that $p \Vdash \varphi$.
Ord_1: For all $\varphi \in \mathsf{Fmla}_1$ and $n \in \mathsf{Num}$, if $\vdash^{\mathsf{b}} \varphi(m)$ for all $m \in \mathsf{Num}$, then $\vdash \forall x \prec n.\, \varphi(x)$.
Ord_2: For all $n \in \mathsf{Num}$, there exists a finite set $M \subseteq \mathsf{Num}$ such that $\vdash \forall x.\, x \in M \vee n \prec x$.
$\mathsf{Repr}_{\neg}$:
(1) $\vdash \dot{\neg}(\langle\varphi\rangle, \langle\neg\,\varphi\rangle)$ for all $\varphi \in \mathsf{Sen}$.
(2) $\vdash^{\mathsf{b}} \forall x, y.\, \dot{\neg}(\langle\varphi\rangle, x) \wedge \dot{\neg}(\langle\varphi\rangle, y) \to x \equiv y$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{Repr}_{\mathsf{S}}$:
(1) $\vdash \dot{\mathsf{S}}(\langle\varphi\rangle, \langle\varphi\langle\varphi\rangle\rangle)$ for all $\varphi \in \mathsf{Fmla}_1$.
(2) $\vdash^{\mathsf{b}} \forall x, y.\, \dot{\mathsf{S}}(\langle\varphi\rangle, x) \wedge \dot{\mathsf{S}}(\langle\varphi\rangle, y) \to x \equiv y$ for all $\varphi \in \mathsf{Fmla}_1$.
$\mathsf{Repr}_{\Vdash}$:
(1) $p \Vdash \varphi$ implies $\vdash^{\mathsf{b}} \dot{\Vdash}(\langle p\rangle, \langle\varphi\rangle)$ for all $p \in \mathsf{Proof}$ and $\varphi \in \mathsf{Sen}$.
(2) $p \nVdash \varphi$ implies $\vdash^{\mathsf{b}} \neg\, \dot{\Vdash}(\langle p\rangle, \langle\varphi\rangle)$ for all $p \in \mathsf{Proof}$ and $\varphi \in \mathsf{Sen}$.
$\mathsf{Clean}_{\Vdash}$: $\vdash^{\mathsf{b}} \neg\, \dot{\Vdash}(n, \langle\varphi\rangle)$ for all $\varphi \in \mathsf{Sen}$ and $n \in \mathsf{Num}$ such that $n \neq \langle p\rangle$ for all $p \in \mathsf{Proof}$.
HBL_1: $\vdash \varphi$ implies $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
HBL_2: $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle \wedge \dot{\vdash}\langle\varphi \to \psi\rangle \to \dot{\vdash}\langle\psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
HBL_3: $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle \to \dot{\vdash}\langle\dot{\vdash}\langle\varphi\rangle\rangle$ for all $\varphi \in \mathsf{Sen}$.
HBL_4: $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle \wedge \dot{\vdash}\langle\psi\rangle \to \dot{\vdash}\langle\varphi \wedge \psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
$\mathsf{HBL}_1^{\Leftarrow}$: $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{HBL}_{1,\vdash}^{\Leftarrow}$: $\vdash \dot{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.
SHBL_3 (Jeroslow's formulation): $\vdash \dot{\vdash}(\tau) \to \dot{\vdash}\langle\dot{\vdash}(\tau)\rangle$ for all closed pseudo-terms τ.
SHBL_3 (our simplified formulation): $\vdash \dot{\vdash}(t) \to \dot{\vdash}\langle\dot{\vdash}(t)\rangle$ for all closed terms t.
WHBL_2: $\vdash^{\mathsf{b}} \varphi \leftrightarrow \psi$ implies $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle \to \dot{\vdash}\langle\psi\rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
$\mathsf{Rel}^{\mathsf{Pf}}_{\dot{\vdash}}$: $\vdash^{\mathsf{b}} \dot{\vdash}\langle\varphi\rangle \;\leftrightarrow\; \exists x.\, \mathsf{Pf}(x, \langle\varphi\rangle)$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{Compl}_{\mathsf{Pf}}$: $\models \mathsf{Pf}(n, \langle\varphi\rangle)$ implies $\vdash^{\mathsf{b}} \mathsf{Pf}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.
$\mathsf{Compl}_{\neg\mathsf{Pf}}$: $\models \neg\, \mathsf{Pf}(n, \langle\varphi\rangle)$ implies $\vdash^{\mathsf{b}} \neg\, \mathsf{Pf}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.
$\mathsf{LCQ}_{\models}$:
(1) $\not\models \bot$; (2) for all $\varphi, \psi \in \mathsf{Sen}$, $\models \varphi$ and $\models \varphi \to \psi$ imply $\models \psi$;
(3) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \varphi(n)$ for all $n \in \mathsf{Num}$ then $\models \forall x.\, \varphi(x)$;
(4) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \exists x.\, \varphi(x)$ then $\models \varphi(n)$ for some $n \in \mathsf{Num}$;
(5) for all $\varphi \in \mathsf{Sen}$, $\models \varphi$ or $\models \neg\, \varphi$.
$\mathsf{Sound}^{\vdash^{\mathsf{b}}}_{\models}$: $\vdash^{\mathsf{b}} \varphi$ implies $\models \varphi$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{TIP}^{\vdash}_{\models}$: $\models \dot{\vdash}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.
$\mathsf{Repr}_{\mathcal{F}}$ (Jeroslow's formulation): For all $f \in \mathcal{F}_m$, there exists $\dot{f} \in \mathsf{PTerm}_m$ such that $\vdash \dot{f}(n_1, \ldots, n_m) \equiv f(n_1, \ldots, n_m)$.
Note that, since $\dot{f}$ is a pseudo-term, the above condition is equivalent to the conjunction of the following two conditions (which express representability, as defined in Section 4.4):
(1) $\vdash \dot{f}(n_1, \ldots, n_m, f(n_1, \ldots, n_m))$ for all $n_1, \ldots, n_m \in \mathsf{Num}$.
(2) $\vdash \forall x, y.\, \dot{f}(n_1, \ldots, n_m, x) \wedge \dot{f}(n_1, \ldots, n_m, y) \to x \equiv y$ for all $n_1, \ldots, n_m \in \mathsf{Num}$.
Indeed, (1) is exactly $\vdash \dot{f}(n_1, \ldots, n_m) \equiv f(n_1, \ldots, n_m)$ after expanding our introduced notation for pseudo-terms, and (2) follows from the uniqueness part of the pseudo-term condition.
$\mathsf{Repr}_{\mathcal{F}}$ (our corrected and simplified formulation):
(1) For all $f \in \mathcal{F}_1$, there exists $\dot{f} \in \mathsf{Ops}$ such that $\vdash \dot{f}(n) \equiv f(n)$ for all $n \in \mathsf{Num}$.
(2) $\mathsf{FVars}(g(t)) = \mathsf{FVars}(t)$ and $(g(t))[s/x] = g(t[s/x])$ for all $g \in \mathsf{Ops}$, $s, t \in \mathsf{Term}$ and $x \in \mathsf{Var}$.
CapN: $\mathsf{N} \in \mathcal{F}_1$ and $\mathsf{N}\langle\varphi\rangle = \langle\neg\,\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
CapSS: $\mathsf{ssub}\,\psi \in \mathcal{F}_1$ and $\mathsf{ssub}\,\psi\,\langle\dot{f}\rangle = \langle\psi(\dot{f}\langle\dot{f}\rangle)\rangle$ for all $\psi \in \mathsf{Fmla}_1$ and $f \in \mathcal{F}_1$.