The Provability of Consistency

Sergei Artemov

The Graduate Center, the City University of New York
365 Fifth Avenue, New York City, NY 10016
sartemov@gc.cuny.edu

February 26, 2019

Abstract

Provability semantics suggests well-principled notions of constructive truth and constructive falsity of classical sentences in Peano arithmetic \(\text{PA}\): \(F\) is constructively true iff \(\text{PA}\) proves \(F\); \(F\) is constructively false iff \(\text{PA}\) proves ‘for each \(x\), there is a proof that \(x\) is not a proof of \(F\).’ We also consider an associated notion of constructive consistency of \(\text{PA}\), \(\text{CCons}_{\text{PA}}\), ‘for each \(x\), there is a proof that \(x\) is not a proof of \(0=1\).’ We show that \(\text{PA}\) proves \(\text{CCons}_{\text{PA}}\) hence there is no a Gödel-style impossibility barrier for case-by-case consistency proofs. Furthermore, we prove a finitary version of constructive consistency directly:

for any \(\text{PA}\)-derivation \(S\) we find a finitary proof that \(S\) does not contain \(0=1\).

This proves consistency of \(\text{PA}\) by finitary means and appears to vindicate Hilbert’s program of establishing consistency of formal theories.

1 Consistency problem: lost in translation

Hilbert’s consistency program of the 1920s (cf. [28]) was asking for a finitary demonstration that in a formal theory such as \(\text{PA}\)

\[
\text{any finite sequence } S \text{ of formulas is not a derivation of a contradiction. (1)}
\]

This is a mathematical problem (provided “finitary” is given a mathematical meaning) depending on a parameter \(S\) ranging over finite sequences of formulas.

Impossibility of finitary proofs of consistency, usually attributed to Gödel’s Second Incompleteness Theorem, \(\text{G2}\), is largely a myth based on an uncritical metamathematical interpretation of its mathematical result. Indeed, derivations in \(\text{PA}\) are finite sequences
of formulas. Gödel’s arithmetization encodes derivations by numbers, which is fine due to constructive and epistemically unproblematic coding/decoding procedures.

However, the next step of arithmetization, that of using numeric quantifiers to represent universal properties of derivations, introduces serious aberrations. Arithmetical formula $\text{Con}_{PA}$,

$$\text{for all } x, x \text{ is not a code of a proof of } 0 = 1,$$

does not represent Hilbert’s notion of consistency of $PA$-derivations fairly. Due to well-known expressivity limitations in $PA$, the universal quantifier spills over to nonstandard proof codes, and hence $\text{Con}_{PA}$ stating consistency of both standard and nonstandard proof codes is just too strong for (1). If $\text{Con}_{PA}$ is true in all models of $PA$, then (1) holds. The converse, however, does not hold. So, the result of $G2$, that $PA$ cannot prove $\text{Con}_{PA}$, does not actually block finitary consistency proofs for standard $PA$-derivations as formulated in Hilbert’s consistency program.

Moreover, as we show in this paper, once standard derivations are represented fairly by an input parameter $S$ as in (1) rather than by an indiscriminate universal quantifier over proof codes as in $\text{Con}_{PA}$, no problems with finitary consistency proofs in $PA$ occur. We establish by a direct finitary reasoning that for any given finite sequence of arithmetical sentences $S$

$$\text{there is a finitary proof } p(S) \text{ that } S \text{ is not a derivation of a contradiction.} \tag{2}$$

Namely, for any $PA$-derivation $S$, a standard partial truth definition analysis bounded by complexity of formulas from $S$ establishes that $S$ does not contain $0 = 1$. This demonstrates $PA$-consistency by finitary means.

An obvious feature of (2) is that we are producing a finitely generated series of finitary proofs $p(S)$ and, by $G2$, these proofs do not collapse into one uniform $PA$-proof of consistency.

So what? “Finitary” does not mean “finite.” In limited languages such as propositional, first-order, arithmetical languages, etc., we are fated to live with some well-defined infinite collections of finite constructive objects. Logical axioms and induction axioms in $PA$ constitute constructively generated infinite families of formulas, and for over a hundred years we naturally accept these objects as finitary. We do not dismiss $PA$ as not well defined just because it requires an infinite series of induction axioms.

The case-by-case finitary proofs $p(S)$ from (2) have the same status: they are constructively described finite objects each of which can be shaped as a formal mathematical proof formalizable in $PA$. From this perspective, solution (2) fits Hilbert’s desiderata (1).

We establish a connection of (2) with the theory of constructive truth and constructive falsity of arithmetical sentences inspired by Brouwer-Heyting-Kolmogorov semantics [11, 25] and its formalization in the Logic of Proofs [3, 5, 6]. This connection provides a bigger picture and explains how constructive consistency and its finitary version (2) were able to circumvent some limitations on consistency proofs imposed by Gödel’s Second Incompleteness Theorem.
It appears that despite a long history of suggestions to bypass Gödel’s Second Incompleteness Theorem (cf. [13, 14, 15, 28]), this case-by-case approach to proving consistency when embedded into a general theory of constructive truth, falsity, and consistency is well-principled and offers a new substance. If so, this vindicates, to some extent, Hilbert’s program of establishing consistency of formal theories and reopens the door to the study of similar consistency proofs for other theories.

2 Findings

By no means are we casting doubt upon Gödelian consistency studies, ordinal analysis, etc.; these are the classics of mathematical logic.

However, as far as foundations are concerned, viewing Gödel’s Incompleteness Theorems as blocking the very possibility of proving consistency internally is an unfortunate misconception which we suggest resisting.

Gödel’s consistency formula $\text{Con}_{\text{PA}}$ is (much) stronger than the consistency property (1) of a given theory. The fact that $\text{Con}_{\text{PA}}$ is not provable does in $\text{PA}$ not rule out another perfectly fine possibility reflected in the constructive consistency idea: to have a case-by-case internal consistency proof for each derivation. We show that this possibility can be realized for Peano arithmetic: we offer a finitary consistency proof for $\text{PA}$ based on partial truth definitions.

We did not start with inspecting the conditions of Gödel’s Second Incompleteness Theorem and trying to weaken some of them in order to dodge its impossibility conclusion. Our staring point was the foundational problem in its entirety:

Can mathematical theories establish consistency of their own formalizations? (3)

The prevailing wisdom so far has been “No, by Gödel’s Second Incompleteness Theorem, unless such a theory is inconsistent.” We offer a new foundational answer to (3):

Mathematical theories can prove consistency of their own formalizations. (4)

In particular, this dismantles the reflection tower of theories of increasing metamathematical strength as the ‘only true’ foundational picture.

3 Brouwer-Heyting-Kolmogorov semantics

According to traditional Brouwer-Heyting-Kolmogorov semantics, cf. [11],

• a proof of $A \rightarrow B$ is a construction which, given a proof of $A$, returns a proof of $B$.

Notoriously, this yields weird properties of negation:

1. trivialization: $\neg A$ holds for any non-provable $A$, no real witness of $\neg A$ is required;
2. irrelevance of witnesses: if \( \neg A \) holds, then any \( p \) is a dummy BHK-proof of \( \neg A \).

These features are counterintuitive and undermine the very basic principles of constructive semantics; they were subject of criticism by Kreisel, so-called Kreisel’s second clause (cf. [12]).

Gödel in [16] endorsed classical modal logic \( \mathbf{S4} \) as the calculus of provability in which \( \Box A \) informally represents ‘\( A \) is provable’:

- Axioms and rules of classical propositional logic,
- \( \Box (F \rightarrow G) \rightarrow (\Box F \rightarrow \Box G) \),
- \( \Box F \rightarrow F \),
- \( \Box F \rightarrow \Box \Box F \),
- Rule of Necessitation: \( \vdash F \rightarrow \Box F \).

Gödel connected classical provability with intuitionistic logic \( \mathbf{IPC} \) in a way that respects the provability reading of the latter: \( \mathbf{IPC} \vdash F \) iff \( \mathbf{S4} \vdash \text{tr}(F) \), where \( \text{tr}(F) \) is obtained by ‘boxing’ each subformula of \( F \). At that stage, a provability semantics for \( \mathbf{IPC} \) seemed to reduce to a provability semantics for \( \mathbf{S4} \). However, as it was noticed by Gödel, \( \mathbf{S4} \) endorses reflection principle

\[
\Box F \rightarrow F,
\]

not compatible with the straightforward reading of \( \Box \) as formal provability.

The main idea to overcome this difficulty was to use the language of explicit proof terms and a logic of proofs in lieu of modal language for provability and \( \mathbf{S4} \). Indeed, if \( p:F \) is the proof formula ‘\( p \) is a proof of \( F \)’, then the explicit version of factivity is internally provable

\[
\mathbf{PA} \vdash p:F \rightarrow F.
\]

In the Logic of Proofs \( \mathbf{LP} \), cf. [3, 5, 6], proofs are represented by proof terms constructed from proof variables and proof constants by means of functional symbols for elementary computable operations on proofs, binary \( \cdot, +\), and unary \( !\). Formulas are built by Boolean connectives from propositional atoms and those of the form \( t:F \) where \( t \) is a proof term and \( F \) is a formula.

\( \mathbf{LP} \) has the axioms and rules of classical logic along with:

- \( t:A \rightarrow A \) reflection
• \( t(A \rightarrow B) \rightarrow (s: A \rightarrow [t \cdot s]:B) \)  
  **application**

• \( t:A \rightarrow [t + s]:A \)  
  **sum**

• \( t:A \rightarrow !t:t:A \)  
  **proof checker**

• **Axiom Necessitation:** if \( A \) is an axiom and \( c \) a proof constant, derive \( cA \).

The principal feature of \( \text{LP} \) is its natural arithmetical semantics, according to which \( t:F \) is interpreted as ‘\( t \) is a proof of \( F \).’

Furthermore, \( \text{LP} \) has the ability to realize all \( \text{S4} \) theorems by restoring corresponding proof terms inside occurrences of modality. A **forgetful projection** of an \( \text{LP} \)-formula \( F \) is a modal formula obtained by replacing all assertions \( t(\cdot) \) in \( F \) by \( \square(\cdot) \).

**Realization Theorem** [2] [3]: \( \text{S4} \) is the forgetful projection of \( \text{LP} \).

That the forgetful projection of \( \text{LP} \) is \( \text{S4} \)-compliant is a straightforward observation. The converse has been first established by presenting an algorithm which substitutes proof terms for all occurrences of modalities in a cut-free Gentzen-style \( \text{S4} \)-derivation of a formula \( F \), thereby producing a formula \( F^r \) derivable in \( \text{LP} \). The resulting realization respects Skolem’s idea that negative occurrences of existential quantifiers over proofs (hidden in the modality of provability) are realized by free proof variables whereas positive occurrences are realized by functions of those variables.

Realization Theorem provides an exact semantics for \( \text{S4} \) in \( \text{LP} \). To complete building a provability BHK semantics for \( \text{IPC} \) it is now sufficient to note that \( \text{LP} \) has a natural interpretation as a logic of formal proofs in Peano arithmetic \( \text{PA} \) or a similar system capable of encoding its own proofs:

\[
\text{IPC} \leftrightarrow \text{S4} \leftrightarrow \text{LP} \leftrightarrow \text{PA},
\]

where each \( \leftrightarrow \) is an embedding.

Realization Theorem prompts reasoning in \( \text{S4} \) and automatically produces the corresponding \( \text{LP} \)-reasoning. Let us try to view the negation problem in the original BHK through \( \text{S4} \). Let \( A \) be atomic. In the original BHK setting, \( \neg A \) which is \( A \rightarrow \bot \) holds iff \( \square A \rightarrow \square \bot \), i.e., equivalently in \( \text{S4} \),

\[
\neg \square A.
\]

This means ‘\( A \) has no proofs’ and yields the aforementioned **trivialization** and **irrelevance of witnesses** distortions. Coincidentally, \( \text{S4}/\text{LP} \)-based BHK semantics of negation and implication is compliant with the Kreisel second clause criticism (cf. [6] [12]) which requires an additional provability condition

\[
\square \neg \square A.
\]

In a general BHK setting, the suggested refining of the BHK clause for implication is
• a proof of $A \rightarrow B$ is a pair of constructions $(p, v)$ such that $v(x)$ verifies that if $x$ is a proof of $A$, then $p(x)$ is a proof of $B$.

Let $A$ be atomic. In the $S4/LP$-based BHK, we first Gödel-translate $\neg A$:

$$\Box \neg \Box A.$$ 

Then we realize the result in LP respecting polarities:

$$v(x) : \neg x : A$$

for some proof term $v(x)$. This reads

$$v(x) \text{ is a proof that } x \text{ is not a proof of } A.$$ 

Such adjusted semantics cures trivialization and irrelevance of witnesses defects.

## 4 Constructive truth, falsity and consistency

Can intuitionistic approach provide new insights on what is constructive in classical mathematics?

As usual, we will ignore difference between a syntactic object $X$, its Gödel number $\langle X \rangle$ and the corresponding arithmetical numeral $\overline{\langle X \rangle}$ when safe. Let $tY$ be a shorthand for the standard formula $\text{Proof}(t, Y)$ stating that ‘$t$ is a proof of $Y$ in PA,’ $\Box Y$ stand for $\text{Provable}(Y)$, i.e., $\exists x(x : Y)$. For details, cf. [9, 17, 18, 23, 24].

We use $\bot$ for the propositional constant “false” which in the arithmetical context can be equivalently read as $0 = 1$. We stay on the common sense mathematical ground which assumes soundness of PA with respect to the standard model of arithmetic.

Traditionally, intuitionism reads an arithmetical sentence $F$ in a constructive manner. This however could alter the meaning of $F$. Our idea is to preserve the classical meaning of $F$ so we don’t venture “inside” $F$ and treat $F$ as atomic for a BHK-style interpretation.

**Definition 1** An arithmetical sentence $F$ is **constructively true** iff

$$\text{PA} \vdash F.$$ 

Informally, $F$ should be BHK-true as an intuitionistic atom. By $S4/LP$ realization this means that PA proves $tF$ for some $t$ which is equivalent to $\text{PA} \vdash F$.

$F$ is **constructively false** iff

$$\text{PA} \vdash \forall x \Box \neg x : F.$$ 

(5)

Conceptually, we read ‘$F$ yields $\bot$’ intuitionistically in a BHK fashion and conclude that for some $v(x)$,

$$\text{PA} \vdash \forall x \ v(x) : \neg x : F.$$ 

(6)
By some proof theory, (5) is equivalent to ‘(6) holds for some provably total computable term \(v(x)\).’ Indeed, (6) obviously yields (5). Assume (5). We can describe \(v(x)\) informally. Since \(u:F\) is decidable, given \(x\), enumerate proofs in \(\text{PA}\) until a proof of \(\neg x:F\) is met. Since \(\text{PA} \vdash \forall x \exists y(y \neq x:F)\), \(v(x)\) is provably total.

Gödel’s consistency formula, \(\text{Con}_{\text{PA}}\), is

\[
\text{Con}_{\text{PA}} = \forall x \neg x: \bot.
\]

**Definition 2** **Constructive consistency** is a formula \(\text{CCon}_{\text{PA}}\) stating that *for each number it is provable that it is not a proof of a contradiction*:

\[
\text{CCon}_{\text{PA}} = \forall x \Box \neg x: \bot.
\]

**Theorem 1** \(\text{PA}\) proves its own constructive consistency:

\[
\text{PA} \vdash \text{CCon}_{\text{PA}}.
\]

**Proof.** First, we check that

\[
\text{PA} \vdash \Box \bot \rightarrow \text{CCon}_{\text{PA}}.
\]

Indeed, note that \(\text{PA} \vdash \Box \bot \rightarrow \Box \neg x: \bot\). By generalization,

\[
\text{PA} \vdash \Box \bot \rightarrow \forall x \Box \neg x: \bot \ (= \text{CCon}_{\text{PA}}).
\]

Furthermore,

\[
\text{PA} \vdash \neg \Box \bot \rightarrow \text{CCon}_{\text{PA}}.
\]

Indeed, by first-order logic, \(\text{PA} \vdash x: \bot \rightarrow \exists x (x: \bot)\), hence

\[
\text{PA} \vdash \neg \Box \bot \rightarrow \neg x: \bot.
\]

By \(\Sigma_1\)-completeness of \(\text{PA}\), cf. \[24\], \(\text{PA} \vdash \neg x:F \rightarrow \Box \neg x:F\), hence \(\text{PA} \vdash \neg \Box \bot \rightarrow \Box \neg x:F\). By generalization,

\[
\neg \Box \bot \rightarrow \forall x \Box \neg x: \bot \ (= \text{CCon}_{\text{PA}}).
\]

\(\square\)

To the extent to which constructive consistency \(\text{CCon}_{\text{PA}}\) can be accepted as a possible formalization of the notion of consistency of \(\text{PA}\), Theorem 1 removes Gödel’s impossibility spell from the idea of proving consistency internally. However, Theorem 1 alone does not provide a finitary consistency proof for \(\text{PA}\). We discuss these foundational matters in the Section 5.
5 Holy Grail consistency and Hilbert’s program

Holy Grail consistency (HG-consistency) of PA is the property

\[ \text{any finite sequence } S \text{ of formulas is not a derivation of a contradiction} \]

which is identical with Hilbert’s specification (11) for PA, we are just giving it a name.

HG-consistency for PA is not directly formalizable by a single arithmetical formula in which derivations in PA are represented by their codes since standard natural numbers are not definable in PA.

\[ \text{So, is HG-consistency of PA provable by finitary means?} \] (7)

5.1 Unprovability of Con\(_{\text{PA}}\) is not an answer

Despite widespread opinion, Gödel’s Second Incompleteness Theorem does not answer (7). The consistency formula for PA

\[ \text{Con}_{\text{PA}} = \forall x \neg \neg x : \bot, \]

yields HG-consistency but not the other way around. Semantically, \( \forall x \neg \neg x : \bot \) claims that in no model (possibly nonstandard) of PA is there any number (possibly nonstandard) that encodes a proof of a contradiction. So, Con\(_{\text{PA}}\) is a kind of uniform consistency which looks stronger than HG-consistency, and IS in fact stronger, since HG-consistency can be established by finitary means, cf. Section 6, but Con\(_{\text{PA}}\) cannot.

5.2 Provability of CCon\(_{\text{PA}}\) in PA is not an answer either

If the consistency formula \( \forall x \neg \neg x : \bot \) were internally provable, this would yield that

\[ \text{PA } \vdash p \forall x \neg x : \bot \]

for some proof \( p \). By G2, this is not possible. Constructive consistency offers a more flexible approach: it allows the aforementioned certification \( p \) to depend on \( x \), \( p = p(x) \) and asks whether

\[ \text{PA } \vdash \forall x p(x) : \neg x : \bot. \]

In a general form this is a question of whether

\[ \text{PA } \vdash \forall x \exists y (y : \neg x : \bot), \]

i.e.

\[ \text{PA } \vdash \text{CCon}_{\text{PA}} \]

which was answered affirmatively in Theorem 1. However, the argument

\[ \text{HG-consistency of PA is finitarily provable because } \text{PA } \vdash \text{CCon}_{\text{PA}} \]
is circular since it relies on soundness of PA and hence does not actually prove HG-consistency of PA. After all, any inconsistent T proves its own constructive consistency.

Theorem [1] does not produce a real mathematical proof of HG-consistency. In a formalization process we take a mathematical proof and formalize it as a formal derivation in a given theory; logicians have learned how to do this. If such a formalization is possible, then a correct mathematical proof yields a correct formal derivation.

With \( PA \vdash \text{CCon}_{PA} \) we are facing the opposite problem, deformalization: given that a sentence is formally provable in a theory T, produce a rigorous mathematical proof of this statement. This does not necessarily work, e.g., when T is inconsistent, or T is not sound, like \( PA + \neg \text{Con}_{PA} \), etc. Deformalization can work when T is sound, but the assumption of soundness of T is stronger than the goal, consistency of F. So deformalization is useless for proving consistency and we have to do it in the reverse order: a mathematical proof of HG-consistency of PA first, and its formalization in PA second.

### 5.3 Sufficient conditions for finitary HG-consistency proofs

To establish HG-consistency of PA by finitary tools one has

1. for each PA-proof S to provide a mathematically rigorous proof that S does not prove \( 0 = 1 \);

2. check that all constructions and their properties used in the proof are finitary. In the current context, this requirement is often interpreted as “formalizable in PA.”

Alone, neither of 1 nor 2 is sufficient for claiming a finitary proof of HG-consistency. Indeed, the provability of constructive consistency as demonstrated in Theorem [1] secures 2, but not 1. It does not provide a mathematical proof that no derivation in PA derives a contradiction; it rather shows that constructive consistency holds in PA regardless to whether PA is consistent or not.

The classical soundness-in-the-standard-model argument proves HG-consistency and hence satisfies 1, but uses tools not formalizable in PA (hence does not satisfy 2). Indeed, all axioms of PA are true in the standard model of arithmetic, the logical rules respect arithmetical truth, hence \( 0 = 1 \) being not true cannot be derived in PA. This is a valid mathematical argument which is quite sufficient for a “normal” mathematician. However, it speaks about “truth in the standard model” which is not formalizable in PA due to limited expressiveness of the first-order language (which for a “normal” mathematician might look like a mere technicality).

### 6 Finitary proof of Holy Grail Consistency

Although the internal provability of constructive consistency alone does not settle the question of finitary demonstration of HG-consistency, it suggests looking for finitary
case-by-case proofs. Such case-by-case approach is native for PA; after all, the induction principle there is represented by case-by-case schemas of first-order formulas rather than by a single formula.

Once we know what to look for, finitary case-by-case proofs of Holy Grail consistency readily suggest themselves. Here is one.

In metamathematics of first-order arithmetic, there is a well-known construction called partial truth definitions, cf. [10, 17, 18, 21, 24]. Namely, for each \( n = 0, 1, 2, \ldots \) we inductively build a \( \Sigma_{n+1} \) formula

\[
Tr_n(x, y)
\]
called truth definition for \( \Sigma_n \) formulas which satisfies natural properties of a truth predicate. When \( \varphi \) is a \( \Sigma_n \)-formula and \( y \) is a sequence encoding values of the parameters in \( \varphi \) then \( Tr_n(\varphi, y) \) defines the truth value of \( \varphi \) on \( y \).

Let \( y \) be a code of a finite sequence of numbers and \( y_i \) denote the \( i \)-th number in \( y \). Then the following conditions hold ([10, 17, 18, 21, 24]):

**Lemma 1**

- \( Tr_n(\varphi^\top, y) \) satisfies the usual properties of truth with respect to boolean connectives, quantifiers, and rule Modus Ponens for each \( \varphi \in \Sigma_n \), and these properties are naturally derivable using \( \Sigma_{n+1} \) induction.
- \( PA \) naturally proves Tarski’s condition for any \( \Sigma_n \)-formula \( \varphi \):

\[
Tr_n(\varphi^\top, y) \equiv \varphi(y_1, y_2, \ldots, y_k).
\]

*In particular, \( \neg Tr_n(\varphi^{0=1^\top}) \) is naturally provable.*

- \( Tr_n(\varphi^A, y) \) is naturally provable for any axiom \( A \) of PA of depth \( \leq n \).

Note that all the proofs in Lemma 1 are valid finitary arguments, which are mathematically rigorous by their own natural merits. So, Lemma 1 does not make any metamathematical assumptions about PA, and just uses a formal language of PA for rigorous bookkeeping.

Given a finite sequence \( S \) of formulas which is a legitimate PA-derivation, we first calculate \( n \) such that all formulas from \( S \) have depth \( \leq n \). Then, by a step by step induction on the length of \( S \) we check that for any formula \( \varphi \) in \( S \) with parameters \( y \), the property \( Tr_n(\varphi, y) \) holds. This is an immediate corollary of Lemma 1 since all PA-axioms satisfy \( Tr_n \) and each instance of Modus Ponens respects \( Tr_n \) as well. So, \( Tr_n \) serves as an invariant for all formulas from \( S \). Since, by Lemma 1 \( 0=1 \) does not satisfy \( Tr_n \), \( 0=1 \) cannot occur in \( S \).

We argue that this proof satisfies the desiderata for a finitary proof of Holy Grail consistency.
1. This is a mathematical proof by normal standards of rigor acceptable for a general mathematician.

2. The constructions and required properties used in this argument are formalizable in PA: partial truth definitions, compliance of truth definitions with PA-derivation rules, etc. Hence for each PA-derivation $S$, we have proved in PA that $S$ does not contain $0=1$.

Note that this argument cannot be strengthened to establish provability of Gödel’s (uniform) consistency $\text{Con}_{PA} = \forall x \neg x: \bot$ because, by Tarski’s Theorem, there is no one truth definition in PA that covers all formulas.

7 Proof theory of constructive truth and falsity

7.1 Normal forms

First, we find a provably equivalent quantifier-free formulation of constructive falsity which we call “normal forms.”

The negation of the constructive falsity sentence,

$$\exists x \neg \Box \neg x: F,$$

is a kind of a provability predicate which is true iff $F$ is provable. The details will be clear after the following

Lemma 2 $\text{PA} \vdash (8) \iff \neg \Box \bot \land \Box F$.

Proof. Argue in PA. $(8) \rightarrow \neg \Box \bot$ is straightforward. To check $(8) \rightarrow \Box F$ assume $\neg \Box F$, i.e., $\forall x \neg x: F$. By $\Sigma_1$-completeness of PA,

$$\neg x: F \rightarrow \Box \neg x: F,$$

by generalization and some first-order reasoning,

$$\forall x \neg x: F \rightarrow \forall x \Box \neg x: F.$$

Hence $\forall x \Box \neg x: F$ which is $\neg (8)$. This proves the “$\rightarrow$” direction.

Now assume $\neg \Box \bot$ and $\Box F$. Then $\exists x (x: F)$ and, by $\Sigma_1$-completeness, $\exists x \Box x: F$. Let $t$ be such an $x$, i.e., $\Box t: F$. We claim that $\neg \Box \neg t: F$, since otherwise we would have $\Box t: \neg F$ and $\Box t: F$ which yields $\Box \bot$. So, $\exists x \neg \Box \neg x: F$. □

Theorem 2 [Normal Form Theorem] $F$ is constructively false iff

$$\text{PA} \vdash \text{Con}_{PA} \rightarrow \neg \Box F.$$
**Proof.** By definition, \( F \) is constructively false iff \( \text{PA} \vdash \neg (\Box) \) which, by Lemma 2, is equivalent to \( \text{PA} \vdash \neg \Box \bot \rightarrow \neg \Box F \), i.e., \( \text{PA} \vdash \text{Con}_{\text{PA}} \rightarrow \neg \Box F \).

\( \Box \)

Equivalent forms of constructive falsity of \( F \) are \( \text{PA} \vdash \neg \Box \bot \rightarrow \neg \Box F \), \( \text{PA} \vdash \Box F \rightarrow \Box \bot \).

### 7.2 Sanity Theorem

The following Sanity Theorem demonstrates that constructive truth/falsity satisfy natural desired properties. The main idea of these notions is to provide constructive BHK-style refinement of the classical truth values of arithmetical formulas which respects arithmetical provability and refutability\(^1\). The list of these natural properties corresponds to 1–5 of Sanity Theorem. Items 1–3 are straightforward, item 4 is a non-triviality requirement, item 5 shows that constructive truth/falsity respect arithmetical provability internally, at the level of provable implications.

Note that other natural BHK-inspired formalizations of constructively true/false do not seem to pass this sanity test. For example, taking \( \text{PA} \vdash \Box \neg \Box F \) for “\( F \) is constructively false” does not satisfy 2 with \( \bot \) as \( F \).

**Theorem 3** [Sanity Theorem]

1. \( \text{PA} \vdash F \) yields ‘\( F \) is constructively true’;
2. \( \text{PA} \vdash \neg F \) yields ‘\( F \) is constructively false’;
3. ‘constructively true’ and ‘constructively false’ are mutually exclusive;
4. ‘constructively true/false’ do not coincide with ‘provable/refutable’;
5. ‘constructively true’ and ‘constructively false’ are monotone in the Lindenbaum algebra of \( \text{PA} \): if \( \text{PA} \vdash F \rightarrow G \), then
   - ‘\( F \) is constructively true’ yields ‘\( G \) is constructively true,’
   - ‘\( G \) is constructively false’ yields ‘\( F \) is constructively false.’

**Proof.**

1. By definitions, \( \text{PA} \vdash F \) iff ‘\( F \) is constructively true.’
2. Let \( \text{PA} \vdash \neg F \). Then \( \text{PA} \vdash \Box \neg F \) and, by modal-style reasoning, \( \text{PA} \vdash \Box F \rightarrow \Box \bot \).

Note that if \( F \) is constructively true, then, by 2, \( \neg F \) is constructively false. However, if \( F \) is constructively false, then \( \neg F \) can be either constructively true (e.g., when \( F \) is \( 0 = 1 \)), or constructively false (e.g., when \( F \) is \( \neg R \) from Theorem 5), or neither (e.g., \( F \) is \( \text{Con}_{\text{PA}} \), by Lemma 3 and Theorem 4).

\(^1\)Formula \( F \) is refutable if \( \text{PA} \vdash \neg F \).
3. Suppose \( F \) is constructively true and false. Then \( \text{PA} \vdash F \) and \( \text{PA} \vdash \Box F \rightarrow \Box \bot \), hence \( \text{PA} \vdash \Box F \) and \( \text{PA} \vdash \Box \bot \) which contradicts soundness of \( \text{PA} \) with respect to the standard model.

4. It suffices to find a formula which is true (hence not refutable) but constructively false.

**Lemma 3** Consistency formula \( \text{Con}_{\text{PA}} = \neg \Box \bot \) is true and constructively false.

**Proof.** \( \text{Con}_{\text{PA}} \) is true in the standard model since \( \text{PA} \) is sound, hence consistent. Furthermore, since, by the formalized L"ob Theorem (cf. [9, 24]),

\[
\text{PA} \vdash \Box \neg \Box \bot \rightarrow \Box \bot,
\]

\( \text{Con}_{\text{PA}} \) is constructively false. 

So, \( \text{Con}_{\text{PA}} \) is constructively false but not refutable.

5. In the Lindenbaum algebra of \( \text{PA} \),

\[
[G] \preceq [F] \iff \text{PA} \vdash F \rightarrow G,
\]

constructive truth is closed downward (immediate) and constructive falsehood is closed upward. Indeed, suppose \( \text{PA} \vdash F \rightarrow G \), then \( \text{PA} \vdash \Box F \rightarrow \Box G \). If, in addition, \( G \) is constructively false, then \( \text{PA} \vdash \Box G \rightarrow \Box \bot \) which yields \( \text{PA} \vdash \Box F \rightarrow \Box \bot \) as well. 

\[\square\]

7.3 Inconsistency is not constructively false.

**Theorem 4** Inconsistency \( \neg \text{Con}_{\text{PA}} = \Box \bot \) is not constructively false.

**Proof.** Immediate from Normal Form Theorem 2, since \( \text{PA} \not\vdash \Box \Box \bot ightarrow \Box \bot \): otherwise, by L"ob’s Theorem, \( \text{PA} \vdash \Box \bot \) which is not the case.

So, inconsistency formula \( \neg \text{Con}_{\text{PA}} \) is neither constructively false, nor constructively true.

7.4 Rosser sentences

By Rosser’s Theorem, there is a sentence \( R \), for which independence in \( \text{PA} \) follows from simple consistency of \( \text{PA} \): if \( \text{PA} \) is consistent, then neither \( R \) nor its negation \( \neg R \) is provable, cf. [23].

**Theorem 5** Rosser sentence \( R \) and its negation \( \neg R \) are both constructively false.

**Proof.** The proof of Rosser’s Theorem is syntactic and can be formalized in \( \text{PA} \), cf. [27]:

\[
\text{PA} \vdash \neg \Box \bot \rightarrow (\neg \Box R \land \neg \Box \neg R).
\]

By Normal Form Theorem 2 both \( R \) and \( \neg R \) are constructively false. 

\[\square\]
Constructive liar sentence

Theorem 6 There is a true independent in PA sentence which is not constructively false.

Proof. Using the fixed-point lemma, find an arithmetic sentence $L$ such that

$$PA \vdash L \leftrightarrow 'L \text{ is constructively false}'. \tag{9}$$

Formally,

$$PA \vdash L \leftrightarrow (\Box L \rightarrow \Box \bot).$$

If $PA \vdash L$, then $PA \vdash \Box L$ and, by (9), $PA \vdash \Box \bot$ which is not the case.

If $PA \vdash \neg L$, then, by Sanity Theorem item 2, $L$ is constructively false, hence, $PA \vdash \Box L \rightarrow \Box \bot$. By the fixed point (9), $PA \vdash L$ - a contradiction in PA. So, $L$ is independent and not constructively false.

Note that $L$ is classically true: otherwise $\Box L$ is false and $\Box L \rightarrow \Box \bot$ is vacuously true. By the fixed point (9), $L$ ought to be true as well.

\[\square\]

Summary table of classical and constructive truth/falsity

Here is the summary table of possible overlaps of classical and constructive truth/falsity.

| Intersection of classes                  | Example          |
|-----------------------------------------|------------------|
| True and constructively true            | $0 = 0$          |
| True and constructively false           | $\text{Con}_{PA}, R$ |
| True and neither                        | Constructive Liar $L$ |
| False and constructively true           | $\emptyset$      |
| False and constructively false          | $0 = 1, \neg R$  |
| False and neither                       | $\neg \text{Con}_{PA}$ |

Constructive truth/falsity of dual pairs

Consider dual pairs of arithmetical sentences $F$ and $\neg F$. If one of them is constructively true, hence provable, then the other one is refutable, hence constructively false.

We show that any combinations ‘constructively false’ (we call it case $f$) and ‘neither constructively true nor constructively false’ (case $n$) are possible for dual pairs of arithmetical sentences.

Case $\{f, f\}$ is realized by Rosser sentences $R$ and $\neg R$, cf. Theorem 5.

Case $\{f, n\}$, subcase $F$ is true is realized by $F = \text{Con}_{PA}$, cf. Lemma 3 and Theorem 4.

Let us do case $\{f, n\}$, subcase $F$ is false.
Lemma 4 There is an arithmetical sentence $F$ which is false and constructively false whereas $\neg F$ is neither constructively true nor constructively false.

Proof. Consider $F = \neg \square \bot \land \square \bot$. In a different notation, $F$ is nothing but

$$\text{Con}_{\text{PA}} \land \neg \text{Con}_{\text{PA} + \text{Con}_{\text{PA}}}.$$  

$F$ is false, since $\square \bot$ is false.

$F$ is constructively false. By Normal Form Theorem 2, it suffices to check that $\text{PA} \vdash \square F \to \bot$. Argue in $\text{PA}$: $\square F$ implies $\square \neg \bot$ which, by the formalized L"ob Theorem, yields $\bot$.

$\neg F$ is neither constructively true nor constructively false. Indeed, in $\text{PA}$, $\neg F$ is equivalent to $\square \neg \bot \to \bot$ which is not provable in $\text{PA}$, since otherwise, by L"ob Theorem $\text{PA}$ would prove $\bot$. Therefore, $\neg F$ is not constructively true.

To check that $\neg F$ is not constructively false, it suffices to prove that $\text{PA} \not\vdash \square \neg F \to \bot$. In $\text{PA}$, $\square \neg F$ is equivalent to $\square (\square \bot \to \bot)$, which, by the formalized L"ob Theorem and some modal-style reasoning in $\text{PA}$ is equivalent to $\square \bot$. So, the problem has been reduced to checking that $\square \bot \to \bot$ is not derivable in $\text{PA}$. If it were, then, by L"ob Theorem, $\text{PA}$ would derive $\bot$ which is not the case.

Let us now do case \{n, n\}.

Lemma 5 There is an arithmetical sentence $F$ such that both $F$ and $\neg F$ are neither constructively true nor constructively false.

Proof. It suffices to find $F$ such that both $F$ and $\neg F$ are not constructively false, by the aforementioned discussion in this section, then neither $F$ nor $\neg F$ can be constructively true. So, by Normal Form Theorem 2, we need to find an $F$ such that $\text{PA} \not\vdash \square F \to \bot$ and $\text{PA} \not\vdash \square \neg F \to \bot$.

To find such an $F$, we use the technique developed within the framework of the Provability Logic $\text{GL}$, cf. [4, 9, 24]. In particular, we will need the uniform arithmetical completeness theorem for $\text{GL}$ established independently in [1, 7, 8, 20, 26].

Lemma 6 [Uniform Arithmetical Completeness of Provability Logic] There is an arithmetical interpretation $\ast$ such that for any modal formula $M$,

$$\text{GL} \vdash M \iff \text{PA} \vdash M^\ast.$$  

Lemma 7 $\text{GL} \not\vdash \square p \to \bot$ and $\text{GL} \not\vdash \square \neg p \to \bot$ for a propositional letter $p$.

Proof. By soundness of $\text{GL}$ with respect to arithmetical interpretations, it suffices to deliver arithmetical sentences $X$ and $Y$ such that $\text{PA} \not\vdash \square X \to \bot$ and $\text{PA} \not\vdash \square \neg Y \to \bot$. Obviously, $X = \bot \to \bot$ and $Y = \neg X$ work: they both reduce to showing that $\text{PA} \not\vdash \square (\bot \to \bot) \to \bot$ which is equivalent to $\text{PA} \not\vdash \square \bot$ and obvious.
By Lemma 6, there is an arithmetical sentence \( p^* \) such that both \( \text{PA} \nvdash \Box p^* \rightarrow \Box \bot \) and \( \text{PA} \nvdash \Box \neg p^* \rightarrow \Box \bot \). 

7.8 Beyond arithmetic

What about case-by-case consistency of other theories containing \( \text{PA} \)? Each of these theories proves its own constructive consistency (as in Theorem 1). That being said, a more nuanced \( a \ priori \) analysis can be provided.

Assume that theory \( T \) provably contains \( \text{PA} \) and let \( \Box_T \) and \( \vdash_T \) denote provability and proof predicates respectively for \( T \). Consider formulas

- Consistency: \( \text{Con}_T = \forall x \neg x_T \bot; \)
- Finitary consistency: \( \text{FCon}_T = \forall x \Box_{PA} \neg x_T \bot; \)
- Constructive consistency: \( \text{CCon}_T = \forall x \Box_T \neg x_T \bot. \)

It is easy to check that \( \text{PA} \) proves

\[
\text{Con}_T \rightarrow \text{FCon}_T \rightarrow \text{CCon}_T.
\]  

(10)

Obviously, \( \text{PA} \nvdash \text{Con}_T \).

**Theorem 7** \( \text{PA} \vdash \text{CCon}_T \).

**Proof.** Indeed, argue in \( \text{PA} \).

If \( \Box_T \bot \), then vacuously \( \Box_T \neg x_T \bot \) and hence \( \forall x \Box_T \neg x_T \bot \). Therefore,

\[
\Box_T \bot \rightarrow \text{CCon}_T.
\]

If \( \neg \Box_T \bot \), then \( \forall x \neg x_T \bot \) and, by \( \Sigma_1 \)-completeness of \( \text{PA} \), \( \forall x \Box_{PA} \neg x_T \bot \). Therefore,

\[
\neg \Box_T \bot \rightarrow \text{CCon}_T.
\]

The question of whether \( \text{PA} \) proves \( \text{FCon}_T \) is related to the question of whether there is a finitary proof of case-by-case consistency of \( T \). In particular, if for each \( S \) we can provide a finitary proof that \( S \) is not a \( T \)-proof of a contradiction and this procedure is formalizable in \( \text{PA} \), \( \text{PA} \) should be able to prove \( \forall x \Box_{PA} \neg x_T \bot \).

The following Theorem 8 and Corollary 1 are due to Morgan Sinclaire [22] (private communication on Feb 22, 2019, discussing the first arxiv posting of this paper). They show limits of how far we can go with proving finitary consistency. 

16
Theorem 8 \( \text{PA} \not\vdash \text{FCon}_{\text{PA} + \text{Con}_{\text{PA}}} \).

Proof. Define \( \text{PA}' \) as \( \text{PA} + \text{Con}_{\text{PA}} \). Since, by \( \text{G2} \) applied to \( \text{PA}' \),
\[
\text{PA} \not\vdash \text{Con}_{\text{PA}} \rightarrow \text{Con}_{\text{PA}'},
\]
it suffices to establish
\[
\text{PA} \vdash \text{FCon}_{\text{PA}'} \rightarrow (\text{Con}_{\text{PA}} \rightarrow \text{Con}_{\text{PA}'}).
\]
Argue in \( \text{PA} \). By contrapositive, assume \( \text{Con}_{\text{PA}} \) and \( \neg \text{Con}_{\text{PA}'} \). This yields \( p_{\text{PA}'} \perp \) for some \( p \). By \( \Sigma_1 \)-completeness, \( \Box(p_{\text{PA}'} \perp) \). From \( \text{Con}_{\text{PA}} \), \( \neg \Box \neg (p_{\text{PA}'} \perp) \) which yields \( \neg \text{FCon}_{\text{PA}'} \):
\[
\exists x \neg \Box \neg (x_{\text{PA}'} \perp).
\]

Corollary 1 \( \text{PA} \not\vdash \text{FCon}_T \) for any \( T \supseteq \text{PA} + \text{Con}_{\text{PA}} \).

Furthermore, Morgan Sinclaire [22] and Taishi Kurahashi [19] in independent private communications have pointed out that \( \text{FCon}_T \) is \( \text{PA} \)-provably equivalent to
\[
\text{Con}_{\text{PA}} \rightarrow \text{Con}_T
\]
which can be established by the same reasoning as in the proof of Normal Form Theorem [2]. This observation also immediately implies Theorem 8 and Corollary 1.

As a corollary, we conclude that, generally speaking, neither of converse implications from (10) holds. Indeed, for \( T = \text{PA} \), \( \text{PA} \vdash \text{FCon}_T \) (by Theorem 7), but \( \text{PA} \not\vdash \text{Con}_T \) (by \( \text{G2} \)). For \( T = \text{PA} + \text{Con}_{\text{PA}} \), \( \text{PA} \vdash \text{CCon}_T \) (by Theorem 7), but \( \text{PA} \not\vdash \text{FCon}_T \) (by Theorem 8).

The results of this section show how much we can get by switching to case-by-case consistency. In Section 6 we demonstrated case-by-case consistency of \( \text{PA} \) by finitary tools. By Theorem 7 it is not impossible for any theory \( T \) containing \( \text{PA} \) to prove its case-by-case consistency, and this fact itself, in the form of \( \text{CCon}_T \), is provable at the low, \( \text{PA} \)-level. So, there are no Gödelian roadblocks on the path to internal proofs of consistency of theories; this can be regarded as a main message of this paper.

However, within the current circle of ideas, there are Gödel-style limitations on finitary consistency proofs \( p(S) \) for each derivation \( S \) in a theory \( T \) provably containing \( \text{PA} + \text{Con}_{\text{PA}} \): they need methods spilling over \( \text{PA} \).

8 Acknowledgements

These results have been delivered, in part, at the Festkolloquium on the occasion of Gerhard Jäger’s retirement, University of Bern, on December 13, 2018.

The author is grateful to Melvin Fitting, Gerhard Jäger, Vladimir Krupski, Hirohiko Kushida, Taishi Kurahashi, Eoin Moore, Elena Nogina, Vincent Peluce, and Morgan Sinclaire for helpful suggestions.
References

[1] S. Artemov. Extensions of Arithmetic and Modal Logics (in Russian). Ph.D. Thesis, Moscow State University - Steklov Mathematical Institute, 1979.

[2] S. Artemov. Operational Modal Logic. Technical Report MSI 95-29, Cornell University, 1995.

[3] S. Artemov. Explicit provability and constructive semantics. Bulletin of Symbolic Logic, 7(1):1–36, 2001.

[4] S. Artemov and L. Beklemishev. Provability Logic. In Handbook of Philosophical Logic, 2nd Edition, pp. 189–360. Springer, Dordrecht, 2005.

[5] S. Artemov and M. Fitting. Justification Logic. In: Zalta, Edward N. (ed), The Stanford Encyclopedia of Philosophy, 2012.

[6] S. Artemov and M. Fitting. Reasoning With Reasons. A Study of Justification Logic. Cambridge University Press, 2019.

[7] A. Avron. On modal systems having arithmetical interpretations. The Journal of Symbolic Logic, 49(3):935–942, 1984.

[8] G. Boolos. Extremely undecidable sentences. The Journal of Symbolic Logic 47(1): 191–196, 1982.

[9] G. Boolos. The Logic of Provability. Cambridge University Press, Cambridge, 1993.

[10] S. Buss. First-order proof theory of arithmetic. Pages 79–147 of: Buss, S. (ed), Handbook of Proof Theory. Elsevier, 1998.

[11] D. van Dalen and A.S. Troelstra. Constructivism in Mathematics An Introduction, vol. I. Studies in Logic and the Foundations of Mathematics vol. 121. Elsevier, 1988.

[12] W. Dean and H. Kurokawa. Kreisel’s Theory of Constructions, the Kreisel-Goodman paradox, and the second clause. Advances in Proof-Theoretic Semantics, pp. 27–63. Springer, Cham, 2016.

[13] M. Detlefsen. On interpreting Gödel’s second theorem. Journal of Philosophical Logic, 8:297–313, 1979.

[14] M. Detlefsen. Hilbert’s Program. Dordrecht: Reidel, 1986.

[15] M. Detlefsen. What does Gdel’s second theorem say? Philosophia Mathematica, 9:37–71, 2001.
[16] K. Gödel. Eine Interpretation des intuistionistischen Aussagenkalkuls. *Ergebnisse eines mathematischen Kolloquiums*, 4, 39–40, 1933.

[17] P. Hájek and P. Pudlak. *Metamathematics of First-Order Arithmetic*, Vol 3. Cambridge University Press, 2017.

[18] R. Kaye. *Models of Peano arithmetic*. Oxford: Clarendon Press, 1991.

[19] T. Kurahashi. On your new recent paper ‘The Provability of Consistency.’ Private Communication. February 24, 2019.

[20] F. Montagna. On the diagonalizable algebra of Peano arithmetic. *Bollettino della Unione Matematica Italiana*, 16(5):795–812, 1979.

[21] P. Pudlak. The Lengths of Proofs. Pages 547–638 of: Buss, S. (ed), *Handbook of Proof Theory*. Elsevier, 1998.

[22] M. Sinclaire. FConT. Private Communication. February 22, 2019.

[23] C. Smoryński. The incompleteness theorems. Pages 821–865 of: Barwise, J. (ed), *Handbook of mathematical logic 4*. North Holland, 1977.

[24] C. Smoryński. *Self-Reference and Modal Logic*. New York: Springer-Verlag, 1985.

[25] A.S. Troelstra and H. Schwichtenberg. *Basic Proof Theory*. No. 43. Cambridge University Press, 2000.

[26] A. Visser. *Aspects of Diagonalization and Provability*. PhD dissertation, Drukkerij Elinkwijk, 1981.

[27] A. Visser. Peano’s smart children: A provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic*, 30(2):161–196, 1989.

[28] R. Zach. *Hilbert’s Program*, The Stanford Encyclopedia of Philosophy (Spring 2016), URL = https://plato.stanford.edu/archives/spr2016/entries/hilbert-program/