D.1

The Incompleteness Theorems

C. SMORYNSKI

Contents

* This subsection covers an advanced topic

HANDBOOK OF MATHEMATICAL LOGIC
Edited by J. Barwise

1. Hilbert's Program

Mathematics, at the turn of the century, was plagued by various difficulties ranging from antinomies and paradoxes to inconsistencies both formal and personal. There had been difficulties earlier in mathematics, but these had been removed or detoured: The Greeks eventually shrugged their shoulders and admitted irrational numbers; the analysts avoided paradoxes involving infinitesimals by finally isolating and rigorizing the concepts of limit and continuity. Even in set theory, the solution to the problem of the paradoxes had been offered as early as 1908 by Zermelo: One must first know what one is talking about before one can axiomatize a subject. Thus, instead of taking as axioms for set theory some intuitively obvious properties of finite sets, some obvious properties of the set of all subsets of a given set, and yet some other obvious properties of a third entity — a process that almost guarantees contradictions — Zermelo first described the cumulative hierarchy and then listed axioms for this *single* entity. Until recent work on large cardinals, axioms later added were merely further properties obviously true for this hierarchy but which were formally underivable.

Sociologists would describe what transpired next in terms of "culture lag". Despite the fact that a consistent set theory was available, mathematicians continued to worry about consistency. Some even doubted the consistency of arithmetic itself! To make matters worse. L.E.J. Brouwer was making the rounds in a bizarre attempt to turn mathematics into a religion.

When, in 1920, Hermann Weyl fell prey to Brouwer's lunacy, David Hilbert decided to intervene. He observed that (REID [1970] p. 155) "What Weyl and Brouwer do comes to the same thing as to follow in the footsteps of Kronecker! They seek to save mathematics by throwing overboard all that which is troublesome.... They would chop up and mangle the science. If we would follow such a reform as the one they suggest, we would run the risk of losing a great part of our most valuable treasures!"

The vehemence with which Hilbert made the above declaration is most readily understood when one remembers that Hilbert made his name by the use of non-constructive techniques. His solution to Gordon's problem in the theory of invariants (REID [1970], Chapter V) had elicited the charge of "theology" from Gordon. Kronecker refused to believe that the theorem, which asserted the existence of objects satisfying some condition, had been proven as the objects had not been explicitly constructed. Lindemann called the technique "unheimlich". Thus, it is not surprising

that Hilbert continued (REID [1970], p. 157) "I believe that as little as Kronecker was able to abolish the irrational numbers ... just as little will Weyl and Brouwer today be able to succeed. Brouwer is not, as Weyl believes him to be, the Revolution — only the repetition of an attempted Putsch".

Even if Hilbert had faith in Zermelo's set theory, he could not use it: For, he had not to secure mathematics but to stop a Putsch. So Hilbert proposed his Conservation Program: To justify the use of abstract techniques, he would show — by as simple and concrete a means as possible — that the use of abstract techniques was *conservative* — i.e. that any concrete assertion one could derive by means of such abstract techniques would be derivable without them.

To clarify these matters, we introduce some Hilbertian jargon whose exact meaning was never delineated by Hilbert. First, in the domain of concrete mathematics, there are *finitistically meaningful* statements and *finitistic* means of proof. The finitistically meaningful statements are called *real* statements and are (say) identities of the form

$$\forall x\,(fx = gx),$$

where f, g are reasonably simple functions (e.g. primitive recursive). Finitistic proofs correspond roughly to computations or combinatorial manipulations. More complicated statements are merely *ideal* ones and, as such, have no meaning; but they can be manipulated abstractly — just as i is not a real number, but can be dealt with algebraically, freely using the fact that $i^2 = -1$. Hilbert's contention was that, just as the use of i leads to no new algebraic identities, the use of ideal statments and abstract reasoning about them would not allow one to derive any new real statements — i.e. none which were not already derivable finitistically. To refute Weyl and Brouwer, Hilbert required that this latter conservation property itself be finitistically provable.

To avail itself of a finitistic treatment, the ideal statements and abstract reasoning would have to be codified in some formal system. Then the abstract reasoning would be codified by simple combinatorial manipulations and similar simple combinatorial manipulations could be used to demonstrate this conservation.

At this point, one could try to analyze either the reasons for Hilbert's belief that this could be done or the assumptions that necessarily underly such a Program. The author does not find these topics particularly interesting and so we skip them.

The question probably on the reader's mind is: This is all very nice, but

where does *consistency* come in? For, as everyone knows, this chapter is supposed to be about consistency. Hilbert's Consistency Program is a natural outgrowth of and successor to Hilbert's Conservation Program. There are two reasons for this:

(i) Consistency is merely the assertion that some string of symbols is not derivable. Since derivations are simple combinatorial manipulations, this is a finitistically meaningful statement and ought to have a finitistic proof.

(ii) Proving consistency of the formal system encoding the abstract concepts already establishes the conservation result!

Reason (i) is straightforward and we do not discuss it. Reason (ii) is particularly important and we should comment on it. Let **R**, **I** denote formal systems encoding real statements with their finitistic proofs and ideal systems with their abstract reasoning, respectively. Let φ be a real statement $\forall x\,(fx = gx)$. Now, if $\mathbf{I} \vdash \varphi$, then there is a derivation, d, of φ from **I**. But, derivations are concrete objects and, for some real formula $P(x, y)$ encoding derivations in **I**,

$$\mathbf{R} \vdash P(d, \ulcorner \varphi \urcorner),$$

where $\ulcorner \varphi \urcorner$ is some code for φ. Now, if φ were false, one would have $fa \neq ga$ for some a and hence,

$$\mathbf{R} \vdash P(c, \ulcorner \neg \varphi \urcorner)$$

for some c. In fact, one would have the stronger assertion

$$\mathbf{R} \vdash fx \neq gx \rightarrow P(c_x, \ulcorner \neg \varphi \urcorner)$$

for some c_x depending on x. But, if **R** proves consistency of **I**, we see

$$\mathbf{R} \vdash \neg (P(d, \ulcorner \varphi \urcorner) \wedge P(c, \ulcorner \neg \varphi \urcorner)),$$

whence $\mathbf{R} \vdash fx = gx$, with free variable x, i.e. $\mathbf{R} \vdash \forall x\,(fx = gx)$.

[The above argument is a bit vague and is rife with additional assumptions. To make it rigorous, we would have to get down to the basics of encoding — which is more than we intend to do in this section. The assumptions on P are brought out in Sections 2 and 3. A formal version of the above argument appears in Section 4.]

The argument of the above paragraph clearly invited Hilbert to establish his Consistency Program: To devise a finitistic means of proving the consistency of various formal systems encoding abstract reasoning with ideal statements.

Since the Consistency Program was as broad as the general Conservation Program and, since it looked more tractable, Hilbert fixed on it, asserting (MESCHKOWSKI [1973], p. 56):

> If the arbitrarily given axioms do not contradict each other through their consequences, then they are true, then the objects defined through the axioms exist. That, for me, is the criterion of truth and existence.

In summary, Hilbert's Consistency Program had as its goal the proof, by finitistic means, of the consistency of strong systems. The solution would completely justify the use of abstract concepts. The proof would successfully repudiate Brouwer and bring Weyl back into the fold.

It's a shame that it couldn't work.

2. Gödel's theorems

In 1930, while in his twenties, Kurt Gödel made a major announcement: Hilbert's Consistency Program could not be carried out. For, he had proven two theorems which were then considered moderately devastating and which still induce nightmares among the infirm. Loosely stated, these theorems are:

FIRST INCOMPLETENESS THEOREM. *Let* **T** *be a formal theory containing arithmetic. Then there is a sentence* φ *which asserts its own unprovability and is such that*:

(i) *If* **T** *is consistent,* $\mathbf{T} \not\vdash \varphi$.

(ii) *If* **T** *is* ω*-consistent,* $\mathbf{T} \not\vdash \neg\varphi$.

SECOND INCOMPLETENESS THEOREM. *Let* **T** *be a consistent formal theory containing arithmetic. Then*

$$\mathbf{T} \not\vdash \mathrm{Con}_{\mathbf{T}},$$

where $\mathrm{Con}_{\mathbf{T}}$ *is the sentence asserting the consistency of* **T**.

The Second Theorem clearly destroys the Consistency Program. For, if **R** cannot prove its own consistency, how can it hope to prove the consistency of **I**? (**R** and **I** are as in Section 1.) Even the First Theorem does this since (1) the statement φ is real; and (2) φ is easily seen to be true. ((1) requires looking at the construction of φ; (2) is seen by observing that φ asserts its unprovability and is indeed unprovable.) Thus, the First Theorem shows that the Conservation Program cannot be carried out and, hence, that the same must hold for the Consistency Program.

Let us consider the proofs of these remarkable theorems.

2.1. Preliminaries

The clause in each theorem that **T** contain arithmetic is just a means of avoiding the problem of stating explicitly what conditions must be met. These conditions are encodability conditions and, as Gödel showed, one can do a great deal of encoding on natural numbers. We defer until Section 3 the discussion of *how* the encoding is handled and discuss here *what* is to be encoded and *where* it is to be encoded.

Throughout this chapter, **T** will be some fixed, but unspecified, consistent formal theory. For later convenience, we assume that the encoding is done in some fixed formal theory **S** and that **T** contains **S**. We do not specify **S** — it is usually taken to be a formal system of arithmetic, although a weak set theory is often more convenient. The sense in which **S** is contained in **T** is better exemplified than explained: If **S** is a formal system of arithmetic and **T** is, say, **ZF**, then **T** contains **S** in the sense that there is a well-known embedding, or *interpretation*, of **S** in **T**. It is this sort of embedding that we have in mind.

Since encoding is to take place in **S**, it will have to have a large supply of constants and closed terms to be used as codes. (E.g. in formal arithmetic, one has $\bar{0}, \bar{1}, \ldots$.) **S** will also have certain function symbols to be described shortly.

To each formula, φ, of the language of **T** is assigned a closed term, $\ulcorner\varphi\urcorner$, called the *code* of φ. [N.B. If φx is a formula with free variable x, then $\ulcorner\varphi x\urcorner$ is a closed term encoding the formula φx, with x viewed as a *syntactic object* and not as a parameter.] Corresponding to the logical connectives and quantifiers are function symbols, neg, imp, etc., such that, for all formulae φ, ψ, $\mathbf{S} \vdash \mathrm{neg}(\ulcorner\varphi\urcorner) = \ulcorner\neg\varphi\urcorner$, $\mathbf{S} \vdash \mathrm{imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) = \ulcorner\varphi \to \psi\urcorner$, etc.

Of particular importance is the substitution operator, represented by the function symbol sub. For formulae φx, terms t with codes $\ulcorner t\urcorner$,

$$\mathbf{S} \vdash \mathrm{sub}(\ulcorner\varphi x\urcorner, \ulcorner t\urcorner) = \ulcorner\varphi t\urcorner.$$

Iteration of sub allows one to define $\mathrm{sub}_3, \mathrm{sub}_4, \ldots$, such that

$$\mathbf{S} \vdash \mathrm{sub}_n(\ulcorner\varphi x_1 \cdots x_n\urcorner, \ulcorner t_1\urcorner, \ldots, \ulcorner t_n\urcorner) = \ulcorner\varphi t_1 \cdots t_n\urcorner.$$

Finally, we also encode derivations and have a binary relation $\mathrm{Prov}_{\mathbf{T}}(x, y)$ (read "x proves y" or "x is a proof of y") such that for closed t_1, t_2: $\mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(t_1, t_2)$ iff t_1 is the code of a derivation in **T** of the formula with code t_2. It follows that $\mathbf{T} \vdash \varphi$ iff $\mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(t, \ulcorner\varphi\urcorner)$ for some closed term t.

If one defines

$$\mathrm{Pr}_{\mathbf{T}}(y) \leftrightarrow \exists x\, \mathrm{Prov}_{\mathbf{T}}(x, y),$$

then one obtains a predicate asserting provability. However, it is *not* always the case that

$$(*) \qquad \mathbf{T} \vdash \varphi \quad \text{iff} \quad \mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner),$$

unless **S** is fairly *sound* (a term to be defined later). The reason is that the existential quantifier in $\mathrm{Pr}_{\mathbf{T}}$ makes it essentially an ideal statement: While a consistent theory cannot prove false real statements, $\forall x\,(fx = gx)$, it can prove false existential ones, $\exists x\,(fx = gx)$. Thus $(*)$ can fail.

The above encoding can be carried out, however, in such a way that the following important conditions are met for all sentences φ,

D1 $$\mathbf{T} \vdash \varphi \quad \text{implies} \quad \mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner).$$

D2 $$\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \urcorner).$$

D3 $$\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \wedge \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \urcorner).$$

Conditions D1–D3 are called the *Derivability Conditions.*

2.2. Proof of the Incompleteness Theorems

The Incompleteness Theorems depend on the following.

2.2.1. THEOREM (Diagonalization Lemma). *Let φx in the language of* **T** *have only the free variable indicated. Then there is a sentence ψ such that*

$$\mathbf{S} \vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner).$$

[N.B. If φ or ψ is not in the language of **S**, then by "$\mathbf{S} \vdash \cdots$", we mean that the equivalence is proven in the theory **S′** in the language of **T** whose only non-logical axioms are those of **S**. **S′** is conservative over **S**.]

PROOF. Given φx, let $\theta x \leftrightarrow \varphi(\mathrm{sub}(x, x))$ be the diagonalization of φ. Let $m = \ulcorner \theta x \urcorner$ and $\psi = \theta m$. Then we claim

$$\mathbf{S} \vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner).$$

For, in **S**, we see that

$$\begin{aligned} \psi &\leftrightarrow \theta m \leftrightarrow \varphi(\mathrm{sub}(m, m)) \\ &\leftrightarrow \varphi(\mathrm{sub}(\ulcorner \theta x \urcorner, m)) \quad (\text{since } m = \ulcorner \theta x \urcorner) \\ &\leftrightarrow \varphi(\ulcorner \theta m \urcorner) \leftrightarrow \varphi(\ulcorner \psi \urcorner). \quad \square \end{aligned}$$

We apply 2.2.1 to $\neg \mathrm{Pr}_{\mathbf{T}}(x)$.

2.2.2. THEOREM (First Incompleteness Theorem). *Let* $\mathbf{T} \vdash \varphi \leftrightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$.

Then:

(i) $\mathbf{T} \nvdash \varphi$,

(ii) *under an additional assumption*, $\mathbf{T} \nvdash \neg \varphi$.

PROOF. (i) Observe $\mathbf{T} \vdash \varphi$ implies $\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$, by D1, which implies $\mathbf{T} \vdash \neg \varphi$, contradicting the consistency of $\mathbf{T}$.

(ii) The additional assumption is a strengthening of the converse to D1, namely $\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$ implies $\mathbf{T} \vdash \varphi$.

We have $\mathbf{T} \vdash \neg \varphi$, hence $\mathbf{T} \vdash \neg \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$ so that $\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$ and, by the additional assumption, $\mathbf{T} \vdash \varphi$, again contradicting the consistency of $\mathbf{T}$. □

2.2.3. THEOREM (Second Incompleteness Theorem). *Let* $\mathrm{Con}_{\mathbf{T}}$ *be* $\neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner)$, *where* Λ *is any convenient contradictory statement. Then*

$$\mathbf{T} \nvdash \mathrm{Con}_{\mathbf{T}}.$$

PROOF. Let φ be as in the statement of Theorem 2.2.2. We show: $\mathbf{S} \vdash \varphi \leftrightarrow \mathrm{Con}_{\mathbf{T}}$.

Observe that $\mathbf{S} \vdash \varphi \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$ implies $\mathbf{S} \vdash \varphi \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner)$, since $\mathbf{T} \vdash \Lambda \rightarrow \varphi$ implies $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \rightarrow \varphi \urcorner)$, by D1, which implies $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$, by D3.

But $\varphi \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner)$ is just $\varphi \rightarrow \mathrm{Con}_{\mathbf{T}}$ and we have proven half of the equivalence.

Conversely, by D2, $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \urcorner)$, which implies $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \urcorner)$, by D1, D3, since $\varphi \leftrightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$. This yields $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \wedge \neg \varphi \urcorner)$, by D1, D3, and logic, which implies $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner)$, by D1, D3, and logic. By contraposition, $\mathbf{S} \vdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner) \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$, which is $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \varphi$, by definitions. □

2.2.4. COROLLARY. $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \mathrm{Con}_{\mathbf{T}+\neg \mathrm{Con}_{\mathbf{T}}}$.

PROOF. By the proof of Theorem 2.2.3,

(i) $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$,

(ii) $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \leftrightarrow \varphi$.

Using D2, D3, it follows that $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Con}_{\mathbf{T}} \urcorner)$, so that

$$\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \mathrm{Con}_{\mathbf{T}} \rightarrow \Lambda \urcorner),$$

which gives $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \mathrm{Con}_{\mathbf{T}+\neg \mathrm{Con}_{\mathbf{T}}}$. □

Corollary 2.2.4 is the Formalized Second Incompleteness Theorem.

Let us finish this exposition of the proofs with two remarks:

2.2.5. REMARK. By the proof of the Second Theorem, the self-referential sentence which asserts its own unprovability is equivalent to the sentence asserting consistency. Hence, this sentence is unique up to provable equivalence and one may correctly speak of *the* sentence that asserts its own unprovability.

2.2.6. REMARK. If the reader compares the loose statement of the First Incompleteness Theorem given earlier with that of Theorem 2.2.2(ii), he will notice that we dropped the reference to ω-consistency. We will discuss this concept in Section 4.2.

2.3. Things to come

Except for the discussion of the mechanics of the encoding, we have finished proving the Incompleteness Theorems. This seems to be a good place to insert a brief description of the sequel.

In Section 3, we discuss the encoding and some related topics. Section 4 concerns metamathematical properties other than consistency and presents some generalizations of the Incompleteness Theorems. In Section 5, we present two applications of the notions and results of Sections 2 and 4. The Incompleteness Theorems are obtained by formalizing Syntax — in Section 6, we discuss what happens when one formalizes Semantics.

3. Encoding

The details of an encoding are fascinating to work out and boring to read. The author wrote the present section for his own benefit and his feelings will not be hurt if the reader chooses to skip it.

In expositions, one often replaces precise statements by imprecise ones, or by precise but false ones. As an example of the latter, it is commonly asserted that one proves the Second Incompleteness Theorem by formalizing the proof of the First. A more correct statement would be that one formalizes D1 by D2 and then reduces the Second Theorem to the First. In Section 2.1, we have been guilty of cheating in two places:

(i) in our vague formulation of the sense in which **T** contains **S**, and

(ii) in our remark on sub.

We discuss (i) now and (ii) in 3.2.2.

In Section 2.1 we blithely remarked that **T** contains **S** in the sense that there is the same sort of embedding of **S** into **T** as there is of arithmetic into

set theory. In ordinary mathematical practice, the details of an embedding can be important: Is it continuous? A homomorphism? The same holds here. We must know, e.g., what we mean by $\mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)\urcorner)$, where φ is in the language of **T** and $\mathrm{Pr}_{\mathbf{T}}$ in that of **S**.

We propose to sweep all of the difficulties under the rug by strengthening the assumption to:

(i) the language of **S** is contained in that of **T**;

(ii) the axioms of **S** are among those of **T**.

While (i) and (ii) will make life easy for us in general, they will really grease the wheels when we discuss D2.

The usual cases considered do not satisfy these conditions; but, if one defines **T'** to be the conservative extension of **T** by the addition of the symbols and axioms of **S**, one can usually show

$$(*) \qquad \mathbf{S} \vdash \forall x\, [\mathrm{Pr}_{\mathbf{T}}(x) \leftrightarrow \mathrm{Pr}_{\mathbf{T}'}(x)],$$

thus reducing the case in question to the one treated here. Theories in which (*) does not hold are pathological and we are not interested in them.

Now that this is settled, we make some additional inessential assumptions. Their only use is to reduce the number of cases that need to be considered when we define various functions representing syntactic operations (cf. 3.2). They are:

(i) The only logical connectives and quantifiers are $\neg$, $\rightarrow$, $\forall$.

(ii) **S** and **T** contain as constants only the numerals: $\bar{0}, \bar{1}, \ldots$.

(iii) Only numerical variables occur.

(iv) **T** contains infinitely many n-ary function and relation symbols for each n.

Thus, the language of **T** consists of:

numerals: $\bar{0}, \bar{1}, \ldots$,

numerical variables: $v_0, v_1, \ldots$,

n-ary function symbols: $f_0^n, f_1^n, \ldots$,

n-ary relation symbols: $R_0^n, R_1^n, \ldots$,

connectives: $\neg$, $\rightarrow$.

quantifier: $\forall$.

The other connectives and quantifier are considered to be abbreviations.

We assume that **S** has a pairing function $\langle\ ,\ \rangle$ with inverses π_1, π_2. Using them, we assign codes, which are closed terms, to the basic syntactic objects as follows:

$$\bar{i} \mapsto \langle \bar{0}, \bar{i} \rangle \qquad \neg \mapsto \langle \bar{4}, \bar{4} \rangle$$

$$v_i \mapsto \langle \bar{1}, \bar{i} \rangle \qquad \rightarrow \mapsto \langle \bar{5}, \bar{5} \rangle$$

$$f_i^n \mapsto \langle \bar{2}, \langle \bar{n}, \bar{i} \rangle \rangle \qquad \forall \mapsto \langle \bar{6}, \bar{6} \rangle$$

$$R_i^n \mapsto \langle \bar{3}, \langle \bar{n}, \bar{i} \rangle \rangle$$

Terms and formulae are finite sequences of these symbols and derivations are finite sequences of formulae. Thus, **S** will have to be able to encode and manipulate finite sequences. In the following subsection, we introduce a nice class of functions and discuss their use for such encoding. In 3.2, we assume these functions are "in" **S** and finish encoding syntax. 3.3, 3.4, and 3.5 discuss some generalizations of the First Incompleteness Theorem that one can prove once one has an awareness of the encoding opportunities available.

3.1. Primitive recursive encoding of finite sequences

Loosely put, the primitive recursive functions are those functions of natural numbers that are obtained by recursion. Of course, to be obtained by recursion, they must be obtained *from* something and, to avoid minor unpleasantries, they must also be closed under explicit definition:

3.1.1. DEFINITION. A function f on natural numbers is *primitive recursive* if it can be generated after finitely many steps by means of the following rules:

i	$f(x) = 0,$	Zero
ii	$f(x) = x + 1,$	Successor
iii	$f(\boldsymbol{x}) = x_i,$	Projection
iv	$f(\boldsymbol{x}) = g(h_1(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x})),$	Composition
v	$\begin{cases} f(0, \boldsymbol{x}) = g(\boldsymbol{x}), \\ f(x+1, \boldsymbol{x}) = h(f(x, \boldsymbol{x}), x, \boldsymbol{x}). \end{cases}$	Recursion

3.1.2. DEFINITION. A relation $R \subseteq N^n$ is primitive recursive if its *representing function*,

$$\chi_R(\boldsymbol{x}) = \begin{cases} 0 & \text{if } R(\boldsymbol{x}), \\ 1 & \text{if } \neg R(\boldsymbol{x}) \end{cases}$$

is primitive recursive.

To facilitate the discussion of the encoding of finite sequences, we

establish a few simple closure properties of the classes of primitive recursive functions and relations. To do this, we need a few functions. Trivially, starting with the Zero function and iterating composition of the Successor function, we see that all constant functions are primitive recursive. Elementary school mathematics tells us that iterating Recursion shows that addition, multiplication, and exponentiation are primitive recursive Subtraction takes us out of the domain of natural numbers; however, *cut-off subtraction*,

$$x \dot{-} y = \begin{cases} x - y & \text{if } x \geq y, \\ 0 & \text{if } x < y, \end{cases}$$

is primitive recursive. For we can define it by Recursion by

$$x \dot{-} 0 = x, \qquad x \dot{-} (y + 1) = \mathrm{pd}(x \dot{-} y),$$

where pd is defined by Recursion by

$$\mathrm{pd}(0) = 0, \qquad \mathrm{pd}(x + 1) = x.$$

Two more handy functions are the sign function, sg, and its complement, $\overline{\mathrm{sg}}$:

$$\mathrm{sg}(0) = 0, \qquad \mathrm{sg}(x + 1) = 1;$$

$$\overline{\mathrm{sg}}(0) = 1, \qquad \overline{\mathrm{sg}}(x + 1) = 0.$$

3.1.3. LEMMA (definition by cases). *Let g_1, g_2, h be primitive recursive and define f by*

$$f(\boldsymbol{x}) = \begin{cases} g_1(\boldsymbol{x}) & \text{if } h(\boldsymbol{x}) = 0, \\ g_2(\boldsymbol{x}) & \text{if } h(\boldsymbol{x}) \neq 0. \end{cases}$$

Then f is primitive recursive.

PROOF. $f(\boldsymbol{x}) = g_1(\boldsymbol{x}) \cdot \overline{\mathrm{sg}}(h(\boldsymbol{x})) + g_2(\boldsymbol{x}) \cdot \mathrm{sg}(h(\boldsymbol{x}))$. □

3.1.4. COROLLARY. *The relation of equality is primitive recursive.*

PROOF. Let $h(x, y) = |x - y| = (x \dot{-} y) + (y \dot{-} x)$. □

Note that sg and $\overline{\mathrm{sg}}$ were used in somewhat of a logical manner in the above proof. To further illustrate this, let χ_R, χ_S be the representing functions of relations R, S. Observe:

$$\chi_{\neg R}(\boldsymbol{x}) = \overline{\mathrm{sg}}(\chi_R(\boldsymbol{x})),$$

$$\chi_{R \wedge S}(\boldsymbol{x}) = \mathrm{sg}(\chi_R(\boldsymbol{x}) + \chi_S(\boldsymbol{x})),$$

$$\chi_{R \vee S}(\boldsymbol{x}) = \chi_R(\boldsymbol{x}) \cdot \chi_S(\boldsymbol{x}).$$

If we define bounded quantifiers, $\exists y \leq x$, $\forall y \leq x$, then for $R(y, \boldsymbol{x})$:

$$\chi_{\exists y \leq x R}(x, \boldsymbol{x}) = \prod_{y \leq x} \chi_R(y, \boldsymbol{x}),$$

$$\chi_{\forall y \leq x R} = \chi_{\neg \exists y \leq x \neg R}.$$

3.1.5. LEMMA. (i) *Let $g(x, \boldsymbol{x})$ be primitive recursive. Then f_1 and f_2 are primitive recursive, where*

$$f_1(x, \boldsymbol{x}) = \sum_{y \leq x} g(y, \boldsymbol{x}), \qquad f_2(x, \boldsymbol{x}) = \prod_{y \leq x} g(y, \boldsymbol{x}).$$

(ii) *If $R(y, \boldsymbol{x})$ is a primitive recursion relation, then the relations S, T are also primitive recursive, where*

$$S(x, \boldsymbol{x}) \leftrightarrow \exists y \leq x\, R(y, \boldsymbol{x}), \qquad T(x, \boldsymbol{x}) \leftrightarrow \forall y \leq x\, R(y, \boldsymbol{x}).$$

The proof of this lemma is left to the reader.

3.1.6. DEFINITION (bounded μ-operator). Let $g(y, \boldsymbol{x})$ be a function. We define

$$f(x, \boldsymbol{x}) = \mu y < x\, [g(y, \boldsymbol{x}) = 0]$$

by $f(x, \boldsymbol{x}) =$ the least $y < x$ such that $g(y, \boldsymbol{x}) = 0$, if such a y exists, and $f(x, \boldsymbol{x}) = x$, otherwise.

3.1.7. LEMMA. *If $g(y, \boldsymbol{x})$ is primitive recursive, then so is $\mu y < x\, [g(y, \boldsymbol{x}) = 0]$.*

PROOF. Define

$$f_1(x, \boldsymbol{x}) = \begin{cases} 0 & \text{if } \exists y \leq x\, [g(y, \boldsymbol{x}) = 0], \\ 1 & \text{if } \neg \exists y \leq x\, [g(y, \boldsymbol{x}) = 0]. \end{cases}$$

f_1 is primitive recursive and we may define

$$f(x, \boldsymbol{x}) = \sum_{y < x} f_1(y, \boldsymbol{x}). \quad \square$$

We have enough tools at hand to show the primitive recursiveness of the well-known pairing function,

$$\langle x, y\rangle = \tfrac{1}{2}((x+y)^2 + 3x + y),$$

and its inverses. We simply use the bounded μ-operator:

$$\langle x, y\rangle = \mu z < (x+y)^2 + 3x + y + 1\,[2z = (x+y)^2 + 3x + y],$$

$$\pi_1 z = \mu x < z + 1\,[\exists y \leq z\,(\langle x, y\rangle = z)],$$

$$\pi_2 z = \mu y < z + 1\,[\langle \pi_1 z, y\rangle = z],$$

where we use the fact that $x, y \leq \langle x, y\rangle$.

To encode finite sequences, we use the Fundamental Theorem of Arithmetic, whereby every natural number ≥ 2 has a unique representation:

$$a = p_{i_0}^{n_0} \cdots p_{i_k}^{n_k},$$

where $p_{i_0}, \ldots, p_{i_k}$ are distinct primes and all n_i are positive. We have the following definitions:

i $$x \mid y \leftrightarrow \exists z \leq y\,(xz = y).$$

ii $$x < y \leftrightarrow \exists z \leq y\,(y = x + z + 1).$$

iii $$x \text{ is prime} \leftrightarrow x \neq 0 \wedge x \neq 1 \wedge \forall z \leq x\,[z \mid x \rightarrow z = x \vee z = 1].$$

iv $$p_n = n\text{-th prime:}\quad p_0 = 2,$$

$$p_{n+1} = \mu x < p_n! + 1\,[p_n < x \wedge x \text{ is prime}].$$

v $$a \in \mathrm{Seq} \leftrightarrow a = 1 \vee a > 1 \wedge \forall x \leq a\,[p_{x+1} \mid a \rightarrow p_x \mid a].$$

vi $$\mathrm{lh}(a) = \begin{cases} 0 & \text{if } a \notin \mathrm{Seq} \vee a = 1, \\ \mu x \leq a\,[p_x \mid a \wedge p_{x+1} \nmid a] & \text{if } a \in \mathrm{Seq} \wedge a \neq 1. \end{cases}$$

vii $$(a)_x = \mu y \leq x + 1\,[p_x^{y+1} \mid a \wedge p_x^{y+2} \nmid a].$$

viii $$a * b = \begin{cases} a \cdot \prod_{x \leq \mathrm{lh}(b)} p_{\mathrm{lh}(a)+x+1}^{(b)_x+1} & \text{if } a \neq 1 \wedge b \neq 1, \\ a & \text{if } b = 1 \wedge a \neq 1, \\ b & \text{if } a = 1. \end{cases}$$

We should comment on v–viii. Seq denotes the set of sequence numbers, i.e. those numbers with no gaps in their list of prime divisors. For such numbers, we have

$$a = \prod_{i \le \mathrm{lh}(a)} p_i^{(a)_i + 1}.$$

If a, b are sequence numbers encoding $(a_0, \ldots, a_m)$, $(b_0, \ldots, b_n)$, respectively, then $a * b$ is a sequence number encoding the concatenation $(a_0, \ldots, a_m, b_0, \ldots, b_n)$.

We write $(a_0, \ldots, a_n)$ for $2^{a_0+1} \cdots p_n^{a_n+1}$. In particular, $(a) = 2^{a+1}$ and $(\ ) = 1$.

An immediate application of these notions is the following.

3.1.8. LEMMA (course-of-values recursion). *Let g, h be primitive recursive and define f by*

$$f(0, \boldsymbol{x}) = g(\boldsymbol{x}), \qquad f(x+1, \boldsymbol{x}) = h(\bar{f}(x, \boldsymbol{x}), x, \boldsymbol{x}),$$

where $\bar{f}(x, \boldsymbol{x}) = (f(0, \boldsymbol{x}), \ldots, f(x, \boldsymbol{x}))$ is the course-of-values function associated with f. Then f is primitive recursive.

PROOF. Observe that $\bar{f}$ is primitive recursive:

$$f(0, \boldsymbol{x}) = 2^{g(\boldsymbol{x})}, \qquad \bar{f}(x+1, \boldsymbol{x}) = \bar{f}(x, \boldsymbol{x}) * (h(\bar{f}(x, \boldsymbol{x}), x, \boldsymbol{x})).$$

But $f(x, \boldsymbol{x}) = (\bar{f}(x, \boldsymbol{x}))_x$. □

We will use this lemma in the next subsection to finish our discussion of encoding.

3.2. Primitive recursive encoding of syntax

So far we have codes for basic syntactic objects (variables, numerals, etc.) and a primitive recursive technique of encoding finite sequences We now combine what we have to encode more complicated syntactic objects.

3.2.1. DEFINITION. We generate codes for complex terms and formulae as follows:

(i) If $t_1, \ldots, t_n$ have codes $\ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner$, then

$$\ulcorner f_i^n t_1 \cdots t_n \urcorner = (\ulcorner f_i^n \urcorner, \ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner),$$

$$\ulcorner R_i^n t_1 \cdots t_n \urcorner = (\ulcorner R_i^n \urcorner, \ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner),$$

where $\ulcorner f_i^n \urcorner$, $\ulcorner R_i^n \urcorner$ are the codes assigned in the introduction.

(ii) If φ, ψ, have codes $\ulcorner \varphi \urcorner$, $\ulcorner \psi \urcorner$, respectively, then

$$\ulcorner \neg \varphi \urcorner = (\ulcorner \neg \urcorner, \ulcorner \varphi \urcorner). \qquad \ulcorner \varphi \to \psi \urcorner = (\ulcorner \to \urcorner, \ulcorner \varphi \urcorner, \ulcorner \psi \urcorner).$$

(iii) If φ has code $\ulcorner\varphi\urcorner$ and v_i is a variable, then

$$\ulcorner\forall v_i\,\varphi\urcorner = (\ulcorner\forall\urcorner, \ulcorner v_i\urcorner, \ulcorner\varphi\urcorner).$$

[*Note.* This gives us the functions neg, imp: $\mathrm{neg}(x) = (\ulcorner\neg\urcorner, x)$; $\mathrm{imp}(x, y) = (\ulcorner\rightarrow\urcorner, x, y)$.]

We now show that the complex syntactic notions are primitive recursive:

(i) The representing function for terms is defined by course-of-values recursion:

$$T(x) = \begin{cases} 0 & \text{if } \exists i \leq x\,[x = \langle 0, i\rangle \vee x = \langle 1, i\rangle], \\ 0 & \text{if } x \in \mathrm{Seq} \wedge \exists n i \leq x\,[(x)_0 = \langle 2, \langle n, i\rangle\rangle \wedge \mathrm{lh}(x) = n \wedge \\ & \qquad \wedge \forall y < n\,(T((x)_{y+1}) = 0], \\ 1 & \text{otherwise.} \end{cases}$$

(ii) Similarly, one defines the representing function for formulae:

$$F(x) = \begin{cases} 0 & \text{if } x \in \mathrm{Seq} \wedge \exists n i \leq x\,[(x)_0 = \langle 3, \langle n, i\rangle\rangle \wedge \mathrm{lh}(x) = n \wedge \\ & \qquad \wedge \forall y < n\,(T((x)_{y+1}) = 0], \\ 0 & \text{if } x \in \mathrm{Seq} \wedge (x)_0 = \ulcorner\neg\urcorner \wedge F((x)_1) = 0 \wedge \mathrm{lh}(x) = 1, \\ 0 & \text{if } x \in \mathrm{Seq} \wedge (x)_0 = \ulcorner\rightarrow\urcorner \wedge F((x)_1) = F((x)_2) = 0 \wedge \mathrm{lh}(x) = 2, \\ 0 & \text{if } x \in \mathrm{Seq} \wedge (x)_0 = \ulcorner\forall\urcorner \wedge \exists i \leq x\,((x)_1 = \langle 1, i\rangle) \\ & \qquad \wedge F((x)_2) = 0 \wedge \mathrm{lh}(x) = 2, \\ 1 & \text{otherwise.} \end{cases}$$

3.2.2. Sub

In Section 2.1 we spoke of *a* substitution function. As it is, we need two: one to replace a (code for a) free variable by a (code for a) term and one to replace a (code for a) free variable by a (code for a) numeral which supposedly designates the same number designated by a given term. The former is needed, e.g., to recognize axioms such as $\forall x\,\varphi x \rightarrow \varphi t$. Both could be used for the Diagonalization Lemma; but the latter is needed if one wants a free-variable form of the Diagonalization Lemma. Other uses are hard to describe here and we leave the function to "speak for itself" when we apply it later. We define the first syntactic substitution function by course-of-values recursion, first treating terms and then formulae:

$$\mathrm{sub}(\ulcorner v_i \urcorner, i, y) = y,$$

$$\mathrm{sub}(\ulcorner v_j \urcorner, i, y) = \ulcorner v_j \urcorner, \qquad j \neq i,$$

$$\mathrm{sub}(\ulcorner f_j^n t_1 \cdots t_n \urcorner, i, y) = (\ulcorner f_j^n \urcorner, \mathrm{sub}(\ulcorner t_1 \urcorner, i, y), \ldots, \mathrm{sub}(\ulcorner t_n \urcorner, i, y)),$$

$$\mathrm{sub}(\ulcorner R_j^n t_1 \cdots t_n \urcorner, i, y) = (\ulcorner R_j^n \urcorner, \mathrm{sub}(\ulcorner t_1 \urcorner, i, y), \ldots, \mathrm{sub}(\ulcorner t_n \urcorner, i, y)),$$

$$\mathrm{sub}(\ulcorner \neg \varphi \urcorner, i, y) = (\ulcorner \neg \urcorner, \mathrm{sub}(\ulcorner \varphi \urcorner, i, y)),$$

$$\mathrm{sub}(\ulcorner \varphi \to \psi \urcorner, i, y) = (\ulcorner \to \urcorner, \mathrm{sub}(\ulcorner \varphi \urcorner, i, y), \mathrm{sub}(\ulcorner \psi \urcorner, i, y)),$$

$$\mathrm{sub}(\ulcorner \forall v_i \varphi \urcorner, i, y) = \ulcorner \forall v_i \varphi \urcorner,$$

$$\mathrm{sub}(\ulcorner \forall v_j \varphi \urcorner, i, y) = (\ulcorner \forall \urcorner, \ulcorner v_j \urcorner, \mathrm{sub}(\ulcorner \varphi \urcorner, i, y)),$$

$$\mathrm{sub}(x, i, y) = 0, \quad x \text{ not of the above forms.}$$

With this definition, one easily sees that, for any term t,

$$\mathrm{sub}(\ulcorner \varphi v_i \urcorner, i, \ulcorner t \urcorner) = \ulcorner \varphi t \urcorner.$$

If we observe that v_i occurs freely in φ iff $\mathrm{sub}(\ulcorner \varphi \urcorner, i, \langle 0, i \rangle) \neq \ulcorner \varphi \urcorner$, then we can primitive recursively define the function sub referred to in Section 2.1 by

$$\mathrm{sub}(x, y) = \begin{cases} \mathrm{sub}(x, i, y) & \text{if } i = \mu j < x \,[v_j \text{ occurs free in } x], \\ x & \text{if } i \text{ does not exist.} \end{cases}$$

sub_3, sub_4, etc. are defined by iteration.

A second important substitution function is

$$s(x, y) = \mathrm{sub}(x, \langle 0, y \rangle).$$

This satisfies: If φv_i has only v_i free, t is a closed term denoting n, then $s(\ulcorner \varphi v_i \urcorner, t) = \ulcorner \psi \urcorner$, where ψ is equivalent to $\varphi \bar{n}$. (To prove $\psi \leftrightarrow \varphi \bar{n}$, one need only show $\vdash t = \bar{n}$. For then, substitutivity of equality yields $\vdash \ulcorner \varphi \bar{n} \urcorner = \ulcorner \psi \urcorner$.) Moreover, if φ has only v_i free, $s(\ulcorner \varphi \urcorner, x)$ is formally the code of a sentence. We often abbreviate $s(\ulcorner \varphi x \urcorner, y)$ by $\ulcorner \varphi \dot{y} \urcorner$.

3.2.3. $\mathrm{Prov}_{\mathbf{T}}$

The next step is to define $\mathrm{Prov}_{\mathbf{T}}(x, y)$. A derivation of φ is a sequence of formulae such that each element of the sequence is either an axiom of **T**, a logical axiom, or a consequence by some logical rule of earlier members of the sequence. The logical axioms can be taken to be primitive recursive in the sense that the set of codes of such axioms is primitive recursive. We leave it to the reader to check this for his favorite axiomatization. Further,

by clever choice of axioms, we can assume that the only rule of inference is modus ponens:

$$\text{MP:}\quad \frac{\varphi \quad \varphi \to \psi}{\psi}.$$

We *assume* that the set of (codes of) axioms of **T** is primitive recursive. Thus, we get:

$$\begin{aligned}\mathrm{Prov}_{\mathbf{T}}(x, y) \leftrightarrow x \in \mathrm{Seq} \wedge \forall i \leq \mathrm{lh}(x)\,[(x)_i \text{ is a logical axiom} \vee \\ \vee\ (x)_i \text{ is an axiom of } T \vee \\ \vee\ \exists jk < i\,((x)_k = \mathrm{imp}((x)_j, (x)_i))] \wedge \\ \wedge\ y = (x)_{\mathrm{lh}(x)}.\end{aligned}$$

$$\mathrm{Pr}_{\mathbf{T}}(y) \leftrightarrow \exists x\ \mathrm{Prov}_{\mathbf{T}}(x, y).$$

3.2.4. S, I: Numeralwise representability

A relation $\mathbf{R} \subseteq N^n$ is said to be *numeralwise represented* or *numerated* by a formula φ in **S** if one has, for all $m_1 \cdots m_n$,

$$Rm_1 \cdots m_n \text{ is true} \quad \text{iff} \quad \mathbf{S} \vdash \varphi \bar{m}_1 \cdots \bar{m}_n.$$

R is *binumerated* by φ if one also has

$$Rm_1 \cdots m_n \text{ is false} \quad \text{iff} \quad \mathbf{S} \vdash \neg\, \varphi \bar{m}_1 \cdots \bar{m}_n.$$

We assume that **S** binumerates every primitive recursive relation. To do this, it suffices to find a formula φ_f, for every primitive recursion function f, such that all numerical instances of the defining equations for f are provable. E.g. if f is defined by Recursion from g, h, then for all $m_1, \ldots, m_n, m$,

$$\mathbf{S} \vdash \exists! x\, \varphi_f(\bar{0}, \bar{m}_1, \ldots, \bar{m}_n, x) \wedge \forall x\, [\varphi_f(\bar{0}, \bar{m}_1, \ldots, \bar{m}_n, x) \to \varphi_g(\bar{m}_1, \ldots, \bar{m}_n, x)],$$

$$\begin{aligned}\mathbf{S} \vdash \exists! x\, \varphi_f(\bar{m} + \bar{1}, \bar{m}_1, \ldots, \bar{m}_n, x) \\ \wedge \forall x, y\, [\varphi_f(\bar{m}, \bar{m}_1, \ldots, \bar{m}_n, y) \wedge \\ \wedge\ \varphi_f(\bar{m} + \bar{1}, \bar{m}_1, \ldots, \bar{m}_n, x) \to \varphi_h(y, \bar{m}, \bar{m}_1, \ldots, \bar{m}_n, x)].\end{aligned}$$

Then, a metamathematical induction on the number of steps it takes to generate f (and on the first argument of f in the case of definition by recursion) shows that φ_f binumerates the graph of f.

Given that **S** binumerates all primitive recursive functions (and hence all primitive recursive relations), it follows that all of the primitive recursive

encoding functions are binumerated: neg, imp, sub, s, and $\mathrm{Prov}_{\mathbf{T}}$. This last fact allows us to verify D1:

$$\mathbf{T} \vdash \varphi \Leftrightarrow \mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(t, \ulcorner \varphi \urcorner) \quad \text{for some closed term } t,$$

$$\Rightarrow \mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner).$$

A quick rereading of the proof of the First Incompleteness Theorem will show that we did not need D2 and D3 to establish it. Thus we have given (modulo only a little handwaving) a complete proof of this Theorem.

3.2.5. S, II: D2 and D3

The Second Incompleteness Theorem does not come so cheaply. For D3, one must show

$$\mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(a, \ulcorner \varphi \urcorner) \wedge \mathrm{Prov}_{\mathbf{T}}(b, \ulcorner \varphi \to \psi \urcorner) \to \mathrm{Prov}_{\mathbf{T}}(a * b * (\ulcorner \psi \urcorner), \ulcorner \psi \urcorner),$$

with *free* variables a, b. To do this requires much more than mere binumerability of primitive recursive relations: The representations must be correct with free variables. For this, **S** must be able to prove not merely each instance of the defining equations for a given primitive recursive function, but must also prove the equations with free variables. It will also be necessary for **S** to prove induction on primitive recursive relations.

The latter necessity is clear if one considers D2. D2 is a formal version of the following sharpening of D1: If f is primitive recursive,

$$(*) \qquad fm_1 \cdots m_n = m \Rightarrow \mathbf{T} \vdash f\bar{m}_1 \cdots \bar{m}_n = \bar{m}$$

$$\Rightarrow \mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(d(\bar{m}_1, \ldots, \bar{m}_n), \ulcorner f\bar{m}_1 \cdots \bar{m}_n = \bar{m} \urcorner),$$

for some primitive recursive function d. To see that such a d exists, observe that a proof that $f\boldsymbol{m} = m$ (in **S**, and hence in **T**) is almost just a computation — it will be given by a sequence of equations and implications of equations. Thus, the existence of d satisfying (*) is not too surprising; nor is the fact that we claim:

$$(**) \qquad \mathbf{S} \vdash f\boldsymbol{x} = y \to \mathrm{Prov}_{\mathbf{T}}(d\boldsymbol{x}, \ulcorner f\dot{x} = \dot{y} \urcorner).$$

The proof, which is too long to be included here, proceeds by metamathematical induction on the number of steps needed to generate f and (when the recursion clause is used) formal induction on the primitive recursive relation (**). Once one has accepted (**), the primitive recursiveness of $\mathrm{Prov}_{\mathbf{T}}$ yields

$$\mathbf{S} \vdash \mathrm{Prov}_{\mathbf{T}}(x, y) \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Prov}_{\mathbf{T}}(\dot{x}, \dot{y}) \urcorner),$$

and one needs only apply D1, D3 to the valid formula $\varphi x \to \exists x \varphi x$, to conclude the validity of D2.

3.2.6. S, III: Choice of S

By the preceding discussion, one can adequately encode syntax in **S** if **S** admits a representation of primitive recursive functions in such a way that

(i) the defining equations for primitive recursive functions are provable with free variables;

(ii) induction on primitive recursive relations is provable;

(iii) computations are almost derivations of the equations they establish.

We list three examples of such theories.

(a) **PRA** = Primitive Recursive Arithmetic. **PRA** contains the numerals $\bar{0}, \bar{1}, \ldots$ and there is a function symbol in **PRA** for each (definition via the rules for the generation of primitive recursive functions of a) primitive recursive function. In addition to some trivial axioms concerning the constants and the successor function, the axioms of **PRA** are the defining equations of the functions and induction on quantifier-free formulae.

(b) **PA** = Peano's Arithmetic. **PA** also has the numerals as its constants, but it only has function symbols for successor, addition, and multiplication. The axioms consist of trivial axioms concerning the constants and the successor function, the recursion equations for addition and multiplication, and induction on all formulae of the language. Even proving that **PA** binumerates the primitive recursive functions requires another encoding trick. The most famous technique uses the Chinese Remainder Theorem to encode finite sequences. Conditions (i) and (ii) are proven by formalizing the use of the encoding of finite sequences and is non-trivial insofar as very few texts give the details. Condition (iii) can be bypassed by observing that the representation of a PR function can be written in the form $\exists x \varphi$, where φ is much simpler syntactically than the corresponding primitive recursive function. E.g., by Matiyasevich's Theorem, φ can be taken to be an equation involving two polynomials. Thus, the formalization of $\varphi x \to \mathrm{Pr}_{\mathbf{PA}}(\ulcorner \varphi \dot{x} \urcorner)$ is much simpler.

(c) **ZF** = Zermelo–Fraenkel set theory. This is both a good and a bad example. It is bad because the whole encoding problem is more easily solved in a set theory than in an arithmetic theory. By the same token, it is a good example.

3.3. Rosser's Theorem

By Section 3.2, binumerability of primitive recursive relations in **S**

suffices for the First Incompleteness Theorem — as a condition on **S**. There is still the necessity of assuming something about **T** —

(i) that **T** contain **S**,

(ii) that **T** be consistent, and

(iii) (for the second half of the theorem) that theorems of **T** of the form $\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$ be true.

Rosser's Theorem allows one to drop the last soundness condition on **T** by using a modification of $\mathrm{Prov}_{\mathbf{T}}$: Define

$$\mathrm{Prov}^R_{\mathbf{T}}(x, y) \leftrightarrow \mathrm{Prov}_{\mathbf{T}}(x, y) \wedge$$
$$\wedge \forall z w \leq x\, [\mathrm{Prov}_{\mathbf{T}}(z, w) \rightarrow y \neq \mathrm{neg}(w) \wedge w \neq \mathrm{neg}(y)],$$
$$\mathrm{Pr}^R_{\mathbf{T}}(y) \leftrightarrow \exists x\, \mathrm{Prov}^R_{\mathbf{T}}(x, y),$$
$$\mathrm{Con}^R_{\mathbf{T}} \leftrightarrow \neg\, \mathrm{Pr}^R_{\mathbf{T}}(\ulcorner\Lambda\urcorner).$$

3.3.1. ROSSER'S THEOREM. *Let* $\mathbf{T} \vdash \varphi \leftrightarrow \neg\, \mathrm{Pr}^R_{\mathbf{T}}(\ulcorner\varphi\urcorner)$. *Then*

(i) $\mathbf{T} \nvdash \varphi$;

(ii) $\mathbf{T} \nvdash \neg\varphi$;

(iii) $\mathbf{T} \vdash \mathrm{Con}^R_{\mathbf{T}}$.

PROOF. (i) By the consistency of **T**, $\mathrm{Prov}_{\mathbf{T}}$ and $\mathrm{Prov}^R_{\mathbf{T}}$ binumerate the same relation. Hence $\mathrm{D1}^R$ holds: $\mathbf{T} \vdash \psi \Rightarrow \vdash \mathrm{Pr}^R_{\mathbf{T}}(\ulcorner\psi\urcorner)$. Thus, the proof of the first part of the First Incompleteness Theorem yields the result.

(ii) This follows from (iii).

(iii) We leave this to the reader along with the remark that **T** is consistent and $\mathbf{T} \vdash \neg\Lambda$. □

*3.4. Recursion theory

(The reader is referred to Chapter C.1 for a full discussion of recursion theory.)

Historically, recursion theory developed out of the incompleteness theorems. Once one knows a little recursion theory, however, it is natural to look back.

3.4.1. DEFINITION. A set $\mathbf{S} \subseteq \mathbb{N}$ of natural numbers is *recursively enumerable* (r.e.) iff for some primitive recursive relation R,

$$\mathbf{S}x \Leftrightarrow \exists y\, Rxy.$$

An equivalent definition is:

3.4.2. DEFINITION. A set $\mathbf{S} \subseteq \mathbb{N}$ is r.e. iff $\mathbf{S} = \emptyset$ or, for some primitive recursive function f, $\mathbf{S} = \mathrm{ran}(f)$.

Another useful concept is given by the following definition.

3.4.3. DEFINITION. A set $\mathbf{S} \subseteq \mathbb{N}$ is *recursive* iff $\mathbf{S}$ and $\mathbb{N} - \mathbf{S}$ are both r.e. A function $f : \mathbb{N} \to \mathbb{N}$ is recursive iff its graph (viewed as a subset of $\mathbb{N}$ by means of a primitive recursive pairing function) is recursive.

The recursion-theoretic counterpart of the First Incompleteness Theorem is:

3.4.4. THEOREM. *There is an r.e. non-recursive set.*

One proves this by finding an enumeration, $W_0, W_1, \ldots$, of r.e. sets and an r.e. set K_1 such that

$$\forall xy\, (\langle x, y\rangle \in K_1 \Leftrightarrow x \in W_y).$$

Then $K = \{x : \langle x, x\rangle \in K_1\}$ is r.e. with a non-r.e. complement. One then proves the First Incompleteness Theorem by showing that K can be numeralwise represented in $\mathbf{T}$. If $\mathbf{T}$ is sound enough, this is not too difficult — one uses the numeralwise representation of K in $\mathbf{S}$ that arises from the binumeration of the primitive recursive relation that K is the projection of. If $\mathbf{T}$ is not very sound, the numeralwise representation of K in $\mathbf{S}$ may not be one in $\mathbf{T}$ as $\mathbf{T}$ may simply prove more numbers to be in K. One usually avoids difficulties with numeralwise representations in unsound theories by means of the following:

3.4.5. DEFINITION. Let $A, B \subseteq \mathbb{N}$ be disjoint r.e. sets. A and B are *effectively inseparable* iff there is a recursive function f such that for all r.e. sets W_i, W_j, if

(i) $W_i \cap W_j = \emptyset$,

(ii) $A \subseteq W_i$, $B \subseteq W_j$,

then $f(i, j) \notin W_i \cup W_j$.

3.4.6. THEOREM. *Effectively inseparable r.e. sets exist.*

We shall accept this on faith.

To prove that part of Rosser's Theorem that corresponds to the First Incompleteness Theorem, one constructs φ, ψ which numeralwise represent A, B, respectively, in $\mathbf{S}$ and for which

$$(*) \qquad \mathbf{S} \vdash \forall x \neg(\varphi x \wedge \psi x).$$

Then, if **T** is a consistent formal theory, the set of (codes of) its theorems can be shown to be r.e. and one defines

$$W_i = \{n\colon \mathbf{T} \vdash \varphi \bar{n}\}, \qquad W_j = \{n\colon \mathbf{T} \vdash \neg \varphi \bar{n}\}.$$

By (*), $W_i \cap W_j = \emptyset$. Since **T** contains **S** and φ numerates A in **S**, it follows that $A \subseteq W_i$, $B \subseteq W_j$ and $n_0 = f(i,j) \notin W_i \cup W_j$. Thus $\mathbf{T} \nvdash \varphi \bar{n}_0, \neg \varphi \bar{n}_0$.

*3.5. The formula hierarchy

The purpose of the present subsection is mainly to establish some notation for several more advanced sections below.

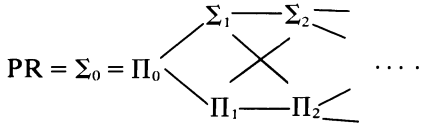
Recall that **S** is assumed to have, for each primitive recursive function f, a formula φ_f representing it in the strong sense of 3.2.5. φ_f is called a *primitive recursive formula*, or a PR formula.

3.5.1. DEFINITION. A formula φ is Σ_n (Π_n) iff for some PR formula ψ,

$$\varphi = Q_1 x_1 \cdots Q_n x_n \psi,$$

where $Q_1 = \exists$ ($\forall$) and the quantifiers alternate in type. We write $\varphi \in \Sigma_n$ (Π_n), ambiguously, as φ is Σ_n (Π_n) or provably equivalent (in **S**) to a Σ_n (Π_n) formula.

Thus, one has the inclusions:



3.5.2. THEOREM. *There is a Σ_n truth definition for Σ_n formulae. I.e., for each n, there is a formula $\mathrm{Tr}_{\Sigma_n} \in \Sigma_n$ with only the numerical variable x free such that, for $\varphi_x \in \Sigma_n$,*

$$\mathbf{S} \vdash \varphi x \leftrightarrow \mathrm{Tr}_{\Sigma_n}(\ulcorner \varphi \dot{x} \urcorner).$$

A similar result holds for Π_n.

We omit the proof and note the following.

3.5.3. COROLLARY. *The formula $\mathrm{Tr}_{\Sigma_n}(s(x,x))$ is Σ_n, non-Π_n.*

The proof is left as an easy exercise to the reader.

It follows that all the inclusions indicated above are proper. Thus we have a genuine hierarchy.

From our point of view, there are two uses of this Hierarchy: First, since it is a hierarchy, we can use it to measure the complexity of formulae or of sets of formulae. Such use is made in Section 4. A second use is not of the Hierarchy itself taken as a hierarchy, but rather of Theorem 3.5.2: Many set-theoretic proofs could be carried out in arithmetic *if* one had a truth definition. Tarski's Theorem asserts that there is no truth definition for the entire language [Exercise]. By Theorem 3.5.2, there are partial truth definitions $\mathrm{Tr}_0, \mathrm{Tr}_1, \ldots$ such that Tr_n works up to the n-th level of the Hierarchy. This allows the formalization of certain outwardly set-theoretic constructions within arithmetic. An application is discussed in Section 6.

Before proceeding, it is worth noting the following.

3.5.4. FACT (demonstrable Σ_1 completeness). *If* $\varphi \in \Sigma_1$, *then*

$$\mathbf{S} \vdash \varphi \boldsymbol{x} \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{\boldsymbol{x}} \urcorner).$$

This follows from the discussion of 3.2.5.

4. Metamathematical properties other than consistency

Metamathematically, consistency is a minimal assumption on a theory. One might wish for stronger properties to hold — e.g. ω-consistency. If **T** is a theory about a particular structure, as PA is a theory about the semiring of natural numbers, one might wish for even more — *soundness* (anything provable is true) or *completeness* (anything or its negation is provable — hence anything true is provable).

In this section, we discuss these properties. In 4.1 we consider a soundness scheme — the *Reflection Principle*. ω-consistency is discussed in 4.2 and its relation to the more intuitive Reflection Principle is presented. Completeness, which we know to be false, nonetheless gives rise to consistent schemata. These are discussed in 4.3.

4.1. Reflection principles

The First Incompleteness Theorem is proven by considering the sentence that asserts its own *un*provability. Under minimal assumptions, it is clear that the sentence must be true — and hence unprovable. But what about the sentence that asserts its own *provability*? Is it true? false? It was precisely this problem that led to the following important theorem characterizing provable instances of the Reflection Principle:

4.1.1. LÖB'S THEOREM. *Let φ be closed. Then*

$$\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi \quad \textit{iff} \quad \mathbf{T} \vdash \varphi.$$

PROOF. The one direction is obvious. For the other, assume that $\mathbf{T} \nvdash \varphi$. Then $\mathbf{T} + \neg\varphi$ is consistent and we may appeal to the Second Incompleteness Theorem to conclude that $\mathbf{T} + \neg\varphi$ does not yield $\mathrm{Con}_{\mathbf{T}+\neg\varphi}$, hence not $\neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg\varphi \to \Lambda \urcorner)$. Thus

$$\mathbf{T} + \neg\varphi \nvdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner).$$

Contraposition yields $\mathbf{T} \nvdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi$. □

As hinted above, this solves the problem of sentences asserting their own provability — such sentences are provable (and hence equivalent — cf. also 5.1). This also focuses our attention on the following schemata:

Local Reflection Principle

Rfn(**T**): $\mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi$, φ closed.

First Uniform Reflection Principle

RFN(**T**): $\forall x\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner) \to \forall x\, \varphi x$, φ has only x free.

Second Uniform Reflection Principle

RFN′(**T**): $\forall x\, [\mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner) \to \varphi x]$, φ has only x free.

[A stipulation must be inserted here: As indicated by the notation $\ulcorner \varphi \dot{x} \urcorner$, the variable x must range over elements which can be named by constants. Thus, we insist that the x in the uniform versions of the Reflection Principle be a numerical variable. In fact, throughout the following, we shall assume that all variables explictly exhibited are numerical variables, although non-numerical variables may occur unexhibited in the formulae.]

The reflection principles are clearly schematic assertions of soundness — anything provable is true. As such, they immediately imply consistency and, thus, we see that they are underivable in **T**. Of course, Theorem 4.1.1 tells us more than this: It characterizes the provable instances of Rfn(**T**). Nonetheless, it may be instructive to restate the First and Second Incompleteness Theorems in terms of the reflection principles:

4.1.2. FIRST INCOMPLETENESS THEOREM. *For some true, unprovable φ,*

$$\mathbf{T} \nvdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi.$$

4.1.3. SECOND INCOMPLETENESS THEOREM. *For any refutable* φ,

$$\mathbf{T} \nvdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi.$$

Let us check that these statements are equivalent to the more familiar versions. The First Incompleteness Theorem is no problem if we specify that the true unprovable sentence we have in mind is the one asserting its own unprovability. Then Theorem 4.1.1 simply yields

$$\mathbf{T} \nvdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi \quad \text{iff} \quad \mathbf{T} \nvdash \varphi,$$

where the two equivalents are the two versions of the First Incompleteness Theorem. For the Second Incompleteness Theorem, observe that, for refutable φ, the following are equivalent over $\mathbf{T}$:

$$\mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi \quad \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \quad \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner) \quad \mathrm{Con}_{\mathbf{T}}.$$

Theorem 4.1.1 again yields the equivalence of the two versions.

Thus, Löb's Theorem is a generalization of the incompleteness theorems. While this alone would justify taking a closer look at the reflection principles, it might be worth our while to mention another motivating factor. Recall that the main impetus behind Hilbert's Consistency Program was the fact that consistency was equivalent to soundness for real statements:

4.1.4. THEOREM. *Over* $\mathbf{S}$, *the following are equivalent*:

(i) $\mathrm{Con}_{\mathbf{T}}$;
(ii) $\mathrm{Rfn}_{\Pi_1}(\mathbf{T})$;
(iii) $\mathrm{RFN}_{\Pi_1}(\mathbf{T})$;
(iv) $\mathrm{RFN}'_{\Pi_1}(\mathbf{T})$;

where the subscript "Π_1" *indicates restriction of the schemata to* $\varphi \in \Pi_1$.

PROOF. The implications (iv)$\to$(iii)$\to$(ii) are fairly direct. (ii)$\to$(i) follows from the above observation that $\mathrm{Con}_{\mathbf{T}} \leftrightarrow (\mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi)$ for any refutable φ — one merely chooses such a $\varphi \in \Pi_1$.

(i)$\to$(iv). Let $\varphi \in \Pi_1$ have only x free. Then $\neg \varphi x \in \Sigma_1$ and, by demonstrable Σ_1-completeness (see 3.5),

$$(*) \qquad \mathbf{S} \vdash \neg \varphi x \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner).$$

But

$$(**) \qquad \mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner) \to \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner),$$

whence (*) and (**) combine to give

$$\mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner) \to \varphi x. \quad \square$$

The interested reader is referred to 5.2 for some applications of Theorem 4.1.4. For the moment, we simply use it as our second reason to justify our interest in reflection principles: *Consistency is equivalent to a restricted Reflection Principle.*

Having decided that reflection principles are worth studying, we may as well begin. First, let us observe that the schemata are listed in full generality. For one thing, we must restrict ourselves to schemata since the sentence

$$\forall x\,[\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\urcorner)\rightarrow \mathrm{Tr}(\ulcorner\varphi\dot{x}\urcorner)],$$

where $\mathrm{Tr}(\ulcorner\psi\urcorner)$ asserts "ψ is true", cannot be asserted in **T**. For, by Tarski's Theorem on Truth Definitions, there can be no truth definition for **T** within **T** itself (cf. 3.5). Further, extra variables in either version of the uniform scheme can be contracted by means of a pairing function, reducing the general scheme to the two listed. A hybrid, e.g. $\forall x\,[\forall y\,\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\dot{y}\urcorner)\rightarrow \forall y\,\varphi xy]$, is clearly implied by the several variable Second Uniform Reflection Principle. Finally, we have the following theorem.

4.1.5. THEOREM (Feferman). RFN(**T**) *and* RFN′(**T**) *are equivalent over* **S**.

PROOF. Obviously, the instance, $\forall x\,\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\urcorner)\rightarrow\forall x\,\varphi x$, of RFN(**T**) is implied by the corresponding instance of RFN′(**T**). REN′(**T**). The converse requires a minor (but often useful) lemma:

4.1.6. LEMMA. $\mathbf{S}\vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner\mathrm{Prov}_{\mathbf{T}}(\dot{y},\ulcorner\varphi\dot{x}\urcorner)\rightarrow\varphi\dot{x}\urcorner)$.

PROOF. (a) By D1 and D3,

$$\mathbf{S}\vdash \mathrm{Prov}_{\mathbf{T}}(y,\ulcorner\varphi\dot{x}\urcorner)\rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\urcorner)$$
$$\rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\mathrm{Prov}_{\mathbf{T}}(\dot{y},\ulcorner\varphi\dot{x}\urcorner)\rightarrow\varphi\dot{x}\urcorner).$$

(b) Since $\neg\,\mathrm{Prov}_{\mathbf{T}}\in\mathrm{PR}$, we similarly have

$$\mathbf{S}\vdash\neg\,\mathrm{Prov}_{\mathbf{T}}(y,\ulcorner\varphi\dot{x}\urcorner)\rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\neg\,\mathrm{Prov}_{\mathbf{T}}(\dot{y},\ulcorner\varphi\dot{x}\urcorner)\urcorner)$$
$$\rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\mathrm{Prov}_{\mathbf{T}}(\dot{y},\ulcorner\varphi\dot{x}\urcorner)\rightarrow\varphi\dot{x}\urcorner).$$

Combining (a) and (b) yields the lemma. □

To complete the proof of Theorem 4.1.5, let φ be given and observe

$$\mathbf{S}\vdash\forall x\,[\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\urcorner)\rightarrow\varphi x]\leftrightarrow\forall xy\,[\mathrm{Prov}_{\mathbf{T}}(y,\ulcorner\varphi\dot{x}\urcorner)\rightarrow\varphi x].$$

But the right-hand side of this equivalence is derivable in $\mathbf{S}+\mathrm{RFN}(\mathbf{T})$ by Lemma 4.1.6. □

Before proceeding further, it is amusing to note the following provable instance of reflection:

$$(*) \qquad \mathbf{S} \vdash \exists y\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Prov}_{\mathbf{T}}(\dot{y}, \ulcorner \varphi \urcorner) \urcorner) \rightarrow \exists y\, \mathrm{Prov}_{\mathbf{T}}(y, \ulcorner \varphi \urcorner),$$

i.e. $\mathbf{S} \vdash \exists y\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Prov}_{\mathbf{T}}(\dot{y}, \ulcorner \varphi \urcorner) \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$. (*) follows immediately from Lemma 4.1.6 and condition D3 and we leave its derivation to the reader. By Theorem 4.1.4 and the Second Incompleteness Theorem, the variable y on the right side of (*) cannot in general denote the same code for a derivation as the y on the left. We can also see this by appeal to the following free-variable form of Löb's Theorem:

4.1.7. THEOREM. *Let φ have only x free. Then*

$$\mathbf{T} \vdash \forall x\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner) \rightarrow \varphi x] \quad \textit{iff} \quad \mathbf{T} \vdash \forall x\, \varphi x.$$

We omit the proof.

So far, we have shown that the use of reflection principles allows a generalization of the incompleteness theorems, that $\mathrm{Con}_{\mathbf{T}}$ is equivalent to a restriction of the Reflection schemata, and that $\mathrm{RFN}(\mathbf{T})$ is as general as $\mathrm{RFN}'(\mathbf{T})$ and further schemata with additional variables. We ought to ask ourselves the simple question: How much of an improvement is Reflection over Consistency? Obviously, consistency does not imply soundness — e.g. $\mathbf{T} = \mathbf{PA} + \neg \mathrm{Con}_{\mathbf{PA}}$ is consistent but not sound as $\neg \mathrm{Con}_{\mathbf{PA}}$ is a false theorem of $\mathbf{T}$. (For this same $\mathbf{T}$, however, $\mathbf{T} + \mathrm{Con}_{\mathbf{T}} \vdash \mathrm{RFN}(\mathbf{T})$ — for $\mathbf{T} \vdash \neg \mathrm{Con}_{\mathbf{T}}$.) A first step is given by the following simple lemma:

4.1.8. LEMMA. *Let φ be closed. Then*

(i) $\mathbf{T} + \varphi + \mathrm{Rfn}(\mathbf{T}) \vdash \mathrm{Rfn}(\mathbf{T} + \varphi)$,

(ii) $\mathbf{T} + \varphi + \mathrm{RFN}(\mathbf{T}) \vdash \mathrm{RFN}(\mathbf{T} + \varphi)$.

PROOF. Observe that for any ψ, we have the following over $\mathbf{S}$:

$$\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}+\varphi}(\ulcorner \psi \urcorner) \leftrightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \rightarrow \psi \urcorner). \quad \square$$

4.1.9. COROLLARY. *Let φ be closed.*

(i) *Let $\mathbf{T} + \varphi$ be consistent. Then $\mathbf{T} + \varphi \not\vdash \mathrm{Rfn}(\mathbf{T})$.*

(ii) *Let $\mathbf{T}'$ be a consistent finite extension of $\mathbf{T}$. Then $\mathbf{T}' \not\vdash \mathrm{Rfn}(\mathbf{T})$.*

(iii) *If $\mathbf{T} + \mathrm{Con}_{\mathbf{T}}$ is consistent, then $\mathbf{T} + \mathrm{Con}_{\mathbf{T}} \not\vdash \mathrm{Rfn}(\mathbf{T})$.*

4.1. Hierarchy considerations

By Corollary 4.1.9, neither Rfn(**T**) nor RFN(**T**) is implied by any finite set of axioms consistent with **T**. We devote the rest of this subsection to discussing an often useful improvement of this in the uniform case. For this purpose, we must use the notions and notations of the Formula Hierarchy (discussed in 3.5). This material is less detailed and may be omitted on first reading.

Let $\mathrm{RFN}_{\Pi_k}(\mathbf{T})$ denote the restriction of the scheme RFN(**T**) to formulae in Π_k. Similarly, one defines $\mathrm{RFN}_{\Sigma_k}(\mathbf{T})$, $\mathrm{RFN}'_{\Pi_k}(\mathbf{T})$, and $\mathrm{RFN}'_{\Sigma_k}(\mathbf{T})$. A first result is:

4.1.10. THEOREM. *Over* **S**, *the following are equivalent* ($k \geq 0$):

(i) $\mathrm{RFN}_{\Sigma_k}(\mathbf{T})$,

(ii) $\mathrm{RFN}_{\Pi_{k+1}}(\mathbf{T})$,

(i.a) $\mathrm{RFN}'_{\Sigma_k}(\mathbf{T})$,

(ii.a) $\mathrm{RFN}'_{\Pi_{k+1}}(\mathbf{T})$.

The equivalences (x) $\leftrightarrow$ (x.a) follow by taking a closer look at the proof of Theorem 4.1.5. The implications (ii) $\rightarrow$ (i), (ii.a) $\rightarrow$ (i.a) are trivial as $\Sigma_k \subseteq \Pi_{k+1}$. For the converses, one uses provable closure under numerical substitution: $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \forall x\, \varphi x \urcorner) \rightarrow \forall x\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner)$.

In terms of the Hierarchy, Lemma 4.1.8 can be restated:

4.1.11. THEOREM. *Let* $\varphi \in \Pi_k$ *be closed and let* $n \geq k$. *Then*

(i) $\mathbf{S} + \varphi + \mathrm{Rfn}_{\Sigma_n}(\mathbf{T}) \vdash \mathrm{Rfn}_{\Sigma_n}(\mathbf{T} + \varphi)$,

(ii) $\mathbf{S} + \varphi + \mathrm{RFN}_{\Sigma_n}(\mathbf{T}) \vdash \mathrm{RFN}_{\Sigma_n}(\mathbf{T} + \varphi)$.

This is seen by observing that, if $\psi \in \Sigma_n$, then $\varphi \rightarrow \psi \in \Sigma_n$.

Using a Π_k truth definition for Π_k formulae, it can be shown that $\mathrm{RFN}'_{\Pi_k}(\mathbf{T})$ can be written as a single Π_k sentence. Bearing this in mind, we have:

4.1.12. COROLLARY. (i) $\mathbf{T} + \mathrm{RFN}_{\Sigma_{k+1}}(\mathbf{T}) \vdash \mathrm{RFN}_{\Sigma_{k+1}}(\mathbf{T} + \mathrm{RFN}_{\Pi_{k+1}}(\mathbf{T}))$, $k \geq 0$.

(ii) $\mathbf{T} + \mathrm{RFN}_{\Pi_{k+1}}(\mathbf{T}) \vdash \mathrm{RFN}_{\Pi_{k+1}}(\mathbf{T} + \mathrm{RFN}_{\Pi_k}(\mathbf{T}))$, $k \geq 1$.

(iii) *If* $\mathbf{T} + \mathrm{RFN}_{\Pi_k}(\mathbf{T})$ *is consistent, then*

$$\mathbf{T} + \mathrm{RFN}_{\Pi_k}(\mathbf{T}) \not\vdash \mathrm{Rfn}_{\Sigma_k}(\mathbf{T}), \quad k \geq 1.$$

(iii′) *If* $\mathbf{T} + \mathrm{Con}_{\mathbf{T}}$ *is consistent, then*

$$\mathbf{T} + \mathrm{Con}_{\mathbf{T}} \not\vdash \mathrm{Rfn}_{\Sigma_1}(\mathbf{T}).$$

PROOF. Parts (i) and (ii) follow from Theorems 4.1.10 and 4.1.11. Parts (iii) and (iii′) are simple applications of the incompleteness theorems in the forms of Theorems 4.1.2 and 4.1.3. □

By Corollary 4.1.12, RFN is not implied by any (consistent) bounded set of its instances. The following theorem of Kreisel and Levy improves this immensely:

4.1.13. ESSENTIAL UNBOUNDEDNESS THEOREM. *Let n be given. Let* **U** *be an r.e. theory (not necessarily containing* **S**) *in the language of* **T**. *If* $\mathbf{T} \vdash \mathrm{RFN}(\mathbf{U})$, *then no consistent extension of* **U** *by a set* Δ *of* Σ_n *sentences implies all theorems of* **T**. *In particular,* **T** *cannot be axiomatized over* **U** *by any set of* Σ_n *axioms.*

PROOF. Let Tr_n be a Σ_n truth definition for Σ_n formulae. Define ψ by

$$S \vdash \psi \leftrightarrow \forall x\, [\mathrm{Tr}_n(x) \rightarrow \neg\, \mathrm{Pr}_{\mathbf{U}}(\mathrm{imp}(x, \ulcorner \psi \urcorner))].$$

Intuitively, ψ asserts its unprovability from any true Σ_n sentence.

(a) Since $\mathbf{T} \vdash \mathrm{RFN}(\mathbf{U})$, it follows that

$$\mathbf{T} \vdash \forall x\, [\mathrm{Pr}_{\mathbf{U}}(\mathrm{imp}(x, \ulcorner \psi \urcorner)) \rightarrow (\mathrm{Tr}_n(x) \rightarrow \psi)].$$

Thus

$$\mathbf{T} \vdash \neg\, \psi \rightarrow \forall x\, [\neg\, \mathrm{Tr}_n(x) \vee \neg\, \mathrm{Pr}_{\mathbf{U}}(\mathrm{imp}(x, \ulcorner \psi \urcorner))]$$

$$\rightarrow \forall x\, [\mathrm{Tr}_n(x) \rightarrow \neg\, \mathrm{Pr}_{\mathbf{U}}(\mathrm{imp}(x, \ulcorner \psi \urcorner))] \rightarrow \psi.$$

Thus $\mathbf{T} \vdash \psi$.

(b) Suppose $\mathbf{U} + \Delta$ extends T; $\Delta \subseteq \Sigma_n$. Then $\mathbf{U} + \Delta \vdash \psi$. We now show that this implies the inconsistency of $\mathbf{U} + \Delta$: Since $\mathbf{U} + \Delta \vdash \psi$, it follows that $\mathbf{U} + X \vdash \psi$ for some finite $X \subseteq \Delta$. Let $\varphi = \wedge X$. Then $\mathbf{U} \vdash \varphi \rightarrow \psi$. But this implies

(i) $$\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{U}}(\ulcorner \varphi \rightarrow \psi \urcorner).$$

Since $\mathbf{T} \vdash \psi$,

(ii) $$\mathbf{T} \vdash \mathrm{Pr}_{\mathbf{U}}(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow \neg\, \mathrm{Tr}_n(\ulcorner \varphi \urcorner).$$

But $\varphi \in \Sigma_n$ and

(iii) $$\mathbf{T} \vdash \varphi \leftrightarrow \mathrm{Tr}_n(\ulcorner \varphi \urcorner),$$

and (i)–(iii) yield $\mathbf{T} \vdash \neg\, \varphi$. □

The Essential Unboundedness Theorem is a useful tool for proving unboundedness theorems — results of the form: Axiomatizing $\mathbf{T}_1$ over $\mathbf{T}_2$ requires arbitrarily complicated formulae. The most basic example is that in which $\mathbf{T}_2 = \mathrm{PC}$:

4.1.14. DEFINITION. $\mathbf{T}$ is said to be *reflexive* if $\mathbf{T} \vdash \mathrm{RFN}(\mathrm{PC})$, where PC denotes the predicate calculus (as formulated in the language of $\mathbf{T}$).

We remark that, by Lemma 4.1.8, if $\mathbf{T}$ is reflexive, $\mathbf{T} \vdash \mathrm{RFN}(\mathbf{U})$ for all finite subsystems $\mathbf{U}$ of $\mathbf{T}$. In particular, $\mathbf{T} \vdash \mathrm{Con}_{\mathbf{U}}$ for such $\mathbf{U}$.

4.1.15. REFLEXIVENESS THEOREM. **PA** *and* **ZF** *are reflexive.*

The usual proof of this theorem is too long to be given here. For ZF, we can give the following simple proof: By formalized induction on the length of a derivation,

$$\mathbf{ZF} \vdash \forall x\, [\mathrm{Pr}_{\mathbf{PC}}(\ulcorner \varphi \dot{x} \urcorner) \rightarrow \forall a\, (\mathrm{Trans}(a) \wedge x \in a \rightarrow \varphi^{(a)} x)],$$

where $\varphi^{(a)}$ denotes the relativization of φ to a and $\mathrm{Trans}(a)$ asserts that a is transitive. By the set-theoretic reflection principle,

$$\mathbf{ZF} \vdash \forall x\, [\neg \varphi x \rightarrow \exists a\, [\mathrm{Trans}(a) \wedge x \in a \wedge \neg \varphi^{(a)} x]],$$

whence

$$\mathbf{ZF} \vdash \forall x\, [\mathrm{Pr}_{\mathrm{PC}}(\ulcorner \varphi \dot{x} \urcorner) \rightarrow \varphi x].$$

As corollaries, we see that (i) the induction scheme of **PA** is not implied by any bounded set of its instances, and (ii) one cannot bound all the schemata of **ZF**.

*4.2. ω-consistency

The concept of ω-consistency was introduced by Gödel for the purpose of stating the hypotheses needed for the First Incompleteness Theorem. The ω-consistency of $\mathbf{T}$ is neither the optimal nor the most intuitive condition sufficient for the theorem. Nonetheless, its use here is so firmly entrenched in the literature that we are obligated to comment on it.

Informally, ω-consistency is the property that holds of $\mathbf{T}$ if the following two conditions are *not* simultaneously satisfied for any φ:

(i) $\mathbf{T} \vdash \exists x\, \varphi x$;

(ii) $\mathbf{T} \vdash \neg \varphi \bar{0}, \neg \varphi \bar{1}, \ldots$.

Formally, ω-consistency can be represented (in varying degrees of generality) by (modifications of) the following scheme:

4.2.1.

$$\mathrm{Pr}_{\mathbf{T}}(\ulcorner \exists x\, \varphi x \urcorner) \to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner).$$

Let 1-$\mathrm{Con}_{\mathbf{T}}$ denote the restriction of 4.2.1 to $\varphi \in \mathrm{PR}$ possessing only one free variable.

4.2.2. FORMALIZED FIRST INCOMPLETENESS THEOREM. *Let* φ *be* $\neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$. *Then*:

(i) $\mathbf{T} + \mathrm{Con}_{\mathbf{T}} \vdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$,

(ii) $\mathbf{T} + 1\text{-}\mathrm{Con}_{\mathbf{T}} \vdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \urcorner)$.

PROOF. Part (i) was shown in the course of the proof of the Second Incompleteness Theorem.

For part (ii), let $\varphi = \forall x\, \psi x$, $\psi \in \mathrm{PR}$. Then

$$\mathbf{T} + 1\text{-}\mathrm{Con}_{\mathbf{T}} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \urcorner) \to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \dot{x} \urcorner) \to \exists x \neg \psi x,$$

by demonstrable Σ_1-completeness (3.5.4). Thus,

$$(*) \qquad \mathbf{T} + 1\text{-}\mathrm{Con}_{\mathbf{T}} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \urcorner) \to \neg \varphi$$

$$(**) \qquad \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner).$$

But 1-$\mathrm{Con}_{\mathbf{T}}$ yields $\mathrm{Con}_{\mathbf{T}}$ since it asserts the unprovability of something. Thus, by (i),

$$\mathbf{T} + 1\text{-}\mathrm{Con}_{\mathbf{T}} \vdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner),$$

which, with (**), yields (ii). □

Probably the most important thing to notice about the above proof is that 1-$\mathrm{Con}_{\mathbf{T}}$ was used only to derive (*): $\mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \urcorner) \to \neg \varphi$, for closed $\varphi \in \Pi_1$ — i.e., 1-$\mathrm{Con}_{\mathbf{T}}$ was used only to derive $\mathrm{Rfn}_{\Sigma_1}(\mathbf{T})$. Conversely, $\mathrm{Rfn}_{\Sigma_1}(\mathbf{T})$ can be used to derive 1-$\mathrm{Con}_{\mathbf{T}}$:

$$\mathbf{S} + \mathrm{Rfn}_{\Sigma_1}(\mathbf{T}) \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \exists x\, \varphi x \urcorner) \to \exists x\, \varphi_x$$

$$\to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner),$$

since $\neg \varphi x \leftrightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner)$ by demonstrable PR-completeness and the fact that $\mathrm{Rfn}_{\Sigma_1}(\mathbf{T})$ implies $\mathrm{Con}_{\mathbf{T}}$.

By the preceding paragraph, $\mathrm{Rfn}_{\Sigma_1}(\mathbf{T})$ and 1-$\mathrm{Con}_{\mathbf{T}}$ are equivalent. Since the statement 4.2.1 of ω-consistency is easily seen to be Σ_2, Corollary 4.1.12 shows that we cannot expect such behavior to hold for more than some very special cases.

4.2.3. DEFINITION. We define the following formal schemata representing ω-consistency:

Local ω-consistency

ω-$\mathrm{Con}_{\mathbf{T}}$: $\quad \mathrm{Pr}_{\mathbf{T}}(\ulcorner \exists x\, \varphi x \urcorner) \to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner)$, $\quad \varphi$ has only x free,

Uniform ω-consistency

ω-$\mathrm{CON}_{\mathbf{T}}$: $\forall y\, [\mathrm{Pr}_{\mathbf{T}}(\ulcorner \exists x\, \varphi x \dot{y} \urcorner) \to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \dot{y} \urcorner)]$, $\quad \varphi$ has only x, y free,

Global ω-consistency

ω-$\mathrm{CON}^{G}_{\mathbf{T}}$: $\quad \forall \varphi\, [\mathrm{Pr}_{\mathbf{T}}(\ulcorner \exists x\, \varphi x \urcorner) \to \exists x \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg \varphi \dot{x} \urcorner)]$,

where $\forall \varphi$ indicates quantification over codes of formulae possessing only one free variable.

We hasten to emphasize the fact that, unlike the case with reflection principles, we have here a *global* representation of the given concept as well as the local and uniform ones. The reason is simply that we only use φ in 4.2.1 in the form of a *code* and not, as with reflection, as a subformula of some larger formula. Thus, we can quantify over all φ in the present context.

It is not hard to see that ω-$\mathrm{CON}^{G}_{\mathbf{T}} \vdash \omega$-$\mathrm{CON}_{\mathbf{T}} \vdash \omega$-$\mathrm{Con}_{\mathbf{T}}$ over $\mathbf{S}$. Local schemata are usually difficult to deal with and so we ignore ω-$\mathrm{Con}_{\mathbf{T}}$. Thus, we are interested in ω-$\mathrm{CON}^{G}_{\mathbf{T}}$ and ω-$\mathrm{CON}_{\mathbf{T}}$ — and in their hierarchical restrictions:

4.2.4. DEFINITION. Let $k \geq 1$. The restriction of ω-$\mathrm{CON}_{\mathbf{T}}$ to formulae $\varphi \in \Sigma_{k-1}$ is termed k-$\mathrm{CON}_{\mathbf{T}}$ and the corresponding informal concept is called *k-consistency*.

Observe that, via a Σ_k truth definition for Σ_k formulae, the corresponding restriction of ω-$\mathrm{CON}^{G}_{\mathbf{T}}$ ($\forall \varphi \mapsto \forall \varphi \in \Sigma_k$) is equivalent to $(k+1)$-$\mathrm{CON}_{\mathbf{T}}$. Further, as in Theorem 4.1.10, k-$\mathrm{CON}_{\mathbf{T}}$ is equivalent to the corresponding restriction for $\varphi \in \Pi_k$.

The following theorem characterizes these notions in terms of the more intuitive reflection principles:

4.2.5. THEOREM. *Over* $\mathbf{S}$, *we have*

i $\quad k$-$\mathrm{CON}_{\mathbf{T}} \leftrightarrow \mathrm{RFN}_{\Pi_{k+1}}(\mathbf{T})$, $\qquad (k = 1,2)$

ii $\quad k$-$\mathrm{CON}_{\mathbf{T}} \leftrightarrow \mathrm{RFN}_{\Pi_3}(\mathbf{T} + \mathrm{RFN}_{\Pi_k}(\mathbf{T}))$, $\qquad (k \geq 2)$

iii $\omega\text{-CON}_{\mathbf{T}}^{G} \leftrightarrow \mathrm{RFN}_{\Pi_3}(\mathbf{T} + \mathrm{RFN}(\mathbf{T}))$,

iii' $\omega\text{-CON}_{\mathbf{T}}^{G} \leftrightarrow \mathrm{RFN}_{\Pi_3}(\mathbf{T} + \omega\text{-CON}_{\mathbf{T}})$.

The proof is rather long and we omit it. Some related results are covered in Chapter D.2.

[As an aside, we would like to mention the following: The formula hierarchy can, as an obvious use, be applied to obtain quantitative refinements of various results. See e.g. 4.1[a]. Theorem 4.2.5 gives an application of a different order: The explication of ω-consistency as a Reflection Principle (iii) presupposes an understanding of the expression "Π_3". I.e. one must know something about the formula hierarchy even to state the relation between ω-consistency and soundness.]

4.3. Completeness properties

Somewhat loosely, the First Incompleteness Theorem asserts that consistent strong formal theories are incomplete. Nonetheless, there are consistent schemata asserting completeness. We (need) consider only the local versions:

Syntactic completeness

$\mathrm{SynComp}_{\mathbf{T}}$: $\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner) \vee \mathrm{Pr}_{\mathbf{T}}(\ulcorner\neg\varphi\urcorner)$, closed φ.

Semantic completeness

$\mathrm{SemComp}_{\mathbf{T}}$: $\varphi \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$, closed φ.

ω-completeness

$\omega\text{-Comp}_{\mathbf{T}}$: $\forall x\, \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\dot{x}\urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\forall x\, \varphi x\urcorner)$, φ has only x free.

Without further ado, we state:

4.3.1. THEOREM. *The following are equivalent over* **S**:

(i) $\neg\mathrm{Con}_{\mathbf{T}}$;
(ii) $\mathrm{SynComp}_{\mathbf{T}}$;
(iii) $\mathrm{SemComp}_{\mathbf{T}}$;
(iv) $\omega\text{-Comp}_{\mathbf{T}}$.

PROOF. Obviously (i)$\rightarrow$(ii), (iii), (iv).

(ii)$\rightarrow$(i) We appeal to a Formalized Rosser's Theorem: If φ is $\neg\mathrm{Pr}_{\mathbf{T}}^{R}(\ulcorner\varphi\urcorner)$, then

$$\mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash \neg\mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner),\ \neg\mathrm{Pr}_{\mathbf{T}}(\ulcorner\neg\varphi\urcorner),$$

whence

$$\mathbf{S} + \mathrm{SynComp}_{\mathbf{T}} \vdash \neg\, \mathrm{Con}_{\mathbf{T}}.$$

(iii)→(i) Here, one takes $\varphi = \mathrm{Con}_{\mathbf{T}}$ and applies the Second Incompleteness Theorem. We omit the details in favor of the following case:

(iv)→(i) Let $\varphi x = \neg\, \mathrm{Prov}_{\mathbf{T}}(x, \ulcorner \Lambda \urcorner)$. By Lemma 4.1.6,

$$\mathbf{S} \vdash \forall x\ \mathrm{Pr}_{\mathbf{T}}(\ulcorner \neg\, \mathrm{Prov}_{\mathbf{T}}(\dot{x}, \ulcorner \Lambda \urcorner) \urcorner),$$

whence

$$\begin{aligned} \mathbf{S} + \omega\text{-}\mathrm{Comp}_{\mathbf{T}} &\vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \forall x \neg\, \mathrm{Prov}_{\mathbf{T}}(x, \ulcorner \Lambda \urcorner) \urcorner) \\ &\vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner) \to \Lambda \urcorner) \\ &\vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner), \end{aligned}$$

by the Formalized Löb's Theorem: $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi \urcorner) \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$. □

*4.3[a]. Kent's Theorem

By Theorem 4.3.1(iii), the scheme $\varphi \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$, is equivalent to $\neg\, \mathrm{Con}_{\mathbf{T}}$ and, hence, is not in general derivable — not even when restricted to $\varphi \in \Pi_1$ (namely $\varphi = \mathrm{Con}_{\mathbf{T}}$). We also know, from 3.5.4, that the subscheme $\varphi x \to \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \dot{x} \urcorner)$, $\varphi \in \Sigma_1$, is derivable. The following Theorem of Kent shows that the situation is even more complicated yet:

4.3.2. THEOREM. *For any n, there is a sentence φ such that*

(i) $\mathbf{S} \vdash \varphi \to \mathrm{Pr}_{\mathbf{S}}(\ulcorner \varphi \urcorner)$;

(ii) *For no $\psi \in \Sigma_n$ does* $\mathbf{S} \vdash \varphi \leftrightarrow \psi$.

PROOF. First, let χ be such that for *no* $\psi \in \Sigma_n$ consistent with $\mathbf{S}$ do we have

$$(*) \qquad \mathbf{S} + \psi \vdash \chi \quad \text{or} \quad \mathbf{S} + \psi \vdash \neg\, \chi.$$

To construct such a χ, we take a hint from the Essential Unboundedness Theorem and let

$$\chi \leftrightarrow \forall x\ [\mathrm{Tr}_n(x) \to \neg\, \mathrm{Pr}_{\mathbf{S}}^{R}(\mathrm{imp}(x, \ulcorner \chi \urcorner))],$$

where Tr_n is a Σ_n truth definition for Σ_n formulae. Mimicking the proof of Rosser's Theorem, we see that (*) fails for all $\psi \in \Sigma_n$ consistent with $\mathbf{S}$.

Now let φ be $\chi \wedge \mathrm{Pr}_{\mathbf{S}}(\ulcorner \Lambda \urcorner)$. Clearly (i) holds. To see (ii), suppose $\mathbf{S} \vdash \varphi \leftrightarrow \psi$ for $\psi \in \Sigma_n$. Then $\mathbf{S} \vdash \psi \to \varphi$, so $\mathbf{S} \vdash \neg\, \psi$, since otherwise (*) is true. Since $\neg\, \varphi$ is $\neg\, (\chi \wedge \mathrm{Pr}_{\mathbf{S}}(\ulcorner \Lambda \urcorner))$, $\mathbf{S} \vdash \neg\, \varphi$, so $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{S}}(\ulcorner \Lambda \urcorner) \to \neg\, \chi$, contradicting $\neg$(*) since $\mathbf{S} + \mathrm{Pr}_{\mathbf{S}}(\ulcorner \Lambda \urcorner)$ is consistent. □

5. Two applications

5.1. A fix-point theorem

By diagonalization, one easily finds sentences ψ, χ such that

$$\mathbf{S} \vdash \psi \leftrightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \urcorner), \qquad \mathbf{S} \vdash \chi \leftrightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \chi \urcorner).$$

The proof of the Second Incompleteness Theorem and Löb's Theorem, respectively, yielded the interesting facts that ψ and χ were not only unique, but explicitly definable:

$$\mathbf{S} \vdash \psi \leftrightarrow \mathrm{Con}_{\mathbf{T}} \leftrightarrow \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \Lambda \urcorner),$$

$$\mathbf{S} \vdash \chi \leftrightarrow \mathbf{t},$$

where $\mathbf{t}$ = truth. An older proof of Löb's Theorem uses the fix-point,

$$\theta \leftrightarrow (\mathrm{Pr}_{\mathbf{T}}(\ulcorner \theta \urcorner) \rightarrow \varphi),$$

for any given φ. A little algebra soon reveals

$$\theta \leftrightarrow (\mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \varphi),$$

whence θ too is explicitly definable — this time from the remaining variable. These turn out not to be isolated examples, but rather instances of a general result.

To obtain a simple statement of this result, we consider a propositional language with propositional variables $p, q, r, \ldots$; the usual connectives, $\wedge$, $\vee$, $\neg$, $\rightarrow$; propositional constants $\mathbf{t}$, $\mathbf{f}$ for truth and falsity; and a modal operator $\Box$ to stand for provability. This will be an interpreted system rather than a deductive one: Given an assignment, $p \mapsto \varphi_p$, of sentences to propositional variables, we obtain a translation φ_α for each formula α of the propositional language:

$$\varphi_{\alpha \circ \beta} = \varphi_\alpha \circ \varphi_\beta, \qquad \circ = \wedge, \vee, \rightarrow;$$

$$\varphi_{\neg \alpha} = \neg \varphi_\alpha; \qquad \varphi_{\Box \alpha} = \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi_\alpha \urcorner).$$

If $\alpha(p_1, \ldots, p_n)$ is a formula of the propositional language and we assign ψ_i to p_i, then we write $\alpha(\psi_1, \ldots, \psi_n)$ for $\varphi_{\alpha(p_1, \ldots, p_n)}$, i.e. for the result of substituting each ψ_i for the corresponding p_i and $\mathrm{Pr}_{\mathbf{T}}$ for $\Box$.

5.1.1. THEOREM (De Jongh's Fix-Point Theorem[1]). *Let $\alpha(p, q)$ be such that p occurs only inside the scope of $\Box$. Then, for some $\beta(q)$ and all sentences $\psi_1, \ldots, \psi_n$ of the language of* $\mathbf{T}$,

[1] A proof of Theorem 5.1.1 has now been published in SAMBIN [0000].

$$\mathbf{S} \vdash \beta(\psi) \leftrightarrow \alpha(\beta(\psi), \psi).$$

Further, $\beta(\psi)$ is, up to provable equivalence, the only fix-point of α.

The proof of this is too complicated to be included here. However, we can prove the following special case:

5.1.2. THEOREM. *Let $\alpha(p, q) = \alpha'(\Box\gamma(p, q), q)$, where in $\alpha'(x, y)$ the variable x does not occur inside the scope of a $\Box$. Then the fix-points of α are determined parametrically by*

$$\beta(q) = \alpha'[\Box\gamma(\alpha'(\mathbf{t}, q), q), q].$$

PROOF. Although we are interested in establishing the result with sentences ψ replacing the variables q, the propositional notation is more convenient. An expression $\vdash \delta(q) \leftrightarrow \delta'(q)$ is understood accordingly.

Since the Diagonalization Lemma holds, we may assume that we have a p such that $\vdash p \leftrightarrow \alpha(p, q)$. It will follow from this that $\vdash p \leftrightarrow \beta(q)$.

5.1.3. LEMMA. *For all r, t, δ, $r \leftrightarrow t, \Box(r \leftrightarrow t) \vdash \delta(r) \leftrightarrow \delta(t)$.*

This follows from the derivability conditions by induction on the length of δ. We omit the proof.

To prove Theorem 5.1.2, we first show:

$$(*) \qquad \vdash \Box\gamma(p, q) \leftrightarrow \Box\gamma(\alpha'(\mathbf{t}, q), q).$$

By the fix-point assumption, $\vdash p \leftrightarrow \alpha'(\Box\gamma(p, q), q)$. Thus, since $\Box\gamma(p, q)$ does not occur inside the scope of $\Box$ in α',

$$(**) \qquad \Box\gamma(p, q) \vdash \Box\gamma(p, q) \leftrightarrow \mathbf{t}$$

$$\vdash \alpha'(\Box\gamma, q) \leftrightarrow \alpha'(\mathbf{t}, q) \vdash p \leftrightarrow \alpha'(\mathbf{t}, q).$$

The derivability conditions yield

$$(***) \qquad \Box\gamma(p, q) \vdash \Box\Box\gamma(p, q) \vdash \Box(p \leftrightarrow \alpha'(\mathbf{t}, q)).$$

Applying the lemma to (**), (***), we immediately have

$$\Box\gamma(p, q) \vdash \Box\gamma(\alpha'(\mathbf{t}, q), q),$$

i.e. half of (*).

For the converse, assume $\Box\gamma(p, q) \wedge \neg\gamma(p, q)$. Then

$$\Box\gamma(p, q) \wedge \neg\gamma(p, q) \vdash \neg\gamma(\alpha'(\mathbf{t}, q), q),$$

by the same reasoning as above. Contraposition yields

$$\gamma(\alpha'(\mathbf{t}, q), q) \vdash \Box\gamma(p, q) \rightarrow \gamma(p, q),$$

whence

$$\Box\gamma(\alpha'(\mathbf{t}, q), q) \rightarrow \Box(\Box\gamma(p, q) \rightarrow \gamma(p, q)) \rightarrow \Box\gamma(p, q),$$

by the Formalized Löb's Theorem,

$$\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner) \rightarrow \varphi\urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner).$$

This completes the proof of (*).

To conclude the proof, observe

$$\beta(q) = \alpha'[\Box\gamma(\alpha'(\mathbf{t}, q), q), q] \leftrightarrow \alpha'(\Box\gamma(p, q), q) \text{ by (*)}$$
$$\leftrightarrow p,$$

the latter by choice of p. □

5.1.4. EXAMPLE. We list: (i) $\alpha(p, q)$, (ii) $\alpha'(p, q)$, (iii) $\gamma(p, q)$, (iv) $\Box\gamma(\alpha'(\mathbf{t}, q), q)$, (v) $\beta(q)$, and (vi) a final simplification of (v):

	(a)	(b)	(c)	(d)
(i)	$\neg\Box p$	$\Box p$	$\Box p \rightarrow q$	$\Box(p \rightarrow q)$
(ii)	$\neg p$	p	$p \rightarrow q$	p
(iii)	p	p	p	$p \rightarrow q$
(iv)	$\Box\neg\mathbf{t}$	$\Box\mathbf{t}$	$\Box(\mathbf{t} \rightarrow q)$	$\Box(\mathbf{t} \rightarrow q)$
(v)	$\neg\Box\neg\mathbf{t}$	$\Box\mathbf{t}$	$\Box(\mathbf{t} \rightarrow q) \rightarrow q$	$\Box(\mathbf{t} \rightarrow q)$
(vi)	$\neg\Box\mathbf{f}$	$\mathbf{t}$	$\Box q \rightarrow q$	$\Box q$

5.2. Conservation results

In this section, we present some conservation results of Kreisel.

Recall that Hilbert's Conservation Program called for a proof that the use of ideal statements and abstract reasoning led to no new real theorems. While the incompleteness theorems showed that this is *in general* impossible, they do not rule out the possibility of success in special cases. In fact, we will even use the Second Incompleteness Theorem in establishing one conservation result

First, we present the main result:

5.2.1. CONSERVATION THEOREM. *Let* $\varphi \in \Pi_1$. *Then* $\mathbf{T} \vdash \varphi \Rightarrow \mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash \varphi$.

PROOF. Let $\varphi \in \Pi_1$ and suppose $\mathbf{T} \vdash \varphi$. D1 yields $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$. But $\mathrm{Con}_{\mathbf{T}}$ is equivalent to $\mathrm{RFN}_{\Pi_1}(\mathbf{T})$ over $\mathbf{S}$ by Theorem 4.1.4, whence $\mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash \varphi$. □

The next two results are corollaries:

5.2.2. THEOREM. *Let* $\varphi \in \Pi_1$. *Then* $\mathbf{T} + \neg \mathrm{Con}_{\mathbf{T}} \vdash \varphi \Rightarrow \mathbf{T} \vdash \varphi$.

PROOF.

$$\mathbf{T} + \neg \mathrm{Con}_{\mathbf{T}} \vdash \varphi \Rightarrow \mathbf{T} + \mathrm{Con}_{\mathbf{T}+\neg \mathrm{Con}_{\mathbf{T}}} \vdash \varphi, \quad \text{by Theorem 5.2.1}$$

$$\Rightarrow \mathbf{T} + \mathrm{Con}_{\mathbf{T}} \vdash \varphi,$$

by the Formalized Second Incompleteness Theorem. But we have

$$\mathbf{T} + \mathrm{Con}_{\mathbf{T}} \vdash \varphi, \qquad \mathbf{T} + \neg \mathrm{Con}_{\mathbf{T}} \vdash \varphi,$$

whence $\mathbf{T} \vdash \varphi$. □

5.2.3. THEOREM. *Let* $\mathbf{T}$, $\mathbf{T}'$ *contain* $\mathbf{S}$ *and let*

$$(*) \qquad \mathbf{T} \vdash \forall x\, [\mathrm{Prov}_{\mathbf{T}'}(x, \ulcorner \varphi \urcorner) \rightarrow \mathrm{Prov}_{\mathbf{T}}(tx, \ulcorner \psi \urcorner)]$$

for some primitive recursive term t. *Then* $\mathbf{S} \vdash \mathrm{Pr}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \urcorner)$.

PROOF. By the Conservation Theorem,

$$\mathbf{S} + \mathrm{Con}_{\mathbf{T}} \vdash (*) \vdash \neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \urcorner) \rightarrow \neg \mathrm{Pr}_{\mathbf{T}'}(\ulcorner \varphi \urcorner),$$

by contraposition. Now absorb $\mathrm{Con}_{\mathbf{T}}$ into $\neg \mathrm{Pr}_{\mathbf{T}}(\ulcorner \psi \urcorner)$ and contrapose again. □

5.2.4. COROLLARY (Relative Consistency Theorem). *Let* $\mathbf{T}$, $\mathbf{T}'$ *contain* $\mathbf{S}$ *and let*

$$\mathbf{S} \vdash \forall x\, [\mathrm{Prov}_{\mathbf{T}'}(x, \ulcorner \Lambda \urcorner) \rightarrow \mathrm{Prov}_{\mathbf{T}}(tx, \ulcorner \Lambda \urcorner)]$$

for some primitive recursive term t. *Then* $\mathbf{S} \vdash \mathrm{Con}_{\mathbf{T}} \rightarrow \mathrm{Con}_{\mathbf{T}'}$.

The corollary is worth commenting on. By the Second Incompleteness Theorem, one cannot prove consistency of strong theories within weak theories. Sometimes, the consistency of a strong theory is genuinely in doubt and one can give a relative consistency result, e.g.

$$(**) \qquad \mathrm{Con}_{\mathbf{ZF}} \rightarrow \mathrm{Con}_{\mathbf{ZF}+\neg \mathrm{CH}}.$$

A throwback to Hilbert's Consistency Program is the demand that the proof of (**) be carried out within as weak a system as possible. Epistemologically, there is no need to give a proof of (**) in, say, **PRA**: For, the value of (**) depends entirely on the acceptance of ZF and one might as

well prove (**) in the latter theory — a technically easier undertaking. However, we can avoid philosophical bickering; for, by Corollary 5.2.4, if this is done at all nicely, it automatically follows that (**) can be proven in the weaker theory — namely **PRA**. Thus, there is nothing to argue about here.

In the last paragraph, we pointed out how one conservation result caused a potential philosophical problem to vanish. Usually, the value of conservation results is that they allow one to use stronger techniques to shorten proofs and conserve one's energy. We refer the reader to Chapter D.3 for quantitative information.

*6. The formalized completeness theorem

There are several possible advanced topics that one could discuss. The close relation between induction principles and reflection principles (often bearing the misnomer, "consistency proofs") is discussed in Chapter D.2. One could discuss efforts to complete a formally incomplete theory by the iterated addition of reflection principles. Another topic concerns proof-theoretic applications of the reflection principles.

We shall discuss the formalized completeness theorem and use it to give model-theoretic proofs of the incompleteness theorems.

In this section, we set $\mathbf{S} = \mathbf{PA}$.

*6.1. The Hilbert–Bernays Completeness Theorem

Formalizing the Henkin completeness proof within **PA** yields:

6.1.1. HILBERT–BERNAYS COMPLETENESS THEOREM. *Let* **U** *have a primitive recursive set of axioms. There is a* Δ_2 *set of formulae,* Tr_M, *such that in* $\mathbf{PA} + \mathrm{Con}_U$ *one can prove that this set defines a model of* **U**:

$$\mathbf{PA} + \mathrm{Con}_U \vdash \forall x\,(\mathrm{Pr}_U(x) \rightarrow \mathrm{Tr}_M(x)).$$

Let us explain this: A formula φ is said to be Δ_n if it can be written both as a Σ_n and a Π_n formula. Theorem 6.1.1 asserts that, modulo Con_U, one can prove in **PA** the existence of a model of **U** whose truth definition is Δ_2.

The meaning of this is best understood by a description of the proof, which is just an arithmetization of the set-theoretic one: One adds to the language of **U** an infinite primitive recursive set of new constants $c_0, c_1, \ldots$, and adds the axiom

$$(*) \qquad \exists x\, \varphi x \to \varphi(c_{\ulcorner \varphi \urcorner})$$

for each formula φ. One then enumerates all sentences $\varphi_0, \varphi_1, \ldots$ of this augmented language and defines a complete theory by starting with **U** and adding, at step n, φ_n or $\neg \varphi_n$ — according to whether φ_n is consistent with what has been chosen before or not. The construction is readily described within **PA**. Assuming $\mathrm{Con}_{\mathbf{U}}$, one can also prove that the construction never terminates. The resulting set of sentences forms a complete theory which, by virtue of the axioms $(*)$, forms a model of **U**. Inspection shows that the truth definition of the model is Δ_2.

*6.2. The incompleteness theorems

Scott was the first to observe that one can give a model-theoretic proof of the First Incompleteness Theorem:

6.2.1. First Incompleteness Theorem. *There is a sentence φ such that* (i) $\mathbf{PA} \not\vdash \varphi$ *and* (ii) $\mathbf{PA} \not\vdash \neg \varphi$.

Proof. Assume **PA** is complete. Then, since **PA** is true, $\mathbf{PA} \vdash \mathrm{Con}_{\mathbf{PA}}$ and we can apply Theorem 6.1.1 to obtain a formula Tr_M which gives a truth definition for a model of **PA**. Choose φ by

$$\mathbf{PA} \vdash \varphi \leftrightarrow \neg \mathrm{Tr}_M(\dot{\ulcorner \varphi \urcorner}).$$

We claim $\mathbf{PA} \not\vdash \varphi$, $\mathbf{PA} \not\vdash \neg \varphi$. For if $\mathbf{PA} \vdash \varphi$, then $\mathbf{PA} \vdash \mathrm{Tr}_M(\ulcorner \varphi \urcorner)$ so $\mathbf{PA} \vdash \neg \varphi$. Similarly, $\mathbf{PA} \vdash \neg \varphi$ implies $\mathbf{PA} \vdash \varphi$. □

We shall discuss this proof a little later. First, we wish to prove the Second Incompleteness Theorem. For this, we need some notation:

6.2.2. Definition. Let $\mathfrak{M}, \mathfrak{N}$ be models of **PA**. If $\mathfrak{M}$ is definable in $\mathfrak{N}$ (even in the weak sense that the atomic relations of $\mathfrak{M}$ are $\mathfrak{N}$-definable), we write $\mathfrak{M} <_{\mathrm{d}} \mathfrak{N}$.

The Hilbert–Bernays Completeness Theorem yields immediately the fact: If $\mathfrak{N} \models \mathbf{PA} + \mathrm{Con}_{\mathbf{PA}}$, then there is an $\mathfrak{M}$ such that $\mathfrak{M} <_{\mathrm{d}} \mathfrak{N}$.

The usefulness of this notion is given by the following.

6.2.3. Lemma. *Let $\mathfrak{M}, \mathfrak{N}$ be models of* **PA**, $\mathfrak{M} <_{\mathrm{d}} \mathfrak{N}$. *Then $\mathfrak{M}$ is definably an end-extension of $\mathfrak{N}$ — i.e. there is a unique $\mathfrak{N}$-definable isomorphic embedding of $\mathfrak{N}$ into $\mathfrak{M}$ as an initial segment of $\mathfrak{M}$.*

PROOF. The proof is straightforward: $0_{\mathfrak{N}}$ obviously maps onto $0_{\mathfrak{M}}$. Extend the map, say F, by

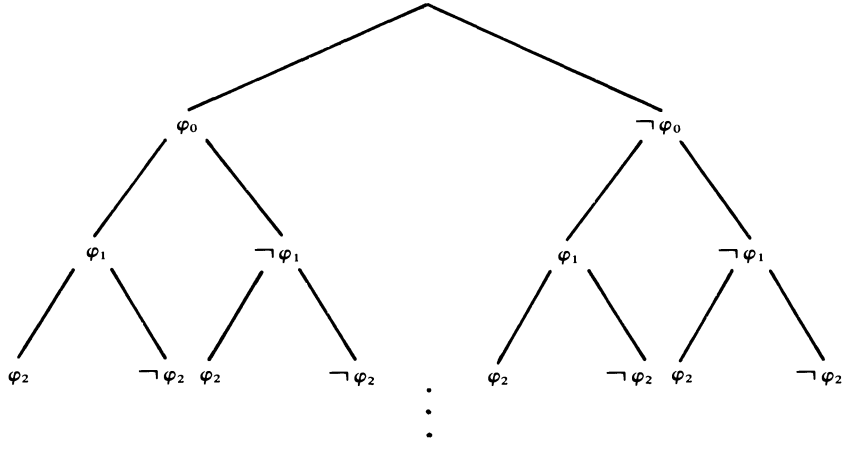
$$F(S_{\mathfrak{N}}x) = S_{\mathfrak{M}}(Fx),$$

where S denotes the successor functions of the models. The recursion equations in $\mathfrak{M}$ and induction in $\mathfrak{N}$ verify that F is an isomorphism of $\mathfrak{N}$ onto an initial segment of $\mathfrak{M}$. □

Using this lemma, we may present Kreisel's proof of the following.

6.2.4. SECOND INCOMPLETENESS THEOREM. $\mathbf{PA} \not\vdash \mathrm{Con}_{\mathbf{PA}}$.

PROOF. Let $\varphi_0, \varphi_1, \ldots$ be an enumeration of sentences of the language described in the proof of 6.1.1. That proof can be viewed as an attempt to choose an infinite consistent path through the following tree:



We may assume for definiteness that the construction proceeds by taking the *leftmost* consistent path. Choose φ such that $\mathbf{PA} \vdash \varphi \leftrightarrow \neg \mathrm{Tr}_{\mathfrak{M}}(\ulcorner \varphi \urcorner)$, for the truth definition, Tr_M, of the model constructed. Let $\varphi = \varphi_{n_0}$ in the enumeration. The tree, as defined to the n_0-th level, is absolute. I.e. it is the same in every model. (Note: This is not true for the infinite tree simply because any non-standard model $\mathfrak{M}$ will encode a level for φ_N for non-standard integers N. But the *finite* trees are fixed.)

Assume $\mathbf{PA} \vdash \mathrm{Con}_{\mathbf{PA}}$. Let $\mathfrak{N}_0 \models \mathbf{PA}$. Then there is a model, $\mathfrak{N}_1$, definable in $\mathfrak{N}_0$: $\mathfrak{N}_1 <_{\mathrm{d}} \mathfrak{N}_0$. But $\mathfrak{N}_1$ is also a model of $\mathrm{Con}_{\mathbf{PA}}$ and there is an $\mathfrak{N}_2 <_{\mathrm{d}} \mathfrak{N}_1$. Repeating, we get an infinite sequence,

$$\mathfrak{N}_0 \succ_d \mathfrak{N}_1 \succ_d \mathfrak{N}_2 \succ_d \cdots,$$

such that (say)

$$\mathfrak{N}_0 \models \varphi_{n_0}, \quad \mathfrak{N}_1 \models \neg\varphi_{n_0}, \quad \mathfrak{N}_2 \models \varphi_{n_0}, \ldots.$$

We now use construction to derive a contradiction. Given $\mathfrak{N}_i$, let $\boldsymbol{\varphi}^i = \langle \varphi_0^{e_{0i}}, \varphi_1^{e_{1i}}, \ldots, \varphi_{n_0 \cdot i}^{e_{n_0 \cdot i}} \rangle$ denote the portion of the path used in constructing $\mathfrak{N}_{i+1}$ — where $\varphi_j^0 = \varphi_j$, $\varphi_j^1 = \neg\varphi_j$, and $e_{ji} \in \{0, 1\}$. Recall that $\boldsymbol{\varphi}^i$ is the leftmost consistent path (as viewed) in $\mathfrak{N}_i$.

Either using facts that $\mathrm{Pr}_{\mathbf{PA}}$ is Σ_1 and that Σ_1 sentences are preserved under end-extensions, or using D2 and the fact that

$$\mathbf{PA} + \mathrm{Con}_{\mathbf{PA}} \vdash \forall\psi\,(\mathrm{Pr}_{\mathbf{PA}}(\ulcorner\psi\urcorner) \rightarrow \mathrm{Tr}_M(\ulcorner\psi\urcorner)),$$

one sees that $\boldsymbol{\varphi}^{i+1}$ can never lie to the left of $\boldsymbol{\varphi}^i$. For, once $\mathfrak{N}_i$ says that a sequence is inconsistent, every resulting $\mathfrak{N}_j$ will also assert its inconsistency. Put differently, larger models can allow new proofs (even of inconsistencies) e.g. by means of non-standard axioms encoded by infinite integers; but they cannot erase old proofs.

Thus $\boldsymbol{\varphi}^{i+1}$ cannot lie to the left of $\boldsymbol{\varphi}^i$. Furthermore, $\boldsymbol{\varphi}^{i+1} \neq \boldsymbol{\varphi}^i$ since

$$\varphi_{n_0 \cdot i+1}^{e_{n_0 \cdot i+1}} = \varphi_{n_0}^{1-e_{n_0 \cdot i}} \leftrightarrow \neg\varphi_{n_0 \cdot i}^{e_{n_0 \cdot i}}.$$

Thus the path $\boldsymbol{\varphi}^{i+1}$ lies properly to the right of $\boldsymbol{\varphi}^i$.

But the tree determined by $\varphi_0, \ldots, \varphi_{n_0}$ is finite and there are only 2^{n_0+1} different paths through this tree. This contradicts the assumption $\mathbf{PA} \vdash \mathrm{Con}_{\mathbf{PA}}$ by which we obtained an infinite sequence of paths: $\boldsymbol{\varphi}^0, \boldsymbol{\varphi}^1, \ldots$. □

*6.3. Comments

With Theorems 6.2.1 and 6.2.4, we have gone full circle: We have gotten back to the results with which we began this chapter. The present proofs differ somewhat from the originals and it is worth making a few comparisons.

Let us first comment on the forms of the independent sentences given by the two proofs of the First Incompleteness Theorem. "The" sentence which asserts its own unprovability is

(i) unique up to provable equivalence;

(ii) Π_1 and hence true.

"The" sentence asserting its falsity in the model constructed is

(i') not unique — for, if $\varphi \leftrightarrow \neg\,\mathrm{Tr}_M(\ulcorner\varphi\urcorner)$, then

$$\neg\varphi \leftrightarrow \neg\,\mathrm{Tr}_M(\ulcorner\neg\varphi\urcorner);$$

(ii′) Δ_2 and, by (i′), there is no obvious way of deciding its truth or falsity.

(ii′) can be gotten around as follows: One of φ, $\neg\varphi$ is true. Let $\exists x\,\forall y\,\psi xy$ be the Σ_2 form of the true statement. Then, for some n, $\chi = \forall y\,\psi\bar{n}y$ is true. But $\mathbf{PA} \not\vdash \chi$ as one would then have $\mathbf{PA} \vdash \exists x\,\forall y\,\psi xy$. But $\mathbf{PA} \not\vdash \neg\chi$ since χ is true.

The model-theoretic proof of the First Incompleteness Theorem given here is similar to the classical one — for, once one assumes completeness, Tr_M is the same as $\mathrm{Pr}_{\mathbf{PA}}$. The model-theoretic proof of the Second Incompleteness Theorem differs radically from the classical one and we note some differences in the sort of information they yield:

(i) The classical proof readily yields the formalized version, $\mathbf{PA} \vdash \mathrm{Con}_{\mathbf{PA}} \rightarrow \mathrm{Con}_{\mathbf{PA}+\neg\mathrm{Con}_{\mathbf{PA}}}$. Further, it applies directly to weaker theories like **PRA.**

(ii) While the classical proof yields the existence of *some* model in which $\mathrm{Con}_{\mathbf{PA}}$ fails, the model-theoretic one shows that, for any presentation of the Henkin construction (as given by the encoding, the enumeration $\varphi_0, \varphi_1, \ldots$, etc.), there is a number m such that, for *any* model $\mathfrak{N}$ of **PA**, the sequence

$$(*) \qquad \mathfrak{N} \succ_d \mathfrak{N}_1 \succ_d \cdots,$$

determined by the given presentation, must stop after fewer than m steps with a model in which $\mathrm{Con}_{\mathbf{PA}}$ is false. (Of course, by the classical proof, there is a presentation of the Henkin construction with a very short sequence (*) — simply let $\varphi_0 = \neg\mathrm{Con}_{\mathbf{PA}}$. The present proof works for all enumerations $\varphi_0, \ldots$.)

References

The following is a very biased selection of the many papers on the topic of this chapter. It includes some papers whose contents were not discussed

FEFERMAN, S.

[1962] Transfinite recursive progressions of axiomatic theories, *J. Symbolic Logic*, **27**, 259–316.

GÖDEL, K.

[1931] Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I, *Monatsh. Math. Phys.*, **38**, 173–198.

HASENJÄGER, G.

[1953] Eine Bemerkung zu Henkin's Beweis für die Vollständigkeit des Prädikatenkalküls der ersten Stufe, *J. Symbolic Logic*, **18**, 42–48.

HILBERT, D. and P. BERNAYS
[1970] *Grundlagen der Mathematik*, I (Springer, Berlin, 2nd ed.).
JEROSLOW, R.G.
[1973] Redundancies in the Hilbert–Bernays derivability conditions for Gödel's second incompleteness theorem, J. Symbolic Logic, **38**, 359–367.
KENT, C.F.
[1973] The relation of A to Prov$\ulcorner A \urcorner$ in the Lindenbaum sentence algebra, *J. Symbolic Logic*, **38**, 295–298.
KREISEL, G. and A. LEVY
[1968] Reflection principles and their use for establishing the complexity of axiomatic systems, *Z. Math. Logik Grundlagen Math.*, **14**, 97–142.
KREISEL, G. and G. TAKEUTI
[1974] Formally self-referential propositions in cut-free classical analysis and related systems, *Dissertationes Math.*, **118**, 1–50.
LÖB, M.H.
[1955] Solution of a problem of Leon Henkin, *J. Symbolic Logic*, **20**, 115–118.
MESCHKOWSKI, H.
[1973] *Hundert Jahre Mengenlehre* (Deutscher Taschenbuch Verlag; München).
REID, C.
[1970] *Hilbert* (Springer, Berlin).
ROSSER, J.B.
[1936] Extensions of some theorems of Gödel and Church, *J. Symbolic Logic*, **1**, 87–91.
SAMBIN, G.
[0000] An effective fixed-point theorem in intuitionistic diagonalizable algebras, *Studia Logica*, to appear.
SMORYNSKI, C.
[0000] ω-consistency and reflection, in: *Proceedings of the* 1975 *Logic Colloquium at Clermont-Ferrand*, to appear.
SOLOVAY, R.
[1976] Provability interpretations of modal logic, *Isr. J. Math.*, **25**, 287–304.