Gödel's Incompleteness Theorems

Omar

November 13, 2026

Contents

About this report **3**

1 Setting the stage **3**
1.1 Effectively axiomatized theories . . . 3
1.2 The language L_A . . . 5
1.3 Expressing and capturing . . . 5
1.4 A preview: sufficiently strong theories . . . 6

2 The theories Q and PA **7**
2.1 Robinson Arithmetic . . . 7
2.2 Δ_0, Σ_1 and Π_1 wffs . . . 8
2.3 What Q can prove . . . 8

3 Primitive recursive functions **9**

4 L_A expresses and Q captures every p.r. function **10**
4.1 Gödel's β-function . . . 10
4.2 Expressing p.r. functions . . . 11
4.3 Capturing functions . . . 12
4.4 Q is p.r. adequate . . . 12

5 Arithmetization of syntax **14**
5.1 Gödel numbering . . . 14
5.2 Syntax is primitive recursive . . . 14
5.3 Diagonalization . . . 15

6 The First Incompleteness Theorem **15**
6.1 The Gödel sentence . . . 15
6.2 The semantic argument . . . 16
6.3 The syntactic argument . . . 16
6.4 Provability predicates and the Diagonalization Lemma . . . 17
6.5 Rosser's improvement . . . 18

7 The Second Incompleteness Theorem **18**
7.1 The consistency sentence . . . 18
7.2 The idea . . . 19
7.3 The proof . . . 19

8 Why the derivability conditions hold **20**

9 What was taken on trust **21**

About this report

This report proves Gödel's two incompleteness theorems for formal theories of arithmetic. I follow Peter Smith's *An Introduction to Gödel's Theorems* (2nd ed.) closely. Results with no tag are small lemmas I needed to make a step explicit.

Some lemmas are proved in full and some only sketched. The rule I used is this: anything that does real work in the argument gets a full proof, and anything that is a long but unsurprising check ("this relation can be computed with bounded loops", "Q proves this obvious fact about $\leq$") gets a sketch. Whenever a proof is only a sketch I say so, and Section 9 lists all of them in one place.

Here is roughly what gets proved. Call a theory *nice* if it is consistent, its axioms and proofs can be recognized by a primitive recursive test, and it proves everything the weak arithmetic Q proves (precise definitions come later).

- **First Theorem** (in Rosser's form). If T is nice, there is a sentence φ of the language of arithmetic such that $T \nvdash \varphi$ and $T \nvdash \neg\varphi$.
- **Second Theorem.** If T is nice and also has a little induction (enough to contain $\mathsf{I}\Sigma_1$, which PA does), then T cannot prove the sentence Con_T that says T is consistent.

The plan is as follows. Section 1 sets up the basic notions and gives an easy incompleteness argument which already has the shape of the real one, but rests on an assumption we can't yet justify. Sections 2 to 4 replace that assumption with something we can prove: the very weak theory Q can correctly compute every primitive recursive function. Section 5 codes syntax as arithmetic, so that "m codes a proof of the sentence coded by n" becomes a primitive recursive relation between numbers. Section 6 puts the pieces together for the First Theorem, and Sections 7 and 8 do the Second.

1 Setting the stage

1.1 Effectively axiomatized theories

An algorithm here is a finite list of deterministic instructions that works on finite inputs (numbers, or finite strings over a finite alphabet). A total function $f : \mathbb{N} \to \mathbb{N}$ is *effectively computable* if some algorithm, given any n, halts after finitely many steps with output $f(n)$.

Definition 1.1 (Smith §§3.1, 3.3)**.** Let $\Sigma \subseteq \mathbb{N}$.

- Σ is *effectively decidable* if its characteristic function is effectively computable.
- Σ is *effectively enumerable* if $\Sigma = \emptyset$ or there is an effectively computable $f : \mathbb{N} \to \mathbb{N}$ whose range is exactly Σ, so that $f(0), f(1), f(2), \ldots$ lists Σ (repetitions allowed).

The same definitions apply to sets of finite strings over a finite alphabet, with strings in place of numbers.

Both notions only make sense for sets of finite objects, since an algorithm has to be able to read its input. They are not given a precise mathematical definition here (that would need Turing machines or μ-recursive functions, together with the Church–Turing thesis). This doesn't affect the proof. The informal notions are only used in this first section, and from Section 3 on everything is done with primitive recursive functions, which are defined precisely.

My completeness notes define a logic (a language with a proof calculus) but not a theory. A theory is a logic together with a choice of axioms.

Definition 1.2 (cf. Smith §§4.1–4.3)**.** A (formal) *theory* T consists of

(a) a first-order language L, usually with an intended interpretation;

(b) a set Ax_T of sentences of L, the *axioms* of T;

(c) a proof system for first-order logic over L.

A *T-proof* of φ is a proof of φ in the proof system from axioms in Ax_T, and we write $T \vdash \varphi$ if one exists. In the notation of my completeness notes, $T \vdash \varphi$ is $\mathrm{Ax}_T \vdash \varphi$, i.e. $\Gamma \vdash \varphi$ is derivable for some finite $\Gamma \subseteq \mathrm{Ax}_T$. The *theorems* of T are the sentences φ with $T \vdash \varphi$.

Some books define a theory as a set of sentences closed under consequence, which would be the set of theorems here. I follow Smith in keeping the axioms as part of the theory, because whether T is effectively (or p.r.) axiomatized depends on how the axioms are presented, not only on what follows from them.

Definition 1.3 (Smith §4.3)**.** A theory T is *effectively axiomatized* if

(i) its language has a finite alphabet and it is effectively decidable which strings are wffs and which are sentences;

(ii) it is effectively decidable which wffs are axioms of T;

(iii) it has a proof system in which it is effectively decidable whether a given array of wffs is a correctly built proof from the axioms of T.

Definition 1.4 (Smith §4.4)**.** Let T be a theory and φ range over the sentences of its language.

- T is *consistent* if there is no φ with $T \vdash \varphi$ and $T \vdash \neg\varphi$.
- T is *negation-complete* if for every φ, $T \vdash \varphi$ or $T \vdash \neg\varphi$.
- T *decides* φ if $T \vdash \varphi$ or $T \vdash \neg\varphi$. It *correctly decides* φ if $T \vdash \varphi$ when φ is true and $T \vdash \neg\varphi$ when φ is false.
- T is *decidable* if the set of its theorems is effectively decidable.
- T is *sound* if every theorem of T is true on the intended interpretation of its language.

Theorem 1.5 (Smith 4.1)**.** If T is effectively axiomatized, then its wffs, its sentences, its proofs and its theorems can each be effectively enumerated.

Proof. List all strings over the finite alphabet: the strings of length 1 in alphabetical order, then those of length 2, and so on. Checking whether a string is a wff (or a sentence) is decidable, so throwing away the failures leaves an effective enumeration of the wffs (or sentences).

Treating finite sequences of wffs as strings over a slightly bigger alphabet (add a separator symbol), we can enumerate them the same way. Proof-hood is decidable, so filtering gives an enumeration of the proofs. Finally, run through the proofs and record the last line of each one that ends in a sentence. That lists exactly the theorems. ⊠

Being able to list the theorems is not the same as being able to decide whether a given sentence is one. If φ is not a theorem, the listing just never produces it, and we never find out. Negation-completeness gives this.

Theorem 1.6 (Smith 4.2)**.** Any consistent, effectively axiomatized, negation-complete theory is decidable.

Proof. Given a string, first check whether it is a sentence; if not, it is not a theorem. If it is a sentence φ, start enumerating the theorems. By negation-completeness, one of φ and $\neg\varphi$ eventually turns up. If φ turns up, answer yes. If $\neg\varphi$ turns up, answer no; this is correct because T is consistent, so φ can't also be a theorem. Either way the procedure halts with the right answer. ⊠

1.2 The language L_A

The language of basic arithmetic is the interpreted language $L_A = \langle \mathcal{L}_A, \mathcal{I}_A \rangle$. Following Smith, I write formal expressions in sans-serif and informal mathematics in italics, so $1 + 2 = 3$ is a fact about numbers while $\mathsf{1} + \mathsf{2} = \mathsf{3}$ is a sentence of L_A. Greek letters are metavariables for formal expressions and are never part of the formal language.

Definition 1.7 (Smith §5.2)**.** *Syntax.* $\mathcal{L}_A$ is a first-order language with identity whose non-logical vocabulary is the constant $\mathsf{0}$, the one-place function symbol S, and the two-place function symbols $+$ and $\times$ (written infix). Terms are built from $\mathsf{0}$ and variables using S, $+$ and $\times$. The *standard numeral* for n is $\mathsf{SS}\ldots\mathsf{S0}$ with n occurrences of S, written $\overline{\mathsf{n}}$. Atomic wffs have the form $\sigma = \tau$.

Semantics. $\mathcal{I}_A$ has the natural numbers as domain. Closed terms get values by $val[\mathsf{0}] = 0$, $val[\mathsf{S}\tau] = val[\tau] + 1$, $val[(\sigma + \tau)] = val[\sigma] + val[\tau]$ and $val[(\sigma \times \tau)] = val[\sigma] \times val[\tau]$, so $val[\overline{\mathsf{n}}] = n$. A sentence $\sigma = \tau$ is true iff its two sides have the same value, connectives are read truth-functionally, and $\exists\xi\,\varphi(\xi)$ is true iff $\varphi(\overline{\mathsf{n}})$ is true for some n (dually for $\forall$).

The quantifier clause can be this simple because every element of the domain is named by a numeral. With a general structure we would need assignments and the full Tarskian satisfaction relation, but here the two give the same answers.

1.3 Expressing and capturing

There are two different senses in which a wff can "be about" a numerical property.

Definition 1.8 (Smith §5.4: expressing)**.** A wff $\varphi(\mathsf{x})$ with one free variable *expresses* the numerical property P in L_A if for every n: if n has P then $\varphi(\overline{\mathsf{n}})$ is true, and if n lacks P then $\neg\varphi(\overline{\mathsf{n}})$ is true. Likewise for relations, and a one-place function f is expressed by $\varphi(\mathsf{x}, \mathsf{y})$ if for all m, n: $f(m) = n$ implies $\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is true, and $f(m) \neq n$ implies $\neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is true.

Definition 1.9 (Smith §5.5: capturing)**.** A theory T *captures* the property P by the wff $\varphi(\mathsf{x})$ if for every n: if n has P then $T \vdash \varphi(\overline{\mathsf{n}})$, and if n lacks P then $T \vdash \neg\varphi(\overline{\mathsf{n}})$. Likewise for relations. (Capturing *functions* needs a little more care; see Section 4.3.)

Expressing depends only on the richness of the language. Capturing depends on what the theory can prove. If T is sound and φ captures P in T, then φ also expresses P, since everything T proves is true. The converse fails: we will see in Theorem 6.14 that provability in T is expressible but not capturable in T.

Characteristic functions. I use Smith's convention, which is the opposite of the one in my earlier notes: the characteristic function c_P of a property P takes the value 0 when n has P and 1 when it doesn't.

1.4 A preview: sufficiently strong theories

Before the real proof, here is an argument that already has the same diagonal shape. It requires on an assumption we can't justify yet, and the rest of the report is mostly about replacing that assumption with something we can prove.

Definition 1.10 (Smith §7.1)**.** A theory of arithmetic T is *sufficiently strong* if it captures every effectively decidable numerical property.

This is a natural thing to ask of a theory. If we can already decide whether n has P by a mechanical check, we'd want the formal theory to be able to prove the answer case by case.

Theorem 1.11 (Smith 7.1)**.** No consistent, sufficiently strong, effectively axiomatized theory of arithmetic is decidable.

Proof. Suppose T is all four: consistent, sufficiently strong, effectively axiomatized and decidable. By Theorem 1.5 we can effectively enumerate its wffs, and keeping only those with exactly the variable x free gives an effective list

$$\varphi_0(\mathsf{x}),\ \varphi_1(\mathsf{x}),\ \varphi_2(\mathsf{x}),\ \ldots$$

Define the *diagonal* property D by

$$n \text{ has } D \iff T \vdash \neg\varphi_n(\overline{\mathsf{n}}).$$

D is decidable: given n, run the enumeration to find $\varphi_n(\mathsf{x})$, substitute $\overline{\mathsf{n}}$ for x, put $\neg$ in front, and use the decision procedure for T on the result.

So D is captured by some wff in the list, say $\varphi_d(\mathsf{x})$:

$$n \text{ has } D \implies T \vdash \varphi_d(\overline{\mathsf{n}}), \qquad n \text{ lacks } D \implies T \vdash \neg\varphi_d(\overline{\mathsf{n}}).$$

Now look at $n = d$. If d lacked D, then $T \vdash \neg\varphi_d(\overline{\mathsf{d}})$, which by the definition of D means d has D. So d has D. Then the definition of D gives $T \vdash \neg\varphi_d(\overline{\mathsf{d}})$, while capturing gives $T \vdash \varphi_d(\overline{\mathsf{d}})$, so T is inconsistent. ⊠

Theorem 1.12 (Smith 7.2)**.** If T is a consistent, sufficiently strong, effectively axiomatized theory of arithmetic, then T is not negation-complete.

Proof. If it were negation-complete, it would be decidable by Theorem 1.6, contradicting Theorem 1.11. ⊠

Remark. The argument also shows incompletability. Add an undecided sentence (or its negation) as a new axiom and the result is still consistent, sufficiently strong and effectively axiomatized, so it is still incomplete.

Two things stop this from being Gödel's theorem. First, we have no proof that any actual theory is sufficiently strong; that would need a precise account of *all* decidable properties. Second, the property D is decided by a procedure that uses the hypothetical decision procedure for T, so the argument only works by contradiction from decidability.

Gödel's proof gets around both problems. In place of "all decidable properties" it uses the primitive recursive ones, and we can prove that even the weak theory Q captures all of those (Theorem 4.10). In place of the abstract list $\varphi_0, \varphi_1, \ldots$ it uses Gödel numbers, which let a wff talk about wffs. And it diagonalizes directly on provability, which is primitive recursive to check (given a proof) even though theoremhood is not decidable.

2 The theories Q and PA

2.1 Robinson Arithmetic

Definition 2.1 (Smith §10.3)**.** *Robinson Arithmetic* Q is the theory in $\mathcal{L}_A$ with classical first-order logic with identity and these seven axioms:

Axiom 1	$\forall x(0 \neq Sx)$
Axiom 2	$\forall x \forall y(Sx = Sy \rightarrow x = y)$
Axiom 3	$\forall x(x \neq 0 \rightarrow \exists y(x = Sy))$
Axiom 4	$\forall x(x + 0 = x)$
Axiom 5	$\forall x \forall y(x + Sy = S(x + y))$
Axiom 6	$\forall x(x \times 0 = 0)$
Axiom 7	$\forall x \forall y(x \times Sy = (x \times y) + x)$

The axioms are true on $\mathcal{I}_A$ and first-order logic is sound, so Q is sound and therefore consistent.

Remark (on the proof system)**.** Smith officially gives Q and PA a Hilbert-style logic, so that proofs are plain finite sequences of wffs; this makes them easy to code by numbers in Section 5. My completeness notes used Hermes' sequent calculus instead. Nothing here depends on the choice. Both systems are sound and complete, so both prove exactly the semantic consequences of the axioms, and the set of theorems of T is the same either way. Only the coding of proofs changes, and checking a sequent-calculus derivation is just as mechanical as checking a Hilbert-style one.

Theorem 2.2 (Smith 10.1–10.6)**.**

(i) If $m \neq n$ then $Q \vdash \overline{m} \neq \overline{n}$.

(ii) If τ is a closed term with value t, then $Q \vdash \tau = \overline{t}$.

(iii) Q correctly decides every quantifier-free sentence.

Proof sketch. (i) Say $m < n$ and put $k = n - m - 1$. Axiom 1 gives $0 \neq S\overline{k}$. Axiom 2, used m times in contrapositive form, adds an S to both sides each time, ending at $\overline{m} \neq \overline{n}$. (ii) By induction on the construction of τ. The cases $S\tau$, $+$ and $\times$ use Axioms 4–7 to evaluate, e.g. $\overline{m} + \overline{n} = \overline{m + n}$ by peeling successors off the right-hand argument with Axiom 5, n times. (iii) By (i) and (ii), Q correctly decides every equation between closed terms, and propositional logic then decides every truth-functional combination of them correctly. ⊠

Q is very weak at proving generalizations. Here is the standard example.

Theorem 2.3 (Smith 10.8)**.** $Q \nvdash \forall x(0 + x = x)$, although $Q \vdash 0 + \overline{n} = \overline{n}$ for every n.

Proof sketch. The instances follow from Theorem 2.2(ii). For the generalization, it is enough (by soundness of first-order logic) to find a structure where Q's axioms are true but $\forall x(0 + x = x)$ is false. Take the natural numbers plus two extra objects a, b. Let S act normally on numbers and fix a and b. Let $+$ be normal on numbers, with $a + n = a$, $b + n = b$, and $x + a = b$, $x + b = a$ for every x. Axioms 1–5 can be checked to hold, and $0 + a = b \neq a$. Multiplication can be extended in a similar way so that Axioms 6 and 7 hold too. ⊠

Definition 2.4 (Smith §13.2)**.** *Peano Arithmetic* PA is Q plus the universal closure of every instance of the Induction Schema

$$\{\varphi(0) \wedge \forall \mathsf{x}(\varphi(\mathsf{x}) \to \varphi(\mathsf{Sx}))\} \to \forall \mathsf{x}\, \varphi(\mathsf{x}),$$

for every open wff $\varphi(\mathsf{x})$ of $\mathcal{L}_A$. The theory $\mathsf{I\Sigma_1}$ is Q plus the instances of the schema where φ is Σ_1 (defined below).

A theory T *extends* Q if every L_A wff is a wff of T's language and T proves every theorem of Q. So PA and $\mathsf{I\Sigma_1}$ extend Q, and Q counts as extending itself.

2.2 Δ_0, Σ_1 and Π_1 wffs

Definition 2.5 (Smith §11.2)**.** $\xi \leq \zeta$ abbreviates $\exists \nu(\nu + \xi = \zeta)$, where ν is a variable not occurring in ξ or ζ. The *bounded quantifiers* are

$$(\forall \xi \leq \kappa)\varphi \;\text{ for }\; \forall \xi(\xi \leq \kappa \to \varphi), \qquad (\exists \xi \leq \kappa)\varphi \;\text{ for }\; \exists \xi(\xi \leq \kappa \wedge \varphi),$$

where κ is a numeral or a variable different from ξ.

Definition 2.6 (Smith §§11.4, 11.5)**.** The Δ_0 *wffs* are built from atomic wffs $\sigma = \tau$ and $\sigma \leq \tau$ using connectives and bounded quantifiers only. The Σ_1 *wffs* are built from Δ_0 wffs using $\wedge$, $\vee$, bounded quantifiers and unbounded $\exists$. The Π_1 *wffs* are built from Δ_0 wffs using $\wedge$, $\vee$, bounded quantifiers and unbounded $\forall$. A wff is Σ_1*-equivalent* (Π_1*-equivalent*) if elementary logical manipulations turn it into a Σ_1 (Π_1) wff.

Negation is missing from the Σ_1 clauses on purpose: negating an unbounded $\exists$ would give an unbounded $\forall$. Pushing negations inwards with de Morgan's laws and the quantifier dualities shows that the negation of a Δ_0 wff is Δ_0, the negation of a Σ_1 wff is Π_1-equivalent, and vice versa. A typical Π_1 sentence is Goldbach's conjecture: "every even number above two is a sum of two primes" has one unbounded $\forall$ in front of a Δ_0 wff. So Smith calls Π_1 sentences sentences *of Goldbach type*.

2.3 What Q can prove

Theorem 2.7 (Smith §11.1 and Thm 11.1, in part)**.** Q captures the relation $m \leq n$ by the wff $\mathsf{x} \leq \mathsf{y}$. Moreover, for every n and every wff $\varphi(\mathsf{x})$:

(O3) $\mathsf{Q} \vdash \forall \mathsf{x}(\mathsf{x} \leq \overline{\mathsf{n}} \to \{\mathsf{x} = \mathsf{0} \vee \mathsf{x} = \overline{\mathsf{1}} \vee \cdots \vee \mathsf{x} = \overline{\mathsf{n}}\})$;

(O4) if $\mathsf{Q} \vdash \varphi(\overline{\mathsf{0}}), \ldots, \mathsf{Q} \vdash \varphi(\overline{\mathsf{n}})$, then $\mathsf{Q} \vdash (\forall \mathsf{x} \leq \overline{\mathsf{n}})\varphi(\mathsf{x})$;

(O5) if $\mathsf{Q} \vdash \varphi(\overline{\mathsf{k}})$ for some $k \leq n$, then $\mathsf{Q} \vdash (\exists \mathsf{x} \leq \overline{\mathsf{n}})\varphi(\mathsf{x})$;

(O8) $\mathsf{Q} \vdash \forall \mathsf{x}(\mathsf{x} \leq \overline{\mathsf{n}} \vee \overline{\mathsf{n}} \leq \mathsf{x})$.

The same holds for any T extending Q.

Proof sketch. If $m \leq n$, then $\mathsf{Q} \vdash \overline{\mathsf{k}} + \overline{\mathsf{m}} = \overline{\mathsf{n}}$ with $k = n - m$, so $\mathsf{Q} \vdash \exists \mathsf{v}(\mathsf{v} + \overline{\mathsf{m}} = \overline{\mathsf{n}})$. If $m > n$, suppose $\mathsf{a} + \overline{\mathsf{m}} = \overline{\mathsf{n}}$ inside Q. Axioms 5 and 2 strip one S off each side at a time until we reach $\mathsf{a} + \overline{\mathsf{m} - \mathsf{n}} = \mathsf{0}$, i.e. $\mathsf{S}(\mathsf{a} + \overline{\mathsf{m} - \mathsf{n} - \mathsf{1}}) = \mathsf{0}$, contradicting Axiom 1. So $\mathsf{Q} \vdash \neg\exists \mathsf{v}(\mathsf{v} + \overline{\mathsf{m}} = \overline{\mathsf{n}})$.

(O3) goes by induction on n outside Q, using Axiom 3 to split into the cases $\mathsf{a} = \mathsf{0}$ and $\mathsf{a} = \mathsf{Sa'}$ and then Axioms 5 and 2 to pass from $\mathsf{Sa'} \leq \mathsf{S}\overline{\mathsf{n}}$ to $\mathsf{a'} \leq \overline{\mathsf{n}}$. (O4) follows from (O3): given $\mathsf{a} \leq \overline{\mathsf{n}}$, (O3) gives a disjunction $\mathsf{a} = \overline{\mathsf{0}} \vee \cdots \vee \mathsf{a} = \overline{\mathsf{n}}$, and in each case $\varphi(\mathsf{a})$ follows from the corresponding $\varphi(\overline{\mathsf{k}})$. (O5) is immediate from $\mathsf{Q} \vdash \overline{\mathsf{k}} \leq \overline{\mathsf{n}}$. (O8) is another outside induction on n. Smith works these out in his §11.8; they are routine and I won't repeat them. ⊠

Theorem 2.8 (Smith 11.3)**.** Q correctly decides every Δ_0 sentence.

Proof. Say a Δ_0 sentence has *degree* k if it is built from atomic wffs by k applications of connectives and bounded quantifiers. We argue by (informal, outside Q) course-of-values induction on k.

Degree 0. The sentence is $\sigma = \tau$ or $\sigma \leq \tau$ with σ, τ closed. Equations are handled by Theorem 2.2. For $\sigma \leq \tau$, Q proves $\sigma = \overline{\mathsf{s}}$ and $\tau = \overline{\mathsf{t}}$ where s, t are the values, and since Q captures $\leq$ it correctly decides $\overline{\mathsf{s}} \leq \overline{\mathsf{t}}$; substituting identicals, it correctly decides $\sigma \leq \tau$.

Degree $k+1$. If the sentence is a truth-functional compound of lower-degree sentences, these are correctly decided by the induction hypothesis, and classical propositional logic then correctly decides the compound. Suppose it is $(\forall \xi \leq \overline{\mathsf{n}})\varphi(\xi)$. (The bound must be a numeral, since the whole thing is a sentence.) If it is true, each of $\varphi(\overline{\mathsf{0}}), \ldots, \varphi(\overline{\mathsf{n}})$ is a true sentence of lower degree, so Q proves each, and (O4) gives the bounded generalization. If it is false, some $\varphi(\overline{\mathsf{m}})$ with $m \leq n$ is false, so $\mathsf{Q} \vdash \neg\varphi(\overline{\mathsf{m}})$, and (O5) gives $(\exists \xi \leq \overline{\mathsf{n}})\neg\varphi(\xi)$, which is equivalent to the negation of our sentence. The bounded $\exists$ case is similar. ⊠

Theorem 2.9 (Smith 11.5: Q is Σ_1-complete)**.** If φ is a true Σ_1 sentence, then $\mathsf{Q} \vdash \varphi$. So the same holds for any theory extending Q.

Proof. Induction on the number of steps needed to build φ from Δ_0 wffs. The base case is Theorem 2.8. If φ is a true conjunction, both conjuncts are true and provable; if it is a true disjunction, one disjunct is true and provable. If φ is a true $(\forall \xi \leq \overline{\mathsf{n}})\psi(\xi)$, each $\psi(\overline{\mathsf{k}})$ with $k \leq n$ is true and provable, so (O4) applies; the bounded $\exists$ case uses (O5). Finally, if φ is a true $\exists \xi\, \psi(\xi)$, then some $\psi(\overline{\mathsf{n}})$ is true, hence provable, and existential introduction gives φ. ⊠

Q proves every true Σ_1 sentence, but it doesn't prove every true Π_1 sentence: Theorem 2.3 is a counterexample. That asymmetry is what makes room for incompleteness. Here is the useful consequence.

Corollary 2.10 (cf. Smith 11.6, 11.7)**.** Let T be consistent and extend Q, and let φ be a Π_1 (or Π_1-equivalent) sentence. If $T \nvdash \neg\varphi$, then φ is true.

Proof. If φ were false, $\neg\varphi$ would be true and Σ_1-equivalent, so Q, and hence T, would prove it. ⊠

3 Primitive recursive functions

Definition 3.1 (Smith §14.2)**.** The *initial functions* are the successor S, the zero function $Z(x) = 0$, and the projections $I^k_i(x_1, \ldots, x_k) = x_i$.

- f is defined by *composition* from g and h if $f(\vec{x}, \vec{y}, \vec{z}) = h(\vec{x}, g(\vec{y}), \vec{z})$.
- f is defined by *primitive recursion* from g and h if

$$f(\vec{x}, 0) = g(\vec{x}), \qquad f(\vec{x}, Sy) = h(\vec{x}, y, f(\vec{x}, y)).$$

  (When $\vec{x}$ is empty, g is just a number.)

The *primitive recursive* (p.r.) functions are the initial functions and everything obtainable from them by finitely many compositions and primitive recursions. A property or relation is p.r. if its characteristic function is.

For example, $x{+}0 = x$, $x{+}Sy = S(x{+}y)$ defines addition from the successor, then multiplication is defined from addition, and so on.

Theorem 3.2 (Smith 14.1)**.** Every p.r. function is effectively computable, using only (possibly nested) bounded `for` loops.

Proof sketch. The initial functions are trivial and composition just feeds one program's output into another. For primitive recursion, compute $f(\vec{m}, n)$ by setting a register to $g(\vec{m})$ and running `for` $y = 0$ `to` $n - 1$: register $:= h(\vec{m}, y, \text{register})$. The number of iterations is fixed before the loop starts. ⊠

The converse fails [Smith 14.2]: a diagonal argument over an enumeration of p.r. definitions gives a computable function that is not p.r. We won't need this.

Write $(\mu x \le n)P(x)$ for the least $x \le n$ with $P(x)$, or n if there is none.

Theorem 3.3 (Smith 14.3, 14.4)**.**

(A) If f is p.r., so is the relation $f(\vec{x}) = y$.

(B) Truth-functional combinations of p.r. properties and relations are p.r.

(C) Bounded quantifications $(\exists x \le g(n))P(x)$ and $(\forall x \le g(n))P(x)$ of a p.r. P are p.r. when g is.

(D) If P and g are p.r., so is $f(n) = (\mu x \le g(n))P(x)$.

(E) Definition by cases from p.r. functions and p.r. conditions gives a p.r. function.

In particular the following are p.r.: $m = n$, $m < n$, $m \le n$, $m \mid n$, $\mathit{Prime}(n)$, the function $\pi(n)$ giving the $(n{+}1)$-th prime, the function $\mathit{exf}(n, i)$ giving the exponent of π_i in the prime factorization of n, and $\mathit{len}(n)$, the number of distinct prime factors of n.

Proof sketch. Each claim says that something computable with bounded loops can be checked or computed with bounded loops, and Smith gives explicit p.r. definitions. For instance, for (C), if p is the characteristic function of P, then $k(0) = p(0)$, $k(Sy) = p(Sy) \times k(y)$ defines a p.r. k with $k(n) = 0$ exactly when some $x \le n$ has P. For the examples: $m \mid n \iff (\exists y \le n)(0 < y \wedge 0 < m \wedge m \times y = n)$; the next prime after k is at most $k! + 1$, so π is defined by a recursion using a bounded μ; and $\mathit{exf}(n, i) = (\mu x \le n)\{\pi_i^x \mid n \wedge \neg(\pi_i^{x+1} \mid n)\}$. ⊠

4 L_A expresses and Q captures every p.r. function

This section is the technical core of the First Theorem. It is where the vague "arithmetic can talk about computation" becomes a theorem. The difficulty is that L_A has nothing but $\mathsf{0}$, S, $+$ and $\times$. A recursion says "the value at Sy is computed from the value at y", and to say that in L_A we have to talk about the whole finite sequence of values $f(0), f(1), \ldots, f(y)$. So we need a way to code finite sequences by numbers using only addition and multiplication.

4.1 Gödel's β-function

Coding by prime powers (sending $k_0, \ldots, k_n$ to $2^{k_0}3^{k_1}\cdots$) needs exponentiation, which L_A doesn't have. Gödel's trick uses remainders instead. Write $rm(c, m)$ for the remainder when c is divided by m, and put

$$\beta(c, d, i) =_{\text{def}} rm(c,\ d(i + 1) + 1).$$

Lemma 4.1. If x and $y > 1$ are coprime, there is some $b < y$ with $rm(bx, y) = 1$.

Proof. Suppose $b_1 < b_2 < y$ and $rm(b_1x, y) = rm(b_2x, y)$. Then $y \mid (b_2 - b_1)x$, and since $\gcd(x, y) = 1$, Euclid's lemma gives $y \mid b_2 - b_1$, which is impossible as $0 < b_2 - b_1 < y$. So the y numbers $rm(bx, y)$ for $b < y$ are all different. There are only y possible remainders, so every remainder occurs, including 1. ⊠

Lemma 4.2 (Chinese Remainder Theorem)**.** Let $m_0, \ldots, m_n > 1$ be pairwise coprime and let $k_i < m_i$ for each i. Then there is a c with $rm(c, m_i) = k_i$ for every $i \le n$.

Proof. Let M_i be the product of all the m_j with $j \neq i$. Each such m_j is coprime to m_i, so M_i is too, and Lemma 4.1 gives b_i with $rm(b_iM_i, m_i) = 1$. Put $c = \sum_{j \le n} k_jb_jM_j$. For $j \neq i$, m_i divides M_j, so modulo m_i everything vanishes except $k_ib_iM_i \equiv k_i$. Since $k_i < m_i$, the remainder is exactly k_i. ⊠

Lemma 4.3 (Gödel's β-function lemma, Smith §15.2)**.** For any finite sequence $k_0, k_1, \ldots, k_n$ there are c, d such that $\beta(c, d, i) = k_i$ for every $i \le n$.

Proof. Let $u = \max(n + 1, k_0, \ldots, k_n)$, put $d = u!$ and $m_i = d(i + 1) + 1$ for $i \le n$. Each $m_i > 1$, and $k_i \le u \le d < m_i$.

The m_i are pairwise coprime. Suppose a prime p divides both $da + 1$ and $db + 1$ where $1 \le a < b \le n + 1$. Then p divides the difference $d(b - a)$. Since $b - a \le n < u$, $b - a$ divides $u! = d$, so in either case ($p \mid d$ or $p \mid b - a$) we get $p \mid d$. But then $p \mid da$, so p can't divide $da + 1$. Contradiction.

Now Lemma 4.2 gives c with $rm(c, m_i) = k_i$, i.e. $\beta(c, d, i) = k_i$, for all $i \le n$. ⊠

"Remainder" is definable with $+$ and $\times$ alone, so β is expressed in L_A by the Δ_0 wff

$$\mathsf{B}(\mathsf{c}, \mathsf{d}, \mathsf{i}, \mathsf{y}) =_{\text{def}} (\exists \mathsf{u} \le \mathsf{c})[\mathsf{c} = \{\mathsf{S}(\mathsf{d} \times \mathsf{Si}) \times \mathsf{u}\} + \mathsf{y} \wedge \mathsf{y} \le (\mathsf{d} \times \mathsf{Si})].$$

This says $c = (d(i + 1) + 1)u + y$ with $y < d(i + 1) + 1$, i.e. y is the remainder.

4.2 Expressing p.r. functions

Theorem 4.4 (Smith 15.1, 15.2)**.** Every p.r. function can be expressed in L_A by a Σ_1 wff. Hence (taking $\psi(\vec{\mathsf{x}}, \mathsf{0})$ where ψ expresses the characteristic function) every p.r. property and relation is expressed by a Σ_1 wff [Smith 15.3].

Proof. It is enough to show three things: the initial functions are expressible by Σ_1 wffs, and expressibility by Σ_1 wffs survives composition and primitive recursion. Then we follow the definition of any p.r. function step by step.

Initial functions. $Sx = y$ is expressed by $\mathsf{Sx} = \mathsf{y}$, $Z(x) = y$ by $\mathsf{x} = \mathsf{x} \wedge \mathsf{y} = \mathsf{0}$, and, for example, $I^3_2(x, y, z) = u$ by $\mathsf{x} = \mathsf{x} \wedge \mathsf{y} = \mathsf{u} \wedge \mathsf{z} = \mathsf{z}$. All are Δ_0.

Composition. If $\mathsf{G}(\mathsf{x}, \mathsf{y})$ expresses g and $\mathsf{H}(\mathsf{x}, \mathsf{y})$ expresses h, then $\exists \mathsf{z}(\mathsf{G}(\mathsf{x}, \mathsf{z}) \wedge \mathsf{H}(\mathsf{z}, \mathsf{y}))$ expresses $h(g(x))$, and it is Σ_1 if G and H are. Many-place cases are the same.

Primitive recursion. Suppose $f(\vec{x}, 0) = g(\vec{x})$ and $f(\vec{x}, Sy) = h(\vec{x}, y, f(\vec{x}, y))$, with g expressed by $\mathsf{G}(\vec{\mathsf{x}}, \mathsf{k})$ and h by $\mathsf{H}(\vec{\mathsf{x}}, \mathsf{u}, \mathsf{v}, \mathsf{w})$. Then $f(\vec{x}, y) = z$ iff there is a sequence $k_0, \ldots, k_y$ with $k_0 = g(\vec{x})$, $k_{u+1} = h(\vec{x}, u, k_u)$ for $u < y$, and $k_y = z$. By Lemma 4.3 this holds iff there are c, d with $\beta(c, d, 0) = g(\vec{x})$, $\beta(c, d, Su) = h(\vec{x}, u, \beta(c, d, u))$ for $u < y$, and $\beta(c, d, y) = z$. (Left to right is the lemma. Right to left: the values $\beta(c, d, 0), \ldots, \beta(c, d, y)$ satisfy the recursion, so by induction on u they are the values $f(\vec{x}, u)$.) That is rendered in L_A by

$$\exists \mathsf{c} \exists \mathsf{d}\{\exists \mathsf{k}[\mathsf{B}(\mathsf{c}, \mathsf{d}, \mathsf{0}, \mathsf{k}) \wedge \mathsf{G}(\vec{\mathsf{x}}, \mathsf{k})] \wedge (\forall \mathsf{u} \le \mathsf{y})[\mathsf{u} = \mathsf{y} \vee \exists \mathsf{v} \exists \mathsf{w}\{(\mathsf{B}(\mathsf{c}, \mathsf{d}, \mathsf{u}, \mathsf{v}) \wedge \mathsf{B}(\mathsf{c}, \mathsf{d}, \mathsf{Su}, \mathsf{w})) \wedge \mathsf{H}(\vec{\mathsf{x}}, \mathsf{u}, \mathsf{v}, \mathsf{w})\}] \wedge \mathsf{B}(\mathsf{c}, \mathsf{d}, \mathsf{y}, \mathsf{z})\}.$$

Since B is Δ_0, this is Σ_1 if G and H are. ⊠

4.3 Capturing functions

Capturing a function needs a bit more than capturing the relation $f(x) = y$: we also want the theory to know the value is unique.

Definition 4.5 (Smith §16.1)**.** T *captures* the one-place function f by $\varphi(\mathsf{x},\mathsf{y})$ if for all m, n: (i) if $f(m) = n$ then $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, and (ii) $T \vdash \exists!\mathsf{y}\,\varphi(\overline{\mathsf{m}}, \mathsf{y})$. Likewise for many-place functions. T *weakly captures* f by φ if φ captures the relation $f(x) = y$, i.e. (i) holds and (ii′) if $f(m) \neq n$ then $T \vdash \neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.

Lemma 4.6 (Smith §16.1(b))**.** If φ captures f in T and $f(m) = n$, then $T \vdash \forall\mathsf{y}(\varphi(\overline{\mathsf{m}}, \mathsf{y}) \leftrightarrow \mathsf{y} = \overline{\mathsf{n}})$.

Proof. For arbitrary a: if $\varphi(\overline{\mathsf{m}}, \mathsf{a})$, then together with $\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ the uniqueness clause gives $\mathsf{a} = \overline{\mathsf{n}}$. Conversely, $\mathsf{a} = \overline{\mathsf{n}}$ and $\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ give $\varphi(\overline{\mathsf{m}}, \mathsf{a})$ by Leibniz's Law. ⊠

Theorem 4.7 (Smith 16.1, one direction)**.** Suppose T extends Q. If $\psi(\mathsf{x},\mathsf{y})$ captures the characteristic function c_P, then $\psi(\mathsf{x}, \mathsf{0})$ captures P.

Proof. If n has P, then $c_P(n) = 0$, so $T \vdash \psi(\overline{\mathsf{n}}, \mathsf{0})$. If n lacks P, then $c_P(n) = 1$, so $T \vdash \psi(\overline{\mathsf{n}}, \overline{\mathsf{1}})$; together with uniqueness and $T \vdash \mathsf{0} \neq \overline{\mathsf{1}}$ this gives $T \vdash \neg\psi(\overline{\mathsf{n}}, \mathsf{0})$. ⊠

The next construction turns weak capturing into capturing. It looks fussy, but it is used in the key step of Theorem 4.10.

Theorem 4.8 (Smith 16.2(2), 16.3)**.** Let T extend Q, and suppose φ weakly captures f in T. Put

$$\widetilde{\varphi}(\mathsf{x},\mathsf{y}) =_{\text{def}} \varphi(\mathsf{x},\mathsf{y}) \wedge (\forall\mathsf{z} \leq \mathsf{y})(\varphi(\mathsf{x},\mathsf{z}) \rightarrow \mathsf{z} = \mathsf{y}).$$

Then $\widetilde{\varphi}$ captures f in T. Moreover, for every n,

$$T \vdash \forall\mathsf{x}\forall\mathsf{y}(\widetilde{\varphi}(\mathsf{x}, \overline{\mathsf{n}}) \wedge \widetilde{\varphi}(\mathsf{x},\mathsf{y}) \rightarrow \mathsf{y} = \overline{\mathsf{n}}).$$

Proof. Suppose $f(m) = n$. Then $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, and for each $k < n$ we have $f(m) \neq k$, so $T \vdash \neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{k}})$. Hence for each $k \leq n$, $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{k}}) \rightarrow \overline{\mathsf{k}} = \overline{\mathsf{n}}$, and (O4) gives $T \vdash (\forall\mathsf{z} \leq \overline{\mathsf{n}})(\varphi(\overline{\mathsf{m}}, \mathsf{z}) \rightarrow \mathsf{z} = \overline{\mathsf{n}})$. So $T \vdash \widetilde{\varphi}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.

The "moreover" claim: argue in T from $\widetilde{\varphi}(\mathsf{a}, \overline{\mathsf{n}}) \wedge \widetilde{\varphi}(\mathsf{a},\mathsf{b})$. By (O8), $\mathsf{b} \leq \overline{\mathsf{n}}$ or $\overline{\mathsf{n}} \leq \mathsf{b}$. If $\mathsf{b} \leq \overline{\mathsf{n}}$, the bounded clause of $\widetilde{\varphi}(\mathsf{a}, \overline{\mathsf{n}})$ gives $\varphi(\mathsf{a},\mathsf{b}) \rightarrow \mathsf{b} = \overline{\mathsf{n}}$, and $\varphi(\mathsf{a},\mathsf{b})$ is the first conjunct of $\widetilde{\varphi}(\mathsf{a},\mathsf{b})$, so $\mathsf{b} = \overline{\mathsf{n}}$. If $\overline{\mathsf{n}} \leq \mathsf{b}$, the same argument with the roles swapped gives $\overline{\mathsf{n}} = \mathsf{b}$.

Uniqueness $T \vdash \exists!\mathsf{y}\,\widetilde{\varphi}(\overline{\mathsf{m}}, \mathsf{y})$ now follows: existence from $\widetilde{\varphi}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, and uniqueness from the "moreover" claim with $\mathsf{x} := \overline{\mathsf{m}}$. ⊠

4.4 Q is p.r. adequate

Definition 4.9 (Smith §17.1)**.** T is *p.r. adequate* if every p.r. function is captured in T by some wff.

Theorem 4.10 (Smith 17.1, 17.2)**.** Q is p.r. adequate, and the capturing can always be done by a Σ_1 wff. Hence every theory extending Q captures every p.r. function, and (by Theorem 4.7) every p.r. property and relation, by Σ_1 wffs.

Since Q is sound, any wff that captures a function or relation in Q also expresses it. I'll use this repeatedly.

Proof. As before we show that the initial functions are captured by Σ_1 wffs and that this is preserved by composition and primitive recursion.

Initial functions. The same wffs as in Theorem 4.4. The uniqueness clauses are immediate from the identity rules.

Composition. Let G, H capture one-place g, h; we show $\exists \mathsf{z}(\mathsf{G}(\mathsf{x},\mathsf{z}) \wedge \mathsf{H}(\mathsf{z},\mathsf{y}))$ captures $h(g(x))$. If $h(g(m)) = n$, let $k = g(m)$; then $\mathsf{Q} \vdash \mathsf{G}(\overline{\mathsf{m}},\overline{\mathsf{k}})$ and $\mathsf{Q} \vdash \mathsf{H}(\overline{\mathsf{k}},\overline{\mathsf{n}})$, so $\mathsf{Q} \vdash \exists \mathsf{z}(\mathsf{G}(\overline{\mathsf{m}},\mathsf{z}) \wedge \mathsf{H}(\mathsf{z},\overline{\mathsf{n}}))$. For uniqueness, argue in Q from $\mathsf{G}(\overline{\mathsf{m}},\mathsf{b}) \wedge \mathsf{H}(\mathsf{b},\mathsf{a})$: uniqueness for G gives $\mathsf{b} = \overline{\mathsf{k}}$, so $\mathsf{H}(\overline{\mathsf{k}},\mathsf{a})$, and uniqueness for H gives $\mathsf{a} = \overline{\mathsf{n}}$. Many-place cases are the same with more bookkeeping.

Primitive recursion. I do the simplest nontrivial case,

$$f(0) = g, \qquad f(Sx) = h(f(x)),$$

with h captured by $\mathsf{H}(\mathsf{x},\mathsf{y})$. The general case has the same proof with extra parameters.

First, B weakly captures β: it is Δ_0 and expresses β, so by Theorem 2.8, Q proves $\mathsf{B}(\overline{\mathsf{c}},\overline{\mathsf{d}},\overline{\mathsf{i}},\overline{\mathsf{n}})$ when $\beta(c,d,i) = n$ and its negation otherwise. So by Theorem 4.8, $\widetilde{\mathsf{B}}$ captures β, and it is still Δ_0. Now take the wff from the proof of Theorem 4.4, with $\widetilde{\mathsf{B}}$ in place of B:

$$\begin{aligned}\mathsf{F}(\mathsf{x},\mathsf{y}) =_{\text{def}}\ & \exists \mathsf{c}\exists \mathsf{d}\{\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},0,\overline{\mathsf{g}}) \wedge \\ & (\forall \mathsf{u} \le \mathsf{x})[\mathsf{u} = \mathsf{x} \vee \exists \mathsf{v}\exists \mathsf{w}\{(\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\mathsf{u},\mathsf{v}) \wedge \widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\mathsf{Su},\mathsf{w})) \wedge \mathsf{H}(\mathsf{v},\mathsf{w})\}] \wedge \widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\mathsf{x},\mathsf{y})\}.\end{aligned}$$

This is Σ_1 if H is. We show it captures f.

(i) If $f(m) = n$, then $\mathsf{F}(\overline{\mathsf{m}},\overline{\mathsf{n}})$ is true: Lemma 4.3 supplies c, d coding $f(0),\dots,f(m)$, and $\widetilde{\mathsf{B}}$ and H express β and h. It is a true Σ_1 sentence, so $\mathsf{Q} \vdash \mathsf{F}(\overline{\mathsf{m}},\overline{\mathsf{n}})$ by Theorem 2.9. This is exactly where the β-function lemma gets used.

(ii) Existence in $\exists!\mathsf{y}\,\mathsf{F}(\overline{\mathsf{m}},\mathsf{y})$ follows from (i). For uniqueness, write $k_j = f(j)$ for $j \le m$ (so $k_0 = g$, $k_m = n$), and argue in Q. Assume $\mathsf{F}(\overline{\mathsf{m}},\mathsf{e})$ and fix witnesses c,d for its two existential quantifiers; call the resulting conjunction (D). I claim that for each $j \le m$, (D) implies $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{j}},\overline{\mathsf{k_j}})$. This is proved by an induction on j carried out *outside* Q: for each particular j we write down a Q-derivation.

- $j = 0$: the first conjunct of (D) is $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},0,\overline{\mathsf{g}})$.
- $j \to j+1$ (where $j < m$): instantiate the bounded universal in (D) at $\mathsf{u} = \overline{\mathsf{j}}$ (legitimate since $\mathsf{Q} \vdash \overline{\mathsf{j}} \le \overline{\mathsf{m}}$). Since $\mathsf{Q} \vdash \overline{\mathsf{j}} \ne \overline{\mathsf{m}}$, the first disjunct is ruled out, so there are a,b with $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{j}},\mathsf{a})$, $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\mathsf{S}\overline{\mathsf{j}},\mathsf{b})$ and $\mathsf{H}(\mathsf{a},\mathsf{b})$. The "moreover" clause of Theorem 4.8, together with the induction hypothesis $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{j}},\overline{\mathsf{k_j}})$, gives $\mathsf{a} = \overline{\mathsf{k_j}}$, so $\mathsf{H}(\overline{\mathsf{k_j}},\mathsf{b})$. Since $h(k_j) = k_{j+1}$ and H captures h, we have $\mathsf{H}(\overline{\mathsf{k_j}},\overline{\mathsf{k_{j+1}}})$, and uniqueness for H gives $\mathsf{b} = \overline{\mathsf{k_{j+1}}}$. As $\mathsf{S}\overline{\mathsf{j}}$ is the numeral $\overline{\mathsf{j+1}}$, we get $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{j+1}},\overline{\mathsf{k_{j+1}}})$, and $\exists$-elimination discharges a,b.

At $j = m$ we have $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{m}},\overline{\mathsf{n}})$. The last conjunct of (D) is $\widetilde{\mathsf{B}}(\mathsf{c},\mathsf{d},\overline{\mathsf{m}},\mathsf{e})$, so Theorem 4.8 gives $\mathsf{e} = \overline{\mathsf{n}}$. Discharging the witnesses, $\mathsf{F}(\overline{\mathsf{m}},\mathsf{e}) \to \mathsf{e} = \overline{\mathsf{n}}$, and generalizing on e finishes uniqueness. ⊠

The wffs built by following a p.r. definition step by step in this way are what Smith calls *canonical*. When I later pick a wff to capture a p.r. relation like Prf_T, I mean a canonical one: it is Σ_1, it captures, and it expresses. In Section 8 it will matter that it was built this way and isn't just any wff that happens to capture the relation.

5 Arithmetization of syntax

5.1 Gödel numbering

Definition 5.1 (Smith §§19.1, 19.3)**.** Give each symbol of $\mathcal{L}_A$ a basic code:

$\neg$	$\wedge$	$\vee$	$\to$	$\leftrightarrow$	$\forall$	$\exists$	$=$	$($	$)$	$\mathsf{0}$	S	$+$	$\times$	x	y	z	...
1	3	5	7	9	11	13	15	17	19	21	23	25	27	2	4	6	...

(the k-th variable gets $2(k+1)$). An expression $s_0 s_1 \dots s_k$ with basic codes $c_0, \dots, c_k$ gets the *Gödel number* (g.n.) $\pi_0^{c_0}\pi_1^{c_1}\cdots\pi_k^{c_k}$, where π_i is the $(i+1)$-th prime. A sequence of expressions $e_0, \dots, e_n$ with g.n.s $g_0, \dots, g_n$ gets the *super g.n.* $\pi_0^{g_0}\pi_1^{g_1}\cdots\pi_n^{g_n}$.

For example, $\mathsf{SS0}$ has g.n. $2^{23}\cdot 3^{23}\cdot 5^{21}$. Decoding is done by *exf*: if n is the g.n. of $s_0 \dots s_k$, then $exf(n,i)$ is the basic code of s_i. For other languages, extend the table of basic codes. Any other "sensible" coding, one that can be translated to and from this one by p.r. functions, would serve equally well [Smith 19.2].

Notation [Smith 19.5]. If φ is an expression, $\ulcorner\varphi\urcorner$ denotes its g.n.; inside a formal wff, $\ulcorner\varphi\urcorner$ abbreviates the standard numeral for that g.n. So $\ulcorner\mathsf{SS0}\urcorner$ is a number outside the formal language, and a very long numeral inside it.

5.2 Syntax is primitive recursive

Definition 5.2 (Smith §19.4)**.** *Term*(n), *Atom*(n), *Wff*(n), *Sent*(n) hold when n is the g.n. of a term, an atomic wff, a wff, a sentence of L_A. For a theory T, $Prf_T(m,n)$ holds when m is the super g.n. of a T-proof of the sentence with g.n. n. (*Prf* without subscript is Prf_{PA}.)

Definition 5.3 (Smith §§22.1, 22.3)**.** T is *p.r. axiomatized* if the properties of being (the g.n. of) a T-wff, a T-sentence, an axiom of T, and (the super g.n. of) a correctly built T-proof are all p.r. A theory is *nice* if it is consistent, p.r. axiomatized, and extends Q.

Theorem 5.4 (Smith 19.1, 20.7)**.** *Term*, *Atom*, *Wff*, *Sent* and *Prf* are p.r. More generally, Prf_T is p.r. for any p.r. axiomatized T.

Proof sketch. To test *Term*(n): decode n with *exf* and *len*, then check whether the string is $\mathsf{0}$, a variable, or built from $\mathsf{0}$ and variables by S, $+$, $\times$. A clean way to do this is to ask whether there is a *term-sequence* ending in it, i.e. a list of expressions each of which is $\mathsf{0}$, a variable, or S, $+$ or $\times$ applied to earlier entries. The length of such a list and the size of its code can be bounded by a p.r. function of n, so the search is bounded. *Atom*, *Wff* and *Sent* are similar.

To test $Prf_T(m,n)$: decode m into a sequence of expressions; check each is a wff; check each is an axiom or follows from earlier lines by a rule (for modus ponens this is a bounded search for two earlier lines of the right shape); check the last line has g.n. n and *Sent*(n). Every step is a bounded loop, so the whole check is p.r.

Smith's Chapter 20 turns this into explicit p.r. definitions. I haven't gone through that, and I don't think it changes how believable the claim is: once one believes that checking a proof is a bounded mechanical task, this is a programming exercise. ⊠

5.3 Diagonalization

Definition 5.5 (Smith §19.6)**.** The *diagonalization* of a wff φ is $\exists \mathsf{y}(\mathsf{y} = \ulcorner \varphi \urcorner \wedge \varphi)$.

If $\varphi = \varphi(\mathsf{y})$ has just y free, its diagonalization is logically equivalent to $\varphi(\ulcorner \varphi \urcorner)$: the wff applied to its own code.

Theorem 5.6 (Smith 19.2, 20.2)**.** The function *diag* taking the g.n. of a wff to the g.n. of its diagonalization is p.r.

Proof sketch. Let $m \star n$ be the g.n. of the expression coded by m followed by the one coded by n; this concatenation function is p.r. [Smith 20.1] since it only shifts prime exponents along. The function $num(n) = \ulcorner \overline{\mathsf{n}} \urcorner$ is p.r. by $num(0) = \ulcorner \mathsf{0} \urcorner$ and $num(Sn) = \ulcorner \mathsf{S} \urcorner \star num(n)$. Then

$$diag(n) = \ulcorner \exists \mathsf{y}(\mathsf{y} = \ulcorner \urcorner \star num(n) \star \ulcorner \wedge \urcorner \star n \star \ulcorner) \urcorner. \qquad \boxtimes$$

For a theory T whose language includes L_A, $diag_T$ is defined the same way on T's wffs and is p.r. by the same argument.

6 The First Incompleteness Theorem

6.1 The Gödel sentence

Fix a nice theory T. By Theorems 4.10, 5.4 and 5.6:

- Prf_T is p.r., so it is canonically captured in T by a Σ_1 wff $\mathsf{Prf}_T(\mathsf{x}, \mathsf{y})$, which also expresses it;
- $diag_T$ is p.r., so it is captured in T by a Σ_1 wff $\mathsf{Diag}_T(\mathsf{x}, \mathsf{y})$, which also expresses it.

Define the numerical relation $Gdl_T(m, n)$ to hold iff $Prf_T(m, diag_T(n))$: m codes a T-proof of the diagonalization of the wff with g.n. n. Following Smith §24.3, put

$$\mathsf{Gdl}_T(\mathsf{x}, \mathsf{y}) =_{\text{def}} \exists \mathsf{z}(\mathsf{Prf}_T(\mathsf{x}, \mathsf{z}) \wedge \mathsf{Diag}_T(\mathsf{y}, \mathsf{z})).$$

Lemma 6.1. Gdl_T is Σ_1, and it both expresses and captures Gdl_T in T.

Proof. It is Σ_1 because Prf_T and Diag_T are. It expresses Gdl_T: $\mathsf{Gdl}_T(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is true iff there is k with $Prf_T(m, k)$ and $diag_T(n) = k$, i.e. iff $Prf_T(m, diag_T(n))$.

For capturing, let $k = diag_T(n)$. If $Gdl_T(m, n)$, then $T \vdash \mathsf{Prf}_T(\overline{\mathsf{m}}, \overline{\mathsf{k}})$ and $T \vdash \mathsf{Diag}_T(\overline{\mathsf{n}}, \overline{\mathsf{k}})$, so $T \vdash \mathsf{Gdl}_T(\overline{\mathsf{m}}, \overline{\mathsf{n}})$. If not, then $T \vdash \neg \mathsf{Prf}_T(\overline{\mathsf{m}}, \overline{\mathsf{k}})$, and Lemma 4.6 gives $T \vdash \forall \mathsf{z}(\mathsf{Diag}_T(\overline{\mathsf{n}}, \mathsf{z}) \to \mathsf{z} = \overline{\mathsf{k}})$. So any z with $\mathsf{Diag}_T(\overline{\mathsf{n}}, \mathsf{z})$ is $\overline{\mathsf{k}}$ and fails $\mathsf{Prf}_T(\overline{\mathsf{m}}, \mathsf{z})$, i.e. $T \vdash \neg \mathsf{Gdl}_T(\overline{\mathsf{m}}, \overline{\mathsf{n}})$. $\boxtimes$

Now define

$$\mathsf{U}(\mathsf{y}) =_{\text{def}} \forall \mathsf{x} \neg \mathsf{Gdl}_T(\mathsf{x}, \mathsf{y}), \qquad \mathsf{G}_T =_{\text{def}} \exists \mathsf{y}(\mathsf{y} = \ulcorner \mathsf{U} \urcorner \wedge \mathsf{U}(\mathsf{y})).$$

So G_T is the diagonalization of U, which means $diag_T(\ulcorner \mathsf{U} \urcorner) = \ulcorner \mathsf{G}_T \urcorner$, and G_T is logically equivalent to $\mathsf{U}(\ulcorner \mathsf{U} \urcorner)$, i.e. to $\forall \mathsf{x} \neg \mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$. Informally, $\mathsf{U}(\mathsf{y})$ says "the diagonalization of the wff with code y has no T-proof", and G_T applies this to U itself.

Theorem 6.2 (Smith 21.1, generalized)**.** G_T is true if and only if $T \nvdash \mathsf{G}_T$.

Proof. G_T is true iff $\forall \mathsf{x} \neg \mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$ is true iff (since Gdl_T expresses Gdl_T) there is no m with $Gdl_T(m, \ulcorner \mathsf{U} \urcorner)$ iff there is no m with $Prf_T(m, diag_T(\ulcorner \mathsf{U} \urcorner))$, i.e. with $Prf_T(m, \ulcorner \mathsf{G}_T \urcorner)$, iff $T \nvdash \mathsf{G}_T$. $\boxtimes$

Theorem 6.3 (Smith 21.4)**.** G_T is Π_1-equivalent.

Proof. $\mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$ is Σ_1, so its negation is Π_1-equivalent, and putting $\forall \mathsf{x}$ in front keeps it Π_1. G_T is logically equivalent to the result. $\boxtimes$

6.2 The semantic argument

Theorem 6.4 (Smith 22.1)**.** If T is nice and sound, then $T \nvdash \mathsf{G}_T$ and $T \nvdash \neg\mathsf{G}_T$.

Proof. If $T \vdash \mathsf{G}_T$, then G_T is true by soundness, so $T \nvdash \mathsf{G}_T$ by Theorem 6.2; contradiction. So $T \nvdash \mathsf{G}_T$, hence G_T is true, hence $\neg\mathsf{G}_T$ is false, so it isn't a theorem of the sound T. ⊠

This uses only that Gdl_T *expresses* Gdl_T, which needs only Theorem 4.4, not the harder Theorem 4.10. Gödel cared about avoiding the assumption of soundness, which is a semantic assumption and arguably as doubtful as anything we might want to prove. The next argument replaces it with purely syntactic conditions, and that is where capturing gets used.

6.3 The syntactic argument

Theorem 6.5 (Smith 21.5, generalized)**.** If T is nice (in particular consistent), then $T \nvdash \mathsf{G}_T$.

Proof. Suppose $T \vdash \mathsf{G}_T$, and let m be the super g.n. of a proof. Then $Prf_T(m, \ulcorner \mathsf{G}_T \urcorner)$, which is to say $Gdl_T(m, \ulcorner \mathsf{U} \urcorner)$. Since Gdl_T captures Gdl_T, it follows that $T \vdash \mathsf{Gdl}_T(\overline{\mathsf{m}}, \ulcorner \mathsf{U} \urcorner)$. But $T \vdash \mathsf{G}_T$ gives $T \vdash \forall \mathsf{x} \neg \mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$, and instantiating at $\overline{\mathsf{m}}$ gives $T \vdash \neg \mathsf{Gdl}_T(\overline{\mathsf{m}}, \ulcorner \mathsf{U} \urcorner)$. So T is inconsistent. ⊠

For the other half we need a stronger condition than consistency.

Definition 6.6 (Smith §21.6)**.** T is *ω-incomplete* if for some $\varphi(\mathsf{x})$, T proves every $\varphi(\overline{\mathsf{m}})$ but not $\forall \mathsf{x}\, \varphi(\mathsf{x})$. T is *ω-inconsistent* if for some $\varphi(\mathsf{x})$, T proves every $\varphi(\overline{\mathsf{m}})$ and also proves $\neg \forall \mathsf{x}\, \varphi(\mathsf{x})$; equivalently, if for some $\varphi'(\mathsf{x})$, $T \vdash \exists \mathsf{x}\, \varphi'(\mathsf{x})$ while $T \vdash \neg\varphi'(\overline{\mathsf{m}})$ for every m. T is *ω-consistent* if it is not ω-inconsistent.

An ω-consistent theory is consistent, because an inconsistent theory proves everything. A sound theory is ω-consistent [Smith 21.5]: if every $\varphi(\overline{\mathsf{m}})$ is true then $\forall \mathsf{x} \varphi(\mathsf{x})$ is true, so its negation is not a theorem. Q is ω-incomplete by Theorem 2.3, and Theorem 6.5 together with the proof of the next theorem shows that every nice theory is: it proves each $\neg \mathsf{Gdl}_T(\overline{\mathsf{m}}, \ulcorner \mathsf{U} \urcorner)$ but not their generalization G_T.

Theorem 6.7 (Smith 21.7, generalized)**.** If T is nice and ω-consistent, then $T \nvdash \neg\mathsf{G}_T$.

Proof. Suppose $T \vdash \neg\mathsf{G}_T$, which amounts to $T \vdash \exists \mathsf{x}\, \mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$. T is consistent, so $T \nvdash \mathsf{G}_T$, so no m codes a proof of G_T, i.e. for every m, not $Gdl_T(m, \ulcorner \mathsf{U} \urcorner)$. Since Gdl_T captures Gdl_T, $T \vdash \neg \mathsf{Gdl}_T(\overline{\mathsf{m}}, \ulcorner \mathsf{U} \urcorner)$ for every m. Together with $T \vdash \exists \mathsf{x}\, \mathsf{Gdl}_T(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$ this makes T ω-inconsistent. ⊠

Theorem 6.8 (Smith 22.2: Gödel's First Incompleteness Theorem)**.** If T is nice, there is a Π_1 sentence φ of L_A (namely G_T, up to logical equivalence) such that $T \nvdash \varphi$, and if T is also ω-consistent, $T \nvdash \neg\varphi$. Moreover φ is true.

Proof. Theorems 6.5, 6.7 and 6.3. It is true by Theorem 6.2, since it is unprovable. ⊠

Remark. The undecidable sentence is not some strange new kind of statement. It is an arithmetical Π_1 sentence, of the same logical form as Goldbach's conjecture, though astronomically long if written out. And T is *incompletable*. If T is sound, then $T + \mathsf{G}_T$ is sound too (as G_T is true), and it is still nice, so it has its own undecidable sentence $\mathsf{G}_{T+\mathsf{G}_T}$, and so on. Note also that ω-consistency is genuinely stronger than consistency: $\mathsf{PA} + \neg\mathsf{G}$ is consistent (if PA is) by Theorem 6.5, but it proves $\exists \mathsf{x}\, \mathsf{Gdl}(\mathsf{x}, \ulcorner \mathsf{U} \urcorner)$ while refuting each instance, so it is ω-inconsistent [Smith 21.6].

6.4 Provability predicates and the Diagonalization Lemma

Definition 6.9 (Smith §24.1)**.** $\mathsf{Prov}_T(\mathsf{x}) =_{\text{def}} \exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v},\mathsf{x})$. It is Σ_1 and expresses $Prov_T(n)$, the property of being the g.n. of a T-theorem.

Theorem 6.10 (Smith 24.1)**.** Let T be nice. For any sentence φ:

(C1) if $T \vdash \varphi$, then $T \vdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner)$;

(Cω) if T is ω-consistent and $T \vdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner)$, then $T \vdash \varphi$.

Proof. (C1) If m codes a proof of φ, then $T \vdash \mathsf{Prf}_T(\overline{\mathsf{m}}, \ulcorner\varphi\urcorner)$, so $T \vdash \exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v}, \ulcorner\varphi\urcorner)$. (Alternatively: $\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$ is then a true Σ_1 sentence, so even Q proves it.) (Cω) If $T \nvdash \varphi$, then for every m, $T \vdash \neg\mathsf{Prf}_T(\overline{\mathsf{m}}, \ulcorner\varphi\urcorner)$; together with $T \vdash \exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v}, \ulcorner\varphi\urcorner)$ this makes T ω-inconsistent. ⊠

Theorem 6.11 (Smith 24.3)**.** If T is nice, then $T \vdash \mathsf{G}_T \leftrightarrow \neg\mathsf{Prov}_T(\ulcorner\mathsf{G}_T\urcorner)$.

Proof. Pure logic turns $\mathsf{U}(\mathsf{y}) = \forall\mathsf{x}\neg\exists\mathsf{z}(\mathsf{Prf}_T(\mathsf{x},\mathsf{z}) \wedge \mathsf{Diag}_T(\mathsf{y},\mathsf{z}))$ into the equivalent $\forall\mathsf{z}(\mathsf{Diag}_T(\mathsf{y},\mathsf{z}) \rightarrow \neg\exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v},\mathsf{z}))$, i.e. $\forall\mathsf{z}(\mathsf{Diag}_T(\mathsf{y},\mathsf{z}) \rightarrow \neg\mathsf{Prov}_T(\mathsf{z}))$. So

$$\vdash \mathsf{G}_T \leftrightarrow \forall\mathsf{z}(\mathsf{Diag}_T(\ulcorner\mathsf{U}\urcorner,\mathsf{z}) \rightarrow \neg\mathsf{Prov}_T(\mathsf{z})).$$

Since $diag_T(\ulcorner\mathsf{U}\urcorner) = \ulcorner\mathsf{G}_T\urcorner$, Lemma 4.6 gives $T \vdash \forall\mathsf{z}(\mathsf{Diag}_T(\ulcorner\mathsf{U}\urcorner,\mathsf{z}) \leftrightarrow \mathsf{z} = \ulcorner\mathsf{G}_T\urcorner)$. Substituting, $T \vdash \mathsf{G}_T \leftrightarrow \forall\mathsf{z}(\mathsf{z} = \ulcorner\mathsf{G}_T\urcorner \rightarrow \neg\mathsf{Prov}_T(\mathsf{z}))$, and the right side is equivalent to $\neg\mathsf{Prov}_T(\ulcorner\mathsf{G}_T\urcorner)$. ⊠

So not only is G_T true iff unprovable; T itself proves this. Nothing in the proof used anything special about $\neg\mathsf{Prov}_T$, which gives the general result.

Theorem 6.12 (Smith 24.4: the Diagonalization Lemma)**.** If T is nice and $\varphi(\mathsf{x})$ is any wff of its language with one free variable, there is a sentence γ with $T \vdash \gamma \leftrightarrow \varphi(\ulcorner\gamma\urcorner)$.

Proof. Put $\psi(\mathsf{y}) =_{\text{def}} \forall\mathsf{z}(\mathsf{Diag}_T(\mathsf{y},\mathsf{z}) \rightarrow \varphi(\mathsf{z}))$ and let γ be the diagonalization of ψ. By logic, $T \vdash \gamma \leftrightarrow \forall\mathsf{z}(\mathsf{Diag}_T(\ulcorner\psi\urcorner,\mathsf{z}) \rightarrow \varphi(\mathsf{z}))$. Since $diag_T(\ulcorner\psi\urcorner) = \ulcorner\gamma\urcorner$, Lemma 4.6 gives $T \vdash \forall\mathsf{z}(\mathsf{Diag}_T(\ulcorner\psi\urcorner,\mathsf{z}) \leftrightarrow \mathsf{z} = \ulcorner\gamma\urcorner)$. Hence $T \vdash \gamma \leftrightarrow \forall\mathsf{z}(\mathsf{z} = \ulcorner\gamma\urcorner \rightarrow \varphi(\mathsf{z}))$, i.e. $T \vdash \gamma \leftrightarrow \varphi(\ulcorner\gamma\urcorner)$. ⊠

Such a γ is called a *fixed point* of φ. If φ is Π_1-equivalent, so is γ, because Diag_T is Σ_1 [Smith 24.5]. The argument for the First Theorem now runs for any fixed point of $\neg\mathsf{Prov}_T$.

Theorem 6.13 (Smith 24.6)**.** Let T be nice and $T \vdash \gamma \leftrightarrow \neg\mathsf{Prov}_T(\ulcorner\gamma\urcorner)$. Then $T \nvdash \gamma$, and if T is ω-consistent, $T \nvdash \neg\gamma$.

Proof. If $T \vdash \gamma$, then $T \vdash \neg\mathsf{Prov}_T(\ulcorner\gamma\urcorner)$ by the biconditional and $T \vdash \mathsf{Prov}_T(\ulcorner\gamma\urcorner)$ by (C1). If $T \vdash \neg\gamma$, then $T \vdash \mathsf{Prov}_T(\ulcorner\gamma\urcorner)$ by the biconditional, so $T \vdash \gamma$ by (Cω), and T is inconsistent. ⊠

Here is an application that explains an asymmetry we'll meet again.

Theorem 6.14 (Smith 24.8)**.** If T is nice, no wff captures $Prov_T$ in T. In particular $Prov_T$ is not p.r.

Proof. Suppose $\mathsf{Pr}(\mathsf{x})$ captured $Prov_T$. By the Diagonalization Lemma there is γ with $T \vdash \gamma \leftrightarrow \neg\mathsf{Pr}(\ulcorner\gamma\urcorner)$. If $T \nvdash \gamma$, capturing gives $T \vdash \neg\mathsf{Pr}(\ulcorner\gamma\urcorner)$, so $T \vdash \gamma$. So $T \vdash \gamma$. Then capturing gives $T \vdash \mathsf{Pr}(\ulcorner\gamma\urcorner)$ and the biconditional gives $T \vdash \neg\mathsf{Pr}(\ulcorner\gamma\urcorner)$, so T is inconsistent. The last claim follows because T captures every p.r. property. ⊠

So Prov_T expresses provability but doesn't capture it. By (C1) the theory always proves $\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$ when φ is a theorem, but it can't always prove $\neg\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$ when φ isn't. Theorem 7.6 will show that, under mild conditions, it never can.

6.5 Rosser's improvement

The assumption of ω-consistency in Theorem 6.8 is a bit unsatisfying. Rosser removed it by changing the sentence. Where G_T says "I am not provable", Rosser's sentence says "if I am provable, then my negation already has a smaller proof".

Definition 6.15 (Smith §25.3)**.** $\overline{Prf}_T(m,n)$ holds when m codes a T-proof of the negation of the sentence with g.n. n. It is p.r. (since $n \mapsto \ulcorner\neg\urcorner \star n$ is), so it is captured by a Σ_1 wff $\overline{\mathsf{Prf}}_T(\mathsf{x},\mathsf{y})$. Put

$$\mathsf{RProv}_T(\mathsf{x}) =_{\text{def}} \exists \mathsf{v}(\mathsf{Prf}_T(\mathsf{v},\mathsf{x}) \wedge (\forall \mathsf{w} \leq \mathsf{v})\neg\overline{\mathsf{Prf}}_T(\mathsf{w},\mathsf{x})).$$

By the Diagonalization Lemma there is a sentence R_T with $T \vdash \mathsf{R}_T \leftrightarrow \neg\mathsf{RProv}_T(\ulcorner \mathsf{R}_T \urcorner)$.

Theorem 6.16 (Smith 25.4, 25.5: Gödel–Rosser)**.** Let T be nice and let γ be any fixed point of $\neg\mathsf{RProv}_T$. Then $T \nvdash \gamma$ and $T \nvdash \neg\gamma$. So every nice theory is negation-incomplete.

Proof. $T \nvdash \gamma$. Suppose $T \vdash \gamma$. Then (i) $T \vdash \neg\mathsf{RProv}_T(\ulcorner\gamma\urcorner)$. Let m code a proof of γ, so $T \vdash \mathsf{Prf}_T(\overline{\mathsf{m}},\ulcorner\gamma\urcorner)$. Since T is consistent, $\neg\gamma$ has no proof, so for every n, $T \vdash \neg\overline{\mathsf{Prf}}_T(\overline{\mathsf{n}},\ulcorner\gamma\urcorner)$; by (O4), $T \vdash (\forall \mathsf{w} \leq \overline{\mathsf{m}})\neg\overline{\mathsf{Prf}}_T(\mathsf{w},\ulcorner\gamma\urcorner)$. Combining and existentially generalizing, (ii) $T \vdash \mathsf{RProv}_T(\ulcorner\gamma\urcorner)$. (i) and (ii) contradict each other.

$T \nvdash \neg\gamma$. Suppose $T \vdash \neg\gamma$. Then (iii) $T \vdash \mathsf{RProv}_T(\ulcorner\gamma\urcorner)$. Let m code a proof of $\neg\gamma$, so $T \vdash \overline{\mathsf{Prf}}_T(\overline{\mathsf{m}},\ulcorner\gamma\urcorner)$. By consistency γ has no proof, so $T \vdash \neg\mathsf{Prf}_T(\overline{\mathsf{n}},\ulcorner\gamma\urcorner)$ for every n, and by (O4) $T \vdash (\forall \mathsf{v} \leq \overline{\mathsf{m}})\neg\mathsf{Prf}_T(\mathsf{v},\ulcorner\gamma\urcorner)$, i.e. $T \vdash \forall \mathsf{v}(\mathsf{Prf}_T(\mathsf{v},\ulcorner\gamma\urcorner) \rightarrow \neg\, \mathsf{v} \leq \overline{\mathsf{m}})$. By (O8) that gives $T \vdash \forall \mathsf{v}(\mathsf{Prf}_T(\mathsf{v},\ulcorner\gamma\urcorner) \rightarrow \overline{\mathsf{m}} \leq \mathsf{v})$. With $\overline{\mathsf{Prf}}_T(\overline{\mathsf{m}},\ulcorner\gamma\urcorner)$ this yields $T \vdash \forall \mathsf{v}(\mathsf{Prf}_T(\mathsf{v},\ulcorner\gamma\urcorner) \rightarrow (\exists \mathsf{w} \leq \mathsf{v})\overline{\mathsf{Prf}}_T(\mathsf{w},\ulcorner\gamma\urcorner))$, which is (iv) $T \vdash \neg\mathsf{RProv}_T(\ulcorner\gamma\urcorner)$. (iii) and (iv) contradict each other. ⊠

The trick is that both halves only need facts about proofs below a fixed bound m, and Q can handle bounded facts thanks to (O4) and (O8). The price is that $\neg\mathsf{RProv}_T$ has an unbounded universal quantifier buried inside a negated $\overline{\mathsf{Prf}}_T$, so R_T isn't obviously Π_1. With a bit of induction ($\mathsf{I}\Sigma_1$) one can choose things so that it is [Smith 25.6].

7 The Second Incompleteness Theorem

7.1 The consistency sentence

Let $\bot =_{\text{def}} \mathsf{0} = \overline{\mathsf{1}}$. Since $\mathsf{Q} \vdash \mathsf{0} \neq \overline{\mathsf{1}}$, a theory extending Q is consistent iff it doesn't prove $\bot$.

Definition 7.1 (Smith §§31.1, 33.1)**.** $\mathsf{Con}_T =_{\text{def}} \neg\mathsf{Prov}_T(\ulcorner\bot\urcorner)$. We abbreviate $\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$ as $\Box_T\varphi$, or just $\Box\varphi$ when T is clear, so Con_T is $\neg\Box\bot$.

Con_T is Π_1-equivalent and is true iff T is consistent. One warning about the box: $\Box\varphi$ looks like a wff built from φ, but it isn't. $\ulcorner\varphi\urcorner$ is just a numeral, so however complicated φ is, $\Box\varphi$ is the same Σ_1 wff Prov_T applied to some numeral.

One half of "Con_T is undecidable" is easy, and needs nothing new.

Theorem 7.2 (Smith 31.2, generalized)**.** If T is nice and ω-consistent, then $T \nvdash \neg\mathsf{Con}_T$.

Proof. T is consistent, so no n codes a proof of $\bot$, so $T \vdash \neg\mathsf{Prf}_T(\overline{\mathsf{n}},\ulcorner\bot\urcorner)$ for every n. If also $T \vdash \neg\mathsf{Con}_T$, i.e. $T \vdash \exists \mathsf{v}\, \mathsf{Prf}_T(\mathsf{v},\ulcorner\bot\urcorner)$, T would be ω-inconsistent. ⊠

The other half, $T \nvdash \mathsf{Con}_T$, is the Second Theorem. Note it is another case of ω-incompleteness: T proves $\neg\mathsf{Prf}_T(\overline{\mathsf{n}},\ulcorner\bot\urcorner)$ for each n but can't prove $\forall \mathsf{v}\neg\mathsf{Prf}_T(\mathsf{v},\ulcorner\bot\urcorner)$.

7.2 The idea

Theorem 6.5 says: if T is consistent then $T \nvdash \mathsf{G}_T$. Inside T's language that is the sentence $\mathsf{Con}_T \to \neg\Box\mathsf{G}_T$. If T could *prove* it (the "Formalized First Theorem"), we'd be done quickly: if $T \vdash \mathsf{Con}_T$, then $T \vdash \neg\Box\mathsf{G}_T$, and by Theorem 6.11 that gives $T \vdash \mathsf{G}_T$, contradicting Theorem 6.5.

So the job is to show that T can carry out the argument of Theorem 6.5 about itself. That argument used three things: the fixed point property (D1) below, the fact (C1) that proofs give provability, and some logic. Hilbert, Bernays and Löb isolated conditions on $\Box$ which say that T "knows" these facts.

Definition 7.3 (Smith §33.2: the HBL derivability conditions)**.** For all sentences φ, ψ:

(C1) if $T \vdash \varphi$, then $T \vdash \Box\varphi$;

(C2) $T \vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$;

(C3) $T \vdash \Box\varphi \to \Box\Box\varphi$.

(C1) holds for every nice T (Theorem 6.10). (C2) says T knows its provability is closed under modus ponens. (C3) says T knows (C1): that whatever is provable is provably provable. In this section I assume the conditions hold and derive the Second Theorem; Section 8 discusses why they hold. From Theorem 6.11 we have

(D1) $T \vdash \mathsf{G} \to \neg\Box\mathsf{G}$,

(D2) $T \vdash \neg\Box\mathsf{G} \to \mathsf{G}$,

dropping the subscript T from G_T.

7.3 The proof

Theorem 7.4 (Smith 33.2: the Formalized First Theorem)**.** If T is nice and the derivability conditions hold, then $T \vdash \mathsf{Con}_T \to \neg\Box\mathsf{G}$.

Proof. For any φ, logic gives $T \vdash \neg\varphi \to (\varphi \to \bot)$, so by (C1) and (C2),

$$\text{(A)} \qquad T \vdash \Box\neg\varphi \to \Box(\varphi \to \bot).$$

Now:

1.	$T \vdash \mathsf{G} \to \neg\Box\mathsf{G}$	(D1)
2.	$T \vdash \Box(\mathsf{G} \to \neg\Box\mathsf{G})$	from 1 by (C1)
3.	$T \vdash \Box\mathsf{G} \to \Box\neg\Box\mathsf{G}$	from 2 by (C2)
4.	$T \vdash \Box\neg\Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	instance of (A)
5.	$T \vdash \Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	from 3, 4
6.	$T \vdash \Box\mathsf{G} \to (\Box\Box\mathsf{G} \to \Box\bot)$	from 5 by (C2)
7.	$T \vdash \Box\mathsf{G} \to \Box\Box\mathsf{G}$	(C3)
8.	$T \vdash \Box\mathsf{G} \to \Box\bot$	from 6, 7
9.	$T \vdash \neg\Box\bot \to \neg\Box\mathsf{G}$	contraposing 8

and line 9 is $T \vdash \mathsf{Con}_T \to \neg\Box\mathsf{G}$. ☒

Line 7 is the only place (C3) is used, and it is what lets T go from "G is provable" to "it is provable that G is provable", mirroring the step in Theorem 6.5 where we went from a proof of G to the theorem $\mathsf{Gdl}_T(\overline{\mathsf{m}}, \ulcorner\mathsf{U}\urcorner)$.

Theorem 7.5 (Smith 33.3: Gödel's Second Incompleteness Theorem)**.** If T is nice and the derivability conditions hold, then $T \nvdash \mathsf{Con}_T$.

Proof. By Theorem 7.4 and (D2), $T \vdash \mathsf{Con}_T \to \mathsf{G}$. If $T \vdash \mathsf{Con}_T$ then $T \vdash \mathsf{G}$, contradicting Theorem 6.5. ⊠

Two more consequences, both short.

Theorem 7.6 (Smith 33.4, 33.5)**.** If T is nice and the derivability conditions hold, then for every sentence φ, $T \vdash \neg\Box\varphi \to \mathsf{Con}_T$. Hence there is no sentence φ with $T \vdash \neg\Box\varphi$.

Proof. $T \vdash \bot \to \varphi$, so by (C1) and (C2) $T \vdash \Box\bot \to \Box\varphi$; contrapose. If $T \vdash \neg\Box\varphi$, then $T \vdash \mathsf{Con}_T$, contradicting Theorem 7.5. ⊠

So T proves $\Box\varphi$ for all its theorems but never proves $\neg\Box\varphi$ for any sentence at all, even the ones it can't prove. To prove that something is unprovable, it would have to prove its own consistency.

Theorem 7.7 (Smith 34.4: Löb's Theorem)**.** If T is nice and the derivability conditions hold, then $T \vdash \Box\varphi \to \varphi$ implies $T \vdash \varphi$.

Proof. Assume $T \vdash \Box\varphi \to \varphi$. Apply the Diagonalization Lemma to $\mathsf{Prov}_T(\mathsf{x}) \to \varphi$ to get γ with $T \vdash \gamma \leftrightarrow (\Box\gamma \to \varphi)$. Then

1.	$T \vdash \gamma \to (\Box\gamma \to \varphi)$	from the fixed point
2.	$T \vdash \Box\gamma \to \Box(\Box\gamma \to \varphi)$	from 1 by (C1), (C2)
3.	$T \vdash \Box\gamma \to (\Box\Box\gamma \to \Box\varphi)$	from 2 by (C2)
4.	$T \vdash \Box\gamma \to \Box\Box\gamma$	(C3)
5.	$T \vdash \Box\gamma \to \Box\varphi$	from 3, 4
6.	$T \vdash \Box\gamma \to \varphi$	from 5 and the assumption
7.	$T \vdash \gamma$	from 6 and the fixed point
8.	$T \vdash \Box\gamma$	from 7 by (C1)
9.	$T \vdash \varphi$	from 6, 8.

⊠

Löb's Theorem gives the Second Theorem again: if $T \vdash \mathsf{Con}_T$, i.e. $T \vdash \neg\Box\bot$, then $T \vdash \Box\bot \to \bot$, so $T \vdash \bot$ by Löb. It also answers Henkin's question about the sentence H that "says" it is provable ($T \vdash \mathsf{H} \leftrightarrow \Box\mathsf{H}$): such an H is always a theorem.

8 Why the derivability conditions hold

This is the part of the Second Theorem I take on trust. (C1) is already proved. Here is what (C2) and (C3) need, following Smith's Chapter 35.

(C2). Outside the theory this is easy. If m codes a proof of $\varphi \to \psi$ and n codes a proof of φ, then $m \star n \star 2^{\ulcorner\psi\urcorner}$ codes a proof of ψ: write the first proof, then the second, then ψ by modus ponens. So

$$Prf(m, \ulcorner\varphi \to \psi\urcorner) \wedge Prf(n, \ulcorner\varphi\urcorner) \implies Prf(m \star n \star 2^{\ulcorner\psi\urcorner}, \ulcorner\psi\urcorner)$$

holds for all m, n. What (C2) needs [Smith 35.1] is that PA can prove the general version, with variables in place of m and n. Because Prf was built canonically, by following the p.r. definition

of *Prf*, one can follow the informal reasoning step by step inside PA. Smith says the details are tedious and unilluminating and doesn't give them, and neither do I.

There is a real point hidden here. (C2) is a fact about the *wff* Prf, not just about the relation it captures. If we replaced Prf by a "noisy" wff that captures the same relation (Smith's example is Prf conjoined with some irrelevant theorem), the informal facts would be unchanged but PA might no longer be able to follow the reasoning, and (C2) could fail. This is why the canonical construction in Section 4 matters.

(C3). Since $\Box\varphi$ is always Σ_1, (C3) follows from the following result with $\psi := \Box\varphi$.

Theorem 8.1 (Smith 35.2, 35.3; not proved here)**.** For every Σ_1 sentence ψ, $\mathsf{PA} \vdash \psi \to \Box\psi$. Hence $\mathsf{PA} \vdash \Box\varphi \to \Box\Box\varphi$ for every φ.

Theorem 2.9 says every *true* Σ_1 sentence is provable. This theorem says PA *knows* that: it proves "if ψ then ψ is provable". The plan is to formalize the proof of Theorem 2.9 inside PA. That proof was an induction, outside Q, on the construction of Σ_1 wffs, and we run the same induction again, now showing at each stage that $\mathsf{PA} \vdash \psi \to \Box\psi$. The catch is that the induction passes through wffs with free variables (to get to $\exists \mathsf{x}\, \psi(\mathsf{x})$ we need $\psi(\mathsf{x}) \to \Box\psi(\mathsf{x})$ for open ψ), so we need a version of $\Box$ that applies to open wffs, and proving the base cases (like $\mathsf{x}{+}\mathsf{y} = \mathsf{z} \to \Box(\mathsf{x}{+}\mathsf{y} = \mathsf{z})$) uses the induction schema inside PA. Smith sketches this and points to Boolos, *The Logic of Provability*, ch. 2, for the details.

Of everything in the report, this is the step I'd most like to see written out, because it's where the Second Theorem really depends on what the theory can prove about itself. I believe it for two reasons. Its informal content ("if a Σ_1 sentence is true, a proof of it can be found by following the construction") is exactly the proof of Theorem 2.9, which we have seen. And that proof only uses elementary manipulations of codes, which is the kind of reasoning PA is good at.

Only a little induction is needed, so the argument works for more than PA.

Definition 8.2 (Smith §35.3)**.** T is *nice** if it is nice and extends $\mathsf{I}\Sigma_1$.

Theorem 8.3 (Smith 35.4, 35.5; not proved here)**.** The derivability conditions hold for every nice* theory. Hence if T is nice*, then $T \nvdash \mathsf{Con}_T$.

In particular, if PA is consistent then $\mathsf{PA} \nvdash \mathsf{Con}_{\mathsf{PA}}$. (The Second Theorem also holds for Q itself, but the proof is different, since Q is too weak for the derivability conditions [Smith §35.4].)

9 What was taken on trust

Every proof above is complete except for the following, which are sketched or cited.

1. **Q's facts about $\leq$** (Theorem 2.7: (O3), (O8)). Proved in Smith §11.8 by simple outside inductions. I gave the idea for (O3) and the derivations of (O4), (O5).

2. **Closure properties of p.r. functions** (Theorem 3.3). Each is a short explicit definition; a sample is given.

3. **The general case of primitive recursion** in Theorem 4.10. I did $f(0) = g$, $f(Sx) = h(f(x))$ in full; extra parameters $\vec{x}$ and the dependence of h on y only add bookkeeping.

4. ***Prf*$_T$ and *diag* are p.r.** (Theorems 5.4, 5.6). Argued via bounded loops; Smith's Chapter 20 has the explicit definitions.

5. **(C2) for** PA. The informal fact is clear; transcribing it into PA is long.

6. **(C3) for** PA**, i.e. formalized** Σ_1**-completeness.** The main one: the Second Theorem rests on it.

Items 1 to 4 are routine checks, and I'm comfortable relying on them. Items 5 and 6 are different. They are claims about what PA can prove about its own proofs, and the Second Theorem depends on them in a way that the First Theorem depends on nothing comparable. If there is one place to go through a full proof, it is item 6 (Boolos, ch. 2).