For we know in part, and we prophesy in part.
But when that which is perfect is come, then
that which is in part shall be done away.
(I. Corinthians 13, 9–10)

Part B

Incompleteness

Chapter III

Self-Reference

Introduction. The preceding Part A was devoted to positive results in fragments of arithmetic: recall that we already showed that Robinson's Q proves all true Σ_1 sentences (which was rather easy) and then devoted considerable space to the development of mathematics in fragments ($I\Sigma_n$ or $B\Sigma_{n+1}$). We now know how to develop a theory of finite sets and definable infinite sets in $I\Sigma_1$ and we have arithmetized important parts of logic, recursion theory and combinatorics, sometimes using fragments stronger than $I\Sigma_1$. The present Part B will deal with *incompleteness* of systems of axioms in arithmetic: we shall prove the celebrated Gödel's incompleteness theorems (saying among other things that no consistent axiomatized theory containing Q is complete, i.e. each such theory has a sentence φ such that $(T+\varphi)$ and $(T+\neg\varphi)$ is consistent. Moreover, we shall be interested in comparing theories (containing arithmetic) with respect to their strength. One possibility is just to investigate the inclusion of theories (T is a subtheory of S, i.e. each formula provable in T is provable in S), but we shall study two related notions, namely interpretability (one can define basic notions of T in S such that T becomes a subtheory of S modulo these definitions) and partial conservativity (for some class Γ of formulas, e.g. Σ_k-formulas, each $\varphi \in \Gamma$ provable in T is provable in S). The study will be confined to theories containing a certain fragment of arithmetic.

The present Chapter III is devoted to Gödel's method of self-reference (which is of proof-theoretical character); the next Chapter IV will deal with models of fragments. Note that self-reference has interesting philosophical aspects; but they will be entirely disregarded. Chapter III has the following structure:

Section 1 contains preliminares, in particular it presents a definition of a theory *containing* (some) *arithmetic.* This is important since the results of this chapter are not confined to theories in the language of arithmetic (and apply, for example, also to systems of set theory). Section 2 contains Gödel's incompleteness theorems and related topics, as well as a characterization of interpretability among theories like *PA* (having induction for all formulas). In Sect. 3 we shall deal with theories not having induction for all formulas, mainly with finitely axiomatized theories; we present a strengthening of Gödel's second incompleteness theorem and a characterization of interpretability for such theories. Finally in Sect. 4 we shall systematically study interpretability and partial conservativity with a special emphasis to finitely axiomatizable theories. Among other theories we shall study the system ACA_0 – finitely axiomatizable second order arithmetic extending PA conservatively – and show how these theories differ with respect to interpretability. (This will be obtained as a corollary to general theorems on interpretability and partial conservativity.)

1. Preliminaries

(a) Interpretability and Partial Conservativity

This subsection may be viewed as a continuation of Sect. 0 (Preliminaries) at the beginning of the book. We collect here several definitions and state some obvious facts. Recall the notion of a theory T (0.11) and its language L. It is clear what we mean by saying that T and L are Δ_1.

1.1 Definition. (1) T is a *subtheory* of S (or: S is an *extension* of T) if the language L_T of T is a sublanguage of the language L_S of S and if each formula of L_T provable in T is provable in S.

(2) Let Γ be a class of formulas. T is *Γ-conservative* over S if each Γ-formula φ of L provable in T is provable in S.

(3) S is a *Γ-conservative extension* of T if S is an extension of T, i.e. T is a subtheory of S, and, in addition, S is Γ-conservative over T; i.e. if *any* formula is T-provable then it is S-provable; and if a Γ-formula is S-provable then it is T-provable. (Thus S is stronger then T but for Γ-formulas T-provability coincides with S-provability.)

1.2 Definition. (1) Let L_1 be a language and let S be a theory in a language L_2. To *interpret* L_1 in S means to define the following in S:

- the range of variables of L_1-formulas in S and
- for each predicate P, each function symbol F and each constant c of L_1, its translation P^*, F^*, c^* into S.

In more details, such an interpretation $*$ is given by:

- a formula $\chi(x)$ such that $S \vdash (\exists x)\chi(x)$,
- for each n-ary predicate P of L_1, a formula $\psi_P(x_1 \dots x_k)$ of L_2 with exactly n free variables,
- for each n-ary function symbol F of L_1, a formula $\psi_F(x_1 \dots x_k, y)$ of L_2 with exactly $k+1$ free variables such that

$$S \vdash \bigwedge_{i=1}^{k} \chi(x_i) \to (\exists! y)(\chi(y) \,\&\, \psi_F(x_1 \ldots x_{k,y}))$$

and similarly for ψ_c ($S \vdash (\exists! y)(\chi(y) \,\&\, \psi_c(y))$).

Using ψ_P, ψ_F and ψ_c we may define in S a predicate P^*, a function F^* (defining $F^*(x_1 \ldots x_k)$ by ψ_F if $\bigwedge_i \chi(x_i)$ and putting e.g. $F^*(x_1 \ldots) = x_1$ otherwise) and a constant c^*; this extends to a translation of each L_1-formula $\varphi(x_1 \ldots x_k)$ into an L_2-formula $\varphi^*(x_1, \ldots, x_k)$ (more precisely: into a formula of L_2 enriched by the $*$-symbols); φ^* results from φ by replacing each predicate, function, constant by its starred counterpart. The formula φ^* may be called "φ in the sense of the interpretation" or "the translation of φ".

(3) If T is a theory in the language L_1, S a theory in L_2 and $*$ is an interpretation of L_1 in S then $*$ is an *interpretation* of T in S if, for each axiom $\varphi(x_1, \ldots, x_k)$ of T,

$$S \vdash \bigwedge_{i=1}^{k} \chi(x_i) \to \varphi^* .$$

We have the following evident theorem:

1.3 Theorem. If $*$ is an interpretation of T in S then for each L_1 formula φ, $T \vdash \varphi$ implies $S \vdash \bigwedge \chi(x_i) \to \varphi^*$; in particular, for each closed L_1-formula φ, $T \vdash \varphi$ implies $S \vdash \varphi^*$.

Proof by induction on the length of a proof. □

1.4 Definition. This generalizes to a *parametrical interpretation* of L_1 in S: it consists of a formula $\vartheta(\mathbf{z})$ such that $S \vdash (\exists \mathbf{z})\vartheta(\mathbf{z})$ (range of parameters), a formula $\chi(x, \mathbf{z})$ such that $S \vdash \vartheta(\mathbf{z}) \to (\exists x)\chi(x, \mathbf{z})$ (range of L_1-variables given parameters) and for each P, F, c as above formulas $\psi_P(\mathbf{x}, \mathbf{z})$, $\psi_F(\mathbf{x}, y, \mathbf{z}))$, $\psi_c(y, \mathbf{z})$ of appropriate arities such that $S \vdash \vartheta(\mathbf{z}) \,\&\, \bigwedge_i \chi(x_i, \mathbf{z}) \to (\exists! y)(\chi(y, \mathbf{z}) \,\&\, \psi_F(\mathbf{x}, y, z))$ and similarly for ψ_c. The definition of φ^* for a formula $\varphi(x_1 \ldots x_n)$ (not contining the variables $\mathbf{z}$) is clear; $*$ is a *parametrical interpretation of T in S* if, for each axiom $\varphi(\mathbf{x})$ of T,

$$S \vdash (\vartheta(\mathbf{z}) \,\&\, \bigwedge_i \chi(x_i, \mathbf{z})) \to \varphi^*(\mathbf{x}, \mathbf{z}) .$$

1.5 Remark. (1) A generalization of 1.3 for parametrical interpretation is evident.

(2) The (parametrical) interpretation $*$ has *absolute equality* if $S \vdash x =^* y \to x = y$. We restrict ourselves to interpretations with absolute equality except when stated otherwise.

(3) Note that if $*$ is parametrical and S can define an object satisfying ϑ (i.e. for some $\vartheta'(\mathbf{z})$, $S \vdash \vartheta'(\mathbf{z}) \to \vartheta(z)$ and $S \vdash (\exists!\mathbf{z})\vartheta'(\mathbf{z})$ then $*$ may be replaced by a non-parametrical interpretation $**$ such that, for each closed T-formula φ, $S \vdash \varphi^* \to \varphi^{**}$ (just postulate that $\mathbf{z}$ is the unique tuple satisfying ϑ' throughout). This is in particular the case if S contains a fragment of arithmetic and $\vartheta(z)$ is an arithmetical formula such that S proves $L\vartheta$ (least number principle).

(4) The reader may observe that if $*$ is an interpretation of T in S then $*$ defines in each model $M \models S$ a model $M' \models T$; similarly for a parametrical interpretation.

1.6 Corollary. If T is (parametrically) interpretable in S and S is consistent then T is also consistent.

1.7 Remarks. (1) Clearly, each theory T is interpretable in itself by means of the *identical interpretation.*

(2) If T is interpretable in S and S is interpretable in U then T is interpretable in U by means of the *composed interpretation.*

1.8 Lemma. If T is interpretable in $(S + \{\varphi\})$ and also in $(S + \{\neg\varphi\})$ then T is interpretable in S.

Proof. Let $(\chi_i, \psi_{P,i} \dots)$ be the interpretation in question; $(i = 1, 2)$. Put $\chi(x) \equiv (\varphi \,\&\, \chi_1(x)) \vee (\neg\varphi \,\&\, \chi_2(x))$,

$$\psi_P(\mathbf{x}) \equiv (\varphi \,\&\, \psi_{P,1}(\mathbf{x})) \vee (\neg\varphi \,\&\, \psi_{P,2}(\mathbf{x}))$$

etc. Similarly for parametrical interpretations. $\square$

1.9 Definition. An interpretation $*$ of T in S is *Γ-faithful* if, for each closed $\varphi \in \Gamma$, $T \vdash \varphi$ is equivalent to $S \vdash \varphi^*$. It is *faithful* if it is Γ-faithful for Γ being the set of all L_1-formulas.

(b) Theories Containing Arithmetic; Sequential Theories; *PA* and ACA_0

1.10 Convention. Saying that S *contains* T we shall mean that T is interpretable in S and a certain interpretation of T in S has been fixed. Thus, in particular, if T is a subtheory of S then S contains T; but in general axioms of T are assumed to hold only on a subdomain of the universe of S. A typical example is Zermelo-Frankel set theory *ZF*: in *ZF* we can define a set N and prove that its elements satisfy axioms of *PA* with respect to appropriately defined operations of successor, addition and multiplication. In other words, *PA* is interpreted in *ZF* by these definitions.

1.11 Remark. It is possible that there are two substantially different interpretations of T in S, so we have to fix one.

1.12 Definition. A theory T containing Q is *sequential* if there are predicates $SEQ(z,u)$ and $\beta(x,v,z)$ coding sequences of arbitrary objects of T in the following weak sense (read $SEQ(z,u)$ "z codes a sequence of length at least u" and $\beta(x,v,z)$ "x is the v-th element of z"; let $Number(u)$ be the domain of the chosen interpretation of Q in T): T proves

(1) $SEQ(z,u) \to Number(u) \,\&\, (\forall v < u)(Number(v) \to (\exists! x)\beta(x,v,z))$;
(2) $SEQ(z,u) \to (\forall y)(\exists z')(SEQ(z',u+1) \,\&\, (\forall v < u)(Number(v) \to (\forall x)(\beta(x,v,z) \equiv \beta(x,v,z')) \,\&\, \beta(y,u,z')$.

(Here $u+1$ means the successor of u in the sense of the interpretation.)

1.13 Remark. (1) says that if z codes a sequence of length u then for each $v < u$, the v-th element of z is uniquely determined; and (2) is a prolongation axiom. Note that we do not define the length of a sequence. Nor do we have any extensionality etc. This notion is particularly importnat for the case that T has also objects that are not numbers, as the theory ACA_0 below.

Note that a theory T in the language of arithmetic stronger than $I\Sigma_0^{exp}$ *is* sequential; we may take $Seq(s) \,\&\, lh(s) = u$ for $SEQ(s,u)$ and $(s)_v = x$ for $\beta(x,v,s)$. In Chap. V we show that each theory in the langauge of arithmetic containing $I\Sigma_0$ is sequential.

In the next definition we introduce a particular system ACA_0 of second order arithmetic and later we show that it contains PA. Results of this chapter will have various corollaries concerning relations of PA and ACA_0. The reader uninterested in ACA_0 may skip the rest of the subsection.

1.14 Definition. The language of second order arithmetic consists of
(1) variables of two sorts: number variables $x, y, \ldots$, and set variables $X, Y, \ldots$.
(2) predicates $=$, $\leq$, $\in$ (binary), function symbols S, $+$, $*$ (usual arities) and a constant $\overline{0}$.

First order terms coincide with terms of the language of (first order) arithmetic, i.e. are elements of the free algebra generated by number variables, the constant $\overline{0}$ and function symbols. *Second order terms* are set variables.

Atomic formulas are formulas of the following form: $t = s$, $t \in X$, $X = Y$ where t, s are first order terms and X, Y are second order variables. *Formulas* result from atomic formulas using connectives and first order and second order quantifiers $(\forall x)$, $(\forall X)$.

The second order arithmetic ACA_0 has the language of second order arithmetic and the following axioms:

(1) Axioms of Robinson's arithmetic Q,
(2) Arithmetical comprehension scheme: for each formula $\varphi(x, \mathbf{y}, \mathbf{X})$ *not containing any second order quantifiers*, and not containing the variable Z, the axiom

$$(\exists Z)(\forall x)(x \in Z \equiv \varphi(x, \mathbf{y}, \mathbf{X}))$$

(3) The following (single) *induction axiom:*

$$\overline{0} \in X \,\&\, (\forall x)(x \in X \to S(x) \in X) \to (\forall x)(x \in X).$$

(4) $X = Y \equiv (\forall x)(x \in X \equiv x \in Y)$ (extensionality).

1.15 Remarks. (1) How does this fit into our notion of an axiomatic theory? It is indeed possible to generalize the notion of an axiomatic theory as presented in Sect. 0 to allow various sorts of variables (see e.g. [Kreisel-Krivine]); but, on the other hand, we may consider any many-sorted theory as one-sorted with an unary predicate for each sort. This means for ACA_0: Instead of variables of two sorts we have two predicates $Number(x)$, $Set(x)$ and axioms that there are numbers, there are sets and $Number(x) \equiv \neg Set(x)$; we have predicates and function symbols of the language L_0 of first order arithmetic and in addition $\in$. We formulate axioms saying that numbers are closed under S, $+$, $*$ and satisfy Q (e.g. $Number(x) \,\&\, x \neq 0 \to (\exists y)(Number(y) \,\&\, x = S(y))$ etc.). Similarly we rewrite the comprehension.

$$\begin{aligned} Number(y) \,\&\, Set(w) \to (\exists z)(&Set(z) \\ &\& (\forall x)(Number(x) \to (x \in z \equiv \varphi'(x, y, w)))) \end{aligned}$$

(where φ' results from a formula φ above by the obvious changes) and the induction axiom. Finally we add extensionality:

$$Set(x) \,\&\, Set(y) \to (\forall u)(Number(u) \to (u \in x \equiv u \in y)) \equiv x = y.$$

Thus we understand the two-sorted formulation as a shorthand for the one-sorted system just described. (The interested reader could show that this is equivalent to a consequent many-sorted approach; models of both formulations are in an obvious correspondence.)

(2) How does ACA_0 fit into various systems of second order arithmetic? And why is ACA_0 introduced and studied here? A much stronger theory results from ACA_0 if we postulate comprehension for *all* φ; this is *the full* second order arithmetic. It has important particular subsystems studied in Simpson's book [Simpson]. We shall not go into this; but we shall show below that ACA_0 extends *PA* conservatively and ACA_0 is finitely axiomatizable. We shall see later in this chapter that *PA* is *not* finitely axiomatizable; but it is the union of the hierarchy of theories $I\Sigma_k$, $k \in N$, each being finitely axiomatizable (see Chap. I, Sect. 2) and has a finitely axiomatizable

conservative extension (in a richer, second-order language), namely ACA_0. Note in passing that the relation of ACA_0 to PA is very similar to the relation of Gödel-Bernays set theory GB to Zermelo-Fraenkel set theory ZF (and, of course, both ZF and GB contain both PA and ACA_0).

1.16 Theorem. ACA_0 is a conservative extension of PA; more precisely, the obvious interpretation of the language of PA in ACA_0 is a faithful interpretation of PA in ACA_0.

Remark. If we treat ACA_0 consequently as a two-sorted theory then formulas of PA are particular formulas of ACA_0 and it makes sense to say that ACA_0 is a conservative extension of PA. If we understand ACA as a one sorted theory, we have to interpret PA-formulas as ACA_0-formulas by restricting all quantifiers to $Number(x)$. But if there is no danger of misunderstanding we shall identify both approaches.

Proof. First we show that ACA_0 extends PA, i.e. that each induction axiom I_φ is provable in ACA_0. But this is easy: by comprehension, let $Z = \{x \mid \varphi(x, \mathbf{y})\}$. Assume $\varphi(0, \mathbf{y})$ and $(\forall x)(\varphi(x, \mathbf{y}) \rightarrow \varphi(S(x), \mathbf{y}))$. Then $0 \in Z$ and $(\forall x)(x \in Z \rightarrow S(x) \in Z)$. We get $((\forall x)(x \in Z)$, i.e. $(\forall x)\varphi(x, \mathbf{y})$. Now we prove conservativity.

A model-theoretic proof is easy: take any countable model M of PA; without loss of generality you may assume that no subset of M is an element of M. Interpret elements of M as numbers and take all parametrically definable subsets of M for sets: Let S be the set of all such subsets. The new model M' has the domain $M \cup S$, operations and ordering are as in M (trivially extended to S), and $\in$ is interpreted as the restriction of actual membership to $M \times S$. Checking that $M' \models ACA_0$ is not difficult and is left to the reader. (One may consult 0.9 and/or Chap. IV, Sect. 1 (a) if necessary). Thus: if a PA-sentence φ is consistent with PA it is consistent with ACA_0; thus ACA_0 is conservative over PA. ☐

Remark. In Chap. IV, Sect. 4 we shall formalize this proof in $I\Sigma_1$ and show that the assertion "ACA_0 extends PA conservatively" is provable in $I\Sigma_1$. By the results of Chap. IV, this will imply that there is a *primitive recursive* function associating to each ACA_0-proof of a PA-formula its PA-proof. (Better results are known.)

1.17 Theorem. ACA_0 is sequential.

Proof. We have to define coding of sequences of numers and sets. We shall define a coding of sequences of sets, and give indications how to modify the definition to code sequences of both numbers and sets.

Define $SEQ(Z, u)$ iff Z consists of ordered pairs; and put $\beta(X, v, Z)$ iff $(\forall x)(x \in X \equiv (v, x) \in Z)$. Note that if $SEQ(Z, u)$ then $SEQ(Z, u')$ for any

u'. Thus we write just $SEQ(Z)$. If $SEQ(Z)$ and Y is a class then define $G(Z,Y,u)$ as follows:

$$(x,v) \in G(Z,Y,u) \equiv (v \neq u \,\&\, (x,v) \in Z. \vee .v = u \,\&\, y \in Y).$$

Existence follows by comprehension; and verification of necessary provabilities is easy.

If we want to code sequences of both sets and numbers we may consider sets Z of triples $((x,\varepsilon),v)$ such that for each v either for all (x,ε) such that $((x,\varepsilon),v) \in Z$ we have $\varepsilon = 1$ (and then $\beta(X,v,Z) \equiv X = \{x \mid ((x,1),v) \in Z\}$ or there is exactly one x such that $(x,0),v) \in Z$ (and then $\beta(x,v,Z) \equiv ((x,0),v) \in Z$). □

1.18 Theorem. ACA_0 is finitely axiomatizable.

Proof. We are interested in a quick proof, not in a polished finite axiom system. Thus we shall start with $I\Sigma_1$, which we know to be finitely axiomatizable. We add the induction axiom (A1), which is a single axiom. Observe that ACA_0 proves $(\forall X)(X$ is piecewise coded) (consider the class of all x such that there is a piece of X of length x); thus take the axiom (A2) saying that each X is p.c. Recall the relativized satisfaction $Sat_{0,X}$ (see I.2.55); take the axiom (A3) $(\forall f \in \Sigma_0^\bullet(X))(\exists Z)(\forall e)(e \in Z \equiv Sat_{0,X}(f,e))$ (we may desire that free variables of f are the first u free variables and e is an u-tuple of numbers). In fact, it is enough to quantify f over all *open* $\Sigma_0(X)$-formulas. Note that our theory proves Tarski's truth conditions for $Sat_{0,X}$ so that we have "snowing"-snowing for $\Sigma_0^\bullet$(X)-formulas. (A4) and (A5) will describe universal and existential projection:

(A4) $(\forall X)(\exists Z)(\forall s)(s \in Z \equiv (\forall x)(\langle x\rangle \frown s \in X)$

(A5) $(\forall X)(\exists Z)(\forall s)(s \in Z \equiv (\exists x)(\langle x\rangle \frown s \in X)$

(A6) says $(\forall X)(\exists Z)(\forall x)(x \in Z \equiv \langle x\rangle \in X)$.

Clearly, $I\Sigma_1$ + (A2) – (A6) proves each instance of comprehension for a formula $\varphi(x,X)$ without second order quantifier: Apply (A3) to φ_0, where φ_0 is the open part of φ (φ assumed prenex) and then apply (A4) and (A5) according to the quantifier prefix of φ. Thus in particular we have comprehension for all formulas not containing any set variables at all.

We add one more axiom that reduces the general case to the subcase $\varphi(x,X)$ just described.

We take the axiom stating the property of the function $G(Z,Y,u)$:

$$(\forall u)(\forall Y)(\forall Z)(SEQ(Z) \to (\exists W)(\forall x,v)((x,v) \in W \equiv (v \neq u \,\&\, (x,v) \in Z \vee v = u \,\&\, x \in Y).$$

This makes possible to replace k sets by just one; e.g. X_0, X_1, X_2 are replaced by $G(G(G(0,X_0,0),X_1,1)),X_2,2)$. □

(c) Numerations and Binumerations

We have already mentioned the notion of a binumeration of a set of natural numbers in a theory $T \supseteq Q$ (see I.1.65). In this short subsection we define a more general notion of a numeration and prove some easy facts on them. Deeper theorems on numerations will be proved (and used) in Sect. 3. At the end of this subsection we shall show that our present knowledge is sufficient for a proof of a weak form of Gödel's first incompleteness theorem. A strong form will be proved by means of self-reference in Sect. 2.

1.19 Definition. Let $R \subseteq N^n$ be a relation. A formula $\varphi(x_1, \ldots, x_n)$ *numerates* R in a theory T containing Q if, for each $k_1, \ldots, k_n \in N$, we have the following:

$$(*) \qquad R(k_1, \ldots, k_n) \Leftrightarrow T \vdash \varphi(\overline{k}_1, \ldots, \overline{k}_n).$$

1.20 Remark. Note that φ binumerates R in T if φ numerates R in T and $\neg\varphi$ numerates the complement of R in T, i.e. besides $(*)$ we have

$$\text{not } R(k_1, \ldots, k_n) \Leftrightarrow T \vdash \neg\varphi(k_1, \ldots, k_n).$$

1.21 Definition. A theory T containing Q is *Γ-sound* (where Γ is a class of formulas of L_0) if each L_0-formula provable in T is true in the standard model N. T is *sound* if it is Γ-sound for Γ being the class of all L_0-formulas. (Thus PA is sound and so are its subsystems.)

1.22 Remark. We shall be particularly interested in Σ_1-sound theories. They are also called 1-consistent; T is 1-consistent if $(T + Tr(\Pi_1))$ is consistent, where $Tr(\Pi_1)$ is the set of all Π_1-formulas true in N. (Show that the two definitions are equivalent.) Note that Σ_1-soundness implies consistency.

1.23 Theorem. (1) If A is defined by a Σ_1-formula $\varphi(x_1, \ldots, x_n)$ in N then φ numerates A in Q and also in each Σ_1-sound theory T containing Q.

(2) If A is Δ_1 (in N) then there is a Σ_1-formula φ which binumerates A in Q and also in each consistent T containing Q.

Proof. (1) follows immediately from Σ_1-completeness and Σ_1-soundness. (See 1.1.8). To prove (2) first assume $A \subseteq N$; take a Σ_1-definition $(\exists y)\sigma(x, y)$ and a Π_1-definition $(\forall y)\pi(x, y)$ of A ($\sigma, \pi \in \Sigma_0$). Consider the following formula $\varphi(x)$:

$$\varphi(x) \equiv (\exists y)(\sigma(x, y) \,\&\, (\forall z \leq y)\pi(x, y)).$$

(This is our first example of *witness comparison* that will play an extremely important role in the sequel: The formula says that there is a witness y for

$(\exists u)\sigma(x,u)$ such that beneath y there is no witness for $(\exists u)\neg\pi(x,u)$.) If $k \in A$ then for some m, $N \models \sigma(\overline{k},\overline{m})$, trivially, for each $m' \leq m$, $N \models \pi(\overline{k},\overline{m})$ since $N \models (\forall y)\pi(k,y)$. Thus Q proves $\neg\pi(\overline{k},\overline{m})\,\&\,(\forall y \leq \overline{m})\neg\sigma(k,y)$. But then Q proves $\neg\varphi(\overline{k})$: Work in Q and assume $\varphi(\overline{k})$; let y_0 be such that $\sigma(\overline{k},y_0)$ and $(\forall z \leq y_0)\pi(k,z)$. We have either $y_0 \leq \overline{m}$ or $y_0 \geq \overline{m}$ (cf. I.1.6). But $y_0 \leq \overline{m}$ implies $\neg\sigma(\overline{k},y_0)$, a contradiction, and $y_0 \geq \overline{m}$ implies $\pi(\overline{k},\overline{m})$, also a contradiction. This completes the proof. □

Now assume that R is an n-ary relation. Let ML_n be the maximo-lexicographical ordering of N^n, i.e. $(k_1,\ldots,k_n)ML_n(q_1,\ldots,q_n)$ iff $\max(k_1,\ldots,k_n) < \max(q_1,\ldots,q_n)$ or [the maxima are equal and for the first i such that $k_i \neq q_i$ we have $k_i < q_i$]). Define ML_n by a Σ_0 formula $(x_1,\ldots,x_n) \leq_n (y_1,\ldots,y_n)$ by copying the definition (caution: we use a *disjunction* over $i = 1,\ldots,n$). Analogously as for $<$, Q proves $(x_1,\ldots,x_n) \leq_n (\overline{q}_1,\ldots,\overline{q}_n) \equiv \bigvee\{\bigwedge_i x_i = \overline{k}_i \mid (\overline{k}_1,\ldots,\overline{k}_n)ML_n(\overline{q}_1,\ldots,\overline{q}_n)\}$. (Cf. 5.1.6.). Having this construct a $\varphi(x_1,\ldots,x_n)$ from $(\exists u)\sigma(x_1\ldots x_n,u)$ and $(\forall u)\pi(x_1,\ldots,x_n,u)$ as above but using $\leq_n$ instead of $<$.

1.24 Remark. (1) The preceding theorem implies that each Δ_1 relation A also has a Π_1 binumeration in $T \supseteq Q$ (since the complement of A is also Δ_1)

(2) Recall the simple remark in I.1.65 which states that if A is defined by a formula φ which is Δ_1 in T then A is binumerated by φ in T.

(3) In Sect. 3 we remove the assumption of Σ_1 soundness from 1.23 (1).

(4) Functions are particular relations; thus if F is a Δ_1 mapping of N into N then there is a Σ_1 formula φ such that

$$k = F(m) \Leftrightarrow Q \vdash \varphi(\overline{k},\overline{m})$$
$$k \neq F(m) \Leftrightarrow Q \vdash \neg\varphi(\overline{k},\overline{m})$$

But for functions we can say even more (generalization for functions of several variables being left to the reader):

1.25 Theorem. Let $F : N \to N$ be Σ_1 (possibly partial). Then there is a Σ_1 formula $\varphi(x,y)$ such that, for each $m \in dom(F)$,

$$Q \vdash \varphi(\overline{m},y) \equiv y = \overline{F(m)}\,.$$

Proof. Let F be defined by a Σ_1-formula $(\exists z)\sigma(x,y,z)$, σ being Σ_0 (i.e. $F(m) = k$ iff $N \models (\exists z)\sigma(\overline{m},\overline{k},z)$). Let $\varphi(x,y)$ be the following formula $(\exists z)\varphi_0(x,y,z)$:

$$\begin{aligned}(\exists z)[\sigma(x,y,z)\,\&\,(\forall u,v \leq y)(u \neq y \to \neg\sigma(x,u,v)&\\ \&\,(\forall u,v \leq z)(u \neq y \to \neg\sigma(x,u,v))]\,.&\end{aligned}$$

(Note that in general we cannot speak on $\max(y, z)$ in Q.) First let $F(m) = k$, $N \models \sigma(\overline{m}, \overline{k}, \overline{q})$; then it is easy to show $Q \vdash \varphi_0(\overline{m}, \overline{k}, \overline{q})$. Thus $Q \vdash y = \overline{k}_1 \to \varphi(\overline{m}, y)$. Second, work in Q and assume $u \neq \overline{k}$ and let $\varphi_0(\overline{m}, u, v)$. For k and q we do have a maximum; denote it h.

Case 1. $\overline{h} \leq u$ or $\overline{h} \leq v$. Then start from $\varphi_0(\overline{m}, u, z)$ and show $\neg\sigma(\overline{m}, \overline{k}, \overline{q})$ – contradiction.

Case 2. $u, v \leq \overline{h}$. Then start from $\varphi_0(\overline{m}, \overline{k}, \overline{q})$ and show $\neg\sigma(\overline{m}, u, v)$, again a contradiction. This shows (in Q) $y \neq \overline{k} \to \neg\varphi(\overline{m}, y)$.

Remark. The reader may formulate and prove a generalization to functions with several arguments as an exercise.

1.26 Remark. In Sect. 2 we shall study Gödel's celebrated incompleteness theorem; our main method will be Gödel's self-reference technique. In the following remark we shall show how to give a non-constructive proof of a weak form of Gödel's first incompleteness theorem using means available so far. We claim the following:

If T is a Σ_1-sound Σ_1 theory containing Q then T is incomplete; there is a Σ_1 sentence ν such that both ν and $\neg\nu$ are unprovable in T.

Proof. Recall that there is a Σ_1 set K_0 of natural numbers which is not Δ_1 (the reader may find a proof of this in the next remark).

Now let $\alpha(x)$ be a Σ_1 formula numerating K_0 in T, i.e. $K_0 = \{n \mid T \vdash \alpha(\overline{n})\}$. Let $K_1 = \{n \mid T \vdash \neg\alpha(\overline{n})\}$. Since K_0 is not Δ_1 and T is consistent we have $K_0 \cup K_1 \neq N$; for $n \in N - K_0 - K_1$ both $\alpha(\overline{n})$ and $\neg\alpha(\overline{n})$ are unprovable in T. □

1.27 Remark. For the reader's convenience we prove here that there is a Σ_1 set of natural numbers which is not Δ_1. For example, let K_0 be the set of all φ such that φ is a Σ_1-formula with just one free variable and is satisfied on N by φ (itself). Then K_0 is Σ_1 (it is defined by

$$x \in \Sigma_1^\bullet \,\&\, x \text{ has one free variable}^\bullet \,\&\, Sat_{\Sigma,1}(x, [x]))\,.$$

If $N - K$ were Σ_1 then we could produce a Σ_1-formula $\varphi(x)$ defining the same set of natural numbers as the formula

$$x \in \Sigma_1^\bullet \,\&\, x \text{ has one free variable}^\bullet \,\&\, \neg Sat_{\Sigma,1}(x, [x])\,.$$

By "snowing"-snowing.

$$N \models \overline{\varphi} \in \Sigma_1^\bullet \,\&\, \overline{\varphi} \text{ has one free variable}^\bullet\,;$$

now it is easy to see that we get the liar's paradox:

$$N \models Sat_{\Sigma,1}(\overline{\varphi}, [\overline{\varphi}]) \equiv \neg Sat_{\Sigma,1}(\overline{\varphi}, [\overline{\varphi}])\,,$$

which is a contradiction. We have shown that K_0 is not Σ_1.

2. Self-Reference and Gödel's Theorems, Reflexive Theories

We now come to the substance of Gödel's method of proof of incompleteness of arithmetic, namely to the use of arithmetization for *self-reference*, which is often roughly described as the existence of arithmetical sentences "speaking of themselves". Since in our representation, sentences are particular numbers, we are not surprised if we encounter sentences speaking of (other) sentences. Take, for example, the case of "snowing"-snowing: we know that $N \models \varphi$ iff $N \models Sat_{\Sigma,n}(\overline{\varphi}, \emptyset)$ (φ being Σ_n). Given any formula $\psi(x)$ with exactly one free variable, a sentence φ is *self-referential* with respect to ψ in N (or: is a *fixed-point* of ψ in N) if $N \models \varphi \equiv \psi(\overline{\varphi})$, thus $N \models \varphi$ is equivalent to $N \models \psi(\overline{\varphi})$. We can say rather suggestively that φ says "I have the property ψ". As usual, we are more interested in provability in a theory than in truth in N; we say that φ is a *fixed-point of* $\psi(x)$ *in* T if $T \vdash \varphi \equiv \psi(\overline{\varphi})$.

In subsection (a) we prove various theorems on the existence of fixed points; in (b) we prove Gödel's incompleteness theorems and related results, and in (c) we derive consequences for theories similar to full Peano arithmetic *PA*; among other things we show that *PA* is not finitely axiomatizable. We shall pay much attention to finitely axiomatizable theories in Sect. 3.

(a) Existence of Fixed Points

2.1 Fixed-point Theorem (Or Diagonal Lemma).

(1) ((Non-parametric version). Let T be a theory containing Q and let $\psi(x)$ be a formula with exactly one free variable. Then there is a sentence φ such that $T \vdash \varphi \equiv \psi(\overline{\varphi})$.

(2) (Parametric version). Let $T \supseteq Q$ and let $\psi(x, \mathbf{y})$ be a formula with free variables as indicated. Then there is a formula $\varphi(\mathbf{y})$ such that $T \vdash \varphi(\mathbf{y}) \equiv \psi(\overline{\varphi(\mathbf{y})}, \mathbf{y})$.

Remark. We prove (1) and indicate how to generalize to (2). The rest of the subsection contains some corollaries and related results as well as definitions of some properties that fixed points may have or not have.

Proof of (1). Let $\psi(x)$ be given. Let F be the Δ_1 function associating with each formula $\delta(x)$ (in the language of T having exactly one free variable x) the closed formula $\delta(\overline{\delta})$, i.e. the result of substitution of $\overline{\delta}$ into δ for x. (Let F be 0 for other arguments.) By 1.25 there is a formula $\alpha(x, v)$ such that, for each δ, $T \vdash \alpha(\overline{\delta}, v) \equiv v = \overline{F(\delta)}$. Let $\chi(x)$ be the formula $(\exists v)(\alpha(x, v) \,\&\, \psi(v))$ and let φ be $F(\chi)$, i.e. $\chi(\overline{\chi})$. Q proves the following equivalences:

$$\varphi \equiv \chi(\overline{\chi}) \equiv (\exists v)(\alpha(\overline{\chi}, v) \,\&\, \psi(y)) \equiv (\exists v)(\nu = \overline{F(\chi)} \,\&\, \psi(v))$$
$$\equiv (\exists v)(v = \overline{\varphi} \,\&\, \psi(v)) \equiv \psi(\overline{\varphi}).$$

To prove (2) modify F. F associates with each formula $\delta(x,\mathbf{y})$ the formula $\delta(\overline{\delta},\mathbf{y})$. Let α be as above, for the new F. If $\psi(x,\mathbf{y})$ is given, let $\chi(x,\mathbf{y})$ be the formula $(\exists v)(\alpha(x,v)\,\&\,\psi(v,y))$, let φ be $F(\chi)$. We have

$$Q \vdash \varphi(\mathbf{y}) \equiv \psi(\overline{\varphi},\mathbf{y})$$

as above. □

2.2. Corollary. (1) Let T contain Q and let $\psi(x,z)$ be a T-formula. Then there is a T-formula $\varphi(z)$ such that for each $k \in N$, $T \vdash \varphi(\overline{k}) \equiv \psi(\overline{\varphi(\overline{k})},\overline{k})$.

(2) More generally, if $\psi((x,z,\mathbf{y})$ is a T-formula then there is a T-formula $\varphi(z,\mathbf{y})$, such that, for each k,

$$T \vdash \varphi((\overline{k},\mathbf{y}) \equiv \psi(\overline{\varphi(\overline{k},\mathbf{y})},\overline{k},\mathbf{y})\,.$$

Proof. We prove (1). Let $F(\delta,k)$ be the Δ_1 function associating with each k the formula $\delta(\overline{k})$; let $\beta(u,v,w)$ be such that $Q \vdash \beta(\overline{\delta},\overline{k},\overline{w}) \equiv w = \overline{F(\delta,k)}$. Let $\psi'(x,z)$ be $(\exists w)(\beta(x,z,w)\,\&\,\psi(w,z))$; let $\varphi(z)$ be such that

$$T \vdash \varphi(z) \equiv \psi'(\overline{\varphi},z)\,.$$

Then $T \vdash \varphi(\overline{k}) \equiv \psi'(\overline{\varphi},\overline{k}) \equiv (\exists w)(\beta(\overline{\varphi},\overline{k},w)\ \&\ \psi(w,\overline{k}) \equiv \psi(\overline{\varphi(\overline{k})},\overline{k})$. □

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

2.3 Corollary. Let T be a consistent theory containing Q. Then there is no formula $Tr(x)$ such that for each φ, $T \vdash \varphi \equiv Tr(\overline{\varphi})$.

Proof. If we had $Tr(x)$ we could reproduce the liar's paradox: let φ be such that $T \vdash \varphi \equiv \neg Tr(\overline{\varphi})$ (by 2.1). Then $T \vdash \varphi \equiv \neg\varphi$, i.e. T is inconsistent. □

2.4 Remark. This means that T cannot define its own truth. In particular, if $T \supseteq I\Sigma_1$ is a consistent theory in the language of arithmetic we have, for each k, partial satisfactions $Sat_{\Sigma,k}$ and $Sat_{\Pi,k}$ for the universe but there is *no* formula $Sat(z,e)$ such that T proves Tarski's properties of satisfaction for *all* formulas.

2.5 Discussion. Let $T \vdash (\forall x)\psi(x)$; then φ is a fixed point of ψ in T iff $T \vdash \varphi$. (We can say that *ψ has only provable fixed points.*). If $T \vdash (\forall x)\neg\psi(x)$ then *ψ has only refutable fixed points.* In both cases, all fixed points of ψ are T-provably equivalent, i.e. if φ_1 and φ_2 are fixed points of φ then $T \vdash \varphi_1 \equiv \varphi_2$. We say that ψ has a *unique fixed point* up to an equivalence. These are trivial examples; we shall construct non-trivial examples in the next subsection.

2.6 Theorem (Existence of Self-referential Pairs). Let T contain Q and let $\psi_1(x,y)$, $\psi_2(x,y)$ be two T-formulas. Then there are T-sentences φ_1, φ_2 such that $T \vdash \varphi_1 \equiv \psi_1(\overline{\varphi}_1, \overline{\varphi}_2)$ and $T \vdash \varphi_2 \equiv \psi_2(\overline{\varphi}_1, \overline{\varphi}_2)$.

Proof. Let $F_i(\delta_1(x,y), \delta_2(x,y))$ be $\delta_i(\overline{\delta}_1, \overline{\delta}_2)$ $(i = 1,2)$; let $\alpha_i(u,v,w)$ be such that $Q \vdash \alpha_i(\overline{\delta}_1, \overline{\delta}_2, w) \equiv w = F_i(\delta_1, \delta_2)$. Let $\chi(x,y)$ be $(\exists w_1, w_2))(\alpha_1(x,y,w_1)$ & $\alpha_2(x,y,w_2)$ & $\psi_i(w_1,w_2))$; let φ_i be $\chi_i(\overline{\chi}_1, \overline{\chi}_2)$. The proof of $T \vdash \varphi_i \equiv \psi_i(\overline{\varphi}_1, \overline{\varphi}_2)$ is now routine. □

(b) Gödel's First Incompleteness Theorem and Related Topics

Fixed points theorems are particularly fruitful when applied to formulas related to the set of all T-*proofs* or the set of all T-provable formulas (where T is a theory containing Q). If T is rich enough to make arithmetization possible ($T \supseteq I\Sigma_1$ suffices) then it is natural to work with the formalized proof predicate $Proof^\bullet$ (see Chap. I, Sect. 4); but for Gödel's first incompleteness theorem it suffices that T contain Q and that the set of all proofs is Δ_1 and hence has a Σ_1 binumeration. (Gödel's second incompleteness theorem is more delicate and to prove it we have to guarantee some provability conditions, see below.) The theorems are proved in a constructive way, i.e. we give *examples* of sentences that are independent (neither provable nor refutable). Since these examples are both famous and useful we give their names (Gödel's formula, Rosser's formula).

2.7 Definition. Let T be a Σ_1 theory containing Q and let $\pi(x)$ be a Σ_1 formula defining the set of all T-provable sentences (i.e. $T \vdash \varphi$ iff $N \models \pi(\overline{\varphi})$). A *Gödel sentence* based on π is a fixed point of $\neg\pi$, i.e. a sentence ν such that

$$T \vdash \nu \equiv \neg\pi(\overline{\nu}).$$

(We may describe ν by saying that it asserts its own unprovability.)

2.8 Theorem. Let T, π be as above and let ν be a Gödel sentence given by π. (1) If T is consistent then T does not prove ν. (2) If T is Σ_1-sound then T does not prove $\neg\nu$.

Proof. (1) Assume $T \vdash \nu$, then $N \models \pi(\overline{\nu})$, thus $T \vdash \pi(\overline{\nu})$ by Σ_1-completeness; thus $T \vdash \neg\nu$ and T is inconsistent.

(2) Assume $T \vdash \neg\nu$, i.e. $T \vdash \pi(\overline{\nu})$. If T were Σ_1-sound then π would *numerate* the set of all provable sentences, thus we would have $T \vdash \nu$ and T would be inconsistent. Thus T cannot be Σ_1-sound. □

We present a technique that makes it possible to get rid of Σ_1-soundness.

2.9 Definition. Let $(\exists u)\alpha(x,u)$ and $(\exists u)\beta(x,u)$ be two Σ_1-sentences; thus α, β are Σ_0. A *witness comparison sentence* given by these sentences is the sentence

$$(*) \qquad (\exists u)[\beta(x,u) \,\&\, (\forall v < u)\neg\alpha(x,u)]$$

(Observe that this is again a Σ_1 sentence; if we call each u such that $\alpha(x,u)$ a *witness* for $(\exists u)\alpha(x,u)$ and similarly for $(\exists u)\beta(x,u)$ then $(*)$ says "there is a witness u for $(\exists u)\beta(x,u)$ such that no witness for $(\exists u)\alpha(x,u)$ is $< u$". The formula $(*)$ will be often denoted $(\exists u)\beta \preccurlyeq (\exists u)\alpha$. In particular, let $(\exists u)\alpha(x,u)$ be a Σ_1-formula defining the set of all T-provable sentences and let $(\exists u)\beta(x,u)$ be a Σ_1-formula defining the set of all T-refutable sentences, i.e. $T \vdash \varphi \Leftrightarrow N \models (\exists u)\alpha(\overline{\varphi},u)$ and $T \vdash \neg\varphi \Leftrightarrow N \models (\exists u)\beta(\overline{\varphi},u)$. A *Rosser* sentence given by α and β is a formula ρ such that

$$T \vdash \rho \equiv (\exists u)\beta(\overline{\rho},u) \preccurlyeq (\exists u)\alpha(\overline{\rho},u)\,,$$

i.e.

$$T \vdash \rho \equiv (\exists u)(\beta(\overline{\rho},u) \,\&\, (\forall v < u)\neg\alpha(\overline{\rho},u))\,.$$

Calling for a moment an element u such that $\alpha(x,u)$ a *demonstration* of x and u such that $\beta(x,u)$ a *refutation* of x, a Rosser sentence says: "there is a refutation of me beneath of which there is no demonstration of me".

2.10 Gödel-Rosser Incompleteness Theorem. Let T be a Σ_1 theory containing Q and let ρ be a Rosser sentence for T. If T is consistent then T proves neither ρ nor $\neg\rho$.

Proof. Assume $T \vdash \rho$, then for some d, $N \models \alpha((\overline{\rho},\overline{d})$ and hence $T \vdash \alpha(\overline{\rho},\overline{d})$. Now work in Q and let y be a witness for $(\exists y)\beta(\overline{\rho},y)$ such that beneath y there is no witness for $(\exists y)\alpha(\overline{\rho},y)$. Then $y \leq \overline{\rho} \vee \overline{d} \leq y$ (cf. I.1.6) and $\overline{d} \leq y$ is impossible since $\alpha(\overline{\rho},d)$. We have proved $T \vdash (\exists y < \overline{d})\beta(\overline{\rho},y)$, i.e. $T \vdash \bigvee_{e<d}\beta(\overline{\rho},\overline{e})$. But since T is consistent T does not prove $\neg\rho$ and thus for each e, $N \models \neg\beta(\overline{\rho},\overline{e})$, thus $T \vdash \neg\beta(\overline{\rho},\overline{e})$ by Σ_1-completeness. Thus $T \vdash \bigwedge_{e<d}\neg\beta(\overline{\rho},\overline{e})$ and T is inconsistent.

Second, assume $T \vdash \neg\rho$, thus

$$T \vdash (\forall y)(\beta(\overline{\rho},y) \rightarrow ((\exists z < y)\alpha(\overline{\rho},z))\,.$$

(Beneath each refutation of me there is a demonstration of me.) There is a d such that $T \vdash \beta(\overline{\rho})$; thus $T \vdash (\exists z < \overline{d})\alpha(\overline{\rho},z)$. Similarly as above we get $T \vdash \bigvee_{e<d}\alpha(\overline{\rho},\overline{e})$ and $T \vdash \bigwedge_{e<d}\neg\alpha(\overline{\rho},\overline{e})$, thus T is inconsistent. □

2.11 Theorem (Essential Undecidability of Arithmetic). Each consistent theory T containing Q is undecidable; i.e. the set of its theorems is not Δ_1.

Proof. This is because if T were decidable we could extend it to a Δ_1 theory T' which is complete; but this is impossible by 2.10. (Indeed, let $\varphi_0, \varphi_1, \ldots$ be a Δ_1 enumeration of sentences of T; define $\alpha_n = \varphi_n$ if $Con(T \cup \{\alpha_0, \ldots, \alpha_{n-1}\})$ and $\alpha_n = \neg\varphi_n$ otherwise; axioms of T' are all α_n. This usual completion procedure is easily seen to be Δ_1 if T is decidable.) □

2.12 Remark. There are many ways in which Gödel's incompleteness theorems can be generalized, strengthened or modified. We shall present some of them in this book. To close this subsection we prove a theorem showing the existence of a formula which is "as independent as possible" (called a flexible formula).

2.13 Definition. Let T contain Q. A formula $\varphi(x)$ with just one free variable is *flexible* over T if each elementary conjunction of formulas $\varphi(\overline{0}), \varphi(\overline{1}), \varphi(\overline{2}), \ldots$ is consistent with T. In more details, let $(0)\alpha$ be $\neg\alpha$ and $(1)\alpha$ be α; then for each string $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n$, the theory

$$T, (\varepsilon_0)\varphi(\overline{0}), (\varepsilon_1)\varphi(\overline{1}), \ldots, (\varepsilon_n)\varphi(\overline{n})$$

is consistent.

2.14 Lemma. Let T be a Σ_1 theory containing Q. Then there is a Σ_1 formula $\varphi(x)$ such that for each k, $(T + (\forall x)(\varphi(x) \equiv x = \overline{k})$ is consistent. (The additional axiom says that k is the unique element satisfying φ.)

Proof. Define a function F as follows: $F(\psi) = k$ iff ψ is a T-formula with just one free variable and there is a T-proof d of $\neg(\forall x)(\psi(x) \equiv x = \overline{k})$ such that for each $d' < d$ and each $k' \leq d'$, d' is not a T-proof of $\neg(\forall x)(\psi(x) \equiv x = \overline{k'})$ (thus d is minimal possible). Clearly, F is a partial Σ_1 function. By 1.25, let $\alpha(x, y)$ be such that if $F(\psi) = k$ then $T \vdash \alpha(\overline{\psi}, y) \equiv y = \overline{k}$. By the diagonal lemma, let $\varphi(y)$ be such that $T \vdash \varphi(y) \equiv \alpha(\overline{\varphi}, y)$. We claim that $F(\varphi)$ is undefined, which means that for each k, $(T + (\forall x)(\varphi(x) \equiv x = \overline{k}))$ is consistent. Assume not and let d be the minimal possible proof as above, d proving $\neg(\forall x)(\varphi(x) \equiv x = \overline{k})$. Then $T \vdash \alpha(\overline{\varphi}, y) \equiv y = \overline{k}$, thus $T \vdash (\forall x)(\varphi(x) \equiv x = \overline{k})$ and T is inconsistent. □

2.15 Theorem (on flexible formulas). Let T be Δ_1 theory containing Q. Then theree is a flexible Σ_1 formula.

Proof. Let $\beta(s, i)$ be a Σ_1 formula defining the i-th element of a sequence s; let $\varphi(x)$ be the formula from 2.14 and let $\psi(x)$ be

$$(\exists s)(\varphi(s) \,\&\, \beta(s, x) = 1).$$

Now let ε be a string $(\varepsilon_0, \varepsilon_1 \ldots, \varepsilon_n)$ and take $T + (\forall x)(\varphi(x) \equiv x = \overline{\varepsilon})$. This theory is consistent and proves $\psi(\overline{k}) \equiv \overline{\varepsilon}_k = 1$, i.e. proves $(\varepsilon_k)\psi(\overline{k})$ $(k = 0, \ldots, n)$. This completes the proof. □

(c) Gödel's Second Incompleteness Theorem

Our formulation of Gödel's first incompleteness theorem was rather general; it concerned any Σ_1 definition π of all T-provable formulas. In formulating the second Gödel's incompleteness theorem we shall first present a rather general formulation (2.21) but then shall immediately present a particular case (2.22) concerning theories in which syntax has been developed (i.e. theories containing $I\Sigma_1$). We derive the following consequences: Löb's theorem (2.25), the fact that *PA* is not finitely axiomatizable (2.24) and a stronger fact saying that, for each $n \geq 1$, $I\Sigma_n$ is not axiomatizable using Σ_{n+2} formulas (2.27).

2.16 Definition. Let T be a theory containing Q and let π be a Σ_1 definition of the set of all T-provable formulas. The ***provability conditions*** for π are the following conditions:

$$(1) \qquad T \vdash \varphi \text{ implies } T \vdash \pi(\overline{\varphi}),$$

$$(2) \qquad T \vdash \pi(\overline{\varphi}) \rightarrow \pi(\overline{\pi(\overline{\varphi})}),$$

$$(3) \qquad T \vdash \pi(\overline{\varphi}) \mathbin{\&} \pi(\overline{\varphi \rightarrow \psi}) \rightarrow \pi(\overline{\psi})$$

2.17 Lemma. Let T contain $I\Sigma_1$, let τ be a Σ_1 definition of T and let the predicates $Proof^{\bullet}_{\tau}(s,x)$ (s is a τ-proof of x) and $Pr^{\bullet}_{\tau}(x)$ (x is τ-provable) be as in I.4.3. Then $Pr^{\bullet}_{\tau}$ satisfies the provability conditions 2.16.

Proof. Observe that if τ is a Σ_1 definition of T then $Pr^{\bullet}_{\tau}$ is a definition of the set of all T-provable formulas and is Σ_1 in T. Thus if $T \vdash \varphi$ then the formula $Pr^{\bullet}_{\tau}(\overline{\varphi})$ is true and therefore T-provable, by Σ_1-completeness (see I.1.8). This proves (1). To prove (2) reason inside T and repeat the proof of (1), now using formalized Σ_1-completeness (see I.4.32). (3) is immediate from the definition of $Proof^{\bullet}_{\tau}$. □

2.18 Definition. Let T be a theory containing Q and let π be a definition of the set of all T-provable formulas. The ***consistency statement*** given by π is the formula $\neg\pi(\overline{0=1})$. We denote the last formula by Con_{π} or $Con(\pi)$.

2.19 Lemma. Let T and π be as above and assume that π satisfies the provability conditions (1) and (3) (see 2.16). Then, for any T-sentence φ, T proves $Con(\pi) \equiv (\neg\pi(\overline{\varphi}) \vee \neg\pi(\overline{\neg\varphi}))$.

Proof. By (1), $T \vdash \pi(\overline{\neg 0 = 1})$. Thus

$$T \vdash \neg Con(\pi) \rightarrow \pi(\overline{0=1}) \mathbin{\&} \pi(\overline{\neg 0 = 1}),$$
$$T \vdash \pi(\overline{0 = 1 \rightarrow (\neg 0 = 1 \rightarrow \varphi)})$$

(by (1)) and hence, by double use of (3), $T \vdash \neg Con(\pi) \rightarrow \pi(\overline{\varphi})$. Similarly, $T \vdash \neg Con(\pi) \rightarrow \pi(\overline{\neg\varphi})$.

Conversely, note $T \vdash \pi(\overline{\varphi \rightarrow (\neg\varphi \rightarrow 0 = 1)})$ by (1), thus double use of (3) gives $T \vdash (\pi(\overline{\varphi}) \,\&\, \pi(\overline{\neg\varphi})) \rightarrow \neg Con(\pi)$. □

2.20 Lemma. Now assume T includes $I\Sigma_1$ and let $Pr^{\bullet}_{\tau}$ be as above. If Con is the consistency statement given by $Pr^{\bullet}_{\tau}$ then

$$T \vdash Con \equiv Con^{\bullet}_{\tau},$$

where $Con^{\bullet}_{\tau}$ is as in I.4.7.

Proof. Fully analogous to the proof of 2.19, but now not as a schema but as a single theorem. □

2.21 Theorem. (Gödel's Second Incompleteness Theorem). Let T be a theory containing Q and let π be a Σ_1 definition of the set of all T-provable formulas satisfying the provability conditions.

Let ν be Gödel's sentence based on π. Then

$$T \vdash \nu \equiv Con(\pi)$$

and hence T does not prove $Con(\pi)$ provided T is consistent.

Proof. $T \vdash \nu \rightarrow \neg\pi(\overline{\nu})$, thus $T \vdash \nu \rightarrow Con(\pi)$. Conversely, $T \vdash \neg\nu \rightarrow \pi(\overline{\nu})$ and, by (2), $T \vdash \pi(\overline{\nu}) \rightarrow \pi(\overline{\pi(\overline{\nu})})$, i.e. $T \vdash \pi(\overline{\nu}) \rightarrow \pi(\overline{\neg\nu})$. Therefore we have $T \vdash \neg\nu \rightarrow (\pi(\overline{\nu}) \;\&\; \pi(\overline{\neg\nu}))$, i.e. $T \vdash \neg\nu \rightarrow \neg Con(\pi)$. □

2.22 Corollary. Let T be a theory containing $I\Sigma_1$ and let ν be Gödel's sentence based on $Pr^{\bullet}_{\tau}$, where τ is a Σ_1 definition of axioms of T. Then $T \vdash \nu \equiv Con^{\bullet}_{\tau}$; thus T does not prove $Con^{\bullet}_{\tau}$, provided T is consistent.

2.23 Remark. In the terminology of 2.5, Gödel's second incompleteness theorem gives us an example of a formula having a unique fixed point and such that this fixed point is an independent formula (neither provable nor refutable) – assuming that our theory T is Σ_1-sound.

On the other hand, if $\lambda(x)$ is a flexible formula then for any two different fixed points φ, ψ of λ, T does not prove $\varphi \equiv \psi$ (and it is easy to see from the proof of the diagonal lemma that each formula has infinitely many different fixed points).

2.24 Corollary. Peano arithmetic PA is not finitely axiomatizable.

Proof. If PA were finitely axiomatizable then for some k, PA would be equivalent to $I\Sigma_k$; but $PA \vdash Con(I\Sigma^{\bullet}_k)$ and, by 2.22, $I\Sigma_k$ does not prove $Con(I\Sigma^{\bullet}_k)$. □

2.25 Löb's Theorem. Let T be as in 2.22, i.e. T contains $I\Sigma_1$, τ is a Σ_1 definition of T. Assume that φ is a sentence such that $T \vdash Pr^\bullet_\tau(\overline{\varphi}) \rightarrow \varphi$. Then $T \vdash \varphi$.

Proof. By our assumption, $(T + \neg\varphi) \vdash \neg Pr^\bullet_\tau(\overline{\varphi})$, thus $(T + \neg\varphi) \vdash Con^\bullet_{\tau+\neg\overline{\varphi}}$ (where $(\tau + \neg\varphi)(x) \equiv \tau(x) \vee x = \overline{\varphi}$), thus by 2.22, $(T + \neg\varphi)$ is inconsistent. □

2.26 Corollary. Let T contain $I\Sigma_1$ and let κ be a fixed point of the formula $Pr^\bullet_\tau(x)$ (where τ is a Σ_1 definition of T), i.e. $T \vdash \kappa \equiv Pr^\bullet_\tau(\overline{\kappa})$. The formula κ is called Henkin's formula (and says "I am provable"). By Löb's theorem, $T \vdash \kappa$.

2.27 Theorem. For each $n \geq 1$, $I\Sigma_n$ is not axiomatizable by Σ_{n+2} formulas.

Proof. First show that $I\Sigma_n$ is not axiomatizable by Π_{n+1} formulas. If it were we would have a finite set S of Π_{n+1} formulas axiomatizing $I\Sigma_n$ (since the last theory is finitely axiomatizable, see I.2.52, but by I.4.33, $I\Sigma_n$ proves the consistency of the set of all true Π_{n+1} formulas and hence, by "snowing"-snowing, we would have $I\Sigma_n \vdash S^\bullet \subseteq Tr(\Pi_{n+1})$ ($S^\bullet$ is $\{\overline{\varphi}_1, \ldots, \overline{\varphi}_n\}^\bullet$ where S is $\{\varphi_1, \ldots, \varphi_n\}$). Furthermore, since finite axiomatizability of $I\Sigma_n$ is provable in $I\Sigma_1$, we would get $I\Sigma_1 \vdash Con^\bullet(S^\bullet) \rightarrow Con^\bullet(I\Sigma_n)$ and, by the above, $I\Sigma_n \vdash Con^\bullet(S)$. Thus we would get $I\Sigma_n \vdash Con^\bullet(I\Sigma^\bullet_n)$ which contradicts Gödel's second incompleteness theorem.

Now assume that S is a finite set of Σ_{n+2} formulas axiomatizing $I\Sigma_n$, $S = \{\varphi_1, \ldots, \varphi_n\}$, $\varphi_i = (\exists x)\psi_i(x)$, where ψ_i is Π_{n+1}. Let us work in $I\Sigma_n$. We have $Tr_{\Sigma_{n+1}}(S^\bullet)$ and $Con^\bullet(S^\bullet) \rightarrow Con^\bullet(I\Sigma^\bullet_n)$; thus for some $x_1, \ldots, x_n$ we have $\bigwedge_i Tr_{\Pi_n}(\overline{\psi}_i(x_i))$ and for $S^\bullet_1 = \{\overline{\psi}_1(x_1), \ldots, \overline{\psi}_n(x_n)\}$ we have $Con^\bullet(S^\bullet_1) \rightarrow Con^\bullet(S^\bullet)$. But $Con^\bullet(S^\bullet_1)$ follows by I.4.33. We have proved $I\Sigma_n \vdash Con^\bullet(I\Sigma^\bullet_n)$ and have again a contradiction with Gödel's second incompleteness theorem. □

2.28 Discussion. Till now we have worked with arbitrary theories containing Q or $I\Sigma_1$ and either have assumed nothing about the complexity of T as a set of formulas (thus as a set of natural number) or have assumed that T is Σ_1 (has a Σ_1 definition). We shall often need stronger assumptions on T; we shall now list and discuss four most frequent possibilities.

(a) T is Σ_1;
(b) T is Δ_1;
(c) T is $I\Sigma_1$-provably Δ_1;
(d) T is Σ^{exp}_0

Let us stress the fact that we identify a theory with the set of its axioms (not the set of provable formulas). It is easy to see that (d) $\Rightarrow$ (c) $\Rightarrow$ (b) $\Rightarrow$ (a); let us comment on this.

For Σ_0^{exp} formulas cf. I.1.28 and I.2.73; we allow bounded quantifiers of the form $(\forall x \leq y)$, $(\forall x \leq 2^y)$ and similarly for $\exists$. For provably Δ_1 sets cf. I.1.51; by I.1.50 and I.1.52 each provably Δ_1 set is Δ_1 and each Δ_1 set is Σ_1.

Σ_1 sets are often called recursively enumerable and Δ_1 sets are called recursive (cf. Sect. 0). We also know that each primitive recursive set is $I\Sigma_1$-provably Δ_1 and have promised to show the converse in Chap. IV by model-theoretic means (cf. I.1.54). Concerning Σ_0^{exp} sets, it is easy to prove that they coincide with $\Sigma_0^{exp}(exp)$ sets and with elementary recursive sets as defined e.g. in Grzegorczyk's book (but we shall not need this fact).

Define two theories T, S to be ***deductively equivalent*** if they prove the same theorems. (Thus deductively equivalent theories are just two different axiomatizations of a deductively closed set of formulas).

2.29 Craig's Theorem. Each Σ_1 theory T is deductively equivalent to a Σ_0^{exp} theory.

Proof. We show that T is deductively equivalent to a provably Δ_1 theory; the reader may show that the same construction gives in fact a Σ_0^{exp} theory. Let τ be a Σ_1 formula defining the set T of axioms, let τ be $(\exists y)\tau(x,y)$ where τ_o is Σ_o. Craig's trick is to replace each axiom φ by a sufficiently long conjunction

$$\underbrace{\varphi \,\&\, \varphi \,\&\, \ldots \varphi}_{n \text{ times}}$$

(denoted by $Repeat(n,\varphi)$ if the number of conjuncts is n) which makes it possible to bound the quantifier $(\exists y)$. Thus put

$$\delta(x) \equiv (\exists y \leq x)(\exists u, v \leq x)(x = Repeat^\bullet(v,u) \,\&\, \tau_o(u,y))$$

where $Repeat^\bullet$ is defined in the obvious way and is Δ_1 in $I\Sigma_1$, thus $\delta(x)$ is also Δ_1 in $I\Sigma_1$. The set S defined by $\delta(x)$ is an axiom system related to T as follows: $\varphi \in T$ iff, for some n, $Repeat(n,\varphi) \in S$. Clearly, S is deductively equivalent to T. $\square$

2.30 Remarks. (1) Observe that in fact we can prove more: if τ is a Σ_1-definition of T in $I\Sigma_1$ and δ is as above and defines S then $I\Sigma_1$ proves that the theories$^\bullet$ τ and δ are deductively$^\bullet$ equivalent$^\bullet$.

(2) A formalized version of Craig's theorem reads as follows: $I\Sigma_1$ proves that if T is a $\Sigma_1^\bullet$ theory$^\bullet$ then there is a $\Delta_1^\bullet$ theory S such that T and S are deductively$^\bullet$ equivalent$^\bullet$. (Proof obvious).

(3) Observe that if δ is Δ_1 in $I\Sigma_1$ and defines T then τ binumerates T in $I\Sigma_1$ (and hence in T). (This is because, by 1.23, if A is any set of natural

numbers and τ is its definition and is Δ_1 in a sound $S \supseteq Q$ then τ binumerates A in S.)

2.31 Further Remarks. (1) We know that each Δ_1 set of natural numbers has a Σ_1 binumeration (and a Π_1 binumeration) in T where T is any consistent Δ_1 theory containing Q. Note that a Δ_1 set may have, and in fact always has different Σ_1 binumerations whose equivalence is not provable in T: let $C \subseteq N$ and let $\gamma(x)$ be any Σ_1 binumeration of C. Let ρ be a Rosser sentence of T, let $(\exists y)\beta(x,y)$ be a Σ_1 definition of refutable formulas; put

$$\gamma_1(x) \equiv \gamma(x) \,\&\, (\forall y < x)\neg\beta(\overline{\rho}, y)$$
$$\gamma_2(x) \equiv \gamma(x) \vee (\exists y < x)\beta(\overline{\rho}, y)\,.$$

Clearly, $T \vdash \gamma_1(x) \rightarrow \gamma(x) \rightarrow \gamma_2(x)$; we show that both δ_1 and γ_2 binumerate C but T does not prove $\gamma_1(x) \equiv \gamma_2(x)$. Observe that for each k, $T \vdash \gamma_1(\overline{k}) \equiv \gamma(\overline{k}) \equiv \gamma_2(\overline{k})$ since for each k

$$T \vdash \neg(\exists y \leq k)\beta(\overline{\rho}, k)$$

($\neg\rho$ being unprovable), thus all three formulas binumerate the same set; but $(T + \rho)$ is consistent, i.e. $T + (\exists y)\beta(\overline{\rho}, y)$ is consistent; and in the last theory we prove $(\exists y)(\gamma_2(y) \,\&\, \neg\gamma_1(y))$. Thus T does not prove $\gamma_2(y) \rightarrow \gamma_1(y)$.

(2) Let us show that for each Σ_1-sound Δ_1 theory T containing Q there is a Δ_1 set A such that no formula δ which is Δ in T_1 binumerates A in T (thus the claim of (1) above cannot be improved to a binumeration which is Δ_1 in T). Let T be given and observe the following:

If T proves $(\forall x)(\delta(x) \equiv \pi(x))$ where $\delta \in \Sigma_1$ and $\pi \in \Pi_1$, then for each $k \in N$, $N \models (\delta(\overline{k}) \equiv \pi(\overline{k}))$ (i.e. $N \models \delta(\overline{k}) \vee \neg\pi(\overline{k})$ and $N \models \neg\delta(\overline{k}) \vee \pi(\overline{k})$ – Σ_1 soundness suffices), thus $N \models (\forall x)(\delta(x) \equiv \pi(x))$. Thus for each $k \in N$, either there is a witness for $\delta(\overline{k})$ or there is a witness for $\neg\pi(\overline{k})$.

Now define A as follows: if n is a T-proof of $(\forall x)(\delta(x) \equiv \pi(x))$ (where $\delta \in \Sigma_1, \pi \in \Pi_1$) then look for witnesses for $\delta(\overline{n})$ and $\neg\pi(\overline{n})$. If you find a witness for $\delta(\overline{n})$, define $n \notin A$; if you find a witness for $\neg\pi(\overline{n})$, define $n \in A$. For remaining n define e.g. $n \in A$.

The reader may check that this is a Δ_1 definition of a set A and that if n, δ, π are as above then A differs from the set defined by δ (and π) at least in n.

(3) By 1.23, if φ is a Σ_1 definition of a set A then φ numerates A in each Σ_1-sound theory $S \supseteq Q$. Let us show that the assumption of Σ_1-soundness in essential; in general, φ numerates a superset of A. (For example, let A be the set of axioms of PA, let $\pi(x)$ be the formula $x \in PA^\bullet$ (where $PA^\bullet$ is defined by copying the definition of PA in $I\Sigma_1$) and let $\varphi(x)$ be the formula $\pi(x) \vee (\neg Con^\bullet(PA^\bullet) \,\&\, x = \overline{0=1})$. Then π is Σ_1 in PA, numerates PA in PA but numerates $(PA + \overline{0=1})$ in the consistent (but Σ_1-ill) theory $(PA + \neg Con^\bullet(PA^\bullet))$.

(4) Returning again to Craig's theorem, let us observe the following: If τ is a Σ_1 definition of a theory T and δ is the Δ_1-in-$I\Sigma_1$ definition of the corresponding Craig theory S then (i) T is deductively equivalent to S, (ii) $I\Sigma_1$ proves that τ is deductively$^\bullet$ equivalent$^\bullet$ to δ, (iii) δ binumerates S in T (even in $I\Sigma_1$) but (iv) τ may numerate a proper extension of T in T.

(d) Pure Extensions of *PA*

In this and next subsection we shall deal with extensions of *PA* having the same language as *PA*. Here we show that each such theory T is essentially reflexive (2.35) and if T is Δ_1 then it has a Δ_2 definition π such that T proves $Con^\bullet_\pi$ (2.37). This last result shows some limitations of Gödel's second incompleteness theorem but it should not be overestimated; its main use is in technical proofs.

2.32 Definition. (1) For each theory T, let $T \restriction k$ be the set of all axioms of T less than k.

(2) As above, for each finite set $S = \{\varphi_1, \ldots, \varphi_n\}$, let $S^\bullet$ denote $\{\overline{\varphi}_1, \ldots, \overline{\varphi}_n\}^\bullet$.

(3) Let T be a theory extending $I\Sigma_1$. T is *reflexive* if for each k, T proves $Con^\bullet((T \restriction k)^\bullet)$ (i.e. for each finite subtheory T_o of T, T proves the consistency of T_o).

2.33 Definition. (1) A theory T' is a *pure* extension of T if T' extends T and has the same language as T.

(2) Let $T \supseteq I\Sigma_1$. T is *essentially reflexive* if each pure extension of T is reflexive.

2.34 Lemma. Let $T \supseteq I\Sigma_1$. T is essentially reflexive iff for each sentence φ, $T \vdash \varphi \rightarrow Con^\bullet(\{\varphi\}^\bullet)$.

Proof. If T is essentially reflexive and φ is a sentence then $(T+\varphi)$ is reflexive, thus $(T+\varphi) \vdash Con^\bullet(\{\varphi\}^\bullet)$, hence $T \vdash \varphi \rightarrow Con^\bullet\{\varphi\}^\bullet)$. Conversely, let $T \vdash \varphi \rightarrow Con^\bullet(\{\varphi\}^\bullet)$ for each φ and let T' be a pure extension of T; let φ be the conjunction of all axioms of $T' \restriction k$. Then the last provability immediately gives $T' \vdash Con((T' \restriction k)^\bullet)$. □

2.35 Theorem. *PA* is essentially reflexive.

Proof. Immmediately from I.4.34. □

2.36 Theorem. (1) Let T be any consistent pure extension of *PA*. Then for any k, T is not axiomatizable using only Σ_k formulas.

(2) For each $k \geq 1$, no consistent pure extension of $I\Sigma_k$ is axiomatizable using only Σ_{k+2} formulas.

Proof. Clearly, (1) follows from (2). The proof of (2) is a variant of the proof of 2.27: Let T be a pure extension of $I\Sigma_k$. First, if S is a finite set of Π_{k+1} formulas such that S proves $I\Sigma_k$ then we get $S \vdash Con^\bullet(S^\bullet)$ using I.4.34, hence S is inconsistent by Gödel's second incompleteness theorem, thus T is not Π_{k+1} axiomatizable (and (1) follows). The generalization to Σ_{k+2} formulas is as in 2.27, again getting $S \vdash Con^\bullet(S^\bullet)$, a contradiction. □

2.37 Theorem. Let $T \in \Delta_1$ be a pure extension of PA. Then there is a binumeration τ of T in T such that τ is Δ_2 in T and $T \vdash Con^\bullet(\tau)$.

Proof. By 1.23, T has a Σ_1 binumeration σ in T and a Π_1 binumeration π in T; let $\delta(x)$ be $\sigma(x)$ & $(\forall y \leq x)(\sigma(y) \equiv \pi(y))$. Evidently, δ is a Δ_2 binumeration of T in T. Let $\tau(x)$ be the formula $\delta(x)$ & $Con^\bullet(\pi \restriction x)$ (where $\pi \restriction x(y)$ is $\pi(y)$ & $y \leq x$). Then τ is another Δ_2 binumeration of T in T (since, thanks to reflexivity, for each k we have $T \vdash \tau(\overline{k}) \equiv \delta(\overline{k}) \equiv \sigma(\overline{k}) \equiv \pi(\overline{k})$). We prove $T \vdash Con^\bullet(\tau)$. Let us work in T. We distinguish two cases. First assume $Con^\bullet(\pi)$; then evidently $Con^\bullet(\tau)$ (since τ implies π). Second, assume $\neg Con^\bullet(\pi)$ and let z be the least number such that $\neg Con^\bullet(\pi \restriction (z+1))$. Then $Con^\bullet(\pi \restriction z)$ and for all x, $\tau(x)$ implies $(\pi \restriction z)(x)$. Thus $Con^\bullet(\tau)$ and the proof is complete. □

2.38 Remark. The reader has observed that τ just describes the largest initial segment of δ which is consistent; thus it is not much surprising that $Con^\bullet(\tau)$ is provable. Reflexivity guarantees that τ defines (and T-binumerates) T. Observe also that the formula $Con^\bullet(\tau)$ is Π_2 in T.

(e) Interpretability in Pure Extensions of *PA*

Our aim is to prove the following two theorems.

2.39 Theorem. Let $T \in \Delta_1$ be a pure extension of PA, let $S \in \Delta_1$ be another theory. Then the following are equivalent:
(i) S is interpretable in T.
(ii) S is locally interpretable in T, i.e. each finite subtheory of S is interpretable in T.
(iii) For each k, $T \vdash Con^\bullet((S \restriction k)^\bullet)$
(iv) There is a binumeration τ of S in T such that $T \vdash Con^\bullet(\tau)$.

2.40 Theorem. If T and S are as in 2.39 and S is also a pure extension of PA then S is interpretable in T iff S is Π_1 conservative for T.

2.41 *Proof of* 2.39. We prove (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) $\Rightarrow$ (i). (i) $\Rightarrow$ (ii) is evident. To prove (ii) $\Rightarrow$ (iii) it suffices to show that if S_o is finite and interpretable in T then $T \vdash Con^\bullet(S_o^\bullet)$. We apply the following

2.42 Lemma. Let S_o, T_o be finite theories, let T_o contain $I\Sigma_1$. If S_o is interpretable in T_o then $T_o \vdash Con^\bullet(T_o^\bullet) \rightarrow Con^\bullet(S_o^\bullet)$.

Proof. Let $intp(d)$ mean that d is a tuple containing T_o-formulas $\chi, \psi_P, \ldots, \psi_F, \ldots, \psi_c \ldots$ of respective arities (given by the language of S_o) as well as T_o-proofs of $(\exists x)\chi(x)$ and of the translations of the axioms of S_o. This is a finite sequence since S_o is finite. Clearly, $intp(d)$ is Δ_1; moreover a little checking shows that it is defined by a formula $intp^\bullet(x)$ which is Δ_1 in $I\Sigma_1$ (just formalize!). Similarly, let for each S_o-formula φ, $i(\varphi)$ be the translation of φ into a T_o-formula given by the above. Similarly, we have a definition $i^\bullet$, Δ_1 in $I\Sigma_1$, of i. Using Σ_1-induction we get $T_o \vdash Pr^\bullet_{S_o}(x) \rightarrow Pr^\bullet_{T_o}(i^\bullet(x))$; and similarly, T_o proves the properties of i concerning connectives. Thus $T_o \vdash Con^\bullet(T_o^\bullet) \rightarrow Con^\bullet(S_o^\bullet)$, as desired. □

2.43 *Proof of* 2.39 *continued.* The lemma gives $T \vdash Con^\bullet((S \restriction k)^\bullet)$ since if $S_o = S \restriction k$ and S_o is interpretable in T then S_o is interpretable in a finite subtheory T_o of T; and $T \vdash Con^\bullet(T_o^\bullet)$ due to reflexivity. This completes the proof of (ii) $\Rightarrow$ (iii).

Now we prove (iii) $\Rightarrow$ (iv). Assume $(\forall k) T \vdash Con^\bullet((S \restriction k)^\bullet)$ and let σ, π be Σ_1 and Π_1 binumerations of S in $I\Sigma_1$ (thus in T); let $\delta(x)$ be $\sigma(x)$ & $(\forall y \leq x)(\sigma(y) \equiv \pi(y))$ (cf. 2.37). Finally, let $\tau(x)$ be $\delta(x)$ & $Con^\bullet(\delta \restriction x)$. Then $T \vdash Con^\bullet(\tau)$ and τ binumerates S in T.

(iv) $\Rightarrow$ (i). We prove the arithmetized completeness theorem appropriately relativized. Let $T \vdash Con^\bullet(\tau)$ where τ is Σ_m (say), τ binumerates S in T. Then τ is Δ_{m+1} in T and therefore T proves that S has a full low Δ_{m+2} model. In fact, the proof of Low basis theorems gives as formulas $(\chi, \psi_S, \psi_+, \psi_*, \ldots)$ defining a low Δ_{m+2} model M of S in T. These formulas can be directly taken to define the desired interpretation i. (To see this prove a small "snowing"-snowing lemma for each S-formula ψ:

$$T \vdash x, \ldots, y \in M \rightarrow (\psi^i(x, \ldots, y) \equiv M \models \overline{\psi}[x, \ldots, y]))\,.$$

This completes the proof of 2.39. □

2.44 *Proof of* 2.40. First assume that S is interpretable in T; we prove Π_1-conservation. Let i be an interpretation of S in T. Thanks to the least number principle we may assume that i is absolute with respect to equality, i.e.

$$T \vdash \chi(x) \,\&\, \chi(y) \rightarrow x =^i y \equiv x = y\,.$$

(Otherwise define a factorization).

The interpretation defines, in T, a model $(M, S^i, +^i, *^i, 0^i, \leq^i)$ of the language of arithmetic (nothing is claimed on the existence of a full satisfaction!). Let us work in T. We show that there is an isomorphic embedding F of the universe onto an initial segment of M. (Cf. IV.1.3.) Let $pism(s)$ (partial isomophism) mean $Seq(s)$ & $(s)_o = \overline{0}_M$ & $(\forall_i < lh(s) - 1)((s)_{i+1} = S_M((s)_i))$. Show, by induction, that there are (uniquely determined) partial isomorphisms of arbitrary length; their union is the desired F. Verify by induction, that F preserves $+$ and $*$ (better: F is an isomorphic embedding of the universe into M) and that the range of F is an initial segment of M w.r.t. $\leq_M$. Then, outside T, show for each bounded formula we have

$$T \vdash \varphi(x, \ldots) \equiv \varphi^i(F(x), \ldots).$$

Thus if $(\forall x)\varphi(x)$ is Π_1 and $S \vdash (\forall x)\varphi(x)$ then $T \vdash ((\forall x)\varphi(x))^i$, which gives $T \vdash (\forall x)\varphi^i(F(x))$, thus $T \vdash (\forall x)\varphi(x)$.

Conversely, let T be Π_1-conservative for S. Since S is reflexive, we have, for each k, $S \vdash Con^\bullet((S \restriction k)^\bullet)$, thus $T \vdash Con^\bullet((S \restriction k)^\bullet)$, thus S is interpretable in T. □

3. Definable Cuts

Pure extensions of PA, studied in the preceding section, are never finitely axiomatized. In this section we shall investigate mainly finitely axiomatized theories containing $I\Sigma_1$ but possibly having a richer language than the language of arithmetic. These may be e.g. fragments $I\Sigma_n$, $B\Sigma_{n+1}$ $(n > 0)$ or ACA_o, Gödel-Bernays set theory and many others. Besides other things, we prove (in (c)) that a consistent finitely axiomatized sequential theory cannot prove full induction, i.e. there is a formula $\varphi(x)$ such that $T \vdash \varphi(\overline{0})$ & $(\forall x)(\varphi(x) \rightarrow \varphi(x+1))$ but T does not prove $(\forall x)\varphi(x)$. (Here x is a variable ranging over numbers, but φ may contain other variables and symbols of the language of T, e.g. in the case of ACA_0 φ may contain quantified set variables.) It follows then that there is a formula $I(x)$ such that T proves that $I(x)$ defines an initial segment (i.e. T proves $I(\overline{0})$, $(\forall x)(I(x) \rightarrow I(x+1)$, $(\forall x, y)(I(x)$ & $y \leq x \rightarrow I(y))$ but T does not prove $(\forall x)I(x)$. (Take $I(x)$ to be $(\forall y \leq x)\varphi(y)$.) We call such an I a proper definable cut. Definable cuts will play an important role in the present and the following section. Here, after proving an important theorem on shortening cuts (subsection (a)) we shall strengthen Gödel's second incompleteness theorem to a theorem saying (roughly) that if $T \supseteq I\Sigma_1$ is consistent and finitely axiomatizable and I is a definable T-cut then T cannot disprove the existence of a $T^\bullet$-proof$^\bullet$ z of contradiction, z belonging to I (subsection (b)). In subsection (c) we shall study an alternative notion of provability, called Herbrand provability (or direct provability) and show, for reasonable finitely axiomatized theories T, that there is a definable cut I such that T *does* prove that in I there is no Herbrand proof of a contradiction in T. Finally, in subsection (d) we prove a very useful criterion of interpretability of finitely axiomatized theories using definable cuts and Herbrand provability.

(a) Definable Cuts and Their Properties

3.1 Definition. Let T contain Q (the language of T may properly extend the language of arithmetic). A formula $I(x)$ with one free variable (understood as a number variable) is a *definable cut in* T (in short, a T-cut) if T proves $I(0)$, $(\forall x)(I(x) \to I(x+1))$ and $(\forall x, y)(y < x \;\&\; I(x) \to I(y))$. I is *proper* if T does not prove $(\forall x)I(x)$. (In the case of ACA_0, be aware of the fact that I is a formula, possibly with bound set variables; I is not a set variable.)

3.2 Remark. (1) Clearly, if I is a T-cut then, for each natural k, $T \vdash I(\overline{k})$.

(2) If $T \supseteq I\Sigma_n$ then clearly no Σ_n formula (and no Π_n formula) is a proper T-cut. But $I\Sigma_n$ has a proper Σ_{n+1} cut: let $n \geq 1$ and let $(\exists y)\psi(x, y)$ be a Σ_{n+1} formula for which T does not prove induction. Take the formula

$$(\exists s)(Seq(s) \;\&\; lh(s) = x \;\&\; (\forall i < x)\psi(i, (s)_i)\,.$$

It is an $I\Sigma_n$-cut and is Σ_{n+1} in $I\Sigma_n$. For $n = 0$ a similar proof works but one has to use a coding of sequences developed in Chap. V.

3.3 Definition. Define 2^x_y in $I\Sigma_1$ by the following recursion: for each x,

$$2^x_0 = x, \quad 2^x_{y+1} = 2^{2^x_y}\,.$$

Similarly, we define $\omega_y(0) = 0$ and for $x > 0$, $\omega_o(x) = 2x$, $\omega_{y+1}(x) = 2^{\omega_y(|x|-1)}$, where $|x|$ is the upper integral part of the binary logarithm of $x+1$, i.e. $y = |x|$ iff y is the least z such that $2^z > x$.

3.4 Lemma. $I\Sigma_1$ proves the following:

(1) $$(\forall y)(\omega_y(x) \leq 2^x)\,,$$

(2) $$(\forall y)2^{x+1}_{y+1} = \omega_y(2^x_{y+1})\,,$$

(3) $$\omega_y(x) \leq \omega_y(x+1)\,,$$

(4) $$\omega_y(x) \leq \omega_{y+1}(x)\,,$$

Proof elementary. Note that the present use of the symbol ω differs from its use in Chap. II (where we dealt with ordinals). Both uses are common in the literature and we prefer to be in accordance with it. □

3.5 Theorem. Let $T \supseteq I\Sigma_1$. For each $k \geq 0$ and each T-cut I there is a T-cut J such that

$$(\forall x \in J)(2_k^x \in I),$$
$$(\forall x \in J)(\omega_k(x) \in J).$$

Proof. (a) First we prove that for each T-cut I there is a T-cut J such that, in T, $J \subseteq I$ and J is closed under addition: just put

$$J(x) \equiv . I(x) \,\&\, (\forall y)(I(x+y)).$$

Clearly, J is a cut, and $J \subseteq I$; if $x, z \in J$ then, for each $y \in I$, $z + y \in I$ and therefore $x + z + y \in I$, thus $x + z \in J$. (Consequently, J is closed under ω_o.).

(b) For each n, there is a T-cut J_n such that $T \vdash J_n \subseteq T$ and $T \vdash (\forall x)(x \in J_n \rightarrow 2_n^x \in I)$. This is proved by induction on n: Assume J_n given. By shortening we may assume that J_n is closed under addition. Define $x \in J_{n+1} \equiv 2^x \in J_n$. Then in T, $x \in J_{n+1} \rightarrow 2^x \in J_n \rightarrow (2)_n^{2^x} \in I$ and J_{n+1} is a cut: if $x \in J_{n+1}$ then $2^{x+1} = 2^x + 2^x \in J_n$ thus $x + 1 \in J_{n+1}$.

(c) For each n, there is a T-cut K_n such that $T \vdash K_n \subseteq I$ and $T \vdash (\forall x)(x \in K_n \rightarrow \omega_n(x) \in K_n)$ (i.e. K_n is closed under ω_n). First let J_n be as in (c) and put

$$x \in K_n \equiv (\exists y \in J_n)(x \leq 2_n^y)$$

Then clearly $T \vdash J_n \subseteq K_n \subseteq I$; in T, if $x \in K_n$ and $x \leq 2_n^y$ for some $y \in J_n$ then $\omega_n(x) \leq \omega_n(2_n^y) = 2_n^{y+1}$ and $y + 1 \in J_n$, thus $\omega_n(x) \in K_n$. Finally, K_n is a cut: in T, if $x \in K_n$ and $x \leq 2_n^y$ for $y \in J_n$ then $x + 1 \leq 2_n^{y+1}$ (for each n, $I\Sigma_1$ proves $2_n^{y+1} \geq 2_n^y + 1$ – show by induction on n).

(d) Theorem 3.5 follows by (b) and (c). □

(b) A Strong Form of Gödel's Second Incompleteness Theorem

We shall investigate existence of proofs of a contradiction in a definable cut. Theorem 3.9 is a general theorem for $T \supseteq Q$ (not mentioning cuts), 3.11 is a consequence for $T \supseteq I\Sigma_1$ saying that, if consistent, T does not prove the consistency of T over a cut. Note that using the apparatus of Chap. V, 3.11 can be generalized for $T \supseteq Q$ (but a careful formulation is necessary, see V.5.28 (ii)). The key device for 3.9 is simultaneous use of two provability predicates.

3.6. Let $T \supseteq Q$ and let π_1, π_2 be two definitions of T-provable formulas. Generalized provability conditions for π_1 and π_2 over T are the following three conditions: For all φ,

(1) $$T \vdash \varphi \text{ implies } T \vdash \pi_1(\overline{\varphi}),$$

(2) $$T \vdash \pi_1(\overline{\varphi}) \to \pi_2(\overline{\pi_1(\overline{\varphi})}),$$

(3) $$T \vdash \pi_1(\overline{\varphi}) \,\&\, \pi_1(\overline{\varphi \to \psi}) \to \pi_2(\overline{\psi}).$$

3.7 Remark. Observe that (1) and (3) imply $T \vdash \pi_1(\overline{\varphi}) \to \pi_2(\overline{\varphi})$: take φ for ψ.

3.8 Definition. The consistency statement for π based on φ is the formula $\neg\pi(\overline{\varphi}) \vee \neg\pi(\overline{\neg\varphi})$. We may denote the last formula by $Conb(\pi, \varphi)$ (cf. 2.18).

3.9 Generalization of Gödel's Second Incompleteness Theorem. Let $T \supseteq Q$ be consistent and let π_1, π_2 satisfy the generalized provability conditions over T. Then there is a sentence φ such that T does not prove the consistency statement $Conb(\pi_2, \varphi)$.

Proof. Let ν be a fixed point of $\neg\pi_1(x)$, i.e. $T \vdash \nu \equiv \neg\pi_1(\overline{\nu})$. This is like Gödel's fixed point but now π_1 need not be Σ_1 and need not define T-provable formulas. Nevertheless, condition (1) is sufficient to show that T does not prove ν (cf. 2.8). Let φ be $\pi_1(\nu)$. We show $T \vdash \neg\pi_2(\overline{\varphi}) \to \nu$ and $T \vdash \neg\pi_2(\overline{\neg\varphi}) \to \nu$; thus T does not prove $\neg\pi_2(\overline{\varphi}) \vee \neg\pi_2(\overline{\neg\varphi})$.

First we prove $T \vdash \neg\nu \to \pi_2(\overline{\pi_1(\overline{\nu})})$. This is because $T \vdash \neg\nu \to \pi_1(\nu)$ and $T \vdash \pi_1(\nu) \to \overline{\pi_1(\overline{\nu})}$ (by (2)).

Second, we prove $T \vdash \neg\nu \to \pi_2(\overline{\neg\pi_1(\overline{\nu})})$. Indeed, $T \vdash \nu \to \neg\pi_1(\overline{\nu})$, thus $T \vdash \pi_1(\overline{\nu \to \neg\pi_1(\overline{\nu})})$ by (1); furthermore, $T \vdash \neg\nu \to \pi_1(\overline{\nu})$, which, together with the preceding provability gives $T \vdash \neg\nu \to \pi_2(\overline{\neg\pi_1(\overline{\nu})})$ by (3). This completes the proof. □

3.10 Definition. Let $T \supseteq I\Sigma_1$, let I be a T-cut and let τ be a Σ_0^{exp} definition of T. Then $Pr_\tau^{\bullet I}(x)$ is the formula

$$(\exists z)(I(z) \,\&\, Proof_\tau^\bullet(z, x))$$

(saying that there is a τ-proof$^\bullet$ of x in I).

(2) $Con_\tau^{\bullet I}$ is the formula $\neg(\exists z)(I(z) \,\&\, Proof_\tau^\bullet(z, \overline{0 = 1}))$.

3.11 Another Generalization of Gödel's Second Incompleteness Theorem. Let $T \supseteq I\Sigma_1$, let I be a T-cut and τ a Σ_0^{exp}-definition of T. Then T does not prove $Con_\tau^{\bullet I}$.

3.12 Remark. The rest of this subsection contains a proof of 3.11. Without any loss of generality we shall assume that T contains a unary function symbol for exponentiation (2^x) and the corresponding axioms. (It is easy to check that all uses of 2^x may be understood as abbreviations.)

3.13 Definition and Discussion. A function F of one argument is multi-exponentially bounded in $T \supseteq I\Sigma_1$ of there is a k such that $T \vdash (\forall x)(F(\alpha) < 2^x_k)$.

Let us indicate that natural functions describing syntax (as concatenation of two sequences, concatenation of a sequence of sequences, substitution etc.) are multi-exponentially bounded in $I\Sigma_1$. To simplify matters, we shall just show existence of multieponential bounds in the standard model N; the formalization is obvious.

(1) Call a function $F(m, \ldots, n)$ of several arguments multi-exponentially bounded if there is a k since that, for each q,

$$(m < q \,\&\, \ldots \,\&\, n < q) \rightarrow F(m, \ldots, n) < 2^q_k .$$

Observe that each term of the language $L_0(exp)$ (i.e. having the constant $\overline{0}$ and function symbols $S, +, *, 2^x$) defines a monotone and multi-exponentially bounded function. Write m.e.b. for "multi-exponentially bounded".

(2) Recall that the pairing function is polynominally bounded: $(x, y) \leq (x + y + 1)^2$. Furthermore, recall that our coding of finite sets of numbers (introduced in $I\Sigma_1$) satisfies $x \subseteq (< y) \rightarrow x < 2^y$ and $x \subseteq (< y) \times (< 2) \rightarrow x < 2^{(y+2+1)^2}$. This implies that if s is a sequence ($Seq(s)$), $lh(s) \leq x$ and each member of s is $< y$ then $s \leq 2^{(x+y+1)^2}$. If s, t are two such sequences then their concatenation $s \frown t$ satisfies $s \frown t \leq 2^{(2x+y+1)^2}$; if q is a sequence of sequences, $lh(q) = z$ and each member of q is a sequence satisfying the assumptions above then the concatenation of q (denoted $Concseq(q)$ in Chap. I) satisfies $Concseq(q) < 2^{(2x+y+1)^2}$. The concrete formulas are not too important; what is important is the fact that the bounds are multi-exponential. Now it is easy to see that the substitution function $Subst(\varphi, x, \tau)$ is m.e.b. and similarly for other syntactic notions.

3.14 Lemma. Let T and τ be as in 3.11, i.e $T \subseteq I\Sigma_1$, τ a Σ^{exp}_o-definition of T.

(a) If φ is Σ^{exp}_o then there is a k such that

$$T \vdash x < u \,\&\, \ldots \,\&\, y < u . \\ \rightarrow [\varphi(x, \ldots, y) \rightarrow (\exists z < 2^u_k)\ Proof^\bullet_\tau(z, \overline{\varphi}(\dot{x}, \ldots, \dot{y})] .$$

(b) Similarly, if I is a T-cut then there is a k such that

$$T \vdash I(x) \rightarrow (\exists z < 2^x_k)\ Proof^\bullet_\tau(z, \overline{I}(\dot{x})) .$$

Proof. (a) The proof is a re-examination and generalization of the proof of Σ_1-completeness of Q (I.1.8). For simplicity prove again a non-formalized version saying the following:

if $m < p \& \ldots \& n < p$ and $N \models \varphi(\overline{m}, \ldots, \overline{n})$ then threre is a T-proof of $\varphi(\overline{m}, \ldots, \overline{n})$ beneath 2^p_k (where k is a constant depending only on φ).

(1) First assume that φ is $\overline{m} + \overline{n} = \overline{m+n}$. The usual proof of this formula looks as follows:

$$
\begin{aligned}
&x + \overline{0} = x \\
&x + Sy = S(x + y) \\
&\overline{m} + \overline{0} = \overline{m} \\
&\overline{m} + S\overline{0} = S(m + \overline{0}) \\
&\overline{m} + S\overline{0} = \overline{m+1} \\
&\ldots \\
&\overline{m} + \overline{n} = S(\overline{m} + \overline{n-1}) \\
&\overline{m} + \overline{n} = \overline{m+n}
\end{aligned}
$$

It depends on details of your Hilbert-style formalism whether this is a proof as it stands or if you have to make some inessential modifications; but in any case, the length of the proof (i.e. the number of proof lines) is polynomial (here linear) in $\max(m, n)$ and the length of each row is polynomial (linear) in $\max(m, n)$. Thus, by the above the whole proof is m.e.b. in $\max(m, n)$.

Similarly for $\overline{m} * \overline{n} = \overline{m * n}$, $\overline{m} \neq \overline{n}$ (if $m \neq n$), and other cases (cf. I.1.8). Also the proof of $\overline{2^m} = 2^{\overline{m}}$ is easily estimated.

(2) Let $Val(t(m, \ldots, n))$ be the value of a closed term (possibly containing exponentiation). For a given t there is a k such that for $m, \ldots, n < q$ there is a $d < 2^q_k$ such that d is a T-proof of $t(\overline{m}, \ldots, \overline{n}) = \overline{Val(t(\overline{m}, \ldots, \overline{n}))}$.

To see this, first note that, by (1) in the proof of 3.13, $Val(t(\overline{m}, \ldots, \overline{n})$ is m.e.b., i.e., for some h, whenever $m, \ldots, n < q$ then $Val(t(\overline{m}, \ldots, \overline{n})) < 2^q_h$. Put $r = 2^q_h$. For simplicity, just take one example: let $t(\overline{m}, \overline{p}, \overline{n})$ be $(2^{\overline{m}} + \overline{n}) * \overline{p}$; we want to estimate a proof of $(2^{\overline{m}} + \overline{n}) * \overline{p} = \overline{(2^m + n) * p}$. By (1) here, we can successively produce proofs of

$$
\begin{gathered}
2^{\overline{m}} = \overline{2^m}, \\
\overline{2^m} + \overline{n} = \overline{2^m + n}, \\
(\overline{2^m + n}) * \overline{p} = \overline{(2^m + n) * p}.
\end{gathered}
$$

There is a common j such that there is a proof of each of these equalities beneath 2^r_j (since each argument involved is $\leq r$). A proof of the desired equality results by concatenating proofs of the equalities above and adding some few lines (instances of transitivity of $=$). Clearly, the whole proof is a m.e.b. function of r, i.e. of the initial arguments $m, \ldots, n$.

(3) It follows that if $\varphi(x, \ldots, y)$ is a true atomic or negated atomic formula then the function assigning to $m, \ldots, n$ the least T-proof of $\varphi(\overline{m}, \ldots, \overline{n})$ is m.e.b.

(4) Prove by induction the following: for each Σ_0^{exp}-formula $\varphi(x, \ldots, y)$, the function assigning to each $m, \ldots, n$ such that $\varphi(\overline{m}, \ldots, \overline{n})$ is true, a T-proof of $\varphi(\overline{m}, \ldots, \overline{n})$, and to each $m, \ldots, n$ such that $\varphi(\overline{m}, \ldots, \overline{n})$ is false, a T-proof of $\neg\varphi(\overline{m}, \ldots, \overline{n})$ is m.e.b. This is true for φ atomic (see (3)); the induction step for connectives is easy. It remains to handle bounded quantifiers $(\forall x \leq y)$ and $(\forall x \leq 2^y)$.

(5) Observe that T proves

$$(\forall x \leq \overline{m})\varphi(x, \overline{m}, \ldots, \overline{n}) \equiv \bigwedge_{j \leq m} \varphi(\overline{j}, \overline{m}, \ldots, \overline{n})$$

and we can find a proof of this equivalence by m.e.b. function (with arguments $m, \ldots, n$). Indeed, analysing I.1.6 (4) we see that finding a proof of

$$x \leq \overline{m} \equiv x = \overline{0} \vee \ldots \vee x = \overline{m}$$

is m.e.b. (in the argument m) and so are functions witnessing the following T-provabilities:

$$\begin{aligned}
&(\forall x \leq \overline{m})\varphi(x, \overline{m}, \ldots) \rightarrow \varphi(\overline{k}, \overline{m}, \ldots), \quad (k \leq m) \\
&(\forall x \leq \overline{m})\varphi(x, \overline{m}, \ldots) \rightarrow \bigwedge_{k \leq m} \varphi(\overline{k}, \overline{m}, \ldots), \\
&\varphi(\overline{k}, \overline{m}, \ldots) \rightarrow (x = k \rightarrow \varphi(x, \overline{m}, \ldots)), \\
&\bigwedge_{k \leq m} \varphi(\overline{k}, \overline{m}, \ldots) \rightarrow \bigvee_{k \leq m} x = b. \rightarrow \varphi(x, \overline{m}, \ldots), \\
&\bigwedge_{k \leq m} \varphi(\overline{k}, \overline{m}, \ldots) \rightarrow (x \leq \overline{m} \rightarrow \varphi(x, \overline{m}, \ldots)), \\
&\bigwedge_{b \leq m} \varphi(\overline{k}, \overline{m}, \ldots) \rightarrow (\forall x \leq \overline{m})\varphi(x, \overline{m}, \ldots)).
\end{aligned}$$

Similarly for $(\forall x \leq \overline{2^m})$, i.e. for $(\forall x \leq 2^{\overline{m}})$: provably a function m.e.b. in the argument 2^m (or even 2^m_k) is m.e.b. in m. This completes the proof of 3.14 (a).

Now we prove (b); this is much easier.

Assume that $I(x)$ is a T-cut; let d be a T-proof of $I(\overline{0})$ & $(\forall x)(I(x) \rightarrow I(x+1))$. Let m be given; a proof of $I(\overline{m})$ consists of proofs of $I(\overline{0}), I(\overline{1}), \ldots, I(\overline{m})$:

$$\begin{aligned}
&\ldots\\
&I(\overline{0})\\
&I(x) \to I(x+1)\\
&I(\overline{0}) \to I(\overline{1})\\
&I(\overline{1})\\
&\ldots\\
&I(\overline{m-1}) \to I(\overline{m})\\
&I(\overline{m})\,.
\end{aligned}$$

It is easily seen that this is a m.e.b. function of m, I, d, i.e. for fixed I and d, a m.e.b. function of m. This completes the proof of 3.14, except for the following pedantical

Remark: In proving (a) we in fact assumed $T \supseteq I\Sigma_1(exp)$, i.e. used explicitly the power-of-two operation. If $T \supseteq I\Sigma_1$ but does not have 2^x in the language then we may extend it by adding the definition of 2^x; this is a particular conservative extension and one can check that there is a m.e.b. function F assigning to each $T(exp)$-proof of a T-formula φ a T-proof of φ. This shows how to get rid of 2^x. □

3.15 Lemma. Let T, τ be as above. For each T-cut I there is a T-cut J such that $T \vdash J \subseteq I$ and $Pr_\tau^{\bullet J}$, $Pr_\tau^{\bullet I}$ satisfy the provability conditions 3.6.

Proof. Let us write Pr^I instead of $Pr_\tau^{\bullet I}$ and similarly for Pr^J. The condition (1) is evidently satisfied: if d is a T-proof of φ then $T \vdash Proof_\tau^\bullet(\overline{d}, \overline{\varphi})$ and $T \vdash I(\overline{d})$. To get (3) it suffices to have $T \vdash J(x) \to I(2_n^x)$ where, provably in T, $u, v, w \subseteq x \to u \frown v \frown w \le 2_n^x$. We prove (2). Let us work in T.

Assume $Pr^J(\overline{\varphi})$, i.e., $J(x)$ & $Proof^\bullet(\overline{\varphi}, x)$. Then, by 3.14, for an appropriate j given from outside there are $z_1, z_2 < 2_j^x$ such that $Proof^\bullet(\overline{J(\dot{x})}, z_1)$ and $Proof^\bullet(\overline{Proof^\bullet(\overline{\varphi}, \dot{\overline{x}})}, z_2)$. Then $z = z_1 \frown z_2 \frown \langle \overline{Pr^J(\overline{\varphi})} \rangle$ (or something very similar, details are unimportant) is a τ-proof$^\bullet$ of $Pr^J(\overline{\varphi})$; thus it suffices that z is in I.

We see that for example putting $k = h + j$ and choosing J such that $T \vdash J(x) \to I(2_k^x)$ (by 3.5), we have (1), (2), (3). This completes the proof. □

3.16 *Proof of* 3.11-*conclusion.* Let T, τ, I be as in 3.11 and let J_o be the cut constructed in 3.15. Then for some particular φ T does not prove $Conb(\tau, \varphi)$. To get the unprovability of $\overline{0=1}$ in the desired cut J, we shorten J_o again in such a way that T proves the following: if $x \in J$ is a τ-proof of φ and $y \in J$ is a τ-proof of $\overline{\neg\varphi}$ then the concatenation of x, y, and the (standardly long) propositional proof of $\overline{0=1}$ from φ and $\neg\varphi$ is in J_o (use e.g. the h above). This completes the proof of 3.11. □

(c) Herbrand Provability and Herbrand Consistency

In this subsection we are going to investigate an alternative notion of provability called Herbrand provability (since it is based on Herbrand's theorem). Even if Herbrand provability is equivalent to the usual provability (provably in $I\Sigma_1$), Herbrand proofs are much "slower" (or: longer) than the usual Hilbert style proofs; we shall show that under some conditions on a theory T, we may always find a cut I in T such that T proves that in I there is no Herbrand proof of inconsistency in T. This is interesting at its own; but the methods we present here will be very useful in studying interpretability and partial conservativity in the next section.

Recall Herbrand's theorem (0.21 and, formalized, I.4.15): it says that a formula Φ is provable (in predicate calculus) iff there is a disjunction D of instances of the open part of the Herbrand form $He(\Phi)$ of Φ such that Φ is a propositional tautology. This leads to the following

3.17 Definition ($I\Sigma_1$). A *H-proof* (Herbrand proof) of a formula x is a propositional proof z of a disjunction of instances of the open part of $He^\bullet(x)$. (Notation: $HProof^\bullet(z,x)$.) If T is a finite theory consisting of closed formulas then a H-proof of x in T is a H-proof z of $(\bigwedge T \to x)$ ($\bigwedge$ standing for a finite conjunction); notation $HProof^\bullet_T(z,x)$. A formula x is H-provable in T ($HPr^\bullet_T(x)$) if there is a H-proof of x in T. T is H-consistent ($HCon^\bullet(T)$) if there is no H-proof of $\neg \bigwedge T$.

3.18 Remark. We know that $I\Sigma_1$ proves Herbrand's theorem, i.e. $I\Sigma_1 \vdash HPr^\bullet_T(x) \equiv Pr^\bullet_T(x)$ and $I\Sigma_1 \vdash HCon^\bullet(T) \equiv Con^\bullet(T)$. But for weaker theories not proving Herbrand's theorem, these notions may differ. This will be discussed in Chap. V; here we pay attention to the fact that the notions of provability and H-provability may differ on a definable cut, i.e. there may be a cut provably not containing a H-proof of a given formula (expressing consistency) but the same cannot be proved for usual Hilbert-style proofs.

3.19 Definition. Let T be a finite theory containing $I\Sigma_1$ (i.e. we assume that T proves all the axioms of $I\Sigma_1$) and let I be a T-cut. We make the following definitions:

$$HPr^{\bullet I}_T(x) \equiv (\exists y)(I(y) \,\&\, HProof^\bullet_T(x,y))\,,$$
$$HCon^{\bullet I}_T \equiv \neg HPr^{\bullet I}(\neg \bigwedge T^\bullet)\,.$$

We also write $HCon^{\bullet I}(T^\bullet)$.

3.20 Theorem. Let T be a finite theory in the language of arithmetic extending $I\Sigma_1$. (1) There is a T-cut I such that

$$T \vdash HCon^{\bullet I}(T^\bullet)\,.$$

(2) More generally, there is a T-cut I such that T proves

$$(\forall u\ \Sigma_1\text{-sentence}^\bullet)(\mathit{Tr}_{\Sigma,1}(u) \rightarrow \mathit{HCon}^{\bullet I}(T^\bullet + u)).$$

3.21 Theorem. Still more generally, let T be a finitely axiomatized sequential theory containing $I\Sigma_1$ (i.e. an interpretation of $I\Sigma_1$ in T has been fixed). Then there is a T-definable cut I such that T proves

$$(\forall u\ \Sigma_1\text{-sentence}^\bullet)(\mathit{Tr}_{\Sigma,1}(u) \rightarrow \mathit{HCon}^{\bullet I}(T^\bullet + u)).$$

Let us make a definition and describe the method of proof; then we shall elaborate the proof in details.

3.22 Definition. For each k we define Q_k-formulas as certain formulas of the language of arithmetic. Q_0-formulas are Σ_0-formulas; Q_{k+1}-formulas are Q_k-formulas and boolean combinations of formulas of the form $(\exists x)\varphi$ where φ is Q_k (say, boolean combinations of $\exists Q_k$-formulas). Thus a Q_k-formula has nesting of quantifiers at most k, Σ_0-formulas being disregarded. A *Q_k-proof* is a (Hilbert style) proof consisting only of Q_k-formulas.

3.23 Remark. Even if a detailed proof of our theorem is rather long, the idea is easily comprehensible. We shall prove 3.20 (2); then we indicate how to get 3.21. First we show that for each k, satisfaction and truth for Q_k-formulas is definable in $I\Sigma_1$. It follows that there is a T-cut J such that Q_k-proofs lying in J preserve truth. Consequently, no Q_k-proof of $\neg \bigwedge T^\bullet$ is in J. Finally, it can be shown that each H-proof of a Q_k-formula can be transformed, in a multi-exponentially bounded way, to a Q_{k+2}-proof of the same formula. Thus if I is a suitable shortening of J then I has the desired properties.

3.24 Corollary. Let $T \supseteq I\Sigma_1$ be consistent, finite and sequential. Then T does not prove full induction.

Proof. If T proves full induction then for each T-cut I, T proves $(\forall x)I(x)$, i.e. there are no proper T-cuts. On the other hand, if T is as above, then $T \vdash \mathit{HCon}^{\bullet I}(T^\bullet)$ for some T-cut I, thus $T \vdash \mathit{HCon}^\bullet(T^\bullet)$. But since $T \supseteq I\Sigma_1$, T proves Herbrand's theorem and hence $T \vdash \mathit{Con}^\bullet(T^\bullet)$,which contradicts Gödel's second incompleteness theorem. □

The rest of the subsection elaborates 3.23.

*

3.25 Lemma. For each k, the following can be constructed in $I\Sigma_1$: a definable cut I_k and a satisfaction for Q_k-formulas from I_k.

Proof. We know that there is a Δ_1 satisfaction for Σ_0-formulas (see I.1.75); thus let I_0 be an improper cut. Assume $Sat_{Q,k}$ is a satisfaction for Q_k-formulas from I_k (i.e. Tarski's conditions are provable and, consequently, "it's-snowing"-it's-snowing lemma is provable). In the usual way we get the satisfaction $Sat'(z,e)$ for $\exists Q_k$-formulas from I_k. It remains to extend the satisfaction to boolean combinations of such formulas, but this is easy. Define $Sat_{Q,k+1}(z,e)$, where z is a boolean combination of $\exists Q_k$-formulas, iff there exists a "constructing sequence" c of z from $\exists Q_k$-formulas (each member is either a $\exists Q_k$-formula or results from some preceding members using connectives; the last member is z) and a corresponding "evaluating sequence" s of zeros and ones such that, if $(c)_i$ is $\exists Q_k$ then $(s)_i = 1$ iff $Sat'(z,e)$, s obeys truth tables and the last member of c is 1. Define $I_{k+1}(x)$ iff $I_k(x)$ and for each Q_{k+1}-formula $z \leq y$ there exists a constructing sequence c with a unique evaluating sequence s. Clearly, I_{k+1} is an $I\Sigma_1$-cut. □

3.26 Lemma. For each k, there is an $I\Sigma_1$-cut $I(x)$ that is a shortening of I_k and such that $I\Sigma_1$ proves the following: If T is a finite theory in the language of arithmetic whose axioms are true Q_k-formulas$^\bullet$, if s is a Q_k-proof$^\bullet$ from T and $I(d)$ then each member of d is true.

Proof. This is trivial: let $I(x)$ say that $I_k(x)$ and each Q_k-proof from true Q_k-formulas is truth-preserving. This is an inductive property since deduction rules are truth preserving and logical axioms are true. □

3.27 Remark. If $T \supseteq I\Sigma_1$ is a finite theory in the language of arithmetic then for some k, all axioms of T are Q_k and T proves all its axioms true (by "it's-snowing"-it's-snowing). Thus it follows that in the cut I from 3.26 there is no Q_k-proof of $\neg \bigwedge T$. More than that: if u is $\Sigma_1^\bullet$ and true then it is a true Q_k formula (since $\Sigma_1 \subseteq Q_1$ and thus I does not contain any Q_k-proof of $\neg \bigwedge T \vee u$). To complete the proof of 3.20 it suffices to clarify the relation of H-proofs to Q_k-proofs. Before doing that let us introduce a technical device, analogous to a Henkin extension.

3.28 Definition ($I\Sigma_1$). Take the predicate calculus with the language of arithmetic; using the method of I.4, add infinitely many constants in such a way that for each formula $\varphi(x)$ of the enriched langauge with just one variable we have a constant $c_{(\forall x)\varphi(x)}$ (associated to φ by a total Δ_1 function) and call the formula

$$\varphi(c_{(\forall x)\varphi(x)}) \to (\forall x)\varphi(x)$$

the *special axiom* for $\varphi(x)$.

3.29 Remark. (1) Analogously to I.4.9 one shows that this is a conservative extension of the predicate calculus, see also below.

(2) The following theorem is in fact a strengthening of one implication from Herbrand's theorem; namely, it shows that Herbrand provability implies provability. But we need more: Q_k-provability and multi-exponential growth.

3.30 Theorem. For each $k \geq 1$, there is a q such that $I\Sigma_1$ proves the following: There is a Δ_1 function $F(d)$, majorized by 2^d_q, such that if d is a H-proof$^\bullet$ of a Q_k-formula$^\bullet$ φ then $F(d)$ is a Q_{k+2}-proof$^\bullet$ of φ.

The rest of this subsection contains a proof. We shall carefully describe the construction of the proof $F(d)$ and check that it is a Q_{k+2}-proof$^\bullet$; the tiresome task of checking that F is m.e.b. is largely left to the reader. The proof is an inspection and elaboration of a proof from Shoenfield's book. (Alternatively, the reader may apply Theorem V.5.14)

3.31. First, let us analyse a little bit the construction of a prenex normal form of a given formula (such a construction is the initial part of the construction of the Herbrand variant, cf. 0.18). Let Φ be given; we may assume that negation occurs only before atomic formulas, distinct quantifiers bind distinct variables and no variable is both free and bound in Φ. (If not then the corresponding changes as well as the proof of equivalence of both forms are given by a m.e.b. Δ_1 function.) Call, for a moment, a formula *clean* if it satisfies our assumptions. A clean formula is a boolean combination (using &, $\vee$, $\rightarrow$) of atoms, negated atoms and quantifies formulas, say $\beta(\varphi_1, \ldots, \varphi_n)$. If φ_1 is $(\Box x)\alpha$ (where $\Box$ is $\forall$ or $\exists$) then $\beta(\varphi_1, \ldots, \varphi_n)$ is provably equivalent to $(\Box' x)\beta(\alpha, \varphi_2, \ldots, \varphi_n)$ where $\Box'$ is $\forall$ or $\exists$; there is a m.e.b. function finding $\Box'$ and giving a proof of the equivalence. Call the transition from $\beta(\varphi_1, \ldots, \varphi_n)$ to $(\Box' x)\beta(\alpha, \ldots, \varphi_n)$ the *extraction* of $\Box$ from $(\Box x)\alpha$ in $\beta(\varphi_1, \ldots, \varphi_n)$, the inverse transition is the *insertion* of $\Box'$ into $\beta(\alpha, \ldots, \varphi_n)$ by quantifying α. A prenex normal form (*PNF*) of Φ is constructed by a sequence of extractions of quantifiers:

$$\Phi = \Phi_0 = (\Box_0 \ldots)\Psi_0 \quad \text{(where } \Box_0 \text{ is an empty block of quantifiers)}$$

$$\ldots$$

$$\Phi_h = (\Box_h \ldots)\Psi_h \quad (\Box_h \text{ is a block of quantifiers and } \Psi_h \text{ is open});$$

here for each $i < k$, $(\Box_{i+1} \ldots)$ is either of the form $(\Box_i \ldots)(\forall x)$ or of the form $(\Box_i \ldots)(\exists x)$ and the formula $(\forall x)\Phi_{i+1}$ (or $(\exists x)\Psi_{i+1}$) results from Ψ_i by an extraction of a quantifier. Φ_h is a *PNF* of Φ; write $(\Box \ldots)K$ instead of $(\Box \ldots)\Psi_h$.

3.32. The Herbrand normal form $He(\Phi)$ can be written as $(\exists \ldots)K^*$ where $(\exists \ldots)$ is the block of existential quantifiers resulting from $(\Box \ldots)$ in $(\Box \ldots)K$ by deleting universal quantifiers and K^* results from K by substitution of certain terms (containing new function symbols) for variables that were universally quantified in $(\Box \ldots)K$ (call these variables $\forall$-variables and the others $\exists$-variables). Now, by 3.17, a Herbrand proof of Φ is a propositional proof of a disjunction $\bigvee_i K(\mathbf{t}_i)$, where each $\mathbf{t}_i$ is a tuple of terms and $K(\mathbf{t}_i)$ is in fact an instance of K^*. We may assume that the disjunction $\bigvee_i K(\mathbf{t}_i)$ contains no variables.

3.33. We shall successively eliminate new function symbols, replacing them by special constants. Write $\mathbf{t}_i^0$ for $\mathbf{t}_i$; we shall construct a sequence of propositional tautologies

$$\bigvee_i K(\mathbf{t}_i^0),$$
$$\dots$$
$$\bigvee_i K(\mathbf{t}_i^p);$$

each $K(\mathbf{t}_i^j)$ will be an instance of K (but not necessarily of K^*). The transition from j to $j+1$ proceeds as follows:

Take a term $f(\mathbf{a})$ occuring in $\bigvee_i K(\mathbf{t}_i^j)$ where f is a Herbrand function (not occuring in Φ) and the terms $\mathbf{a}$ contain no Herbrand functions. Replace $f(\mathbf{a})$ (in all occurrences in $\bigvee_i K(\mathbf{t}_i^j)$) by the special constant c defined as follows:

The function f corresponds to a quantifier $(\forall x)$ in $PNF(\Phi)$ that was extracted when going from $(\Box_u \dots)\Psi_u$ to $(\Box_u \dots)(\forall x)\Psi_{u+1}(x, \mathbf{y})$; c is the special constant for $(\forall x)\Psi_{u+1}(x, \mathbf{a})$, i.e. the corresponding special axiom is

$$\Psi_{u+1}(c, \mathbf{a}) \to (\forall x)\Psi_{u+1}(x, \mathbf{a}).$$

This change clearly preserves the property of being a tautology (since it commutes with all connectives). We finally arrive at a tautology $\bigvee_i K(\mathbf{t}_i^p)$ which does not contain any new function. Its propositional proof is obtained from the original propositional proof by a m.e.b. function.

3.34. We want to show that each disjunct $K(\mathbf{t}_i^p)$ implies our original formula Φ by a Q_{k+2}-proof using possibly special axioms for constants used in the construction. Write $K(\mathbf{s})$ for $K(\mathbf{t}_i^p)$. Recall that $K(\mathbf{s})$ is in fact $\Psi_h(\mathbf{s})$ (see the construction of $PNF(\Phi)$). Now Φ_h is $(\Box_h \dots)\Psi_h$ and

$$(\Box_h)\Psi_h = \begin{cases} (\Box_{h-1})(\forall x)\Psi_{h-1}(x, \mathbf{y}) & \text{(Case 1)} \\ (\Box_{h-1})(\exists x)\Psi_{h-1}(x, \mathbf{y}) & \text{(Case 2)} \end{cases}$$

where $\Psi_{h-1}(\mathbf{y})$ results from $(\Box x)\Psi_h(x, \mathbf{y})$ by inserting $(\Box x)$ in its place.

Case 1. The variable x corresponds in $\mathbf{s}$ to a special constant c belonging to the special axiom $\Psi_h(c, \mathbf{s}') \to (\forall x)\Psi_h(x, \mathbf{s}')$ (where $\mathbf{s}'$ is the rest of $\mathbf{s}$, i.e. $\mathbf{s}$ is $(c, \mathbf{s}')$). Denote this axiom by $Sp(c)$. Thus we have a proof of

$$Sp(c) \,\&\, \Psi_h(c, \mathbf{s}') \to \Psi_{h-1}(\mathbf{s}')$$

and the proof is a Q_{k+1}-proof (since each Ψ_i is a Q_k-formula; Sp has one more unbounded quantifier and a proof of the equivalence of $(\forall x)\Psi_h(x, \mathbf{s}')$ and $\Psi_{h-1}(\mathbf{s}')$ uses only Q_{k+1}-formulas).

Case 2. Here let $\mathbf{s}$ be $(d, \mathbf{s}')$ where d is some term corresponding to x; clearly, the formula $\Psi_h(d, \mathbf{s}') \to (\exists x)\Psi_h(x, \mathbf{s}')$ has a Q_{k+1}-proof and the same holds for $\Psi_h(d, \mathbf{s}') \to \Psi_{h-1}(\mathbf{s}')$.

3.35. Iterating this we finally construct a Q_{k+1}-proof of

$$\bigwedge_j Sp(c_j) \,\&\, K(\mathbf{s}) \to \Phi$$

(since Φ is Ψ_0). Since $K(\mathbf{s})$ was an arbitrary disjunct of the tautology $\bigvee_i K(\mathbf{t}_i^p)$, we have a Q_{k+1}-proof of

$$\bigwedge_j Sp(c_j) \to \Phi\,.$$

(The reader should to check the fact that the construction is a m.e.b. function of the original Herbrand proof.) It remains to eliminate of the special constants.

3.36. Assume that the constants c_j are ordered according to decreasing complexity, i.e. c_1 does not occur in axioms $Sp(c_2), \ldots$ etc. Let the special axiom for c_1 be $\varphi(c_1) \to (\forall x)\varphi(x)$ and replace c_1 in the formula

$$Sp(c_1) \to (Sp(c_2) \to \cdots \to \Phi)$$

by a new variable y; we get a Q_{k+2}-proof of each of the following:

$$\begin{gathered}
(\varphi(y) \to (\forall x)\varphi(x)) \to (Sp(c_2) \to \cdots \Phi)\,,\\
(\exists y)(\varphi(y) \to (\forall x)\varphi(x)) \to (Sp(c_2) \to \cdots \Phi)\,,\\
((\forall y)\varphi(y) \to (\forall x)\varphi(x)) \to (Sp(c_2) \to \cdots \Phi)\,,\\
(Sp(c_2) \to \cdots \Phi)\,;
\end{gathered}$$

the last formula is Q_{k+1}. By iterating this we get a Q_{k+2}-proof of Φ, q.e.d.

3.37 Remark. Note that our main trick (not necessary in Shoenfield's original proof) was to keep the complexity of the formulas involved low; $PNF(\Phi)$ need not be Q_k, but we avoid reaching $PNF(\Phi)$ by carefully inserting each quantifier in its place immediately after it has been introduced. This makes the quantifier bounded as soon as possible.

3.38 Remark. Now our proof of 3.20 has been completed (cf. 3.23). Let us indicate how to modify the whole proof in order to get a proof of Theorem

3.21. The problem is that our fixed interpretation of $I\Sigma_1$ in T may be *relative*, i.e. there may be object being non-numbers. But thanks to sequentiality, we can code sequences of arbitrary objects.

We shall sketch the general proof, the reader may elaborate details.

(1) First assume that the language of T is rich enough and contains a unary predicate $N(x)$ ranging over number (in the sense of the fixed interpretation of $I\Sigma_1$ in T), a constant $\overline{0}$, function symbols S, $+$, $*$ and a predicate $\leq$ having the obvions meaning in the sense of the interpretation. (If this is not the case, replace T by a conservative extension T' using the respective definitions; show that each T-cut I has a shortening J such that

$$T \vdash HCon^{I}(T') \rightarrow HCon^{J}(T).)$$

(2) Define Q_o-formulas as boolean combinations of arithmetical Σ_o-formulas and arbitrary atomic formulas; then define Q_{k+1} formulas from Q_k formulas as above.

(3) We have two notions of sequences: the arithmetical notion for sequences of numbers and the notions of sequences of arbitrary objects given by the fact that T is sequential. Observe that there is a T-cut I_o such that, roughly, T proves that for lengths from I_o, the two notions of sequences of numbers coincide; in more detail, let $Seq(x)$, $(x)_y$ have its usual meaning and let $SEQ(z)$, $\beta(u,v,z)$ be as in 1.12.

Put

$$\begin{aligned} I_o(v) \equiv N(v) \,\&\, [((\forall x)Seq(x) &\rightarrow (\exists z)(SEQ(z) \\ &\& (\forall w < v)(\forall u)((x)_w = u \equiv \beta(u,w,z)) \\ &\& (\forall z)(SEQ(z) \,\&\, (\forall v < w)(\forall u)(\beta(u,w,z) \rightarrow N(u)) \\ &\rightarrow (\exists x)(Seq(x) \,\&\, (\forall w < v)(\forall u)((x)_w = u \equiv \beta(u,w,z))] \,. \end{aligned}$$

Then I_o is the desired cut.

(4) We can shorten I_o to get a cut I_1 such that, for T-terms$^\bullet$ from I_1, their value is uniquely determined.

Define evaluation$^\bullet$ of variables of a term$^\bullet$ as a particular SEQuence of objects (somehow assigning meaning to the variables$^\bullet$ of the term$^\bullet$) and define a corresponding evaluating sequence for the derivation of our term$^\bullet$; I_1 is the collection of all $u \in I_o$ such that, for each term$^\bullet$ $t \leq u$ and each evaluation$^\bullet$ of its variables, each derivation$^\bullet$ of t has a unique evaluating sequence. I_1 is a cut.

(5) By a possible further shortening we get a cut I_2 such that T proves: for terms$^\bullet$ of the language of arithmetic and their evaluations by numbers, the new notion of the value of a term coincides with the old one (cf. I.1.64).

(6) We define satisfaction for atomic formulas $z \in I_2$ of the whole language and for $\Sigma_o^\bullet$-formulas$^\bullet$ $z \in I_2$ of the language of arithmetic in the obvious manner. There is a shortening J_o of I_2 such that T proves that for Q_o-formulas $z \in J_o$, satisfaction is uniquely determined and satisfies Tarski's

truth conditions. This can be iterated: for each k, there is a cut J_k and a satisfaction for Q_k-formulas$^\bullet$ such that T proves its usual properties.

(7) Given k and Q_k, J_k has a shortening K_k such that T proves that Q_k-proofs from true formulas are truth-preserving. Then we can continue as in the proof 3.20: there is a m.e.b. function transforming each H-proof$^\bullet$ of a Q_k-formula$^\bullet$ into a Q_{k+2}-proof of that formula. This completes our proof sketch of 3.21.

(d) Cuts and Interpretations

This subsection contains a proof of the following theorem.

3.39 Theorem. Let S, T be theories, finitely axiomatized and $T \supseteq I\Sigma_1$ and sequential. Then S is interpretable in T iff there is a T-cut I such that $T \vdash HCon^I(S^\bullet)$.

Proof. First let us discuss the implication $\Rightarrow$. The proof is a generalization of the proof of 3.20, thus a modification of our proof of 2.21. We describe the necessary changes. First, simplify the notion of Q_k-formulas: since we do not assume that the language of S contains the language of arithmetic, let Q_o-formulas be all atomic S-formulas; Q_{k+1}-formulas are boolean combinations of $\exists Q_k$-formulas. Thus Q_k-formulas are particular S-formulas. Show by induction on k, using the given interpretation, that in T there is a cut I_k and a satisfaction $Sat_{Q,k}$ for Q_k-formulas such that T proves Tarski's truth conditions and thus T proves the following "snowing"-snowing-lemma: for each S-formula $\varphi(x, \ldots)$ which is a Q_k-formula,

$$T \vdash (\chi(x) \,\&\, \ldots) \rightarrow (\varphi^*(x, \ldots) \equiv Sat_{Q,k}(\varphi, [x, \ldots])\,.$$

(Here $*$ is the interpretation).

There is a shortening J_k of I_k such that $T \vdash$ (Q_k-proofs $z \in I$ preserve $Sat_{Q,k}$-truth) and there is a shortening K_k of J_k such that T proves that each H-proof$^\bullet$ (from no special axioms) $u \in K_{k+2}$ of a Q_k- formula φ determines a Q_{k+2}-proof $z \in J_{k+2}$ of φ. Consequently T proves the following: if (the concatenation of all elements of) $S^\bullet$ is Q_k then K_{k+2} contains no H-proof$^\bullet$ of $\neg S^\bullet$ (recall that $T \vdash$ ($S^\bullet$ is true) since $T \vdash S^*$). Thus the proof of $\Rightarrow$ is complete. The rest of the subsection contains a proof of $\Leftarrow$. $\square$

3.40 Lemma. Let $T \supseteq I\Sigma_1$. For each T-cut I there is a T-cut J such that $T \vdash I \subseteq J$ and $T \vdash$ (J is closed under concatenation of sequences).

Proof. Let us work in T. Let $s, t \leq x$; then $lh(s), lh(t) \leq |x|$ and, for each i, $(s)_i, (t)_i \leq |x|$. If w is the concatenation of s, t, i.e. $w = s \frown t$, then

$lh(u) \leq 2|x|$ and each member is $\leq |x|$, thus, by 3.13, $w \leq 2^{(3(x)+1)^2}$. An elementary computation shows that

$$\omega_2(x) = 2^{1/4*(|x|-1)^2};$$

thus it suffices to assume that J is closed under ω_2 (cf. 3.5). $\square$

3.41 Corollary. Let L be a finite language, let $T \supseteq I\Sigma_1$, let I be a T-cut. Then there is a T-cut J such that $T \vdash J \subseteq I$ and T proves that for each function symbol F and predicate symbol P of $L^\bullet$, of arity n, whenever $t_1, \ldots, t_n$ are in J then so is $F(t_1, \ldots, t_n)$ and $P(t_1, \ldots, t_n)$.

3.42 *Proof of* 3.39 $\Leftarrow$. Our starting situation is: S is a finite theory (assume S is a one-element set), $T \supseteq I\Sigma_1$ is sequential, I is a T-cut such that $T \vdash HCon^\bullet(S^\bullet)$. Thus $T \vdash$ (in I, there is no propositional *proof*$^\bullet$ of any disjunction of instantions of $\neg Sk(S^\bullet)$), cf. the definition I.4.11 of $He^\bullet(\neg S^\bullet)$).

Define, in T, a function F described as follows: for each x, first take the conjunction κ_x of all closed instances of $Sk(S^\bullet)$ less than x, then construct the truth table for κ_x. If there is an evaluation of atoms making κ_x propositionally true then $F(x)$ is such an evaluation (e.g. the least one). Otherwise $F(x)$ is a propositional proof of $\neg\kappa_x$. Check that F is Δ_1 and m.e.b. in T. Let J be a T-cut shorter than I and such that $T \vdash x \in J \to F(x) \in I$. Then T proves that, for each $x \in J$, the conjunction κ_x (of all closed instances of $Sk^\bullet(S^\bullet)$ less than x) is propositionally satisfable.

(2) What follows resembles König's lemma; but since we do not have enough induction we have to replace infinity in the sense of the universe by infinity in the sense of a cut. Namely, let us work in T, let L_1 be the language of $Sk(S)$ and let $V(e)$ mean that, for some $x \in J$, e is an evaluation of closed $L_1^\bullet$-atoms less than x, (say, an x-evaluation) making κ_x propositionally true. V is, in the obvious sense, a dyadic tree and is, so to speak, J-infinite: for each $x \in J$, some x-evaluation is in V. An evaluation $e \in V$ for κ_x is said to have J-unboundedly many prolongations in V if, for each $y \in J$, $y > x$, there is an evaluation $e' \in V$ for κ_y such that $e \subseteq e'$. Clearly, if $e \in V$ is an x-evaluation and has J-unboundedly many prolongations in V then there is a least $(x+1)$-evaluation $e' \in V$ extending e and having J-unboundedly many prolongations in V (since there are at most two $(x+1)$-evaluations prolonging e). Write, for a moment, $LP(e)$ for e'. An x-evaluation $e \in V$ is leftmost (i.e. leftmost having J-unboundedly many prolongations) if, for each $y < x$, $e \upharpoonright (y+1) = LP(e \upharpoonright y)$. Clearly, if an x-evaluation $e \in V$ is leftmost $(x+1)$-evaluations prolonging e.

Put $J_1(x) \equiv .\ J(x)$ & there is a unique leftmost x-evaluation $e \in V$.

Clearly, J_1 is a T-cut and $T \vdash J_1 \subseteq J$; we may define a function B on J_1 assigning to each closed L_1-atom in J_1 its truth (0 or 1) in accordance with the unique leftmost x-evaluation (for any x satisfactorily large).

(3) Shorten J_1 to J_2 such that $T \vdash$ (J_2 is closed under concatenation) and restrict B to J_2. Thus, T-provably, if $t_1, \ldots, t_n$ are closed terms in J_2 and $\overline{F}$, is a function symbol from L_1 (of arity n) then the term $\overline{F}(t_1, \ldots, t_n)$ is also in J_2. Similarly for a predicate.

(4) We put

$$\chi(x) \equiv x \text{ is a closed } L_1^\bullet\text{-term in } J_2\,;$$

for each function symbol F of L, let

$$\psi_F(x_1, \ldots, x_n, y) \equiv y \text{ is the term } \overline{F}(x_1, \ldots x_n)\,,$$

(application$^\bullet$ of $\overline{F}$ to $x_1, \ldots, x_n$) and for each predicate P of L, let

$$\psi_P(x_1, \ldots, x_n) \equiv B \text{ assigns the value 1 to the closed atom } \overline{P}(x_1, \ldots, x_n)\,.$$

This defines our interpretation; it suffices to verify that it is an interpretation of $Sk(S)$ in T. But this is now clear: each ψ_F defines a total n-ary function on the set of all closed terms (which is non-empty). Verify (by induction outside T) that, for each L_1-term $t(x_o, \ldots, x_n)$, T proves

$$\chi(u_o)\,\&\,\ldots\,\&\,\chi(u_n) \rightarrow [t^*(u_o, \ldots, u_n) = \bar{t}(n_o/x_o, \ldots, u_n/x_n)]$$

(cf. I.1.67), and, for each open L_1-formula $\varphi(x_o, \ldots, x_n)$, T proves

$$\begin{aligned}\chi(u_o)\,\&\,\ldots\,\&\,\chi(u_n) \rightarrow [\varphi^*(u_o, \ldots, u_n) \equiv{} & \text{ with respect to } B,\\ & \varphi(u_o/x_e, \ldots, u_n/x_n) \text{ has truth value 1}]\,.\end{aligned}$$

(Note that you always need a piece of B which is a finite set.) Since T proves that each closed instance of $Sk(S^\bullet)$ has truth value 1 under B, we get $T \vdash [(\forall \ldots)Sk(S)]^\bullet$ this completes the proof of 3.39. □

3.43 Remark. If S contains an equality predicate we may be interested in an interpretation with absolute equality, i.e. such that $T \vdash \chi(x)\ \&\ \chi(y) \rightarrow (x = y \equiv x =^* y)$. The above construction is easily refined to get an interpretation with absolute equality: in T, call a closed term u B-minimal if for each $u' < u$ such that $\chi(u')$, B gives value 0 to the formula$^\bullet$ $u =^\bullet u'$. Note that with each u such that $\chi(u)$ we may associate an $u' = MT(u)$ such that u' is minimal and $B(u =^\bullet u') = 1$. (Again, this because to decide $u =^\bullet u'$ for all $u \leq u'$ we need to know only a set-piece of $B^\bullet$). Thus replace $\chi(u)$ by ($\chi(u)$ & u is B-minimal) and make appropriate changes – details left to the reader.

3.44 Remark. For PA, we get the following as a corollary: A finite theory S is interpretable in PA iff $PA \vdash Con^\bullet(S^\bullet)$. Clearly, this could be obtained directly from other previous results.

However, we can now generalize investigations of Sect. 2 (e), concerning pure extensions of PA, i.e. extensions of PA in the language of PA to arbitrary sequential theories T containing PA (i.e. PA is interpreted in T, the interpretation being fixed) and having full induction, i.e. if $N(x)$ in the predicate "x is a number" and $\varphi(x)$ is any formula of the language of T then T proves

$$\varphi(\overline{0}) \,\&\, (\forall x)(N(x) \,\&\, \varphi(x). \rightarrow \varphi(x+1). \rightarrow (\forall x)(N(x) \rightarrow \varphi(x)).$$

Note that if T is sequential and has full induction (with respect to the given interpretation of Q in N) then T contains PA.

3.45 Theorem. Let $T \in \Delta_1$ be a sequential theory having full induction, let $S \in \Delta_1$ be another theory. Then the following are equivalent:

(i) S is interpretable in T.
(ii) S is locally interpretable in T.
(iii) For each k, $T \vdash Con^\bullet((S \upharpoonright k)^\bullet)$.
(iv) There is a binumeration σ of S such that $T \vdash Con^\bullet(\sigma)$.

3.46 Theorem. If T and S are as above and S is sequential and has full induction then S is interpretable in T iff S is Π_1-conservative for T.

To prove 3.46, just check proofs of 2.39 and 2.40. You have to use the following two facts:

3.47 Lemma. If T is sequential and has full induction then T is reflexive.

Proof. By 3.21, for each finite $T_o \supseteq T$ we have $T \vdash HCon^\bullet(T_o^\bullet)$ (since there are no proper T-cuts); but T contains $I\Sigma_1$ and therefore $T \vdash Con^\bullet(T_o^\bullet) \equiv HCon^\bullet(T_o^\bullet)$. □

3.48 Remark. In the proof of (iv) $\Rightarrow$ (i) observe that we get an interpretation of S in T such that all S-objects are interpreted as some numbers in T, thus we may apply the least number principle and get an interpretation absolute with respect to equality.

4. Partial Conservativity and Interpretability

4.1. In the present section we shall investigate the notion of Γ-conservativity (Γ being a class of formulas) and interpretability as means of comparing thesies. Both notions were defined in the present chapter, Sect. 1 (a); and for pure extensions of PA (more generally, for sequential theories with full induction) we already have a result saying that S is interpretable in T iff

S is Π_1-conservative for T (see 2.40 and 3.46). Now we want to remove the assumption of full induction and, in particular, get results for finitely axiomatized theories. Recall that we have the characterization 3.39 of interpretability of a finite S in a sequential $T \supseteq I\Sigma_1$ ($T \vdash HCon^{\bullet I}(S^{\bullet})$ for some T-cut I; this will be used repeatedly.) We shall see that results concerning partial conservativity do not depend on any assumption of finite axiomatization, but properties of interpretability in finite theories differ drastically from properties of interpretability in theories with full induction.

We shall particularly focus our attention on pairs S, T of theories (containing $I\Sigma_1$, say) such that S results from T by adding one axiom, thus S is $(T+\varphi)$ for some φ. If φ is independent of T, i.e. neither provable nor refutable in T, it is natural to ask whether S is interpretable in T and how conservative S is over T. Instead of saying that $(T + \varphi)$ is Γ-conservative over T we say that φ is Γ-conservative over T; similarly for interpretability. In subsection (a) we shall ask these questions for some prominent formulas (Gödel's and Rosser's formula); subsection (b) contains some general theorems on partial conservativity and in subsection (c) these theorems are applied and related to interpretability. All this is a possible answer to the question what more we can say concerning axiomatic systems of arithmetic than that they are all incomplete.

(a) Some Prominent Examples

4.2. In the whole subsection, T denotes a theory containing $I\Sigma_1$ and τ is a formula Δ_1 in $I\Sigma_1$ defining T (thus τ binumerates T in T). We shall often write $Con^{\bullet}_T$ instead of $Con^{\bullet}_{\tau}$. $Con^{\bullet}_T$ is called Gödel's formula; recall that, by 2.22, $Con^{\bullet}_T$ is equivalent to the self-referential formula ν such that $T \vdash \nu \equiv \neg Pr_T(\overline{\nu})$. i.e. to Gödel's fixed point. Clearly, $Con^{\bullet}_T$ is a Π_1-sentence. Rosser's formula ρ is the self-referential formula such that (cf. 2.9).

$$T \vdash \rho \equiv (\exists y)(Proof^{\bullet}_T(y, \overline{\neg\rho}) \,\&\, (\forall z \leq y)\neg Proof^{\bullet}_T(z, \overline{\rho})),$$

i.e. ρ says "there is a proof y of my negation such that there is no proof of me beneath y". Thus ρ is a Σ_1-sentence. These formulas played a prominent role in Gödel's incompleteness theorems (cf. 2.8, 2.10, 2.22); now we shall discuss their properties concerning partial conservativity and interpretability.

4.3. Recall the notion of H-provability and H-proofs (H for Herbrand) in a finite theory T and the sentence $HCon^{\bullet}(T)$ expressing consistency with respect to Herbrand proofs.

4.4 Definition. Let T be a finite theory. An *H-Rosser formula* for T is a self-referential formula ρ such that

$$T \vdash \rho \equiv (\exists y)(HProof^{\bullet}_T(y, \overline{\neg\rho}) \,\&\, (\forall z \leq y)(\neg HProof^{\bullet}(z, \overline{\rho})).$$

This is like the Rosser formula but with Herbrand proofs instead of (Hilbert) proofs.

Our results are summarized in the following

4.5 Theorem. Let $T \in \Delta_1$ be a consistent theory, $T \supseteq I\Sigma_1$.

(1) Gödel's formula $Con_T^\bullet$ is not interpretable in T; its negation $\neg Con_T^\bullet$ is interpretable in T.

(2) $\neg Con_T^\bullet$ is Π_1-conservative over T; $Con_T^\bullet$ is Σ_1-conservative over T iff T is Σ_1-sound (i.e. each provable Σ_1-sentence is true).

(3) Rosser's formula is Π_1-conservative; $\neg\rho$ is Σ_1-conservative iff T is Σ_1-sound. (The same for H-Rosser's formula assuming T finite.)

(4) If T is sequential and has induction for all formulas then neither the Rosser's formula ρ nor $\neg\rho$ is interpretable in T.

(5) But if T is sequential and finite and $H\rho$ is H-Rosser's formula then both $H\rho$ and $\neg H\rho$ are interpretable in T.

4.6 Remark. (1) There is an open problem if for T as in (5) the usual Rosser's formula ρ and/or its negation is interpretable in T.

(2) The rest of the subsection contains a proof of 4.5.

*

4.7 Lemma. Let $S, T \in \Delta_1$, $S, T \supseteq I\Sigma_1$, let S have a finite language and let S be interpretable in T. Then, for each Σ_1 definition τ of T there is a Σ_1 definition σ of S such that

$$T \vdash Con_\tau^\bullet \to Con_\sigma^\bullet .$$

Proof. Let i be the mapping of S-formulas into T-formula induced by the interpretation; i is Δ_1. Copy the definition in T: we get a function $i^\bullet$, Δ_1 in T, such that, for each $\psi \in S$, T proves $i(\psi)$ and hence $T \vdash Pr_\tau^\bullet(i^\bullet(\overline{\psi}))$. Let σ be a Σ_1 definition of S. We cannot claim

$$T \vdash (\forall x)(\sigma_o(x) \to Pr_T^\bullet(i^\bullet(x)) ;$$

but take $\sigma(x)$ to be $\sigma_o(x)$ & $Pr_T^\bullet(i^\bullet(x))$. It is easy to check that σ is the desired definition of S. (Note that if τ is a Σ_1 binumeration of T in T then σ is a Σ_1 binumeration of S in T.) □

4.8 Lemma. Let T be as above and consistent; then the formula $Con_T^\bullet$ is not interpretable in T.

Proof. Assume that the theory $(T \vdash Con_T^\bullet)$ is interpretable in T; then, by the preceding lemma, there is a Σ_1-definition σ of $T + Con_T^\bullet$ such that

$T \vdash Con^\bullet_T \to Con^\bullet_\sigma$, i.e. $(T + Con^\bullet_T) \vdash Con^\bullet_\sigma$, which contradicts Gödel's second incompleteness theorem (see 2.22). $\square$

4.9 Lemma. Let T be as above; then $\neg Con^\bullet_T$ is interpretable in T and is Π_1-conservative over T.

Proof. We show that $(T + \neg Con^\bullet_T)$ is interpretable in, and Π_1 conservative for, $(T + Con^\bullet_T)$. This gives interpretability in T by 1.8 and conservativity over T using the rule of proof by cases.

By Gödel's second incompleteness theorem, 2.22, copied inside T, we get $T \vdash Con^\bullet_T \to \neg Pr_T(\overline{Con^\bullet_T})$, i.e.

$$T \vdash Con^\bullet(T) \to Con^\bullet(T + \overline{\neg Con^\bullet_T})$$

(or, more pedantically, if τ is a Σ_1 definition of T and $\tau'(x)$ is $\tau(x) \vee x = \overline{\neg Con^\bullet_\tau}$ then $T \vdash Con^\bullet_\tau \to Con^\bullet_{\tau'}$). Thus take $(T + Con^\bullet_T)$. By the low arithmetized completeness theorem, in $(T + Con^\bullet_T)$ we may define a full low model M of $T + \overline{\neg Con^\bullet_T}$; seen from outside, we just get an interpretation of $T + \neg Con^\bullet_T$ in $T + Con^\bullet_T$. This proves interpretability. Moreover, the interpretation is very well behaved, since everything is low $\Sigma^*_o(\Sigma_1)$; thus we may imitate the proof in 2.44 and construct inside $T + Con^\bullet_T$ a $\Sigma^*_o(\Sigma_1)$-embedding F of the universe onto an initial segment of M. (Induction for $\Sigma^*_o(\Sigma_1)$ used!) Then conclude as in 2.44. $\square$

4.10 Lemma. If $T \supseteq I\Sigma_1$ is Σ_1-sound then each non-refutable Π_1-sentence is Σ_1-conservative over T.

Proof. Let $\pi \in \Pi_1, \sigma \in \Sigma_1, \neg\pi$ unprovable in T. If $(T+\pi) \vdash \sigma$ then $T \vdash \neg\pi \vee \sigma$ and $\neg\pi \vee \sigma$ is Σ_1 in T. From the non-refutability of π we get $N \models \pi$ by Σ_1-completeness; on the other hand, from $T \vdash \neg\pi \vee \sigma$ we get $N \models \neg\pi \vee \sigma$ by Σ_1-soundness. This gives $N \models \sigma$ and $T \vdash \sigma$. $\square$

4.11 Corollary. If $T \in \Delta_1$ contains $I\Sigma_1$ and is Σ_1-sound then $Con^\bullet_T$ is Σ_1-conservative.

We shall prove the converse implication later. Now we turn to Rosser's sentences.

4.12 Lemma. Let $T \in \Delta_1$, $T \supseteq I\Sigma_1$, T consistent. Then Rosser's sentence ρ_T is Π_1-nonconservative.

Proof. ρ evidently implies the following Π_1-sentence

$$(\forall z)(Proof^\bullet(\overline{\rho}, z) \to (\exists y < z)Proof^\bullet(\overline{\neg\rho}, z)).$$

Denote it by π. If $T \vdash \pi$ then $T \vdash Pr^{\bullet}_{T}(\overline{\rho}) \to \rho$, since, by Löb's theorem 2.25, $T \vdash \rho$, which is impossible. Thus $(T + \rho)$ proves π but T does not. □

4.13 Remark. The same holds for H-Rosser's formula (provided T is finite).

4.14 Corollary. If T is a consistent sequential theory with full induction then neither ρ nor $\neg\rho$ are interpretable (i.e. T interprets neither $(T+\rho)$ nor $T+\neg\rho$).

This follows by 3.46, 4.12 and the fact that $\neg\rho$ is an unprovable Π_1 formula. Similarly for H-Rosser. On the other hand, we have the following:

4.15 Lemma. If $T \supseteq I\Sigma_1$ is consistent, finitely axiomatized and sequential then both the H-Rosser sentence and its negation are interpretable in T.

Proof. For simplicity assume T to be just a sentence. Let ρ be the Rosser sentence. We use Theorem 3.39. It suffices to interpret $(T + \rho)$ in $(T + \neg\rho)$ and conversely. We show that $(T+\rho)$ is interpretable in $(T+\neg\rho)$, the existence of the converse interpretation is proved analogously.

Let J be a $(T + \neg\rho)$-cut such that $(T + \neg\rho) \vdash HCon^{\bullet J}(\overline{T + \neg\rho})$. Let us work in $(T + \neg\rho)$. In J there is no H-proof$^{\bullet}$ of $\neg(\overline{T + \neg\rho})$; and $\neg\rho$ says that for each H-proof$^{\bullet}$ of $\overline{\neg\rho}$ in T (i.e. for each H-proof$^{\bullet}$ $\overline{\neg(T\&\neg\neg\rho)}$) there is a smaller H-proof$^{\bullet}$ of $\overline{\rho}$ in $\overline{T}$ (i.e. a H-proof$^{\bullet}$ of $\overline{\neg(T\&\neg\rho)}$). Consequently, J does not contain any H-proof$^{\bullet}$ of $\overline{\neg(T\&\neg\neg\rho)}$ and we get $HCon(T + \neg\neg\rho)$.

The provability of the last formula in $(T + \neg\rho)$ gives an interpretation of $(T + \neg\neg\rho)$ in $(T + \neg\rho)$, hence an interpretation of $(T + \rho)$ in $(T + \neg\rho)$ by 3.39. □

4.16 Corollary and Remark. If $T \supseteq I\Sigma_1$ is consistent, finite and sequential then there is a Π_1 formula φ such that both φ and $\neg\varphi$ are interpretable in T (once more, this means that the theory $(T + \varphi)$ is interpretable in T and so is $(T + \neg\varphi)$).

Note that this is not true for sequential theories with full induction (like *PA* and *ZF*) since in such a theory each interpretable Π_1-sentence is provable (cf. 2.40). But one can find a $\varphi \in \Delta_2$ such that both φ and $\neg\varphi$ is interpretable.

On the other hand, take ACA_0 for T; assumptions of our corollary are satisfied. The corresponding formula φ has the following properties: $(ACA_0 + \varphi)$ is interpretable in ACA_0 but $(PA+\varphi)$ is not interpretable in *PA* (since φ is unprovable). Similarly for Gödel-Bernays and Zermelo-Fraenkel set theories *GB* and *ZF*. In the next section we shall obtain a Σ_1 sentence φ with converse properties: $(PA + \varphi)$ is interpretable in *PA* but $(ACA_0 + \varphi)$ is not in ACA_0.

4.17 Lemma. Let $T \supseteq I\Sigma_1$ be consistent.

(1) $Con^{\bullet}_{T}$ is Σ_1-conservative over T iff T is Σ_1-sound.

(2) Let ρ be the Rosser (or H-Rosser) Σ_1-sentence. Then $\neg\rho$ is Σ_1-conservative iff T is Σ_1-sound.

Proof. We shall prove (1). The proof will be a preparation for generalized Rosser sentences studied in the next subsection; there we also prove assertion (2).

The implication $\Leftarrow$ is evident: if $\sigma \in \Sigma_1$, T is Σ_1-sound and $T \vdash Con^\bullet_T \to \rho$ then $N \models Con^\bullet_T \to \rho$, $N \models \rho$ and $T \vdash \rho$.

Now assume T not to be Σ_1-sound and let $\varphi(y)$ be a Σ_0-formula such that $N \models \neg(\exists y)\varphi(y)$ but $T \vdash (\exists y)\varphi(y)$. Take the following self-referential formula:

$$T \vdash \xi \equiv (\exists y)(Proof^\bullet(\overline{\neg\xi}, y) \vee \varphi(y) \,\&\, (\forall z \leq y)\neg Proof^\bullet(\overline{\xi}, z)).$$

Note that $T \vdash \neg\varphi(\overline{k})$ for each k.

(i) We show that T does not prove ξ. Assume $T \vdash \xi$; we show that T is inconsistent. Let d be a proof of ξ in T; then $T \vdash Proof^\bullet(\overline{\xi}, \overline{d})$ and for each $k < d$, $T \vdash \neg\varphi(\overline{k})$.

From $T \vdash \xi$ we get

$$T \vdash (\exists y < \overline{d}) Proof^\bullet(\overline{\neg\xi}, y),$$
$$T \vdash \bigvee_{k<d} Proof^\bullet(\overline{\neg\xi}, \overline{k})$$

and since each formula $Proof^\bullet(\overline{\neg\xi}, \overline{k})$ is decidable in T there is a k such that

$$T \vdash Proof^\bullet(\overline{\neg\xi}, \overline{k}).$$

Thus k is a proof of $\neg\xi$ in T, thus $T \vdash \neg\xi$ and T is inconsistent. This shows that ξ is unprovable in T provided T is consistent.

(ii) We now prove $T + Con^\bullet_T \vdash \xi$ or, equivalently, $T + \neg\xi \vdash \neg Con^\bullet_T$. Let us work in $(T + \neg\xi)$. Since $(\exists y)\varphi(y)$, let y_0 be the least y such that $\varphi(y)$. From $\neg\xi$ we get $(\exists z \leq y_0) Proof^\bullet(\overline{\xi}, z)$; let $Proof^\bullet(\overline{\xi}, z_0)$. Then $Proof^\bullet(\overline{\xi}, z_0)$ & $(\forall y < z_0)\neg\varphi(y)$, hence inside T we are in the same situation as we were in (i) outside. Copy the reasoning from (i) into T; we get

$$Pr^\bullet(\overline{Proof^\bullet}(\dot{\overline{\xi}}, \dot{z}_0)) \,\&\, (\forall y < z_0) Pr^\bullet(\overline{\neg\varphi}(\dot{y})).$$

From $Pr^\bullet(\xi)$ we get

$$Pr^\bullet(\bigvee_{\dot{y}<z_0} \overline{Proof^\bullet}(\dot{\overline{\neg\xi}}, \dot{y})),$$

hence

$$(\exists y < z_0) Pr^\bullet(\overline{Proof^\bullet}(\dot{\overline{\neg\xi}}, \dot{y})),$$
$$(\exists y < z_0) Proof^\bullet(\neg\overline{\xi}, y),$$

thus we get $Pr^\bullet(\overline{\xi})$ and $Pr^\bullet(\neg\overline{\xi})$, thus $\neg Con^\bullet$. $\square$

We have presented a rather detailed proof; similar proofs below will mostly be less detailed.

(b) General Theorems on Partial Conservativity; Some Fixed-Point Theorems

In this subsection we shall state and prove some basic facts on partially conservative sentences. Our proofs will use self-reference as a very basic means. Some methods of construction of self-referential sentences will be isolated and properties of constructed formulas will be stated in general theorems called usually fixed-point theorems. It is remarkable that most of our constructions are some generalizations of Rosser's formula, i.e. are based on witness comparisons. We first introduce some notations and formulate a simple technical lemma, then we formulate our basic results (inclusive the fixed-point theorems) and finally we elaborate proofs. In the whole subsection, T is a consistent theory, $T \in \Delta_1$, $T \supseteq I\Sigma_1$.

4.18 Definition Let $\alpha(u)$, $\beta(u)$ be T-formulas, let $\triangle$ be $(\exists u)\alpha(u)$ and let ∇ be $(\exists u)\beta(u)$. In T, call each a satisfying $\alpha(u)$ a witness for $\triangle$ and similarly for ∇ (cf. 2.9). By $\triangle \prec \nabla$ we denote the witness comparison formula

$$(\exists u)(\alpha(u) \,\&\, (\forall \nu \leq u)\neg\beta(v))$$

(there is a witness for $\triangle$ less than each witness for ∇); similarly for $\triangle \preccurlyeq \nabla$ (replace $(\forall v \leq u)$ by $(\forall v < u)$). The formula $\triangle$ is called the *antecedent* of $\triangle \prec \nabla$ and $\triangle$ its *succedent*. Similarly for $\triangle \preccurlyeq \nabla$. Note that $\triangle$ and ∇ may contain free variables as parameters. (Thus, for example, Rosser's formula is a formula ρ such that

$$T \vdash \rho \equiv Pr^{\bullet}(\overline{\neg\rho}) \prec Pr^{\bullet}(\overline{\rho}).)$$

4.19 Remark. Rosser's formula is Σ_1 in T, since we assume $T \supseteq I\Sigma_1$. Recall that in $B\Sigma_n$, Σ_n formulas are closed under bounded universal quantifers. We shall deal with formulas comparing witnesses of Σ_n-formulas but continue to assume only $I\Sigma_1$. Therefore we shall have to pay some attention to the arithmetical complexity of the resulting formulas. The following simple lemma will be extremely useful.

4.20 Lemma. Let $\varphi(x)$ be a Σ_n-formula. There is a Σ_n-formula $\psi(y)$ such that

(1) $$(\forall k) I\Sigma_1 \vdash \psi(\overline{k}) \equiv (\forall x \leq \overline{k})\varphi(x)\,,$$

(2) $$I\Sigma_1 \vdash \psi(y) \rightarrow (\forall x \leq y)\varphi(x)\,.$$

Proof. For $n = 0$ the assertion is trivial. For $n \geq 1$ and $\varphi(x)$ of the form $(\exists u)\alpha(x, u)$, where α is Π_{n-1} let $\psi(y)$ be

$$(\exists s)(Seq(s) \,\&\, (\forall x \leq y)\alpha(x, (s)_x)\,. \qquad \square$$

4.21 Definition. (1) If $\varphi(x)$ and $\psi(y)$ are as above then we denote $\psi(y)$ by $[(\forall x \leq y)\varphi(x))]^{\#,\Sigma,n}$ or $[(\forall x \leq y)\varphi(x)]^{\#}$ (if Σ and n is clear from the context). Dually we define $[(\exists x \leq y)\varphi(x)]^{\#,\Pi,n}$ for φ being Π_n (or, briefly, $[(\exists x \leq y)\varphi(x)]^{\#}$).

(2) Let α and β be as by 4.18 and assume that β is Π_n. Let β' be the natural Σ_n-form of $\neg\beta$. Then $\Delta \prec^{\#} \nabla$ is the formula

$$(\exists u)(\alpha(u) \,\&\, [(\forall v \leq u)\beta'(v)]^{\#,\Sigma,n})\,;$$

the notation $\Delta \prec^{\#} \nabla$ will be used only in situations where n is clear from the context. Similarly for $\Delta \preccurlyeq^{\#} \nabla$; the words "antecedent" and "succedent" have the obvious meaning.

4.22 Remark. We shall deal with self-referential formulas ξ such that

$$T \vdash \xi \equiv \Delta(\overline{\neg\xi}) \prec^{\#} \nabla(\overline{\xi})\,;$$

more generally, there may be another parameter:

$$T \vdash \xi \equiv \Delta(\overline{\neg\xi}, \overline{k}) \prec^{\#} \nabla(\overline{\xi}, \overline{k})$$

(k natural number). Such formulas will be extremely useful; almost always the antecedent and/or the succedent will contain some proof predicate. Let us mention $Pr^{\bullet}_T$ and $HPr^{\bullet}_T$ (Herband provability, assuming T finite.) Our general theorems will be obtained by combining self-reference and witness comparison with definable cuts and partial truth definitions. Recall that satisfaction for $\Sigma^{\bullet}_n$-formulas$^\bullet$ and $\Pi^{\bullet}_n$-formulas ($Sat_{\Sigma,n}$, $Sat_{\pi,n}$) were defined in $I\Sigma_1$ already in I.1.75–76; in Chap. I, Sect. 2 we used them to prove finite axiomatizability of $I\Sigma_n$ and $B\Sigma_{n+1}$ and studied satisfaction for relativized hierarchies. In Chap. I, Sect. 4 we showed that these satisfaction relations are just a particular case of the general notion of partial satisfaction, namely, satisfaction for the standard model (defined inside $I\Sigma_1$). In the present chapter we already met Σ_1-satisfaction in Sect. 3 (Theorem 3.21 saying that for suitable (finite) theories T, there is a T-cut J such that T proves $(\forall u \in Tr(\Sigma_1))HCon^{\bullet J}(T + u))$ (for any true $\Sigma^{\bullet}_1$-sentence, in J there is no Herhand-proof$^\bullet$ of $\neg(\bigwedge T \,\&\, u)$). We shall now make systematic use of provability (H-provability) in extensions of T resulting by adding a true Γ-formula$^\bullet$ where Γ is a class of formulas$^\bullet$, say $\Sigma^{\bullet}_n$ or $\Pi^{\bullet}_n$. This leads us to the following.

4.23 Definition and Convention. Let $n \in N$, let Γ be Σ_n or Π_n. We make the following definition in $I\Sigma_1$: $Pr^{\bullet}_{T,\Gamma}(x)$ (x is provable in T from a true Γ-sentence) if

$$(\exists y \in \Gamma)(Tr_{\Gamma}(y) \,\&\, Pr^{\bullet}_{(T+y)}(x)).$$

If T is finite we define $HPr^{\bullet}_{T,\Gamma}(x)$ in the obvious way.

Convention. (1) To simplify notation, in the rest of this section we drop the index T and the dot; thus we shall write $Pr(x)$, $Pr_{\Gamma}(x)$ instead of $Pr^{\bullet}_{T}(x)$, $Pr^{\bullet}_{T,\Gamma}(x)$ etc.

(2) If Δ_i is $(\exists u)\alpha_i(u)$ and ∇_i is $(\exists v)\beta_i(v)$ ($i = 1,2$) then $(\Delta_1 \vee \nabla_2) \prec (\Delta_1 \vee \nabla_2)$ means the formula saying "there is a witness for $\alpha_1 \vee \alpha_2$ less than each witness for $\beta_1 \vee \beta_2$", similarly for $\prec^{\#}$ instead of $\prec$.

4.24 Definition. (1) φ is *hereditarily Γ-conservative* over T if, for each T_0 such that $I\Sigma_1 \subseteq T_0 \subseteq T$, φ is Γ-conservative over T_0.

(2) φ is *doubly Γ-conservative* over T if φ is Γ-conservative over T and $\neg\varphi$ is $\tilde{\Gamma}$-conservative over T (where $\tilde{\Gamma}$ is the dual class or Γ).

We shall now formulate three general theorems on partial conservativity. Note that we continue to assume that T is consistent, $T \in \Delta_1$ and $T \supseteq I\Sigma_1$.

4.25 Theorem. For each $n \geq 1$ there is (1) a hereditarily Π_n-conservative Σ_n-sentence, (2) a hereditarily Σ_n-conservative Π_n-sentence, (3) a doubly Π_n-conservative Σ_n-sentence (its negation is thus a doubly Σ_n conservative Π_n-sentence).

Examples (Γ is Σ_n, Λ is Π_n):

(1) ξ such that $I\Sigma_1 \vdash \xi \equiv Pr_{\Gamma}(\neg\overline{\xi}) \prec^{\#} Pr(\overline{\xi})$,
(2) $(\neg\xi)$ such that $I\Sigma_1 \vdash \xi \vdash Pr(\neg\overline{\xi}) \prec^{\#} Pr_{\Lambda}(\overline{\xi})$,
(3) ξ such that $I\Sigma_1 \vdash \xi \subseteq Pr_{\Gamma}(\neg\overline{\xi}) \prec^{\#} Pr_{\Lambda}(\xi)$.

If T is Σ_n-sound we may take in (1) a ξ such that $I\Sigma_1 \vdash \xi \equiv Pr_{\Gamma}(\neg\overline{\xi})$.

4.26 Theorem (on non-separability). Let Γ be Σ_n or Π_n ($n \geq 1$), let Th be the set of all theorems of T, $Consv(\Gamma)$ and $hConsv(\Gamma)$ the set of all Γ-conservative and hereditarily Γ-conservative sentences respectively, $NRef$ the set of all the sentences non-refutable in T. Then obviously

$$Th \subseteq hConsv(\Gamma) \subseteq Consv(\Sigma_1) \cap Consv(\Pi_1) \subseteq NRef$$

and there is no set X such that

(1) X is Δ_1 and $Th \subseteq X \subseteq NRef$, *or*
(2) X is Π_1 and $Th \subseteq X \subseteq Consv(\Gamma)$, *or*
(3) X is Σ_1 and $hConsv(\Gamma) \subseteq X \subseteq NRef$, *or*
(4) X is Σ_2, $\Gamma \supseteq \Sigma_1$ and $hConsv(\Gamma) \subseteq X \subseteq Consv(\Sigma_1)$, *or*
(4') X is Σ_2, $\Gamma \supseteq \Pi_1$ and $hConsv(\Gamma) \subseteq X \subseteq Consv(\Pi_1)$.

4.27 Theorem (Π_2-completeness). For each $n \geq 1$ and $\Gamma = \Sigma_n$ or Π_n, both $Consv(\Gamma)$ and $hConsv(\Gamma)$ is Π_2-complete.

4.28 Remark. (1) Concerning 4.26, observe that Th is Σ_1, $NRef$ is Π_1 and $Consv(\Gamma)$ is Π_2. Thus the result is optimal: in 4.26 (1) Δ_1 can be replaced by neither Σ_1 nor Π_1, in 4.26 (2) Π_1 cannot be replaced by Σ_1, in 4.26 (3) Σ_1 cannot be replaced by Π_1, and in 4.26 (4,4$'$) Σ_2 cannot be replaced by Π_2.

(2) Recall from recursion theory that a set $X \subseteq N$ is Π_2-complete if $X \in \Pi_2$ and each Π_2 set Y is Δ_1-reducible to X, i.e. for some Δ_1 total function F, $Y = \{n \mid F(n) \in X\}$.

(3) We shall present two general fixed point theorems that form the main means of proofs of the preceding theorems.

4.29 Sheperdson-Smoryński's Fixed Point Theorem. Let Φ, Ψ be Σ_1 formulas.

(1) Let $I\Sigma_1 \vdash \xi \equiv [(Pr(\neg\overline{\xi}) \vee \Phi) \prec (Pr(\xi) \vee \Psi)]$. Then
   (i) $T \vdash \xi$ iff $N \models \Phi \prec \Psi$ iff $N \models \xi$;
   (ii) $T \vdash \neg\xi$ iff $N \models \Psi \preccurlyeq \Phi$.

(2) More generally, let for $i = 1, 2$, $T_i \supseteq I\Sigma_1$, let Pr_i be the proof predicate based on a fixed Δ_1 definition of T_i. Let

$$I\Sigma_1 \vdash \xi \equiv [(Pr_1(\neg\xi) \vee Pr_2(\neg\xi) \vee \Phi) \prec (Pr_1(\xi) \vee Pr_2(\xi) \vee \Psi)]\,.$$

Then
   (i) $T_1 \vdash \xi$ iff $T_2 \vdash \xi$ iff $N \models \Phi \prec \Psi$ iff $N \models \xi$;
   (ii) $T_1 \vdash \neg\xi$ iff $T_2 \vdash \neg\xi$ iff $N \models \Psi \preccurlyeq \Phi$.

4.30 Lindström's Fixed Point Theorem. (Let T be as above.)

(1) Let $\chi(y)$ be Σ_n and let $I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\neg\overline{\xi}) \prec^{\#} (\exists y)\neg\chi(y)$. Then
   (i) for each m, $(T + \xi) \vdash \chi(m)$,
   (ii) for each $I\Sigma_1 \subseteq T_0 \subseteq T$ and each Π_n-sentence π, $(T_0 + \xi) \vdash \pi$ implies $T_0 + \{\chi(m) \mid m\} \vdash \pi$.

(2) Let $\chi(y)$ be Π_n and let $I\Sigma_1 \vdash \xi \equiv (\exists y)\neg\chi(y) \prec^{\#} Pr_{\Pi_n}(\xi)$. Then
   (i) for each m, $(T + \neg\xi) \vdash \chi(m)$,
   (ii) for each $I\Sigma_1 \subseteq T_0 \subseteq T$ and each Σ_n-sentence β, $(T_0 + \neg\xi) \vdash \beta$ implies $T_0 + \{\chi(m) \mid m\} \vdash \beta$.

4.31 Remark. From these fixed point theorems we derive our three main results (4.25–4.27) as well as other results of independent interest (in particular, Theorem 4.34 claiming that each Σ_1 set has a sound Σ_1 numeration in T; this was promised in 1.24). An elaboration follows.

4.32 (*Proof of Smoryński's Fixed Point Theorem* 4.29). We are going to prove part (2) of the theorem, (1) being a particular case. Recall that we have two consistent theories $T_i \subseteq I\Sigma_1$ $(i = 1, 2)$ and for both T_i we have a fixed definition of T_i which is Δ_1 in $I\Sigma_1$ (i.e. binumerates T_i in $I\Sigma_1$). Our diagonal formula ξ satisfies

$$I\Sigma_1 \vdash \xi \equiv [(Pr_1(\overline{\neg\xi}) \vee Pr_2(\overline{\neg\xi}) \vee \Phi) \prec (Pr_1(\overline{\xi}) \vee Pr_2(\overline{\xi}) \vee \Psi)]$$

where Φ, Ψ are fixed Σ_1 formulas. Thus ξ is Σ_1 in $I\Sigma_1$.

(1) First observe that if any of the formulas ξ, $Pr_1(\overline{\xi})$, $Pr_2(\overline{\xi})$, $Pr_1(\overline{\neg\xi})$, $Pr_2(\overline{\neg\xi})$, Φ, Ψ is true in N (i.e. has a standard witness d) then ξ becomes Δ_1 in $I\Sigma_1$ since the existential quantitier in the witness comparison formula can be bounded by $\overline{d}$.

(2) Thus we get

$$T_i \vdash \xi \Rightarrow N \models \xi \Rightarrow T_j \vdash \xi\,,$$
$$T_i \vdash \neg\xi \Rightarrow (N \models \neg\xi \text{ and } T_i \vdash \neg\xi)$$

$(i = 1, 2)$ – note that $T_i \vdash \xi$ iff $N \models Pr_i(\overline{\xi})$.

(3) Furthermore, $T_i \vdash \xi$ implies $N \models \Phi \prec \Psi$ and similarly, $T_i \vdash \neg\xi$ implies $N \models \Psi \preccurlyeq \Phi$. We prove the former claim: Let d be a T_i-proof of ξ; thus it is both true and T_i-provable that there is an $y \leq \overline{d}$ which witnesses the antecedent $(Pr_1(\overline{\neg\xi}) \vee Pr_2(\overline{\neg\xi}) \vee \Phi)$ and $(\forall z \leq y)(z$ does not witness the succedent). Let $e \in N$ be such a y. But e cannot witness $Pr_j(\overline{\neg\xi})$ since this would make T_1 or T_2 contradictory; thus e witnesses Φ and hence $\Phi \prec \Psi$. We have $N \models \Phi \preccurlyeq \Psi$.

(4) Conversely, $N \models \Phi \preccurlyeq \Psi$ implies $T_i \vdash \xi$ since if $N \models \Phi \preccurlyeq \Psi$ then $N \models \Phi$, which makes $\xi \Delta_1$ in $I\Sigma_1$ and hence decidable (provable or refutable) in T_i. But T_i cannot refute ξ since, by (3), this would imply $N \models \Psi \preccurlyeq \Phi$. Similarly, $N \models \Psi \preccurlyeq \Phi$ implies $T_i \vdash \neg\xi$. This gives all we need. □

The theorem can be parametrized (and again called Sheperdson-Smoryński's fixed point theorem):

4.33 Theorem. Let $\varphi(x,y)$, $\psi(x,y)$ be Σ_0-formulas (or: formulas Δ_1 in $I\Sigma_1$), put $\Phi(x) \equiv (\exists y)\varphi(x,y), \Psi(x) \equiv (\exists y)\psi(x,y)$.

(1) Let T be as in 4.29 (1). Assume that for each k the formula $\xi(\mathrm{x})$ satisfies the following:

$$T \vdash \xi(\overline{k}) \equiv Pr(\overline{\neg\xi(\dot{k})}) \vee \Phi(\overline{k}) \prec Pr(\overline{\xi(\dot{k})}) \vee \Psi(\overline{k})\,.$$

Then

$$T \vdash \xi(\overline{k}) \text{ iff } N \models \Phi(\overline{k}) \prec \Psi(\overline{k}) \text{ iff } N \models \xi(\overline{k})\,;$$
$$T \vdash \neg\xi(\overline{k}) \text{ iff } N \models \Psi(\overline{k}) \preccurlyeq \Phi(\overline{k})\,.$$

(2) More generally, let T_i, Pr_i be as in 4.29 (2) (i.e. as in 4.32) and let for each k, $I\Sigma_1$ prove the following:

$$\xi(\overline{k}) \equiv (Pr_1(\overline{\neg\xi}(\dot{k})) \vee Pr_2(\overline{\neg\xi}(\dot{k})) \vee \Phi) \prec (Pr_1(\overline{\xi}(\dot{k})) \vee Pr_2(\overline{\xi}(\dot{k})) \vee \Psi).$$

Then

$$T_1 \vdash \xi(\overline{k}) \text{ iff } T_2 \vdash \xi(\overline{k}) \text{ iff } N \models \xi(\overline{k}) \text{ iff } N \models \Phi(\overline{k}) \prec \Psi(\overline{k}),$$
$$T_1 \vdash \neg\xi(\overline{k}) \text{ iff } T_2 \vdash \neg\xi(\overline{k}) \text{ iff } N \models \Psi(\overline{k}) \preccurlyeq \Phi(\overline{k}).$$

The proof is fully analogous to 4.32; the existence of ξ is obvious. Part (1) of the parametric version will now be used to get several important consequences.

4.34 Theorem. If X, Y are disjoint Σ_1 sets then there is a Σ_1-formula $\xi(x)$ such that, for each k,

$$k \in X \text{ iff } T \vdash \xi(\overline{k}) \text{ iff } N \models \xi(\overline{k}),$$
$$k \in Y \text{ iff } T \vdash \neg\xi(\overline{k}).$$

Proof. Take Σ_1-definitions of X and Y for $\Phi(x)$ and $\Psi(x)$. □

4.35 Corollary. (1) Each Σ_1 set X has a sound numeration in T, i.e. a formula which both defines X and numerates X in T.

(2) Each Σ_1 set has a Π_1 numeration in T.

4.36 Theorem (= 4.26 (1)). There is no Δ_1 set X such that $Th \subseteq X \subseteq NRef$.

Proof. By 4.34, let ξ be such that $k \in X$ iff $T \vdash \xi(\overline{k})$ and $k \notin X$ iff $T \vdash \neg\xi(\overline{k})$; let $T \vdash \varphi \equiv \neg\xi(\overline{\varphi})$. Clearly, $\varphi \in X$ iff $\varphi \notin X$, a contradiction. □

4.37 Theorem. For each Π_1 set X there is a formula $\xi(x)$ such that, for each k,

(1) $k \in X$ iff $\xi(\overline{k})$ is neither Σ_1-conservative nor Π_1-conservative over T,

(2) $k \notin X$ iff $T \vdash \xi(\overline{k})$.

Proof. Observe that the formula $\xi(x)$ constructed for the Σ_1-set $-X$ as in 4.29 is Σ_1, thus Σ_1-nonconservative whenever unprovable. Its Π_1-nonconservativeness is proved similarly as the Π_1-nonconservativeness of Rosser's formula. □

4.38 Corollary (= 4.26 (2)). There is no Π_1-set X that $Th \subseteq X \subseteq Consv(\Sigma_1) \cup Consv(\Pi_1)$.

Proof. Assume we have such an X and take the corresponding formula $\xi(x)$ from 4.37. Let $T \vdash \varphi \equiv \xi(\overline{\varphi})$; if $\varphi \notin X$ then $T \vdash \xi(\overline{\varphi})$, thus $T \vdash \varphi$ and $\varphi \in X$, a contradiction. On the other hand, if $\varphi \in X$ then $\xi(\overline{\varphi})$ (and hence φ) is Σ_1-nonconservative as well as Π_1-nonconservative, i.e. $\varphi \notin X$, a contradiction. □

We shall need the following corollary of Sheperdson-Smoryński's fixed point theorem.

4.39 Theorem. Let $T_1, T_2 \supseteq I\Sigma_1$ be Δ_1 and consistent, let X, Y be disjoint Σ_1-sets. Then there is a Σ_1-formula $\xi(x)$ such that, for each k,

$$k \in X \text{ iff } T_1 \vdash \xi(\overline{k}) \text{ iff } T_2 \vdash \xi(\overline{k}) \text{ iff } N \models \xi(\overline{k}),$$
$$k \in Y \text{ iff } T_1 \vdash \neg\xi(\overline{k}) \text{ iff } T_2 \vdash \neg\xi(\overline{k}).$$

In particular, $\xi(x)$ is a sound numeration of X both in T_1 and in T_2.

Proof evident. □

We now turn to proofs from true Γ-formulas (cf. 4.23). We write $Proof^\bullet_\Gamma(z,x)$ instead of $Proof^\bullet_{T,\Gamma}(z,x)$.

4.40 Lemma. For each φ and d,

$$T \vdash Proof^\bullet_\Gamma(\overline{d}, \overline{\varphi}) \to \varphi;$$

consequently if for some d, T proves that $\overline{d}$ is a T-proof$^\bullet$ of $\overline{\varphi}$ from true Γ-formulas$^\bullet$ then T proves φ.

Proof. Let d and φ be given: assume that d is a sequence $\varphi_1, \ldots, \varphi_n$ of T-formulas which is a $(T+\psi)$-proof of φ for some $\psi \in \Gamma$, $\psi \le d$. (If this is not the case then $T \vdash \neg Proof^\bullet_\Gamma(\overline{d}, \overline{\varphi})$ and we are done). Let us work in T.

Assume $Proof^\bullet_\Gamma(\overline{d}, \overline{\varphi})$. Then for some $y \in \Gamma$; $y \le \overline{d}$, we have $Proof^\bullet_{(\Gamma+x)}(\overline{d}, \overline{\varphi})$ and $Tr_\Gamma(y)$, i.e. for some $\psi \le d$ we have $Tr_\Gamma(\overline{\psi})$ and $Proof^\bullet_{(\Gamma+\overline{\varphi})}(\overline{d}, \overline{\psi})$. Thus we get ψ. Having this we successively prove $\varphi_1, \varphi_2, \ldots, \varphi_n$, hence φ. □

4.41. We now prove Lindström's fixed point theorem (4.30).

(1) Let $I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\neg\overline{\xi}) \prec^{\#} (\exists y)\neg\chi(y)$, $\chi \in \Sigma_n$; recall the convention 4.23. Let $I\Sigma_1 \subseteq T_0 \subseteq T$. We shall prove the following:
   (a) $T + \xi \vdash \chi(\overline{m})$,
   (b) $T_0 + \xi$ is Π_n-conservative for $T_0 + \{\chi(\overline{m}) \mid m\}$.

Let us work in $T+\xi+\neg\chi(\overline{m})$; then m is a witness for the succedent of ξ, thus $(\exists y < \overline{m}) Proof^\bullet_{\Sigma_n}(y, \neg\overline{\xi})$; this gives $\neg\overline{\xi}$ by 4.40. We see that $(T+\xi+\neg\chi(\overline{m}))$ is contradictory. This proves (a).

Now assume $I\Sigma_1 \subseteq T_0 \subseteq T$ and let π be a Π_n-formula such that $(T_0+\xi) \vdash \pi$, thus $(T_0+\neg\pi) \vdash \neg\xi$; let d be a proof of $\neg\xi$ from the natural Σ_n equivalent δ of $\neg\pi$. Let us work in $T_0 + \{\chi(\overline{m}) \mid m\} + \neg\xi$, we want to prove π. Assume the contrary, i.e. δ; then $Tr(\overline{\delta})$, i.e. $Proof^{\bullet}_{\Sigma_n}(\overline{d}, \neg\overline{\xi})$, hence d is a witness of the antecedent of ξ. This gives $(\exists y \leq \overline{d})\neg\chi(y)$, i.e. $\bigvee_{k\leq\overline{d}} \neg\chi(\overline{m})$, which is a contradiction in our theory. This proves (1).

(2) Now assume $I\Sigma_1 \vdash \xi \equiv (\exists y)\neg\chi(y) \prec^{\#} Pr_{\Pi_n}(\xi)$, $\chi \in \Pi_n$.
Let T_0 be as above. We shall prove
 (a) $(\forall m)(T + \neg\xi \vdash \chi(\overline{m}))$,
 (b) $(T_0 + \neg\xi)$ is Σ_n-conservative for $T_0 + \{\chi(m) \mid m\}$.

Let us work in $T_0 + \neg\xi + \chi(\overline{m})$. Then $\overline{m}$ witnesses the antecedent of ξ, and $\neg\xi$ implies $(\exists y \leq \overline{m})Proof^{\bullet}_{\Pi_n}(\overline{\xi}, y)$, hence ξ by 4.40, a contradiction. This proves (a).

Now assume $I\Sigma_1 \subseteq T_0 \subseteq T$ and let π be a Π_n-formula such that $(T_0+\xi) \vdash \pi$, thus $(T_0 + \neg\pi) \vdash \neg\xi$; let d be a proof $\neg\xi$ from the natural Σ_n equivalent δ of $\neg\pi$. Let us work in $T_0 + \{\chi(\overline{m}) \mid m\} + \neg\xi$, we want to prove π. Assume the contrary, i.e. δ; then $Tr(\overline{\delta})$, i.e. $Proof^{\bullet}_{\Sigma_n}(\overline{d}, \overline{\neg\xi})$, hence d is a witness of the antecedent of ξ. This gives $(\exists y \leq \overline{d})\neg\chi(y)$, i.e. $\bigvee_{k\leq d} \neg\chi(\overline{m})$, which is a contradiction in our theory. This proves (1).

(2) Now assume $I\Sigma_1 \vdash \xi \equiv (\exists y)\neg\chi(y) \prec^{\#} Pr_{\Pi_n}(\overline{\xi})$, $\chi \in \Pi_n$. Let T_0 be as above. We shall prove
 (a) $(\forall m)(T + \neg\xi \vdash \chi(\overline{m}))$,
 (b) $(T_0 + \neg\xi)$ is Σ_n-conservative for $T_0 + \{\chi(m) \mid m\}$.

Let us work in $T_0 + \neg\xi + \chi(\overline{m})$. Then $\overline{m}$ witnesses the antecedent of ξ, and $\neg\xi$ implies $(\exists y \leq \overline{m})Proof^{\bullet}_{\Pi_n}(\overline{\xi}, y)$, hence ξ by 4.40, a contradiction. This proves (a).

Now let $\delta \in \Sigma_n$ be such that $(T_0 + \neg\xi) \vdash \delta$, hence $(T_0 + \neg\delta) \vdash \xi$; let d be a corresponding proof. Work in $T_0 + \{\chi(\overline{m}) \mid m\} + \xi$; it suffices to prove δ. Assume the contrary, i.e. $Tr(\overline{\neg\delta})$; then $Proof^{\bullet}_{\Pi_n}(\overline{d}, \overline{\xi})$, i.e. ξ gives $(\exists y \leq \overline{d})\neg\chi(y)$, a contradiction. This completes our proof. □

4.42 Generalizing Lindström's Fixed Point Theorem. There are two ways of generalization: first, we may let both the antecedent and the succedent depend on ξ; second, we may parametrize the whole formula. We obtain the following four cases (T, Pr_{Σ_n}, Pr_{Π_n} as above, $I\Sigma_1 \subseteq T_0 \subseteq T$, $\chi(x,y)$ is Σ_n in (1), (2) and is Π_n in (3), (4).

$$(1) \qquad I\Sigma_1 \vdash \xi \equiv Pr^{\bullet}_{\Sigma_n}(\neg\overline{\xi}) \prec^{\#} (\exists y)\neg\chi(\overline{\xi}, y).$$

Properties: ξ is Σ_n in $I\Sigma_1$; $(T + \xi) \vdash \chi(\overline{\xi}, \overline{m})$ for all m; $(T_0 + \xi)$ is Π_n-conservative for $T_0 + \{\chi(\overline{\xi}, \overline{m}) \mid m\}$.

$$(2) \qquad I\Sigma_1 \vdash \xi(\overline{k}) \equiv Pr^{\bullet}_{\Sigma_n}(\overline{\neg\xi(\overline{k})}) \prec^{\#} (\exists y)\neg\chi(\overline{k}, y).$$

Properties: $\xi(x)$ is Σ_n in $I\Sigma_1$; for all k, m, $(T + \xi(\overline{k}))\chi(\overline{k}, \overline{m})$; for all k, $(T_0 + \xi(\overline{k})$ is Π_n-conservative for $T + \{\chi(\overline{k}, \overline{m}) \mid m\}$.

(3) $$I\Sigma_1 \vdash \xi \equiv (\exists y)\neg\chi(\overline{\neg\xi}, y) \prec^{\#} Pr^{\bullet}_{\Pi_n}(\overline{\xi})\,.$$

Properties: $\neg\xi$ is Π_n in $I\Sigma_1$; for all m, $(T + \neg\xi) \vdash \chi(\overline{\neg\xi}, \overline{m})$; $(T_0 + \neg\xi)$ is Σ_n-conservative for $T_0 + \{\chi(\overline{\neg\xi}, \overline{m}) \mid m\}$.

(4) $$I\Sigma_1 \vdash \xi(\overline{k}) \equiv (\exists y)\neg\chi(\overline{k}, y) \prec^{\#} Pr^{\bullet}_{\Pi_n}(\overline{\xi(\overline{k})})\,.$$

Properties: $\neg\xi(x)$ is Π_n in $I\Sigma_1$; for all k, m, $(T + \neg\xi(\overline{k})) \vdash \chi(\overline{k}, \overline{m})$; $(T_0 + \neg\xi(\overline{k}))$ is Σ_n-conservative for $T_0 + \{\chi(\overline{k}, \overline{m}) \mid m\}$.

Proofs are obvious modifications of 4.41. □

4.43 Theorem. Let Γ be Σ_n or Π_n, $n \geq 1$. For each $X \in \Sigma_1$ there is a Γ-formula ξ such that, for each k,

$$k \in X \text{ iff } T \vdash \neg\xi(\overline{k})\,,$$

$k \in X$ iff $\xi(\overline{k})$ is hereditarily $\tilde{\Gamma}$-conservative over T. ($\tilde{\Gamma}$ is the dual of Γ).

Proof. In the preceding theorem, let $(\exists y)\neg\chi(x, y)$ define X. For $\Gamma = \Sigma_n$ use (2). Then: if $k \in X$ and m is a witness for $(\exists y)\neg\chi(\overline{k}, y)$ then

$$T + \xi(\overline{k}) \vdash \chi(\overline{k}, \overline{m}) \,\&\, \neg\chi(\overline{k}, \overline{m})\,;$$

if $k \notin X$, i.e. for each m $I\Sigma_1 \vdash \chi(\overline{k}, \overline{m})$, then $T + \xi(\overline{k})$ is hereditarily Π_n-conservative over T.

Similarly for $\Gamma = \Pi_n$ use 4.42 (4). □

4.44 Corollary (= 4.26 (3)). There is no $X \in \Sigma_1$ such that

$$hConsv(\Gamma) \subseteq X \subseteq NRef\,.$$

Proof. Assume we have such an X and in the preceding theorem (with Γ and $\tilde{\Gamma}$ interchanged) let φ be such that $I\Sigma_1 \vdash \varphi \equiv \xi(\overline{\varphi})$. We show $\varphi \notin X$; but then 4.43 gives $\varphi \in hConsv(\Gamma)$, in contradiction to the inclusion assumed.

Thus let $\varphi \in X$; then $\varphi \in NRef$ (since $X \subseteq NRef$), but $T \vdash \neg\xi(\overline{\varphi})$, thus $T \vdash \neg\varphi$ (since $\varphi \in X$). A contradiction. □

4.45. We prove theorem 4.25 by inspecting the respective self-referential formulas.

(1) In 4.42 (1) let $(\exists y)\neg\chi(x,y)$ define the empty set, i.e. $I\Sigma_1 \vdash \chi(x,y) \equiv \overline{0} = \overline{0}$. Then $I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\overline{\neg\xi})$ (thus ξ says: my negation is provable in T from a true Σ_n-formula; in other words: I imply in T a false Π_n-formula), $I\Sigma_1 \vdash \chi(\overline{\xi}, \overline{k})$ for each k, thus ξ is hereditarily Π_n-conservative over T and clearly ξ is unprovable in T provided T is Σ_n-sound.

Lemma. Assume $T \vdash Pr(x) \to \Delta(x) \to Pr_\Gamma(x)$, and $T \vdash Pr(x) \to \nabla(x) \to Pr_\Lambda(x)$, $T \vdash \xi \equiv \Delta(\overline{\neg\xi}) \prec^{\#} \nabla(\overline{\xi})$. Then neither ξ nor $\neg\xi$ is T-provable.

A standard proof (using 4.40) is left to the reader as an exercise.

(2) In 4.42 (1) put $\chi(x,y) \equiv \neg Proof^\bullet(y,x)$, i.e. $(\exists y)\neg\chi(x,y)$ is equivalent to $Pr^\bullet(x)$ and

$$I\Sigma_1 \vdash \xi \equiv Pr^\bullet_{\Sigma_n}(\overline{\neg\xi}) \prec^{\#} Pr^\bullet(\overline{\xi})\,.$$

Then $I\Sigma_1 \vdash \chi(\overline{\xi}, \overline{m})$ for each m, $\neg\xi$ is unprovable and hereditarily Σ_n-conservative over T.

(3) In 4.42 (3) let $\chi(x,y)$ be as above, thus

$$I\Sigma_1 \vdash \xi \equiv Pr^\bullet(\overline{\neg\xi}) \prec^{\#} Pr^\bullet_{\Pi_n}(\overline{\xi})\,.$$

Then $I\Sigma_1 \vdash \chi(\overline{\neg\xi}, \overline{m})$ for each m, $\neg\xi$ is unprovable and hereditarily Σ_n-conservative over T.

(4) In 4.42(1) take $\neg Proof^\bullet_{\Pi_n}(y,x)$ for $\chi(x,y)$ and, at the same time, in 4.42 (3) take $\neg Proof^\bullet_{\Sigma_n}(y,x)$ for $\chi(x,y)$. In either case we get

$$I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\overline{\neg\xi}) \prec^{\#} Pr_{\Pi_n}(\overline{\xi})\,.$$

Thus $T + \xi \vdash \neg Proof_{\Pi_n}(\overline{m}, \overline{\xi})$ by 4.42 (1), $T + \neg\xi \vdash \neg Proof^\bullet_{\Pi_n}(\overline{m}, \overline{\xi})$ by 4.40, thus 4.42 (1) gives Π_n-conservativity of ξ over T. (Nothing is claimed on hereditary conservativity: for a general T_0 between $I\Sigma_1$ and T we cannot guarantee $T_0 \vdash \neg Proof^\bullet_{\Pi_n}(\overline{m}, \overline{\xi})$.)

Similarly, 4.42 (3) and 4.40 gives $T \vdash \neg Proof^\bullet_{\Sigma_n}(\overline{m}, \overline{\xi})$ for each m and consequently $\neg\xi$ is Σ_n-conservative over T.

4.46 Theorem. Let T be Σ_n or Π_n, $n \geq 2$. Let X be Π_2. Then there is a Γ-formula $\xi(x)$ such that, for each k,

$$\begin{aligned} k \in X &\equiv \xi(\overline{k}) \text{ is hereditarily } \tilde{\Gamma}\text{-conservative}\,, \\ k \notin X &\equiv \xi(\overline{k}) \text{ is } \Sigma_1\text{-nonconservative}\,. \end{aligned}$$

The same is true for Σ_1 replaced by Π_1 (with another ξ).

Proof. Let X be $\{k \mid (\forall m)R(k,m)\}$ where R is Σ_1. By 4.39, let $\rho(x,y)$ be a Σ_1-numeration (Π_1-numeration) of R both in $I\Sigma_1$ and in T. In 4.42 (2), (4) let $\chi(x,y)$ be $\rho(x,y)$.

If $\Gamma = \Sigma_n$ let $\xi(x)$ be the formula $\xi(x)$ from 4.42 (2); if $\Gamma = \Pi_n$ let $\xi(x)$ be the formula $\neg\xi(x)$ from 4.42 (4). Then $\xi(x)$ is Γ; if $k \in X$ then $I\Sigma_1 \vdash \rho(\overline{k}, \overline{m})$ for all m, hence $\xi(\overline{k})$ is hereditarily Γ-conservative over T. If $k \notin X$ then, for some m_0, T does not prove $\rho(\overline{k}, \overline{m}_0)$; thus $\xi(\overline{k})$ is not Σ_1-conservative (Π_1-conservative) over T. □

4.47 Theorem (= 4.26). Let $\Gamma = \Sigma_n$ or Π_n, $n \geq 2$. There is no $Y \in \Sigma_2$ such that $hConsv(\Gamma) \subseteq Y \subseteq Consv(\Sigma_1)$ or $hConsv(\Gamma) \subseteq Y \subseteq Consv(\Pi_1)$.

Proof. Assume that Y is such set. Then 4.46 (with $X = N - Y$) gives ($k \notin Y$ iff $\xi(\overline{k})$ is hereditarily Γ-conservative). Let $I\Sigma_1 \vdash \varphi \equiv \xi(\overline{\varphi})$; then $\varphi \in Y$ implies that φ is not Σ_1-conservative but $Y \subseteq Consv(\Sigma_1)$ a contradiction. On the other hand, $\varphi \notin Y$ implies that φ is hereditarily Γ-conservative but $hConsv(\Gamma) \subseteq Y$, a contradiction. □

4.48 Theorem (= 4.27). For each $\Gamma = \Sigma_n, \Pi_n$ ($n \geq 1$), both $Consv(\Gamma)$ and $hConsv(\Gamma)$ is Π_2-complete.

Proof. Clearly, $Consv(\Gamma)$ is Π_2; to see that $hConsv(\Gamma)$ is Π_2, observe that φ is hereditarily Γ-conservative over T iff for each *finite* $T_0 \subseteq T$ containing $I\Sigma_1$, φ is Γ-conservative over T_0. For each Γ (including Σ_1 and Π_1), 4.46 gives a formula $\xi(x)$ such that

$$k \in X \Rightarrow \xi(\overline{k}) \text{ is hereditarily } \Gamma\text{-conservative},$$
$$k \notin X \Rightarrow \xi(\overline{k}) \text{ is not } \Gamma\text{-conservative}.$$
□

4.49. Now we present Švejdar's proof of the fact (claimed in 4.5 (3)) that for a T not Σ_1-sound, the negation of Rosser's formula is Σ_1-nonconservative over T.

Let $\delta(x)$ be a Σ_0 formula such that $T \vdash (\exists x)\delta(x)$ but $N \models (\forall x)\neg\delta(x)$; let ρ be Rosser's formula, thus $T \vdash \rho \equiv (Pr^\bullet(\overline{\neg\rho}) \prec Pr^\bullet(\overline{\rho}))$. We know the following: both ρ and $\neg\rho$ are unprovable in T;

$$T + Con^\bullet \vdash \neg Pr^\bullet(\rho) \,\&\, \neg Pr^\bullet(\overline{\neg\rho})$$

(this is just a formalized version of the preceding assertion). Let

$$T \vdash \varphi \equiv ((\exists x)\delta(x) \prec Pr^\bullet(\overline{\rho \rightarrow \varphi}))\,.$$

(a) $\rho \rightarrow \varphi$ is T-unprovable. Indeed, if d is a T-proof of $\rho \rightarrow \varphi$ then $T \vdash \bigvee_{k \leq d} \delta(\overline{k})$; but for each k, $T \vdash \neg\delta(\overline{k})$.
(b) $T + Con^\bullet \vdash \varphi$. Indeed, $T \vdash (\exists x)\delta(x)$, thus $(T + \neg\varphi)$ proves $Pr^\bullet(\overline{\rho \rightarrow \varphi}) \preccurlyeq (\exists x)\delta(x)$. (This is a Σ_1-formula, call it σ), thus $T + \neg\varphi \vdash Pr^\bullet(\overline{\sigma})$.

But $T \vdash Pr^\bullet(\overline{\sigma \to \neg\varphi})$, thus $T + \neg\varphi \vdash Pr^\bullet(\overline{\neg\varphi})$ & $Pr^\bullet(\overline{\rho \to \varphi})$, hence $T + \neg\varphi \vdash Pr^\bullet(\overline{\neg\rho})$, which implies $T + \neg\varphi \vdash \neg Con^\bullet$.

(c) $T + \neg\rho \vdash \varphi \vee (Pr^\bullet(\overline{\rho}) \preccurlyeq Pr^\bullet(\overline{\neg\rho}))$. This because $(T + \neg\rho + Con^\bullet)$ proves ρ and $(T + \neg\rho + \neg Con^\bullet)$ proves $(Pr^\bullet(\overline{\rho}) \preccurlyeq Pr^\bullet(\overline{\neg\rho}))$.

(d) But $\varphi \vee (Pr^\bullet(\overline{\rho}) \preccurlyeq Pr^\bullet(\overline{\neg\rho}))$ is unprovable in T. Assume T proves this disjunction; then $T + \rho$ would prove φ (since $T + \rho$ disproves the second disjunct); this contradicts (a). Thus $\varphi \vee (Pr^\bullet(\overline{\rho}) \preccurlyeq Pr^\bullet(\overline{\neg\rho}))$ is the desired Σ_1 formula.

(c) Applications, Mainly to Interpretability

The subsection has the following structure: main results are formulated in 4.50–4.59 and compare (or, better, contrast) (i) interpretability in PA with interpretability in its conservative second order extension ACA_0 (4.55) and (ii) interpretability with partial conservativity for finitely axiomatized theories $T \supseteq I\Sigma_1$ (4.56–4.57). Proofs are elaborated in 4.50–4.65; this part starts with Lindström's second fixed point theorem, which is rather technical but of independent interest. The proofs combine tricky self-reference, partial truth definitions and use of definable cuts. The rest of the subsection (4.66–4.69) contains some additional results on partial conservativity.

4.50 Definition. Put $Intp_T = \{\varphi \mid (T + \varphi) \text{ is interpretable in } T\}$. Recall that we say "$\varphi$ is interpretable in T" instead of "$(T + \varphi)$ is interpretable in T". Further recall Theorem 3.46 telling us that, for T sequential and with full induction, $Intp_T = Consv_T(\Pi_1)$; thus Theorem 4.27 gives inmediately the following

4.51 Corollary. If T is sequential and with full induction, then $Intp_T$ is Π_2-complete.

4.52 Lemma. If $T \supseteq I\Sigma_1$ is finitely axiomatized then $Intp_T$ is Σ_1.

Proof. This is more or less evident; φ is interpretable in T iff there are definitions of all (finitely many) symbols of T (in the sense of the interpretation) in T and T-proofs of translations of all (finitely many) axioms of the theory $(T+\varphi)$. Using finite sequences we may write down a formula $Interp^\bullet_T(x, y)\Delta_1$ in $I\Sigma_1$ such that

$$\varphi \in Intp_T \text{ iff } N \models (\exists y) Interp^\bullet_T(\overline{\varphi}, y);$$

$Interp^\bullet_T$ just says that y is a sequence consisting of all those finitely many definitions and proofs. (cf. 2.42). □

4.53 Theorem. If $T \supseteq I\Sigma_1$ is finitely axiomatized, $n \geq 1$ and Γ is Σ_n or Π_n then $Consv_T(\Gamma) - Intp_T \neq \emptyset$; there is a $\tilde{\Gamma}$-sentence φ which is (hereditarily) Γ-conservative but is not interpretable in T.

Proof. This follows directly by 4.26, since clearly $Intp_T \subseteq NRef_T$ and $Intp_T \in \Sigma_1$, i.e. $hConsv_T(\Gamma)$ is not a subset of $Intp_T$. Checking the proof of 4.26 (3) (i.e. 4.44) we can see that the sentence constructed there is $\tilde{\Gamma}$. □

4.54 Discussion. Take ACA_0 for T and Π_1 for Γ. We get a Σ_1 formula φ that is Π_1-conservative over ACA_0 but $(ACA_0 + \varphi)$ is not interpretable in ACA_0. Since φ is Π_1-conservative over ACA_0 and is Σ_1, i.e. a formula in the language of PA, φ is Π_1-conservative over PA and hence, by 2.40, is interpretable in PA. (Similarly for $T = GB$, i.e. Gödel-Bernays set theory and ZF-Zermels-Frankel set theory, using 3.46). Note that we can take the following self-referential formula for φ:

$$T \vdash \varphi \equiv Pr^{\bullet}_{\Sigma_1}(\overline{\neg\varphi}) \prec (\exists y) Interp^{\bullet}(\overline{\varphi}, y)\,.$$

This is slightly simpler than the formula obtained directly from 4.53, i.e. from 4.26.

Compare the present result with 4.16 and summarize:

4.55 Corollary. (1) There is a Σ_1-formula φ such that $(PA+\varphi)$ is interpretable in PA but $(ACA_0 + \varphi)$ is not interpretable in ACA_0.

(2) There is a Π_1-formula φ such that $(ACA_0+\varphi)$ is interpretable in ACA_0 but $(PA + \varphi)$ is not interpretable in PA.

(Similarly for ZF and GB instead of PA and ACA_0.)

4.56 Remark. We shall analyze the present situation more deeply; we shall present a classification of independent Σ_1-sentences φ (over an arbitrary consistent finitely axiomatized sequential theory $T \supseteq I\Sigma_1$) – with respect to the following questions. (i) is φ interpretable in T? (ii) is $\neg\varphi$ interpretable in T? (iii) is φ Π_1-conservative over T? We shall again get corollaries for $T = ACA_0$, GB and their relation to PA and ZF respectively.

Thus throughout the subsection, T is a consistent finitely axiomatized sequential theory containing $I\Sigma_1$. The question (i), (ii), (iii) admit eight possible combinations of answers and give eight possible types of independent Σ_1-sentences. Till now, we have got some examples; e.g. for $\neg Con^{\bullet}$ the answers are (yes, yes, no) and for the H-Rosser formula $H\rho$ (Rosser with respect to Herbrand proofs) the answers are (no, yes, yes). We are going to prove the following.

4.57 Theorem. There are independent Σ_1-sentences of all eight possible types.

The theorem is an inmediate consequence of Theorem 4.64 below. We shall construct examples of sentences of all eight types in a rather uniform way, using substantially results from Sect. 3.

4.58 Corollary. Take again $T = ACA_0$; recall that for each PA-sentence φ, φ is Π_1-conservative over PA iff φ is Π_1-conservative over ACA_0 iff $(PA+\varphi)$ is interpretable in PA. Consequently, if σ is an independent Σ_1-sentence then $(PA + \neg\sigma)$ is never interpretable in PA. Hence we have eight possible types of dependent Σ_1-sentences according be the following questions:

- is $(PA + \sigma)$ interpretable in PA?
- is $(ACA_0 + \sigma)$ interpretable in ACA_0?
- is $(ACA_0 + \neg\sigma)$ interpretable in ACA_0?

Similarly for GB and ZF.

4.59 Remark. (1) We leave open the question which possibilities we have for Σ_1(non)conservativity of $\neg\sigma$ in the case of T being Σ_1-ill; in combination with the eight types above there at most 16 possible types. The reader may investigate this as an exercise.

(2) This is the end of the statement of results of the present subsection. The rest contains an elaboration, including Lindström's second fixed point theorem.

*

4.60 Lindström's Second Fixed Point Theorem. Let $T \supseteq I\Sigma_1$ and let $m, n \geq 0$. Let $\varphi(x, y)$ be a Σ_n-formula and $\theta(x, y)$ a Π_k-formula. Let ξ satisfy

$$T \vdash \xi \equiv [Pr^{\bullet}_{\Sigma_n}(\overline{\neg\xi}) \vee (\exists y)\varphi(\overline{\neg\xi}, y) \prec^{\#} Pr^{\bullet}_{\Pi_k}(\overline{\xi}) \vee (\exists y)\theta(\overline{\xi}, y)]$$

(where $Pr^{\bullet}_{\Sigma_n}$, $Pr^{\bullet}_{\Pi_k}$ are as in 4.23). Then

(1) $(T + \xi) \vdash (\exists y \leq \overline{m})\theta(\overline{\xi}, y) \to (\exists y \leq \overline{m})\varphi(\overline{\neg\xi}, y)$ for each m,
(2) $(T + \neg\xi) \vdash (\exists y \leq \overline{m})\varphi(\overline{\neg\xi}, y) \to (\exists y \leq \overline{m})\theta(\overline{\xi}, y)$ for each m;
(3) $(T + \xi)$ is Π_n-conservative for $T + \{\neg\theta(\overline{\xi}, \overline{m}) \mid m\}$,
(4) $(T + \neg\xi)$ is Σ_k-conservative for $T + \{\neg\varphi(\overline{\neg\xi}, \overline{m}) \mid m\}$.

Furthermore, ξ is $\Sigma_{\max(k,n,1)}$ in T.

The assertion remains valid if $Pr^{\bullet}$ is replaced by $HPr^{\bullet}$ (Herbrand provability).

Proof. (1) Let us work in $(T + \xi + (\exists y \leq \overline{m})\theta(\overline{\xi}, y))$. The succedent of ξ has a witness $\leq \overline{m}$; this implies in the usual way that the antecedent of ξ has also such a witness; thus $(\exists y < \overline{m})(Proof^{\bullet}_{\Sigma_n}(\overline{\neg\xi}) \vee \varphi(\overline{\neg\xi}, y))$. But this y is not a witness for $Pr^{\bullet}_{\Sigma_n}(\overline{\neg\xi})$ (otherwise we would get ξ by 4.40 or by the variant of 4.40 using $HPr^{\bullet}$). Thus we get $(\exists y \leq \overline{m})\varphi(\overline{\neg\xi}, y)$.

(2) is proved analogously.

(3) Assume $(T+\xi) \vdash \pi$ where π is Π_n, let d be a T-proof of $(\neg\pi \rightarrow \neg\xi)$. Let us work in $(T + \{\neg\theta(\overline{\xi}, \overline{m}) \mid m\} + \neg\xi)$; we want to prove π. Assume $\neg\pi$, thus $Tr(\overline{\neg\pi})$, i.e. $\overline{\neg\pi}$ is a true Σ_n-sentence; but then $Proof^{\bullet}_{\Sigma_n}(\overline{d}, \neg\overline{\xi})$ and, since we assume $\neg\xi$, beneath $\overline{d}$ there is a witness for the succedent of ξ. But beneath $\overline{d}$ there is no witness for $(\exists n)\theta(\overline{\xi}, u)$ (since $\neg\theta(\overline{\xi}, \overline{0}), \neg\theta(\overline{\xi}, \overline{1}), \ldots, \neg\theta(\overline{\xi}, \overline{d})$). Neither is there a witness for $Pr^{\bullet}_{\Pi_k}(\overline{\xi})$ beneath $\overline{d}$ since this would imply ξ by 4.40. This gives a contradiction.

(4) similarly. □

4.61 Remark. For $n = 0$ we may replace $Pr^{\bullet}_{\Sigma_0}$ by $Pr^{\bullet}$ and/or allow φ to be Δ_1 (in $I\Sigma_1$, say) instead of being Σ_0. Similarly for $k = 0$.

4.62 Corollary. In 4.60, let m, n be arbitrary but let φ, θ be Σ_1-in-$I\Sigma_1$. (By 4.61, this does not exclude the possibility $m = 0$ and /or $n = 0$). Let X, Y be Σ_1-sets defined by $(\exists y)\theta(x, y)$ and $(\exists y)\varphi(x, y)$ respectively and assume $X, Y \subseteq NRef$ (non-refutable formulas). Then the sentence ξ from 4.60 is Π_n-conservative, $\neg\xi$ is Σ_k-conservative (over T), $\xi \notin X$ and $(\neg\xi) \notin Y$.

Proof. Without loss of generality we may assume that, for each k, no m witnesses both $k \in X$ and $k \in Y$, i.e. $N \models (\forall x, y)(\neg\varphi(x, y) \vee \theta(x, y))$ (assume e.g. that all witnesses for X are even and all witnesses for Y are odd). Take ξ from 4.60, i.e.

$$I\Sigma_1 \vdash \xi \equiv Pr^{\bullet}_{\Sigma_n}(\neg\overline{\xi}) \vee (\exists y)\varphi(\neg\overline{\xi}, y) \prec^{\#} Pr^{\bullet}_{\Pi_k}(\overline{\xi}) \vee (\exists y)\theta(\overline{\xi}, y) .$$

First show $\xi \notin X$ and $(\neg\xi) \notin Y$. Assume the contrary and let m be the least witness for $\xi \in X \vee (\neg\xi) \in Y$. Thus either N satisfies (and T proves)

$$\theta(\overline{\xi}, \overline{m}) \mathbin{\&} (\forall x < \overline{m})\neg\theta(\overline{\xi}, x) \mathbin{\&} (\forall x \leq \overline{m})\neg\varphi(\neg\overline{\xi}, x) ,$$

or N satisfies and T proves the formula resulting from the last one by interchanging θ, $\overline{\xi}$ with φ, $\overline{\neg\xi}$ respectively. In the former case we get a contradiction in $(T + \xi)$ using 4.60 (1), which contradicts our assumption $X \subseteq NRef$; in the latter case 4.60 (2) gives a contradiction in $(T + \neg\xi)$, which contradicts $Y \subseteq NRef$. Thus we have $\xi \notin X$ and $\xi \notin Y$. This implies $T \vdash \neg\theta(\overline{\xi}, \overline{m})$ for each m and therefore 4.60 (3), (4) give the desired conservation. □

4.63 Discussion and Definition. Now we shall formulate and prove a theorem implying immediately our Theorem 4.57. We shall deal with formulas $HPr^{\bullet}(x)$, $HPr^{\bullet}_{\Sigma_1}(x)$ (Herbrand provability in T; Herbrand provability in $T^{\bullet}$ from a true Σ_1-sentence) as well as with the formula $Intp^{\bullet}(x) \equiv (\exists y)Interp^{\bullet}(x, y)$ from 4.52 (interpretability of $(T^{\bullet} + x)$ in $T^{\bullet}$). We shall investigate self-referential Σ_1-sentences ξ such that

$$I\Sigma_1 \vdash \xi \equiv \Delta(\overline{\neg\xi}) \prec \nabla(\overline{\xi}) ,$$

where the antecedent $\Delta(x)$ has one of the forms

$$HPr^{\bullet}(x),\quad HPr^{\bullet}_{\Sigma_1}(x),\quad HPr^{\bullet}(x) \vee Intp^{\bullet}(x),\quad HPr^{\bullet}_{\Sigma_1}(x) \vee Intp^{\bullet}(x),$$

and the succedent $\nabla(x)$ has one of the forms

$$HPr^{\bullet}(x),\quad HPr^{\bullet}(x) \vee Intp^{\bullet}(x).$$

This will give eight cases, which are examples of our eight types of independent Σ_1-sentences. For some of the following assertions it is immaterial whether we use $HPr^{\bullet}$ or $Pr^{\bullet}$, but for some (using Theorem 3.20) it is not.

Observe that all eight cases are particular cases of 4.62 (with the convention 4.61 applied) for $n = 0$ or 1, $k = 0$ and X, Y being either empty or $Intp_T$; we use Herbrand provability.

4.64 Theorem. (1) All cases give independent Σ_1-sentences;
(2) ξ is interpretable iff $X = \emptyset$ (the succedent does not contain $Intp^{\bullet}$);
(3) $\neg\xi$ is interpretable iff $Y = \emptyset$ (the antecedent does not contain $Intp^{\bullet}$);
(4) ξ is Π_1-conservative iff $n = 1$ (the antecedent contains $HPr^{\bullet}_{\Sigma_1}$).
Remark. This obviously implies Theorem 4.57.

Proof. 4.62 gives directly the following: ξ is independent (since ξ is Π_n-conservative and $\neg\xi$ is Σ_k-conservative); if the antecedent contains $HPr^{\bullet}_{\Sigma_1}$ then ξ is Π_1-conservative; if $X = Intp_T$ then ξ is not interpretable; if $Y = Intp_T$ then $\neg\xi$ is not interpretable. Thus it remains to prove the converse implications in (2)–(4).

(2) Assume $X = \emptyset$; we prove $\xi \in Intp_T$. (Here it is vital to work with Herbrand provability.) We have $T \vdash \xi = \Delta(\overline{\neg\xi}) \prec HPr^{\bullet}(\overline{\xi})$. It suffices to interpret $(T + \xi)$ in $(T + \neg\xi)$. By 3.20 there is a $(T + \neg\xi)$-cut J such that $(T + \neg\xi) \vdash HCon^{\bullet J}(T^{\bullet} + \overline{\neg\xi})$. Let us work in $(T + \neg\xi)$. In J there is no H-proof$^{\bullet}$ of $\overline{\xi}$ in $T^{\bullet}$. But $\neg\xi$ implies that beneath each witness for $\Delta(\neg\xi)$, in particular, beneath each H-proof$^{\bullet}$ of $\overline{\neg\xi}$ in $T^{\bullet}$ there is a H-proof$^{\bullet}$ of $\overline{\xi}$ in $T^{\bullet}$, thus in J there is no H-proof$^{\bullet}$ of $\neg\overline{\xi}$ in $T^{\bullet}$ and hence $HCon^{\bullet J}(T^{\bullet} + \overline{\xi})$. By 3.39, $(T + \overline{\xi})$ is interpretable $(T + \overline{\neg\xi})$.

(3) Assume $Y = \emptyset$; we prove $(\neg\xi) \in Intp_T$. We have $T \vdash \xi \equiv HPr^{\bullet}_{\Gamma}(\overline{\neg\xi}) \prec \nabla(\overline{\xi})$ where $\Gamma = \Sigma_0$ or Σ_1. It suffices to interpret $(T + \neg\xi)$ in $(T + \xi)$. Let us work in $(T + \xi)$. By 3.30 we have a cut J such that $(\forall u \in Tr(\Sigma_1))HCon^{\bullet J}(T^{\bullet} + \overline{\xi} + u)$, i.e. J does not contain any H-proof$^{\bullet}$ of $(\neg T^{\bullet} \vee \neg\overline{\xi} \vee u)$. But ξ implies that there is a H-proof$^{\bullet}$ y of $\overline{\neg\xi}$ from T and from a true Σ_1-formula u_0 (i.e. a H-proof$^{\bullet}$ of $\neg T^{\bullet} \vee \neg u_0 \vee \overline{\neg\xi}$) such that beneath y there is no witness for $\nabla(\xi)$, in particular, no H-proof$^{\bullet}$ of $\overline{\xi}$ from $T^{\bullet}$, hence we have $HCon^{\bullet J}(T' + \overline{\neg\xi})$; by 3.39, $(T + \neg\xi)$ is interpretable in $(T + \xi)$.

(4) Assume $n = 0$, i.e. $HPr^{\bullet}_{\Sigma_1}$ does not occur in ∇. We prove that ξ is Π_1-nonconservative. Clearly, ξ implies (in T) the formula saying "beneath

each witness for $\nabla(\overline{\xi})$ there is a witness for $\nabla(\overline{\neg\xi})$" (call it π) and under our assumptions π is Π_1. It suffices to show that π is unprovable in T. This is done analogously to the corresponding proof concerning Rosser's formula. If π were T-provable we would have

$$T \vdash (\text{there exists a witness for } \nabla(\overline{\xi}) \to \xi)$$

(consider the least witness for $\nabla(\overline{\neg\xi})$, which exists by $I\Sigma_1$). Thus we would have

$$\begin{aligned}
&T \vdash \neg\xi \to (\nabla(\overline{\xi}) \text{ has no witness}),\\
&T \vdash \neg\xi \to \neg HPr^\bullet(\overline{\xi}),\\
&T \vdash \neg\xi \to HCon^\bullet(T^\bullet + \overline{\neg\xi}),\\
&T \vdash \neg\xi \to Con^\bullet(T^\bullet + \overline{\neg\xi})
\end{aligned}$$

(due to the provability of Herbrand's theorem in $I\Sigma_1$), which contradicts Gödels's second incompleteness theorem. This completes the proof of 4.66 and of 4.57. □

*

To close the present section we add some few other applications of partial conservativity. $T \in \Delta_1$ is a consistent theory containing $I\Sigma_1$.

4.66 Theorem (speed up). Let f be a Δ_1 function mapping N into N, let $\Gamma = \Sigma_k$ or Π_k ($k \geq 1$). There is a φ Γ-conservative over T and such that, for each n, there is a proof $d > n$ of a Γ-sentence γ in $(T + \varphi)$ such that each T-proof of γ is bigger than $f(d)$.

Proof. Otherwise $Consv(\Gamma)$ would be Σ_2, which contradicts 4.27. (Since a Π_2-complete set is not Σ_2.) □

4.67 Definition and discussion. A sentence φ is a *self-prover* if $T \vdash \varphi \to Pr^\bullet(\varphi)$. Clearly, each Σ_1-sentence is a self-prover. We show that there are other self-provers as well. It is easily seen that for each sentence ψ, the formula ψ & $Pr^\bullet(\overline{\psi})$ is a self-prover; it φ is a self-prover then $T \vdash \varphi \equiv (\varphi \,\&\, Pr^\bullet(\overline{\varphi}))$.

4.68 Theorem. For each $\Gamma = \Sigma_n$ ($n \geq 1$) or $\Gamma = \Pi_m$ ($m \geq 2$) there is a self-prover which is exactly of complexity Γ.

Proof. By 4.26 (3), let ψ be such that $\neg\psi$ is Γ-conservative over T and not provable in $(T + \neg Con^\bullet)$ (i.e. take $X = \{\alpha \mid (T + \neg Con) \vdash \alpha\}$). Let φ be ψ & $Pr^\bullet(\overline{\psi})$. By our assumption on Γ, φ is Γ in T. We prove that φ is not $\tilde{\Gamma}$ (dual class). Assume $T \vdash \varphi \equiv \sigma$, $\sigma \in \tilde{\Gamma}$. Then $T \vdash \varphi \to \neg\sigma$, hence $T \vdash \neg\psi \to \neg\sigma$, and since $\neg\psi$ is Γ-conservative we get $T \vdash \neg\sigma$, thus $T \vdash \neg\varphi$,

$T \vdash \neg(\psi \,\&\, Pr^\bullet(\overline{\psi}))$, $T \vdash Pr^\bullet(\overline{\psi}) \to \neg\psi$, $T \vdash \neg Con^\bullet \to \neg\psi$, which contradicts our assumption. $\square$

4.69 Remark. Let Γ be Σ_n or Π_n, let Λ be Σ_m or Π_m, assume that Λ does not include Γ. Then there is a formula φ exactly Γ (not $\tilde{\Gamma}$ in T) and hereditarily exactly Λ-conservative (not $\tilde{\Lambda}$-conservative). For example, let $\Gamma = \Sigma_8$, let Λ be Σ_5; let π be Π_5, independent and hereditarily Σ_5-conservative over T and let σ be Σ_8, independent, hereditarily Π_8-conservative over $(T + \pi)$. Then take $\varphi \equiv \pi \,\&\, \sigma$.