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 $\Sigma_1$ 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 $\varphi$ 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 $\Gamma$ of formulas, e.g. $\Sigma_k$-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 preliminaries, 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 $\text{ACA}_0$ – finitely axiomatizable second order arithmetic extending $\text{PA}$ 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 $\Delta_1$.

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 $\Gamma$ be a class of formulas. $T$ is $\Gamma$-conservative over $S$ if each $\Gamma$-formula $\varphi$ of $L$ provable in $T$ is provable in $S$.

(3) $S$ is a $\Gamma$-conservative extension of $T$ if $S$ is an extension of $T$, i.e. $T$ is a subtheory of $S$, and, in addition, $S$ is $\Gamma$-conservative over $T$; i.e. if any formula is $T$-provable then it is $S$-provable; and if a $\Gamma$-formula is $S$-provable then it is $T$-provable. (Thus $S$ is stronger then $T$ but for $\Gamma$-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 $\psi_P(x_1 \ldots x_k)$ of $L_2$ with exactly $n$ free variables,
- for each $n$-ary function symbol $F$ of $L_1$, a formula $\psi_F(x_1 \ldots x_k, y)$ of $L_2$ with exactly $k + 1$ free variables such that
1. Preliminaries

\[ S \vdash \bigwedge_{i=1}^{k} \chi(x_i) \rightarrow (\exists y)(\chi(y) \& \psi_F(x_1 \ldots x_k, y)) \]

and similarly for \( \psi_c \) (\( S \vdash (\exists y)(\chi(y) \& \psi_c(y)) \)).

Using \( \psi_P \), \( \psi_F \) and \( \psi_c \) we may define in \( S \) a predicate \( P^* \), a function \( F^* \) (defining \( F^*(x_1 \ldots x_k) \) by \( \psi_F \) if \( \bigwedge \chi(x_i) \) 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); \( \varphi^* \) results from \( \varphi \) by replacing each predicate, function, constant by its starred counterpart. The formula \( \varphi^* \) may be called "\( \varphi \) in the sense of the interpretation" or "the translation of \( \varphi \)."

(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 \( T \),

\[ S \vdash \bigwedge_{i=1}^{k} \chi(x_i) \rightarrow \varphi^*. \]

We have the following evident theorem:

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

\[ \text{Proof by induction on the length of a proof.} \]

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

\[ S \vdash (\vartheta(z) \& \bigwedge \chi(x_i, z)) \rightarrow \varphi^*(x, z). \]

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

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

(3) Note that if \( * \) is parametrical and \( S \) can define an object satisfying \( \vartheta \) (i.e. for some \( \vartheta'(z) \), \( S \vdash \vartheta'(z) \rightarrow \vartheta(z) \) and \( S \vdash (\exists ! z) \vartheta'(z) \) then \( * \) may be replaced by a non-parametrical interpretation \( ** \) such that, for each closed \( T \)-formula \( \varphi \), \( S \vdash \varphi^* \rightarrow \varphi^{**} \) (just postulate that \( z \) is the unique tuple satisfying \( \vartheta' \) 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 \((\chi_i, \psi_{P,i}, \ldots)\) be the interpretation in question; \((i = 1, 2)\). Put \( \chi(x) = (\varphi \wedge \chi_1(x)) \vee (\neg \varphi \wedge \chi_2(x)) \),

\[ \psi_P(x) = (\varphi \wedge \psi_{P,1}(x)) \vee (\neg \varphi \wedge \psi_{P,2}(x)) \]

etc. Similarly for parametrical interpretations. \( \square \)

1.9 Definition. An interpretation \( * \) of \( T \) in \( S \) is \( \Gamma \)-faithful if, for each closed \( \varphi \in \Gamma \), \( T \vdash \varphi \) is equivalent to \( S \vdash \varphi^* \). It is faithful if it is \( \Gamma \)-faithful for \( \Gamma \) 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) $SEQ(z, u) \rightarrow Number(u) \& (\forall v < u)(Number(v) \rightarrow (\exists !x)\beta(x, v, z))$;
(2) $SEQ(z, u) \rightarrow (\forall y)(\exists z')(SEQ(z', u + 1) \& (\forall v < u)(Number(v) \rightarrow (\forall x)(\beta(x, v, z) \equiv \beta(x, v, z')) \& \beta(y, u, z'))$.

(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 important 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 $IΣ^\text{exp}_0$ is sequential; we may take $Seq(s)$ if $lh(s) = u$ for $SEQ(s, u)$ and $(s)_0 = x$ for $\beta(x, v, s)$. In Chap. V we show that each theory in the language of arithmetic containing $IΣ_0$ 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 $0$.

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 $0$ 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 $\varphi(x, y, X)$ *not containing any second order quantifiers*, and not containing the variable $Z$, the axiom

\[
(\exists Z)(\forall x)(x \in Z \equiv \varphi(x, y, X))
\]

(3) The following (single) induction axiom:

\[
\bar{0} \in X \& (\forall x)(x \in X \rightarrow S(x) \in X) \rightarrow (\forall x)(x \in X).
\]

(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 $\text{Number}(x)$, $\text{Set}(x)$ and axioms that there are numbers, there are sets and $\text{Number}(x) \equiv \neg \text{Set}(x)$; 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 $S$, $+$, $\ast$ and satisfy $Q$ (e.g. $\text{Number}(x) \& x \neq 0 \rightarrow (\exists y)(\text{Number}(y) \& x = S(y))$ etc.). Similarly we rewrite the comprehension.

\[
\text{Number}(y) \& \text{Set}(w) \rightarrow (\exists z)(\text{Set}(z) \& (\forall x)(\text{Number}(x) \rightarrow (x \in z \equiv \varphi'(x, y, w))))
\]

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

\[
\text{Set}(x) \& \text{Set}(y) \rightarrow (\forall u)(\text{Number}(u) \rightarrow (u \in x \equiv u \in y)) \equiv x = y.
\]

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* $\varphi$; 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
1. Preliminaries

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 G"odel-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 $I_\varphi$ is provable in $ACA_0$. But this is easy: by comprehension, let $Z = \{ x \mid \varphi(x,y) \}$. Assume $\varphi(0,y)$ and $(\forall x)(\varphi(x,y) \to \varphi(S(x),y))$. Then $0 \in Z$ and $(\forall x)(x \in Z \to S(x) \in Z)$. We get $((\forall x)(x \in Z)$, i.e. $(\forall x)\varphi(x,y)$. 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 $\varphi$ 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 numbers 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
Thus we write just $SEQ(Z)$. If $SEQ(Z)$ and $Y$ is a class then define $G(Z,Y,u)$ as follows:

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

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 $(x,\varepsilon)$ 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 $Sat_{\varphi,X}$ (see I.2.55); take the axiom (A3) $(\forall f \in \Sigma_0^*(X))(\exists Z)(\forall e)(e \in Z \equiv Sat_{\varphi,f,e})$ (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 $Sat_{\varphi,X}$ so that we have "snowing"-snowing for $\Sigma_0^*(X)$-formulas. (A4) and (A5) will describe universal and existential projection:

(A4) $(\forall X)(\exists Z)(\forall s)(s \in Z \equiv (\forall x)((x) \& s \in X)$

(A5) $(\forall X)(\exists Z)(\forall s)(s \in Z \equiv (\exists x)((x) \& s \in X)$

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

Clearly, $I\Sigma_1 + (A2) - (A6)$ proves each instance of comprehension for a formula $\varphi(x,X)$ without second order quantifier: Apply (A3) to $\varphi_0$, where $\varphi_0$ is the open part of $\varphi$ ($\varphi$ assumed prenex) and then apply (A4) and (A5) according to the quantifier prefix of $\varphi$. 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 $G(Z,Y,u)$:

$$(\forall u)(\forall Y)(\forall Z)(SEQ(Z) \rightarrow (\exists W)(\forall x,v)((x,v) \in W$$

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

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:

\[ R(k_1, \ldots, k_n) \iff T \vdash \varphi(k_1, \ldots, k_n). \]

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

\[ \neg R(k_1, \ldots, k_n) \iff T \vdash \neg \varphi(k_1, \ldots, k_n). \]

1.21 Definition. A theory $T$ containing $Q$ is $\Gamma$-sound (where $\Gamma$ 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 $\Gamma$-sound for $\Gamma$ 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 $\Sigma_1$-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 $\Pi_1$-formulas true in $N$. (Show that the two definitions are equivalent.) Note that $\Sigma_1$-soundness implies consistency.

1.23 Theorem. (1) If $A$ is defined by a $\Sigma_1$-formula $\varphi(x_1, \ldots, x_n)$ in $N$ then $\varphi$ numerates $A$ in $Q$ and also in each $\Sigma_1$-sound theory $T$ containing $Q$.

(2) If $A$ is $\Delta_1$ (in $N$) then there is a $\Sigma_1$-formula $\varphi$ which binumerates $A$ in $Q$ and also in each consistent $T$ containing $Q$.

Proof. (1) follows immediately from $\Sigma_1$-completeness and $\Sigma_1$-soundness. (See 1.1.8). To prove (2) first assume $A \subseteq N$; take a $\Sigma_1$-definition $(\exists y)\sigma(x, y)$ and a $\Pi_1$-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
(∃u)σ(x,u) such that beneath y there is no witness for (∃u)¬σ(x,u). If k ∈ A then for some m, N ⊨ σ(k,m), trivially, for each m′ ≤ m, N ⊨ π(k, m) since N ⊨ (∀y)π(k, y). Thus Q proves ¬π(k, m) & (∀y ≤ m)¬σ(k, y). But then Q proves ¬φ(k): Work in Q and assume φ(k); let y₀ be such that σ(k, y₀) and (∀z ≤ y₀)π(k, z). We have either y₀ ≤ m or y₀ ≥ m (cf. I.1.6). But y₀ ≤ m implies ¬σ(k, y₀), a contradiction, and y₀ ≥ m implies π(k, m), also a contradiction. This completes the proof.

Now assume that R is an n-ary relation. Let MLₙ be the maximolexicographical ordering of Nⁿ, i.e. (k₁,...,kₙ)MLₙ(q₁,...,qₙ) iff max(k₁,...,kₙ) < max(q₁,...,qₙ) or [the maxima are equal and for the first i such that kᵢ ≠ qᵢ we have kᵢ < qᵢ)]. Define MLₙ by a Σ₀ formula (x₁,...,xₙ) ≤ₙ (y₁,...,yₙ) by copying the definition (caution: we use a disjunction over i = 1,...,n). Analogously as for <, Q proves (x₁,...,xₙ) ≤ₙ (q₁,...,qₙ) = V{A; xᵢ = kᵢ | (k₁,...,kₙ)MLₙ(q₁,...,qₙ) (Cf. 5.1.6.). Having this construct φ(x₁,...,xₙ) from (∃u)σ(x₁...xₙ,u) and (∀u)π(x₁,...,xₙ,u) as above but using ≤ₙ instead of <.

1.24 Remark. (1) The preceding theorem implies that each Δ₁ relation A also has a Π₁ binumeration in T ⊇ Q (since the complement of A is also Δ₁)
(2) Recall the simple remark in 1.1.65 which states that if A is defined by a formula φ which is Δ₁ in T then A is binumerated by φ in T.
(3) In Sect. 3 we remove the assumption of Σ₁ soundness from 1.23 (1).
(4) Functions are particular relations; thus if F is a Δ₁ mapping of N into N then there is a Σ₁ formula φ such that

\[ k = F(m) \iff Q \vdash \varphi(k,m) \]
\[ k \neq F(m) \iff Q \vdash \neg \varphi(k,m) \]

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 → N be Σ₁ (possibly partial). Then there is a Σ₁ formula φ(x,y) such that, for each m ∈ dom(F),

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

Proof. Let F be defined by a Σ₁-formula (∃z)σ(x,y,z), σ being Σ₀ (i.e. F(m) = k iff N ⊨ (∃z)σ(μ,k,z)). Let φ(x,y) be the following formula (∃z)[φ(x,y,z) & (∀u,v ≤ y)(u ≠ y → ¬σ(x,u,v) & (∀u,v ≤ z)(u ≠ y → ¬σ(x,u,v)].
(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 \( Q \vdash y = \overline{k} \rightarrow \varphi(\overline{m}, y) \). 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 \( \lnot \sigma(\overline{m}, \overline{k}, \overline{q}) \) — contradiction.

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

**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 \( \Sigma_1 \)-sound \( \Sigma_1 \) theory containing \( Q \) then \( T \) is incomplete; there is a \( \Sigma_1 \) sentence \( \nu \) such that both \( \nu \) and \( \lnot \nu \) are unprovable in \( T \).

**Proof.** Recall that there is a \( \Sigma_1 \) set \( K_0 \) of natural numbers which is not \( \Delta_1 \) (the reader may find a proof of this in the next remark).

Now let \( \alpha(x) \) be a \( \Sigma_1 \) formula numerating \( K_0 \) in \( T \), i.e. \( K_0 = \{ n \mid T \vdash \alpha(n) \} \). Let \( K_1 = \{ n \mid T \vdash \lnot \alpha(n) \} \). Since \( K_0 \) is not \( \Delta_1 \) and \( T \) is consistent we have \( K_0 \cup K_1 \neq N \); for \( n \in N - K_0 - K_1 \) both \( \alpha(n) \) and \( \lnot \alpha(n) \) are unprovable in \( T \).

1.27 **Remark.** For the reader's convenience we prove here that there is a \( \Sigma_1 \) set of natural numbers which is not \( \Delta_1 \). For example, let \( K_0 \) be the set of all \( \varphi \) such that \( \varphi \) is a \( \Sigma_1 \)-formula with just one free variable and is satisfied on \( N \) by \( \varphi \) (itself). Then \( K_0 \) is \( \Sigma_1 \) (it is defined by

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

If \( N - K \) were \( \Sigma_1 \) then we could produce a \( \Sigma_1 \)-formula \( \varphi(x) \) defining the same set of natural numbers as the formula

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

By "snowing"-snowing.

\[ N \models \varphi \in \Sigma_1^* \& \varphi \text{ has one free variable}^*; \]

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

\[ N \models Sat_{\Sigma,1}(\varphi, [\varphi]) \equiv \lnot Sat_{\Sigma,1}(\varphi, [\varphi]), \]

which is a contradiction. We have shown that \( K_0 \) is not \( \Sigma_1 \).
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 $N \models \text{Sat}_{\Sigma,n}(\varphi,\emptyset)$ ($\varphi$ being $\Sigma_n$). Given any formula $\psi(x)$ with exactly one free variable, a sentence $\varphi$ is self-referential with respect to $\psi$ in $N$ (or: is a fixed-point of $\psi$ in $N$) if $N \models \varphi \equiv \psi(\varphi)$, thus $N \models \varphi$ is equivalent to $N \models \psi(\varphi)$. We can say rather suggestively that $\varphi$ says "I have the property $\psi$". As usual, we are more interested in provability in a theory than in truth in $N$; we say that $\varphi$ is a fixed-point of $\psi(x)$ in $T$ if $T \vdash \varphi \equiv \psi(\bar{\varphi})$.

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 $\varphi$ such that $T \vdash \varphi \equiv \psi(\bar{\varphi})$.

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

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 $\Delta_1$ function associating with each formula $\delta(x)$ (in the language of $T$ having exactly one free variable $x$) the closed formula $\delta(\bar{\delta})$, i.e. the result of substitution of $\bar{\delta}$ into $\delta$ for $x$. (Let $F$ be 0 for other arguments.) By 1.25 there is a formula $\alpha(x,\bar{v})$ such that, for each $\delta$, $T \vdash \alpha(\bar{\delta},v) \equiv v = F(\bar{\delta})$. Let $\chi(x)$ be the formula $(\exists v)(\alpha(x,v) \& \psi(v))$ and let $\varphi$ be $F(\chi)$, i.e. $\chi(\bar{x})$. $Q$ proves the following equivalences:

\[ \varphi \equiv \chi(\bar{x}) \equiv (\exists v)(\alpha(x,v) \& \psi(v)) \equiv (\exists v)(\psi(F(\bar{x}) \& \psi(v))) \]

\[ \equiv (\exists v)(v = \varphi \& \psi(v)) \equiv \psi(\bar{\varphi}). \]
To prove (2) modify $F$. $F$ associates with each formula $\delta(x, y)$ the formula $\delta(\delta, y)$. Let $\alpha$ be as above, for the new $F$. If $\psi(x, y)$ is given, let $\chi(x, y)$ be the formula $(\exists y)(\alpha(x, v) \& \psi(v, y))$, let $\varphi$ be $F(\chi)$. We have

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

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 $k \in N$, $T \vdash \varphi(k) \equiv \psi(\overline{\varphi(k)}, k)$.

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

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

Proof. We prove (1). Let $F(\delta, k)$ be the $\Delta_1$ function associating with each $k$ the formula $\delta(\overline{k})$; let $\beta(u, v, w)$ be such that $Q \vdash \beta(\overline{\delta, k}, w) \equiv w = F(\delta, k)$. 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 $T \vdash \varphi(k) \equiv \psi'(\overline{\varphi}, k) \equiv (\exists w)(\beta(\overline{\varphi}, k, w) \& \psi(w, k) \equiv \psi(\overline{\varphi(k)}, k).$  

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 $\varphi$, $T \vdash \varphi \equiv Tr(\varphi)$.

Proof. If we had $Tr(x)$ we could reproduce the liar's paradox: let $\varphi$ 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 $T \supset I\Sigma_1$ is a consistent theory in the language of arithmetic we have, for each $k$, partial satisfactions $Sat_{\Sigma, k}$ and $Sat_{II, k}$ 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 $\varphi$ is a fixed point of $\psi$ in $T$ iff $T \vdash \varphi$. (We can say that $\psi$ has only provable fixed points.) If $T \vdash (\forall x)\neg \psi(x)$ then $\psi$ has only refutable fixed points. In both cases, all fixed points of $\psi$ are $T$-provably equivalent, i.e. if $\varphi_1$ and $\varphi_2$ are fixed points of $\varphi$ then $T \vdash \varphi_1 \equiv \varphi_2$. We say that $\psi$ 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 $\phi_1, \phi_2$ such that $T \vdash \phi_1 \equiv \psi_1(\phi_1, \phi_2)$ and $T \vdash \phi_2 \equiv \psi_2(\phi_1, \phi_2)$.

Proof. Let $F_i(\delta_1(x, y), \delta_2(x, y))$ be $\delta_i(\delta_1, \delta_2)$ ($i = 1, 2$); let $\alpha_i(u, v, w)$ be such that $Q \vdash \alpha_i(\delta_1, \delta_2, w) \equiv w = F_i(\delta_1, \delta_2)$. Let $\chi(x, y)$ be $(\exists w_1, w_2)((\alpha_1(x, y, w_1) \& \alpha_2(x, y, w_2) \& \psi_1(w_1, w_2))$; let $\phi_i$ be $\chi_i(\bar{x}_1, \bar{x}_2)$. The proof of $T \vdash \phi_i \equiv \psi_i(\phi_1, \phi_2)$ 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 ($T \supset I \Sigma_1$ suffices) then it is natural to work with the formalized proof predicate $Proof^*$ (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 $\Delta_1$ and hence has a $\Sigma_1$ 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 $\Sigma_1$ theory containing $Q$ and let $\pi(x)$ be a $\Sigma_1$ formula defining the set of all $T$-provable sentences (i.e. $T \vdash \varphi$ iff $N \vdash \pi(\varphi)$). A Gödel sentence based on $\pi$ is a fixed point of $\neg \pi$, i.e. a sentence $\nu$ such that

$$T \vdash \nu \equiv \neg \pi(\nu).$$

(We may describe $\nu$ by saying that it asserts its own unprovability.)

2.8 Theorem. Let $T, \pi$ be as above and let $\nu$ be a Gödel sentence given by $\pi$.
(1) If $T$ is consistent then $T$ does not prove $\nu$. (2) If $T$ is $\Sigma_1$-sound then $T$ does not prove $\neg \nu$.

Proof. (1) Assume $T \vdash \nu$, then $N \vdash \pi(\nu)$, thus $T \vdash \pi(\nu)$ by $\Sigma_1$-completeness; thus $T \vdash \neg \nu$ and $T$ is inconsistent.

(2) Assume $T \vdash \neg \nu$, i.e. $T \vdash \pi(\nu)$. If $T$ were $\Sigma_1$-sound then $\pi$ would enumerate the set of all provable sentences, thus we would have $T \vdash \nu$ and $T$ would be inconsistent. Thus $T$ cannot be $\Sigma_1$-sound. □

We present a technique that makes it possible to get rid of $\Sigma_1$-soundness.
2.9 Definition. Let \((\exists u)\alpha(x,u)\) and \((\exists u)\beta(x,u)\) be two \(\Sigma_1\)-sentences; thus \(\alpha, \beta\) are \(\Sigma_0\). A witness comparison sentence given by these sentences is the sentence

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

(Observe that this is again a \(\Sigma_1\) 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 \((\exists u)\beta \preceq (\exists u)\alpha\). In particular, let \((\exists u)\alpha(x,u)\) be a \(\Sigma_1\)-formula defining the set of all \(T\)-provable sentences and let \((\exists u)\beta(x,u)\) be a \(\Sigma_1\)-formula defining the set of all \(T\)-refutable sentences, i.e. \(T \vdash \varphi \iff N \models (\exists u)\alpha(\varphi, u)\) and \(T \vdash \neg \varphi \iff N \models (\exists u)\beta(\varphi, u)\). A Rosser sentence given by \(\alpha\) and \(\beta\) is a formula \(\rho\) such that

\[T \vdash \rho \equiv (\exists u)\beta(\bar{\rho}, u) \preceq (\exists u)\alpha(\bar{\rho}, u),\]

i.e.

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

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 \(\Sigma_1\) theory containing \(Q\) and let \(\rho\) be a Rosser sentence for \(T\). If \(T\) is consistent then \(T\) proves neither \(\rho\) nor \(\neg \rho\).

Proof. Assume \(T \vdash \rho\), then for some \(d\), \(N \models \alpha((\bar{\rho}, \bar{d})\) and hence \(T \vdash \alpha(\bar{\rho}, \bar{d})\).

Now work in \(Q\) and let \(y\) be a witness for \((\exists y)\beta(\bar{\rho}, y)\) such that beneath \(y\) there is no witness for \((\exists y)\alpha(\bar{\rho}, y)\). Then \(y \leq \bar{\rho} \lor d \leq y\) (cf. I.1.6) and \(d \leq y\) is impossible since \(\alpha(\bar{\rho}, \bar{d})\). We have proved \(T \vdash (\exists y < \bar{d})\beta(\bar{\rho}, y)\), i.e. \(T \vdash \forall e < d \beta(\bar{\rho}, e)\). But since \(T\) is consistent \(T\) does not prove \(\neg \rho\) and thus for each \(e\), \(N \models \neg \beta(\bar{\rho}, e)\), thus \(T \vdash \neg \beta(\bar{\rho}, e)\) by \(\Sigma_1\)-completeness. Thus \(T \vdash \bigwedge_{e < d} \neg \beta(\bar{\rho}, e)\) and \(T\) is inconsistent.

Second, assume \(T \vdash \neg \rho\), thus

\[T \vdash (\forall y)(\beta(\bar{\rho}, y) \rightarrow ((\exists z < y)\alpha(\bar{\rho}, z))).\]

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

2.11 Theorem (Essential Undecidability of Arithmetic). Each consistent theory \(T\) containing \(Q\) is undecidable; i.e. the set of its theorems is not \(\Delta_1\).
III. Self-Reference

Proof. This is because if $T$ were decidable we could extend it to a $\Delta_1$ theory $T'$ which is complete; but this is impossible by 2.10. (Indeed, let $\varphi_0, \varphi_1, \ldots$ be a $\Delta_1$ enumeration of sentences of $T$; define $\alpha_n = \varphi_n$ if $Con(TU\{\alpha_0, \ldots, \alpha_{n-1}\})$ and $\alpha_n = \neg \varphi_n$ otherwise; axioms of $T'$ are all $\alpha_n$. This usual completion procedure is easily seen to be $\Delta_1$ 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 $\varphi(0), \varphi(1), \varphi(2), \ldots$ is consistent with $T$. In more details, let $(0)\alpha$ be $\neg \alpha$ and $(1)\alpha$ be $\alpha$; then for each string $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n$, the theory

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

is consistent.

2.14 Lemma. Let $T$ be a $\Sigma_1$ theory containing $Q$. Then there is a $\Sigma_1$ formula $\varphi(x)$ such that for each $k$, $(T + (\forall x)(\varphi(x) \equiv x = k))$ is consistent. (The additional axiom says that $k$ is the unique element satisfying $\varphi$.)

Proof. Define a function $F$ as follows: $F(\psi) = k$ iff $\psi$ is a $T$-formula with just one free variable and there is a $T$-proof $d$ of $\neg (\forall x)(\psi(x) \equiv x = k)$ such that for each $d' < d$ and each $k' \leq d'$, $d'$ is not a $T$-proof of $\neg (\forall x)(\psi(x) \equiv x = k')$ (thus $d$ is minimal possible). Clearly, $F$ is a partial $\Sigma_1$ function. By 1.25, let $\alpha(x, y)$ be such that if $F(\psi) = k$ then $T \vdash \alpha(\overline{\psi}, y) \equiv y = k$. 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 $k$, $(T + (\forall x)(\varphi(x) \equiv x = k))$ is consistent. Assume not and let $d$ be the minimal possible proof as above, $d$ proving $\neg (\forall x)(\varphi(x) \equiv x = k)$. Then $T \vdash \alpha(\overline{\varphi}, y) \equiv y = k$, thus $T \vdash (\forall x)(\varphi(x) \equiv x = k)$ and $T$ is inconsistent. 

2.15 Theorem (on flexible formulas). Let $T$ be $\Delta_1$ theory containing $Q$. Then there is a flexible $\Sigma_1$ formula.

Proof. Let $\beta(s, i)$ be a $\Sigma_1$ 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 $\varepsilon$ 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 $\psi(k) \equiv \overline{\varepsilon}_k = 1$, i.e. proves $(\varepsilon_k)\psi(k)$ ($k = 0, \ldots, n$). This completes the proof.
(c) Gödel's Second Incompleteness Theorem

Our formulation of Gödel’s first incompleteness theorem was rather general; it concerned any \( \Sigma_1 \) definition \( \pi \) 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 \( \Sigma_{n+2} \) formulas (2.27).

2.16 Definition. Let \( T \) be a theory containing \( Q \) and let \( \pi \) be a \( \Sigma_1 \) definition of the set of all \( T \)-provable formulas. The provability conditions for \( \pi \) are the following conditions:

(1) \( T \vdash \varphi \) implies \( T \vdash \pi(\varphi) \),
(2) \( T \vdash \pi(\varphi) \rightarrow \pi(\overline{\varphi}) \),
(3) \( T \vdash \pi(\overline{\varphi}) \land \pi(\varphi \rightarrow \psi) \rightarrow \pi(\psi) \)

2.17 Lemma. Let \( T \) contain \( I\Sigma_1 \), let \( \tau \) be a \( \Sigma_1 \) definition of \( T \) and let the predicates \( Pr^\tau_\pi(s,x) \) (\( s \) is a \( \tau \)-proof of \( x \)) and \( Pr^\tau_\pi(x) \) (\( x \) is \( \tau \)-provable) be as in 1.4.3. Then \( Pr^\tau_\pi \) satisfies the provability conditions 2.16.

Proof. Observe that if \( \tau \) is a \( \Sigma_1 \) definition of \( T \) then \( Pr^\tau_\pi \) is a definition of the set of all \( T \)-provable formulas and is \( \Sigma_1 \) in \( T \). Thus if \( T \vdash \varphi \) then the formula \( Pr^\tau_\pi(\varphi) \) is true and therefore \( T \)-provable, by \( \Sigma_1 \)-completeness (see I.1.8). This proves (1). To prove (2) reason inside \( T \) and repeat the proof of (1), now using formalized \( \Sigma_1 \)-completeness (see I.4.32). (3) is immediate from the definition of \( Pr^\tau_\pi \). \( \square \)

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

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

Proof. By (1), \( T \vdash \pi(\neg 0 = 1) \). Thus

\[
T \vdash \neg Con(\pi) \rightarrow \pi(0 = 1) \land \pi(\neg 0 = 1),
T \vdash \pi(0 = 1 \rightarrow (\neg 0 = 1 \rightarrow \varphi))
\]
III. Self-Reference

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

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

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

$$T \vdash \text{Con} \equiv \text{Con}_r,$$

where $\text{Con}_r$ 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 $\pi$ be a $\Sigma_1$ definition of the set of all $T$-provable formulas satisfying the provability conditions. Let $\nu$ be Gödel's sentence based on $\pi$. Then

$$T \vdash \nu \equiv \text{Con}(\pi)$$

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

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

2.22 Corollary. Let $T$ be a theory containing $I\Sigma_1$ and let $\nu$ be Gödel's sentence based on $Pr^*_r$, where $r$ is a $\Sigma_1$ definition of axioms of $T$. Then $T \vdash \nu \equiv \text{Con}_r$, thus $T$ does not prove $\text{Con}_r$, 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 $\Sigma_1$-sound.

On the other hand, if $\lambda(x)$ is a flexible formula then for any two different fixed points $\phi, \psi$ of $\lambda$, $T$ does not prove $\phi \equiv \psi$ (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 $PA \vdash \text{Con}(I\Sigma_k)$ and, by 2.22, $I\Sigma_k$ does not prove $\text{Con}(I\Sigma_k)$. □
2.25 Löb's Theorem. Let $T$ be as in 2.22, i.e. $T$ contains $I\Sigma_1$, $\tau$ is a $\Sigma_1$ definition of $T$. Assume that $\varphi$ is a sentence such that $T \vdash Pr^*_\tau(\varphi) \rightarrow \varphi$. Then $T \vdash \varphi$.

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

2.26 Corollary. Let $T$ contain $I\Sigma_1$ and let $\kappa$ be a fixed point of the formula $Pr^*_\tau(x)$ (where $\tau$ is a $\Sigma_1$ definition of $T$), i.e. $T \vdash \kappa \equiv Pr^*_\tau(\kappa)$. The formula $\kappa$ 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 $\Sigma_{n+2}$ formulas.

Proof. First show that $I\Sigma_n$ is not axiomatizable by $\Pi_{n+1}$ formulas. If it were we would have a finite set $S$ of $\Pi_{n+1}$ 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 $\Pi_{n+1}$ formulas and hence, by "snowing"-snowing, we would have $I\Sigma_n \vdash S^* \subseteq \text{Tr}(\Pi_{n+1})$ ($S^*$ is $\{\psi_1, \ldots, \psi_n\}^*$ where $S = \{\varphi_1, \ldots, \varphi_n\}$). Furthermore, since finite axiomatizability of $I\Sigma_n$ is provable in $I\Sigma_1$, we would get $I\Sigma_1 \vdash Con^*(S^*) \rightarrow Con^*(I\Sigma_n)$ and, by the above, $I\Sigma_n \vdash Con^*(S)$. Thus we would get $I\Sigma_n \vdash Con^*(I\Sigma_n^*)$ which contradicts Gödel's second incompleteness theorem.

Now assume that $S$ is a finite set of $\Sigma_{n+2}$ formulas axiomatizing $I\Sigma_n$, $S = \{\varphi_1, \ldots, \varphi_n\}$, $\varphi_i = (\exists x)\psi_i(x)$, where $\psi_i$ is $\Pi_{n+1}$. Let us work in $I\Sigma_n$. We have $\text{Tr}_{\Sigma_{n+1}}(S^*)$ and $Con^*(S^*) \rightarrow Con^*(I\Sigma_n^*)$; thus for some $x_1, \ldots, x_n$ we have $\bigwedge_i \text{Tr}_{\Pi_n}(\psi_i(x_i))$ and for $S_i^* = \{\psi_1(x_1), \ldots, \psi_n(x_n)\}$ we have $Con^*(S_i^*) \rightarrow Con^*(S^*)$. But $Con^*(S_i^*)$ follows by I.4.33. We have proved $I\Sigma_n \vdash Con^*(I\Sigma_n^*)$ 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 $\Sigma_1$ (has a $\Sigma_1$ definition). We shall often need stronger assumptions on $T$; we shall now list and discuss four most frequent possibilities.

(a) $T$ is $\Sigma_1$;
(b) $T$ is $\Delta_1$;
(c) $T$ is $I\Sigma_1$-provably $\Delta_1$;
(d) $T$ is $\Sigma_0^{exp}$
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 \( \Sigma_0^{\exp} \) 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 \( \Delta_1 \) sets cf. I.1.51; by I.1.50 and I.1.52 each provably \( \Delta_1 \) set is \( \Delta_1 \) and each \( \Delta_1 \) set is \( \Sigma_1 \).

\( \Sigma_1 \) sets are often called recursively enumerable and \( \Delta_1 \) sets are called recursive (cf. Sect. 0). We also know that each primitive recursive set is \( I\Sigma_1 \)-provably \( \Delta_1 \) and have promised to show the converse in Chap. IV by model-theoretic means (cf. 1.1.54). Concerning \( \Sigma_0^{\exp} \) 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 \textit{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 \textbf{Craig's Theorem.} Each \( \Sigma_1 \) theory \( T \) is deductively equivalent to a \( \Sigma_0^{\exp} \) theory.

\textbf{Proof.} We show that \( T \) is deductively equivalent to a provably \( \Delta_1 \) theory; the reader may show that the same construction gives in fact a \( \Sigma_0^{\exp} \) theory. Let \( \tau \) be a \( \Sigma_1 \) formula defining the set \( T \) of axioms, let \( \tau \) be \( (\exists y)\tau(x, y) \) where \( \tau_0 \) is \( \Sigma_0 \). Craig’s trick is to replace each axiom \( \varphi \) by a sufficiently long conjunction

\[ \varphi \land \varphi \land \ldots \land \varphi \]

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

\[ \delta(x) \equiv (\exists y \leq x)(\exists u, v \leq x)(x = \text{Repeat}^*(v, u) \land \tau_0(u, y)) \]

where \( \text{Repeat}^* \) is defined in the obvious way and is \( \Delta_1 \) in \( I\Sigma_1 \), thus \( \delta(x) \) is also \( \Delta_1 \) 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 \( n, \text{Repeat}(n, \varphi) \in S \). Clearly, \( S \) is deductively equivalent to \( T \).

2.30 \textbf{Remarks.} (1) Observe that in fact we can prove more: if \( \tau \) is a \( \Sigma_1 \)-definition of \( T \) in \( I\Sigma_1 \) and \( \delta \) is as above and defines \( S \) then \( I\Sigma_1 \) proves that the theories* \( \tau \) and \( \delta \) are deductively* equivalent*.

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

(3) Observe that if \( \delta \) is \( \Delta_1 \) in \( I\Sigma_1 \) and defines \( T \) then \( \tau \) 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 $\tau$ is its definition and is $\Delta_1$ in a sound $S \supseteq Q$ then $\tau$ binumerates $A$ in $S$.)

2.3.1 Further Remarks. (1) We know that each $\Delta_1$ set of natural numbers has a $\Sigma_1$ binumeration (and a $\Pi_1$ binumeration) in $T$ where $T$ is any consistent $\Delta_1$ theory containing $Q$. Note that a $\Delta_1$ set may have, and in fact always has different $\Sigma_1$ binumerations whose equivalence is not provable in $T$: let $C \subseteq N$ and let $\gamma(x)$ be any $\Sigma_1$ binumeration of $C$. Let $\rho$ be a Rosser sentence of $T$, let $(\exists y)\beta(x, y)$ be a $\Sigma_1$ definition of refutable formulas; put

$$
\gamma_1(x) \equiv \gamma(x) \& (\forall y < x)\neg \beta(\bar{\rho}, y)
$$

$$
\gamma_2(x) \equiv \gamma(x) \lor (\exists y < x)\beta(\bar{\rho}, y).
$$

Clearly, $T \vdash \gamma_1(x) \rightarrow \gamma(x) \rightarrow \gamma_2(x)$; we show that both $\delta_1$ and $\gamma_2$ binumerate $C$ but $T$ does not prove $\gamma_1(x) \equiv \gamma_2(x)$. Observe that for each $k$, $T \vdash \gamma_1(k) \equiv \gamma(k) \equiv \gamma_2(k)$ since for each $k$

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

($\neg \rho$ being unprovable), thus all three formulas binumerate the same set; but $(T + \rho)$ is consistent, i.e. $T + (\exists y)\beta(\bar{\rho}, y)$ 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 $\Sigma_1$-sound $\Delta_1$ theory $T$ containing $Q$ there is a $\Delta_1$ set $A$ such that no formula $\delta$ which is $\Delta$ in $T_1$ binumerates $A$ in $T$ (thus the claim of (1) above cannot be improved to a binumeration which is $\Delta_1$ 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 $k \in N$, $N \vDash (\delta(k) \equiv \pi(k))$ (i.e. $N \vDash \delta(k) \lor \neg \pi(k)$ and $N \vDash \neg \delta(k) \lor \pi(k)$ - $\Sigma_1$ soundness suffices), thus $N \vDash (\forall x)(\delta(x) \equiv \pi(x))$. Thus for each $k \in N$, either there is a witness for $\delta(k)$ or there is a witness for $\neg \delta(k)$.

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 $\delta(n)$ and $\neg \pi(n)$. If you find a witness for $\delta(n)$, define $n \notin A$; if you find a witness for $\neg \pi(n)$, define $n \in A$. For remaining $n$ define e.g. $n \in A$.

The reader may check that this is a $\Delta_1$ definition of a set $A$ and that if $n, \delta, \pi$ are as above then $A$ differs from the set defined by $\delta$ (and $\pi$) at least in $n$.

(3) By 1.23, if $\varphi$ is a $\Sigma_1$ definition of a set $A$ then $\varphi$ numerates $A$ in each $\Sigma_1$-sound theory $S \supseteq Q$. Let us show that the assumption of $\Sigma_1$-soundness in essential; in general, $\varphi$ numerates a superset of $A$. (For example, let $A$ be the set of axioms of $PA$, let $\pi(x)$ be the formula $x \in PA^*$ (where $PA^*$ is defined by copying the definition of $PA$ in $I\Sigma_1$) and let $\varphi(x)$ be the formula $\pi(x) \lor (\neg \text{Con}^*(PA^*) \& x = 0 = 1))$. Then $\pi$ is $\Sigma_1$ in $PA$, numerates $PA$ in $PA$ but numerates $(PA + \bar{0} = 1)$ in the consistent (but $\Sigma_1$-ill) theory $(PA + \neg \text{Con}^*(PA^*))$. 
(4) Returning again to Craig’s theorem, let us observe the following: If \( \tau \) is a \( \Sigma_1 \) definition of a theory \( T \) and \( \delta \) is the \( \Delta_1 \)-in-\( \Pi_1 \) definition of the corresponding Craig theory \( S \) then (i) \( T \) is deductively equivalent to \( S \), (ii) \( I\Sigma_1 \) proves that \( \tau \) is deductively* equivalent* to \( \delta \), (iii) \( \delta \) bim numerates \( S \) in \( T \) (even in \( I\Sigma_1 \)) but (iv) \( \tau \) 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 \( \Delta_1 \) then it has a \( \Delta_2 \) definition \( \pi \) such that \( T \) proves \( \text{Con}^* \pi \) (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 \( T \upharpoonright k \) 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 \( S^* \) denote \( \{\overline{\varphi}_1, \ldots, \overline{\varphi}_n\}^* \).

(3) Let \( T \) be a theory extending \( I\Sigma_1 \). \( T \) is reflexive if for each \( k \), \( T \) proves \( \text{Con}^*((T \upharpoonright k)^*) \) (i.e. for each finite subtheory \( T_0 \) of \( T \), \( T \) proves the consistency of \( T_0 \)).

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 \( \varphi \), \( T \vdash \varphi \rightarrow \text{Con}^* (\{\varphi\}^*) \).

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

2.35 Theorem. \( PA \) is essentially reflexive.

Proof. Immediately from I.4.34. \( \square \)

2.36 Theorem. (1) Let \( T \) be any consistent pure extension of \( PA \). Then for any \( k \), \( T \) is not axiomatizable using only \( \Sigma_k \) formulas.
(2) For each $k \geq 1$, no consistent pure extension of $I\Sigma_k$ is axiomatizable using only $\Sigma_{k+2}$ 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 $\Pi_{k+1}$ formulas such that $S$ proves $I\Sigma_k$ then we get $S \vdash \text{Con}^*(S^*)$ using I.4.34, hence $S$ is inconsistent by Gödel's second incompleteness theorem, thus $T$ is not $\Pi_{k+1}$ axiomatizable (and (1) follows). The generalization to $\Sigma_{k+2}$ formulas is as in 2.27, again getting $S \vdash \text{Con}^*(S^*)$, a contradiction. □

2.37 Theorem. Let $T \in \Delta_1$ be a pure extension of $PA$. Then there is a binumeration $\tau$ of $T$ in $T$ such that $\tau$ is $\Delta_2$ in $T$ and $T \vdash \text{Con}^*(\tau)$.

Proof. By 1.23, $T$ has a $\Sigma_1$ binumeration $\sigma$ in $T$ and a $\Pi_1$ binumeration $\pi$ in $T$; let $\delta(x)$ be $\sigma(x) \& (\forall y \leq x)(\sigma(y) \equiv \pi(y))$. Evidently, $\delta$ is a $\Delta_2$ binumeration of $T$. Let $\tau(x)$ be the formula $\delta(x) \& \text{Con}^*(\pi \upharpoonright x)$ (where $\pi \upharpoonright x(y) \equiv \pi(y) \& y \leq x$). Then $\tau$ is another $\Delta_2$ binumeration of $T$ in $T$ (since, thanks to reflexivity, for each $k$ we have $T \vdash \tau(k) \equiv \delta(k) \equiv \sigma(k) \equiv \pi(k)$). We prove $T \vdash \text{Con}^*(\tau)$. Let us work in $T$. We distinguish two cases. First assume $\text{Con}^*(\pi)$; then evidently $\text{Con}^*(\tau)$ (since $\tau$ implies $\pi$). Second, assume $\neg \text{Con}^*(\pi)$ and let $z$ be the least number such that $\neg \text{Con}^*(\pi \upharpoonright (z+1))$. Then $\text{Con}^*(\pi \upharpoonright z)$ and for all $x$, $\tau(x)$ implies $(\pi \upharpoonright z)(x)$. Thus $\text{Con}^*(\tau)$ and the proof is complete. □

2.38 Remark. The reader has observed that $\tau$ just describes the largest initial segment of $\delta$ which is consistent; thus it is not much surprising that $\text{Con}^*(\tau)$ is provable. Reflexivity guarantees that $\tau$ defines (and $T$-binumerates) $T$. Observe also that the formula $\text{Con}^*(\tau)$ is $\Pi_2$ 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 $k$, $T \vdash \text{Con}^*((S \upharpoonright k)^*)$
(iv) There is a binumeration $\tau$ of $S$ in $T$ such that $T \vdash \text{Con}^*(\tau)$.

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 $\Pi_1$ conservative for $T$. 
2.41 Proof of 2.39. We prove (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (i). (i) ⇒ (ii) is evident. To prove (ii) ⇒ (iii) it suffices to show that if \( S_0 \) is finite and interpretable in \( T \) then \( T \vdash \text{Con}^*(S_0^*) \). We apply the following

2.42 Lemma. Let \( S_0, T_0 \) be finite theories, let \( T_0 \) contain \( I\Sigma_1 \). If \( S_0 \) is interpretable in \( T_0 \) then \( T_0 \vdash \text{Con}^*(T_0^*) \rightarrow \text{Con}^*(S_0^*) \).

Proof. Let \( \text{intp}(d) \) mean that \( d \) is a tuple containing \( T_0 \)-formulas \( \chi, \psi_p, \ldots, \psi_F, \ldots, \psi_c \ldots \) of respective arities (given by the language of \( S_0 \)) as well as \( T_0 \)-proofs of \( (\exists x)\chi(x) \) and of the translations of the axioms of \( S_0 \). This is a finite sequence since \( S_0 \) is finite. Clearly, \( \text{intp}(d) \) is \( \Delta_1 \); moreover a little checking shows that it is defined by a formula \( \text{intp}^*(x) \) which is \( \Delta_1 \) in \( I\Sigma_1 \) (just formalize!). Similarly, let for each \( S_0 \)-formula \( \phi \), \( i(\phi) \) be the translation of \( \phi \) into a \( T_0 \)-formula given by the above. Similarly, we have a definition \( i^* \), \( \Delta_1 \) in \( I\Sigma_1 \), of \( i \). Using \( \Sigma_1 \)-induction we get \( T_0 \vdash \text{Pr}_{S_0}^*(x) \rightarrow \text{Pr}_{T_0}^*(i^*(x)) \); and similarly, \( T_0 \) proves the properties of \( i \) concerning connectives. Thus \( T_0 \vdash \text{Con}^*(T_0^*) \rightarrow \text{Con}^*(S_0^*) \), as desired. 

2.43 Proof of 2.39 continued. The lemma gives \( T \vdash \text{Con}^*((S \upharpoonright k)^*) \) since if \( S_0 = S \upharpoonright k \) and \( S_0 \) is interpretable in \( T \) then \( S_0 \) is interpretable in a finite subtheory \( T_0 \) of \( T \); and \( T \vdash \text{Con}^*(T_0^*) \) due to reflexivity. This completes the proof of (ii) ⇒ (iii).

Now we prove (iii) ⇒ (iv). Assume \( \forall k)T \vdash \text{Con}^*((S \upharpoonright k)^*) \) and let \( \sigma, \pi \) be \( \Sigma_1 \) and \( \Pi_1 \) binnumerations 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 \( \tau(x) = \delta(x) \& \text{Con}^*(\delta \upharpoonright x) \).

Then \( T \vdash \text{Con}^*(\tau) \) and \( \tau \) binnumerates \( S \) in \( T \).

(iv) ⇒ (i). We prove the arithmetized completeness theorem appropriately relativized. Let \( T \vdash \text{Con}^*(\tau) \) where \( \tau \) is \( \Sigma_m \) (say), \( \tau \) binnumerates \( S \) in \( T \). Then \( \tau \) is \( \Delta_{m+1} \) in \( T \) and therefore \( T \) proves that \( S \) has a full low \( \Delta_{m+2} \) model. In fact, the proof of Low basis theorems gives as formulas \( (\chi, \psi_S, \psi^+, \psi^*, \ldots) \) defining a low \( \Delta_{m+2} \) 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 \( \psi \):

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

This completes the proof of 2.39. 

2.44 Proof of 2.40. First assume that \( S \) is interpretable in \( T \); we prove \( \Pi_1 \)-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.

\[
T \vdash \chi(x) \& \chi(y) \rightarrow x = \overline{i} y \equiv x = y.
\]

(Otherwise define a factorization.)
The interpretation defines, in $T$, a model $(M, S^i, +^i, \star^i, 0^i, \leq^i)$ 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 $\text{pism}(s)$ (partial isomorphism) mean $\text{Seq}(s) \& (s)_{i+1} = 0_M \& (\forall_i < lh(s) - 1)((s)_i = S_M((s)_i))$. 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 $\star$ (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 $\Pi_1$ 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 $\Pi_1$-conservative for $S$. Since $S$ is reflexive, we have, for each $k$, $S \vdash \text{Con}^\ast((S \upharpoonright k)^\ast)$, thus $T \vdash \text{Con}^\ast((S \upharpoonright k)^\ast)$, 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 $ACA_0$, 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 $T \vdash \varphi(0) \& (\forall x)(\varphi(x) \rightarrow \varphi(x + 1))$ but $T$ does not prove $(\forall x)\varphi(x)$. (Here $x$ is a variable ranging over numbers, but $\varphi$ may contain other variables and symbols of the language of $T$, e.g. in the case of $ACA_0 \varphi$ 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 $I(0)$, $(\forall x)(I(x) \rightarrow I(x + 1)$, $(\forall x, y)(I(x) \& y \leq x \rightarrow I(y))$ 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 $T^\ast$-proof $z$ 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 $I(0)$, $(\forall x)(I(x) \rightarrow I(x + 1))$ and $(\forall z,y)(y < x \& I(x) \rightarrow I(y))$. $I$ 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 $k$, $T \vdash I(k)$.

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

$$(\exists s)(\text{Seq}(s) \& lh(s) = x \& (\forall i < x)\psi(i,s)).$$

It is an $I\Sigma_n$-cut and is $\Sigma_{n+1}$ 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 $x$,

$$2^0_x = x, \quad 2^{x+1}_x = 2^{2^x}.$$  

Similarly, we define $\omega_y(0) = 0$ and for $x > 0$, $\omega_0(x) = 2x$, $\omega_{y+1}(x) = 2^{\omega_y(|x| - 1)}$, 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. $(\forall y)2^{x+1}_{y+1} = \omega_y(2^{2^x}_{y+1})$,
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 $\omega$ 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

$$\forall x \in J) (2^x_k \in I),$$

$$\forall x \in J) (\omega_k(x) \in J).$$

**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 $\omega_0$).

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

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

Then clearly $T \vdash J_n \subseteq K_n \subseteq I$; in $T$, if $x \in K_n$ and $x \leq 2^y_n$ for some $y \in J_n$ then $\omega_n(x) \leq \omega_n(2^y_n) = 2^{y+1}_n$ 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 $x \leq 2^y_n$ for $y \in J_n$ then $x + 1 \leq 2^{y+1}_n$ (for each $n$, $I \Sigma_1$ proves $2^{y+1}_n \geq 2^y_n + 1$ – 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 $\pi_1$, $\pi_2$ be two definitions of $T$-provable formulas. Generalized provability conditions for $\pi_1$ and $\pi_2$ over $T$ are the following
three conditions: For all \( \varphi \),

\[
\begin{align*}
(1) & \quad T \vdash \varphi \text{ implies } T \vdash \pi_1(\bar{\varphi}), \\
(2) & \quad T \vdash \pi_1(\bar{\varphi}) \rightarrow \pi_2(\pi_1(\bar{\varphi})), \\
(3) & \quad T \vdash \pi_1(\bar{\varphi}) \& \pi_1(\bar{\varphi} \rightarrow \psi) \rightarrow \pi_2(\psi).
\end{align*}
\]

3.7 Remark. Observe that (1) and (3) imply \( T \vdash \pi_1(\bar{\varphi}) \rightarrow \pi_2(\bar{\varphi}) \): take \( \varphi \) for \( \psi \).

3.8 Definition. The consistency statement for \( \pi \) based on \( \varphi \) is the formula

\[
\neg \pi(\bar{\varphi}) \lor \neg \pi(\neg \varphi).
\]

We may denote the last formula by \( \text{Con}_b(\pi, \varphi) \) (cf. 2.18).

3.9 Generalization of Gödel's Second Incompleteness Theorem. Let \( T \supseteq Q \) be consistent and let \( \pi_1, \pi_2 \) satisfy the generalized provability conditions over \( T \). Then there is a sentence \( \varphi \) such that \( T \) does not prove the consistency statement \( \text{Con}_b(\pi_2, \varphi) \).

Proof. Let \( \nu \) be a fixed point of \( \neg \pi_1(x) \), i.e. \( T \vdash \nu \equiv \neg \pi_1(\bar{\nu}) \). This is like Gödel's fixed point but now \( \pi_1 \) need not be \( \Sigma_1 \) and need not define \( T \)-provable formulas. Nevertheless, condition (1) is sufficient to show that \( T \) does not prove \( \nu \) (cf. 2.8). Let \( \varphi \) be \( \pi_1(\nu) \). We show \( T \vdash \neg \pi_2(\bar{\varphi}) \rightarrow \nu \) and \( T \vdash \neg \pi_2(\neg \varphi) \rightarrow \nu \); thus \( T \) does not prove \( \neg \pi_2(\bar{\varphi}) \lor \neg \pi_2(\neg \varphi) \).

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

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

3.10 Definition. Let \( T \supseteq I \Sigma_1 \), let \( I \) be a \( T \)-cut and let \( \tau \) be a \( \Sigma_0^{\text{exp}} \) definition of \( T \). Then \( P^I_\tau(x) \) is the formula

\[
(\exists z)(I(z) \& \text{Proof}^\ast_\tau(z, x))
\]

(saying that there is a \( \tau \)-proof* of \( x \) in \( I \)).

(2) \( \text{Con}_\tau^I \) is the formula \( \neg (\exists z)(I(z) \& \text{Proof}^\ast_\tau(z, 0 = 1)) \).

3.11 Another Generalization of Gödel's Second Incompleteness Theorem. Let \( T \supseteq I \Sigma_1 \), let \( I \) be a \( T \)-cut and \( \tau \) a \( \Sigma_0^{\text{exp}} \)-definition of \( T \). Then \( T \) does not prove \( \text{Con}_\tau^I \).

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 $0$ and function symbols $\cdot, +, \ast, 2^x$) 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 $x \subseteq (y) \times (2) \rightarrow x < 2(x+y+1)^2$. This implies that if $s$ is a sequence $(\text{Seq}(s))$, $lh(s) \leq x$ 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 $s \circ t$ satisfies $s \circ t \leq 2(2x+y+1)^3$; 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 $\text{Concseq}(q)$ in Chap. 1) satisfies $\text{Concseq}(q) < 2(2x+y+1)^2$. 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 $\text{Subst}(\varphi, x, \tau)$ is m.e.b. and similarly for other syntactic notions.

3.14 Lemma. Let $T$ and $\tau$ be as in 3.11, i.e $T \subseteq I\Sigma_1$, $\tau$ a $\Sigma_0^{exp}$-definition of $T$.

(a) If $\varphi$ is $\Sigma_0^{exp}$ then there is a $k$ such that

$$T \vdash x < u \& \ldots \& y < u.$$

$$\rightarrow [\varphi(x, \ldots, y) \rightarrow (\exists z < 2^u_k) \text{Proof}^*_\tau(z, \varphi(x, \ldots, y)].$$

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

$$T \vdash I(x) \rightarrow (\exists z < 2^x_k) \text{Proof}^*_\tau(z, I(x)).$$

Proof. (a) The proof is a re-examination and generalization of the proof of $\Sigma_1$-completeness of $Q$ (1.1.8). 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 there is a \( T \)-proof of \( \varphi(\overline{m}, \ldots, \overline{n}) \) beneath \( 2^k \) (where \( k \) is a constant depending only on \( \varphi \)).

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

\[
\begin{align*}
x + 0 &= x \\
x + Sy &= S(x + y) \\
\overline{m} + 0 &= \overline{m} \\
\overline{m} + S\overline{0} &= S(\overline{m} + 0) \\
\overline{m} + S\overline{0} &= \overline{m + 1} \\
&\ldots \\
\overline{m} + \overline{n} &= S(\overline{m + n - 1}) \\
\overline{m} + \overline{n} &= \overline{m + n}
\end{align*}
\]

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 \( \overline{m} \cdot \overline{n} = \overline{m \cdot n} \), \( \overline{m} \cdot \overline{n} \) (if \( m \neq n \)), and other cases (cf. I.1.8).

Also the proof of \( 2^m = 2^m \) 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 \( d < 2^q \) such that \( d \) is a \( T \)-proof of \( t(\overline{m}, \ldots, \overline{n}) = \Val(t(\overline{m}, \ldots, \overline{n})) \).

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 \( \Val(t(\overline{m}, \ldots, \overline{n})) < 2^h \).

Put \( r = 2^q_h \). For simplicity, just take one example: let \( t(\overline{m}, p, \overline{n}) \) be \( (2^m + n) \cdot p \); we want to estimate a proof of \( (2^m + n) \cdot p = (2^m + n) \cdot p \). By (1) here, we can successively produce proofs of

\[
\begin{align*}
2^m &= 2^m \\
2^m + \overline{n} &= 2^m + \overline{n} \\
(2^m + n) \cdot \overline{p} &= (2^m + n) \cdot \overline{p}
\end{align*}
\]

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 \( \varphi(m, \ldots, \bar{n}) \) is m.e.b.

(4) Prove by induction the following: for each \( \Sigma_0^{exp} \)-formula \( \varphi(x, \ldots, y) \), the function assigning to each \( m, \ldots, n \) such that \( \varphi(m, \ldots, \bar{n}) \) is true, a \( T \)-proof of \( \varphi(m, \ldots, \bar{n}) \), and to each \( m, \ldots, n \) such that \( \varphi(m, \ldots, \bar{n}) \) is false, a \( T \)-proof of \( \neg \varphi(m, \ldots, \bar{n}) \) is m.e.b. This is true for \( \varphi \) 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

\[
(\forall x \leq m) \varphi(x, m, \ldots, \bar{n}) \equiv \bigwedge_{j \leq m} \varphi(j, m, \ldots, \bar{n})
\]

and we can find a proof of this equivalence by m.e.b. function (with arguments \( m, \ldots, n \)). Indeed, analysing 1.1.6 (4) we see that finding a proof of

\[
x \leq \bar{m} \equiv x = 0 \lor \ldots \lor x = \bar{m}
\]

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

\[
(\forall x \leq m) \varphi(x, m, \ldots, \bar{n}) \rightarrow \varphi(k, m, \ldots), \quad (k \leq m)
\]

\[
(\forall x \leq m) \varphi(x, m, \ldots) \rightarrow \bigwedge_{k \leq m} \varphi(k, m, \ldots),
\]

\[
\varphi(k, m, \ldots) \rightarrow (x = k \rightarrow \varphi(x, m, \ldots)),
\]

\[
\bigwedge_{k \leq m} \varphi(k, m, \ldots) \rightarrow \bigvee_{k \leq m} x = b. \rightarrow \varphi(x, m, \ldots),
\]

\[
\bigwedge_{k \leq m} \varphi(k, m, \ldots) \rightarrow (x \leq m \rightarrow \varphi(x, m, \ldots)),
\]

\[
\bigwedge_{b \leq m} \varphi(k, m, \ldots) \rightarrow (\forall x \leq m) \varphi(x, m, \ldots)).
\]

Similarly for \( (\forall x \leq 2^m) \), i.e. for \( (\forall x \leq 2^m) \): 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 \( I(0) \) & (\( \forall x \)(I(x) \rightarrow I(x + 1))). Let \( m \) be given; a proof of \( I(m) \) consists of proofs of \( I(0), I(1), \ldots, I(m) \):
... 
$I(0)$
$I(x) \rightarrow I(x + 1)$
$I(0) \rightarrow I(1)$
$I(1)$

... 
$I(m - 1) \rightarrow I(m)$
$I(m)$

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 \(T \supset I \Sigma_1(\text{exp})\), i.e. used explicitly the power-of-two operation. If \(T \supset I \Sigma_1\) 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 \(T(\text{exp})\)-proof of a \(T\)-formula \(\varphi\) a \(T\)-proof of \(\varphi\). This shows how to get rid of \(2^x\). 

3.15 Lemma. Let \(T, \tau\) be as above. For each \(T\)-cut \(I\) there is a \(T\)-cut \(J\) such that \(T \vdash J \subseteq I\) and \(Pr^*_r J, Pr^*_r I\) satisfy the provability conditions 3.6.

Proof. Let us write \(Pr^I\) instead of \(Pr^*_r J\) and similarly for \(Pr^J\). The condition (1) is evidently satisfied: if \(d\) is a \(T\)-proof of \(\varphi\) then \(T \vdash \text{Proof}^*_r(d, \varphi)\) and \(T \vdash I(d)\). To get (3) it suffices to have \(T \vdash J(x) \rightarrow I(2^x)\) where, provably in \(T\), \(u, v, w \subseteq x \rightarrow u \cup v \cup w < I^\tau\). We prove (2). Let us work in \(T\).

Assume \(Pr^I(\varphi)\), i.e., \(J(x) \& \text{Proof}^*_r(\varphi, x)\). Then, by 3.14, for an appropriate \(j\) given from outside there are \(z_1, z_2 < 2^x\) such that \(\text{Proof}^*_r(J(x), z_1)\) and \(\text{Proof}^*_r(J(x), z_2)\). Then \(z = z_1 \cup z_2 \cup (Pr^J(\varphi))\) (or something very similar, details are unimportant) is a \(\tau\)-proof* of \(Pr^\tau(\varphi)\); thus it suffices that \(z\) is in \(I\).

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

3.16 Proof of 3.11-conclusion. Let \(T, \tau, I\) be as in 3.11 and let \(J_o\) be the cut constructed in 3.15. Then for some particular \(\varphi\) \(T\) does not prove \(\text{Conb}(\tau, \varphi)\). To get the unprovability of \(0 = 1\) 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 \(\tau\)-proof of \(\varphi\) and \(y \in J\) is a \(\tau\)-proof of \(\neg \varphi\) then the concatenation of \(x, y\), and the (standardly long) propositional proof of \(0 = 1\) from \(\varphi\) and \(\neg \varphi\) is in \(J_o\) (use e.g. the \(h\) above). This completes the proof of 3.11.
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 $\Phi$ is provable (in predicate calculus) iff there is a disjunction $D$ of instances of the open part of the Herbrand form $He(\Phi)$ of $\Phi$ such that $\Phi$ 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 $He^*(x)$. (Notation: $HProof^*(z, x)$.) If $T$ is a finite theory consisting of closed formulas then a H-proof of $x$ in $T$ is a H-proof $z$ of $(\land T \rightarrow x)$ ($\land$ standing for a finite conjunction); notation $HProof^*_T(z, x)$. A formula $x$ is H-provable in $T$ ($HPr^*_T(x)$) if there is a H-proof of $x$ in $T$. $T$ is H-consistent ($HCon^*(T)$) if there is no H-proof of $\neg \land T$.

3.18 Remark. We know that $I\Sigma_1$ proves Herbrand’s theorem, i.e. $I\Sigma_1 \vdash HPr^*_T(x) \equiv Pr^*_T(x)$ and $I\Sigma_1 \vdash HCon^*(T) \equiv Con^*(T)$. 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:

$$HPr^*_T(x) \equiv (\exists y)(I(y) \land HProof^*_T(x, y)),$$

$$HCon^*_T \equiv \neg HPr^*_T(\neg \land T^*).$$

We also write $HCon^*_T(T^*)$.

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

$$T \vdash HCon^*_T(T^*).$$
II. Self-Reference

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

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

3.21 Theorem. Still more generally, let T be a finitely axiomatized sequential theory containing IΣ₁ (i.e. an interpretation of IΣ₁ in T has been fixed). Then there is a T-definable cut I such that T proves

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

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 Qₖ-formulas as certain formulas of the language of arithmetic. Q₀-formulas are Σ₀-formulas; Qₖ₊₁-formulas are Qₖ-formulas and boolean combinations of formulas of the form (∃x)φ where φ is Qₖ (say, boolean combinations of ∃Qₖ-formulas). Thus a Qₖ-formula has nesting of quantifiers at most k, Σ₀-formulas being disregarded. A Qₖ-proof is a (Hilbert style) proof consisting only of Qₖ-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 Qₖ-formulas is definable in IΣ₁. It follows that there is a T-cut J such that Qₖ-proofs lying in J preserve truth. Consequently, no Qₖ-proof of ¬ ∨ T* is in J. Finally, it can be shown that each H-proof of a Qₖ-formula can be transformed, in a multi-exponentially bounded way, to a Qₖ₊₂-proof of the same formula. Thus if J is a suitable shortening of J then J has the desired properties.

3.24 Corollary. Let T ⊃ IΣ₁ 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 (∀x)I(x), i.e. there are no proper T-cuts. On the other hand, if T is as above, then T ⊬ HCon^\bullet I(T^*) for some T-cut I, thus T ⊬ HCon^\bullet I(T^*). But since T ⊃ IΣ₁, T proves Herbrand's theorem and hence T ⊬ Con^\bullet I(T^*), 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 IΣ₁: a definable cut I_k and a satisfaction for Qₖ-formulas from I_k.
3. Definable Cuts

Proof. We know that there is a $\Delta_1$ satisfaction for $\Sigma_0$-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*, if $s$ is a $Q_k$-proof 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 $\neg \land T$. More than that: if $u$ is $\Sigma_k^*$ 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 $\neg \land T \lor u$). 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 language with just one variable we have a constant $c(\forall x)\varphi(x)$ (associated to $\varphi$ by a total $\Delta_1$ function) and call the formula

$$\varphi(c(\forall x)\varphi(x)) \rightarrow (\forall x)\varphi(x)$$

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.
3.30 Theorem. For each $k \geq 1$, there is a $q$ such that $I\Sigma_1$ proves the following: There is a $\Delta_1$ function $F(d)$, majorized by $2^d$, such that if $d$ is a $H$-proof* of a $Q_k$-formula* $\phi$ then $F(d)$ is a $Q_{k+2}$-proof* of $\phi$.

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*; 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 $\Phi$ 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 $\Phi$. (If not then the corresponding changes as well as the proof of equivalence of both forms are given by a m.e.b. $\Delta_1$ function.) Call, for a moment, a formula clean if it satisfies our assumptions. A clean formula is a boolean combination (using $\&$, $\lor$, $\neg$) of atoms, negated atoms and quantifies formulas, say $\beta(\varphi_1, \ldots, \varphi_n)$. If $\varphi_1$ is $(\square x)\alpha$ (where $\square$ is $\forall$ or $\exists$) then $\beta(\varphi_1, \ldots, \varphi_n)$ is provably equivalent to $(\square' x)\beta(\alpha, \varphi_2, \ldots, \varphi_n)$ where $\square'$ is $\forall$ or $\exists$; there is a m.e.b. function finding $\square'$ and giving a proof of the equivalence. Call the transition from $\beta(\varphi_1, \ldots, \varphi_n)$ to $(\forall' x)\beta(\alpha, \ldots, \varphi_n)$ the extraction of $\forall$ from $(\forall x)\alpha$ in $\beta(\varphi_1, \ldots, \varphi_n)$, the inverse transition is the insertion of $\forall'$ into $\beta(\alpha, \ldots, \varphi_n)$ by quantifying $\alpha$. A prenex normal form (PNF) of $\Phi$ is constructed by a sequence of extractions of quantifiers:

$$\Phi = \Phi_0 = (\square_0 \ldots)\Psi_0$$

(where $\square_0$ is an empty block of quantifiers)

$$\ldots$$

$$\Phi_k = (\square_k \ldots)\Psi_k$$

($\square_k$ is a block of quantifiers and $\Psi_k$ is open); here for each $i < k$, $(\square_{i+1} \ldots)$ is either of the form $(\square_i \ldots)(\forall x)$ or of the form $(\square_i \ldots)(\exists x)$ and the formula $(\forall x)\phi_{i+1}$ (or $(\exists x)\psi_{i+1}$) results from $\psi_i$ by an extraction of a quantifier. $\Phi_k$ is a PNF of $\Phi$; write $(\square \ldots)K$ instead of $(\square \ldots)\Psi_k$.

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 $(\square \ldots)$ in $(\square \ldots)K$ 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 $(\square \ldots)K$ (call these variables $\forall$-variables and the
3. Definable Cuts 183

others \( \exists \)-variables). Now, by 3.17, a Herbrand proof of \( \Phi \) is a propositional proof of a disjunction \( \bigvee_i K(t_i) \), where each \( t_i \) is a tuple of terms and \( K(t_i) \) is in fact an instance of \( K^* \). We may assume that the disjunction \( \bigvee_i K(t_i) \) contains no variables.

3.33. We shall successively eliminate new function symbols, replacing them by special constants. Write \( t_i^0 \) for \( t_i \); we shall construct a sequence of propositional tautologies

\[
\bigvee_i K(t_i^0),
\]

\[
\ldots
\]

\[
\bigvee_i K(t_i^p);
\]

each \( K(t_i^j) \) will be an instance of \( K \) (but not necessarily of \( K^* \)). The transition from \( j \) to \( j+1 \) proceeds as follows:

Take a term \( f(a) \) occurring in \( \bigvee_i K(t_i^j) \) where \( f \) is a Herbrand function (not occurring in \( \Phi \)) and the terms \( a \) contain no Herbrand functions. Replace \( f(a) \) (in all occurrences in \( \bigvee_i K(t_i^j) \)) 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 \( (\exists u \ldots)\Psi_u \) to \( (\forall u \ldots)(\forall x)\Psi_{u+1}(x,y) \); \( c \) is the special constant for \( (\forall x)\Psi_{u+1}(x,a) \), i.e. the corresponding special axiom is

\[ \Psi_{u+1}(c,a) \rightarrow (\forall x)\Psi_{u+1}(x,a). \]

This change clearly preserves the property of being a tautology (since it commutes with all connectives). We finally arrive at a tautology \( \bigvee_i K(t_i^p) \) 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 \( K(t_i^p) \) implies our original formula \( \Phi \) by a \( Q_{k+2} \)-proof using possibly special axioms for constants used in the construction. Write \( K(s) \) for \( K(t_i^p) \). Recall that \( K(s) \) is in fact \( \Psi_h(s) \) (see the construction of \( PNF(\Phi) \)). Now \( \Psi_h \) is \( (\Box_h \ldots)\Psi_h \) and

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

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

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

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

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

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

$$\bigwedge_j Sp(c_j) \& K(s) \rightarrow \Phi$$

(since $\Phi$ is $\psi_0$). Since $K(s)$ was an arbitrary disjunct of the tautology $\bigvee_i K(t_i^j)$, we have a $Q_{k+1}$-proof of

$$\bigwedge_j Sp(c_j) \rightarrow \Phi.$$  

(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 $\varphi(c_1) \rightarrow (\forall x)\varphi(x)$ and replace $c_1$ in the formula

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

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

$$\varphi(y) \rightarrow (\forall x)\varphi(x),$$

$$\exists y)(\varphi(y) \rightarrow (\forall x)\varphi(x),$$

$$((\forall y)\varphi(y) \rightarrow (\forall x)\varphi(x)) \rightarrow (Sp(c_2) \rightarrow \cdots \Phi),$$

$$Sp(c_2) \rightarrow \cdots \Phi);$$

the last formula is $Q_{k+1}$. By iterating this we get a $Q_{k+2}$-proof of $\Phi$, 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. Definable Cuts

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 0, function symbols $S, +, \ast$ and a predicate $\leq$ having the obvious 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

$$T \vdash HCon^J(T') \to HCon^J(T).$$

(2) Define $Q_0$-formulas as boolean combinations of arithmetical $\Sigma_0$-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 $I_0$ such that, roughly, $T$ proves that for lengths from $I_0$, 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

$$I_0(v) \equiv N(v) \& [((\forall x)Seq(x) \rightarrow (\exists z)(SEQ(z))$$

$$\& (\forall w < v)(\forall u)((x)_w = u \equiv \beta(u, w, z))$$

$$\& (\forall z)(SEQ(z) \& (\forall v < u)(\forall u)(\beta(u, w, z) \rightarrow N(u))$$

$$\rightarrow (\exists x)(Seq(x) \& (\forall v < u)(\forall u)((x)_w = u \equiv \beta(u, w, z))).$$

Then $I_0$ is the desired cut.

(4) We can shorten $I_0$ to get a cut $I_1$ such that, for $T$-terms from $I_1$, their value is uniquely determined.

Define evaluation of variables of a term as a particular sequence of objects (somehow assigning meaning to the variables of the term) and define a corresponding evaluating sequence for the derivation of our term; $I_1$ is the collection of all $u \in I_0$ such that, for each term $t \leq u$ and each evaluation of its variables, each derivation 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 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. 1.1.64).

(6) We define satisfaction for atomic formulas $z \in I_2$ of the whole language and for $\Sigma_0^*$-formulas $z \in I_2$ of the language of arithmetic in the obvious manner. There is a shortening $J_0$ of $I_2$ such that $T$ proves that for $Q_0$-formulas $z \in J_0$, 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* 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* of a \( Q_k \)-formula* 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

\[
T \vdash HCon^T(S^*).
\]

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 \( Q_0 \)-formulas be all atomic \( S \)-formulas; \( Q_{k+1} \)-formulas are boolean combinations of \( Q_k \)-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 \( T \vdash (Q_k \text{-proofs } z \in I \text{ preserve } Sat_{Q,k} \text{-truth}) \) and there is a shortening \( K_k \) of \( J_k \) such that \( T \) proves that each \( H \)-proof* (from no special axioms) \( u \in K_{k+2} \) of a \( Q_k \)-formula \( \varphi \) determines a \( Q_{k+2} \)-proof \( z \in J_{k+2} \) of \( \varphi \). Consequently \( T \) proves the following: if (the concatenation of all elements of) \( S^* \) is \( Q_k \) then \( K_{k+2} \) contains no \( H \)-proof* of \( \neg S^* \) (recall that \( T \vdash (S^* \text{ is true}) \) since \( T \vdash S^* \)). Thus the proof of \( \Rightarrow \) is complete. The rest of the subsection contains a proof of \( \Leftarrow \). □

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 \( T \vdash (J \text{ is closed under concatenation of sequences}) \).

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. \( w = s \bowtie t \), then
An elementary computation shows that
\[ \omega_2(x) = 2^{1/4*(|x|-1)^2}; \]
thus it suffices to assume that \( J \) is closed under \( \omega_2 \) (cf. 3.5).

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 \( L^* \), 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 \( T \vdash H\text{Con}^*(S^*) \). Thus \( T \vdash (in \ I, there \ is \ no \ propositional \ proof^* \ of \ any \ disjunction \ of \ instantions \ of \ \neg Sk(S^*)) \), cf. the definition 1.4.11 of \( H^e^*(\neg S^*) \).

Define, in \( T \), a function \( F \) described as follows: for each \( x \), first take the conjunction \( \kappa_x \) of all closed instances of \( Sk(S^*) \) less than \( x \), then construct the truth table for \( \kappa_x \). If there is an evaluation of atoms making \( \kappa_x \) 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 \( A^\Delta_1 \) and m.e.b. in \( T \). Let \( J \) be a \( T \)-cut shorter than \( I \) and such that \( T \vdash x \in J \rightarrow F(x) \in I \). Then \( T \) proves that, for each \( x \in J \), the conjunction \( \kappa_x \) (of all closed instances of \( Sk(S^*) \) less than \( x \)) is propositionally satisfiable.

(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 \( L_1^* \)-atoms less than \( x \), (say, an \( x \)-evaluation) making \( \kappa_x \) 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 \( \kappa_x \) 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 \( \kappa_y \) 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).
III. Self-Reference

(3) Shorten \( J_1 \) to \( J_2 \) such that \( T \vdash (J_2 \text{ is closed under concatenation}) \) 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

\[
\chi(x) = x \text{ is a closed } L_1^* \text{-term in } J_2;
\]

for each function symbol \( F \) of \( L \), let

\[
\psi_F(x_1, \ldots, x_n, y) = y \text{ is the term } \overline{F}(x_1, \ldots, x_n),
\]

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

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

This defines our interpretation; it suffices to verify that it is an interpretation of \( \text{Sk}(S) \) in \( T \). But this is now clear: each \( \psi_F \) 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 \( t(x_0, \ldots, x_n) \), \( T \) proves

\[
\chi(u_0) \& \ldots \& \chi(u_n) \rightarrow [t^*(u_0, \ldots, u_n) = \overline{t}(n_0/x_0, \ldots, u_n/x_n)]
\]

(cf. I.1.67), and, for each open \( L_1 \)-formula \( \varphi(x_0, \ldots, x_n) \), \( T \) proves

\[
\chi(u_0) \& \ldots \& \chi(u_n) \rightarrow [\varphi^*(u_0, \ldots, u_n) \equiv \text{ with respect to } B,
\quad \varphi(u_0/x_0, \ldots, u_n/x_n) \text{ has truth value 1}].
\]

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

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\(^*\) \( u =^* u' \). Note that with each \( u \) such that \( \chi(u) \) we may associate an \( u' = MT(u) \) such that \( u' \) is minimal and \( B(u =^* u') = 1 \). (Again, this because to decide \( u =^* u' \) for all \( u \leq u' \) we need to know only a set-piece of \( B^* \)). Thus replace \( \chi(u) \) by \( (\chi(u) & u \text{ is } B \text{-minimal}) \) 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 \( PA \vdash \text{Con}(S^*) \). 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 φ(x) is any formula of the language of T then T proves

\[ \varphi(0) \& (\forall x)(N(x) \& \varphi(x)) \rightarrow \varphi(x + 1). \rightarrow (\forall x)(N(x) \rightarrow \varphi(x)). \]

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 \( k \), \( T \vdash \text{Con}^*(S \upharpoonright k) \).
(iv) There is a binumeration \( \sigma \) of \( S \) such that \( T \vdash \text{Con}^*(\sigma) \).

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 \( \Pi_1 \)-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 \( T_0 \supseteq T \) we have \( T \vdash H\text{Con}^*(T_0^*) \) (since there are no proper \( T \)-cuts); but \( T \) contains \( I\Sigma_1 \) and therefore \( T \vdash \text{Con}^*(T_0^*) \equiv H\text{Con}^*(T_0^*) \).

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 \( \Gamma \)-conservativity (\( \Gamma \) being a class of formulas) and interpretability as means of comparing theses. 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 $\Pi_1$-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 $T \supseteq I\Sigma_1$ ($T \vdash H\text{Con}^*$($S^*$) 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 $\varphi$. If $\varphi$ 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 $\Gamma$-conservative over $T$ we say that $\varphi$ is $\Gamma$-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 $\tau$ is a formula $\Delta_1$ in $I\Sigma_1$ defining $T$ (thus $\tau$ binumerates $T$ in $T$). We shall often write $\text{Con}^*_T$ instead of $\text{Con}^*$. $\text{Con}^*_T$ is called Gödel's formula; recall that, by 2.22, $\text{Con}^*_T$ is equivalent to the self-referential formula $\nu$ such that $T \vdash \nu \equiv \neg \text{Pr}_T(\nu)$, i.e. to Gödel's fixed point. Clearly, $\text{Con}^*_T$ is a $\Pi_1$-sentence. Rosser's formula $\rho$ is the self-referential formula such that

\[ T \vdash \rho \equiv (\exists y)(\text{Proof}^*_T(y, =\bar{\rho}) \& (\forall z \leq y)\neg \text{Proof}^*_T(z, \bar{\rho})) , \]

i.e. $\rho$ says "there is a proof $y$ of my negation such that there is no proof of me beneath $y$". Thus $\rho$ is a $\Sigma_1$-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 $H\text{Con}^*(T)$ 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 $\rho$ such that

\[ T \vdash \rho \equiv (\exists y)(H\text{Proof}^*_T(y, =\bar{\rho}) \& (\forall z \leq y)(\neg H\text{Proof}^*_T(z, \bar{\rho})) . \]
4. Partial Conservativity and Interpretability

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 $\text{Con}_T^\bullet$ is not interpretable in $T$; its negation $\neg\text{Con}_T^\bullet$ is interpretable in $T$.
(2) $\neg\text{Con}_T^\bullet$ is $\Pi_1$-conservative over $T$; $\text{Con}_T^\bullet$ is $\Sigma_1$-conservative over $T$ iff $T$ is $\Sigma_1$-sound (i.e. each provable $\Sigma_1$-sentence is true).
(3) Rosser’s formula is $\Pi_1$-conservative; $\neg\rho$ is $\Sigma_1$-conservative iff $T$ is $\Sigma_1$-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 $\rho$ 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 $\rho$ 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 $\Sigma_1$ definition $\tau$ of $T$ there is a $\Sigma_1$ definition $\sigma$ of $S$ such that

$$T \vdash \text{Con}_\tau^\bullet \rightarrow \text{Con}_\sigma^\bullet$$

Proof. Let $i$ be the mapping of $S$-formulas into $T$-formula induced by the interpretation; $i$ is $\Delta_1$. Copy the definition in $T$: we get a function $i^\bullet, \Delta_1$ in $T$, such that, for each $\psi \in S$, $T$ proves $i(\psi)$ and hence $T \vdash \text{Pr}_T^\bullet(i^\bullet(\psi))$. Let $\sigma$ be a $\Sigma_1$ definition of $S$. We cannot claim

$$T \vdash (\forall x)(\sigma_o(x) \rightarrow \text{Pr}_T^\bullet(i^\bullet(x)))$$

but take $\sigma(x)$ to be $\sigma_o(x) \& \text{Pr}_T^\bullet(i^\bullet(x))$. It is easy to check that $\sigma$ is the desired definition of $S$. (Note that if $\tau$ is a $\Sigma_1$ binumeration of $T$ in $T$ then $\sigma$ is a $\Sigma_1$ binumeration of $S$ in $T$.)

4.8 Lemma. Let $T$ be as above and consistent; then the formula $\text{Con}_T^\bullet$ is not interpretable in $T$.

Proof. Assume that the theory ($T \vdash \text{Con}_T^\bullet$) is interpretable in $T$; then, by the preceding lemma, there is a $\Sigma_1$-definition $\sigma$ of $T + \text{Con}_T^\bullet$ such that
$T \vdash \text{Con}^*_T \rightarrow \text{Con}^*_T$, i.e. $(T + \text{Con}^*_T) \vdash \text{Con}^*_T$, which contradicts Gödel's second incompleteness theorem (see 2.22).

4.9 Lemma. Let $T$ be as above; then $\neg \text{Con}^*_T$ is interpretable in $T$ and is $\Pi_1$-conservative over $T$.

Proof. We show that $(T + \neg \text{Con}^*_T)$ is interpretable in, and $\Pi_1$ conservative for, $(T + \text{Con}^*_T)$. 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 $T \vdash \text{Con}^*_T -\neg \Pr_T(\text{Con}^*_T)$, i.e.

$T \vdash \text{Con}^*(T) \rightarrow \text{Con}^*(T + \neg \text{Con}^*_T)$

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

4.10 Lemma. If $T \supseteq I\Sigma_1$ is $\Sigma_1$-sound then each non-refutable $\Pi_1$-sentence is $\Sigma_1$-conservative over $T$.

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

4.11 Corollary. If $T \supseteq I\Sigma_1$ contains $I\Sigma_1$ and is $\Sigma_1$-sound then $\text{Con}^*_T$ is $\Sigma_1$-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 $\rho_T$ is $\Pi_1$-nonconservative.

Proof. $\rho$ evidently implies the following $\Pi_1$-sentence

$(\forall z)(\text{Proof}^*(\bar{\rho}, z) \rightarrow (\exists y < z)\text{Proof}^*(\neg \bar{\rho}, z))$. 
Denote it by $\pi$. If $T \vdash \pi$ then $T \vdash Pr_T^\pi(p) \rightarrow \rho$, since, by Löb's theorem 2.25, $T \vdash \rho$, which is impossible. Thus $(T + \rho)$ proves $\pi$ 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 $\rho$ 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 $\Pi_1$ 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 $\rho$ 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 $(T + \neg \rho) \vdash HCon^*J(T + \neg \rho)$. Let us work in $(T + \neg \rho)$. In $J$ there is no $H$-proof* of $\neg (T + \neg \rho)$; and $\neg \rho$ says that for each $H$-proof* of $\neg \rho$ in $T$ (i.e. for each $H$-proof* of $(T \& \neg \neg \rho)$ there is a smaller $H$-proof* of $\neg \rho$ in $T$ (i.e. a $H$-proof* of $(T \& \neg \neg \rho)$). Consequently, $J$ does not contain any $H$-proof* of $\neg (T \& \neg \neg \rho)$ 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 $\Pi_1$ formula $\varphi$ such that both $\varphi$ 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 $\Pi_1$-sentence is provable (cf. 2.40). But one can find a $\varphi \in \Delta_2$ such that both $\varphi$ and $\neg \varphi$ is interpretable.

On the other hand, take $ACA_0$ for $T$; assumptions of our corollary are satisfied. The corresponding formula $\varphi$ has the following properties: $(ACA_0 + \varphi)$ is interpretable in $ACA_0$ but $(PA + \varphi)$ is not interpretable in $PA$ (since $\varphi$ is unprovable). Similarly for Gödel-Bernays and Zermelo-Fraenkel set theories $GB$ and $ZF$. In the next section we shall obtain a $\Sigma_1$ sentence $\varphi$ 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. $Con^*_T$ is $\Sigma_1$-conservative over $T$ iff $T$ is $\Sigma_1$-sound.
2. Let $\rho$ be the Rosser (or $H$-Rosser) $\Sigma_1$-sentence. Then $\neg \rho$ is $\Sigma_1$-conservative iff $T$ is $\Sigma_1$-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 \( \Rightarrow \) is evident: if \( \sigma \in \Sigma_1 \), \( T \vdash \text{Con}^* \rightarrow \rho \) then \( N \vdash \text{Con}^*_T \rightarrow \rho \), \( N \vdash \rho \) and \( T \vdash \rho \).

Now assume \( T \) not to be \( \Sigma_1 \)-sound and let \( \varphi(y) \) be a \( \Sigma_0 \)-formula such that \( N \vdash \neg(\exists y) \varphi(y) \) but \( T \vdash (\exists y) \varphi(y) \). Take the following self-referential formula:

\[
T \vdash \xi \equiv (\exists y)(\text{Proof}^*(\neg \xi, y) \lor \varphi(y) \& (\forall z \leq y) \neg \text{Proof}^*(\xi, z)).
\]

Note that \( T \vdash \neg \varphi(k) \) for each \( k \).

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

From \( T \vdash \xi \) we get

\[
T \vdash (\exists y < d)\text{Proof}^*(\neg \xi, y),
\]

\[
T \vdash \bigvee_{k<d} \text{Proof}^*(\neg \xi, k)
\]

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

\[
T \vdash \text{Proof}^*(\neg \xi, k).
\]

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

(ii) We now prove \( T + \text{Con}^*_T \vdash \xi \) or, equivalently, \( T + \neg \xi \vdash \neg \text{Con}^*_T \). 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 \( (\exists z \leq y_0) \text{Proof}^*(\xi, z) \); let \( \text{Proof}^*(\xi, z_0) \). Then \( \text{Proof}^*(\xi, z_0) \)

\& \( (\forall y < z_0) \neg \varphi(y) \), hence inside \( T \) we are in the same situation as we were in (i) outside. Copy the reasoning from (i) into \( T \); we get

\[
Pr^*(\text{Proof}^*(\xi, z_0)) \& (\forall y < z_0)Pr^*(\neg \varphi(y)).
\]

From \( Pr^*(\xi) \) we get

\[
Pr^*\left( \bigvee_{\dot{y} < z_0} \text{Proof}^*(\neg \xi, \dot{y}) \right),
\]

hence

\[
(\exists y < z_0)Pr^*(\text{Proof}^*(\neg \xi, \dot{y})),
\]

\[
(\exists y < z_0)\text{Proof}^*(\neg \xi, y),
\]

thus we get \( Pr^*(\xi) \) and \( Pr^*(\neg \xi) \), thus \( \neg \text{Con}^* \).
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 $\Delta$ be $(\exists u)\alpha(u)$ and let $\nabla$ be $(\exists u)\beta(u)$. In $T$, call each a satisfying $\alpha(u)$ a witness for $\Delta$ and similarly for $\nabla$ (cf. 2.9). By $\Delta \prec \nabla$ we denote the witness comparison formula

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

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

$$T \vdash \rho \equiv \Pr^*(\neg \rho) \prec \Pr^*(\rho).$$

4.19 Remark. Rosser's formula is $\Sigma_1$ in $T$, since we assume $T \supseteq I\Sigma_1$. Recall that in $B\Sigma_n$, $\Sigma_n$ formulas are closed under bounded universal quantifiers. We shall deal with formulas comparing witnesses of $\Sigma_n$-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 $\Sigma_n$-formula. There is a $\Sigma_n$-formula $\psi(y)$ such that

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

$$(2) \quad I\Sigma_1 \vdash \psi(y) \rightarrow (\forall x \leq y)\varphi(x).$$
III. Self-Reference

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

\[
(\exists s)(\text{Seq}(s) \land (\forall x \leq y)\alpha(x, (s)_x)).
\]

\[ \Box \]

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 \( [(\forall x \leq y)\varphi(x)]^\#_\Pi \) (if \( \Sigma \) and \( n \) is clear from the context). Dually we define \( [(\exists x \leq y)\varphi(x)]^\#_{\Pi, n} \) for \( \varphi \) being \( \Pi_n \) (or, briefly, \( [(\exists x \leq y)\varphi(x)]^\# \)).

(2) Let \( \alpha \) and \( \beta \) be as by 4.18 and assume that \( \beta \) is \( \Pi_n \). Let \( \beta' \) be the natural \( \Sigma_n \)-form of \( -\beta \). Then \( \Delta \varphi\neg \# \Box \) is the formula

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

the notation \( \Delta \varphi\neg \# \Box \) will be used only in situations where \( n \) is clear from the context. Similarly for \( \Delta \varphi\Box \); the words "antecedent" and "succedent" have the obvious meaning.

4.22 Remark. We shall deal with self-referential formulas \( \xi \) such that

\[
T \vdash \xi \equiv \Delta(\neg\xi) \varphi\neg \# \Box (\xi);
\]

more generally, there may be another parameter:

\[
T \vdash \xi \equiv \Delta(\neg\xi, k) \varphi\neg \# \Box (\xi, k)
\]

(\( k \) natural number). Such formulas will be extremely useful; almost always the antecedent and/or the succedent will contain some provability predicate. Let us mention \( Pr^*_T \) and \( HPr^*_T \) (Herbrand 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 \( \Sigma_n^* \)-formulas and \( \Pi_n^* \)-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 \( \Sigma_1 \)-satisfaction in Sect. 3 (Theorem 3.21 saying that for suitable (finite) theories \( T \), there is a \( T \)-cut \( J \) such that \( T \) proves \( (\forall u \in Tr(\Sigma_1))HCon^\#_J(T + u) \) (for any true \( \Sigma_1^* \)-sentence, in \( J \) there is no Herhand-provability of \( \neg(\land T & u) \)). We shall now make systematic use of provability (\( H \)-provability) in extensions of \( T \) resulting by adding a true \( \Gamma \)-formula where \( \Gamma \) is a class of formulas, say \( \Sigma_n^* \) or \( \Pi_n^* \). This leads us to the following.
4.23 Definition and Convention. Let $n \in N$, let $\Gamma$ be $\Sigma_n$ or $\Pi_n$. We make
the following definition in $I\Sigma_1$: $Pr^\bullet_{\Gamma, T}(x)$ ($x$ is provable in $T$ from a true
$\Gamma$-sentence) if

$$(\exists y \in \Gamma)(Tr_{\Gamma}(y) \land Pr^\bullet_{T + y}(x)).$$

If $T$ is finite we define $HPr^\bullet_{\Gamma, T}(x)$ 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 $Pr^\bullet_{T}(x)$,
$Pr^\bullet_{\Gamma, T}(x)$ etc.

(2) If $\Delta_i$ is $(\exists u)\alpha_i(u)$ and $\forall_i$ is $(\exists v)\beta_i(v)$ ($i = 1, 2$) then $(\Delta_1 \lor \forall_2) <
(\Delta_1 \lor \forall_2)$ means the formula saying "there is a witness for $\alpha_1 \lor \alpha_2$ less than
each witness for $\beta_1 \lor \beta_2$", similarly for $<\#$ instead of $\lt$.

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

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

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 $\Pi_n$-conservative
$\Sigma_n$-sentence, (2) a hereditarily $\Sigma_n$-conservative $\Pi_n$-sentence, (3) a doubly
$\Pi_n$-conservative $\Sigma_n$-sentence (its negation is thus a doubly $\Sigma_n$ conservative
$\Pi_n$-sentence).

Examples ($\Gamma$ is $\Sigma_n$, $\Lambda$ is $\Pi_n$):

(1) $\xi$ such that $I\Sigma_1 \vdash \xi \iff Pr_{\Gamma}(\neg \xi) \not< \# Pr_{\Gamma}(\xi)$,

(2) $(-\xi)$ such that $I\Sigma_1 \vdash \xi \vdash Pr_{\Gamma}(\neg \xi) \not< \# Pr_{\Lambda}(\xi)$,

(3) $\xi$ such that $I\Sigma_1 \vdash \xi \subseteq Pr_{\Gamma}(\neg \xi) \not< \# Pr_{\Lambda}(\xi)$.

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

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

$$Th \subseteq hConsu(\Gamma) \subseteq Consu(\Sigma_1) \cap Consu(\Pi_1) \subseteq NRef$$

and there is no set $X$ such that

(1) $X$ is $\Delta_1$ and $Th \subseteq X \subseteq NRef$, or

(2) $X$ is $\Pi_1$ and $Th \subseteq X \subseteq Consu(\Gamma)$, or

(3) $X$ is $\Sigma_1$ and $hConsu(\Gamma) \subseteq X \subseteq NRef$, or

(4) $X$ is $\Sigma_2$, $\Gamma \supseteq \Sigma_1$ and $hConsu(\Gamma) \subseteq X \subseteq Consu(\Sigma_1)$, or

(4') $X$ is $\Sigma_2$, $\Gamma \supseteq \Pi_1$ and $hConsu(\Gamma) \subseteq X \subseteq Consu(\Pi_1)$. 
III. Self-Reference

4.27 Theorem (Π₂-completeness). For each \( n \geq 1 \) and \( \Gamma = \Sigma_n \) or \( \Pi_n \), both \( \text{Cons}(\Gamma) \) and \( h\text{Cons}(\Gamma) \) is \( \Pi_2 \)-complete.

4.28 Remark. (1) Concerning 4.26, observe that \( Th \) is \( \Sigma_1 \), \( N\text{Ref} \) is \( \Pi_1 \) and \( \text{Cons}(\Gamma) \) is \( \Pi_2 \). Thus the result is optimal: in 4.26 (1) \( \Delta_1 \) can be replaced by neither \( \Sigma_1 \) nor \( \Pi_1 \), in 4.26 (2) \( \Pi_1 \) cannot be replaced by \( \Sigma_1 \), in 4.26 (3) \( \Sigma_1 \) cannot be replaced by \( \Pi_1 \), and in 4.26 (4, 4') \( \Sigma_2 \) cannot be replaced by \( \Pi_2 \).

(2) Recall from recursion theory that a set \( X \in \Pi_2 \) is incomplete if \( X \in \Pi_2 \) and each \( \Pi_1 \) set \( Y \) is \( \Delta_1 \)-reducible to \( X \), i.e. for some \( \Delta_1 \) 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 \( \Phi, \Psi \) be \( \Sigma_1 \) formulas.

(1) Let \( I\Sigma_1 \vdash \xi \equiv [(Pr(\neg \xi) \lor \Phi) \not<_{\Sigma_1} (Pr(\xi) \lor \Psi)] \). Then

\[(i) \quad T \vdash \xi \iff N \models \Phi \not<_{\Sigma_1} \Psi \iff N \models \xi; \]
\[(ii) \quad T \vdash \neg \xi \iff N \models \Psi \not<_{\Sigma_1} \Phi. \]

(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 \( \Delta_1 \) definition of \( T_i \). Let

\[ I\Sigma_1 \vdash \xi \equiv [(Pr_1(\neg \xi) \lor Pr_2(\neg \xi) \lor \Phi) \not<_{\Sigma_1} (Pr_1(\xi) \lor Pr_2(\xi) \lor \Psi)]. \]

Then

\[(i) \quad T_1 \vdash \xi \iff T_2 \vdash \xi \iff N \models \Phi \not<_{\Sigma_1} \Psi \iff N \models \xi; \]
\[(ii) \quad T_1 \vdash \neg \xi \iff T_2 \vdash \neg \xi \iff N \models \Psi \not<_{\Sigma_1} \Phi. \]

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

(1) Let \( \chi(y) \) be \( \Sigma_n \) and let \( I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\neg \xi) \not< \# (\exists y)\neg \chi(y) \). Then

\[(i) \quad \text{for each } m, (T + \xi) \vdash \chi(m), \]
\[(ii) \quad \text{for each } I\Sigma_1 \subset T_0 \subset T \text{ and each } \Pi_n \text{-sentence } \pi, (T_0 + \xi) \vdash \pi \text{ implies } T_0 + \{ \chi(m) \mid m \} \vdash \pi. \]

(2) Let \( \chi(y) \) be \( \Pi_n \) and let \( I\Sigma_1 \vdash \xi \equiv (\exists y)\neg \chi(y) \not< \# Pr_{\Pi_n}(\xi) \). Then

\[(i) \quad \text{for each } m, (T + \neg \xi) \vdash \chi(m), \]
\[(ii) \quad \text{for each } I\Sigma_1 \subset T_0 \subset T \text{ and each } \Sigma_n \text{-sentence } \beta, (T_0 + \neg \xi) \vdash \beta \text{ implies } T_0 + \{ \chi(m) \mid m \} \vdash \beta. \]

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 \( \Sigma \) set has a sound \( \Sigma \) 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 $T_i \subseteq IΣ_1$ ($i = 1, 2$) and for both $T_i$ we have a fixed
definition of $T_i$ which is $Δ_1$ in $IΣ_1$ (i.e. binumerates $T_i$ in $IΣ_1$). Our diagonal
formula $ξ$ satisfies

$$IΣ_1 \vdash ξ \equiv [(Pr_1(¬ξ) \lor Pr_2(¬ξ) \lor Φ) \land (Pr_1(ξ) \lor Pr_2(ξ) \lor Ψ)]$$

where $Φ$, $Ψ$ are fixed $Σ_1$ formulas. Thus $ξ$ is $Σ_1$ in $IΣ_1$.

(1) First observe that if any of the formulas $ξ$, $Pr_1(ξ)$, $Pr_2(ξ)$, $Pr_1(¬ξ)$,
$Pr_2(¬ξ)$, $Φ$, $Ψ$ is true in $N$ (i.e. has a standard witness $d$) then $ξ$ becomes
$Δ_1$ in $IΣ_1$ since the existential quantitier in the witness comparison formula
can be bounded by $d$.

(2) Thus we get

$$T_i \vdash ξ \Rightarrow N \models ξ \Rightarrow T_j \vdash ξ,$$

$$T_i \vdash ¬ξ \Rightarrow (N \models ¬ξ \text{ and } T_i \vdash ¬ξ)$$

$(i = 1, 2)$—note that $T_i \vdash ξ$ iff $N \models Pr_i(ξ)$.

(3) Furthermore, $T_i \vdash ξ$ implies $N \models Φ \land Ψ$ and similarly, $T_i \vdash ¬ξ$ implies
$N \models Ψ \land ¬Φ$. 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 $y \leq d$ which witnesses the
antecedent ($Pr_1(¬ξ) \lor Pr_2(¬ξ) \lor Φ$) and ($∀z < y)(z$ does not witness the
succeedent). Let $e \in N$ be such a $y$. But $e$ cannot witness $Pr_j(¬ξ)$ since this
would make $T_1$ or $T_2$ contradictory; thus $e$ witnesses $Φ$ and hence $Φ \land Ψ$.
We have $N \models Φ \land Φ$.

(4) Conversely, $N \models Φ \land Ψ$ implies $T_i \vdash ξ$ since if $N \models Φ \land Φ$ then $N \models Φ$,
which makes $ξΔ_1$ in $IΣ_1$ and hence decidable (provable or refutable) in $T_i$.
But $T_i$ cannot refute $ξ$ since, by (3), this would imply $N \models Φ \land ¬Φ$. Similarly,
$N \models Ψ \land ¬Φ$ implies $T_i \vdash ¬ξ$. This gives all we need.

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

**4.33 Theorem.** Let $φ(x, y)$, $ψ(x, y)$ be $Σ_0$-formulas (or: formulas $Δ_1$ in $IΣ_1$),
put $Φ(x) \equiv (∃y)φ(x, y)$, $Ψ(x) \equiv (∃y)ψ(x, y)$.

(1) Let $T$ be as in 4.29 (1). Assume that for each $k$ the formula $ξ(x)$ satisfies
the following:

$$T \vdash ξ(k) \equiv Pr(¬ξ(k)) \lor Φ(k) \land Pr(ξ(k)) \lor Ψ(k).$$

Then

$$T \vdash ξ(k) \text{ iff } N \models Φ(k) \land Ψ(k) \text{ iff } N \models ξ(k);$$
$$T \vdash ¬ξ(k) \text{ iff } N \models Ψ(k) \land ¬Φ(k).$$
(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:

$$\xi(k) \equiv (Pr_1(\overline{\xi(k)}) \lor Pr_2(\overline{\xi(k)}) \lor \Phi) \land (Pr_1(\xi(k)) \lor Pr_2(\xi(k)) \lor \Psi).$$

Then

$$T_1 \vdash \xi(k) \iff T_2 \vdash \xi(k) \iff N \vDash \xi(k) \iff N \vDash \Phi(k) \land \Psi(k),$$

$$T_1 \vdash \overline{\xi(k)} \iff T_2 \vdash \overline{\xi(k)} \iff N \vDash \overline{\Psi(k)} \iff \overline{\Phi(k)}.$$

The proof is fully analogous to 4.32; the existence of $\xi$ 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 $\Sigma_1$ sets then there is a $\Sigma_1$-formula $\xi(x)$ such that, for each $k$,

$$k \in X \iff T \vdash \xi(k),$$

$$k \in Y \iff T \vdash \overline{\xi(k)}.$$  

Proof. Take $\Sigma_1$-definitions of $X$ and $Y$ for $\Phi(x)$ and $\Psi(x)$.

4.35 Corollary. (1) Each $\Sigma_1$ set $X$ has a sound numeration in $T$, i.e. a formula which both defines $X$ and numerates $X$ in $T$.

(2) Each $\Sigma_1$ set has a $\Pi_1$ numeration in $T$.

4.36 Theorem (= 4.26 (1)). There is no $\Delta_1$ set $X$ such that $Th \subseteq X \subseteq NRef$.

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

4.37 Theorem. For each $\Pi_1$ set $X$ there is a formula $\xi(x)$ such that, for each $k$,

(1) $k \in X \iff \xi(k)$ is neither $\Sigma_1$-conservative nor $\Pi_1$-conservative over $T$,

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

Proof. Observe that the formula $\xi(x)$ constructed for the $\Sigma_1$-set $-X$ as in 4.29 is $\Sigma_1$, thus $\Sigma_1$-nonconservative whenever unprovable. Its $\Pi_1$-nonconservativeness is proved similarly as the $\Pi_1$-nonconservativeness of Rosser's formula.

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

4. Partial Conservativity and Interpretability

**Proof.** Assume we have such an $X$ and take the corresponding formula $\zeta(x)$ from 4.37. Let $T \vdash \varphi \equiv \zeta(\overline{\varphi})$; if $\varphi \notin X$ then $T \vdash \zeta(\overline{\varphi})$, thus $T \vdash \varphi$ and $\varphi \in X$, a contradiction. On the other hand, if $\varphi \in X$ then $\zeta(\overline{\varphi})$ (and hence $\varphi$) is $\Sigma_1$-nonconservative as well as $\Pi_1$-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 $\Delta_1$ and consistent, let $X, Y$ be disjoint $\Sigma_1$-sets. Then there is a $\Sigma_1$-formula $\zeta(x)$ such that, for each $k$,

$$
k \in X \text{ iff } T_1 \vdash \zeta(\overline{k}) \text{ iff } T_2 \vdash \zeta(\overline{k}) \text{ iff } N \vdash \zeta(\overline{k}),
$$

$$
k \in Y \text{ iff } T_1 \vdash \neg \zeta(\overline{k}) \text{ iff } T_2 \vdash \neg \zeta(\overline{k}).
$$

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

**Proof** evident.

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

4.40 Lemma. For each $\varphi$ and $d$,

$$
T \vdash \text{Proof}_{T}^\Gamma(\overline{d}, \overline{\varphi}) \rightarrow \varphi;
$$

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

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

Assume $\text{Proof}_{T}^\Gamma(\overline{d}, \overline{\varphi})$. Then for some $y \in \Gamma$, $y \leq \overline{d}$, we have $\text{Proof}_{T}^\Gamma_{(\Gamma + x)}(\overline{d}, \overline{\varphi})$ and $Tr_T(y)$, i.e. for some $\psi \leq d$ we have $Tr_T(\overline{\varphi})$ and $\text{Proof}_{T}^\Gamma_{(\Gamma + \varphi)}(\overline{d}, \overline{\psi})$.

Thus we get $\psi$. Having this we successively prove $\varphi_1, \varphi_2, \ldots, \varphi_n$, hence $\varphi$.

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

(1) Let $I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\neg \overline{\xi}) \leftarrow (\exists y) \neg \chi(y), \chi \in \Sigma_n$;

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 $\Pi_n$-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 $\xi$, thus $(\exists y < \overline{m})\text{Proof}_{\Sigma_n}^\Gamma(y, \neg \overline{\xi})$; 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 $\pi$ be a $\Pi_n$-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 $\Sigma_n$ equivalent $\delta$ of $\neg \pi$. Let us work in $T_0 + \{\chi(m) \mid m\} + \neg \xi$, we want to prove $\pi$. Assume the contrary, i.e. $\delta$; then $Tr(\delta)$, i.e. $Proof_{\Sigma_n}(\bar{d}, \neg \xi)$, hence $d$ is a witness of the antecedent of $\xi$. This gives $(\exists y \leq \bar{d}) \neg \chi(y)$, i.e. $\vee_{\delta \leq d} \neg \chi(m)$, which is a contradiction in our theory. This proves (1).

(2) Now assume $I\Sigma_1 \vdash \xi \equiv (\exists y) \neg \chi(y) \prec^\# Pr_{\Pi_n}(\bar{\xi})$, $\chi \in \Pi_n$.

Let $T_0$ be as above. We shall prove

(a) $(\forall m)(T + \neg \xi \vdash \chi(m))$,

(b) $(T_0 + \neg \xi)$ is $\Sigma_n$-conservative for $T_0 + \{\chi(m) \mid m\}$.

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

Now assume $I\Sigma_1 \subseteq T_0 \subseteq T$ and let $\pi$ be a $\Pi_n$-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 $\Sigma_n$ equivalent $\delta$ of $\neg \pi$. Let us work in $T_0 + \{\chi(m) \mid m\} + \neg \xi$, we want to prove $\pi$. Assume the contrary, i.e. $\delta$; then $Tr(\delta)$, i.e. $Proof_{\Sigma_n}(\bar{d}, \neg \xi)$, hence $d$ is a witness of the antecedent of $\xi$. This gives $(\exists y \leq \bar{d}) \neg \chi(y)$, i.e. $\vee_{\delta \leq d} \neg \chi(m)$, which is a contradiction in our theory. This proves (1).

(2) Now assume $I\Sigma_1 \vdash \xi \equiv (\exists y) \neg \chi(y) \prec^\# Pr_{\Pi_n}(\bar{\xi})$, $\chi \in \Pi_n$. Let $T_0$ be as above. We shall prove

(a) $(\forall m)(T + \neg \xi \vdash \chi(m))$,

(b) $(T_0 + \neg \xi)$ is $\Sigma_n$-conservative for $T_0 + \{\chi(m) \mid m\}$.

Let us work in $T_0 + \neg \xi + \chi(m)$. Then $m$ witnesses the antecedent of $\xi$, and $\neg \xi$ implies $(\exists y \leq m) Proof_{\Pi_n}(\bar{\xi}, y)$, hence $\xi$ 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 $T_0 + \{\chi(m) \mid m\} + \xi$; it suffices to prove $\delta$. Assume the contrary, i.e. $Tr(\neg \delta)$; then $Proof_{\Sigma_n}(\bar{d}, \xi)$, i.e. $\xi$ gives $(\exists y \leq \bar{d}) \neg \chi(y)$, 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 $\xi$; second, we may parametrize the whole formula. We obtain the following four cases ($T$, $Pr_{\Sigma_n}$, $Pr_{\Pi_n}$ as above, $I\Sigma_1 \subseteq T_0 \subseteq T$, $\chi(x, y)$ is $\Sigma_n$ in (1), (2) and is $\Pi_n$ in (3), (4).

(1) $I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\bar{\xi}) \prec^\# (\exists y) \neg \chi(\bar{\xi}, y)$.

Properties: $\xi$ is $\Sigma_n$ in $I\Sigma_1$; $(T + \xi) \vdash \chi(\bar{\xi}, m)$ for all $m$; $(T_0 + \xi)$ is $\Pi_n$-conservative for $T_0 + \{\chi(\bar{\xi}, m) \mid m\}$.

(2) $I\Sigma_1 \vdash \xi(k) \equiv Pr_{\Sigma_n}(\bar{\xi}(k)) \prec^\# (\exists y) \neg \chi(k, y)$.
Properties: \( \xi(x) \) is \( \Sigma_n \) in \( I \Sigma_1 \); for all \( k, m \), \( (T + \xi(k)) \chi(k, m) \); for all \( k \), \( (T_0 + \xi(k)) \) is \( \Pi_n \)-conservative for \( T + \{ \chi(k, m) \mid m \} \).

\[
I \Sigma_1 \vdash \xi \equiv (\exists y) \neg \chi(\overline{\xi}, y) \prec \# P r^*_{\Pi_n}(\overline{\xi}).
\]

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

\[
I \Sigma_1 \vdash \neg \xi(\overline{k}) \equiv (\exists y) \neg \chi(\overline{\neg \xi}, y) \prec \# P r^*_{\Pi_n}(\overline{\neg \xi}).
\]

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

Proofs are obvious modifications of 4.41.

4.43 Theorem. Let \( \Gamma \) be \( \Sigma_n \) or \( \Pi_n \), \( n \geq 1 \). For each \( X \in \Sigma_1 \) there is a \( \Gamma \)-formula \( \xi \) 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 \( \bar{\Gamma} \)-conservative over \( T \). (\( \bar{\Gamma} \) is the dual of \( \Gamma \)).

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 \( (\exists y) \neg \chi(k, y) \) then

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

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

**Lemma.** Assume \(T \vdash Pr(x) \rightarrow \Delta(x) \rightarrow Pr_{\Pi}(x)\), and \(T \vdash Pr(x) \rightarrow \forall(x) \rightarrow Pr_{\Lambda}(x)\), \(T \vdash \xi \equiv \Delta(\neg\xi) \rightarrow \# \forall(\xi)\). Then neither \(\xi\) nor \(\neg\xi\) is \(T\)-provable.

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

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

\[
I\Sigma_1 \vdash \chi(\overline{\xi}, \overline{m})\]

for each \(m\), \(\neg\xi\) is unprovable and hereditarily \(\Sigma_n\)-conservative over \(T\).

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

\[
I\Sigma_1 \vdash \chi(\overline{\xi}, \overline{m})\]

for each \(m\), \(\neg\xi\) is unprovable and hereditarily \(\Sigma_n\)-conservative over \(T\).

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

\[
I\Sigma_1 \vdash \chi(\overline{\xi}, \overline{m})\]

for each \(m\), \(\neg\xi\) is unprovable and hereditarily \(\Sigma_n\)-conservative over \(T\).

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

**4.46 Theorem.** Let \(T\) be \(\Sigma_n\) or \(\Pi_n\), \(n \geq 2\). Let \(X\) be \(\Pi_2\). Then there is a \(\Gamma\)-formula \(\chi(x)\) such that, for each \(k\),

\[
k \in X \equiv \chi(\overline{k})\]

is hereditarily \(\Gamma\)-conservative,

\[
k \notin X \equiv \chi(\overline{k})\]

is \(\Sigma_1\)-nonconservative.

The same is true for \(\Sigma_1\) replaced by \(\Pi_1\) (with another \(\xi\)).

**Proof.** Let \(X\) be \(\{k \mid (\forall m)R(k, m)\}\) where \(R\) is \(\Sigma_1\). By 4.39, let \(\rho(x, y)\) be a \(\Sigma_1\)-numeration (\(\Pi_1\)-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 $\Gamma$; if $k \in X$ then $I\Sigma_1 \vdash \rho(k, \bar{m})$ for all $m$, hence $\xi(k)$ is hereditarily $\Gamma$-conservative over $T$. If $k \notin X$ then, for some $m_0$, $T$ does not prove $\rho(k, \bar{m}_0)$; thus $\xi(k)$ is not $\Sigma_1$-conservative ($\Pi_1$-conservative) over $T$.

4.47 Theorem (= 4.26). Let $\Gamma = \Sigma_n$ or $\Pi_n$, $n \geq 2$. There is no $Y \subseteq \Sigma_2$ such that $h\text{Consv}(\Gamma) \subseteq Y \subseteq \text{Consv}(\Sigma_1)$ or $h\text{Consv}(\Gamma) \subseteq Y \subseteq \text{Consv}(\Pi_1)$.

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

4.48 Theorem (= 4.27). For each $\Gamma = \Sigma_n$, $\Pi_n$ ($n \geq 1$), both $\text{Consv}(\Gamma)$ and $h\text{Consv}(\Gamma)$ is $\Pi_2$-complete.

Proof. Clearly, $\text{Consv}(\Gamma)$ is $\Pi_2$; to see that $h\text{Consv}(\Gamma)$ is $\Pi_2$, observe that $\varphi$ is hereditarily $\Gamma$-conservative over $T$ iff for each finite $T_0 \subseteq T$ containing $I\Sigma_1$, $\varphi$ is $\Gamma$-conservative over $T_0$. For each $\Gamma$ (including $\Sigma_1$ and $\Pi_1$), 4.46 gives a formula $\xi(x)$ such that

$$k \in X \Rightarrow \xi(k) \text{ is hereditarily } \Gamma\text{-conservative},$$

$$k \notin X \Rightarrow \xi(k) \text{ is not } \Gamma\text{-conservative}.$$  

4.49. Now we present Švejdar's proof of the fact (claimed in 4.5 (3)) that for a $T$ not $\Sigma_1$-sound, the negation of Rosser's formula is $\Sigma_1$-nonconservative over $T$.

Let $\delta(x)$ be a $\Sigma_0$ formula such that $T \vdash (\exists x)\delta(x)$ but $N \vDash (\forall x)\neg\delta(x)$; let $\rho$ be Rosser's formula, thus $T \vdash \rho \equiv (\varphi^* (\neg \varphi) < \varphi^* (\bar{\varphi}))$. We know the following: both $\rho$ and $\neg\rho$ are unprovable in $T$;

$$T + \text{Con}^* \vdash \neg\varphi^* (\rho) \& \neg\varphi^* (\bar{\varphi})$$

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

$$T \vdash \varphi \equiv ((\exists x)\delta(x) < \varphi^* (\rho \rightarrow \varphi)).$$

(a) $\rho \rightarrow \varphi$ is $T$-unprovable. Indeed, if $d$ is a $T$-proof of $\rho \rightarrow \varphi$ then $T \vdash \bigvee_{k \leq d} \delta(k)$; but for each $k$, $T \vdash \neg\delta(k)$.

(b) $T + \text{Con}^* \vdash \varphi$. Indeed, $T \vdash (\exists x)\delta(x)$, thus $(T + \neg \varphi)$ proves $\varphi^* (\rho \rightarrow \varphi) \ll (\exists x)\delta(x)$. (This is a $\Sigma_1$-formula, call it $\sigma$), thus $T + \neg \varphi \vdash \varphi^* (\sigma)$. 


But $T \vdash Pr^*(\sigma \rightarrow \neg \varphi)$, thus $T + \neg \varphi \vdash Pr^*(\neg \varphi)$ & $Pr^*(\neg \varphi \rightarrow \varphi)$, hence $T + \neg \varphi \vdash Pr^*(\neg \varphi)$, which implies $T + \neg \varphi \vdash \neg Con^*$.  

c) $T + \neg \rho \vdash \varphi \lor (Pr^*(\neg \rho) \nsucc Pr^*(\neg \rho))$. This because $(T + \neg \rho + Con^*)$ proves $\rho$ and $(T + \neg \rho + \neg Con^*)$ proves $(Pr^*(\neg \rho) \nsucc Pr^*(\neg \rho))$.  

d) But $\varphi \lor (Pr^*(\rho) \nsucc Pr^*(\neg \rho))$ is unprovable in $T$. Assume $T$ proves this disjunction; then $T + \rho$ would prove $\varphi$ (since $T + \rho$ disproves the second disjunct); this contradicts (a). Thus $\varphi \lor (Pr^*(\rho) \nsucc Pr^*(\neg \rho))$ is the desired $\Sigma_1$ 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 ACA0 (4.55) and (ii) interpretability with partial conservativity for finitely axiomatized theories $T \supset I\Sigma_1$ (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 $Intp_T = \{ \varphi \mid (T + \varphi)$ 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, $Intp_T = Cons_T(I\Pi_1)$; thus Theorem 4.27 gives immediately the following

4.51 Corollary. If $T$ is sequential and with full induction, then $Intp_T$ is $\Pi_2$-complete.

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

Proof. This is more or less evident; $\varphi$ 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 $Interp_T^*(x, y)\Delta_1$ in $I\Sigma_1$ such that

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

$Interp_T^*$ 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 \( \Gamma \) is \( \Sigma_n \) or \( \Pi_n \), then \( \text{Cons}_T(\Gamma) - \text{Intp}_T \neq \emptyset \); there is a \( \hat{\Gamma} \)-sentence \( \varphi \) which is (hereditarily) \( \Gamma \)-conservative but is not interpretable in \( T \).

Proof. This follows directly by 4.26, since clearly \( \text{Intp}_T \subseteq N\text{Ref}_T \) and \( \text{Intp}_T \in \Sigma_1 \), i.e. \( h\text{Cons}_T(\hat{\Gamma}) \) is not a subset of \( \text{Intp}_T \). Checking the proof of 4.26 (3) (i.e. 4.44) we can see that the sentence constructed there is \( \hat{\Gamma} \). \( \square \)

4.54 Discussion. Take \( ACA_0 \) for \( T \) and \( \Pi_1 \) for \( \Gamma \). We get a \( \Sigma_1 \) formula \( \varphi \) that is \( \Pi_1 \)-conservative over \( ACA_0 \) but \( (ACA_0 + \varphi) \) is not interpretable in \( ACA_0 \). Since \( \varphi \) is \( \Pi_1 \)-conservative over \( ACA_0 \) and is \( \Sigma_1 \), i.e. a formula in the language of \( PA \), \( \varphi \) is \( \Pi_1 \)-conservative over \( PA \) and hence, by 2.40, is interpretable in \( PA \). (Similarly for \( T = GB \), i.e. G"odel-Bernays set theory and \( ZF \)-Zermel-Frankel set theory, using 3.46). Note that we can take the following self-referential formula for \( \varphi \):

\[
T \vdash \varphi \equiv Pr_{\Sigma_1}(\neg \varphi) \land (\exists y)\text{Interp}\,(\varphi, y).
\]

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 \( \Sigma_1 \)-formula \( \varphi \) such that \( (PA + \varphi) \) is interpretable in \( PA \) but \( (ACA_0 + \varphi) \) is not interpretable in \( ACA_0 \).

(2) There is a \( \Pi_1 \)-formula \( \varphi \) 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 \( \Sigma_1 \)-sentences \( \varphi \) (over an arbitrary consistent finitely axiomatized sequential theory \( T \supseteq I \Sigma_1 \)) with respect to the following questions. (i) is \( \varphi \) interpretable in \( T \)? (ii) is \( \neg \varphi \) interpretable in \( T \)? (iii) is \( \varphi \) \( \Pi_1 \)-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 \( \Sigma_1 \)-sentences. Till now, we have got some examples; e.g. for \( \neg \text{Con}^* \) 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 \( \Sigma_1 \)-sentences of all eight possible types.
III. Self-Reference

The theorem is an immediate 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 $\varphi$, $\varphi$ is $\Pi_1$-conservative over $PA$ iff $\varphi$ is $\Pi_1$-conservative over $ACA_0$ iff $(PA + \varphi)$ is interpretable in $PA$. Consequently, if $\sigma$ is an independent $\Sigma_1$-sentence then $(PA + \neg \sigma)$ is never interpretable in $PA$. Hence we have eight possible types of dependent $\Sigma_1$-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 $\Sigma_1$(non)conservativity of $\neg \sigma$ in the case of $T$ being $\Sigma_1$-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 $\Sigma_n$-formula and $\theta(x, y)$ a $\Pi_k$-formula. Let $\xi$ satisfy

$$T \vdash \xi \equiv [Pr_{\Sigma_n}{\neg \xi} \lor (\exists y)(\varphi(\neg \xi, y) \land \# Pr_{\Pi_k}{\xi}) \lor (\exists y)\theta(\xi, y)]$$

(where $Pr_{\Sigma_n}$, $Pr_{\Pi_k}$ are as in 4.23). Then

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

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

The assertion remains valid if $Pr^*$ is replaced by $HPr^*$ (Herbrand provability).

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

(2) is proved analogously.
(3) Assume \((T + \xi) \vdash \pi\) where \(\pi\) is \(\Pi_n\), let \(d\) be a \(T\)-proof of \((-\pi \rightarrow -\xi)\). Let us work in \((T + \{-\theta(\xi, m) \mid m\} + \neg \xi)\); we want to prove \(\pi\). Assume \(-\pi\), thus \(\mathcal{T} \vdash (\neg \pi)\), i.e. \(\neg \pi\) is a true \(\Sigma_n\)-sentence; but then \(\text{Proof}^*_{\Sigma_n}(\bar{d}, \neg \xi)\) and, since we assume \(-\xi\), beneath \(\bar{d}\) there is a witness for the succedent of \(\xi\). But beneath \(\bar{d}\) there is no witness for \((\exists n)\theta(\xi, n)\) (since \(-\theta(\xi, 0), -\theta(\xi, 1), \ldots, -\theta(\xi, \bar{d}))\). Neither is there a witness for \(\bar{p}^*(\xi)\) beneath \(\bar{d}\) since this would imply \(\xi\) by 4.40. This gives a contradiction.

(4) similarly.

4.61 Remark. For \(n = 0\) we may replace \(\text{Pr}_*^\Sigma_0\) by \(\text{Pr}^*\) and/or allow \(\varphi\) to be \(\Delta_1\) (in \(I\Sigma_1\), say) instead of being \(\Sigma_0\). Similarly for \(k = 0\).

4.62 Corollary. In 4.60, let \(m, n\) be arbitrary but let \(\varphi, \theta\) be \(\Sigma_1\)-in-\(I\Sigma_1\). (By 4.61, this does not exclude the possibility \(m = 0\) and/or \(n = 0\)). Let \(X, Y\) be \(\Sigma_1\)-sets defined by \((\exists y)\theta(x, y)\) and \((\exists y)\varphi(x, y)\) respectively and assume \(X, Y \subseteq \text{NRef}^*\) (non-refutable formulas). Then the sentence \(\xi\) from 4.60 is \(\Pi_n\)-conservative, \(-\xi\) is \(\Sigma_k\)-conservative (over \(T\)), \(\zeta \in X\) and \((-\xi) \notin Y\).

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. \(N \vdash (\forall x, y)(-\varphi(x, y) \vee \theta(x, y))\) (assume e.g. that all witnesses for \(X\) are even and all witnesses for \(Y\) are odd). Take \(\xi\) from 4.60, i.e.

\[ I_{\Sigma_1} \vdash \xi \equiv \text{Pr}^*_{\Sigma_n}(\neg \xi) \vee (\exists y)\varphi(\neg \xi, y) - \# \text{Pr}^*_{\Pi_k}(\xi) \vee (\exists y)\theta(\xi, y). \]

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

\[ \theta(\xi, m) \& (\forall x < m)\neg \theta(\xi, x) \& (\forall x \leq m)\neg \varphi(\neg \xi, x), \]

or \(N\) satisfies and \(T\) proves the formula resulting from the last one by interchanging \(\theta\), \(\xi\) with \(\varphi\), \(-\xi\) respectively. In the former case we get a contradiction in \((T + \xi)\) using 4.60 (1), which contradicts our assumption \(X \subseteq \text{NRef}^*\); in the latter case 4.60 (2) gives a contradiction in \((T + \neg \xi)\), which contradicts \(Y \subseteq \text{NRef}^*\). Thus we have \(\xi \notin X\) and \(\xi \notin Y\). This implies \(T \vdash \neg \theta(\xi, m)\) 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 \(\text{HP}_*^\Sigma(x), \text{HP}_*^\Sigma_1(x)\) (Herbrand provability in \(T\); Herbrand provability in \(T^*\) from a true \(\Sigma_1\)-sentence) as well as with the formula \(\text{Interp}^*(x, y)\) from 4.52 (interpretability of \((T^* + x)\) in \(T^*\)). We shall investigate self-referential \(\Sigma_1\)-sentences \(\xi\) such that

\[ I_{\Sigma_1} \vdash \xi \equiv \Delta(\neg \xi) \prec \forall(\xi), \]
where the antecedent $\Delta(x)$ has one of the forms

$$HPr^* (x), \ HPr^* _{\Sigma_1} (x), \ HPr^* (x) \lor Intp^* (x), \ HPr^* _{\Sigma_1} (x) \lor Intp^* (x),$$

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

$$HPr^* (x), \ HPr^* (x) \lor Intp^* (x).$$

This will give eight cases, which are examples of our eight types of independent $\Sigma_1$-sentences. For some of the following assertions it is immaterial whether we use $HPr^*$ or $Pr^*$, 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 $\Sigma_1$-sentences;

(2) $\xi$ is interpretable iff $X = \emptyset$ (the succedent does not contain $Intp^*$);

(3) $\neg \xi$ is interpretable iff $Y = \emptyset$ (the antecedent does not contain $Intp^*$);

(4) $\xi$ is $\Pi_1$-conservative iff $n = 1$ (the antecedent contains $HPr^* _{\Sigma_1}$).

Remark. This obviously implies Theorem 4.57.

Proof. 4.62 gives directly the following: $\xi$ is independent (since $\xi$ is $\Pi_n$-conservative and $\neg \xi$ is $\Sigma_k$-conservative); if the antecedent contains $HPr^* _{\Sigma_1}$ then $\xi$ is $\Pi_1$-conservative; if $X = Intp_T$ then $\xi$ 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 $T \vdash \xi = \Delta (\neg \xi) \prec HPr^* (\xi)$. It suffices to interpret $(T + \xi)$ in $(T + \neg \xi)$. By 3.20 there is a $(T + \neg \xi)$-cut $J$ such that $(T + \neg \xi) \vdash HCon^* J (T^* + \neg \xi)$. Let us work in $(T + \neg \xi)$. In $J$ there is no $H$-proof of $\xi$ in $T^*$. But $\neg \xi$ implies that beneath each witness for $\Delta (\neg \xi)$, in particular, beneath each $H$-proof of $\neg \xi$ in $T^*$ there is a $H$-proof of $\xi$ in $T^*$, thus in $J$ there is no $H$-proof of $\neg \xi$ in $T^*$ and hence $HCon^* J (T^* + \xi)$. By 3.39, $(T + \xi)$ is interpretable in $(T + \neg \xi)$.

(3) Assume $Y = \emptyset$; we prove $(-\xi) \in Intp_T$. We have $T \vdash \xi \equiv HPr^* (\neg \xi) \prec \forall (\xi)$ where $\Gamma = \Sigma_0$ or $\Sigma_1$. 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 $(\forall u \in Tr (\Sigma_1)) \ HCon^* J (T^* + \xi + u)$, i.e. $J$ does not contain any $H$-proof of $(-T^* \lor \neg \xi \lor u)$. But $\xi$ implies that there is a $H$-proof $y$ of $\neg \xi$ from $T$ and from a true $\Sigma_1$-formula $u_0$ (i.e. a $H$-proof of $-T^* \lor \neg u_0 \lor \neg \xi$) such that beneath $y$ there is no witness for $\forall (\xi)$, in particular, no $H$-proof of $\xi$ from $T^*$, hence we have $HCon^* J (T^* + \neg \xi)$; by 3.39, $(T + \neg \xi)$ is interpretable in $(T + \xi)$.

(4) Assume $n = 0$, i.e. $HPr^* _{\Sigma_1}$ does not occur in $\forall$. We prove that $\xi$ is $\Pi_1$-nonconservative. Clearly, $\xi$ implies (in $T$) the formula saying "beneath
each witness for $\forall(\xi)$ there is a witness for $\forall(-\xi)$" (call it $\pi$) and under our assumptions $\pi$ is $\Pi_1$. It suffices to show that $\pi$ is unprovable in $T$. This is done analogously to the corresponding proof concerning Rosser's formula. If $\pi$ were $T$-provable we would have

$$T \vdash (\text{there exists a witness for } \forall(\xi) \rightarrow \xi)$$

(consider the least witness for $\forall(-\xi)$, which exists by $I\Sigma_1$). Thus we would have

$$T \vdash \neg \xi \rightarrow (\forall(\xi) \text{ has no witness}),$$

$$T \vdash \neg \xi \rightarrow \neg HP^s(\xi),$$

$$T \vdash \neg \xi \rightarrow HCon^s(T^* + \neg \xi),$$

$$T \vdash \neg \xi \rightarrow Con^s(T^* + \neg \xi)$$

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

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 $\Delta_1$ function mapping $N$ into $N$, let $\Gamma = \Sigma_k$ or $\Pi_k$ ($k \geq 1$). There is a $\varphi$ $\Gamma$-conservative over $T$ and such that, for each $n$, there is a proof $d > n$ of a $\Gamma$-sentence $\gamma$ in $(T + \varphi)$ such that each $T$-proof of $\gamma$ is bigger than $f(d)$.

Proof. Otherwise $Conv(\Gamma)$ would be $\Sigma_2$, which contradicts 4.27. (Since a $\Pi_2$-complete set is not $\Sigma_2$.) \qed

4.67 Definition and discussion. A sentence $\varphi$ is a self-prover if $T \vdash \varphi \rightarrow Pr^s(\varphi)$. Clearly, each $\Sigma_1$-sentence is a self-prover. We show that there are other self-provers as well. It is easily seen that for each sentence $\psi$, the formula $\psi \& Pr^s(\psi)$ is a self-prover; it $\varphi$ is a self-prover then $T \vdash \varphi$ $\equiv (\varphi \& Pr^s(\varphi)).$

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 $\Gamma$.

Proof. By 4.26 (3), let $\psi$ be such that $\neg \psi$ is $\Gamma$-conservative over $T$ and not provable in $(T + \neg Con^s)$ (i.e. take $X = \{\alpha \mid (T + \neg Con) \vdash \alpha\}$). Let $\varphi$ be $\psi \& Pr^s(\psi)$. By our assumption on $\Gamma$, $\varphi$ is $\Gamma$ in $T$. We prove that $\varphi$ is not $\bar{\Gamma}$ (dual class). Assume $T \vdash \varphi \equiv \sigma$, $\sigma \in \bar{\Gamma}$. Then $T \vdash \varphi \rightarrow \neg \sigma$, hence $T \vdash \neg \psi \rightarrow \neg \sigma$, and since $\neg \psi$ is $\Gamma$-conservative we get $T \vdash \neg \sigma$, thus $T \vdash \neg \varphi$,
T ⊢ \neg (\psi \& Pr^*(\psi)), T ⊢ Pr^*(\psi) → \neg \psi, T ⊢ \neg Con^* → \neg \psi, which contradicts our assumption. □

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 Γ in T) and hereditarily exactly Λ-conservative (not Λ-conservative). For example, let Γ = Σ_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 + π). Then take φ \equiv π & σ.