TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 254, October 1979

PARTIALLY CONSERVATIVE EXTENSIONS OF ARITHMETIC

BY

D. GUASPARI

ABSTRACT. Let T be a consistent r.e. extension of Peano arithmetic; Σ_n^0, Π_n^0 the usual quantifier-block classification of formulas of the language of arithmetic (bounded quantifiers counting "for free"); and Γ, Γ' variables through the set of all classes Σ_n^0, Π_n^0. The principal concern of this paper is the question: When can we find an independent sentence $\phi \in \Gamma$ which is Γ'-conservative in the following sense: Any sentence χ in Γ' which is provable from $T + \phi$ is already provable from T? (Additional embellishments: Ensure that ϕ is not provably equivalent to a sentence in any class "simpler" than Γ; that ϕ is not conservative for classes "more complicated" than Γ'.) The answer, roughly, is that one can find such a ϕ, embellishments and all, unless Γ and Γ' are so related that such a ϕ *obviously* cannot exist. This theorem has applications to the theory of interpretations, since "ϕ is Γ-conservative" is closely related to the property "$T + \phi$ is interpretable in T"–or to variants of it, depending on Γ. Finally, we provide simple model theoretic characterizations of Γ-conservativeness. Most results extend straightforwardly if extra symbols are added to the language of arithmetic, and most have analogs in the Levy hierarchy of set theoretic formulas (T then being an extension of ZF).

Introduction. The main result of this paper is a theorem which is "obviously true" and of which the following is an instance: There is a Σ_9^0 sentence ϕ in the language of arithmetic such that: ϕ is essentially Σ_9^0 (i.e., not provably equivalent in arithmetic to a Π_9^0 sentence); any Σ_5^0 sentence provable in arithmetic $+\phi$ is already provable in arithmetic; some Π_5^0 sentence provable in arithmetic $+\phi$ is *not* provable in arithmetic. (According to Matijasevic's theorem it does not, up to an equivalence provable in arithmetic, matter whether or not we allow bounded quantifiers "for free" in the definition of Σ_n^0.) The same is true for many extensions of arithmetic, and an analogous result holds for the Levy hierarchy of formulas over set theory.

Consider, however, some obstacles to its proof: Let T be an r.e. extension of Peano arithmetic. If Γ is a set of formulas say that ϕ is Γ-con (over T) if every sentence from Γ provable in $T + \phi$ is provable in T; that ϕ is Γ-non, otherwise. For the present Γ will always be Π_n^0 or Σ_n^0. $\neg\mathrm{Con}(T)$ is a Σ_1^0 sentence which is Π_1^0-con over T, but might also be T provable. A Rosser sentence for T is guaranteed to be independent but will always be Π_1^0-non–as

Presented to the Society, January 28, 1977; received by the editors September 26, 1977 and, in revised form, October 27, 1977.

AMS (MOS) subject classifications (1970). Primary 02D99; Secondary 02H05.

will every independent sentence obtained by an "effectively inseparable sets" construction. So to produce an independent Σ_1^0 sentence which is Π_1^0-con requires, in some sense, another general construction for generating independent sentences.

Sentences which are Π_1^0-non make new *truths* provable. For example, one can produce a Σ_1^0 sentence σ such that arithmetic $+\sigma$ is consistent and proves Π_1^0 sentences (*truths*) not provable in ZFC, while $ZFC + \sigma$ is inconsistent.

The existence of Σ_1^0 sentences which are Π_1^0-con can be equivalently stated: If T is a consistent r.e. extension of arithmetic, there are Π_1^0 truths unprovable in any consistent extension of T obtained by adding only Σ_1^0 sentences. This is a limitation on provability in false extensions of (false) theories.

After this paper was essentially completed Smorynski called attention to the papers **[Hajek, 1]**, **[Hajek, 2]**, **[Hajkova-Hajek]** dealing with the closely related notion of interpretability: Say that ϕ is interpretable in T if $T + \phi$ is interpretable in T. For a broad class of theories "interpretable" is equivalent to "Π_1^0-con", and for such theories our methods immediately establish: There is a Σ_1^0 sentence ϕ which is interpretable in T (so, of course, relatively consistent with T) but whose relative consistency with T cannot be proven in, say, PA. (The point here is that ϕ is so simple.)

The paper is organized as follows: §1 lays out various examples from nature (Con(T), $\neg$Con(T), Rosser sentences) and shows that we will have to look elsewhere for inspiration in proving the main (obviously true) existence theorem. The heart of the paper is §2. It contains the basic existence theorems and some embellishments thereof. §3 compares the notions "Π_n^0-con" with variants of the notion "interpretable." Straightforward generalizations to larger languages and to an analogous theory for the Levy hierarchy of formulas of set theory are noted in §4. The syntactical part of the paper is concluded in §5 by mentioning some open questions, in particular, that of classifying $\{\phi | \phi \text{ is } \Gamma\text{-con}\}$ in the arithmetical hierarchy. (Solovay has solved a particularly interesting special case originally proposed by Hajek.) §6 contains model theoretic characterizations of Γ-con. E.g., ϕ is Σ_1^0-con over PA (peano arithmetic) iff every countable model of PA contains an initial segment which models $PA + \phi$. (Π_1^0-con can be characterized dually.) The proofs of these theorems are independent of the previous sections. All follow easily from theorems of Friedman which characterize the end extensions of countable models of arithmetic. For models of set theory one and only one new twist occurs. Fact: ϕ is Π_1-con ("Π_1" refers to the Levy hierarchy) over ZF iff every countable *non-ω-model* of ZF can be end extended to model $ZF + \phi$. An example shows that the qualification "non-ω-model" cannot be dropped. Open question: If you *do* drop it, what do you get?

Thanks are due Craig Smorynski and Bob Solovay for helpful correspondences and for generously allowing theorems of theirs to be included in this paper. Thanks are also due Tom McLaughlin who helped weed out some proofs of $0 = 1$ from those of $0 = 0$.

0. Notation and preliminaries. Every theory considered will be formulated in a language satisfying the following conditions: There is a special sort, the number sort, with $u, v, \ldots, z$ among the number variables. In addition to both unbounded quantifiers over each sort there are bounded quantifiers–$\forall x < y$, $\exists w < z$, etc.–over the number sort. Among the nonlogical signs are the signs of arithmetic–namely $+$, $\cdot$, $<$, 0, $'$–applicable only to terms and variables of the number sort.

LA is the language made up from the signs of arithmetic, variables of the number sort, and quantification over the number sort. The theory N (see [**Shoenfield**, p. 22]) is a finitely axiomatizable theory of recursive arithmetic. That is, N decides (correctly) every sentence of *LA* containing no unbounded quantifiers. That fact is formalizable in P: (notation explained later in this section) for any sentence σ in Σ_1^0, $PA \vdash \sigma \rightarrow \mathrm{Thm}_N(\ulcorner \sigma \urcorner)$. *PA*, Peano arithmetic, is N + induction for all formulas of *LA*.

Notice that *ZF* and *GB*, though not ordinarily expressed in languages that meet the requirements above, can easily be reformulated in languages that do so.

CONVENTION. "Theory" means "consistent theory containing *PA*".

So, e.g., S is a subtheory of T means that $PA \subseteq S \subseteq T$. This convention is too strong. In the sequel we can almost always get by on just the assumption that our theories contain N. Most of the time that part of the theory outside *LA* will be quite hazy and unimportant.

Define the Σ_n^0, Π_m^0 formulas of *LA* as usual: $\Sigma_0^0 = \Pi_0^0 = \Delta_0^0 =$ the class of formulas containing no unbounded quantifiers; Σ_{n+1}^0 is the class of formulas $\vec{Q}\vec{x}\chi$ with χ in Π_n^0, $\vec{x}$ a sequence of number variables, and $\vec{Q}$ a possibly empty sequence of quantifiers not containing $\forall$. Π_{n+1}^0 is defined dually. Notice that in this way each class of formulas is literally a subset of any class of "ostensibly more complicated" formulas. Every Σ_n^0 formula is equivalent (in N) to a 'prenex' Σ_n^0 formula: one with prefix a strictly alternating sequence of n unbounded quantifiers followed by a 'matrix' containing no unbounded quantifiers. For any theory T, T-Σ_n^0 is the class of formulas T provably equivalent to Σ_n^0 formulas; etc.; and T-Δ_n^0 is T-$\Sigma_n^0 \cap T$-Π_n^0. $\Gamma, \Gamma', \ldots$ will vary over the classes Σ_n^0, Π_m^0; and T-$\Gamma, \ldots$ over T-Σ_n^0, etc. $\breve{\Gamma}$ is the dual class to Γ: i.e., $\breve{\Sigma}_n^0 = \Pi_n^0$ and $\breve{\Pi}_n^0 = \Sigma_n^0$.

DEFINITION 0.1. If ϕ, θ are formulas of form $\exists x\chi(x)$, $\exists x\psi(x)$ respectively, then

$$\phi \preccurlyeq \theta =_{\mathrm{df}} \exists x(\chi(x) \wedge \forall y < x \neg \psi(y)),$$

$$\phi \prec \theta =_{\mathrm{df}} \exists x(\chi(x) \wedge \forall y \leqslant x \neg \psi(y)).$$

Notice that if ϕ and θ are 'prenex' Σ_n^0 formulas then so are $\phi \preccurlyeq \theta$ and $\phi \prec \theta$. More generally, if ϕ is Σ_n^0 and χ is T-Δ_n^0, then $\phi \preccurlyeq \theta$ is T-Σ_n^0. Abuses of this notation will include writing $\phi \prec (\neg\theta)$ when, e.g., ϕ is Σ_1^0 and θ is Π_1^0. Smorynski has pointed out an unfortunate defect in this notation–namely, that it suggests that $\phi \preccurlyeq \phi$ is true, which is not always the case.

Notice also that if ϕ and θ are 'prenex' Σ_1^0 and either one of them is true, then $\phi \preccurlyeq \theta$ is decidable.

Formalizing syntax and semantics. We will use $i, j, \ldots, n$ metamathematically for natural numbers, and $\mathbf{n}$ for the canonical term denoting n. If k is the Gödel number of ϕ (according to some fixed standard numbering) then $\ulcorner\phi\urcorner$ is $\mathbf{k}$. (So the Gödel number (from now on, "g.n.") of ϕ is a natural number, while $\ulcorner\phi\urcorner$ is a term in LA.)

Procedures for obtaining partial truth definitions are well known. Fact 0.2 is stated primarily to establish notation.

Fact 0.2. For any Γ there is a formula $\phi(x)$ in Γ satisfying:

For any sentence $\chi \in \Gamma$, $PA \vdash \phi(\ulcorner x\urcorner) \leftrightarrow \chi$.

Choose one such ϕ and call it Γ-True.

If χ is $\exists x\phi(x)$ we will use "y is a witness for χ" to mean $\phi(y)$.

We want $\mathrm{Proof}_T(y, x)$ to be a (simple) binumeration in T expressing: y codes a proof from T of (the formula with g.n.) x. That may not be possible–e.g., if T is r.e. but nonrecursive. However, if T is r.e. there is a recursive T' equivalent to T and we can therefore take Proof_T to be a $PA - \Delta_1^0$ binumeration (in PA) of: y codes a proof from T' of x. Let $\mathrm{Thm}_T(x)$ be $\exists y\mathrm{Proof}_T(y, x)$ and $\mathrm{Con}(T)$ be $\neg\mathrm{Thm}_T(\ulcorner \mathbf{0} = \mathbf{1}\urcorner)$.

All conventions for denoting formalized operations on syntax are unsatisfactory. Here is another. Suppose we want to formalize a function which, inputing m and n–g.n.'s for ϕ and θ respectively–outputs a g.n. for $\phi \preccurlyeq \theta$. We add a defined function symbol, say g, which formalizes the function and agree to notate gxy by: "$x \preccurlyeq y$". (The quotation marks are part of the notation.) So, e.g., $N \vdash$"$\ulcorner\phi\urcorner \preccurlyeq \ulcorner\theta\urcorner$"$= \ulcorner\phi \preccurlyeq \theta\urcorner$. Another example, of another function: $N \vdash$"$\ulcorner\theta\urcorner \to \ulcorner\chi\urcorner$"$= \ulcorner\theta \to \chi\urcorner$.

Finally, we will make extensive use of a well-known lemma of Gödel.

Fact 0.3 (Self-reference lemma). If $\phi(x) \in \Gamma$ has only x free, a sentence $\chi \in \Gamma$ can be effectively found for which

$$PA \vdash \chi \leftrightarrow \phi(\ulcorner \chi\urcorner).$$

1. Basic definitions and some examples.

DEFINITION 1.1. Let T be a theory, ϕ a sentence, X any set of formulas in LA. Then,

ϕ is X-con over T iff every sentence in X provable from $T + \phi$ is provable from T,

ϕ is X-non over T iff ϕ is not X-con over T.

The next lemma has to go somewhere, so may as well go here. It characterizes Γ-con. Until referred to it is skippable.

LEMMA 1.2. *Let T be an extension of N. Then*:

(1) *A 'prenex' Π_n^0 sentence ϕ is Σ_n^0-con over T if and only if $\forall\sigma$ (σ a 'prenex' Σ_n^0 sentence and $T \vdash \sigma \Rightarrow T \vdash \sigma \prec (\neg\phi)$).*

(2) *A 'prenex' Σ_n^0 sentence ϕ is Π_n^0-con over T if and only if $\forall\sigma$ (σ a 'prenex' Σ_n^0 sentence and $T \vdash \phi \to \phi \prec \sigma \Rightarrow T \vdash \sigma \to \phi$).*

PROOF. (1) Suppose first that ϕ is Σ_n^0-con over T and that $T \vdash \sigma$. Then $T + \phi \vdash \sigma \prec (\neg\phi)$; and therefore $T \vdash \sigma \prec (\neg\phi)$. Suppose conversely that $T \vdash \sigma \Rightarrow T \vdash \sigma \prec (\neg\phi)$ for any 'prenex' Σ_n^0 sentence σ. We want to show ϕ Σ_n^0-con, so let σ' be Σ_1^0 and $T + \phi \vdash \sigma'$. Then T implies the Σ_n^0 sentence $\neg\phi \vee \sigma'$ and so, by assumption, $T \vdash (\neg\phi \vee \sigma') \prec (\neg\phi)$. But $((\neg\phi \vee \sigma') \prec \neg\phi) \to \sigma'$ and therefore $T \vdash \sigma'$.

(2) Suppose ϕ is a Σ_n^0 sentence. For reference, call the statement "for any 'prenex' sentence σ in Σ_n^0, if $T \vdash \phi \to \phi \prec \sigma$ then $T \vdash \sigma \to \phi$" by the name of (+). Suppose (+), and that π is 'prenex' Π_n^0 and $T + \phi \vdash \pi$. Then (*) $T \vdash \phi \to \pi$; and so $T \vdash \neg\phi \to \phi \prec (\neg\pi)$. Applying (+), $T \vdash \neg\pi \to \phi$; i.e. (**) $T \vdash \neg\phi \to \pi$. Combining (*) with (**), $T \vdash \pi$. Now suppose that (+) is false and produce a Σ_n^0 sentence σ such that $T \vdash \phi \to \phi \prec \sigma$ but $T \nvdash \sigma \to \phi$. Consider the Π_n^0 sentence $\neg(\sigma \preccurlyeq \phi)$, which is a consequence of ϕ because $\phi \prec \sigma$ is a consequence of ϕ. We will show that ϕ is Π_n^0-non by showing that $T \nvdash \neg(\sigma \preccurlyeq \phi)$: From $\neg(\sigma \preccurlyeq \phi)$ we can infer in T that $\sigma \to \phi \prec \sigma$ and therefore $\sigma \to \phi$. But σ was chosen so that $T \nvdash \sigma \to \phi$. $\square$

EXAMPLES OCCURRING IN NATURE. For the moment we are principally concerned with Σ_1^0 and Π_1^0.

DEFINITIONS 1.3. T is Γ-*correct* iff whenever ϕ is a Γ-sentence and $T \vdash \phi$, then ϕ is true.

If T is r.e., θ is a *Rosser sentence* for T iff $PA \vdash \theta \leftrightarrow \mathrm{Thm}_T(\ulcorner \neg\theta \urcorner) \prec \mathrm{Thm}_T(\ulcorner \theta \urcorner)$. (The self-reference lemma guarantees the existence of Rosser sentences. A Rosser sentence is guaranteed to be independent of T.)

For the rest of this section T is r.e.

EXAMPLE 1.4. $\neg\mathrm{Con}(T)$ is Π_1^0-con over T.

This has been observed by several people, including the author. The first to do so was Kreisel, several years ago (see **[Kreisel]**). **[Macintyre-Simmons]** independently obtained an abstract version imposing only weak conditions on T and Thm_T.

EXAMPLE 1.5. Any Rosser sentence for T is Π_1^0-non over T.

PROOF. Let σ be Rosser for T, and let ϕ be the sentence $\mathrm{Thm}_T(\ulcorner \neg\sigma \urcorner) \prec \mathrm{Thm}_T(\ulcorner \sigma \urcorner)$, and θ be $\mathrm{Thm}_T(\ulcorner \sigma \urcorner)$. Then $T \vdash \phi \to \phi \prec \theta$, and so, to show that ϕ–hence σ–is Π_1^0-non it will suffice (by Lemma 1.2) to show that $T \nvdash \theta \to \phi$. But if $T \vdash \theta \to \phi$, then $T \vdash \mathrm{Thm}_T(\ulcorner \sigma \urcorner) \to \sigma$; and so by [Löb], $T \vdash \sigma$. That is impossible, because a Rosser sentence for T cannot be T provable.

EXAMPLE 1.6 (SMORYNSKI). $\mathrm{Con}(T)$ is Σ_1^0-con iff T is Σ_1^0-correct.

PROOF. The difficult part is to show: T Σ_1^0-incorrect implies $\mathrm{Con}(T)$ is Σ_1^0-non. (The easy part, besides being easy, is a consequence of Theorem 2.1; whose proof is also skipped.) Let ψ be any Σ_1^0 sentence and ϕ be such that T proves $\phi \leftrightarrow (\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner) \vee \psi) \preccurlyeq \mathrm{Thm}_T(\ulcorner \phi \urcorner)$. As usual,

(i) $T \nvdash \neg\phi$.

Here, for the first and last time, is a detailed elaboration of "as usual": Suppose $T \vdash \neg\phi$. Then there is a $k \in \omega$ such that T proves: $\mathbf{k}$ witnesses $\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner)$. Since T is consistent, $T \nvdash \phi$; and therefore (because Proof_T is a binumeration) for every $n \in \omega$, $T \vdash \mathbf{n}$ does not witness $\mathrm{Thm}_T(\ulcorner \phi \urcorner)$. Reason now in T: $\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner) \preccurlyeq \mathrm{Thm}_T(\ulcorner \phi \urcorner)$; so $(\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner) \vee \psi) \preccurlyeq \mathrm{Thm}_T(\ulcorner \phi \urcorner)$; so ϕ. That contradicts the consistency of T.

Notice that the proof of (i) depends only on the assumption that T is consistent, hence:

(ii) $T + \mathrm{Con}(T) \vdash \neg\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner)$.

The next claim is the crucial one:

(iii) If ψ is true, so is ϕ.

Let σ be the formula $(\mathrm{Thm}_T(\ulcorner \neg\phi \urcorner) \vee \psi) \preccurlyeq \mathrm{Thm}_T(\ulcorner \phi \urcorner)$. If ψ is true then σ is *decidable* (in N). What if σ is false? Then it must be the case that $N \vdash \neg\sigma$, so $T \vdash \neg\sigma$, so $T \vdash \neg\phi$–a contradiction. So σ, hence ϕ, must be true.

That reasoning can also be formalized in $T + \mathrm{Con}(T)$, for it used only (i) and the knowledge that the truth of either Σ_1^0 sentence θ or Δ guarantees the decidability of $\theta \preccurlyeq \Delta$; which fact can be formalized in PA. So:

(iv) $T + \mathrm{Con}(T) \vdash \psi \to \phi$. (In fact, $T + \mathrm{Con}(T) \vdash \psi \leftrightarrow \phi$.)

Similar but easier reasoning shows that:

(v) $(T \vdash \phi)$ iff ψ is true.

Now we are ready to go. Suppose that T is Σ_1^0-incorrect, and let ψ be a false Σ_1^0 sentence proved by T. Let ϕ be as above. By (v), $T \nvdash \phi$. But by (iv), $T + \mathrm{Con}(T) \vdash \psi \to \phi$ and therefore $T + \mathrm{Con}(T) \vdash \phi$. So $\mathrm{Con}(T)$ is Σ_1^0-non.

EXAMPLE 1.7 (EXPLOITING THE WEAKNESS OF PA). There is a Σ_1^0 sentence θ such that $PA + \theta$ is consistent and proves Π_1^0 sentences (truths) unprovable in ZFC; while $ZFC + \theta$ is inconsistent.

PROOF. Let σ be Rosser over ZFC. Then σ is independent of ZFC and a slight extension of Lemma 1.2 (proved by the same arguments) shows that $PA + \sigma$ proves Π_1^0 sentences not provable in ZFC. Let θ be such that PA proves: $\theta \leftrightarrow (\sigma \wedge \mathrm{Thm}_{PA}(\ulcorner \neg\theta \urcorner)) \preccurlyeq \mathrm{Thm}_{PA}(\ulcorner \theta \urcorner)$. Then θ is stronger than σ (so $PA + \theta$ proves Π_1^0 sentences not provable in ZFC) and, since $\theta \to \neg\mathrm{Con}(PA)$, θ is disprovable in ZFC. (That this example is slightly phoney can be seen by asking: Just how is it we know that $PA + \sigma$ is consistent? The Π_1^0 theory of $PA + \theta$ is contained in the Π_1^0 theory of $ZFC + \mathrm{Con}(ZFC)$.)

2. Existence theorems. Here is what the examples leave open: If T is r.e. is there an *independent* Γ sentence which is $\check{\Gamma}$-con over T? ($\neg\mathrm{Con}(T)$ need not be independent.) If T is Σ_1^0-correct is there any independent Π_1^0 sentence which is Σ_1^0-non over T? The answers are, respectively, Yes and No. First, the No.

THEOREM 2.1. *If T is r.e., the following are equivalent*:
(1) *T is Σ_1^0-correct.*
(2) *Every Π_1^0 sentence consistent with T is Σ_1^0-con over T.*
(3) *Every T-Δ_1^0 sentence is decided by T.*

PROOF. (1) $\Rightarrow$ (2) $\Rightarrow$ (3). Easy exercises. (The implications do not, in fact, depend on T being r.e.)

(3) $\Rightarrow$ (1). Prove the contrapositive. Let σ be a 'prenex' Σ_1^0 sentence which is false but T provable. Let θ be a Σ_1^0 sentence such that T proves: $\theta \leftrightarrow \mathrm{Thm}_T(\ulcorner \neg\theta \urcorner) \prec (\sigma \vee \mathrm{Thm}_T(\ulcorner \theta \urcorner))$. As usual, θ is independent of T. From the point of view of T, the truth of σ entails the decidability–i.e., the Δ_1^0-ness–of θ. More exactly: θ literally *is* Σ_1^0; and σ implies that $\neg\theta \leftrightarrow [(\sigma \vee \mathrm{Thm}_T(\ulcorner \theta \urcorner)) \preccurlyeq \mathrm{Thm}_T(\ulcorner \neg\theta \urcorner)]$. So θ is T-Δ_1^0. $\square$

NOTES TO THEOREM 2.1. The implication $\neg(1) \Rightarrow \neg(3)$ is ineffective: For every $e \in \omega$, let $T_e = PA \cup$ all sentences of LA with g.n.'s in the eth r.e. subset of ω. There is no recursive function f such that T_e consistent and Σ_1^0-incorrect implies $f(e)$ codes an independent T_e-Δ_1^0 pair (meaning a pair (σ, π) such that σ is Σ_1^0, π is Π_1^0, $T_e \vdash \sigma \leftrightarrow \pi$, and $T \nvdash \sigma$). *Proof.* Suppose there *is* such an f. Notice that if (σ, π) is an independent T_e-Δ_1^0 pair then $\pi \to \sigma$ is false and T_e provable. So from f we get a map $e \mapsto \sigma_e \in \Sigma_1^0$ such that: T_e consistent and Σ_1^0-incorrect $\Rightarrow \sigma_e$ is *false* and T_e provable. Then "T_e is consistent and Σ_1^0-correct" is, as a predicate of e, a Boolean combination of r.e. predicates, being equivalent to "T_e is consistent and ($T \nvdash \sigma_e$ or σ_e is true)". That is impossible, for $\{e | T_e$ is consistent and Σ_1^0-correct$\}$ can easily be seen to be a complete Π_2^0 subset of ω.

It will be convenient to formulate two lemmas axiomatizing the proof of the existence theorems–at the cost of some annoying notation.

NOTATION AND SETTING FOR LEMMA 2.2. T is r.e. and Proof_T is a binumeration. If ψ is a 'prenex' Σ_1^0 and θ a 'prenex' Σ_n^0 sentence, then

$$\psi * \theta \underset{\mathrm{df}}{=} \exists u(u \text{ witnesses } \psi) \wedge \forall x, \quad y < u\big(\mathrm{Proof}_T(y, x) \to \Sigma_n^0\text{-true}(``x \prec \ulcorner \theta \urcorner")\big).$$

Note that $\psi * \theta$ is T-Σ_n^0. For an explanation of Σ_n^0-True, see Fact 0.2; for "$x \prec \ulcorner \theta \urcorner$" see the end of §0.

LEMMA 2.2. *Suppose that $PA \vdash \theta \leftrightarrow \psi * \theta$. Then,*

(1) *ψ is true $\Rightarrow T \vdash \theta$,*

(2) *ψ is false $\Rightarrow \neg\theta$ is Σ^0_n-con over every subtheory of T.*

PROOF. (1) Suppose that ψ is true, and let k be the least witness to ψ. Then T (in fact, PA) proves:

$(+)$ $\psi * \theta \leftrightarrow \forall x,\ y < \mathbf{k}\ (\mathrm{Proof}_T(y, x) \to \Sigma^0_n\text{-True}(\text{“}x \prec \ulcorner\theta\urcorner\text{”}))$. For any theory containing N, a bounded quantifier is the same as a disjunction. I.e., it is provable that $\forall u < \mathbf{n}\phi(u) \leftrightarrow (\phi(\mathbf{0}) \wedge \cdots \wedge \phi(\mathbf{n}))$. Therefore if $\chi_1, \ldots, \chi_m$ is the list of Σ^0_n sentences T-provable by proofs with g.n.'s less than k, it follows from $(+)$ that T proves:

$(*)$ $\bigwedge\!\!\bigwedge \chi_i$; and also $\bigwedge\!\!\bigwedge \Sigma^0_n\text{-True}(\ulcorner\chi_i \prec \theta\urcorner) \to \psi * \theta$, hence

$(**)$ $\bigwedge\!\!\bigwedge(\chi_i \prec \theta) \to \theta$.

Now reason in T: Suppose $\neg\theta$. By $(**)$, $\neg \bigwedge\!\!\bigwedge (\chi_i \prec \theta)$. Together with $(*)$ this implies $\bigvee\!\!\bigvee(\theta \preccurlyeq \chi_i)$, hence θ. We have shown $\neg\theta \to \theta$; so θ.

(2) Now suppose that ψ is false and $T' \subseteq T$. It will suffice, by Lemma 1.2, to show that for any σ in Σ^0_n: If $T' \vdash \sigma$, then $T' \vdash \sigma \prec \theta$. So suppose $T' \vdash \sigma$. Then $T \vdash \sigma$. Let k be a g.n. of a T-proof of σ. Since ψ is false T' proves:

$\mathrm{Proof}_T(\mathbf{k}, \ulcorner\sigma\urcorner) \wedge \mathbf{k}, \ulcorner\sigma\urcorner <$ any possible witness to ψ. Therefore T' proves $\psi * \theta \to \Sigma_n\text{-True}(\ulcorner\sigma \prec \theta\urcorner)$, so proves $\theta \to \sigma \prec \theta$. From $\theta \to \sigma \prec \theta$ and σ it follows that $\sigma \prec \theta$. So $T' \vdash \sigma \prec \theta$. □

NOTATION AND SETTING FOR LEMMA 2.3. With T, Proof_T, ψ, and θ as before, define

$$\begin{aligned}\psi^+\theta \underset{\mathrm{df}}{=} \exists y, p(&(1)\text{ no witness for } \psi \text{ is } \leq y \text{ or } p;\\ &(2)\ p \text{ is the g.n. of a } \Pi^0_n \text{ sentence;}\\ &(3)\ \mathrm{Proof}_T(y, \text{“}\ulcorner\theta\urcorner \to p\text{”}); \text{ and}\\ &(4)\ \Sigma^0_n\text{-True}(\text{“}\neg p\text{”})).\end{aligned}$$

Again, $\psi^+\theta$ is Σ^0_n.

LEMMA 2.3. *Suppose that $PA \vdash \theta \leftrightarrow \psi^+\theta$. Then,*

(1) *ψ is true $\Rightarrow T \vdash \neg\theta$,*

(2) *ψ is false $\Rightarrow \theta$ is Π^0_n-con over every subtheory of T.*

PROOF. (1) Suppose that ψ is true and let k witness ψ. Then T proves: $\psi^+\theta \leftrightarrow \exists y,\ p < \mathbf{k}\ (p \text{ is } \Pi^0_n \wedge \mathrm{Proof}_T(y, \text{“}\ulcorner\theta\urcorner \to p\text{”}) \wedge \Sigma_n\text{-True}(\text{“}\neg p\text{”}))$. As before there is a list $\chi_1, \ldots, \chi_m$ of sentences (this time each is Π^0_n) such that T proves

$(*)$ $\bigwedge\!\!\bigwedge(\theta \to \chi_i)$; and $\psi^+\theta \leftrightarrow \bigvee\!\!\bigvee\ \Sigma_n\text{-True}(\ulcorner\neg\chi_i\urcorner)$, hence

$(**)$ $\theta \to \bigvee\!\!\bigvee\ \neg\chi_i$.

But $(*)$ and $(**)$ together imply $\neg\theta$.

(2) Suppose ψ is false. Let $T' \subseteq T$ and π be a Π^0_n sentence such that

$T' \vdash \theta \to \pi$. Then $T \vdash \theta \to \pi$, so let k be a g.n. of a T-proof of $\theta \to \pi$. T' proves: $\mathbf{k}$, $\ulcorner \pi \urcorner <$ any possible witness to ψ. Now reason in T'. Suppose $\neg \pi$. Then Σ_n^0-True($\ulcorner \neg \pi \urcorner$); therefore $\psi^+\theta$, and therefore θ. But we know that $\theta \to \pi$. So π. We have shown that $\neg \pi \to \pi$. Therefore π. □

Note to Lemma 2.3. The definition of $\psi^+\theta$ used above is due to Solovay and is considerably simpler and leads to a considerably simpler proof of the lemma than the author's original definition.

Theorem 2.4. *Let T be r.e. Then we can find, effectively from any r.e. index for T, an independent Γ sentence which is $\check{\Gamma}$-con over every subtheory of T.*

Proof. We will prove this for the case $\Gamma = \Sigma_n^0$. The proof for Π_n^0 is similar. By the self-reference lemma we can effectively find a Σ_n^0 sentence θ such that PA proves:

$$\theta \leftrightarrow \mathrm{Thm}_T(\ulcorner \neg \theta \urcorner) * \theta.$$

The sentence we want is $\neg \theta$. It will suffice to show that $T \nvdash \neg \theta$; for that guarantees that $\mathrm{Thm}_T(\ulcorner \neg \theta \urcorner)$ is false, hence by Lemma 2.2(2) that $\neg \theta$ is Σ_n^0-con over any suitable T' (and "nonprovable and Σ_n^0-con" certainly guarantees "independent"). So suppose instead that $T \vdash \neg \theta$. Then $\mathrm{Thm}_T(\ulcorner \neg \theta \urcorner)$ is true; so by Lemma 2.2(1) $T \vdash \theta$, contradicting the consistency of T.

If $\Gamma = \Pi_n^0$, find a $\Sigma_n^0 \phi$ such that $\phi \leftrightarrow \mathrm{Thm}_T(\ulcorner \phi \urcorner)^+\phi$. □

We will call ϕ essentially Γ over T if Γ is its simplest classification in T, and exactly Γ-con if Γ is the most complicated class for which ϕ is conservative. More exactly:

Definition 2.5. ϕ is *essentially* Γ over T iff $\phi \in \Gamma$ but $\phi \notin T - \check{\Gamma}$.

ϕ is *exactly* Γ*-con* over T iff ϕ is Γ-con but $\check{\Gamma}$-non over T.

The next theorem says just what you might expect–that we can obtain any not-obviously-absurd combination of essentially Γ and exactly Γ'-con.

Theorem 2.6. *Let T be r.e. and $\Gamma' \subseteq \Gamma$. Then, effectively from an r.e. index for T we can find a sentence which is essentially Γ and exactly $\check{\Gamma}'$-con over every subtheory of T.*

Proof. To keep the notation in hand take a special case: suppose $\Gamma' = \Pi_5^0$ and $\Gamma = \Sigma_8^0$. Let π be an independent Π_5^0 sentence which is Σ_5^0-con over every subtheory of T; and σ be an independent Σ_8^0 sentence which is Π_8^0-con over every subtheory of $T + \pi$. Let ϕ be $\pi \wedge \sigma$, and T' be any subtheory of T. Then ϕ is Π_5^0-non over T' because ϕ implies π. Further, ϕ is Σ_5^0-con, and therefore exactly Σ_5^0-con: For if θ is Σ_5^0 and $T + \pi + \sigma \vdash \theta$, then (because σ is Π_8^0-con over $T + \pi$) $T + \pi \vdash \theta$; and so (because π is Σ_5^0-con over T) $T \vdash \theta$. To complete the proof we need to show that ϕ is not equivalent in T' to any Π_8^0 sentence. So suppose that θ is Π_8^0 and that $T' \vdash \phi \leftrightarrow \theta$. Then infer successively: $T' \vdash (\pi \wedge \sigma) \leftrightarrow \theta$; $T' + \pi \vdash \theta$–because σ is Π^0_8-con over $T + \pi$; $T' + \pi \vdash \phi$; $T' + \pi \vdash \sigma$. The last statement contradicts the $T + \pi$ independence of σ. $\square$

NOTE TO THEOREMS 2.4 AND 2.6. That the sentences provided in Theorems 2.4 and 2.6 are conservative over so many theories is somewhat surprising, since "ϕ is Γ-con over T" is, at least by its looks, a truly global property of T. It is of course trivial to produce ϕ, T', and T with $T' \subseteq T$ and ϕ Γ-con over T but not over T'.

THEOREM 2.7 (SOLOVAY). *Let T be r.e. Then for any Γ there is a Γ sentence ϕ such that*:

ϕ is $\check{\Gamma}$-con over T and $\neg\phi$ is Γ-con over T.

PROOF. Notice that the "subtheory property" is not claimed for ϕ. The parts of the argument through which that claim cannot pass will be starred (and remarked upon at the end).

Let σ be a Σ^0_n sentence such that PA proves:

$$\sigma \leftrightarrow \exists y, p((1)\, y, p < \text{any } T \text{ proof of } \ulcorner\sigma\urcorner \wedge p \text{ is the g.n. of a } \Sigma^0_n \text{ sentence};$$
$$(2)\ \mathrm{Proof}_T(y, \text{“}\ulcorner\sigma\urcorner \to p\text{”});$$
$$(3)\ \Sigma^0_n\text{-True}(\text{“}\neg p\text{”});$$
$$(4)\ \forall s, t < y\big(\mathrm{Proof}_T(t, s) \to \Sigma^0_n\text{-True}(\text{“}s \prec \ulcorner\sigma\urcorner\text{”})\big)).$$

Denote the formula to the right of the $\leftrightarrow$ sign by 'Δ'. By ignoring (4) we see that, in the notation of Lemma 2.3, $PA \vdash \sigma \to \mathrm{Thm}_T(\ulcorner\sigma\urcorner)^+\sigma$. That fact suffices to carry out the argument of part (1) of Lemma 2.3. So, if $\mathrm{Thm}_T(\ulcorner\sigma\urcorner)$ is true then $T \vdash \neg\sigma$–a contradiction. We have

Fact 1. $T \nvdash \sigma$.

Fact 2. σ is Π^0_n-con over T.

PROOF. Suppose $T \vdash \sigma \to \pi$ and let k be the g.n. of a T-proof of $\sigma \to \pi$. Reason in T: Suppose $\neg\pi$. Then $\neg\sigma$. If we substitute $\mathbf{k}$ for "y" and $\ulcorner\pi\urcorner$ for "p" then clauses (1)–(3) of Δ are true. So the only way that σ can be false is for (4) to fail, i.e.:

(i) $\exists s, t < \mathbf{k}(\mathrm{Proof}_T(t, s) \wedge \neg\Sigma^0_n\text{-True}(\text{“}s \prec \ulcorner\sigma\urcorner\text{”}))$.

(Now step outside of T for a moment and produce $\chi_1, \ldots, \chi_m$–the Σ^0_n sentences provable in T by proofs with g.n.'s $< k$. Go back to T.) Therefore,

($*$) $\bigwedge\!\!\bigwedge \chi_i$; and, because we are assuming $\neg\pi$, $\bigvee\!\!\bigvee \neg(\chi_i \prec \sigma)$.

(That last remark is a consequence of the tail end of (i).) Since $\neg\sigma$, the only way to have $\neg(\chi_i \prec \sigma)$ is to have $\neg\chi_i$. We have shown: $\bigwedge\!\!\bigwedge \chi_i \wedge (\neg\pi \to \bigvee\!\!\bigvee \neg\chi_i)$. I.e., we have shown π.

Fact 3. $\neg\sigma$ is Σ^0_n-con over T.

PROOF. Suppose π is Π^0_n and $T \vdash \pi \to \sigma$. It suffices to show that $T \vdash \neg\pi$. Let χ be Σ^0_n such that $\chi \leftrightarrow (\neg\pi \vee \sigma)$ *and*

(ii) $T \vdash (\neg\pi \vee \sigma) \preccurlyeq \chi$;
this last is easily arranged. Let k be a g.n. of a T-proof of χ. We know that σ is Π_n^0-con over T and hence that for any $\psi \in \Pi_n^0$, *if* $T \vdash \sigma \to \psi$ then $T \vdash \psi$. Therefore

(∗∗) $\forall y, p \leqslant \mathbf{k}(p \text{ is } \Pi_n^0 \wedge \mathrm{Proof}_T(y, \text{“}\ulcorner\sigma\urcorner \to p\text{”}) \to \Pi_n^0\text{-True}(p))$

is equivalent to a finite conjunction of T provable statements, so is itself T provable. What (∗∗) entails, and this entailment is formalizable in T, is that *if* σ is to be true, any candidate to play the role of the “y” in Δ must be $> \mathbf{k}$. Hence: *if* σ is true, $\ulcorner\chi\urcorner$ and $\mathbf{k}$ must fall within the range of the bounded quantifier in clause (4) of Δ, and therefore $\sigma \to \Sigma_n^0\text{-True}(\ulcorner\chi \prec \sigma\urcorner)$. Formalizing in T gives $T \vdash \sigma \to \chi \prec \sigma$. Now reason in T: We know that $\sigma \to \chi \prec \sigma$ and that χ. Therefore $\chi \prec \sigma$. But that, together with $(\neg\pi \vee \sigma) \preccurlyeq \chi$–see (ii)–implies $\neg\pi$. We have proven $\neg\pi$. □

Notes to Theorem 2.7. In the proof of Fact 2, the subtheory property escapes at (∗). Suppose, at that point, that we are trying to reason in $T' \subseteq T$. We know that $\neg\pi \to \bigvee\!\!\!\bigvee \neg(\chi_i \prec \sigma)$, but not that $\bigwedge\!\!\!\bigwedge \chi_i$–for the χ's are consequences of T. There is no way, in T', to bring those sentences into collision. One trouble spot in the proof of Fact 3 is (∗∗). We need all of T to prove (∗∗)–and would still need all of T even if, by magic, σ were Π_n^0-con over T'.

Definition 2.8. ϕ is *essentially* Δ_n^0 over T iff ϕ is T-Δ_n^0 but not T-Σ_{n-1}^0 or T-Π_{n-1}^0.

ϕ is *exactly* Δ_n^0-con over T iff ϕ is Δ_n^0-con, but Σ_n^0-non and Π_n^0-non, over T.

Corollary 2.9. *Let T be r.e. and $n > m \geqslant 1$. Then there is a sentence which is essentially Δ_n^0 and exactly Δ_m^0-con over T.*

Proof. It will suffice to prove the theorem for the case $n = m + 1$, the general result following by iterating in the manner of Theorem 2.6. Say $m = 4$, $n = 5$. Let σ be a Σ_4^0 sentence which is Π_4^0-con over T and for which $\neg\sigma$ is Σ_4^0-con over T. Let π be a Π_4^0 sentence which is Σ_4^0-con and independent over $T + \sigma$. We want $\phi =_{\mathrm{df}} \pi \wedge \sigma$. Clearly ϕ is Δ_5^0, Π_4^0-non, and Σ_4^0-non over T. To see that ϕ is Δ_4^0-con: Suppose P is Π_4^0, S is Σ_4^0, $T \vdash P \leftrightarrow S$, and $T + \phi \vdash P(\bigwedge S)$. Then $T + \sigma \vdash \pi \leftrightarrow S$ and so by choice of π, $T + \sigma \vdash S$. So $T + \sigma \vdash P$, and by choice of σ, $T \vdash P$. To see that ϕ is neither T-Σ_4^0 nor T-Π_4^0: Suppose S is Σ_4^0 and $T \vdash \phi \leftrightarrow S$. Then $T + \sigma \vdash \pi \leftrightarrow S$ contradicting the essential Π_4^0-ness of π over $T + \sigma$. Suppose P is Π_4^0 and $T \vdash \phi \leftrightarrow P$. Then $T \vdash \neg\sigma \to \neg P$, and by choice of σ, $T \vdash \neg P$. So $T \vdash \neg\phi$, which is impossible. □

Notes to Corollary 2.9. Only the last step required use of Theorem 2.7. That is where the subtheory property is lost. Mixing Theorem 2.4 and Corollary 2.9 does *not* yield, e.g., an essentially Σ_4^0 sentence which is exactly

Δ_4^0-con. Whether all reasonable combinations of Π, Σ, and Δ can be obtained I do not know.

Our last existence theorem simply codifies the effectiveness available in the proofs of Lemmas 2.2 and 2.3. It helps partially to classify $\{\phi|\phi$ is Γ-con over $T\}$ for r.e. T by showing that it cannot be an r.e. set of formulas.

LEMMA 2.10. *Let X be an r.e. subset of ω and T an r.e. theory. Then, for any Γ, there is a recursive map $k \mapsto \theta_k \in \Gamma$, such that:*

$k \in X \Rightarrow T \vdash \neg\theta_k$.

$k \notin X \Rightarrow \theta_k$ *is independent and $\check{\Gamma}$-con over every subtheory of T.*

PROOF. Consider the case $\Gamma = \Pi_n^0$. Let $\psi(x)$ be a Σ_1^0 predicate which, in the real world, defines X. For each k produce ϕ_k such that (in the notation of Lemma 2.2), PA proves:

$$\phi_k \leftrightarrow (\psi(\mathbf{k}) \vee \mathrm{Thm}_T(\ulcorner \neg\phi_k \urcorner)) * \phi_k.$$

We are going to apply Lemma 2.2 repeatedly. Since the truth of either $\mathrm{Thm}_T(\ulcorner\neg\phi_k\urcorner)$ or $\psi(\mathbf{k})$ implies that $T \vdash \phi_k$, we have for any k:

(i) $T \nvdash \phi_k$,

(ii) $k \in X \Rightarrow \psi(\mathbf{k})$ is true $\Rightarrow T \vdash \phi_k$;

and by the other half of Lemma 2.2,

(iii) $k \notin X \Rightarrow \psi(\mathbf{k}) \vee \mathrm{Thm}_T(\ulcorner\neg\phi_k\urcorner)$ is false $\Rightarrow \neg\phi_k$ is Σ_n^0-con and independent over any subtheory of T.

Let θ_k be (a Π_n^0 sentence equivalent to) $\neg\phi_k$. The other case is similar. $\square$

By iterating the steps in the lemma as in the proof of Theorem 2.6, we get:

THEOREM 2.11. *Let T be r.e. and $\Gamma' \subseteq \Gamma$. Then every Π_1^0 subset of ω is reducible to either of the sets:*

$\{\phi|\phi$ *is essentially* Γ *and exactly* $\check{\Gamma}'$*-con over* $T\}$.

$\{\phi|\phi$ *is essentially* Γ *and exactly* $\check{\Gamma}'$*-con over every subtheory of* $T\}$.

It seems worth noting one more restatement of the existence lemmas which is cute and sometimes useful. Say that $\psi(x)$ is Γ-disjunctive (over T) iff for every sentence $\chi \in \Gamma$ and any $n \in \omega$, $T \vdash \psi(\mathbf{n}) \vee \chi$ implies $T \vdash \psi(\mathbf{n})$ or $T \vdash \chi$.

THEOREM 2.12. *Let X be an r.e. subset of ω and T an r.e. theory. Then there is a $\psi(x) \in \Gamma$ such that ψ numerates X in T and ψ is Γ-disjunctive.*

PROOF. To be consistent with the notation of 2.10 we will make $\psi \in \check{\Gamma}$ and $\check{\Gamma}$-disjunctive. Let the map $k \mapsto \theta_k \in \Gamma$ be as in 2.10, and let $x \mapsto \dot{\theta}_x$ be the formalization of that map. Set

$$\psi(x) \underset{\mathrm{df}}{=} \check{\Gamma}\text{-True}(``\neg\dot{\theta}_x").$$

Since $T \vdash \psi(\mathbf{n}) \leftrightarrow \neg\theta_n$ for each $n \in \omega$, ψ numerates X. Suppose now that $\chi \in \Gamma$ and $T \vdash \psi(\mathbf{n}) \vee \chi$, but $T \nvdash \psi(\mathbf{n})$. Then $n \notin X$ and therefore θ_n is $\check{\Gamma}$-con over T; and, since $T \vdash \neg\psi(\mathbf{n}) \rightarrow \chi$, $T \vdash \theta_n \rightarrow \chi$. So $T \vdash \chi$. $\square$

3. Interpretability.

DEFINITIONS 3.1. ϕ is *interpretable* in T iff $T + \phi$ is interpretable in T in the sense of [**Shoenfield**, pp. 61–62].

T is *essentially reflexive* iff for every sentence ϕ in the language of T, $T \vdash \phi \to \mathrm{Con}(\{\phi\})$.

Interpretability is discussed in [**Hajek, 1**], [**Hajek, 2**], [**Hajkova-Hajek**] and [**Solovay**].

Whether T is essentially reflexive depends solely on the nonarithmetical part of T. All extensions of PA in the same language as PA are essentially reflexive. ZF (and any extension in the same language) is essentially reflexive. GB is not (for an essentially reflexive theory cannot be finitely axiomatized). (See [**Montague**].)

The connections between interpretability and Π_1^0-con are easily stated:

THEOREM 3.2 (HÁJEK, LARGELY). *Let T be r.e. and essentially reflexive. Then, ϕ is interpretable in T iff ϕ is Π_1^0-con over T.*

PROOF. Say that ϕ is strongly consistent with T if for every finite $F \subseteq T$, $T \vdash \mathrm{Con}(F + \phi)$. [**Hájek,1**] shows that for r.e. essentially reflexive T, ϕ is strongly consistent with T iff ϕ is interpretable in T. Thus it will more than suffice to show:

Claim. If T is essentially reflexive (not necessarily r.e.), then ϕ is Π_1^0-con over T iff ϕ is strongly consistent with T.

PROOF. Suppose first that ϕ is Π_1^0-con over T and let $F \subseteq T$ be finite. Since T is essentially reflexive, $T \vdash \bigwedge\!\!\bigwedge F \wedge \phi \to \mathrm{Con}(\bigwedge\!\!\bigwedge F \wedge \phi)$–i.e., $T \vdash \phi \to \mathrm{Con}(F + \phi)$. Since ϕ is Π_1^0-con over T, $T \vdash \mathrm{Con}(F + \phi)$.

Suppose, conversely, that for every finite $F \subseteq T$, T Con $\vdash (F + \phi)$. Suppose that π is Π_1^0 and $T + \phi \vdash \pi$. Produce a finite $F \subseteq T$ such that $F + \phi \vdash \pi$. We may as well assume $N \subseteq F$. Then $T \vdash \mathrm{Thm}_{F+\phi}(\ulcorner \pi \urcorner)$. Now reason in T: If $\neg\pi$, then $\mathrm{Thm}_N(\ulcorner \neg\pi \urcorner)$, so $\mathrm{Thm}_{F+\phi}(\ulcorner \pi \wedge \neg\pi \urcorner)$; contradicting $\mathrm{Con}(F + \phi)$. Since $\neg\pi$ implies a contradiction, π. □

NOTES TO THEOREM 3.2. [**Hájek, 2**] shows that if T contains induction for *all* formulas of T (no assumption about reflexiveness or recursive enumerability) then every sentence interpretable in T is Π_1^0-conservative over T. A slight elaboration of that argument is used in 6.5.

The hypothesis "T is essentially reflexive" cannot merely be omitted from the hypothesis of 3.2. Consider GB, which is not essentially reflexive (nor does it contain induction for all formulas). Then I, the set of sentences interpretable in GB, is r.e. (because GB is finitely axiomatizable), while C, the set of sentences Π_1^0-con over GB, is not (by 2.11). So $I \neq C$. More is known: Solovay has shown that $I \setminus C \neq 0$; and Hájek has observed that $C \setminus I \neq 0$ follows easily from Lemma 2.3.

I owe thanks to Hájek and Švejdar for clearing up my confusions about these things.

An interpretation of ϕ in T provides a proof of the consistency of $T + \phi$ relative to T. Whether this consistency proof is "elementary" depends on the presentation of that interpretation. Suppose our measure of elementary is this: $\mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$ can be proven in PA. Our reductive attitude about the formulation of $\mathrm{Con}(T)$ will come in handy (and may make the next result appear to say more than it does say). If, as in **[Feferman]**, we let Con_ψ be the consistency statement naturally based on ψ (and let $\psi'(x)$ abbreviate $\psi(x) \vee x = \ulcorner\phi\urcorner$), then the meaning–and presumably the provability–of $\mathrm{Con}_\psi \to \mathrm{Con}_{\psi'}$ depends strongly on ψ. By choosing particular representations of $\mathrm{Con}(T)$ and $\mathrm{Con}(T + \phi)$–which are related in the natural way– such difficulties are dodged. If ϕ is a Rosser sentence for T, then there is an elementary proof of the relative consistency of $T + \phi$, even though $T + \phi$ cannot be interpreted in T. The next theorem says that even if ϕ is Σ^0_1, the existence of an interpretation of $T + \phi$ in T need not guarantee the existence of an elementary relative consistency proof for ϕ. (The point, again, is that ϕ is so simple.)

THEOREM 3.3. *Let T be r.e. and essentially reflexive. Then there is a Σ^0_1 sentence ϕ such that ϕ is interpretable in T but $PA \nvdash \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$.*

PROOF. Choose ϕ so that, in the notation of Lemma 2.3,

$$\phi \leftrightarrow \mathrm{Thm}_{PA}(\mathrm{Con}(T) \to \mathrm{Con}(T + \phi))^+ \,\phi.$$

We will apply Lemma 2.3 repeatedly. If $PA \vdash \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$ then by part (1) of the lemma, $T \vdash \neg\phi$. So $PA \vdash \neg\mathrm{Con}(T + \phi)$ and therefore $PA \vdash \neg\mathrm{Con}(T)$–which is impossible because PA is Σ^0_1-correct. Since $PA \nvdash \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$, part (2) of the lemma guarantees that ϕ is Π^0_1-con and therefore interpretable in T. $\square$

Here is how Π^0_n-con can be understood in terms of interpretability. Say that an interpretation I of T' in $T (\subseteq T')$ is *provably* Γ*-faithful* if for every sentence χ in Γ, $T \vdash \chi_I \to \chi$. (Here χ_I is the interpretation of χ. See **[Shoenfield]**.)

THEOREM 3.4. *Let T be r.e. and essentially reflexive and ϕ any sentence in the language of T. Then, ϕ is Π^0_n-con over T iff there is a provably Π^0_n-faithful interpretation of $T + \phi$ in T.*

PROOF. The implication from right to left is immediate. Suppose conversely, that ϕ is Π^0_n-con over T. For each finite $F \subseteq T$ consider the sentence ϕ_F: $\forall x(\Sigma^0_n - \mathrm{True}(x) \to$ the theory whose axioms are $\ulcorner F \urcorner$, $\ulcorner\phi\urcorner$, and x is consistent). Then $T + \phi \vdash \phi_F$; and since ϕ_F is Π^0_n, $T \vdash \phi_F$. Using that fact and the tricks in **[Feferman]** we can find a formula $\psi_1(x)$ binumerating T in T

such that $T \vdash \mathrm{Con}\ \psi_2$, where $\psi_2(x)$ is $(\Sigma_n^0\text{-True}(x) \vee \psi_1(x) \vee x = \ulcorner \phi \urcorner)$. From the point of view of T, ψ_2 is the theory consisting of T (or at least that fragment of T described by ψ_1, which of course is a numeration of T, but is not T *provably* equal to T), ϕ, and all the true Σ_n^0 sentences. By 5.9 of **[Feferman]**, slightly modified, there is a formula $\chi(x)$ such that T proves: "$\{x|\chi(x)\}$ is a complete Henkin extension of $\{x|\psi_2(x)\}$." (A Henkin theory is one in which every provable existential statement is witnessed by some constant.) In particular, if σ is Σ_n^0 then, since $T \vdash \sigma \to \Sigma_n^0\text{-True}(\ulcorner \sigma \urcorner)$, we have $T \vdash \sigma \to \chi(\ulcorner \sigma \urcorner)$. By imitating the usual construction of a model from a complete Henkin theory we can extract from χ an interpretation I of $T + \phi$ in T such that for any sentence θ of T, $T \vdash \chi(\ulcorner \theta \urcorner) \leftrightarrow \theta_I$. Now suppose that π is Π_n^0 and reason in T: $\neg \pi \to \chi(\ulcorner \neg \pi \urcorner), \to (\neg \pi)_I, \to \neg(\pi_I)$. So $\pi_I \to \pi$. □

The argument in effect constructs inside T a model of $T + \phi$ which extends the "standard" one (standard from the point of view of T) and is elementary for Σ_{n-1}^0 formulas. An attempt to produce some corresponding theorem for Σ_n^0 would seem to require us to build not extensions but submodels; yet T thinks that its number structure is the minimal one.

4. Generalizations. *Extending the language of arithmetic.* If L extends LA define PA_L to be PA + induction for all formulas in L. Define $\Sigma_n^0(L)$ and $\Pi_n^0(L)$ in the obvious way. The self-reference lemma holds for PA_L; and if $L' \subseteq L$ has only *finitely many* symbols other than constant symbols (in particular, if L is finite) we can define $\Sigma_n^0(L')$ truth by a $\Sigma_n^0(L')$ formula. The only important difference between L and LA is this: T might not decide all the $\Delta_0^0(L)$ sentences. If one wants to check the extendability of an argument from LA to L that is the first point to check. In particular, if L is finite the existence theorems 2.4, 2.6, 2.7, 2.9, 2.10 all hold with "$\Gamma(L)$" replacing "Γ" everywhere. (Note: the ψ of Lemmas 2.2 and 2.3 must be Σ_1^0, not $\Sigma_1^0(L)$–indeed, the statement "ψ is true" is probably nonsensical otherwise.) It is easy to construct an r.e. theory T in an infinite language L for which *every* formula is T-$\Delta_0^0(L)$: let $\langle S_n | n \in \omega \rangle$ be a sequence of new symbols such that each S_n defines truth for the formulas in $\Sigma_n^0(PA \cup \{S_0, \ldots, S_{n-1}\})$. To extend the theorems mentioning reflexiveness it is necessary to formulate a stronger notion of reflexive. Roughly, T is essentially-L-reflexive iff for every ϕ, $T \vdash \phi \to \mathrm{Con}(\phi$ + all the true $\Delta_0^0(L)$ sentences). If L is finite all extensions of PA_L having language L are essentially L-reflexive. All the results of §3 go through if "essentially reflexive" is everywhere strengthened to "essentially-L-reflexive."

Set theory. Use Σ_n, Π_n for the classes in the Levy hierarchy. The existence theorems will follow once we have a workable definition of $\phi \preccurlyeq \chi$. Define $x \leqslant {}^*y$ to mean: $x \in$ the transitive closure of $\{y\}$; and $x < {}^*y$ iff $x \leqslant {}^*y$ and $x \neq y$. Define $\phi \preccurlyeq \chi$ as before, using $\leqslant {}^*$ in place of $\leqslant$. If ϕ and χ are 'prenex' Σ_n, so is $\phi \preccurlyeq \chi$. Our only worry: $\preccurlyeq{}^*$ is a partial, but not total, order. One of the arrows in Lemma 1.2 fails, but it is not one that hurts. The proofs of Lemmas 2.2 and 2.3 go through if we simply replace Σ_n^0-True by Σ_n-True. So Theorems 2.6, 2.7, 2.10 generalize. We could beef up the language and talk about $\Sigma_n(L)$–there are no problems.

The theorems on interpretations go through as before, but there is a slight difference in spirit. The models corresponding to the interpretations are not even *extensions* of the originals, let alone true outer models (i.e., *end extensions*). To "correct" this we could try reworking some of the theorems to apply to $L_{\infty\omega}$. Here is how one looks–one which will be useful in §6. (Elementary knowledge of $L_{\infty\omega}$ is assumed. What we need is contained in the first few chapters of **[Barwise, 1]** and **[Keisler]**.) Our language contains $\in$ and, for each set x, a constant $\mathbf{x}$ which will turn out to be a name for x. Let λ_x be the usual infinitary sentence saying that $\mathbf{x}$ denotes x. (A transitive set satisfies λ_x iff it contains x.) Define $\mathrm{Proof}_\infty(y;\, z, x)$ to mean: using the usual axioms and rules for $L_{\infty\omega}$, x is a deduction of z from assumptions $y \cup \{\lambda_x | x \in V\}$. Proof_∞ is ZF-Δ_1. We obtain from Proof_∞ the Σ_1 formulas $\mathrm{Thm}_\infty(y;\, z)$–"$z$ is a theorem of $y \cup \{\lambda_x | x \in V\}$"; and $\mathrm{Con}_\infty(y)$–"$y \cup \{\lambda_x | x \in V\}$ is consistent." The proof of the claim in 3.2 is a roadmap for the proof of

THEOREM 4.1. *A sentence ϕ is Π_1-con over ZF iff for every finite $F \subseteq ZF$, $ZF \vdash \mathrm{Con}_\infty(F + \phi)$.*

5. Questions. *Existence theorems.* (1) Can Theorem 2.4 be made uniform? I.e., if $\langle T_i | i \in \omega\rangle$ is an r.e. sequence of r.e. theories is there a Γ sentence which is independent and $\check{\Gamma}$-con over each T_i? The question is open even for sequences of length 2.

(2) Can Theorem 2.6 be extended to allow all not-obviously-absurd combinations of Π, Σ, and Δ?

Classifying Γ-con. Let T be r.e. Here is what is known:

(a) $\{\phi | \phi$ is Γ-con over $T\}$ is, by inspection, a Π_2^0 subset of ω; and, by 2.11, cannot be r.e.

(b) Solovay has shown that if T is essentially reflexive, then $\{\phi | \phi$ is Π_1^0-con over $T\}$ is a complete Π_2^0 subset of ω. So:

(3) Is $\{\phi | \phi$ is Γ-con over $T\}$ a complete Π_2^0 subset of ω?[1]

6. Partially conservative extensions: Semantics. This section contains model theoretic characterizations of partially conservative sentences–characterizations which have been forshadowed by the syntactical arguments of §3 which mimicked model constructions. Some of the proofs can be carried out very smoothly using the machinery of admissible covers **[Barwise, 1]** but others

[1]Hajek has shown that $\{\phi | \phi$ is Γ-con over $PA\}$ is a Π_2^0 complete subset of ω.

cannot–not, at least, in any obvious way. The problem seems to be that compactness arguments, useful as they are for extending models, are not so good at producing submodels. Instead we will use and/or modify a series of theorems and definitions from **[Friedman]**, few of which will be stated in their most general form.

The results about models of arithmetic can be read independently of those about set theory. We will assume the reader knows some basic facts about models of set theory–in particular, that he knows the definition of "standard part". (See e.g., **[Friedman]** or **[Barwise, 1]**.) We will always identify the standard part of a model of set theory with the transitive set to which it is isomorphic.

From now on say that T is a number theory if for some not necessarily finite language L, T is an extension of PA_L with language L (i.e., the only sort is the number sort). T is a set theory if for some language L, T is an extension of ZF_L with language L. (ZF_L is ZF + comprehension, collection, and foundations for all formulas of L.) We will perpetrate the small abuse of crediting every model of set theory with being a model of PA.

Definition 6.1. X is c-closed (for "completion closed") iff X is a collection of subsets of ω such that: (i) X is closed under Turing join and "recursive in"; (ii) whenever $y \in X$ codes an infinite binary tree, some path through that tree is in X.

The terminology "completion closed" is suggested by the fact that if X is c-closed and $T \in X$ is a consistent theory, then some complete extension of T is an element of X.

Definition 6.2. If $\mathfrak{M}$ is an ω-nonstandard model of arithmetic then $x \in \mathrm{ss}(\mathfrak{M})$ if $x \subseteq \omega$ and for some $\phi(x, \vec{y})$ and some $\vec{b} \in |\mathfrak{M}|$, $x = \{n\varepsilon\omega | \mathfrak{M} \models \phi(\mathbf{n}, \vec{\mathbf{b}})\}$. The sets in $\mathrm{ss}(\mathfrak{M})$–read "the standard system of $\mathfrak{M}$"–are all in fact initial segments of Δ_0^0-definable classes. Define $u\varepsilon v$ to mean that the uth prime divides v. Then standard sets all have the form $\{n\varepsilon\omega | \mathfrak{M} \models \mathbf{n} \in \mathbf{a}\}$ for some suitable (nonstandard) $a \in |\mathfrak{M}|$. The corresponding syntactical notion is:

$$\mathrm{bi}(T) = \{x \subseteq \omega | x \text{ is binumerated in the theory } T\}.$$

Theorem 6.3 (**[Friedman]**, modified). (i) *$\forall\mathfrak{M}(\mathfrak{M}$ models $PA \Rightarrow \mathrm{ss}(\mathfrak{M})$ is c-closed).*

(ii) *If X is countable and c-closed, T a consistent set or number theory, and $\mathrm{bi}(T) \subseteq X$, then $\exists\mathfrak{M}(\mathfrak{M} \models T$ and $\mathrm{ss}(\mathfrak{M}) = X)$.*

These results are more or less contained in Theorems 2.4 and 2.5 of **[Friedman]**–see especially the discussion of Z on pp. 541–542. Part (ii) above does not quite follow from Friedman's 2.5 but is proved by the same argument. Our primary interest in 6.3 lies in the following embedding theorem– which shows that an obviously necessary condition for "$\mathfrak{N}$ is an end extension of $\mathfrak{M}$" is also sufficient.

THEOREM 6.4 ([**Friedman**], MODIFIED). *Let L be a countable language and T a number theory* (*resp. set theory*) *with language L. Suppose that $\mathfrak{M}$ and $\mathfrak{N}$ are countable ω-nonstandard models of T. Then, $\mathfrak{N}$ is* (*isomorphic to*) *an end extension of $\mathfrak{M}$ if and only if* $\mathrm{ss}(\mathfrak{M}) = \mathrm{ss}(\mathfrak{N})$ *and every $\Sigma_1^0(L)$* (*resp.*, $\Sigma_1(L)$) *sentence true in $\mathfrak{M}$ is true in $\mathfrak{N}$.*

Theorem 4.2 of [**Friedman**] is not about end extensions of models for set theory, but rather what are sometimes called "rank extensions"–all the new elements in the larger model having ordinal rank greater than any ordinal of the smaller model. However, a proof of 6.4 (above) is embedded in Friedman's proof.

If $\mathfrak{M}$ models PA, say that $\mathfrak{M}T$-models T if $\mathfrak{M} \models T$ and $T \in \mathrm{ss}(\mathfrak{M})$. This stands in for the property that T be numerable in T. It is trivial to check that if $\mathfrak{M} \models T$ there is some elementary extension $\mathfrak{N}$ of $\mathfrak{M}$ which T-models T.

THEOREM 6.5. *Let T be a number theory in a finite language L and ϕ be any sentence of L. Then,*

(i) *ϕ is $\Pi_1^0(L)$-con over $T \Leftrightarrow$ every T-model of T can be end extended to a model of $T + \phi$.*

(ii) *ϕ is $\Sigma_1^0(L)$-con over $T \Leftrightarrow$ every countable T-model of T is an end extension of a model of $T + \phi$.*

Matijasevic's theorem, hereinafter referred to as [M], says that every Σ_1^0 formula is equivalent in PA to a purely existential formula. (This does not generalize to $\Sigma_1^0(L)$!) From [M], 6.5, and the Lowenheim-Skolem theorem it easily follows that

THEOREM 6.6 [M]. *If $L = LA$, then 6.5 remains true if "end extension" is replaced by "extension"; and even if "countable" is dropped from the statement of* (ii).

PROOF OF 6.5. (i) The implication from right to left is trivial: Suppose that every T-model of T can be end extended to model $T + \phi$, and that π is $\Pi_1^0(L)$ and $T + \phi \vdash \pi$. We will show that every model of T models ϕ. Let $\mathfrak{M} \models T$. There exists a T-model of T, $\mathfrak{M}'$, which is an elementary extension of $\mathfrak{M}$. By assumption we can produce an end extension $\mathfrak{N}$ of $\mathfrak{M}'$ modelling $T + \phi$. Then $\mathfrak{N} \models \pi$; and since $\Pi_1^0(L)$ sentences persist downward to initial segments, $\mathfrak{M}' \models \pi$. Therefore $\mathfrak{M} \models \pi$.

Suppose conversely that ϕ is $\Pi_1^0(L)$-con over T. Examine the proof of 3.4. The recursive enumerability of T was used only to guarantee the existence of a numeration of T in T. So expand T to T' by adding a one place predicate

$\mathfrak{T}$, induction for all new formulas, and $\{\mathfrak{T}(\ulcorner\chi\urcorner)|\chi \in T\} \cup \{\neg\mathfrak{T}(\ulcorner\chi\urcorner)|\chi \notin T\}$. T' is a conservative extension of T, so ϕ is $\Pi_1^0(L)$-con over T' and the proof of 3.4 proceeds as before. It provides an interpretation I with the following properties: (a) There is a T' definable function $u \mapsto u_I$ with meaning "u_I is the interpretation of the uth numeral (hence u is the denotation of u_I)". (b) T' proves that function embeds $<$ as an initial segment of $<_I$ and, for any $R \in L$, that $R(u, \ldots, v) \leftrightarrow R_I(u_I, \ldots, v_I)$. If $\mathfrak{M}$ T-models T we can expand $\mathfrak{M}$ to a model $\mathfrak{M}'$ of T' and then pull out the interpretation I to a structure $\mathfrak{N}$ with domain $= \{b|\mathfrak{M}' \models \mathbf{b}$ is in the universe of $I\}$ and relations of form $R^{\mathfrak{N}} = \{(a, \ldots, b)|\mathfrak{M}' \models R_I(\mathbf{a}, \ldots, \mathbf{b})\}$. The function $f = \{(a, b)|\mathfrak{M}' \models \mathbf{b} = (\mathbf{a})_I\}$ embeds $\mathfrak{M}$ as an initial segment of $\mathfrak{N}$. And since T' proves χ_I for each $\chi \in T + \phi$, $\mathfrak{N} \models T + \phi$.

(ii) The proof of the right to left implication is dual to the proof in (i). So suppose that ϕ is $\Sigma_1^0(L)$-con over T and that $\mathfrak{M}$ is countable and $\mathfrak{M}$ T-models T. Let $S = \{\pi \in \Pi_1^0(L)|\mathfrak{M} \models \pi\}$. Because ϕ is $\Sigma_1^0(L)$-con over T, $T' =_{\mathrm{df}} T \cup S \cup \{\phi\}$ is consistent. Since T admits a $\Pi_1^0(L)$ truth definition $S \in \mathrm{ss}(\mathfrak{M})$–and by the closure properties of $\mathrm{ss}(\mathfrak{M})$, $T' \in \mathrm{ss}(\mathfrak{M})$ and $\mathrm{bi}(T) \subseteq \mathrm{ss}(\mathfrak{M})$. Apply 6.3 to get an $\mathfrak{N} \models T'$ with $\mathrm{ss}(\mathfrak{N}) = \mathrm{ss}(\mathfrak{M})$. Because $\mathfrak{N} \models S$, every $\Sigma_1^0(L)$ sentence true in $\mathfrak{N}$ is true in $\mathfrak{M}$. So $\mathfrak{N}$ is a model of $T + \phi$ which (by 6.4) can be embedded as an initial segment of $\mathfrak{M}$. $\square$

NOTES TO 6.5 AND 6.6. (1) The method of 6.5(i) for constructing end extensions seems to be well known.

(2) For any theory S in any language it is trivial to see that a sentence ϕ is (universal sentences)-con over S iff every model of S can be extended to model $S + \phi$. Accordingly, the characterization of Π_1^0-con provided in 6.6 is an immediate consequence of [M]. The corresponding fact about Σ_1^0 does *not* follow from [M] by "trivial model theory".

(3) Model theoretic equivalents of $\Pi_n^0(L)$-con and $\Sigma_n^0(L)$-con are immediate: every T-model of T can be end extended to (is an end extension of) a model of $T + \phi$ which is elementary for $\Sigma_{n-1}^0(L)$ sentences.

The proofs of corresponding theorems for set theories are identical with (or dual to) the proof of 6.5(ii). (Notice that every model of set theory has an elementary extension which is a non-ω-model.) They are stated separately in order to emphasize: (i) that we consider only ω-nonstandard models of ZF; (ii) that 6.5(i) alone has no cardinality restriction.

THEOREM 6.7. *Let T be a set theory in a countable language L and ϕ any sentence of L. Then,*

(i) *ϕ is $\Pi_1(L)$-con over $T \Leftrightarrow$ every countable ω-nonstandard T-model of T can be end extended to a model of $T + \phi$.*

(ii) *ϕ is $\Sigma_1(L)$-con over $T \Leftrightarrow$ every countable ω-nonstandard T-model of T is an end extension of a model of $T + \phi$.*

The necessity of restricting attention ot ω-nonstandard models of set theory is shown by the following example.

THEOREM 6.8. *There is a Σ_1 sentence ϕ such that ϕ is Π_1-con over ZF but ZF proves that the theory $ZF + \phi$ has no ω-model.*

PROOF. Let ϕ be a Σ_1 sentence such that ZF proves $\phi \leftrightarrow \exists$ finite $x \subseteq ZF(\neg \mathrm{Con}_\infty(x \cup \{\ulcorner \phi \urcorner\}))$. §4 contains the definition of Con_∞. We will show first that $ZF + \phi$ cannot have an ω-model.

Let LST be the (finitary) language of set theory. Consider the following metamathematical fact (proof deferred):

(i) For any sentence ϕ of LST, $ZF \vdash \phi \to \mathrm{Con}_\infty(\{\phi\})$.

The proof of (i) can be arithmetized, to give:

(ii) $PA \vdash$ For any x in LST, Thm_{ZF}("$x \to \mathrm{Con}_\infty(\{x\})$").

Applying (ii) inside ZF, we get a theorem of ZF about models of ZF:

(iii) $ZF \vdash$ For any x in LST, if $\mathfrak{M} \models ZF$, then $\mathfrak{M} \models$ "$x \to \mathrm{Con}_\infty(\{x\})$".

Now reason in ZF: Suppose that $\mathfrak{M}$ is an ω-model of $ZF + \phi$. Produce some $b \in |\mathfrak{M}|$ such that $\mathfrak{M}$ satisfies:

$$\mathbf{b} \text{ is finite} \wedge \mathbf{b} \subseteq ZF \wedge (*) \neg \mathrm{Con}_\infty(\mathbf{b} \cup \{\ulcorner \phi \urcorner\}).$$

Because $\mathfrak{M}$ is an ω-model, b is in the standard part of $\mathfrak{M}$ and really a finite subset of ZF. So $\bigwedge\!\!\!\bigwedge b \wedge \ulcorner \phi \urcorner$ is a sentence of LST satisfied by $\mathfrak{M}$. And by (iii), $\mathfrak{M} \models \mathrm{Con}_\infty(\{\bigwedge\!\!\!\bigwedge \mathbf{b} \wedge \ulcorner \phi \urcorner\})$–contradicting $(*)$.

Proof of (i). We will prove the contrapositive. Suppose $\neg \mathrm{Con}_\infty(\{\phi\})$. Produce a set p such that $\mathrm{Proof}_\infty(0; \ulcorner \neg\phi \urcorner, p)$. By the reflection principle there is a transitive set M such that $p \in M$ and $(*)$ $\phi \leftrightarrow (\phi)^M$. The axioms from which the proof p proceeds are all of form λ_x for $x \in M$. (That each such x is in M is a consequence of the usual procedure for coding proofs; alternately, the reflection principle guarantees that we can choose M large enough to contain them all.) So M satisfies each axiom of p. Since $L_{\infty\omega}$ is sound, M satisfies the conclusion of p: i.e., $(\neg\phi)^M$. (Actually, what we know is $\langle M, \in \rangle \models \ulcorner \neg \phi \urcorner$, which is provably equivalent to $(\neg\phi)^M$.) By $(*)$, then, $\neg\phi$.

What is left is to show that ϕ is Π_1-con over ZF. We will use 4.1. Let F_0 be a fairly large finite chunk of ZF (how much we will need will become evident soon). Let F_1 be any finite subset of ZF. We want to show that ZF proves $\mathrm{Con}_\infty(F_1 + \phi)$. Let $F = F_0 \cup F_1$. Then $F + \neg \mathrm{Con}_\infty(F)$ is a theory at least as strong as $F_1 + \phi$, so it will suffice to prove $\mathrm{Con}_\infty(F + \neg \mathrm{Con}_\infty(F))$. To do that imitate the proof of Gödel's 2nd Incompleteness Theorem. The key step is justified by the following fact:

There is a finitely axiomatizable subtheory F of ZF such that for any Σ_1 sentence σ,

$F \vdash \sigma \to \mathrm{Thm}_\infty(\ulcorner F \urcorner; \ \ulcorner \sigma \urcorner)$; (cf. analogous theorem for N). (A similar argument is carried out in detail in **[Krivine-McAloon]**.) □

FINAL QUESTIONS AND COMMENTS. (1) In Theorems 6.5(ii) and 6.7 can the restriction to countable models be dropped?[2]

(2) Let $EE = \{\phi |$ every countable model of ZF can be end extended to a model of $ZF + \phi\}$. Every sentence whose consistency is provable by forcing is in EE–because generic extensions (even of nonstandard models) are end extensions. It is shown in **[Barwise, 1]** or **[Barwise, 2]** that $V = L$ is an element of EE. (Note: That $V = L$ is both Π_1 and Σ_1-con over ZF follows from the Shoenfield Absoluteness Lemma, which can be formulated as follows: For every Σ_1 *sentence* ϕ, $ZF \vdash \phi \leftrightarrow (\phi)^L$.) It is mildly interesting to note that for Barwise's argument the ω-nonstandard models were the hard ones to deal with–while from the present point of view the ω-nonstandard models are the understandable ones. How understandable is EE? Is it a complete Π^1_2 set?

REFERENCES

[Barwise, 1] *Admissible sets and structures*, Springer-Verlag, Berlin, 1975.

[Barwise, 2] *Infinitary methods in the model theory of set theory*, Logic Colloquium, 1969 (R. O. Gandy and M. Yates, Editors), North Holland, Amsterdam, 1971, pp. 53–66.

[Hajkova-Hajek] *On interpretability in theories containing arithmetic*, Fund. Math. **76** (1972), 131–137.

[Hajek, 1] *On interpretability in set theories*, Comment. Math. Univ. Carolinae **12** (1971), 73–79.

[Hajek, 2] *On interpretability in set theories*. II, Comment. Math. Univ. Carolinae **13** (1972), 445–455.

[Feferman] *Arithmetization of metamathematics in a general setting*, Fund. Math. **49** (1960), 35–92.

[Friedman] *Countable models of set theories*, Proc. Cambridge Summer School in Math Logic, 1971 (A. R. D. Mathis and H. Rogers, Editors), Lectures Notes in Math., vol. 337, Springer, Berlin and New York, 1973, pp. 539–573.

[Keisler] *Model theory for infinitary logic*, North Holland, Amsterdam, 1971.

[Kreisel] *On weak completeness of intuitionistic predicate logic*, J. Symbolic Logic **27** (1962), 139–158.

[Krivine-McAloon] *Some true unprovable formulas for set theory*, Bertrand Russell Memorial Logic Conference, Leeds, 1973, pp. 332–341.

[Levy] *A hierarchy of formulas in set theory*, Mem. Amer. Math. Soc. No. 57 (1965).

[Löb] *Solution of a problem of Leon Henkin*, J. Symbolic Logic **20** (1955), 115–118.

[Macintyre-Simmons] *Gödel's diagonalization technique and related properties of theories*, Colloq. Math. **28** (1973), 165–180.

[Matijasevic] *Diophantine representation of recursively enumerable predicates*, Proc. Second Scandinavian Logic Symposium (J. E. Fenstad, Editor), North Holland, Amsterdam, 1971.

[2] J. Quinsey has pointed out that the following elementary argument establishes 6.7(ii) directly without the restriction to countable models: If ϕ is $\Sigma_1(L)$-con over T then for each finite subset F of T, T proves that there is a transitive model of $F + \phi$. Inside an ω-nonstandard T-model $\mathfrak{M}$ we can now by overspill produce a transitive (in the sense of $\mathfrak{M}$) model of $T + \phi$. He has also pointed out that a similar argument (using Kripke's notion of fulfillability) removes the cardinality restriction from 6.5(ii).

[**Montague**] *Semantical closure and nonfinite axiomatizability*, in Infinitistic Methods, Pergamon Press, New York, 1959, pp. 45–69.

[**Shoenfield**] *Mathematical logic*, Addison-Wesley, Reading, Mass., 1967.

[**Solovay**] *On interpretability in set theories* (to appear).

DEPARTMENT OF MATHEMATICS, TEXAS TECHNICAL UNIVERSITY, LUBBOCK, TEXAS 79409

Current address: St. John's College, Annapolis, Maryland 21404