Notre Dame Journal of Formal Logic
Volume 33, Number 4, Fall 1992

On the Proofs of Arithmetical Completeness for Interpretability Logic

DOMENICO ZAMBELLA

Abstract Visser proved that **ILP** is the interpretability logic of any finitely axiomatizable theory containing $I\Delta_0 + \text{SUPEXP}$, Berarducci and Shravrukov proved that **ILM** is the interpretability logic of PA. But these proofs are not based directly on the natural semantics of interpretability logic (i.e., Veltman models). We give simpler alternative proofs of the arithmetical completeness of **ILP** and **ILM** directly based on finite Veltman models. We will provide a general setup for arithmetical completeness proofs of interpretability logic which is in the style of Solovay's arithmetical completeness proof of provability logic.

1 Introduction Visser [7] introduced the binary modal logic **IL** (interpretability logic) and its extensions **ILM** (interpretability logic with Montagna's axiom) and **ILP** (interpretability logic with a persistent relation in its models) to describe the interpretability logic of PA and the interpretability logic of any sufficiently strong theory T which is finitely axiomatizable and Σ_1 sound. The modal completeness of **IL**, **ILP**, and **ILM** was provided by de Jongh and Veltman [3] using so-called Veltman models. These are a very natural generalization of Kripke models. Visser [8] obtained the arithmetical completeness for **ILP** and, more recently, Berarducci [1] and Shavrukov [5] have shown **ILM** to be complete for arithmetical interpretation over PA. All these proofs of arithmetical completeness do not directly use the Veltman models. Using a bisimulation, Visser [8] showed **ILP** to be modal complete with respect to his so-called Friedman models and then used these to prove arithmetical completeness. Berarducci and Shravrukov also used a bisimulation due to Visser [7] showing that **ILM** is modal complete with respect to the so-called simplified models to prove arithmetical completeness. The use of simplified models in proving arithmetical completeness for **ILM** adds a complication because in general these cannot be taken to be finite.

Received June 6, 1991; revised November 11, 1991

Our aim is to provide simpler and more natural proofs of arithmetical completeness for **ILP** and **ILM.** For both we shall use the original Veltman models. As all proofs of arithmetical completeness known so far, ours are based on the ideas exposed in the pioneering work of Solovay [6] and made explicit in de Jongh, Jumelet, and Montagna [4].

The organization of this paper is the following: in the next section we review the axioms of **ILM** and **ILP** and the corresponding classes of Veltman frames. We refer the reader to the literature (see, e.g., [7], [3], and [1]) both for details and comments as well as for the proofs of soundness of the axioms. In Section 3 we present a general technique inspired by Solovay's work to obtain arithmetical completeness for theories containing **IL**, provided that we already have modal completeness with respect to a certain class of finite frames. The common preparatory work of Section 3 is used in the last two sections for the two arithmetical completeness proofs.

2 Interpretability logics The language of the logic of interpretability contains (atomic) propositional letters $p_0, p_1, \ldots$, logical connectives $\rightarrow$ and $\neg$, and a binary modal operator $\rhd$. All other connectives, such as $\wedge$, $\vee$, and $\leftrightarrow$, are defined in the usual way. We use $\bot$ for false and $\top$ for true. The unary modal operator $\Box$ is defined as $\rhd \bot$. The axioms of **IL** are:

(L0) All tautologies of the propositional calculus.
(L1) $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$.
(L2) $\Box A \rightarrow \Box\Box A$.
(L3) $\Box(\Box A \rightarrow A) \rightarrow \Box A$.
(J1) $\Box(A \rightarrow B) \rightarrow A \rhd B$.
(J2) $(A \rhd B \wedge B \rhd C) \rightarrow A \rhd B$.
(J3) $A \rhd B \rightarrow (\Diamond A \rightarrow \Diamond B)$.
(J4) $\Diamond A \rhd A$.

The deduction rules of **IL** are modus ponens and necessitation. The following two other axioms are the characteristic axioms of **ILP** and **ILM**.

(M) $A \rhd B \rightarrow (A \wedge \Box C \rhd B \wedge \Box C)$.
(P) $A \rhd B \rightarrow \Box(A \rhd B)$.

A *Veltman frame* is a triple $\langle W, S, R \rangle$ where W is a set called *universe*, R and S are respectively a binary and a ternary relation on W. The elements of W are called *nodes*. We shall write xRy for $\langle x, y \rangle \in R$ and yS_xz for $\langle x, y, z \rangle \in S$. It is further required that R is transitive and conversely well founded and that for every $x \in W$, S_x is a reflexive and transitive relation on $\{y \mid xRy\} \subseteq W$. Moreover for every $x, y, z \in W$, $xRyRz$ implies yS_xz.

A *Veltman model* is a Veltman frame together with a *forcing* relation $\Vdash$ between elements of W and the formulas of **IL** commuting with the logical connectives and satisfying the following:

$$x \Vdash \Box A \text{ iff } \forall y (xRy \Rightarrow y \Vdash A),$$

$$x \Vdash A \rhd B \text{ iff } \forall y [(xRy \;\&\; y \Vdash A) \Rightarrow \exists z (yS_xz \;\&\; z \Vdash B)].$$

As usual we shall improperly use the same letter W for the model, the frame, and the underlying universe. If W is a frame, we write $W \vDash A$ iff for all forcing relations on W and all nodes of $W, x \Vdash A$.

We shall consider two other possible properties of Veltman frames:

P: If xS_wy then xS_zy for every z such that $wRzRx$.
M: If xS_wyRz then xRz.

We call W a *P-Veltman model* (resp. *M-Veltman model*) if the underlaying frame satisfies P (resp. M).

The modal completeness of **IL**, **ILP**, and **ILM** has been proved by de Jongh and Veltman. In particular, they proved the following three theorems:

(1) **IL** $\vdash A$ iff for every finite Veltman frame W, $W \vDash A$.
(2) **ILP** $\vdash A$ iff for every finite P-Veltman frame W, $W \vDash A$.
(3) **ILM** $\vdash A$ iff for every finite M-Veltman frame W, $W \vDash A$.

3 A Solovay style strategy We want to find a general strategy for proving the arithmetical completeness of the interpretability logic for various arithmetical theories. Let T be a theory in the language of the arithmetic which is Σ_1 sound and Σ_1 complete and strong enough to formalize syntax. Given two arithmetical sentences α and β, we shall write $\alpha \rhd \beta$ to mean the arithmetical formalization of the statement: "*$T + \alpha$ interprets $T + \beta$*". It will always be clear from the context to which theory T we refer. We will use Latin letters for modal formulas and Greek letters for arithmetical formulas so that no confusion will arise from the fact that we are using the same symbols $\rhd$ and $\Box$ both for the modal and for the arithmetical operators.

An *interpretation* is a mapping ι from modal formulas to sentences of the language of the arithmetic such that:

(1) $\iota(A \to B) = \iota(A) \to \iota(B)$
(2) $\iota(\neg A) = \neg\iota(A)$
(3) $\iota(A \rhd B) = \iota(A) \rhd \iota(B)$.

Let us write **IL**(T) for the set of modal formulas which are provable in T for every interpretation ι, i.e., **IL**(T) $= \{A \mid \forall\iota \mathrm{T} \vdash \iota(A)\}$. Let **ILX** be a modal theory in the language of **IL** containing **IL**. We say that **ILX** is *arithmetically sound* for T if, for every modal formula A, if **ILX** $\vdash A$, then for every interpretation ι, T $\vdash \iota(A)$, i.e., if **IL**(T) $\supseteq$ **ILX**. We say that **ILX** is *arithmetically complete* for T if the reverse inclusion also holds, i.e., whenever A is not a theorem of **ILX** then there is an interpretation ι such that $\iota(A)$ is not provable in T.

Claim Let us suppose there is a class of finite Veltman frames X with respect to which we have modal completeness for the theory **ILX**. Let us suppose also that **IL**(T) $\supseteq$ **IL**. If for any frame $W \in X$, there is a set $\{\lambda_x \mid x \in W\}$ of arithmetical sentences such that (o)-(iv) below are satisfied, then **IL**(T) $\subseteq$ **ILX**.

(o) for every $x, y \in W$ if $x \neq y$ then T $\vdash \neg(\lambda_x \wedge \lambda_y)$
(i) for every $x \in W$, T + λ_x is consistent
(ii) for every $x \in W$, T $\vdash \lambda_x \to \Box\bigvee_{xRy} \lambda_y$

(iii) for every $x, y, z \in W$ such that yS_xz, $\mathrm{T} \vdash \lambda_x \rightarrow \lambda_y \rhd \lambda_z$
(iv) for every $x, y \in W$ such that xRy, $\mathrm{T} \vdash \lambda_x \rightarrow \neg(\lambda_y \rhd \neg\bigvee_{yS_xz} \lambda_z)$.

Proof of the claim: We assume **ILX** $\nvdash C$ and define an interpretation ι such that $\mathrm{T} \nvdash \iota(C)$. By the modal completeness there is a finite model W with frame in X such that $W \nVdash C$. Let $\{\lambda_x \mid x \in W\}$ be a set of arithmetical sentences satisfying conditions (o)–(iv). Let ι be the interpretation which maps the atomic proposition p occurring in C to $\iota(p) := \bigvee\{\lambda_x \mid x \Vdash p\}$. We shall show by induction on the complexity of the modal formula A that for every $x \in W$:

(a) $x \Vdash A \Rightarrow \mathrm{T} \vdash \lambda_x \rightarrow \iota(A)$
(b) $x \nVdash A \Rightarrow \mathrm{T} \vdash \lambda_x \rightarrow \neg\iota(A)$.

This will suffice to prove the arithmetical completeness, because if $W \nVdash C$ then for some forcing relation on W and some $x \in W$, $x \nVdash C$, from which then by (b), $\mathrm{T} \vdash \lambda_x \rightarrow \neg\iota(C)$. By (i), λ_x is consistent with T, as is therefore $\neg\iota(C)$. Hence $\mathrm{T} \nvdash \iota(C)$.

It remains only to prove (a) and (b) by induction on the complexity of the formula A. By condition (o) it is clear that (a) and (b) hold for atomic sentences. The inductive steps for $\rightarrow$ and $\neg$ are straightforward, so let us consider just the inductive steps for $\rhd$.

Let us prove first (a). Assume $x \Vdash A \rhd B$. Then for every y such that xRy, if $y \Vdash A$, there is a node z such that $yS_xz \Vdash B$. By the induction hypothesis we can write: for every y such that xRy, if $y \Vdash A$, there is a node z such that yS_xz and $\mathrm{T} \vdash \lambda_z \rightarrow \iota(B)$. Using (iii) and Σ_1 completeness and the soundness of **IL** (i.e., making few steps of reasoning in **IL**) we get $\mathrm{T} \vdash \lambda_x \rightarrow \bigwedge_{xRy \Vdash A}(\lambda_y \rhd \iota(B))$ and finally $\mathrm{T} \vdash \lambda_x \rightarrow (\bigvee_{xRy \Vdash A} \lambda_y \rhd \iota(B))$. On the other hand, by (ii) and using the induction hypothesis (b) we obtain $\mathrm{T} \vdash \iota(A) \rightarrow \neg\bigvee_{y \nVdash A} \lambda_y$, from which, since we assumed $\mathrm{T} \vdash \lambda_x \rightarrow \Box\bigvee_{xRy} \lambda_y$, we get $\mathrm{T} \vdash \lambda_x \rightarrow \Box(\iota(A) \rightarrow \bigvee_{xRy \Vdash A} \lambda_y)$. Again by the soundness of **IL**, $\mathrm{T} \vdash \lambda_x \rightarrow \iota(A) \rhd \bigvee_{xRy \Vdash A} \lambda_y$. Thus the proof of (a) follows.

We now prove (b). Assume $x \nVdash A \rhd B$. Then there is a y such that xRy and $y \Vdash A$ and for every node z such that $yS_xz, z \nVdash B$. Thus, for some y such that xRy we have: $y \Vdash A \wedge \bigwedge_{yS_xz} z \nVdash B$. By the inductive hypotheses we have $\mathrm{T} \vdash \lambda_y \rightarrow \iota(A)$ and $\mathrm{T} \vdash \bigvee_{yS_xz} \lambda_z \rightarrow \neg\iota(B)$. By Σ_1 completeness we have $\mathrm{T} \vdash \Box[\lambda_y \rightarrow \iota(A)]$ and $\mathrm{T} \vdash \Box[\iota(B) \rightarrow \neg\bigvee_{yS_xz} \lambda_z]$, from which by the soundness of **IL** we get $\mathrm{T} \vdash \lambda_y \rhd \iota(A)$ and $\mathrm{T} \vdash \iota(B) \rhd \neg\bigvee_{yS_xz} \lambda_z$. Reason in T and assume λ_x. Assume for a contradiction that $\iota(A) \rhd \iota(B)$. By the soundness of **IL** we would have $\lambda_y \rhd \neg\bigvee_{yS_xz} \lambda_z$, so from (iv) we obtain the desired contradiction. This completes the proof of the claim.

We conclude this section by remarking that conditions (o)–(iv) are not in general necessary, we believe that with a little additional work one can obtain more general, sufficient and necessary, conditions as is done in [2] for the case of provability logic.

4 The interpretability logic of finitely axiomatizable theories In this section T may be any finitely axiomatizable Σ_1 sound theory extending $\mathrm{I}\Delta_0$ + SUPEXP. The main property which distinguishes interpretability over these theories is that the interpretability predicate in T is Σ_1, from which the soundness of the modal axiom P follows immediately. In T it is possible to characterize interpretability as follows. Let Δ_{EXP} be tableaux provability in $\mathrm{I}\Delta_0 + \mathrm{EXP}$, Δ tableaux provably in T and $\nabla = \neg\Delta\neg$, i.e., the tableaux consistency in T. According to the Friedman–Visser characterization [8], α interprets β iff $\Delta_{\mathrm{EXP}}(\nabla\alpha \rightarrow \nabla\beta)$.

We want to prove that **IL**(T) = **ILP**. We leave, as usual, the proof of soundness to the reader and we shall prove only **IL**(T) $\subseteq$ **ILP**. We shall find sentences (o)–(iv) as in the previous section. The method is as in [6]. We define a function F using the fixed point theorem and let the λ_x be some limit statements concerning F.

Assume for convenience that W has been given as a finite set of nonzero natural numbers. We shall use the symbols x, y, and z only for elements of W. Let λ_x be the sentence $\lim_n F(n) = x$ and $\lambda_0 := \forall n F(n) = 0$. Together with the function F we will also define an auxiliary function G which will aid us in bookkeeping. The function G will always "follow" the function F, i.e., if for some n, $F(n) = x$ then $G(n) = F(m)$ for some $m \leq n$. Speaking informally, $G(n) \neq F(n)$ will warn us of the fact that there is no proof of code less than n of $\neg\lambda_{F(n)}$. This has to be considered a "dangerous signal" since we would like in the end to have $\lambda_x \rightarrow \Box\neg\lambda_x$. When such a situation occurs then only "safe" moves are allowed, i.e., F as well as G will move only to a node y for which there is a proof of $\neg\lambda_y$.

The definitions of F and G are as follows:

(a) $F(0) = G(0) = 0$. If $F(n) = 0$ and for some $x \in W$, n witnesses $\Delta\ \neg\lambda_x$, then $F(n+1) = G(n+1) = x$.
(b) If $F(n) = G(n) = x \in W$ and for some node y such that xRy, n witnesses $\Delta_{\mathrm{EXP}}(\nabla\lambda_y \rightarrow \nabla\neg\bigvee_{yS_xz}\lambda_z)$, then $F(n+1) = y$ and $G(n+1) = G(n)$.
(c) If $F(n) = y$ and $G(n) = x$, for some z, yS_xz and n witnesses $\Delta\neg\lambda_z$, then $F(n+1) = G(n+1) = z$.
(d) In all other cases $F(n+1) = F(n)$ and $G(n+1) = G(n)$.

Let μ_x be the sentence $\lim_n G(n) = x$. We shall eventually prove that the two functions have the same limit, i.e., $\mu_x \leftrightarrow \lambda_x$, but for proving this we need the cut elimination theorem. The formalization of the cut elimination theorem is provable in T since T contains SUPEXP but is surely not provable in EXP. To carry on with our proof we need to know what $\mathrm{I}\Delta_0 + \mathrm{EXP}$ proves about the functions F and G, hence the following:

Lemma 1 $\mathrm{I}\Delta_0 + \mathrm{EXP}$ *proves the following*:
(1.1) *For every* $w \in W$, $\mu_w \rightarrow \Delta\bigvee_{wRx}\lambda_x$.
(1.2) *For every* $w, x \in W$, *if* $x \neq w$ *then* $\mu_w \wedge \lambda_x \rightarrow \Delta\bigvee_{xS_wy}\lambda_y$.
(1.3) *For every* $w, y \in W$ *if* wRy *then* $\mu_w \wedge \lambda_w \rightarrow \nabla\lambda_y$.
(1.4) *For every* $x, y, w \in W$, *if* xS_wy *then* $\mu_w \wedge \lambda_x \rightarrow \nabla\lambda_y$.

Proof: Directly from the definition of F, $\mathrm{I}\Delta_0 + \mathrm{EXP}$ proves that if, for some n, $G(n) = w$ then after stage n the function F remains either in w or in the upper cone above w. Thus the limit of F is either w or some node above w. If $G(n) = w$ then by provable Σ_1 completeness, $\Delta_{\mathrm{EXP}}(G(n) = w)$ and a fortiori

$\Delta(G(n) = w)$. The proof of (1.1) follows by combining all this with the fact that $G(n) = w$ implies $\Delta\neg\lambda_w$. To prove (1.2) assume that for some $x \neq w$ we have $\mu_w \wedge \lambda_x$. Then for some n $\Delta_{\mathrm{EXP}}(G(n) = w \wedge F(n) = x)$. Again, observing the definition of the functions F and G, it is easy to argue that whenever $G(n) = w \wedge F(n) = x$ for some $w \neq x$, the function F never leaves the set of nodes which are in S_w relation with x. This gives (1.2). To prove (1.3) assume wRy, λ_w and μ_w, and let n be such that for all $m > n$, $F(m) = G(m) = w$. If $\neg\lambda_y$ were cut free provable, then some $m > n$ would witness $\Delta\neg\lambda_y$. (Here and in the following it is assumed that a cut free provable theorem has infinitely many cut free proofs.) So $\Delta_{\mathrm{EXP}}(\nabla\lambda_y \to \nabla\neg\bigvee_{yS_xz}\lambda_z)$ and then at stage $m + 1$, F would move to y, against our assumption that at stage n F has already reached its limit. To prove (1.4) assume λ_x, μ_w, and xS_wy. Then wRy, and therefore $w \neq y$. Let n be such that for all $m > n$, $F(m) = x$ and $G(m) = w$. Suppose, by contradiction, that $\Delta\neg\lambda_y$. Let $m > n$ a witness of $\Delta\neg\lambda_y$. Then at stage $m + 1$ both F and G move to y, by condition (c). This contradicts our assumption that at stage n G has already reached its limit. (Note that clearly $y \neq w$ since xS_wy and then wRy.)

For the following lemma we need that the formula $((\nabla\alpha \wedge \alpha \rhd \beta) \to \nabla\beta)$ is provable in T. It is easy to check that T (or even $I\Delta_0 + \mathrm{EXP}$) proves $((\Diamond\alpha \wedge \alpha \rhd \beta) \to \Diamond\beta)$, and since in T the formalization of the cut elimination theorem is provable, we can substitute tableaux consistency with normal consistency, so also the former formula is derivable in T. We can prove the following:

Lemma 2 *For every* $x \in W$, $\mathrm{T} \vdash \mu_x \leftrightarrow \lambda_x$.

Proof: Reason in T and assume for a contradiction that $\lambda_x \wedge \neg\mu_x$. Then for some wRx we have μ_w. This implies $\nabla\lambda_x$, for otherwise the function G would have jump to x. Since $x \neq w$ the last move of the function F has been from w to x using condition (b) and therefore $\lambda_x \rhd \neg\bigvee_{xS_wy}\lambda_y$. By the remark above we get immediately $\neg\Delta\bigvee_{xS_wy}\lambda_y$. From Lemma 1.2 we also get $\Delta\bigvee_{xS_wy}\lambda_y$. Thus we have the desired contradiction.

Lemma 3 *For every* $x, y, z \in W$ *such that* yS_xz, $\mathrm{T} \vdash \lambda_x \to \lambda_y \rhd \lambda_z$.

Proof: Reason in T and assume λ_x. We want to show that for every y, z such that $yS_xz, \lambda_y \rhd \lambda_z$, i.e., $\Delta_{\mathrm{EXP}}(\nabla\lambda_y \to \nabla\lambda_z)$. By Lemma 2 we have μ_x and by provable Σ_1 completeness we have that for some k, $\Delta_{\mathrm{EXP}}(G(k) = x)$. Reason in $I\Delta_0 + \mathrm{EXP}$. Assume $\nabla\lambda_y$ and let w be the limit of the function G. Since $G(k) = x$, the limit w is either x or is above x. By Lemma 1.1, from $\nabla\lambda_y$ we know that w has to be strictly below y. Thus either $x = wRy$ or $xRwRy$ and, by the characteristic property of the P-Veltman frames, from yS_xz we get yS_wz. Let u be the limit of F. If $u = w$ from wRz and Lemma 1.3, the lemma follows immediately. Otherwise by Lemma 1.2 and $\nabla\lambda_y$ one has uS_wy. By the transitivity of S_w we obtain uS_wz and thus finally, by Lemma 1.4, $\nabla\lambda_z$.

Lemma 4 *For every* $x \in W$, $\mathrm{T} \vdash \lambda_x \to \Delta\bigvee_{xRy}\lambda_y$.

Proof: Immediate by Lemmas 1.1 and 2.

We can now easily check that the set of sentences $\{\lambda_x \mid x \in W\}$ satisfies (o)–(iv). In fact (o) is trivial, the proof of (i) is completely standard, (ii) derives from

Lemma 4 and the provability in T of the cut elimination theorem. Condition (iii) is Lemma 3 and (iv) is obvious by the definition of F and Lemma 2. This concludes the proof of the completeness theorem.

5 ***The interpretability logic of PA*** In this section we want to prove that **IL**(PA) = **ILM**. The main characteristic of the interpretability in Peano arithmetic is the Orey–Hajek characterization: let $\Box_k\beta$ be the formalization of the sentence "*there is a proof of β which uses only the first k axioms of PA*", let $\Diamond_k \equiv \neg\Box_k\neg$, then it is provable in PA that α interprets β iff $\forall k\Box(\alpha \to \Diamond_k\beta)$. Another characteristic property of PA is that it proves full reflection for any of its finite subtheories. Moreover, this is formalizable in PA, namely: for every α, PA $\vdash \forall k\Box(\Box_k\alpha \to \alpha)$. These facts would be sufficient to carry out the following proof, but for the sake of better readability we shall, following Berarducci, work in ACA_0 rather than in PA. The second-order theory ACA_0 is a conservative extension of PA; in ACA_0 we can speak of models of PA and easy theorems of basic model theory are formalizable and provable in ACA_0. In particular, in ACA_0 we have the following characterization of the interpretability over PA: "*PA + α interprets PA + β iff every model of PA + α has an end extension to a model of PA + β*". In ACA_0 the *standard model* is the set $\{x \mid x = x\}$ with the obvious choice of operations; any other *nonstandard model* has an initial segment which is isomorphic to it. Numbers belonging to this initial segment are called, as usual, *standard numbers*. Full reflection translates in ACA_0 in the following manner: "*for every model Y of PA and every standard number k, $Y \models \Box_k\alpha \to \alpha$*".

As in the previous section we shall prove only that **IL**(PA) $\subseteq$ **ILM**, leaving the converse to the reader. The sentences which are meant to satisfy (o)–(iv) are defined as limits of a recursive function F exactly as in the previous proof. Define, as in [1] for every $x \in W$, $\mathrm{rank}(x,n) :=$ "*the minimal k such that there is a witness $\leq n$ of $\Box_k\neg\lambda_x$*". If k is a number, $x, y \in W$, xRy then we define the sentence $\alpha_{x,y}(k)$ as $\forall j \geq k[F(j) = x \vee F(j) = y]$. Our definition of the function F resembles Berarducci's as far as it is concerned with the S-jumps, but it differs in the R-jumps. Roughly speaking, we allow the function F to make an R-jump if there is a proof that this will not be the last move. We assume for convenience that W has been coded as a finite set of nonzero natural numbers, and we shall use the symbols $w, x, y, \ldots$ etc. only for elements of W.

Proof:

(a) Let $F(0) = 0$ and if $F(n) = 0$ and for some $x \in W, n$ witnesses $\Box\neg\lambda_x$, then $F(n+1) = x$.

(b) If $F(n) = x$ and for some $y \in W$ and some $k < n$ such that $\forall j \in [k,n] F(j) = x$ and xRy, n witnesses $\Box\neg\alpha_{x,y}(\dot{k})$ ($\dot{k}$ is the numeral of k), then $F(n+1) = y$.

(c) If $F(n) = x$ and for some nodes y and z, xS_zy and $\exists i \leq n[\mathrm{rank}(y,n) \leq i < \mathrm{rank}(x,n) \wedge F(i) = z]$, then $F(n+1) = y$. (If this condition obtains for two different nodes, choose the one with minimal code.)

(d) In all the other cases $F(n+1) = F(n)$.

Note that any two points in the orbit of F are connected by an S- and/or R-arrow. We shall write $Y \models \ldots x \ldots y$ if, according to the model Y, the function

F goes from x to y (possibly in a nonstandard number of steps). We write $Y \models \ldots xRy \ldots$ (resp. $Y \models \ldots xS_zy \ldots$) if, in the model Y, F moves in one step from x to y and xRy (resp. xS_zy). If in a model Y the function F moves at stage n from x to y, then we say F moves with an R-step (resp. with S-step) if at stage n condition (b) (resp. condition (c)) has been applied. If, at stage n, F moves from 0 to some node x, we say that F moves with an (a)-step.

Lemma 1 *In PA it is provable that the function F has a limit.*

Proof: This is not obvious since the S-relations are, in general, not well founded. It is clear that if h is the height of the frame the function cannot make more than h consecutive R-moves. By the property M of the M-frame, F cannot make more than h R-moves whether they are consecutive or not. Thus eventually F is allowed only to make S moves. If F would not have a limit, we could construct a definable infinite decreasing sequence of ranks. This is provably false in PA.

We are eventually going to prove $\lambda_x \to \Box\neg\lambda_x$, but to achieve this goal we first need to prove a weaker form of it.

Lemma 2 *For every $x \in W$ and for every $k \in \omega$*, PA $\vdash F(k) = x \to \Box\exists j > \dot{k}F(j) \neq x$.

Proof: Assume $F(k) = x$. Reasoning in ACA_0 we claim that for every model Y of PA, $Y \models \exists j > kF(j) \neq x$. If F moved to x with an (a)-step or with an S-step, we would have $\Box\neg\lambda_x$ and then $Y \models \neg\lambda_x$ so our claim would hold trivially. So, assume that the last move of F has been an R-step, and that say at stage h, the function F moves from z to x. Then for some $i < h$ such that $\forall j \in [i,h]F(j) = z$, h codes a proof of $\neg\alpha_{z,x}(i)$. So, $Y \models \exists j \geq i[F(j) \neq z \wedge F(j) \neq x]$. We have assumed $\forall j \in [i,k][F(j) = z \vee F(x)]$, this is a Σ_1 statement so, by provable Σ_1 completeness, it is true also in Y. Thus $Y \models \exists j > kF(j) \neq x$ and our claim is proved.

Lemma 3 *For every $x \in W$*, PA $\vdash \lambda_x \to \Box\bigvee_{xRy}\lambda_y$.

Proof: It is sufficient to prove that for every x and y, if $\neg xRy$ then PA $\vdash \lambda_x \to \Box\neg\lambda_y$. Reason in ACA_0 and assume for contradiction that λ_x, $\Diamond\lambda_y$, and $\neg xRy$. Choose k such that $F(k) = x$ and let Y be a model of λ_y. By provable Σ_1 completeness we have $Y \models F(k) = x$. Now in Y let z be the last node through which the function passes before arriving at y. The last step must be an S-step otherwise zRy and, by the M property of the M-Veltman frames, we would have xRy. We shall picture the situation as $Y \models \ldots x \ldots zS_wy$. (We recall that either z or y might be equal to x; the previous lemma only guarantees that after stage k the function has moved at least once.) We assumed $\neg xRy$; thus, since zS_wy implies wRy, we have that $w \neq x$. By the definition of F we have that at some stage n, for some $i \leq n$, $\mathrm{rank}(y,n) \leq i < \mathrm{rank}(z,n)$ and $F(i) = w$. By the reflection principle $\mathrm{rank}(y,n)$ has to be nonstandard in Y, and since we have chosen k standard, $\mathrm{rank}(y,n) \geq k$. Thus also $i \geq k$ and so $Y \models \ldots F(k) \ldots F(i)$ and therefore $Y \models \ldots x \ldots w \ldots zS_wy$. By the M property of the M-Veltman frames from wRy we get xRy. Contradiction.

Lemma 4 *For every $x, y, z \in W$ such that yS_xz*, PA $\vdash \lambda_x \to \lambda_y \rhd \lambda_z$.

Proof: Assume λ_x and yS_xz. We shall prove in ACA_0 that, for arbitrary large k, in any model Y of PA, $\lambda_y \to \Diamond_k\lambda_z$. Let k be such that $F(k) = x$. Suppose for a contradiction that there exists a model $Y \models \lambda_y \wedge \Box_k\neg\lambda_z$. Then for n large enough we have $Y \models \mathrm{rank}(z,n) \leq k < n$. Suppose n is also large enough so that (in Y) F has already reached its limit. By the reflection principle $\mathrm{rank}(y,n)$ must be nonstandard in Y. Then $Y \models \mathrm{rank}(z,n) \leq k < \mathrm{rank}(y,n) \wedge \mathrm{F}(k) = x$. So, $Y \models F(n+1) = z$ which contradicts the fact that F has already reached its limit.

Lemma 5. *For every $x, y \in W$ such that xRy,* $\mathrm{PA} \vdash \lambda_x \to \neg(\lambda_y \rhd \neg\bigvee_{yS_xz}\lambda_z)$.

Proof: Reason in ACA_0 and assume λ_x. To prove $\neg(\lambda_y \rhd \neg\bigvee_{yS_xz}\lambda_z)$ it will suffice to find a model Y of λ_y which has no end extension to a model of $\neg\bigvee_{yS_xz}\lambda_z$. Fix k such that $\forall j \geq kF(j) = x$. Since xRy we have: $\Diamond\alpha_{x,y}(\dot{k})$; otherwise the function would jump from x to y contradicting λ_x. Then we can choose our model Y such that $Y \models \forall j > k[F(j) = x \vee F(j) = y]$; since we have assumed λ_x and therefore (by Lemma 3) $Y \models \neg\lambda_x$, we can conclude that $Y \models \lambda_y$. Let Z be any end extension of such a model Y and let z be such that $Z \models \lambda_z$. The proof is complete if we can show that yS_xz. Let n be the minimal number in Z such that $Z \models F(n+1) = z$. By provable Σ_1 completeness and the fact that Σ_1 formulas are conserved by end extensions, we have $Z \models \ldots xRy \ldots z$. Let w be the last node reached with an R step (i.e., for some $u, Z \models \ldots xRy \ldots uRw \ldots z$ and between w and z only S steps occur). Then the rank of all the steps between w and z is larger than $\mathrm{rank}(z,n)$. By the reflection principle $\mathrm{rank}(z,n)$ is a nonstandard number in Z. If all the steps between w and z are S_x steps, we are done; otherwise let S_t be the last non-S_x step between w and z (i.e., $Z \models \ldots xRy \ldots uRw \ldots S_tvS_x \ldots S_xz$). Let $i \geq \mathrm{rank}(z,n)$ be such that $F(i) = t$. Since $\mathrm{rank}(z,n)$ is nonstandard in Z, t cannot occur in the orbit of F before x; so either $t = y$ or $Z \models \ldots xRy \ldots t \ldots S_tvS_x \ldots S_xz$. In both cases one can conclude that yRv and hence yS_xz.

We can now easily check that the set of sentences $\{\lambda_x \mid x \in W\}$ satisfies (o)–(iv). In Fact (o) is trivial, the proof of (i) is completely standard, (ii) is Lemma 3, (iii) is Lemma 4, and (iv) is Lemma 5. This concludes the proof of the completeness theorem.

Acknowledgment I would like to thank Albert Visser for correcting and simplifying some of my arguments, and Dick de Jongh and Rineke Verbrugge for their continuous and patient help.

REFERENCES

[1] Berarducci, A., "The interpretability logic of Peano arithmetic," *Journal of Symbolic Logic*, vol. 56 (1990), pp. 1059–1089.

[2] Berarducci, A. and R. Verbrugge, "On the metamathematics of weak theories," *ITLI Prepublication Series*, ML-91-02, University of Amsterdam, 1991.

[3] de Jongh, D. and F. Veltman, "Provability logic for relative interpretability," pp. 31–42 in *Mathematical Logic*, edited by P. Petkov, Plenum Press, New York, 1990.

[4] de Jongh, D., M. Jumelet, and F. Montagna, "On the proof of Solovay's theorem," *Studia Logica*, forthcoming.

[5] Shavrukov, V., "The logic of relative interpretability over Peano arithmetic," *Steklov Mathematical Institute Moscow*, 1988.

[6] Solovay, R., "The provability interpretation of modal logic," *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287-304.

[7] Visser, A., "Preliminary notes on interpretability logic," *University of Utrecht Logic Group Preprint Series*, vol. 29 (1988).

[8] Visser, A., "Interpretability logic," pp. 175-209 in *Mathematical Logic*, edited by P. Petkov, Plenum Press, New York, 1990.

Department of Mathematics and Computer Science
University of Amsterdam
1018 TV Amsterdam
The Netherlands