CHAPTER VII

The Logic of Provability

Giorgi Japaridze
Department of Computer and Information Science, University of Pennsylvania
Philadelphia, Pennsylvania 19104-6389, USA

Dick de Jongh
Institute for Logic, Language and Computation, University of Amsterdam
NL-1018 TV Amsterdam, The Netherlands

This chapter is dedicated to the memory of George Boolos. From the start of the subject until his death on 27 May 1996 he was the prime inspirer of the work in the logic of provability.

Contents

HANDBOOK OF PROOF THEORY
Edited by S. R. Buss

1. Introduction, Solovay's theorems

Gödel's incompleteness theorems and Church's undecidability theorem for arithmetic showed that reasonably strong formal systems cannot be complete and decidable, and cannot prove their own consistency. Even at the time though these negative theorems were accompanied by positive results. Firstly, formal systems fare better in reasoning in restricted areas, and this reasoning can be formalized in the theories themselves. In Hilbert and Bernays [1939] one finds the formalization of the completeness theorem for the predicate calculus, i.e., reasoning in the predicate calculus is adequately described in strong enough theories. A fortiori, this is so for the propositional calculus in which reasoning is even (provably) decidable. Secondly, there is a positive component in the incompleteness theorems themselves. The formalized version of the second incompleteness theorem, i.e., if it is provable in **PA** that **PA** is consistent, then **PA** is inconsistent, is provable in **PA** itself. The area here called the logic of provability arose in the seventies when two developments took place almost simultaneously. The two facets mentioned above were, one might say, integrated by showing that propositional reasoning about the formalized provability predicate is decidable and can be adequately described in arithmetic itself. And in the same period the de Jongh-Sambin fixed point theorem (see Sambin [1976], Smoryński [1978,1985]) was proved for modal-logical systems with the provability interpretation in mind. Since that time the main achievements have been to show that similar results mostly fail for predicate logic, to recognize reasoning about more complex notions like interpretability where arithmetic can be shown to reason adequately, and also to strengthen Solovay's results directly. Extensive overviews on the subject can be found in Boolos [1993b] and Smoryński [1985], a short history in Boolos and Sambin [1991].

Let us proceed somewhat farther in formulating Solovay's theorems, and call an *arithmetic realization* of the language of modal logic (see section 2) into the language of the arithmetic theory T (Σ_1-sound and extending $\mathbf{I\Sigma_1}$, sometimes $\mathbf{I\Delta_0}$) a mapping * that commutes with the propositional connectives and such that $(\Box A)^* = \mathrm{Pr}_T(\ulcorner A^* \urcorner)$ (where Pr_T is the formalized *provability* predicate for T, i.e., it is of the form $\exists y \mathrm{Proof}_T(x, y)$ where Proof_T is the formalized *proof* predicate of T). If we want to stress the dependency on T we write $(A)^*_T$ for $(A)^*$. More standard is the term "interpretation" for "realization" but that conflicts somewhat with our terminology with regard to interpretability. The term "realization" is used by Boolos [1993b].

1.1. Theorem. (Solovay's first arithmetic completeness theorem) *The modal formula A is provable in T under all arithmetic realizations iff A is provable in the modal logic* **L** (see sections 2, 3).

1.2. Theorem. (Solovay's second arithmetic completeness theorem) *The modal formula A is true under all arithmetic realizations iff A is provable in the modal logic* **S** (see sections 2, 3).

This chapter is to be thought of as divided into three parts: the first part consists of sections 2-10 and is devoted to propositional provability logic, i.e., the propositional logic of the provability predicate and its direct extensions, the second part consists of sections 11-15 and treats interpretability logic and related areas, the last part consists of section 16 and discusses predicate provability logic.

2. Modal logic preliminaries

The language of the modal propositional calculus consists of a set of propositional variables, connectives $\vee$, $\wedge$, $\rightarrow$, $\leftrightarrow$, $\neg$, $\top$, $\bot$ and a unary operator $\Box$. Furthermore, $\Diamond$ is an abbreviation of $\neg \Box \neg$. The modal logic **K** is axiomatized by the schemes 1 and 2:

1. All propositional tautologies in the modal language,

2. $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$,

together with the rules of modus ponens and *necessitation*, i.e., $A/\Box A$. The modal logic **L** is axiomatized by adding the scheme 3:

3. $\Box(\Box A \rightarrow A) \rightarrow \Box A$,

to **K** and keeping the rules of modus ponens and necessitation. The system is often called **GL**, e.g. in Boolos [1993b], and is named **PrL** in Smoryński [1985]. It is an exercise to show that $\Box A \rightarrow \Box\Box A$ is derivable in **L**, which makes **L** an extension of **K4**, the system axiomatized by the axioms of **K** together with $\Box A \rightarrow \Box\Box A$. Extensions of **K** such as **K4** and **L** that are closed under necessitation are called *normal* modal logics.

We will write $A_1, \ldots, A_n \vdash_{\mathbf{K}} B$ for: B is derivable in **K** from $A_1, \ldots, A_n$ without use of necessitation, or more precisely, B is derivable by modus ponens from theorems of **K** plus $A_1, \ldots, A_n$. Similarly for **K4**, **L**. To this notation the *deduction theorem* obviously applies: $A_1, \ldots, A_n, B \vdash_{\mathbf{K}} C$ iff $A_1, \ldots, A_n \vdash_{\mathbf{K}} B \rightarrow C$. We will write $\boxdot A$ for $A \wedge \Box A$. The results codified in the next proposition are readily proved.

2.1. Proposition.
(a) *If* $A_1, \ldots, A_n \vdash_{\mathbf{K}} B$, *then* $\Box A_1, \ldots, \Box A_n \vdash_{\mathbf{K}} \Box B$ *(also for* **K4**, **L***)*,
(b) *if* $\Box A_1, \ldots, \Box A_n \vdash_{\mathbf{K4}} B$, *then* $\Box A_1, \ldots, \Box A_n \vdash_{\mathbf{K4}} \Box B$ *(also for* **L***)*,
(c) *if* $\Box A_1, \ldots, \Box A_n, \Box B \vdash_{\mathbf{L}} B$, *then* $\Box A_1, \ldots, \Box A_n \vdash_{\mathbf{L}} \Box B$,
(d) $\vdash_{\mathbf{K}} \Box(A \wedge B) \leftrightarrow \Box A \wedge \Box B$,
(e) $\vdash_{\mathbf{K4}} \Box(A \leftrightarrow B) \rightarrow \Box(C(A) \leftrightarrow C(B))$,
(f) $\vdash_{\mathbf{K4}} \Box(A \leftrightarrow B) \rightarrow (\Box C(A) \leftrightarrow \Box C(B))$,
(g) $\vdash_{\mathbf{K4}} \boxdot(A \leftrightarrow B) \rightarrow (C(A) \leftrightarrow C(B))$,
(h) *if* $\vdash_{\mathbf{K},\mathbf{K4},\mathbf{L}} A \leftrightarrow B$, *then* $\vdash_{\mathbf{K},\mathbf{K4},\mathbf{L}} C(A) \leftrightarrow C(B)$,
(i) $\vdash_{\mathbf{K}} \Diamond\bot \rightarrow \bot$,

(j) $\vdash_{\mathbf{L}} \diamondsuit p \to \diamondsuit(p \wedge \Box\neg p)$ *and, hence,* $\vdash_{\mathbf{L}} \diamondsuit p \leftrightarrow \diamondsuit(p \wedge \Box\neg p)$ *and* $\vdash_{\mathbf{L}} p \to (p \wedge \Box\neg p) \vee \diamondsuit(p \wedge \Box\neg p)$.

The modal logic **S** is defined by:

$\vdash_{\mathbf{S}} A$ if and only if $\Box B_1 \to B_1, \ldots, \Box B_k \to B_k \vdash_{\mathbf{L}} A$ for some $B_1, \ldots, B_k$.

The logic **S** is not closed under necessitation, and is therefore a nonnormal modal logic.

2.2. Definition.
(a) A *Kripke-frame* for **K** is a pair $\langle W, R\rangle$ with W a nonempty set of so-called *worlds* or *nodes* and R a binary relation, the so-called *accessibility* relation.
(b) A *Kripke-frame* for **K4** is a Kripke-frame $\langle W, R\rangle$ with R a transitive relation.
(c) A *Kripke-frame* for **L** is a Kripke-frame $\langle W, R\rangle$ with R a transitive relation such that the converse of R is well-founded. (Of course, a finite transitive frame is conversely well-founded iff it is irreflexive.)
(d) A *root* of a Kripke-frame is a node w such that $w\,R\,w'$ for all $w' \neq w$ in the frame. (In the case of **K** put the transitive closure of R here in place of R.) The *depth* (also *height*) of a node w in a conversely well-ordered frame is the maximal m for which there exists a sequence $w = w_0 R\, w_1 \ldots R\, w_m$. The *height* of the model is the maximum of the height of its nodes.
(e) A *Kripke-model* for **K** (**K4**, **L**) is a tuple $\langle W, R, \Vdash\rangle$ with $\langle W, R\rangle$ a Kripke-frame for **K** (**K4**, **L**) together with a *forcing relation* $\Vdash$ between worlds and propositional variables. The relation $\Vdash$ is extended to a relation between worlds and all formulas by the stipulations

$w \Vdash \neg A$ iff $w \nVdash A$,
$w \Vdash A \wedge B$ iff $w \Vdash A$ and $w \Vdash B$,
and similarly for the other connectives,
$w \Vdash \Box A$ iff for all w' such that $wR\,w'$, $w' \Vdash A$.

If $K = \langle W, R, \Vdash\rangle$, then $K \vDash A$ is defined as, $w \Vdash A$ for each $w \in W$, and we say that A is *valid* in M.

It is easy to check that Kripke-models are *sound* in the sense that each A derivable in **K** (**K4**, **L**) is valid in each Kripke-model for **K** (**K4**, **L**). In fact, the Kripke-models for **K4** (resp. **L**) are exactly the ones that validate the formulas derivable, respectively in **K4** and **L**. One says that **K4** and **L** *characterize* these classes of models. (For the main concepts of modal logic, see e.g., Chellas [1980], Hughes and Cresswell [1984].) Something stronger is true: in **K**, **K4** and **L** one can derive all the formulas that are valid in their respective model classes (*modal completeness*). The standard method in modal logic for proving completeness is to construct the necessary countermodels by taking maximal consistent sets of the logic as the worlds of the model and providing this set of worlds with an appropriate accessibility relation R. This method cannot be applied to **L**, since the logic is

not compact: there exist infinite syntactically consistent sets of formulas that are semantically incoherent. We will apply to all three logics a method in which one restricts maximal consistent sets to a finite so-called adequate set of formulas. One obtains finite countermodels by this method and hence, immediately, decidability of the logics.

2.3. Definition. If A is not a negation, then $\sim A$ is $\neg A$, otherwise, if A is $\neg B$, then $\sim A$ is B.
An *adequate* set of formulas is a set Φ with the properties:
(i) Φ is closed under subformulas,
(ii) if $B \in \Phi$, then $\sim B \in \Phi$.

It is obvious that each formula is contained in a finite adequate set.

2.4. Theorem. (Modal completeness of **K**, **K4**, **L**.)
If A is not derivable in **K** (**K4**, **L**)*, then there is a frame for* **K** (**K4**, **L**) *on which A is not valid.*

Proof. (**K**) Suppose $\nvdash_{\mathbf{K}} A$. Let Φ be a finite adequate set containing A. We consider the set W of all maximal **K**-consistent subsets of Φ. We define for $w, w' \in W$,

$$w R\, w' \iff \text{for all } \Box D \in w,\ D \in w'.$$

Furthermore, we define $w \Vdash p$ iff $p \in w$. It now follows that for each $B \in \Phi$, $w \Vdash B$ iff $B \in w$, by induction on the length of B. For propositional letters this is so by definition and the case of the connectives is standard, so let us consider the case that B is $\Box C$.
$\Longrightarrow$: Assume $\Box C \in w$. Then, for all w' such that $w R\, w'$, $C \in w'$. By induction hypothesis, $w' \Vdash C$ for all such w'. So, $w \Vdash \Box C$.
$\Longleftarrow$: Assume $\Box C \notin w$. Consider the set $\{D | \Box D \in w\} \cup \{\sim C\}$. We will show this set to be **K**-consistent which means, by the conditions on adequate sets, that a maximal **K**-consistent superset w' of it exists inside Φ. By induction hypothesis, $w' \nVdash C$, and since $w R\, w'$, this implies that $w \nVdash \Box C$.

To show that $\{D | \Box D \in w\} \cup \{\sim C\}$ is indeed **K**-consistent, suppose that it is not, i.e., $D_1, \ldots, D_k \vdash_{\mathbf{K}} C$ for some $\Box D_1, \ldots, \Box D_k \in w$. Then $\Box D_1, \ldots, \Box D_k \vdash_{\mathbf{K}} \Box C$ immediately follows, but that would make w inconsistent, contrary to what was assumed.

(**K4**) Suppose $\nvdash_{\mathbf{K4}} A$. Proceed just as in the case of **K**, except that now:

$$w R\, w' \iff \text{for all } \Box D \in w,\ \text{both } \Box D \in w' \text{ and } D \in w'.$$

The argument is as for **K**; only the case $B \equiv \Box C$ ($\Rightarrow$) needs special attention. Let $\Box C \notin w$. This time, consider the set $\{\Box D, D | \Box D \in w\} \cup \{\sim C\}$. The only additional fact needed in the argument to show that this set is **K4**-consistent is that $\Box D_i \vdash_{\mathbf{K4}} \Box \Box D_i$.

(**L**) Suppose $\nvdash_{\mathbf{L}} A$. Again proceed as before, except that now:

$$wR\,w' \iff \text{for all } \Box D \in w,\ \text{both } \Box D \in w' \text{ and } D \in w' \text{ and, for some } \Box C \in w',\ \Box C \notin w.$$

The argument is as for **K4**; again, only the case $B \equiv \Box C$ ($\Longleftarrow$) merits some special attention. This time, consider $\{\Box D, D | \Box D \in w\} \cup \{\Box C, {\sim}C\}$. Note that the inclusion of $\Box C$ will insure that w' really is a successor of w. The argument that this set is consistent now rests on the fact that, if $\Box D_1, \ldots, \Box D_k \vdash_{\mathbf{L}} \Box(\Box C \to C)$, then $\Box D_1, \ldots, \Box D_k \vdash_{\mathbf{L}} \Box C$. $\dashv$

2.5. Definition. If Φ is an adequate set of formulas, then the rooted Kripke-model $\mathbf{M}$ is Φ-*sound* if in the root w of $\mathbf{M}$, $w \Vdash \Box B \to B$ for each $\Box B$ in Φ; $\mathbf{M}$ is A-*sound* if $\mathbf{M}$ is Φ-sound for the smallest adequate set Φ containing A.

If K is a Kripke-model with root w, then the *derived model* K' of K is constructed by adding a new root w' below w and giving w' exactly the same forcing relation as w for the atoms.

2.6. Lemma. *If K is a Φ-sound Kripke-model with root w, then w' forces in K' exactly the same formulas from Φ as w in K.*

Proof. Let K be a rooted Φ-sound Kripke-model with root w. We prove by induction on the length of A that $w' \Vdash A$ iff $w \Vdash A$. This is so by definition for the atomic formulas and otherwise obvious except for the $\Box$. If $w' \Vdash \Box A$, then $w \Vdash \Box A$, since $w\,R\,w'$. If $w \Vdash \Box A$, then not only for all w'' such that $w\,R\,w''$, $w'' \Vdash A$, but also, by the Φ-soundness of K, $w \Vdash A$. But then, for all w'' such that $w'\,R\,w''$, $w'' \Vdash A$, i.e., $w' \Vdash \Box A$. $\dashv$

2.7. Theorem. (Modal completeness of **S**) $\vdash_{\mathbf{S}} A$ *iff A is forced in the root of all A-sound* **L***-models.*

Proof. $\Longrightarrow$: Assume K is A-sound and root $w \nVdash A$. If we assume to get a contradiction that $\vdash_{\mathbf{S}} A$, then A is provable in **L** from k applications of the reflection scheme: $\Box B_1 \to B_1, \ldots, \Box B_k \to B_k$. Consider the model $K^{(k+1)}$ obtained from K by taking $k+1$ times the derived model starting with K. Each time that one takes the derived model, one or more of the $\Box B_i$ may change from being forced to not being forced (never the other way around). This implies, by the pigeon hole principle, that one of the times that one has taken the derived model $K^{(m)}$ $(0 \leqslant m \leqslant k+1)$ the forcing value of all the $\Box B_i$ remain the same. It is easy to check that, in the root of that model $K^{(m)}$ for that m, $\Box B_i \to B_i$ is forced for all $i \leqslant k$. By lemma 2.6, A is not forced however, and a contradiction has been reached.

$\Longleftarrow$: Assume $\nvdash_{\mathbf{S}} A$. Then a fortiori A is not derivable in **L** from the reflection principles for its boxed subformulas. The result then immediately follows by applying theorem 2.4. $\dashv$

An elegant formulation of the semantics of **S** in terms of infinite models, so-called *tail models* is given in Visser [1984].

The well-known normal modal system **S4** that is obtained by adding the scheme $\Box A \to A$ to **K4** plays a role in section 10. It can be shown that **S4** is modally complete with respect to the (finite) reflexive, transitive Kripke-models.

3. Proof of Solovay's theorems

We rely mostly on Buss's Chapter II of this Handbook. One can find there an *intensional* arithmetization of metamathematics worked out, the (Hilbert-Bernays)-Löb derivability conditions are given and proofs of the diagonalization lemma, Gödel's theorems and Löb's theorem are presented. An additional fact that we need is some formalization of the recursion theorem.

To repeat the statement of Solovay's first arithmetic completeness theorem (theorem 1.1), for Σ_1-sound r.e. theories T containing $\mathbf{I\Sigma_1}$:

$\vdash_{\mathbf{L}} A$ iff $T \vdash A^*$ for all arithmetic realizations *.

Proof of theorem 1.1. $\Rightarrow$: These are just the Hilbert-Bernays-Löb conditions and Löb's theorem (see Chapter II of this Handbook).
$\Leftarrow$: What we have to do is show that, if $\nvdash_{\mathbf{L}} A(p_1 \ldots, p_k)$, then there exist $\alpha_1, \ldots, \alpha_k$ such that $T \nvdash A^*$, where * denotes the realization generated by mapping $p_1, \ldots, p_k$ to $\alpha_1, \ldots, \alpha_k$.[1] Suppose $\nvdash_{\mathbf{L}} A$. Then, by theorem 2.4, there is a finite **L**-model $\langle W, R, \Vdash \rangle$ in which A is not valid. We may assume that $W = \{1, \ldots, l\}$, 1 is the root, and $1 \nVdash A$. We define a new frame $\langle W', R' \rangle$:

$$W' = W \cup \{0\},$$
$$R' = R \cup \{(0, w) \mid w \in W\}.$$

Observe that $\langle W', R' \rangle$ is a finite **L**-frame.

We are going to embed this frame into T by means of a function $h : \omega \to W'$ (with ω the nonnegative integers) and sentences Lim_w, for each $w \in W'$, which assert that w is the limit of h. This function will be defined in such a way that a basic lemma 3.2 holds about the statements that T can prove about the sentences Lim_w. These statements are tailored to prove the next lemma 3.3 that expresses that provability in T behaves for the relevant formulas on the Kripke-model in the same way as the modal operator $\Box$. This will allow us to conclude the proof.

3.2. Lemma.
(a) *T proves that h has a limit in W', i.e., $T \vdash \bigvee \{\mathrm{Lim}_r \mid r \in W'\}$,*
(b) *If $w \neq u$, then $T \vdash \neg (\mathrm{Lim}_w \wedge \mathrm{Lim}_u)$,*
(c) *If $w R' u$, then $T + \mathrm{Lim}_w$ proves that $T \nvdash \neg \mathrm{Lim}_u$,*
(d) *If $w \neq 0$ and not $w R' u$, then $T + \mathrm{Lim}_w$ proves that $T \vdash \neg \mathrm{Lim}_u$,*
(e) *Lim_0 is true,*
(f) *For each $i \in W'$, Lim_i is consistent with T.*

[1] We will use italic capital letters for modal-logical formulas and Greek letters for arithmetic sentences and formulas, except that we will use Roman letters for descriptive names like "Proof".

We now define a realization * by setting for each propositional letter p_i,

$$p_i^* = \bigvee\{\mathrm{Lim}_w \mid w \in W,\ w \Vdash p_i\}.$$

This p_i^* will then function as the above-mentioned α_i.

3.3. Lemma. *For any $w \in W$ and any $\mathbf{L}$-formula B,*

(a) *if $w \Vdash B$, then $T + \mathrm{Lim}_w \vdash B^*$,*

(b) *if $w \nVdash B$, then $T + \mathrm{Lim}_w \vdash \neg B^*$.*

Proof. By induction on the complexity of B. If B is atomic, then clause (a) is evident, and clause (b) is also clear in view of lemma 3.2(b). The cases when B is a Boolean combination are straightforward. So, only the case that B is $\Box C$ will have to be considered.

(a) Assume that $w \Vdash \Box C$. Then, for each $w' \in W$ with $w\,R\,w'$, $w' \Vdash C$. By induction hypothesis, for each such w', $T + \mathrm{Lim}_{w'} \vdash C^*$, and this fact is then provable in T. On the other hand, by lemma 3.2(a) (proved in T itself) and (c), $T + \mathrm{Lim}_w$ proves that $T \vdash \bigvee\{\mathrm{Lim}_{w'} \mid w\,R'\,w'\}$. Together this implies that T proves that $T \vdash C^*$, i.e., $T \vdash (\Box C)^*$.

(b) Assume that $w \nVdash \Box C$. Then, for some $w' \in W$ with $w\,R'\,w'$, $w' \nVdash C$. By induction hypothesis, $T + \mathrm{Lim}_{w'} \vdash \neg C^*$, i.e., $T \vdash C^* \to \neg\mathrm{Lim}_{w'}$. By the second HBL-condition, $T \vdash (\Box C)^* \to \mathrm{Pr}_T(\neg\mathrm{Lim}_w)$. But lemma 3.2(c) implies that $T + \mathrm{Lim}_w \vdash \neg\mathrm{Pr}_T(\neg\mathrm{Lim}_w)$, i.e., $T + \mathrm{Lim}_w \vdash \neg(\Box C)^*$. $\dashv$

Observe by the way that lemma 3.3 expresses that $T + \mathrm{Lim}_w \vdash$ "$w \Vdash B$" $\leftrightarrow B^*$. Assuming lemma 3.2 we can now complete the proof of theorem 1.1. By the construction of the Kripke-model, $1 \Vdash \neg A$. By lemma 3.3, $T + \mathrm{Lim}_1 \vdash \neg A^*$. Since, by lemma 3.2(f), $T + \mathrm{Lim}_1$ is consistent, $T \nvdash A^*$. $\dashv$

Our remaining duties are to define the function h and to prove lemma 3.2. The recursion theorem enables us to define this function simultaneously with the sentences Lim_w (for each $w \in W'$), which, as we have mentioned already, assert that w is the limit of h.

3.4. Definition. (Solovay function h)
We define $h(0) = 0$.
If x is the code of a proof in T of $\neg\mathrm{Lim}_w$ for some w with $h(x)\,R\,w$, then $h(x+1) = w$.
Otherwise, $h(x+1) = h(x)$.

It is not hard to see that h is primitive recursive.

Proof of lemma 3.2. In each case below, except in (e) and (f), we reason in T.

(a) By induction on the nodes. For end nodes w (i.e., the ones with no R-successors), it can be proved that $T \vdash \forall x\,(h(x) = \bar{w} \to \forall y \geqslant x\, h(y) = \bar{w})$ by induction on x, and hence $T \vdash \exists x\, h(x) = \bar{w} \to \mathrm{Lim}_w$. Next, it is easily seen that, if for all successors w' of a node w, $T \vdash \exists x\, h(x) = \bar{w}' \to \bigvee\{\mathrm{Lim}_{w''} \mid w' = w'' \vee w'\,R\,w''\}$, then

$T \vdash \exists x\, h(x) = \bar{w} \to \bigvee\{\mathrm{Lim}_{w'} \mid w = w' \vee w\, R\, w'\}$. Therefore, this will hold for $w = 0$, which implies (a).

(b) Clearly h cannot have two different limits w and u.

(c) Assume w is the limit of h and $w\, R'\, u$. Let n be such that for all $x \geqslant n$, $h(x) = w$. We need to show that $T \nvdash \neg\, \mathrm{Lim}_u$. Deny this. Then, since every provable formula has arbitrarily long proofs, there is $x \geqslant n$ such that x codes a proof of $\neg \mathrm{Lim}_u$; but then, according to definition 3.4, we must have $h(x+1) = u$, which, as $u \neq w$ (by irreflexivity of R'), is a contradiction.

(d) Assume $w \neq 0$, w is the limit of h and not $wR'\, u$. If $u = w$, then (since $w \neq 0$) there exists an x such that $h(x+1) = w$ and $h(x) \neq w$. Then x codes a proof of $\neg\, \mathrm{Lim}_w$ and $\neg\, \mathrm{Lim}_w$ is provable. Next suppose $u \neq w$. Let us fix a number z with $h(z) = w$. Since h is primitive recursive, T proves that $h(z) = w$. Now argue in $T + \mathrm{Lim}_u$: since u is the limit of h and $h(z) = w \neq u$, there is a number x with $x \geqslant z$ such that $h(x) \neq u$ and $h(x+1) = u$. This contradicts the fact that not $(w =)h(z)R'\, u$, Thus, $T + \mathrm{Lim}_u$ is inconsistent, i.e., $T \vdash \neg\, \mathrm{Lim}_u$.

(e) By (a), as T is sound, one of the Lim_w for $w \in W'$ is true. Since for no w do we have $wR'w$, (d) means that each Lim_w, except Lim_0, implies in T its own T-disprovability and therefore is false. Consequently, Lim_0 is true.

(f) By (e), (c) and the soundness of T. ⊣

To repeat the statement of Solovay's second arithmetic completeness theorem (theorem 1.2):

$\vdash_{\mathbf{S}} A$ iff $\mathbb{N} \vDash A^*$ for all arithmetic realizations *.

Proof of theorem 1.2. Since the $\Box A \to A$ are reflection principles and these are true for a sound theory, the soundness part is clear. So, assume $\nvdash_{\mathbf{S}} A$. Modal completeness of **S** then provides us with an A-sound (see definition 2.5) model in which A is not forced in the root. We can repeat the procedure of the proof of the first completeness theorem, but now directly to the model itself (which we assume to have a root 0) without adding a new root, and again prove lemmas 3.2 and 3.3. But this time we have forcing also for 0 and we can improve lemma 3.3 to apply it also to $w = 0$, at least for subformulas of A.

The proof of the (b)-part of that lemma can be copied. With respect to the (a)-part, restricting again to the $\Box$-case, assume that $0 \Vdash \Box C$. Then, for each $w \in W$ with $w \neq 0$, $w \Vdash C$. But now, by the A-soundness of the model, C is also forced in the root 0. By the induction hypothesis, for all $w \in W$, $T + \mathrm{Lim}_w \vdash C^*$. By lemma 3.2(a) then $T \vdash C^*$, so, $T \vdash (\Box C)^*$ and hence $T \vdash \mathrm{Lim}_0 \to (\Box C)^*$.

Applying the strengthened version of lemma 3.3 to $w = 0$ and A, we obtain $T \vdash \mathrm{Lim}_0 \to \neg\, A^*$, which suffices, since Lim_0 is true (lemma 3.2). ⊣

4. Fixed point theorems

For the provability logic **L** a fixed point theorem can be proved. One can view Gödel's diagonalization lemma as stating that in arithmetic theories the formula $\neg \Box p$

has a fixed point: the Gödel sentence. Gödel's proof of his second incompleteness theorem effectively consisted of the fact that the sentence expressing consistency, the arithmetic realization of $\neg\Box\bot$, is provably equivalent to this fixed point. Actually this fact is provable from the principles codified in the provability logic **L**, which means then that it can actually be presented as a fact about **L**. This leads to a rather general fixed point theorem, which splits into a uniqueness and an existence part. It concerns formulas A with a distinguished propositional variable p that only occurs *boxed* in A, i.e., each occurrence of p in A is part of a subformula $\Box B$ of A.

4.1. Theorem. (Uniqueness of fixed points) *If p occurs only boxed in $A(p)$ and q does not occur at all in $A(p)$, then $\vdash_{\mathbf{L}} \boxdot((p \leftrightarrow A(p)) \wedge (q \leftrightarrow A(q)) \rightarrow (p \leftrightarrow q)$.*

4.2. Corollary. *If p occurs only boxed in $A(p)$, and both $\vdash_{\mathbf{L}} B \leftrightarrow A(B)$ and $\vdash_{\mathbf{L}} C \leftrightarrow A(C)$, then $\vdash_{\mathbf{L}} B \leftrightarrow C$.*

4.3. Theorem. (Existence of fixed points) *If p occurs only boxed in $A(p)$, then there exists a formula B, not containing p and otherwise containing only variables of $A(p)$, such that $\vdash_{\mathbf{L}} B \leftrightarrow A(B)$.*

After the original proofs by de Jongh and Sambin (see Sambin [1976], Smoryński [1978,1985], and, for the first proof of uniqueness, Bernardi [1976]) many other, different, proofs have been given for the fixed point theorems, syntactical as well as semantical ones, the latter e.g., in Gleit and Goldfarb [1990]. It is also worthwhile to remark that theorem 4.3 follows from theorem 4.1 (which can be seen as a kind of implicit definability theorem) by way of Beth's definability theorem that holds for **L**. The latter can be proved from interpolation in the usual manner. Interpolation can be proved semantically in the standard manner via a kind of Robinson's consistency lemma (see Smoryński [1978]), and syntactically in the standard manner by cut-elimination in a sequent calculus formulation of **L** (Sambin and Valentini [1982,1983]).

In an important sense the meaning of the fixed point theorem is negative, namely in the sense that, if in arithmetic one attempts to obtain formulas with essentially new properties by diagonalization, one will not get them by using instantiations of purely propositional modal formulas (except once of course: the Gödel sentence, or the sentence Löb used to prove his theorem). That is one reason that interesting fixed points often use Rosser-orderings (see section 9).

5. Propositional theories and Magari-algebras

A *propositional theory* is a set of modal formulas (usually in a finite number of propositional variables) which is closed under modus ponens and necessitation, but not necessarily under substitution.

We say that such a theory is *faithfully interpretable* in **PA**, if there is a realization * such that $T = \{A \mid \mathbf{PA} \vdash A^*\}$. (This is an adaptation of definition 11.1 to the modal propositional language.) Each sentence α of **PA** generates a propositional theory

which is faithfully interpretable in **PA**, namely $Th_\alpha = \{A(p) \mid \mathbf{PA} \vdash A^*(\ulcorner \alpha \urcorner)\}$. Of course, this theory is closed under **L**-derivability: it is an **L**-*propositional theory*. A question much wider than the one discussed in the previous sections is, which **L**-propositional theories are faithfully interpretable in **PA** and other theories. This question was essentially solved by Shavrukov [1993b]:

5.1. Theorem. *An r.e.* **L**-*propositional theory T is faithfully interpretable in* **PA** *iff T is consistent and satisfies* the strong disjunction property (*i.e., $\Box A \in T$ implies $A \in T$, and $\Box A \vee \Box B \in T$ implies $\Box A \in T$ or $\Box B \in T$*).

Note that faithfully interpretable theories in a finite number of propositional variables are necessarily r.e. The theorem was given a more compact proof and at the same time generalized to all r.e. theories extending $\mathbf{I\Delta_0} + \mathrm{EXP}$ by Zambella [1994]. If one applies the theorem to the minimal **L**-propositional theory, an earlier proved strengthening of Solovay's theorem (Artëmov [1980], Avron [1984], Boolos [1982], Montagna [1979], Visser [1980]) rolls out.

5.2. Corollary. (Uniform arithmetic completeness theorem) *There exists a sequence of arithmetic sentences $\alpha_0, \alpha_1, \ldots$ such that, for any n and modal formula $A(p_0, \ldots, p_n)$, $\vdash_{\mathbf{L}} A$ iff, under the arithmetic realization * induced by setting $p_0^* = \alpha_0, \ldots, p_n^* = \alpha_n$, A^* is provable in* **PA**.

Sets of modal formulas that are the true sentences under some realization are closed under modus ponens, but not necessarily under necessitation; such sets are generally not propositional theories in the above sense. Let us call a set T of modal formulas *realistic* if there exists a realization * such that A^* is true, for every $A \in T$. Moreover, we say that T is *well-specified* if, whenever $A \in T$ and B is a subformula of A, we also have $B \in T$ or $\neg B \in T$. Strannegård [1997] proves a result that generalizes both theorem 5.1 and Solovay's second arithmetic completeness theorem. We give a weak but easy to state version of it.

5.3. Theorem. *Let T be a well-specified r.e. set of modal formulas. Then T is realistic iff T is consistent with* **S**.

An even more general point of view than propositional theories is to look at the Boolean algebras of arithmetic theories with one additional operator representing formalized provability and, more specifically, at the ones generated by a sequence of sentences in the algebras of arithmetic. The algebras can be axiomatized equationally and are called *Magari-algebras* (after the originator R. Magari) or *diagonalizable algebras*. Of course, theorem 5.1 can be restated in terms of Magari-algebras. Shavrukov proved two beautiful and essential additional results concerning the Magari-algebras of formal theories that cannot naturally be formulated in terms of propositional theories.

5.4. Theorem. (Shavrukov [1993a]) *The Magari algebras of* **PA** *and* **ZF** *are not isomorphic, and, in fact not elementarily equivalent* (Shavrukov [1997]).

The proof only uses the fact that **ZF** proves the uniform Σ_1-reflection principle for **PA**. A corollary of the theorem is that there is a formula of the second order propositional calculus that is valid in the interpretation with respect to **PA**, but not in the one with respect to **ZF**. Beklemishev [1996b] gives a different kind of example of such a formula for the two theories **PA** and $\mathbf{I}\Delta_0 + \mathrm{EXP}$.

5.5. Theorem. (Shavrukov [1997]) *The first order theory of the Magari algebra of* **PA** *is undecidable.*

Japaridze [1993] contains some moderately positive results on the decidability of certain fragments of (a special version of) this theory.

6. The extent of Solovay's theorems

An important feature of Solovay's theorems is their remarkable stability: a wide class of arithmetic theories and their provability predicates enjoys one and the same provability logic **L**. Roughly, there are three conditions sufficient for the validity of Solovay's results: the theory has to be (a) sufficiently strong, (b) recursively enumerable (a provability predicate satisfying Löb's derivability conditions is naturally constructed from a recursive enumeration of the set of its axioms), and (c) sound. Let us see what happens if we try to do without these conditions. The situation is fully investigated only w.r.t. the soundness condition.

Consider an arbitrary arithmetic r.e. theory T containing **PA** and a Σ_1 provability predicate $\mathrm{Pr}_T(x)$ for T. *Iterated consistency assertions* for T are defined as follows:

$$\mathrm{Con}^0(T) := \top; \quad \mathrm{Con}^{n+1}(T) := \mathrm{Con}(T + \mathrm{Con}^n(T)),$$

where, as usual, $\mathrm{Con}(T + \varphi)$ stands for $\neg\,\mathrm{Pr}_T(\ulcorner \neg\, \varphi \urcorner)$. In other words, $\mathrm{Con}^n(T)$ is (up to provable equivalence) the unique arithmetic realization of the modal formula $\neg\,\Box^n\bot$. We say that T is of *height* n if $\mathrm{Con}^n(T)$ is true and $\mathrm{Con}^{n+1}(T)$ is false in the standard model. If no such n exists, we say that T has *infinite height*.

In a sense, theories of finite height are close to being inconsistent and therefore can be considered as a pathology. The inconsistent theory is the only one of height 0. All Σ_1-sound theories have infinite height, but there exist Σ_1-unsound theories of infinite height. The theory $T + \neg\,\mathrm{Con}^n(T)$ is of height n, if T has infinite height. Moreover, for each consistent but Σ_1-unsound theory T and each $n > 0$, one can construct a provability predicate for T such that T is precisely of height n with respect to this predicate (Beklemishev [1989a]).

Let us call *the provability logic of* T the set of all modal formulas A such that $T \vdash (A)^*_T$, for all arithmetic realizations * with respect to Pr_T. The *truth provability logic of* T is the set of all A such that $(A)^*_T$ is true in the standard model, for all realizations *.

6.1. Theorem. (Visser [1981]) *For an r.e. arithmetic theory T containing* **PA**, *the provability logic of T coincides with*

1. **L**, *if T has infinite height,*
2. $\{A \mid \Box^n\bot \vdash_{\mathbf{L}} A\}$, *if T is of height $0 \leqslant n < \infty$.*

Proof. By Solovay's construction, using the fact that the formula $\Box^n\bot$ is valid in Kripke-models of height $< n$, and only in such models. ⊣

Generalization of Solovay's second theorem is more interesting. To formulate it, we first introduce a convenient notation. For a set of modal formulas X, let $\mathbf{L}X$ denote the closure under modus ponens and substitution of the set X together with all theorems of **L**. In this notation, Solovay's logic **S** is the same as $\mathbf{L}\{\Box A \to A\}$. The following two logics have been introduced by respectively Artëmov [1980] and Japaridze [1986,1988b] with different provability interpretations in mind (see the next section):

$$\begin{aligned} \mathbf{A} &:= \mathbf{L}\{\neg\,\Box^n\bot \mid n \in \mathbb{N}\} \\ \mathbf{D} &:= \mathbf{L}\{\neg\,\Box\bot,\ \Box(\Box A \vee \Box B) \to (\Box A \vee \Box B)\} \end{aligned}$$

Obviously, $\mathbf{A} \subset \mathbf{D} \subset \mathbf{S}$. The following theorem gives an exhaustive description of all truth provability logics.

6.2. Theorem. (Beklemishev [1989a]) *For an r.e. arithmetic theory T containing* **PA**, *the truth provability logic for T coincides with*

1. **S** *iff T is sound,*
2. **D** *iff T is Σ_1-sound but not sound,*
3. **A** *iff T has infinite height but is not Σ_1-sound,*
4. $\mathbf{L}\{\Box^{n+1}\bot, \neg\,\Box^n\bot\}$ *iff T is of height $0 \leqslant n < \infty$.*

It is known that, at least in some natural cases, the other two sufficient conditions can also be considerably weakened. Boolos [1979] shows that the non-r.e. predicate of *ω-provability* (dual to ω-consistency) over Peano arithmetic has precisely the same provability logic as Peano arithmetic itself, i.e., **L**. The same holds for the natural formalization of the Σ_{n+1}-complete predicate *"to be provable in* **PA** *together with all true Π_n-sentences"* (Smoryński [1985]). Solovay (see Boolos [1993b]) showed that **L** is also the logic of the Π^1_1-complete predicate of provability in analysis together with the ω-rule. However, no results to the effect that Solovay's theorems hold for broad classes of non-r.e. predicates are known. On the other hand, Solovay found an axiomatization of the provability logic of the predicate "to be valid in all transitive models of **ZF**", which happens to be a proper extension of **L** (see Boolos [1993b]).

If one is somewhat careful, Solovay's construction can be adapted to show that Solovay's theorems hold for all (r.e., sound) extensions of $\mathbf{I\Delta_0} + \mathrm{EXP}$ (de Jongh, Jumelet and Montagna [1991]). The two theorems formulated earlier in this section can be similarly generalized. However, the most intriguing question, whether

Solovay's theorems hold for essentially weaker theories, such as Buss's S^1_2 or even S_2, remains open. This problem was thoroughly investigated by Berarducci and Verbrugge [1993], where the authors, in a modification of the Solovay construction, only succeeded in embedding particular kinds of Kripke-models into bounded arithmetic. The main technical difficulty lies in the fact that Solovay's construction, in its known variations, presupposes (at least, sentential) provable $\exists\Pi^b_1$-completeness of the theory in question. This property happens to fail for bounded arithmetic under some reasonable complexity-theoretic assumption (Verbrugge [1993a]). As far as we know, it is not excluded that the answer to the question what is the provability logic of S^1_2 also depends on difficult open problems in complexity theory.

Solovay's theorem does not hold in its immediately transposed form for Heyting's arithmetic **HA** (the intuitionistic pendant of **PA**). The logic of the provability predicate of **HA** with regard to **HA** certainly contains additional principles beyond the obvious intuitionistic version of **L**: the intuitionistic propositional calculus **IPC** plus $\Box(\Box A \to A) \to \Box A$. The situation has been discussed by Visser in several papers (Visser [1985,1994], Visser et al. [1995]). It is unknown what the real logic is, for all we know it may even be complete Π^0_2. In any case it contains the additional principles:

- $\Box\neg\neg\Box A \to \Box\Box A$
- $\Box(\neg\neg\Box A \to \Box A) \to \Box\Box A$
- $\Box(A \vee B) \to \Box(A \vee \Box B)$ (Leivant's Principle[2])

But this is not an exhaustive list. It is well possible that the logic of the binary operator of Σ-*preservativity* over **HA** is better behaved than the logic of provability on its own: **PA**$+A$ Σ-preserves **PA**$+B$, if from each Σ_1-sentence from which A follows in **PA**, B is also **PA**-derivable. In classical systems Σ-preservativity is definable in terms of Π_1-conservativity (see sections 12 to 14 for that concept) and vice versa, but constructively this is the proper version to study (see also Visser [1997]).

7. Classification of provability logics

One of the important methodological consequences of Gödel's second incompleteness theorem is the fact that, in general, it is necessary to distinguish between a *theory* T under study, and a *metatheory* U in which one reasons about the properties of T. Perhaps, the most natural choice of U is the full true arithmetic **TA**, the set of all formulas valid in the standard model, yet this is not the only possibility. Other meaningful choices could be T itself, or the reader's favorite minimal fragment of arithmetic, e.g., $\mathbf{I\Sigma_1}$. The separate role of the metatheory is emphasized in the definition of *provability logic of a theory* T *relative to a metatheory* U that was suggested independently by Artëmov [1980] and Visser [1981].

[2]added as one of the "stellingen" (theses) to Leivant's Ph.D. thesis, Amsterdam, 1979

Let T and U be arithmetic theories extending $\mathbf{I\Delta_0} + \mathrm{EXP}$, T r.e. and U not necessarily r.e. The provability logic of T relative to, or simply *at*, U is the set of all modal formulas φ such that $U \vdash (\varphi)^*_T$, for all arithmetic realizations * (denoted $\mathsf{PRL}_T(U)$). Intuitively, $\mathsf{PRL}_T(U)$ expresses those principles of provability in T that can be verified my means of U. As a set of modal formulas, $\mathsf{PRL}_T(U)$ contains $\mathbf{L}$ and is closed under modus ponens and substitution, i.e., is a (not necessarily normal) modal logic extending $\mathbf{L}$.

Solovay's theorems can be restated as saying that, if T is a sound theory, then $\mathsf{PRL}_T(T) = \mathbf{L}$ and $\mathsf{PRL}_T(\mathbf{TA}) = \mathbf{S}$. A modal logic is called *arithmetically complete*, if it has the form $\mathsf{PRL}_T(U)$, for some T and U. The problem of obtaining a reasonable general characterization of arithmetically complete modal logics has become known as 'the classification problem', and was one of the early motivating problems for the Moscow school of provability logic founded by Artëmov.

The solution to this problem is the joint outcome of the work of several authors Artëmov [1980], Visser [1981,1984], Artëmov [1985b], Japaridze [1986,1988b], Beklemishev [1989a]. Artëmov [1980] (applying the so-called *uniform* version of Solovay's theorem, corollary 5.2) showed that all logics of the form $\mathbf{L}X$, for any set X of variable-free modal formulas, are arithmetically complete. In Artëmov [1985b], he showed that such extensions are exhausted by the following two specific families of logics:

$$\mathbf{L}_\alpha := \mathbf{L}\{F_n \mid n \in \alpha\},$$
$$\mathbf{L}^-_\beta := \mathbf{L}\{\bigvee_{n \notin \beta} \neg F_n\},$$

where $\alpha, \beta \subseteq \mathbb{N}$, β is cofinite, and F_n denotes the formula $\Box^{n+1}\bot \to \Box^n\bot$. Some of the provability logics introduced above have this form: $\mathbf{L} = \mathbf{L}_\emptyset$, $\mathbf{A} = \mathbf{L}_\mathbb{N}$, $\mathbf{L}\{\Box^{n+1}\bot, \neg\,\Box^n\bot\} = \mathbf{L}^-_{\mathbb{N}\setminus\{n\}}$.

The families $\mathbf{L}_\alpha$ and $\mathbf{L}^-_\beta$ are ordered by inclusion precisely as their indices, and $\mathbf{L}_\alpha$ is included in $\mathbf{L}^-_\alpha$ for cofinite α. Note that the logics $\mathbf{L}^-_\beta$ are not contained in $\mathbf{S}$, and therefore correspond to unsound metatheories U if the theory T is sound. Visser [1984] showed that these are the only arithmetically complete logics not contained in $\mathbf{S}$. Artëmov [1985b] improved this by actually reducing the classification problem to the interval between $\mathbf{A}$ and $\mathbf{S}$. Any arithmetically complete logic ℓ from this interval generates a family of different arithmetically complete logics of the form $\ell \cap \mathbf{L}^-_\beta$, for cofinite β, and Artëmov showed that such logics, together with the families $\mathbf{L}_\alpha$ and $\mathbf{L}^-_\beta$, exhaust all arithmetically complete ones.

Japaridze [1986,1988b] found a new provability logic within the interesting interval by establishing that $\mathsf{PRL}_\mathbf{PA}(\mathbf{PA} + \omega\text{-Con}(\mathbf{PA})) = \mathbf{D}$, where ω-Con($\mathbf{PA}$) denotes the formalized ω-consistency of $\mathbf{PA}$. The final step was made by Beklemishev [1989a], who showed that $\mathbf{D}$ is the only arithmetically complete modal logic within the interval between $\mathbf{A}$ and $\mathbf{S}$, thus completing the classification. This result was also crucial for the proof of theorem 6.2 of the previous section. We denote $\mathbf{S}_\beta := \mathbf{S} \cap \mathbf{L}^-_\beta$, $\mathbf{D}_\beta := \mathbf{D} \cap \mathbf{L}^-_\beta$ and formulate the resulting theorem.

7.1. Theorem. (Classification Theorem, Beklemishev [1989a]) *The arithmetically complete modal logics are exhausted by the four families:* $\mathbf{L}_\alpha$, $\mathbf{L}^-_\beta$, $\mathbf{S}_\beta$, $\mathbf{D}_\beta$, *for* $\alpha, \beta \subseteq \mathbf{N}$, β *cofinite.*

From a purely modal-logical point of view, the meaning of the classification theorem is that only very few extensions of $\mathbf{L}$ are arithmetically complete. The word 'few' must not be understood here in terms of cardinality, because the family $\mathbf{L}_\alpha$ has already the cardinality of the continuum, but rather less formally. E.g., there is a continuum of different modal logics containing $\mathbf{A}$ (Artëmov [1985b]), but only four of them are arithmetically complete. Similar observations hold for other natural intervals in the lattice of extensions of $\mathbf{L}$.

All arithmetically complete logics have nice axiomatizations, and are generally well-understood, although most of them are not normal. An adequate Kripke-type semantics is known for all arithmetically complete logics: for $\mathbf{L}_\alpha$ and $\mathbf{L}^-_\beta$ it can be formulated in terms of the height of the tree-like models for $\mathbf{L}$; the so-called *tail models* for $\mathbf{S}$ were suggested independently by Boolos [1981] and Visser [1984]; a similar kind of semantics for $\mathbf{D}$ was produced by Beklemishev [1989b]. A corollary is that all logics of the families $\mathbf{S}_\beta$, $\mathbf{D}_\beta$, and $\mathbf{L}^-_\beta$ are decidable, and a logic of the form $\mathbf{L}_\alpha$ is decidable, iff its index α is a decidable subset of $\mathbb{N}$, i.e., iff it has a decidable axiomatization.

The fact that arithmetically complete logics are scarce tells us that inference 'by arithmetic interpretation' considerably strengthens the usual modal-logical consequence relation. In fact, the classification theorem can be understood as a classification of modally expressible arithmetic schemata. Familiar examples of such schemata are: the *local reflection principle* for T, that is the schema $\mathrm{Pr}_T(\ulcorner A \urcorner) \to A$ for all arithmetic sentences A, which is expressed by the modal formula $\Box p \to p$; ω *times iterated consistency* of T, which is expressed by $\{\neg\, \Box^n \bot \mid n \in \mathbb{N}\}$; the *local* Σ_1*-reflection principle*, which can be expressed by the axioms of $\mathbf{D}$, etc. In general, a schema is *modally expressible* over a theory T, if it is deductively equivalent to the set of all arithmetic realizations with respect to T of a family of modal formulas. The classification theorem gives us a complete description of all modally expressible arithmetic schemata: they precisely correspond to axiomatizations of arithmetically complete modal logics.

It is very surprising that *all* such schemata are built up from instances of the local reflection principle, sometimes a little twisted by axioms of $\mathbf{L}^-_\beta$ type. This can be considered as a theoretical justification of the 'empirical' rule that in the study of provability all reasonable metatheories happen to be equivalent to some version of the reflection principle.

We round up the discussion of the classification theorem by giving some examples for natural pairs (theory, metatheory) of fragments of arithmetic.

$$\begin{aligned} \mathsf{PRL}_{\mathbf{I\Sigma_1}}(\mathbf{PA}) &= \mathbf{S}, \\ \mathsf{PRL}_{\mathbf{I\Sigma_1}}(\mathbf{I\Sigma_n}) &= \mathbf{D}, \quad \text{for } n > 1, \\ \mathsf{PRL}_{\mathbf{I\Delta_0}+\mathrm{EXP}}(\mathbf{PRA}) &= \mathbf{D}, \end{aligned}$$

$$\begin{aligned} \mathsf{PRL}_{\mathbf{I\Sigma_1}}(\mathbf{I\Sigma_1} + \mathrm{Con}(\mathbf{PA})) &= \mathbf{A}, \\ \mathsf{PRL}_{\mathbf{PRA}}(\mathbf{I\Sigma_1}) &= \mathbf{L}. \end{aligned}$$

All such results follow easily from the classification theorem and the usual proof-theoretic information about the provability of reflection principles for the theories in question. E.g., $\mathbf{I\Sigma_1} + \mathrm{Con}(\mathbf{PA})$ obviously contains ω-times-iterated consistency for $\mathbf{I\Sigma_1}$, but, being a finite Π_1-axiomatized extension of $\mathbf{I\Sigma_1}$, cannot contain the local Σ_1-reflection principle for $\mathbf{I\Sigma_1}$ (by Löb's theorem). Hence, $\mathsf{PRL}_{\mathbf{I\Sigma_1}}(\mathbf{I\Sigma_1} + \mathrm{Con}(\mathbf{PA}))$ contains $\mathbf{A}$ but does not contain $\mathbf{D}$. The classification theorem implies that, in this case, it must be $\mathbf{A}$.

8. Bimodal and polymodal provability logics

The fact that all reasonable theories have one and the same — Löb's — provability logic is, in a sense, a drawback: it means that the provability logic of a theory cannot distinguish between most of the interesting properties of theories, such as e.g., finite axiomatizability, reflexivity, etc. In fact, by Visser's theorem 6.1, the only recognizable characteristic of a theory is its height, and the situation does not become much better even if one considers truth provability logics.

One obvious way to increase the expressive power of the modal language is to consider provability operators in several different theories simultaneously, which naturally leads to *bi-* and *polymodal provability logic*. It turns out that the modal description of the joint behaviour of two or more provability operators is, in general, a considerably more difficult task than the calculation of unimodal provability logics. There is no single system that can justifiably be called *the* bimodal provability logic — rather, we know particular systems for different natural pairs of provability operators, and none of those systems occupies any privileged place among the others. Moreover, the numerous isolated results accumulated in this area, so far, give us no clue as to a possible general classification of bimodal provability logics for pairs of sound r.e. theories. This problem remains one of the most challenging open problems in provability logic. A short survey of the state of our knowledge in this field is given below.

The language $\mathcal{L}(\Box, \triangle)$ of bimodal provability logic is obtained from that of propositional calculus by adding two unary modal operators $\Box$ and $\triangle$. Let (T, U) be a pair of arithmetic r.e. theories, taken together with some fixed canonical Σ_1 provability predicates Pr_T and Pr_U. An *arithmetic realization* $(\cdot)^*_{T,U}$ *with respect to* (T, U) is a mapping of modal formulas to arithmetic sentences that commutes with Boolean connectives and translates $\Box$ as provability in T and $\triangle$ as that in U:

$$(\Box A)^*_{T,U} = \mathrm{Pr}_T(\ulcorner (A)^*_{T,U} \urcorner), \qquad (\triangle A)^*_{T,U} = \mathrm{Pr}_U(\ulcorner (A)^*_{T,U} \urcorner).$$

The provability logic for (T, U), denoted $\mathsf{PRL}_{T,U}$, is the collection of all $\mathcal{L}(\Box, \triangle)$-formulas A such that $T \vdash (A)^*_{T,U}$ and $U \vdash (A)^*_{T,U}$, for every arithmetic realization *. In general, as in the unimodal case, one can consider bimodal provability logics

for (T,U) relative to an arbitrary metatheory V (where $\mathsf{PRL}_{T,U}$ corresponds to $V = T \cap U$).

Not too much can a priori be said about $\mathsf{PRL}_{T,U}$, for arbitrary T and U. Clearly, $\mathsf{PRL}_{T,U}$ is closed under modus ponens, substitution and the $\Box$- and $\triangle$-necessitation rules. Moreover, $\mathsf{PRL}_{T,U}$ has to be a (normal) extension of the bimodal system **CS**, given by the axioms and rules of **L** formulated separately for $\Box$ and $\triangle$, and by the obvious mixed principles:

$$\Box A \to \triangle \Box A, \qquad \triangle A \to \Box \triangle A.$$

By Solovay's theorem we know that, whenever both T and U have infinite height, the fragment of $\mathsf{PRL}_{T,U}$ in the language $\mathcal{L}(\Box)$ of $\Box$ alone, as well as the one in the language of $\triangle$, actually coincides with **L**. Using the uniform version of Solovay's theorem, Smoryński [1985] showed that **CS** is the minimal bimodal provability logic, i.e., it coincides with $\mathsf{PRL}_{T,U}$ for a certain pair of finite extensions T, U of Peano arithmetic. Beklemishev [1992] showed that there is even a pair of provability predicates for Peano arithmetic itself for which the corresponding bimodal provability logic coincides with **CS**. Such predicates can be called *independent* in the sense that they 'know' as little about each other as is possible in principle. It should be noted however that, neither the theories in Smoryński's example, nor the independent provability predicates are natural — they are constructed by a tricky diagonalization. Thus, we are in the interesting situation that the bimodal logic **CS**, which structurally occupies a privileged place among the provability logics, does not correspond to any known *natural* pair of theories.

Deeper structural information on bimodal provability logics is provided by the Classification Theorem 7.1 for arithmetically complete modal logics. With every bimodal logic ℓ we can associate its *type*:

$$(\ell)^0 := \{A \in \mathcal{L}(\Box) \mid \ell \vdash \triangle A\}.$$

An easy analysis then shows that $(\cdot)^0$ surjectively maps normal extensions of **CS** onto the lattice of the unimodal logics containing **L**. Under the assumption of Σ_1-soundness of $T \cap U$ we obviously have:

$$\mathsf{PRL}_T(U) = (\mathsf{PRL}_{T,U})^0.$$

The classification theorem not only shows that not every type (of unimodal logic) is materialized as that of a bimodal provability logic, but also gives us a complete description of all such possible types.

Besides the general observations above, a number of particular bimodal provability logics for natural pairs of theories are known. These logics cover most of the examples of pairs of arithmetic theories that come to mind, but, unfortunately, are far from being an exhaustive list of all bimodal provability logics.

The best known system is the logic $\mathsf{PRL}_{\mathbf{PA},\mathbf{ZF}}$ discovered by Carlson [1986], and independently (with a different interpretation in mind) by Montagna [1987]. This

logic can be axiomatized over **CS** by the principle of *essential reflexivity*

$$\triangle(\Box A \to A).$$

It is the *only* bimodal provability logic of type **S** and a maximal one among the bimodal logics for pairs of sound theories. In other words, $\mathsf{PRL}_{T,U} = \mathsf{PRL}_{\mathbf{PA},\mathbf{ZF}}$, whenever the theories T, U are sound and U contains the local reflection principle for T.

Furthermore, we know two natural bimodal provability logics of type **D**, introduced by Beklemishev [1996a]. The first one corresponds to pairs of theories (T, U) such that U is a finite extension of T that proves the local Σ_1-reflection principle for T. Typical examples are the pairs $(\mathbf{I\Delta_0} + \mathrm{EXP}, \mathbf{I\Delta_0} + \mathrm{SUPEXP})$, $(\mathbf{I\Sigma_m}, \mathbf{I\Sigma_n})$, for $n > m \geqslant 1$, etc. The logic can be axiomatized over **CS** by the *monotonicity* axiom $\Box A \to \triangle A$ and the schema

$$\triangle(\Box S \to S),$$

where S is an arbitrary (possibly empty) disjunction of formulas of the form $\Box B$ and $\triangle B$.[3]

The second one corresponds to Π_1-essentially reflexive (see definition 12.3) extensions of theories of bounded arithmetic complexity such as e.g., $(\mathbf{I\Delta_0} + \mathrm{EXP}, \mathbf{PRA})$, $(\mathbf{I\Sigma_n}, \mathbf{I\Sigma^R_{n+1}})$ for $n \geqslant 1$, where $\mathbf{I\Sigma^R_k}$ is defined like $\mathbf{I\Sigma_k}$ but with the induction for Σ_k-formulas formulated as a rule. The corresponding provability logic can be axiomatized over **CSM** by the Π_1-*essential reflexivity* schema

$$\triangle A \to \triangle(\Box(A \to S) \to S),$$

where S is as before.

We also know two natural provability logics of type **A** (Beklemishev [1994]). The first system corresponds to pairs of theories (T, U) such that U is an extension of T by finitely many Π_1-sentences and proves ω-times-iterated consistency of T, such as e.g., the pairs $(\mathbf{PA}, \mathbf{PA} + \mathrm{Con}(\mathbf{ZF}))$, $(\mathbf{I\Sigma_1}, \mathbf{I\Sigma_1} + \mathrm{Con}(\mathbf{I\Sigma_2}))$, etc. This logic can be axiomatized over **CSM** by the principle

$$\triangle A \to \Box(\triangle \bot \vee A), \tag{P}$$

valid for all Π_1-axiomatizable extensions of theories, together with the schema

$$\triangle \neg \Box^n \bot, \qquad n \geqslant 1.$$

The second system corresponds to reflexive Π_1-axiomatizable extensions of theories, such as e.g., $(\mathbf{PA}, \mathbf{PA} + \{\mathrm{Con}^n(\mathbf{PA}) \mid n \geqslant 1\})$, $(\mathbf{I\Sigma_1}, \mathbf{I\Sigma_1} + \{\mathrm{Con}(\mathbf{I\Sigma_n}) \mid n \geqslant 1\})$. It can be axiomatized over **CSM** plus (P) by the *reflexivity* axiom

$$\triangle A \to \triangle \Diamond A.$$

[3] In the following, **CS** together with the monotonicity axiom will be denoted **CSM**.

Finally, we know by Beklemishev [1996a] a natural system of type **L** that corresponds to finite extensions of theories of the form $(T,\, T + A)$, where both $T + \varphi$ and $T + \neg\, \varphi$ are conservative over T with respect to Boolean combinations of Σ_1-sentences. Examples of such pairs are $(\mathbf{PRA}, \mathbf{I\Sigma_1})$, $(\mathbf{I\Sigma_n^R}, \mathbf{I\Sigma_n})$, for $n \geqslant 1$, and others. The logic is axiomatized over **CSM** by the $\mathcal{B}(\Sigma_1)$-*conservativity* schema

$$\triangle B \to \Box B,$$

where B denotes an arbitrary Boolean combination of formulas of the form $\Box C$ and $\triangle C$.

The six bimodal logics described above essentially exhaust all nontrivial cases for which natural provability logics have explicitly been characterized. It is worth mentioning that all these systems are decidable, and a suitable Kripke-style semantics is known for each of them. Smoryński [1985] contains an extensive treatment of $\mathsf{PRL}_{\mathbf{PA},\mathbf{ZF}}$ including proofs of three arithmetic completeness theorems due to Carlson. These theorems are extended by Strannegård [1997] to the setting of r.e. sets of bimodal formulas (as discussed in section 5). Visser [1995] presents a beautiful approach to Kripke semantics for bimodal provability logics. Beklemishev [1994, 1996a] gives a detailed survey of the current state of the field.

Apart from describing the joint behaviour of two 'usual' provability predicates, each of them being separately well enough understood, bimodal logic has been successfully used for the analysis of some nonstandard, not necessarily r.e., concepts of provability. The systems emerging from such an analysis often have not so much in common with **CS**, although different 'bimodal analyses' do share common technical ideas.

As early as 1986, Japaridze [1986,1988b] characterized the bimodal logic of provability and ω-*provability* (dual to ω-consistency) in Peano arithmetic. Later his study was simplified and further advanced by Ignatiev [1993a] and Boolos [1993b,1993a], who, among other things, showed that the same system corresponds to some other, so-called *strong*, concepts of provability (taken jointly with the usual one). Other examples of strong provability predicates are the Σ_{n+1}-complete *provability from all true arithmetic* Π_n*-sentences*, for $n \geqslant 1$, and the Π_1^1-complete *provability under the* ω*-rule in analysis*.

Japaridze's bimodal logic can be axiomatized by the axioms and rules of **L**, formulated separately for $\Box$ and for $\triangle$, the monotonicity principle $\Box A \to \triangle A$, and an additional Π_1-*completeness* principle

$$\Diamond A \to \triangle \Diamond A,$$

which reflects in so far as that is possible that $\triangle$ is strong enough to prove all true Π_1-sentences (if $\Box$ is the usual r.e. provability predicate and $\triangle$ a strong provability predicate). Japaridze's logic is decidable and has a reasonable Kripke semantics. An extensive treatment of Japaridze's logic is given in Boolos [1993b].

Bimodal analysis of other unusual provability concepts has been undertaken by Visser [1989,1995] and Shavrukov [1991,1994]. Using the work of Guaspari and

Solovay [1979], Shavrukov [1991] found a complete axiomatization of the bimodal logic of the usual and *Rosser's provability predicate* for Peano arithmetic (see also section 9). It is worth noting that Rosser's provability predicate, although numerating (externally) the same theory as the usual one, has a very different modal behaviour; e.g., Rosser consistency of **PA** is a provable fact, but on the other hand, Rosser's provability predicate is not provably closed under modus ponens. Shavrukov [1994] characterizes the logic of the so-called *Feferman provability predicate*. This work was preceded by Visser [1989,1995], where the concept of *provability in* **PA** *from 'nonstandardly finitely many' axioms* and some other unusual provability concepts were bimodally characterized. These systems were motivated by their connections with interpretability logic, but another motivation originates with Jeroslow and Putnam who studied the Rosser and Feferman style systems as 'experimental' systems: their self-correcting behaviour is supposed to be closer to the way humans reason. Studying ordinary provability and self-correcting provability can provide a good heuristic for appreciating the differences between both kinds of systems.

A final example of such an analysis of an unusual proof predicate by the development of a bimodal logic was Lindström [1994]'s analysis of *Parikh provability*, i.e., the proof predicate that allows $\Box A/A$ as a rule of inference.

Additional early results in bimodal logic, e.g., a bimodal analysis of the so-called Mostowski operator, can be found in Smoryński [1985].

Many results in bimodal provability logic can be generalized to *polymodal logic*. Such a generalization is particularly natural in the modal-logical study of progressions of theories, a topic in proof theory that goes as far back as the work of Turing [1939]. From the modal-logical point of view, however, such a generalization, in all known cases, does not lead to any essentially new phenomena. Roughly, the resulting systems happen to be direct sums of their bimodal fragments; therefore we shall not go into the details.

Polymodal analogues are known for Japaridze's bimodal logic (modalities, indexed by natural numbers n, correspond to the operators *to be provable from all true* Π_n*-sentences*), and for natural provability logics due to Carlson and Beklemishev. Here, the modal operators correspond to the theories of the original Turing-Feferman progressions of transfinitely iterated reflection principles, and thus, are indexed by ordinals for some constructive system of ordinal notation, say, the natural one up to ϵ_0. Iterating full reflection leads to the polymodal analogue of $\mathsf{PRL}_{\mathbf{PA},\mathbf{ZF}}$, and transfinitely iterated consistency leads to a natural polymodal analogue of **A**-type provability logics (Beklemishev [1991,1994]).

9. Rosser orderings

To discuss Rosser sentences and more generally the so-called Rosser provability predicate in a modal context, Guaspari and Solovay [1979] enriched the modal language by adding, for each $\Box A$ and $\Box B$, the formulas $\Box A \prec \Box B$ and $\Box A \preccurlyeq \Box B$, with as their arithmetic realizations the Σ_1-sentences "A^* is provable by a proof

that is smaller than any proof of B^*", and "A^* is provable by a proof that is smaller than or equal to any proof of B^*" (so-called *witness comparison formulas*). They axiomatized modal logics $\mathbf{R}^-$ and $\mathbf{R} = \mathbf{R}^- +$ the rule $\Box A/A$, and gave an arithmetic completeness result for $\mathbf{R}$. In this arithmetic completeness result they did have to allow arbitrary *standard* provability predicates in the arithmetic realizations however, i.e., arbitrary provability predicates satisfying the three Löb conditions. Shavrukov [1991] (see also the end of section 8) showed that this restriction can be dropped when one restricts the contexts for the new operator to $\Box A \prec \Box \neg A$ (the *Rosser provability predicate*, for short: $\Box^R A$), and de Jongh and Montagna [1991] showed that, allowing formulas with free variables as arithmetic substitutions leads to $\mathbf{R}^-$ as the arithmetically complete system. Guaspari and Solovay [1979] also showed that for some standard provability predicates all *Rosser sentences* (i.e., sentences α such that $\mathbf{PA} \vdash \alpha \leftrightarrow (\mathrm{Pr}_{\mathbf{PA}}(\ulcorner \neg \alpha \urcorner) \prec \mathrm{Pr}_{\mathbf{PA}}(\ulcorner \alpha \urcorner))$ are equivalent, and that for some other standard provability predicates this is not the case. This leaves open the question whether a reasonable notion of *usual* proof predicate can be defined for which the question "Is the Rosser sentence unique?" does have a definite answer. Hence also, uniqueness of fixed points is not provable in $\mathbf{R}$. Finally, they showed that also the existence part of the fixed point theorem fails for $\mathbf{R}$. Simpler proofs for the completeness theorems were given in de Jongh [1987] and Voorbraak [1988].

There are connections between this work in provability logic and speed up. First, de Jongh and Montagna [1988,1989] gave a new simpler proof of Parikh [1971]'s theorem that, for any provably recursive function g there is a sentence α provable in $\mathbf{PA}$ such that $\mathbf{PA}$ proves $\mathrm{Pr}_{\mathbf{PA}}(\ulcorner \alpha \urcorner)$ by a much shorter proof in the sense of g ($a <_g b$ iff $g(a) < b$) than it proves A itself. In de Jongh and Montagna [1988] this was done for g the identity function by showing that $\vdash_{\mathbf{R}} \boxdot(p \leftrightarrow (\Box\Box p \prec \Box p)) \rightarrow p$ ($\Box\Box p \prec \Box p$ *has only provable fixed points in* $\mathbf{R}$). The result shows that any fixed point α in $\mathbf{PA}$ of the arithmetic formula $\mathrm{Pr}_{\mathbf{PA}}[\mathrm{Pr}_{\mathbf{PA}}(x)] \prec \mathrm{Pr}_{\mathbf{PA}}(x)$[4] is provable in $\mathbf{PA}$, and the shortest proof of $\mathrm{Pr}_{\mathbf{PA}}(\ulcorner \alpha \urcorner)$ is shorter than the one of α. In the paper general conditions were given under which formulas have only provable fixed points in $\mathbf{R}$. In de Jongh and Montagna [1989] a Guaspari-Solovay theory of $\prec_g$ is developed for the notion "much shorter in the sense of g". Under reasonable conditions on g the resulting modal theory $\mathbf{R}^{\mathbf{0}}_{\mathbf{g}}$ is not dependent on g. Parikh's theorem can then be proved in this setting. The theory of provable fixed points was extended to this setting. In his review of these and some consecutive papers Beklemishev [1993b] rightly remarked that the changing of the orders of the proofs in Guaspari-Solovay style interferes with the order induced by the function g and makes some of the results somewhat less clear than one might wish.

Montagna [1992] applied the results on provable fixed points in a study of *metamathematical rules*, i.e., rules like $\mathrm{Pr}_T(\ulcorner \alpha \urcorner)/\alpha$ that can be considered as realizations of modal-logical rules (in case: $\Box A/A$). He classified these rules into two types: rules giving only polynomial speed up in proofs in arithmetic, and rules giving a superexponential speed up. In Hájek, Montagna and Pudlak [1993] it was

[4]for the meaning of [...] see notation 12.2.

shown that the rule $\Box A/A$ is maximally powerful among these metamathematical rules in the sense that the use of any of them can be polynomially simulated by $\Box A/A$. Moreover, in that paper natural examples of statements of which the proof is superexponentially shortened by the above rule are given.

10. Logic of proofs

A provability reading of the modality $\Box$ as *"is (informally) provable"* was an intended semantics for the classical system **S4** of propositional modal logic (see end of section 2) since Gödel's paper (Gödel [1933]). However, as we have seen, the straightforward interpretation of $\Box F$ as $\Pr(\ulcorner F \urcorner)$ leads to the logics **L** and **S** which are incompatible with **S4**. The reflexivity principle $\Box F \to F$ fails in **L**, and the necessitation rule fails in **S**. Nevertheless, an interesting interpretation of the **S4**-modality as formal provability is possible. One can have the reflexivity principle as well as the necessitation rule if one incorporates into the modal language machinery to keep all proofs "real", i.e., given by actual natural numbers and not quantifying over them as in the provability predicate. Artëmov succeeded in doing this by replacing the quantifiers by a kind of Skolem functions in his logic of proofs **LP** (Artëmov [1994,1995]).

The language of **LP** contains besides the usual Boolean constants, connectives and sentence variables, *proof variables* $x_0, \ldots, x_n, \ldots$, *proof axiom constants* $a_0, \ldots, a_n, \ldots$, function symbols: monadic !, and binary $+$ and $\times$, and finally the modal operator symbol $[\![\ ]\!](\ )$.

Terms and formulas are defined in the natural way: proof variables and axiom constants are terms; sentence variables and Boolean constants are formulas; whenever s and t are terms $!t, (s+t), (s \times t)$ are again terms, Boolean connectives behave conventionally, and for t a term and F a formula, $[\![t]\!]\, F$ is a formula. We will write $s \cdot t$ or even st instead of $(s \times t)$ and skip parentheses when convenient. A term is *ground* if is does not contain variables. The system $\mathbf{LP}_{AS}$ has as its axioms all formulas of the of the forms below, and as its only rule *modus ponens*:

A0. The tautologies in the language of **LP**,
A1. $[\![t]\!]\, F \to F$ "reflexivity"
A2. $[\![t]\!]\, (F \to G) \to ([\![s]\!]\, F \to [\![ts]\!]\, G)$ "application"
A3. $[\![t]\!]\, F \to [\![!t]\!][\![t]\!]\, F$ "proof checker"
A4. $[\![s]\!]\, F \to [\![s+t]\!]\, F, \quad [\![t]\!]\, F \to [\![s+t]\!]\, F$ "choice"
AS. A finite set of formulas of the form $[\![c]\!]\, A$, where c is an axiom constant, and A is an axiom A0-A4 "axiom specification"

The system **LP** is the generic name for the $\mathbf{LP}_{AS}$'s of the various axiom specifications *AS*. The intended understanding of **LP** is as a logic of operations on proofs, where $[\![t]\!]\, F$ stands for *"t is a code for a proof of F"*. For the usual Gödel proof predicate $\mathrm{Proof}(x, y)$ in **PA** there are provably recursive functions from codes of proofs to codes of proofs corresponding to $\times$ and !: $\times$ stands for an operation on proof sequences which realizes the *modus ponens* rule in arithmetic, and ! is a proof checker operation

as it appears in the proof of the second Gödel Incompleteness theorem. The usual proof predicate has a natural nondeterministic version PROOF(x, y) here called *standard nondeterministic proof predicate*: "*x is a code of a derivation containing a formula with a code y*". The predicate PROOF already has all three operations of the **LP**-language: the operation $s + t$ is in its case just the concatenation of the (nondeterministic) proofs s and t.

The system **LP** reminds one of propositional dynamic logic (see e.g., Harel [1984]), but is really quite different in character, since the modalities $[\![t]\!](\cdot)$ do not satisfy the property $[\![t]\!](p \to q) \to ([\![t]\!]\, p \to [\![t]\!]\, q)$ in **LP**. This makes the logic **LP** nonnormal and not a polymodal logic in the sense of section 8. Nevertheless, the entire variety of labeled modalities in **LP** can simulate **S4**. For example, the necessitation rule $F/\Box F$ of normal modal logics has its constructive counterpart in **LP**: if $\mathbf{LP} \vdash F$, then $\mathbf{LP} \vdash [\![t]\!]F$ for some ground term t. In general, let F^o be the result of substituting $\Box$ for all occurrences of $[\![t]\!]$ in F, and $X^o = \{F^o \mid F \in X\}$ for any set X of **LP**-formulas. It is easy to see that **LP** is sound with respect to **S4**: $(\mathbf{LP})^o \subseteq \mathbf{S4}$. The converse inclusion $\mathbf{S4} \subseteq (\mathbf{LP})^o$ turns out to be valid as well: by an **LP**-*realization* $r = r(AS)$ of a modal formula F we mean

1. An assignment of **LP**-terms to all occurrences of $\Box$ in F,
2. a choice of an axiom specification AS.

Under F^r we denote the image of F under the realization r. Positive and negative occurrences of modality in a formula and a sequent are defined in the usual way. A realization r is *normal* if all negative occurrences of $\Box$ are realized by proof variables.

10.1. Theorem. (Artëmov [1995]) *If $\vdash_{\mathbf{S4}} F$, then $\vdash_{\mathbf{LP}_{AS}} F^r$ for some axiom specification AS and some normal realization $r = r(AS)$.*

The proof of the theorem provides an algorithm which, for a given cutfree derivation $\mathcal{T}$ in **S4**, assigns **LP**-terms to all appearances of the modality in $\mathcal{T}$.

Let us agree to use a new function symbol $\iota z\, \varphi(z)$ for any arithmetic formula $\varphi(z)$. A formula $\psi(\iota z\, \varphi(z))$ is now supposed to be decoded in the usual way (see van Dalen [1994]) as a pure arithmetic formula $\psi(\iota z\, \varphi(z))^-$: for the innermost occurrence of $\iota z\, \varphi(z)$ put $\psi(\iota z\, \varphi(z))^-$ to be $\exists z(\varphi(z) \wedge \psi(z))$, and then iterate this procedure when needed. Under $\mu z\, \varphi$ we understand the ι-term determined by the formula $\varphi(z) \wedge \forall v < z \neg \varphi(v)$. An arithmetic formula φ is *provably* Δ_1 iff both φ and $\neg\, \varphi$ are provably Σ_1. A term $\mu z\, \varphi$ is *provably recursive* iff φ is provably Σ_1 and *provably total* iff $\mathbf{PA} \vdash \exists z\, \varphi(z)$. A *closed recursive term* is a provably total and provably recursive term $\mu z\, \varphi$ such that φ contains no free variables other than z. Closed recursive terms represent all provably recursive names for natural numbers. We have to make these distinctions, since some operations on proofs, e.g., the proof checker !, really depend on the name of the argument, not only on its value.

A *proof predicate* is a provably Δ_1-formula $\mathrm{Prf}(x, y)$ such that, for all φ, if $\mathbf{PA} \vdash \varphi$, then, for some $n \in \omega$, $\mathrm{Prf}(n, \ulcorner \varphi \urcorner)$ holds. A proof predicate $\mathrm{Prf}(x, y)$ is here called *normal* if

1. For every proof k, the set of corresponding theorems is finite and the function $T(k) =$ *the code of the set* $\{l \mid \mathrm{Prf}(k,l)\}$ is provably total recursive,

2. For any finite set X of codes of theorems of **PA** there exists a natural number n such that $X \subseteq T(n)$.

For each normal proof predicate Prf there are provably recursive terms $m(x,y)$, $a(x,y)$, $c(x)$ such that for all closed recursive terms s,t and for all arithmetic formulas φ,ψ the following formulas are valid:

$$\mathrm{Prf}(s,{\ulcorner}\varphi\to\psi{\urcorner})\wedge\mathrm{Prf}(t,{\ulcorner}\varphi{\urcorner})\to\mathrm{Prf}(m(s,t),{\ulcorner}\psi{\urcorner})$$

$$\mathrm{Prf}(s,{\ulcorner}\varphi{\urcorner})\to\mathrm{Prf}(a(s,t),{\ulcorner}\varphi{\urcorner}),\quad \mathrm{Prf}(t,{\ulcorner}\varphi{\urcorner})\to\mathrm{Prf}(a(s,t),{\ulcorner}\varphi{\urcorner})$$

$$\mathrm{Prf}(t,{\ulcorner}\varphi{\urcorner})\to\mathrm{Prf}(c({\ulcorner}t{\urcorner}),{\ulcorner}\mathrm{Prf}(t,{\ulcorner}\varphi{\urcorner}){\urcorner}).$$

As we have noted above, the nondeterministic Gödel proof predicate PROOF is a normal proof predicate.

Let AS be an axiom specification. An arithmetic *AS-realization* * of the **LP**-language has the following parameters: AS, a normal proof predicate Prf, an evaluation of the sentence letters by sentences of arithmetic, and an evaluation of proof letters and axiom constants by closed recursive terms. We put $\top^* \equiv (0=0)$ and $\bot^* \equiv (0=1)$, * commutes with Boolean connectives, $(t\cdot s)^* \equiv m(t^*,s^*)$, $(t+s)^* \equiv a(t^*,s^*)$, $(!t)^* \equiv c({\ulcorner}t^*{\urcorner})$, $([\![t]\!]F)^* \equiv \mathrm{Prf}(t^*,{\ulcorner}F^*{\urcorner})$. We assume also that $\mathbf{PA}\vdash G^*$ for all $G\in AS$.

Under any *AS-interpretation* * an **LP**-term t becomes a closed recursive term t^* (i.e., a recursive name of a natural number), and an **LP**-formula F becomes an arithmetic sentence F^*. Also note that the reflexivity principle is there, since $[\![t]\!]F\to F$ is provable in **PA** under any interpretation *. Indeed, let n be the value of t^*. If $\mathrm{Prf}(\overline{n},{\ulcorner}F^*{\urcorner})$ is true, then $\mathbf{PA}\vdash F^*$, thus $\mathbf{PA}\vdash\mathrm{Prf}(\overline{n},{\ulcorner}F^*{\urcorner})\to F^*$. If $\mathrm{Prf}(\overline{n},{\ulcorner}F^*{\urcorner})$ is false, then $\mathbf{PA}\vdash\neg\mathrm{Prf}(\overline{n},{\ulcorner}F^*{\urcorner})$, and again $\mathbf{PA}\vdash\mathrm{Prf}(\overline{n},{\ulcorner}F^*{\urcorner})\to F^*$.

10.2. Theorem. (Artëmov [1995], arithmetic completeness of **LP**) *If* $\vdash_{\mathbf{LP}_{AS}} F$, *then* $\mathbf{PA}\vdash F^*$ *and hence* $\mathbb{N}\vDash F^*$, *for any AS-interpretation* *.

Combining theorems 10.1 and 10.2 provides arithmetic completeness of **S4**:

10.3. Theorem. *If* $\vdash_{\mathbf{S4}} F$, *then* $\mathbf{PA}\vdash F^r$ *for some realization* r *and some axiom specification AS.*

By Gödel's translation of intuitionistic propositional logic into **S4**, which provides a faithful embedding of intuitionistic propositional logic in to **S4** (Gödel [1933],McKinsey and Tarski [1948]), this automatically includes an arithmetic completeness result for intuitionistic logic as well. If one considers this in the light of the Curry-Howard term interpretation of intuitionistic natural deduction (see e.g., Troelstra and Schwichtenberg [1996]), then one notes that many more terms are used in the **LP**-interpretation. It seems worthwhile to search in this light for a naturally restricted subsystem of **LP**.

The logic **LP** is a version of **S4** presented in a more rich operational language, with no information being lost, since **S4** is the exact term-forgetting projection of **LP**. A transliteration of an **S4**-theorem into **LP**-language may result in an exponential growth of its length, because the **S4**-derivations are included in the **LP**-formulas as proof terms. However, this increase looks much less dramatic if we calculate the complexity of the input **S4**-theorem F in an 'honest' way as the length of a proof of F in **S4**: the proof terms appearing in the realization algorithm have a size linear of the length of the proof, so, the total length of an **LP**-realization of an **S4**-theorem F is bounded by the quadratic function of the length of a given **S4**-derivation of F.

11. Notions of interpretability

In the part on interpretability and its logics (sections 11-15) we are going to investigate a family of concepts like interpretability and partial conservativity, which, in a sense, are generalizations of the notion of provability and for which we use the common name "interpretability". In the first two sections we will explain these concepts and relate them to each other. In the third section we develop an extension of provability logic to so-called interpretability logic with these concepts in mind. In the fourth section we will prove arithmetic completeness of the best-known interpretability logic **ILM** with regard to interpretability in as well as Π_1-conservativity over **PA**. In the fifth section we give a brief survey of the logics induced by some other concepts from the above family.

The concepts discussed in the first two sections are defined in terms of the comparison of the deductive strengths of theories like "one theory is included in another" or "one theory is consistent with another". To compare the strengths of two theories, these theories are not necessarily to be written in the same language, it is enough to organize a translation ("interpretation") of the language of one theory into the language of the other and just consider the translated variant of the first theory. While introducing the notions we will even assume that different theories always have different languages, even if the two languages coincide graphically.

For simplicity we restrict our considerations to theories formalized within the classical first order logic with identity; we suppose that the languages of the theories we consider contain finite or in the worst case countable sets of predicate constants and do not contain functional or individual constants. For a language K, Fm_K denotes the set of formulas of K and St_K the set of sentences, i.e., closed formulas of K. If D is a nonempty set, St_K^D denotes the set of sentences of K with parameters in D. More precisely, the elements of St_K^D are pairs $\langle \varphi, f \rangle$, where $\varphi \in \mathrm{Fm}_K$ and f is a valuation of the free variables of φ in D; we usually write $\varphi(a_1, \ldots, a_n)$ instead of $\langle \varphi(x_1, \ldots, x_n), f \rangle$, if $a_1 = f(x_1)$, $\ldots$, $a_n = f(x_n)$.

By a *theory* we mean a pair $T = \langle A, K \rangle$, where K is a language and $A \subseteq \mathrm{St}_K$. The set A contains the extra-logical axioms of T; *provability in* T means derivability from A in the classical predicate logic with identity. Thus, here we do not identify a theory with the set of its theorems, but rather with the set of its nonlogical axioms

(in particular, we suppose that $\mathbf{I\Sigma_1}$ is finitely axiomatized). However we do say that a theory $T = \langle A, K\rangle$ is a *subtheory* of a theory $T' = \langle A', K'\rangle$ and write $T \subseteq T'$, if $K \subseteq K'$ and the set of theorems of T is a subset of that of T'; if at the same time A is finite, then T is said to be a *finite subtheory* of T'. If $K = K'$, we denote the theory $\langle A \cup A', K\rangle$ by $T + T'$; if $M \subseteq \mathrm{St}_K$ and $\varphi \in \mathrm{St}_K$, we may also use $T + M$ and $T + \varphi$ to denote the theories $\langle A \cup M, K\rangle$ and $\langle A \cup \{\varphi\}, K\rangle$, respectively.

Let us not be too lazy to define the well-known notion of *first order model:* for a language K, a *K-model* is a pair $M = \langle D, G\rangle$, where D is a nonempty set (of D- *"individuals"*) called the *domain* and G is a function that assigns to each n-place predicate constant P of K an n-ary relation G_P on D, such that $G_=$ is the identity relation. The *truth* of $\varphi \in \mathrm{St}_K^D$ in M, in symbols $M \vDash \varphi$, is defined in the standard way: an atom $P(a_1, \ldots, a_n)$ is true in M iff $G_P(a_1, \ldots, a_n)$ holds, truth commutes with the Boolean connectives and $M \vDash \forall x \varphi(x)$ iff for all $a \in D$, $M \vDash \varphi(a)$. *The theory T_M of a K-model M* is defined as $\langle \{\varphi \in \mathrm{St}_K | M \vDash \varphi\}, K\rangle$. And M is said to be *a model of a theory* $\langle A, K\rangle$, if $M \vDash \varphi$ for each $\varphi \in A$ (of course the latter also implies that $M \vDash \varphi$ for any closed theorem φ of T).

Let K and K' be languages. We may suppose that the set of individual variables of K is a subset of that of K' and that there are infinitely many variables of K' not belonging to K. Then a *relative translation* from K into K' is a pair $\langle \ell, \sigma(x)\rangle$, where:

- ℓ is a function which assigns to each n-place predicate constant P of K a formula $P^\ell(v_1, \ldots, v_n)$ of K' whose bounded variables do not belong to K and whose free variables are the first n variables of the alphabetical list of the variables of K',
- $\sigma(x)$ is a formula of K' (called the *relativizing formula*) with precisely x free whose bounded variables do not belong to K.

Henceforth we usually omit the word "relative(ly)" and we call translations from the language of **PA** into the same language *arithmetic translations*. Now, for each formula $\varphi \in \mathrm{Fm}_K$ we define $t\varphi$, the *t-translation* of φ into K', by the following induction on the complexity of φ:

- $t(x = y)$ is $x = y$,
- for any other atom $P(x_1, \ldots, x_n)$, $tP(x_1, \ldots, x_n)$ is $P^\ell(x_1, \ldots, x_n)$,
- t commutes with Boolean connectives: $t(\alpha \to \beta) = t\alpha \to t\beta$, etc.,
- $t(\forall x\, \alpha)$ is $\forall x(\sigma(x) \to t\alpha)$, and thus $t(\exists x\, \alpha)$ is $\exists x\, (\sigma(x) \wedge t\alpha)$.

If T and T' are theories in the languages K and K' and t is a translation from K into K', we define the theories

$$t(T) = \langle \{t\varphi \mid \varphi \in \mathrm{St}_K,\ T \vdash \varphi\}, K'\rangle \text{ and}$$
$$t^{-1}(T') = \langle \{\varphi \mid \varphi \in \mathrm{St}_K,\ T' \vdash t\varphi\}, K\rangle.$$

The notion of translation is a formal analog of that of model: a translation $t = \langle \ell, \sigma(x) \rangle$ from K into K' in fact defines a K-model in the language K', where $\sigma(x)$ plays the role of D and ℓ the role of G; as soon as we have a K'-model $M' = \langle D', G' \rangle$ such that $\{a \in D' \mid M' \vDash \sigma(a)\} \neq \emptyset$, a unique K-model $M = \langle D, G \rangle$ arises by taking $D = \{a \in D' \mid M' \vDash \sigma(a)\}$ and $G_P = \{\langle a_1, \ldots, a_n \rangle \in D^n \mid M' \vDash P^\ell(a_1, \ldots, a_n)\}$ for each n-place predicate letter P of K; we call this model the *K-model induced by* (t, M').

Suppose K and K' are languages and $M = \langle D, G \rangle$ and $M' = \langle D', G' \rangle$ are K- and K'-models, respectively. Then an *interpretation* of M in M' is a translation t from K into K' such that for all $\varphi \in \mathrm{St}_K$, $M \vDash \varphi \iff M' \vDash t\varphi$, i.e., the K-model induced by (t, M') is elementarily equivalent to M. And a *strong interpretation* of M in M' is a pair (t, f), where $t = \langle \ell, \sigma(x) \rangle$ is a translation from K into K' (in fact an interpretation of M in M') and f is an injection of D into $\{a \in D' \mid M' \vDash \sigma(a)\}$ such that for any n-place predicate letter P of K and any $a_1, \ldots, a_n \in D$, we have $M \vDash P(a_1, \ldots, a_n) \iff M' \vDash P^\ell(fa_1, \ldots, fa_n)$; as is easily seen we have then also $M \vDash \varphi(a_1, \ldots, a_n) \iff M' \vDash t\varphi(fa_1, \ldots, fa_n)$ for all $a_1, \ldots, a_n \in D$, $\varphi(x_1, \ldots, x_n) \in \mathrm{Fm}_K$. If M and M' are models of theories T and T' respectively and t (or (t, f) for some f) is an interpretation (or a strong interpretation) of M in M', then M is also a model of $T + t^{-1}(T')$ and M' is a model of $T' + t(T)$, so we have a "truth-preserving" way of enriching both T and T'.

Gödel's method of arithmetization can be mentioned here as an impressive example of a strong interpretation (t, f) of the "standard model" of meta-arithmetic (though the latter is not formal) in the standard model of arithmetic: f is just the Gödel numbering function which injects the "domain" of the "standard model" of meta-arithmetic, the set of finite strings of arithmetic symbols, into the domain ω of the standard model of arithmetic, and t is the function which assigns to each meta-predicate its what we call "arithmetic formalization".

The above-defined notion of interpretability of models can also be considered as a relation between complete and consistent theories (of these models). It is easily seen that, if $T = \langle A, K \rangle$ and $T' = \langle A', K' \rangle$ are two such theories and t is a translation from K into K', then the assertions $t(T) \subseteq T'$, $t^{-1}(T') \subseteq T$, $T' \subseteq t(T)$, $T \subseteq t^{-1}(T')$ are equivalent. But in the general case that is not so, and we get at least the following four natural binary relations between theories (T and T' are arbitrary theories and t ranges over translations from the language of T into the language of T'):

11.1. Definition.

- T is *interpretable* in T', if there exists t, called an *interpretation* of T in T' such that $t(T) \subseteq T'$,
- T' is *cointerpretable* in T, if there is t, called a *cointerpretation* of T' in T, such that $t^{-1}(T') \subseteq T$,
- T is *faithfully interpretable* in T', if there is t, called a *faithful interpretation* of T in T', which is both an interpretation of T in T' and a cointerpretation of T' in T,

- T is *weakly interpretable* in T' if there exists t, called a *weak interpretation* of T in T' such that $T' + t(T)$ is consistent (which is also equivalent to the assertion that $T + t^{-1}(T')$ is consistent).

The binary relation of weak interpretability has a natural many-place generalization. Observe that T is weakly interpretable in T' if and only if T is interpretable in some consistent extension of T' which has the same language as T'. Instead of pairs we can consider arbitrary nonempty finite sequences of theories and say that such a sequence $T_1, \ldots, T_n$ is (linearly) *tolerant*, if there are consistent extensions $T_1^+, \ldots, T_n^+$ of these theories such that for each $1 \leqslant i < n$, T_{i+1}^+ is interpretable in T_i^+. Thus, consistency is the unary case of linear tolerance and weak interpretability the binary case. Further generalization consists in removing linearity and passing from sequences of theories to trees: a finite tree of theories is *tolerant*, if there are consistent extensions of these theories, of which each one is interpretable in its predecessors in the tree. The intuition here is that in a tolerant tree of theories we can add to each theory the translated information contained in its children (which already have been augmented in the similar manner), obtaining this way a consistent "avalanche" of information. Changing in the above definition the word "interpretable" for "cointerpretable", we obtain the notion of *cotolerance* of a tree of theories.

The notions of interpretability and weak interpretability between theories were introduced by Tarski, Mostowski and Robinson [1953]; faithful interpretability was first considered by Feferman, Kreisel and Orey [1960], and cointerpretability, tolerance and cotolerance by Japaridze [1992,1993]. These relations between axiomatic theories can be used, and actually have been used many times, to prove relative consistency results, different kinds of conservativity results, decidability and undecidability of theories.

12. Interpretability and partial conservativity.

In many cases the interpretability relations can be characterized in terms of partial conservativity or consistency. We are going to study these characterizations only for theories in the language of **PA** with primitive recursive sets of axioms which contain the axioms of **PA**. Let us call such theories *superarithmetic theories* (again we consider the variant of **PA** without functional symbols; however, below we speak, without any confusion, about terms for primitive recursive functions in superarithmetic theories). As usual, the theorems proved below for this special class of theories are of a much more general character; actually they hold for all reasonable so-called (locally) essentially reflexive theories (see definition 12.3). The main theorems that we are going to prove in this section establish that for such theories interpretability and cointerpretability are nothing but Π_1- and Σ_1-conservativity, respectively (theorems 12.7 and 12.13); weak interpretability corresponds to what we call Π_1-consistency (theorem 12.8), and faithful interpretability of T in S takes place exactly when we have interpretability of T in S and cointerpretability of S in

T (theorem 12.14). For finitely axiomatizable theories the situation is considerably different. We will make some remarks and give references on this at the end of this section.

12.1. Definition. Let R be an n-ary relation on ω, $\alpha(x_1, \ldots, x_n)$ an arithmetic formula, and T a superarithmetic theory. We say that:

- α *defines* R,

  if for all $k_1, \ldots, k_n \in \omega$, we have $R(k_1, \ldots, k_n) \iff \mathbb{N} \vDash \alpha(\bar{k}_1, \ldots, \bar{k}_n)$, $\mathbb{N}$ the standard model of arithmetic,

- α *numerates* R *in* T,

  if for all $k_1, \ldots, k_n \in \omega$, $R(k_1, \ldots, k_n) \Longrightarrow T \vdash \alpha(\bar{k}_1, \ldots, \bar{k}_n)$,

- α *binumerates* R *in* T,

  if α numerates R and $\neg\alpha$ numerates the complement of R in T.

We need some more terminology and notation. The formula class $\Sigma_1!$ is the set of arithmetic formulas which have an explicit Σ_1 form, i.e., $\exists x\, \varphi$ for some primitive recursive formula φ. Similarly for $\Pi_1!$. Simply Σ_1 (resp. Π_1) denotes the class of formulas which are $\mathbf{I\Sigma_1}$-equivalent to some $\Sigma_1!$- (resp. $\Pi_1!$-) formula.

It is known (see Smoryński [1977]) that the predicate "x codes a true $\Sigma_1!$-sentence" can be formalized by a $\Sigma_1!$-formula, which we will denote by $\mathrm{True}(x)$. This formula is such that ($\mathbf{I\Sigma_1}$ proves that)

$$for\ each\ \Sigma_1!{-}sentence\ \varphi,\ \ \mathbf{I\Sigma_1} \vdash \varphi \leftrightarrow \mathrm{True}(\ulcorner\varphi\urcorner).$$

Next, we denote by $\mathrm{Regwit}(y, x)$ the very primitive recursive formula for which

$$\mathrm{True}(x) \ \equiv\ \exists y\, \mathrm{Regwit}(y, x)$$

and say that k is a *regular witness* of a $\Sigma_1!$-sentence φ, iff $\mathrm{Regwit}(\bar{k}, \ulcorner\varphi\urcorner)$ is true. And k is said to be a *regular counterwitness* of a $\Pi_1!$-sentence $\forall z\, \varphi$, iff k is a regular witness of $\exists z \neg \varphi$.

We write $T \vdash_k \varphi$ to express that k is the code of a T-proof of φ, and denote by $\mathbf{PC}$ the pure predicate calculus (with identity); theorems of $\mathbf{PC}$ will be referred to as *tautologies* and $\mathrm{Pr}(x)$ will denote in this section an intensional formalization of the predicate "x codes a tautology"; dually, $\mathrm{Con}(x)$ expresses that x codes a formula φ such that $\mathbf{PC} \nvdash \neg\, \varphi$. For a theory T and a natural number m, $T{\downarrow}m$ denotes the finite subtheory of T obtained by restricting the set of axioms of T to those whose codes are $\leqslant m$.

Suppose T is a theory in the arithmetic language. Given an arithmetic formula α defining the set of (codes of) axioms of T, we can build in a uniform way (see Feferman [1960]) a formula $\mathrm{Prf}_\alpha(z, x)$ (resp. $\mathrm{Prf}_{\alpha{\downarrow}y}(z, x)$) expressing that x codes some sentence φ and $T \vdash_z \varphi$ (resp. $T{\downarrow}y \vdash_z \varphi$). The formulas $\mathrm{Pr}_\alpha(x)$ and $\mathrm{Pr}_{\alpha{\downarrow}y}(x)$ will abbreviate $\exists z\, \mathrm{Prf}_\alpha(z, x)$ and $\exists z\, \mathrm{Prf}_{\alpha{\downarrow}y}(z, x)$, respectively. If T is a finite theory, there is a

canonical formula defining T, namely the formula $x = \bar{n}$, where n is the code of the conjunction of all axioms of T; in this case we will write Pr_T instead of $\mathrm{Pr}_{x = \bar{n}}$. The sentences Con_α, $\mathrm{Con}_{\alpha\downarrow y}$ and Con_T will abbreviate $\neg \mathrm{Pr}_\alpha(\ulcorner x \neq x \urcorner)$, $\neg \mathrm{Pr}_{\alpha\downarrow y}(\ulcorner x \neq x \urcorner)$ and $\neg \mathrm{Pr}_T(\ulcorner x \neq x \urcorner)$, respectively. Using the formula Pr_α, we can also construct in a standard way the sentence Compl_α expressing that T is syntactically complete, i.e., that for every $\varphi \in \mathrm{St}_T$ we have $T \vdash \varphi$ or $T \vdash \neg\varphi$.

12.2. Notation. For any arithmetic formula φ, let $[\varphi]$ denote the term (with exactly the same free variables as φ) for the primitive recursive function which, if the free variables of φ and $[\varphi]$ are $x_1, \ldots, x_n$, assigns to each n-tuple $k_1, \ldots, k_n$ of numbers the code of the formula $\varphi(\bar{k}_1, \ldots, \bar{k}_n)$.

We prefer this notation to the more common dot notation, because it avoids the need to specify the free variables.

12.3. Definition. A theory T, the language of which contains that of **PA**, is said to be *locally essentially reflexive*, if for any sentence φ of the language of T, $T \vdash \mathrm{Pr}_{T\downarrow n}(\ulcorner \varphi \urcorner) \to \varphi$ for all n. The theory is said to be *globally essentially reflexive* if for any formula φ of the language of T, $T \vdash \mathrm{Pr}_{T\downarrow n}[\varphi] \to \varphi$ for all n.

It is known that superarithmetic theories are globally essentially reflexive. At the same time no consistent finite(ly axiomatized) theory which satisfies the conditions of Gödel's second incompleteness theorem can even be locally essentially reflexive, for otherwise such a theory would prove its own consistency. In fact, it is shown in Visser [1990] that essential reflexivity is equivalent to full induction. (The idea of the proof, by the way, is already present in Kreisel and Lévy [1968].) That local essential reflexivity is much weaker than global essential reflexivity follows from the following observation. For any reasonable theory T, T plus local reflection for T is easily seen to satisfy local essential reflexivity. However, by a result from Feferman [1962], T plus local reflection for T is contained in T plus all true Π_1-sentences, which for weaker theories certainly does not entail full induction. It turns out that for our results we just need local essential reflexivity. This is the reason that, in the following we will with "essential reflexivity", perhaps nonstandardly, refer to its local version.

12.4. Definition. Let $T = \langle A, K\rangle$ and $T' = \langle A', K'\rangle$ be theories and suppose that $\Gamma \subseteq \mathrm{Fm}_K \cap \mathrm{Fm}_{K'}$. Then

- T is Γ-*conservative* over T', if for any $\varphi \in \Gamma \cap \mathrm{St}_K$, we have that $T \vdash \varphi$ implies $T' \vdash \varphi$,
- T is Γ-*consistent* with T', if for any $\varphi \in \Gamma \cap \mathrm{St}_K$, we have that $T \vdash \varphi$ implies $T' \nvdash \neg\varphi$; in other words, if T is Γ-conservative over some consistent extension of T' in the same language.

Note that for sufficiently strong theories the notions of Σ_1!- and Σ_1-conservativity (as well as Π_1!- and Π_1-conservativity) are equivalent.

12.5. Lemma. (**PA** $\vdash$:) *Suppose $t = \langle \ell, \sigma(x) \rangle$ is a translation from a language K into a language K' and $\varphi \in \mathrm{St}_K$. Then* $\mathbf{PA} \vdash \mathrm{Pr}(\ulcorner \varphi \urcorner) \to \mathrm{Pr}(\ulcorner \exists x\, \sigma(x) \to t\varphi \urcorner)$.

Proof. Argue in **PA**. Suppose P is a proof of φ in **PC**, and let $x_1, \ldots, x_n$ be all variables occurring freely in P. Let then $\Delta = \sigma(x_1) \wedge \ldots \wedge \sigma(x_n)$. By induction on the length of P, one can easily verify that $\vdash_{\mathbf{PC}} \Delta \to t\varphi$ and hence (as φ is closed) $\mathbf{PC} \vdash \exists\Delta \to t\varphi$, where $\exists\Delta$ is the existential closure of Δ. On the other hand, $\vdash_{\mathbf{PC}} \exists x\, \sigma(x) \to \exists\Delta$. Consequently, $\vdash_{\mathbf{PC}} \exists x\, \sigma(x) \to t\varphi$. $\dashv$

12.6. Lemma. (**PA** $\vdash$:) *For any formula $\alpha(x)$ defining a set of arithmetic sentences, there is an arithmetic translation t such that for any sentence φ,*

(a) $\mathbf{PA} + \mathrm{Con}_\alpha \vdash \mathrm{Pr}_\alpha(\ulcorner \varphi \urcorner) \to t\varphi$,
(b) $\mathbf{PA} + \mathrm{Con}_\alpha + \mathrm{Compl}_\alpha \vdash \mathrm{Pr}_\alpha(\ulcorner \varphi \urcorner) \leftrightarrow t\varphi$.

It is easier to explain the idea of the proof of this lemma than to give a strict proof. Gödel's completeness theorem for the classical predicate calculus (with identity) says that every consistent theory has a model. An analysis of Henkin's proof of this theorem shows how to construct for a consistent arithmeticly definable (say, superarithmetic) theory $T = \langle A, K \rangle$ a model $M = \langle D, G \rangle$, where both D and each relation G_P are arithmetically defined; the whole proof can be formalized in **PA** (Hilbert and Bernays [1939]). As we noted above, to define a K-model in some language (in our case in the language of **PA**) means to give a translation from K into this language; for each concrete sentence φ, **PA** plus the assumption that T is consistent then proves that as soon as φ is a theorem of T, φ is true in M, and the clause (a) of the lemma expresses just this fact; as for clause (b), it is an immediate consequence of (a), for $\mathbf{PA} + \mathrm{Compl}_\alpha \vdash \neg\, \mathrm{Pr}_\alpha(\ulcorner \varphi \urcorner) \to \mathrm{Pr}_\alpha(\ulcorner \neg\, \varphi \urcorner)$ and $\mathbf{PA} \vdash t\neg\, \varphi \leftrightarrow \neg\, t\varphi$.

12.7. Theorem. (Orey [1961], Hájek [1971,1972]) (**PA** $\vdash$:) *For superarithmetic theories T and S the following are equivalent:*

(i) *T is interpretable in S,*
(ii) *for all m, $S \vdash \mathrm{Con}_{T\downarrow m}$,*
(iii) *T is Π_1-conservative over S.*

Proof. (i)$\Rightarrow$(ii): Suppose $t = \langle \ell, \sigma \rangle$ is an interpretation of T in S. Let φ be the conjunction of all axioms of T with codes $\leqslant m$. Then $S \vdash t\varphi$, i.e., $S \vdash \neg\, t \neg\, \varphi$; as $T \vdash \exists x\, (x = x)$, we also have $S \vdash \exists x\, \sigma(x)$. Then, by lemma 12.5 and since S is essentially reflexive, $S \vdash \neg\, \mathrm{Pr}(\ulcorner \neg\, \varphi \urcorner)$, i.e., $S \vdash \mathrm{Con}_{T\downarrow m}$.

(ii)$\Rightarrow$(i): Let $\tau(x)$ be a primitive recursive formula defining the set of axioms of T, and $\alpha(x)$ the formula $\tau(x) \wedge \mathrm{Con}_{T\downarrow x}$. Then, as soon as condition (ii) holds, $\alpha(x)$ binumerates the set of axioms of T in S and, thus, $\mathrm{Pr}_\alpha(x)$ numerates the set of theorems of T in S. According to lemma 12.6(a), there is a translation t such

that $\mathbf{PA} + \mathrm{Con}_\alpha \vdash \mathrm{Pr}_\alpha(\ulcorner \varphi \urcorner) \to t\varphi$ for all φ, whence, taking into account the obvious fact that $S \vdash \mathrm{Con}_\alpha$, we have $S \vdash \mathrm{Pr}_\alpha(\ulcorner \varphi \urcorner) \to t\varphi$ for each φ. This, together with the fact that $\mathrm{Pr}_\alpha(x)$, numerates the set of theorems of T in S, implies that t is an interpretation of T in S.

(ii)$\Rightarrow$(iii): Suppose φ is a $\Pi_1!$-sentence and $T \vdash \varphi$. This means that $T{\downarrow}m \vdash \varphi$ for some m. We can suppose that m is large enough for $T{\downarrow}m$ to prove all axioms of $\mathbf{Q}$ (Robinson's arithmetic). It is known that the latter disproves all false $\Pi_1!$-sentences. This fact is provable in S and, as $S \vdash \mathrm{Pr}_{T{\downarrow}m}(\ulcorner \varphi \urcorner)$, S proves that the consistency of $T{\downarrow}m$ implies φ. Then, if (ii) is satisfied, we have $S \vdash \varphi$.

(iii)$\Rightarrow$(ii): Suppose T is Π_1-conservative over S and let m be an arbitrary natural number. It is obvious that T being essentially reflexive, proves $\mathrm{Con}_{T{\downarrow}m}$. And as $\mathrm{Con}_{T{\downarrow}m}$ is a $\Pi_1!$-sentence, $S \vdash \mathrm{Con}_{T{\downarrow}m}$. $\dashv$

Taking into account that weak interpretability of T in S is nothing but interpretability of T in some consistent extension of S, we get:

12.8. Corollary. ($\mathbf{PA} \vdash :$) *For superarithmetic theories T and S the following are equivalent:*

(i) *T is weakly interpretable in S,*
(ii) *for all m, $S \nvdash \neg\, \mathrm{Con}_{T{\downarrow}m}$,*
(iii) *T is Π_1-consistent with S.*

Our next goal is to find a similar characterization for cointerpretability. We need some preparatory lemmas.

12.9. Lemma. (Guaspari [1979]) ($\mathbf{PA} \vdash :$) *Suppose S is a superarithmetic theory and Γ is a recursively enumerable set of natural numbers. Then there is a Σ_1-formula $\gamma(x)$ such that:*

(i) *$\gamma(x)$ numerates Γ in S;*
(ii) *for any $k \in \omega$, if $k \notin \Gamma$, then $S + \neg\, \gamma(\bar{k})$ is Σ_1-conservative over S.*

Proof. Let $\theta(x)$ be a Σ_1-formula which defines Γ. Let $\Sigma_1!(x)$ be a primitive recursive formula expressing that x is (the code of) a $\Sigma_1!$-sentence and let $\overset{\bullet}{\to}$ be a term for the primitive recursive function which assigns to each pair m_1, m_2 of numbers, as soon as they code some formulas ϵ_1 and ϵ_2, the code of the formula $\epsilon_1 \to \epsilon_2$. Finally, let $\sigma(x)$ be a primitive recursive formula defining the set of axioms of S. Applying self-reference, we can construct a $\Sigma_1!$-formula $\gamma(x)$ such that

$$\begin{aligned}(1)\quad \mathbf{PA} \vdash \gamma(x) \;\leftrightarrow\; & \exists y \Big(\mathrm{Regwit}(y, [\theta(x)]) \wedge \forall z, t \leq y \\ & (\Sigma_1!(z) \wedge \mathrm{Prf}_\sigma(t, [\neg\, \gamma(x)] \overset{\bullet}{\to} z) \to \\ & \exists r (\mathrm{Regwit}(r, z) \wedge \forall r' \leqslant r \neg\, \mathrm{Regwit}(r', [\gamma(x)]))\Big).\end{aligned}$$

The formula $\gamma(x)$ expresses that there is a regular witness y of $\theta(\bar{x})$ and any $\Sigma_1!$-sentence λ with $S \vdash_{\leq y} \neg\, \gamma(\bar{x}) \to \lambda$ has a regular witness less than any regular witness of $\gamma(\bar{x})$.

(i). Suppose $n \in \Gamma$. Then $\theta(\bar{n})$ is true; let k be the smallest regular witness of $\theta(\bar{n})$. Let $\lambda_1, \ldots, \lambda_m$ be all the Σ_1!-sentences with

(2) $\quad S \vdash_{\leq k} \neg\, \gamma(\bar{n}) \to \lambda_i$.

Then

$$(3) \quad S \vdash \gamma(\bar{n}) \;\leftrightarrow\; \bigwedge\{\exists r(\mathrm{Regwit}(r, \ulcorner \lambda_i \urcorner) \wedge \forall\, r' \leqslant r \neg\, \mathrm{Regwit}(r', [\gamma(\bar{n})])) \mid 1 \leqslant i \leqslant m\}.$$

Argue in S. Suppose $\neg\, \gamma(\bar{n})$. Then, by (3), there is i $(1 \leqslant i \leqslant m)$ such that for any regular witness r of λ_i there is an r' with $r' \leqslant r$ which is a regular witness of $\gamma(\bar{n})$. By (2), λ_i is true and has a regular witness. But $\gamma(\bar{n})$ has no regular witness, because (as we have assumed) it is false, which is a contradiction. Thus $S \vdash \gamma(\bar{n})$, and this proves that $\gamma(x)$ numerates Γ in S.

(ii). Now suppose $n \notin \Gamma$ (i.e., $\theta(\bar{n})$ is false), λ is a Σ_1!-sentence and $S + \neg\, \gamma(\bar{n}) \vdash \lambda$. Then $S \vdash_e \neg\, \gamma(\bar{n}) \to \lambda$ for some e (under the standard Gödel numbering, $e > \ulcorner \lambda \urcorner$) and S proves that if $\theta(\bar{n})$ has a regular witness, the latter is larger than e. Argue in S, and suppose $\gamma(\bar{n})$. Then, by (1) and the above remark, $\theta(\bar{n})$ has a regular witness and the smallest such witness is larger than e. Then, again by (1), λ has a regular witness (smaller than any regular witness of $\gamma(\bar{n})$), so λ is the case. Thus $S + \gamma(\bar{n}) \vdash \lambda$ and, since we have assumed that $S + \neg\, \gamma(\bar{n}) \vdash \lambda$, we have $S \vdash \lambda$. This proves the desired conservativity. $\dashv$

12.10. Lemma. (Lindström [1984]) (**PA** $\vdash$:) *For superarithmetic theories T and S, there is a formula $\beta(x)$ such that:*

(i) *for all λ, if $S \vdash \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$, then, for some m, $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner \lambda \urcorner)$,*
(ii) *$\beta(x)$ binumerates the set of axioms of T in S.*

Proof. Let X be the set of all the sentences ϵ such that $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner \epsilon \urcorner)$ for some m. By lemma 12.9, there is a Σ_1!-formula $\gamma(x)$ such that for all sentences λ,

(4) $\quad$ *if $\lambda \in X$, then $S \vdash \gamma(\ulcorner \lambda \urcorner)$,*

(5) $\quad$ *if $\lambda \notin X$, then $S + \neg\, \gamma(\ulcorner \lambda \urcorner)$ is Σ_1-conservative over S.*

Let $\tau(x)$ and $\sigma(x)$ be primitive recursive formulas defining the sets of axioms of T and S, respectively. Applying self-reference, we define $\beta(x)$ by

$$\beta(x) \;\equiv\; \tau(x) \wedge \forall y, z \leqslant x(\mathrm{Prf}_\sigma(y, [\mathrm{Pr}_\beta(z)]) \to \gamma(z)).$$

To prove (i), suppose $S \vdash_m \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$. Clearly $S \vdash \ulcorner \lambda \urcorner \leqslant \bar{m}$ (unless we have some pathological Gödel numbering) and thus $S + \neg\, \gamma(\ulcorner \lambda \urcorner) \vdash \forall x(\beta(x) \to \tau{\downarrow}\bar{m}(x))$, where $\tau{\downarrow}\bar{m}(x)$ denotes $\tau(x) \wedge x \leqslant \bar{m}$. Hence $S + \neg\, \gamma(\ulcorner \lambda \urcorner) \vdash \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner) \to \mathrm{Pr}_{\tau\downarrow\bar{m}}(\ulcorner \lambda \urcorner)$ and, since $S \vdash_m \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$ and τ is primitive recursive,

(6) $\quad S + \neg\, \gamma(\ulcorner \lambda \urcorner) \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner \lambda \urcorner)$.

Suppose $\lambda \notin X$. Then, by (5) and (6) (as $Pr_{T\downarrow m}(\ulcorner \lambda \urcorner) \in \Sigma_1!$), $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner \lambda \urcorner)$, and (i) is proved.

If x is not the code of an axiom of T, then $S \vdash \neg\, \tau(\bar{x})$ and thus $S \vdash \neg\, \beta(\bar{x})$. If x is a code an axiom of T, then $S \vdash \tau(\bar{x})$, and to show that $s \vdash \beta(\bar{x})$ it is enough to show that for all $y, z\,(\leqslant x)$, $S \vdash \mathrm{Prf}_\sigma(\bar{y}, [\mathrm{Pr}_\beta(\bar{z})]) \to \gamma(\bar{z})$.

This is obvious if y is not the code of an S-proof of $\mathrm{Pr}_\beta(\bar{z})$. And if $S \vdash_y \mathrm{Pr}_\beta(\bar{z})$, then, by (i), z codes an element of X, whence, by (4), $\gamma(\bar{z})$. Hence (ii). $\dashv$

In the sequel we will use the following convention: λ^i is λ if $i = 0$, and is $\neg\, \lambda$ if $i = 1$.

12.11. Lemma. (Scott [1962]) (**PA** $\vdash$:) *Suppose S is a superarithmetic theory and $\nu(x)$ is a $\Sigma_1!$-formula. There is then a formula $\zeta(x)$ such that for all (arithmetically defined) functions $g, h : \omega \to \{0,1\}$, if the set $S_g = S + \{\nu(\bar{n})^{g(n)} \mid n \in \omega\}$ is consistent, then so is the set $S_{g,h} = S_g + \{\zeta(\bar{n})^{h(n)} \mid n \in \omega\}$.*

Proof. Let $\sigma(x)$ be a primitive recursive formula defining the set of axioms of S. We define the formula $\alpha(x)$ by:

$$\sigma(x) \vee \exists y \leqslant x \Big((x = [\nu(y)] \wedge \mathrm{True}([\nu(y)])) \vee (x = [\neg\, \nu(y)] \wedge \neg\, \mathrm{True}([\nu(y)])) \Big).$$

Explanation: Let $t : \omega \to \{0,1\}$ be such that $t(n) = 0$ iff $\nu(\bar{n})$ is true, and let $S^+ = S + \{\nu(\bar{n})^{t(n)} \mid n \in \omega\}$. Then the formula $\alpha(x)$ expresses that x is the code of an axiom of S^+. It is easy to see that

(7) *for each function $g : \omega \to \{0,1\}$, $\alpha(x)$ binumerates the set of axioms $S_g = S + \{\nu(\bar{n})^{g(n)} \mid n \in \omega\}$ in S_g. Consequently, $\mathrm{Prf}_\alpha(x,y)$ binumerates the relation "... is an S_g-proof of ..." in S_g.*

(for, roughly speaking, S_g thinks that $S^+ = S_g$).

Let $\mathrm{Seq}(s,l)$ be a primitive recursive formula expressing that s is the code of a $\{0,1\}$-valued sequence of length l (i.e., of a function: $M \to \{0,1\}$, where $M = \{0, \ldots, l-1\}$ if $l > 0$, and $M = \emptyset$ if $l = 0$). Let $\mathrm{Conj}(s,u)$ be a term for the primitive recursive function that assigns to each pair (s,u), of which s codes a finite (possibly empty) $\{0,1\}$-valued sequence $f = \langle f(0), \ldots, f(m) \rangle$ and u codes a formula λ which contains exactly one free variable, the code of the conjunction $\lambda(\bar{0})^{f(0)} \wedge \ldots \wedge \lambda(\bar{m})^{f(m)}$. Now, applying self-reference, we construct a formula $\zeta(x)$ such that

$$(8) \quad \mathbf{PA} \vdash \quad \zeta(x) \;\leftrightarrow\; \forall y \forall s \Big(\mathrm{Seq}(s,x) \wedge \mathrm{Prf}_\alpha(y, \mathrm{Conj}(s, \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\zeta(x)]) \to \exists t < y\, \exists s' (\mathrm{Seq}(s',x) \wedge \mathrm{Prf}_\alpha(t, \mathrm{Conj}(s', \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\neg\, \zeta(x)])) \Big).$$

The formula $\zeta(x)$ asserts that, if $\zeta(\bar{x})$ is proved by some extension of S^+ of the type $S^+ + \pm\zeta(\bar{0}) \wedge \ldots \wedge \pm\zeta(\overline{x-1})$ (where $\pm$ means the presence or absence of $\neg$), then there is a shorter proof of $\neg\, \zeta(\bar{x})$ in some extension of S^+ of the same type.

Assume that the set $S_g = S + \{\nu(\bar{n})^{g(n)} \mid n \in \omega\}$ is consistent. Let $U_0 = \{S_g\}$ and $U_{l+1} = \{R + \zeta(\bar{l}), R + \neg\zeta(\bar{l}) \mid R \in U_l\}$. To prove the lemma it suffices to show by induction on l that each $R \in U_l$ is consistent. The only element S_g of U_0 is consistent by our assumption. Suppose there is a theory in U_{l+1} which is inconsistent. Then there is a theory in U_l which proves $\zeta(\bar{l})$ or $\neg\zeta(\bar{l})$. Let then k be the smallest number such that, for some $R \in U_l$ and $i \in \{0,1\}$, we have $R \vdash_k \zeta(\bar{l})^i$. More precisely, k is the smallest number such that, for some $\{0,1\}$-valued sequence f of length l and some $i \in \{0,1\}$, we have $S_g \vdash_k \bigwedge\{\zeta(\bar{n})^{f(n)} \mid 0 \leqslant n < l\} \to \zeta(\bar{l})^i$, and R is the theory $S_g + \bigwedge\{\zeta(\bar{n})^{f(n)} \mid 0 \leqslant n < l\}$.

Below we employ, without explicit mention, proposition (7), the primitive recursiveness of Seq(.,.), Conj(.,.), and the fact that the number of $\{0,1\}$-valued sequences of length l is finite.

Case 1: $i = 0$. Then

(9) $\quad S_g \vdash \mathrm{Seq}(\ulcorner f \urcorner, \bar{l}) \wedge \mathrm{Prf}_\alpha(\bar{k}, \mathrm{Conj}(\ulcorner f \urcorner, \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\zeta(\bar{l})])$.

By our choice of k, there is no number $t < k$ and no $\{0,1\}$-valued sequence f' of length l such that $S_g \vdash_k \bigwedge\{\zeta(\bar{n})^{f'(n)} \mid 0 \leqslant n < l\} \to \neg\zeta(\bar{l})$. Therefore we also have:

(10) $\quad S_g \vdash \neg\,\exists t < \bar{k}\,\exists s'(\mathrm{Seq}(s', \bar{l}) \wedge \mathrm{Prf}_\alpha(t, \mathrm{Conj}(s', \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\neg\zeta(\bar{l})]))$.

Now, (9) and (10) imply by (8) that $S_g \vdash \neg\zeta(\bar{l})$, whence the theory R is inconsistent, which is in contradiction with the induction hypothesis.

Case 2: $i = 1$. Then $S_g \vdash \mathrm{Seq}(\ulcorner f \urcorner, \bar{l}) \wedge \mathrm{Prf}_\alpha(\bar{k}, \mathrm{Conj}(\ulcorner f \urcorner, \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\neg\zeta(\bar{l})])$, whence

(11) $\quad S_g \vdash \forall y > \bar{k}\,\forall s\,\exists t < y\,\exists s'(\mathrm{Seq}(s', \bar{l}) \wedge \mathrm{Prf}_\alpha(t, \mathrm{Conj}(s', \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\neg\zeta(\bar{l})]))$.

By our choice of k, for each $y \leqslant k$ and any $\{0,1\}$-valued sequence f' of length l, $S_g \nvdash_y \bigwedge\{\zeta(\bar{n})^{f'(n)} \mid 0 \leqslant n < l\} \to \zeta(\bar{l})$, whence

(12) $\quad S_g \vdash \forall y \leqslant \bar{k}\,\forall s\,\neg\,(\mathrm{Seq}(s, \bar{l}) \wedge \mathrm{Prf}_\alpha(y, \mathrm{Conj}(s, \ulcorner \zeta \urcorner) \overset{\bullet}{\to} [\zeta(\bar{l})]))$.

Now, (11) and (12) immediately imply by (8) that R is inconsistent, which is in contradiction with the induction hypothesis. $\dashv$

12.12. Lemma. (Lindström [1984]) (**PA** $\vdash$:) *Suppose T and S are super-arithmetic theories, $\alpha(x)$ binumerates the set of axioms of T in S, and $S \vdash \mathrm{Con}_\alpha$. There is then an interpretation t of T in S such that, for any sentence λ, $S \vdash t\lambda \;\Rightarrow\; S \vdash \mathrm{Pr}_\alpha(\ulcorner \lambda \urcorner)$.*

Proof. Assume the conditions of the lemma and let us fix an enumeration $\{\Psi_n\}_{n\in\omega}$ of all arithmetic sentences. Consider the following recursive definition, where X is any subset of ω:

$$(13)\quad \Phi_n \;=\; \begin{cases} \Psi_n, \text{ if} & \text{(a) } T \vdash \bigwedge\{\Phi_m \mid m < n\} \to \Psi_n, \text{ or} \\ & \text{(b) } T \nvdash \bigwedge\{\Phi_m \mid m < n\} \to \neg\,\Psi_n \text{ and } n \in X, \\ \neg\,\Psi_n, \text{ otherwise.} & \end{cases}$$

(If $n = 0$, $\bigwedge\{\Phi_m \mid m < n\}$ is identified with $\bar{0} = \bar{0}$.)

Let $\xi(x)$ be the formula given by lemma 12.11 for $\nu(x) = \mathrm{Pr}_\alpha(x)$. Next, let $\chi(x, y)$ be a formalization of the result of converting (13) into an explicit definition in the usual way using $\mathrm{Pr}_\alpha(x)$ and $\xi(x)$ to represent the predicates "$T \vdash \ldots$" and "$\ldots \in X$", respectively, and let $\beta(x)$ be $\exists y\, \chi(x, y)$. Obviously, $S \vdash \mathrm{Con}_\beta$ and $S \vdash \mathrm{Compl}_\beta$, whence by lemma 12.6(b), there is a translation t such that for each sentence λ, $S \vdash t\lambda \leftrightarrow \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$; clearly we also have $\mathbf{PA} \vdash \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner) \leftrightarrow \beta(\ulcorner \lambda \urcorner)$ and thus

(14) $S \vdash t\lambda \leftrightarrow \beta(\ulcorner \lambda \urcorner)$.

Suppose now $S \nvdash \mathrm{Pr}_\alpha(\ulcorner \lambda \urcorner)$. To complete the proof we must show that $S \nvdash t\lambda$. Let $g : \omega \to \{0, 1\}$ be such that $g(\ulcorner \lambda \urcorner) = 1$ and

(15) $S + \mathrm{Y}_g$ is consistent,

where $\mathrm{Y}_g = \{\mathrm{Pr}_\alpha(\bar{n})^{g(n)} \mid n \in \omega\}$. Next we define Φ'_n as follows:

$$\Phi'_n = \begin{cases} \Psi_n, \text{ if} & \text{(a)} \quad \mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < n\} \to \Psi_n \urcorner) \in \mathrm{Y}_g, \text{ or} \\ & \text{(b)} \quad \mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < n\} \to \neg\, \Psi_n \urcorner) \notin \mathrm{Y}_g \text{ and} \\ & \qquad \mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < n\} \wedge \Psi_n \to \lambda \urcorner) \notin \mathrm{Y}_g, \\ \neg\, \Psi_n, \text{ otherwise.} & \end{cases}$$

Let $h : \omega \to \{0, 1\}$ be such that

$$h(n) = 0 \text{ iff } Pr_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < n\} \wedge \Psi_n \to \lambda \urcorner) \notin \mathrm{Y}_g,$$

and let $\mathrm{Y}_{g,h} = \mathrm{Y}_g \cup \{\xi(\bar{n})^{h(n)} \mid n \in \omega\}$. Then, by lemma 12.11 and the choice of ξ, (15) implies that

(16) $S + \mathrm{Y}_{g,h}$ is consistent.

By induction on n we can easily check that $S + \mathrm{Y}_{g,h} \vdash \chi(\ulcorner \Phi'_n \urcorner, \bar{n})$, whence

(17) $S + \mathrm{Y}_{g,h} \vdash \beta(\ulcorner \Phi'_n \urcorner)$.

We now show by induction on n that for every n,

(18) $\mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < n\} \to \lambda \urcorner) \notin \mathrm{Y}_g$.

Observe that $\{\epsilon \mid \mathrm{Pr}_\alpha(\ulcorner \epsilon \urcorner) \in \mathrm{Y}_g\}$ is closed under logical deduction. If $n = 0$, (18) holds by our choice of Y_g. Suppose (18) holds for $n = k$.

Case 1: $\Phi'_k = \Psi_k$. Then either (a) $\mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < k\} \to \Psi_k \urcorner) \in \mathrm{Y}_g$, or (b) $Pr_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < k + 1\} \to \lambda \urcorner) \notin \mathrm{Y}_g$. The subcase (b) just means that (18) holds for $n = k + 1$, and the subcase (a) together with the induction hypothesis also implies (18) for $n = k + 1$.

Case 2: $\Phi'_k = \neg\, \Psi_k$. Then $\mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < k\} \wedge \Psi_k \to \lambda \urcorner) \in \mathrm{Y}_g$, for otherwise $\mathrm{Pr}_\alpha(\ulcorner \bigwedge\{\Phi'_m \mid m < k\} \to \neg\, \Psi_k \urcorner) \notin \mathrm{Y}_g$ and so $\Phi'_k = \Psi_k$. But then (18) for $n = k + 1$ easily follows from the induction hypothesis. This proves (18).

Finally, in view of (18), it follows that for some k, $\Phi'_k = \neg\, \lambda$. Hence, by (17), $S + \mathrm{Y}_{g,h} \vdash \beta(\ulcorner \neg\, \lambda \urcorner)$; clearly the latter implies $S + \mathrm{Y}_{g,h} \vdash \mathrm{Pr}_\beta(\ulcorner \neg\, \lambda \urcorner)$ and, (as $S \vdash \mathrm{Con}_\beta$) $S + \mathrm{Y}_{g,h} \vdash \neg\, \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$, whence, by (16), $S \nvdash \mathrm{Pr}_\beta(\ulcorner \lambda \urcorner)$, and by (14), $S \nvdash t\lambda$. The proof of lemma 12.12 is complete. ⊣

12.13. Theorem. (Japaridze [1993]) (**PA** $\vdash$:) *For superarithmetic theories T and S the following are equivalent:*

(i) *S is cointerpretable in T,*
(ii) *for all λ and m, if $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$, then $T \vdash \lambda$,*
(iii) *S is Σ_1-conservative over T.*

Proof. (i)⇒(ii): Suppose $t = \langle\tau, \delta\rangle$ is a cointerpretation of S in T, and also $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$, i.e., $S \vdash \mathrm{Pr}(\ulcorner \bigwedge T\downarrow m \to \lambda\urcorner)$, where $\bigwedge T\downarrow m$ is the conjunction of all axioms of $T\downarrow m$. Then, by lemma 12.5, $S \vdash \mathrm{Pr}(\ulcorner \exists x\, \delta(x) \to t(\bigwedge T\downarrow m \to \lambda)\urcorner)$, which implies, since S is essentially reflexive, $S \vdash \exists x\, \delta(x) \to t(\bigwedge T\downarrow m \to \lambda)$. Therefore, $S \vdash t\,(\exists x(x = x) \to (\bigwedge T\downarrow m \to \lambda))$, whence (as t is a cointerpretation of S in T) $T \vdash \exists x(x = x) \to (\bigwedge T\downarrow m \to \lambda)$, and $T \vdash \lambda$.

(ii)⇒(i): Let us fix the formula β from lemma 12.10. By clause (ii) of that lemma, β binumerates the set of axioms of T in S. Then, by lemma 12.11, there is a translation t such that for all sentences λ,

(19) if $S + \mathrm{Con}_\beta \vdash t\lambda$, then $S + \mathrm{Con}_\beta \vdash \mathrm{Pr}_\beta(\ulcorner\lambda\urcorner)$.

We claim that if condition (ii) of theorem 12.13 holds, then t is a cointerpretation of S in T. Indeed, suppose $S \vdash t\lambda$. Then, by (19), $S + \mathrm{Con}_\beta \vdash \mathrm{Pr}_\beta(\ulcorner\lambda\urcorner)$; on the other hand, we clearly have $S + \neg\,\mathrm{Con}_\beta \vdash \mathrm{Pr}_\beta(\ulcorner\lambda\urcorner)$. Consequently, $S \vdash \mathrm{Pr}_\beta(\ulcorner\lambda\urcorner)$. Then, by lemma 12.10(i), $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$ for some m, which together with the condition (ii) of our theorem, implies that $T \vdash \lambda$.

(ii)⇒(iii): Assume (ii). Suppose λ is a Σ_1-sentence and $S \vdash \lambda$. Let m be such that $T\downarrow m$ contains Robinson's arithmetic. Then $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$, whence, by (ii), $T \vdash \lambda$.

(iii)⇒(ii): Suppose S is Σ_1-conservative over T and $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$. Since $\mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$ is a Σ_1-sentence, it follows that $T \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$ and T, being essentially reflexive, proves λ. ⊣

12.14. Theorem. (Lindström [1984])(**PA** $\vdash$:) *A superarithmetic theory T is faithfully interpretable in a superarithmetic theory S iff T is Π_1-conservative over S and S is Σ_1-conservative over T.*

Proof. In view of theorems 12.7 and 12.13, the direction (⇒) is straightforward. To prove (⇐), suppose T is Π_1-conservative over S and S is Σ_1-conservative over T. Then by theorems 12.7 and 12.13, we have:

(20) for all m, $S \vdash \mathrm{Con}_{t\downarrow m}$,
(21) for all m and λ, if $S \vdash \mathrm{Pr}_{t\downarrow m}(\ulcorner\lambda\urcorner)$, then $T \vdash \lambda$.

Let β be the formula from lemma 12.10, and let $\alpha(x)$ be the formula $\beta(x) \wedge \mathrm{Con}_{\beta\downarrow x}$. Then (arguing as in the proof of theorem 12.7(ii)⇒(i)), (20) implies that $\alpha(x)$ binumerates the set of axioms of T in S and $S \vdash \mathrm{Con}_\alpha$. Consequently, by lemma 12.10, there is an interpretation t of T in S such that

(22) for all λ, $S \vdash t\lambda \Longrightarrow S \vdash \mathrm{Pr}_\alpha(\ulcorner\lambda\urcorner)$.

To show that t is also a cointerpretation of S in T, suppose $S \vdash t\lambda$. Then, by (22), $S \vdash \mathrm{Pr}_\alpha(\ulcorner\lambda\urcorner)$. It is obvious that $\mathbf{PA} \vdash \mathrm{Pr}_\alpha(\ulcorner\lambda\urcorner) \to \mathrm{Pr}_\beta(\ulcorner\lambda\urcorner)$. Then, by lemma 12.10(i), $S \vdash \mathrm{Pr}_{T\downarrow m}(\ulcorner\lambda\urcorner)$ for some m, whence, by (21), $T \vdash \lambda$. $\dashv$

12.15. Finitely axiomatized theories

In the case of finitely axiomatized theories the interpretability relations have other interesting characterizations. E.g., a theorem due to Harvey Friedman (improved by Visser [1990]) establishes that for finitely axiomatized sequential theories T and S, T is interpretable in S if and only if the weak theory $\mathbf{I\Delta_0} + \mathrm{EXP}$ proves that the consistency of S (with respect to cutfree proofs) implies the consistency of T (with respect to cutfree proofs).

12.16. Feasible interpretability

Visser introduced the notion of feasible interpretability. A theory T is *feasibly interpretable* in a theory T' iff there is a translation t from the language of T into the language of T' and a polynomial function $P(x)$ such that for any λ and x, if $T \vdash_x \lambda$, then $T' \vdash_{\leq P(x)} t\lambda$. In a similar manner we can define the notion of feasible Π_1-conservativity: T is feasibly Π_1-conservative over S iff there is a polynomial $P(x)$ such that for any x and Π_1-sentence λ, if $T \vdash_x \lambda$, then $S \vdash_{\leq P(x)} \lambda$. Verbrugge [1993b] showed that theorem 12.7(i)$\Leftrightarrow$(iii) continues to hold when "interpretable" and "Π_1-conservative" are replaced by "feasibly interpretable" and "feasibly Π_1-conservative".

The main difference between interpretability and feasible interpretability appears when one estimates the arithmetic complexity of the two relations: the relation of interpretability between extensions of arithmetic by finite sets of additional axioms is Π_2-complete (this is originally due to Solovay), whereas the relation of feasible interpretability between such theories turns out to be Σ_2-complete (Verbrugge [1993b]).

13. Axiomatization, semantics, modal completeness of ILM

The idea of interpretability logics arose in Visser [1990] in which they were already developed to a large extent. The modal completeness with respect to the Kripke-semantics due to Veltman was, for the most important systems, proved in de Jongh and Veltman [1990]. Realizing that one cannot cover the concept as well as provability, since interpretability has a more infinitary character, one has to choose primitives of course, and, somewhat surprisingly, it turns out that choosing a binary connective is much more rewarding than choosing a unary connective. The arithmetic realization of $A \rhd B$ in a theory T will be that T plus the realization of B is interpretable in T plus the realization of A (T plus A *interprets* T plus B), or, alternatively (and, as we have seen, in the case of **PA** equivalently), that T plus the realization of B is Π_1-conservative over T plus the interpretation of A. The

unary pendant "T interprets T plus A" is much less expressive and was studied in de Rijke [1992]. For a recent complete overview, see Visser [1997].

We first introduce a basic interpretability logic **IL**: it contains, besides the usual axiom $\Box(\Box A \to A) \to \Box A$ for the provability logic **L** and its rules, modus ponens and necessitation, the axioms:

(1) $\Box(A \to B) \to (A \rhd B)$,
(2) $(A \rhd B) \wedge (B \rhd C) \to (A \rhd C)$,
(3) $(A \rhd C) \wedge (B \rhd C) \to (A \vee B \rhd C)$,
(4) $(A \rhd B) \to (\Diamond A \to \Diamond B)$,
(5) $\Diamond A \rhd A$.

With respect to priority of parentheses $\rhd$ is treated as $\to$. Furthermore, in this section, we will consider the extension $\mathbf{ILM} = \mathbf{IL} + \mathrm{M}$ of **IL** where M is the axiom $(A \rhd B) \to (A \wedge \Box C \rhd B \wedge \Box C)$. We will write $\vdash_{\mathbf{IL}}$ and $\vdash_{\mathbf{ILM}}$ for derivability in **IL** and **ILM**, but sometimes we may leave off the subscript. As will be proved further on, the logic **ILM** is the logic of Π_1-conservativity of **PA**, and therefore also, as shown in the previous section, its interpretability logic. We will not treat here the logic **ILP** which arises by extending **IL** by the scheme $(A \rhd B) \to \Box(A \rhd B)$ that axiomatizes the interpretability logic of the most common finitely axiomatizable theories (Visser [1990], using a modal completeness result of de Jongh and Veltman [1990]).

13.1. Lemma.
(a) $\vdash_{\mathbf{IL}} \Box\neg A \to (A \rhd B)$,
(b) $\vdash_{\mathbf{IL}} A \vee \Diamond A \rhd A$,
(c) $\vdash_{\mathbf{IL}} A \rhd A \wedge \Box\neg A$.

Proof. The parts (a) and (b) are easy. For part (c) use lemma 2.1(j) to obtain $\vdash_{\mathbf{L}} A \to (A \wedge \Box\neg A) \vee \Diamond(A \wedge \Box\neg A)$. Then use the necessitation rule, axiom (1), part (b) and axiom (2). ⊣

13.2. Corollary.
(a) *The formulas* $A \rhd B$, $A \wedge \Box\neg A \rhd B$ *and* $A \rhd B \wedge \Box\neg B$ *are* **IL***-equivalent.*
(b) *The formulas* $A \rhd \bot$ *and* $\Box\neg A$ *are* **IL***-equivalent.*

Proof. (a) By lemma 13.1(c) and its converse, which is derivable from axiom (1), and transitivity of $\rhd$ (axiom (2)).

(b) The direction from right to left follows from lemma 13.1(a). The other direction is obtained by using axiom (4) with $\bot$ for B, lemma 2.1(i) and transitivity of $\rhd$. ⊣

13.3. Definition. An **IL***-frame* (also *Veltman-frame*) is an **L**-frame $\langle W, R\rangle$ with, for each $w \in W$, an additional relation S_w, which has the following properties:

(i) S_w is a relation on $w{\uparrow} = \{w' \in W \mid w\,R\,w'\}$,
(ii) S_w is reflexive and transitive,
(iii) if $w', w'' \in w{\uparrow}$ and $w'R\,w''$, then $w'S_w w''$.

We may write S for $\{S_w | w \in W\}$.

13.4. Definition. An **IL**-*model* is given by an **IL**-frame $\langle W, R, S\rangle$ combined with a forcing relation $\Vdash$ with the clauses:

$u \Vdash \Box A \iff \forall v(uRv \Rightarrow v \Vdash A)$,
$u \Vdash A \rhd B \iff \forall v(uRv \text{ and } v \Vdash A \Rightarrow \exists w(vS_u w \text{ and } w \Vdash B))$.

13.5. Definition.

1. If F is a frame, then we write $F \vDash A$ iff $F = \langle W, R, S\rangle$ and $w \Vdash A$ for every $w \in W$ and every $\Vdash$ on F.
2. If $\mathcal{K}$ is a class of frames, we write $\mathcal{K} \vDash A$ iff $F \vDash A$ for each $F \in \mathcal{K}$.
3. $\mathcal{K}_{\mathrm{M}}$, the class of **ILM**-frames, is the class of **IL**-frames satisfying
   (iv) if $u\,S_w\,v\,R\,z$, then $u\,R\,z$.
4. An **ILM**-model is an **IL**-model on an **ILM**-frame.

The scheme M characterizes (see section 2) the class of frames $\mathcal{K}_{\mathrm{M}}$; that is the content of part (b) of the next soundness lemma.

13.6. Lemma. *For all* **IL***-frames* F,

(a) *For each* A, *if* $\vdash_{\mathbf{IL}} A$, *then* $F \vDash A$.
(b) $F \vDash \mathbf{ILM}$ *iff* $F \in \mathcal{K}_{\mathrm{M}}$.
(c) *For each* A, *if* $\vdash_{\mathbf{ILM}} A$, *then* $\mathcal{K}_{\mathrm{M}} \vDash A$.

As before, in the case of **L**, we work inside a so-called adequate set. It is convenient to use the fact that $\Box$ is definable in **IL** in terms of $\rhd$: $\Box A$ is **IL**-equivalent to $\neg A \rhd \bot$ (corollary 13.2(b)). This means that we can, in constructing countermodels, restrict our attention to formulas that do not contain $\Box$. The entire following discussion will be based on the presumption the formulas discussed do not contain $\Box$.

The other side of the coin is that this will allow us to use $\Box$ as a defined symbol. The most convenient way to this turns out to be the following: $\Diamond A$ will be an abbreviation of $\neg(A \rhd \bot)$ and $\Box A$ will then abbreviate the formula $\sim\Diamond\sim A$ (i.e., $\sim A \rhd \bot$). We need to adapt the concept of adequate set to the new situation.

13.7. Definition. An *adequate* set of formulas is a set Φ that satisfies the following conditions:

1. Φ is closed under taking subformulas,
2. if $A \in \Phi$, then $\sim A \in \Phi$,
3. $\bot \rhd \bot \in \Phi$,
4. $A \rhd B \in \Phi$ if A is an antecedent or succedent of some $\rhd$-formula in Φ, and so is B.

13.8. Lemma. *If* Φ *is an adequate set, then* $A \rhd B \in \Phi$ *iff both* $\Diamond A$ *and* $\Diamond B$ *are in* Φ (*and in case* Φ *contains no doubly negated formulas*) *iff both* $\Box\sim A$ *and* $\Box\sim B$ *are in* Φ.

It is obvious that each formula is contained in a finite adequate set. In proving completeness we can of course restrict our attention to formulas without double negations, and will therefore be able to use adequate sets with formulas without double negations, so that we can apply the last part of lemma 13.8. We will write **ILS** if our remarks apply to both **IL** and **ILM**.

13.9. Definition. Let Γ and Δ be maximal **ILS**-consistent subsets of some finite adequate Φ. Then $\Gamma < \Delta \iff$ for each $\Box A \in \Gamma$, $\Box A, A \in \Delta$, and, for some $\Box A \notin \Gamma$, $\Box A \in \Delta$. In this case we say that Δ is a *successor* of Γ (see the proof for **L** of theorem 2.4).

13.10. Definition. Let Γ and Δ be maximal **ILS**-consistent subsets of some given adequate Φ. Then Δ is a *C-critical successor* of Γ ($\Gamma <_C \Delta$) iff

(i) $\Gamma < \Delta$,
(ii) $\sim A, \Box\sim A \in \Delta$ for each A such that $A \rhd C \in \Gamma$.

13.11. Lemma. *If $\Gamma <_C \Delta$ and $\Delta < \Theta$, then $\Gamma <_C \Theta$.*

13.12. Lemma. *Suppose Γ is maximal* **ILS***-consistent in Φ and $\neg(B \rhd C) \in \Gamma$. Then there exists a C-critical successor Δ of Γ, maximal* **ILS***-consistent in Φ, such that $B \in \Delta$.*

Proof. Let Γ, Φ, B and C satisfy the conditions of the lemma. Take Δ to be a maximal **ILS**-consistent extension of

$$\{D, \Box D | \Box D \in \Gamma\} \cup \{\Box\sim A, \sim A | A \rhd C \in \Gamma\} \cup \{B, \Box\sim B\}.$$

Note first that the adequacy of Φ ensures that all the formulas in Δ are indeed available, and second that such a Δ, if it exists, is a C-critical successor of Γ. (It is a successor, because $\Box\sim B$ **IL**-implies $B \rhd C$ and, hence, cannot be a member of Γ.) To prove that such a Δ exists it is sufficient to prove that the above set is **IL**-consistent. Suppose not. Then there exist $A_1, \ldots, A_m$ and $D_1, \ldots, D_k$ with

$$D_1, \ldots, D_k, \Box D_1, \ldots, \Box D_k, \neg A_1, \ldots \neg A_m, \Box\neg A_1, \ldots, \Box\neg A_m, B, \Box\neg B \vdash \bot$$

or equivalently

$$D_1, \ldots, D_k, \Box D_1, \ldots, \Box D_k \vdash B \wedge \Box\neg B \to A_1 \vee \ldots \vee A_m \vee \Diamond(A_1 \vee \ldots \vee A_m).$$

Applying what we know of **L** gives

$$\Box D_1, \ldots, \Box D_k \vdash \Box(B \wedge \Box\neg B \to A_1 \vee \ldots \vee A_m \vee \Diamond(A_1 \vee \ldots \vee A_m)).$$

Axiom (1) then implies

$$\Box D_1, \ldots, \Box D_k \vdash B \wedge \Box\neg B \rhd A_1 \vee \ldots \vee A_m \vee \Diamond(A_1 \vee \ldots \vee A_m).$$

From lemmas and axiom (2) it follows then that

$$\Box D_1, \ldots, \Box D_k \vdash B \rhd A_1 \vee \ldots \vee A_m.$$

Given that $A_1 \rhd C, \ldots, A_m \rhd C \in \Gamma$, we also have, by using axiom (3) that $\Gamma \vdash A_1 \vee \ldots \vee A_m \rhd C$. So, finally, we obtain $\Gamma \vdash B \rhd C$ which contradicts the consistency of Γ. $\dashv$

13.13. Lemma. *Let $B \rhd C \in \Gamma$. Then, if there exists an E-critical successor Δ of Γ with $B \in \Delta$, there also exists an E-critical successor Δ' of Γ with $C, \Box\neg C \in \Delta'$.*

Proof. Suppose B, C, E, Γ and Δ satisfy the assumptions of the lemma and there is no such Δ'. Then there would be $\Box D_1, \ldots, \Box D_n \in \Gamma$, and $F_1 \rhd E, \ldots, F_k \rhd E \in \Gamma$ such that

$$D_1, \ldots, D_n, \Box D_1, \ldots, \Box D_n, \neg F_1, \ldots, \neg F_k, \Box\neg F_1, \ldots, \Box\neg F_k, C, \Box\neg C \vdash \bot$$

and, therefore,

$$D_1, \ldots, D_n, \Box D_1, \ldots, \Box D_n \vdash C \wedge \Box\neg C \to F_1 \vee \ldots \vee F_k \vee \Diamond(F_1 \vee \ldots \vee F_k),$$

which as before implies $\Gamma \vdash B \rhd E$. Since B and E are respectively an antecedent and a succedent of some $\rhd$-formula in Φ, the adequacy conditions imply then that this can be strengthened to $B \rhd E \in \Gamma$. As Δ is supposed to be an E-critical successor of Γ, this implies $\sim B \in \Delta$ and we have arrived at a contradiction. $\dashv$

13.14. Theorem. (Completeness and decidability of **IL**) *If $\nvdash_{\mathbf{IL}} A$, then there is a finite* **IL***-model K such that $K \not\Vdash A$.*

Proof. Take some finite adequate set Φ containing A, and let Γ be a maximal **IL**-consistent subset of Φ containing $\sim A$. The intuitive idea of the construction of the model is to divide the set of successors of each constructed world w, starting with Γ, into different parts, each part containing the E-critical successors w for some $\rhd$-succedent E in the adequate set. For occurrences of the same maximal consistent set in different parts we use distinct copies. The S_w are defined to be the universal relation inside each part consisting of the E-critical successors for some E, but to be such as to make no other connections between worlds. Then lemmas 13.12 and 13.13 give the theorem rather straightforwardly. With some care this program can be executed, but we take a slightly more complicated road that points the way to the completeness proof for **ILM** where the straightforward manner does not work.

Set W_Γ to be the smallest set of pairs $\langle \Delta, \tau \rangle$ with Δ a maximal consistent subset of Φ and τ a finite sequence of formulas from Φ that satisfy the following requirements:

(i) $\Gamma < \Delta$ or $\Gamma = \Delta$,

(ii) τ is a finite sequence of formulas from Φ, the length of which does not exceed the depth of Γ minus the depth of Δ. (So, e.g. Γ is only paired off with the empty sequence.)

It is clear that W_Γ is finite, since, for any Δ, if Δ' is a successor of Δ, then Δ' has fewer successors than Δ. We define R on W_Γ by $w\,R\,w'$ iff $(w)_0 < (w')_0$ and $(w)_1 \subseteq (w')_1$. The required properties check out easily. Let $u\,S_w v$ apply if (I) and (II) hold (writing $*$ for concatenation):

(I) $u, v \in w{\uparrow}$,

(II) either $(w)_1 = (u)_1 \subseteq (v)_1$, or $(u)_1 = (w)_1 * \langle C\rangle * \tau$ and $(v)_1 = (w)_1 * \langle C\rangle * \tau'$ for some C, τ, τ', and if in the latter case $(u)_0$ is a C-critical successor of $(w)_0$, then so is $(v)_0$.

Let us check that under this definition the S_w will have the properties (i)–(iii) required by definition 13.3:

(i) That S_w is a relation on $w{\uparrow}$ is instantaneous.

(ii) Reflexivity and transistivity of S_w are also easy to check.

(iii) If $w', w'' \in w{\uparrow}$ and $w'R\,w''$, then (I) is immediate. For (II) it suffices to recall that successors of C-critical successors are C-critical (lemma 13.11).

Finally, we define $w \Vdash p$ iff $p \in (w)_0$. We will now prove that, for each $B \in \Phi$ and $w \in W_\Gamma$, $w \Vdash B$ iff $B \in (w)_0$, by induction on the length of B. Of course, the connectives are trivial, so it suffices to prove that

$$B \rhd C \in (w)_0 \iff \forall u(w\,R\,u \wedge B \in (u)_0 \to \exists v(u\,S_w\,v \wedge C \in (v)_0)).$$

$\Longleftarrow$: Suppose $B \rhd C \notin (w)_0$. Then $\neg\,(B \rhd C) \in (w)_0$. We have to show that, for some u with $w\,R\,u$, $B \in (u)_0$ and $\forall v\,(u\,S_w v \to {\sim}C \in (v)_0)$. Let Δ with $(w)_0 <_C \Delta$ be as given by lemma 13.12, and take u to be $\langle \Delta, (w)_1 * \langle C\rangle\rangle$. It is clear that u fulfills the requirements.

$\Longrightarrow$: Suppose $B \rhd C \in (w)_0$. Consider any u such that $B \in (u)_0$ and $w\,R\,u$, and first assume $(u)_1 = (w)_1 * \langle E\rangle * \tau$ and $(u)_0$ is an E-critical successor of $(w)_0$. By lemma 13.13 we can find an E-critical successor Δ' of $(w)_0$ with $C \in \Delta'$. It is clear that $v = \langle \Delta', (w)_1 * \langle E\rangle\rangle$ is a member of W_Γ and fulfills all the requirements to make $u\,S_w v$.

If $(u)_1 = (w)_1 * \langle E\rangle * \tau$ but $(u)_0$ is not an E-critical successor of $(w)_0$, then we find a successor Δ' of $(w)_0$ with $C \in \Delta'$ by using axiom (4) instead of lemma 13.13. Again it is clear that $v = \langle \Delta', (w)_1 * \langle E\rangle\rangle$ is a member of W_Γ and fulfills all the requirements to make $u\,S_w v$. The final case is that $(u)_1 = (w)_1$. In that case also we apply axiom (4) to obtain Δ' with $C \in \Delta'$ and take $v = \langle \Delta', (w)_1\rangle$. $\dashv$

13.15. Theorem. (Completeness and decidability of **ILM**) *If* $\nvdash_{\mathbf{ILM}} A$, *then there is a finite* **ILM***-model* K *such that* $K \not\models A$.

The main problem in the proof of this theorem is the following. To apply the characteristic axiom $(A \rhd B) \to (A \wedge \Box C \rhd B \wedge \Box C)$ we seem to be forced to add the succedent of this formula to the adequate set whenever we have the antecedent. A straightforward definition of adequate set for the case of **ILM** would therefore lead adequate sets to be always infinite, which is of course unacceptable. After some searching we are lead to the following definition.

13.16. Definition. An **ILM**-*adequate* set Φ is an adequate set that satisfies the additional condition:

if $B \rhd C, \Box D \in \Phi$, then there is in Φ a formula $B' \rhd C'$ such that
B' is **ILM**-equivalent to $B \wedge \Box D$ and C' to $C \wedge \Box D$.

Even though we require only equivalents to be present in Φ it is of course no longer evident that each finite set of formulas is contained in a finite **ILM**-adequate set, since each newly constructed $B \wedge \Box D$ gives rise to a new $\Box$-formula: $B \wedge \Box D \rhd \bot$. But we will show that this is nevertheless true. To make it easier on ourselves we assume that in our formula A all antecedents and succedents of $\rhd$-formulas have the form $B \wedge \Box \sim B$, except for $\bot$. In view of corollary 13.2(a) this is not an essential restriction. (The restriction is not really necessary, see Berarducci [1990].)

13.17. Lemma. *Each formula A is contained in an* **ILM**-*adequate set Φ that contains only a finite number of* **ILM**-*equivalence classes.*

Proof. Let Φ be the smallest **IL**-adequate set containing A. Let Ψ be the set of antecedents and succedents of $\rhd$-formulas in Φ including $\bot$. We obtain Ψ^* by closing Ψ off under the operation that forms $D \wedge E$ from each formula D in the class and each formula E that, either is a $\Box$-formula in Φ, or is of the form $\Box \sim F$ for some F in the class. The claim is that Ψ^* contains only a finite number of equivalence classes. Given that claim we can construct a finite **ILM**-adequate set by joining to Φ the subformulas of a finite set of representatives of all equivalence classes in Ψ^*, and finally adding all the interpretability formulas combining two members of this finite set of representatives.

It remains to prove the claim. This will be done by induction on the cardinality of Ψ. If that cardinality is 1 (i.e., $\Psi = \{\bot\}$), the result is obvious. So, we can assume that the cardinality is larger than 1. We note that each element of Ψ^* is of the form $B \wedge \Box \sim B \wedge \Box C_1 \wedge \ldots \wedge \Box C_k$, with $B \wedge \Box \sim B$ from Ψ. That $\Box \sim B$ is a member of this conjunction means that in the C_i's all occurrences of $B \wedge \Box \sim B$ can be replaced by $\bot$. Also one will recognize that $B \wedge \Box \sim B$ will only be thrown in by the operation into the C_i in conjuncts of the form $\neg(B \wedge \Box \sim B \wedge \ldots)$. Replacing those occurrences of $B \wedge \Box \sim B$ by $\bot$ means that one can drop the whole conjunct and keep an equivalent formula. If one drops all those conjuncts containing $B \wedge \Box \sim B$, then the resulting formula is of the form $B \wedge \Box \sim B \wedge \Box D_1 \wedge \ldots \wedge \Box D_m$ with $B \wedge \Box \sim B$ not (relevantly) occurring in the D_i. This means that the D_i have been constructed from the $\Box$-formulas in Φ and the other elements of Ψ. Thus, by the induction hypothesis, there are only a finite number of such D_i (up to equivalence) and hence only a finite number of equivalence classes of elements of Ψ^* that start with $B \wedge \Box \sim B$. The same holds for each of the other elements of Ψ, so that the resulting set is finite. $\dashv$

Proof of theorem 13.15. Take some finite **ILM**-adequate set Φ containing A and some maximal consistent subset Γ of Φ containing $\sim A$. We define both W_Γ and R

as in the previous proof. This time, however, we let $u\,S_w v$ apply if (I) holds as well as (II′) and (III),

(II′) $(u)_1 \subseteq (v)_1$, and if $(u)_1 = (w)_1 * \langle C \rangle * \tau$ and $(v)_1 = (w)_1 * \langle C \rangle * \tau'$ for some C, τ, τ', and $(u)_0$ is a C-critical successor of $(w)_0$, then so is $(v)_0$.

(III) each $\Box A \in (u)_0$ is also a member of $(v)_0$,

That under this definition the S_w will have the properties (i)–(iii) is shown in almost the same manner as before; that the S_w has the property (iv) required by definition 13.5 is shown as follows:

Suppose that $\langle \Delta', \tau' \rangle S_w \langle \Delta'', \tau'' \rangle R \langle \Gamma', \sigma \rangle$. We must show $\langle \Delta', \tau' \rangle R \langle \Gamma', \sigma \rangle$. That $\tau' \subseteq \sigma$, is immediate. That $\Delta' < \Gamma'$, follows from $\Delta'' < \Gamma'$ combined with the fact that, by (III), $\Box$-formulas are preserved from Δ' to Δ''.

Naturally, we again define $w \Vdash p$ iff $p \in (w)_0$, and it will be sufficient to prove that, for each $D \in \Phi$, $w \Vdash D$ iff $D \in (w)_0$. The only interesting case is the one that D is $B \rhd C$, i.e., we have to show that

$$B \rhd C \in (w)_0 \iff \forall u (w\,R\,u \wedge B \in (u)_0 \to \exists v (u\,S_w v \wedge C \in (v)_0)).$$

$\Longleftarrow$: Basically as in the proof for **IL**.

$\Longrightarrow$: Assume that $B \rhd C \in (w)_0$, and that u is such that $w\,R\,u$ and $B \in (u)_0$. Let $\{\Box D_1, \ldots, \Box D_n\}$ be the set of $\Box$-formulas in $(u)_0$. By axiom M (see proposition 2.1(d)) and the adequacy of Φ, $(w)_0$ will contain a formula $B' \rhd C'$ with B' and C' respectively **ILM**-equivalent to $B \wedge \Box D_1 \wedge \ldots \wedge \Box D_n$ and $C \wedge \Box D_1 \wedge \ldots \wedge \Box D_n$.

Let us just treat the case that $(u)_1 = (w)_1 * \langle E \rangle * \tau$ and $(u)_0$ is an E-critical successor of $(w)_0$. (The other cases are easy, given our experience with **IL**.) We can find, by lemma 13.13, with $(w)_0$, $(u)_0$ and $B' \rhd C'$ as input, an E-critical successor Δ' of $(w)_0$ with both C and $\Box D \in \Delta'$ for each $\Box D \in (u)_0$. It suffices to take $v = \langle \Delta', (u)_1 \rangle$. Given that each $\Box$-formula in $(u)_0$ appears also in Δ', the depth of Δ' cannot be larger than the depth of $(u)_0$. Therefore, $v \in W_\Gamma$ and v fulfills all requirements. $\dashv$

Visser (see Berarducci [1990]) showed that, from the models constructed in the above proof, one can construct models with an S relation that is independent of the world w (see also definition 15.4). These models may have to be infinite however. The first arithmetic completeness proofs used these models instead of the finite models constructed in the above proof, but we will not introduce them in this section, since our arithmetic completeness proof (section 14) uses the finite models directly.

The fixed point theorem of **L** can be extended to **IL** and hence to **ILM** and **ILP** (de Jongh and Visser [1991]).

14. Arithmetic completeness of ILM

We fix a theory T containing $\mathbf{I\Sigma_1}$. For safety we assume that T is in the language of arithmetic and T is sound, i.e., all its axioms are true (in the standard model

of arithmetic), although in fact it is easy to adjust our proof of the completeness theorem to the weaker condition of Σ_1-soundness of T.

14.1. Definition. The definition of a *realization* given in section 1 is extended to the language of **ILM** by stipulating that $(A \rhd B)^* = \mathrm{Conserv}(\ulcorner A^* \urcorner, \ulcorner B^* \urcorner)$, where $\mathrm{Conserv}(\ulcorner A^* \urcorner, \ulcorner B^* \urcorner)$ is an intensional formalization (see Chapter II of this Handbook) of "$T + B^*$ is Π_1-conservative over $T + A^*$".

If $T = \mathbf{PA}$, then, in view of theorem 12.7, the interpretability and Π_1-conservativity relations over its finite extensions are the "same" in all reasonable senses, so we can take $\mathrm{Conserv}(\ulcorner A^* \urcorner, \ulcorner B^* \urcorner)$ to be a formalization of "$T + B^*$ is interpretable in $T + A^*$". Below we prove the completeness of **ILM** as the logic of Π_1-conservativity over T and thus at the same time the completeness of **ILM** as the logic of interpretability over $T = \mathbf{PA}$. The fact that **ILM** is the logic of interpretability over **PA** was proven more or less simultaneously and independently by Berarducci [1990] and Shavrukov [1988]. Later, Hájek and Montagna [1990,1992] proved that **ILM** is the logic of Π_1-conservativity over $T = \mathbf{I\Sigma_1}$ and stronger theories.

14.2. Theorem. $\vdash_{\mathbf{ILM}} A$ *iff for every realization* *, $T \vdash A^*$.

Proof. The ($\Longrightarrow$) part can be verified by induction on **ILM** proofs. Since the soundness of **L** is already known, we only need to verify that if D is an instance of one of the additional 6 axiom schemata of **ILM**, then, for any realization *, $T \vdash D^*$. All the arguments below are easily formalizable in T:

Axiom (1): $\Box(A \to B) \to (A \rhd B)$. If $T \vdash A \to B$, then clearly $T + B^*$ is conservative over $T + A^*$.

Axiom (2): $(A \rhd B) \wedge (B \rhd C) \to (A \rhd C)$. Evidently, the relation of conservativity is transitive.

Axiom (3): $(A \rhd C) \wedge (B \rhd C) \to A \vee B \rhd C$. It is easy to see that if $T + C^*$ is (Π_1-) conservative over $T + A^*$ and $T + B^*$, then so is it over $T + A^* \vee B^*$.

Axiom (4): $(A \rhd B) \to (\Diamond A \to \Diamond B)$. Clearly, if $T + B^*$ is Π_1-conservative over $T + A^*$ and $T + A^*$ is consistent, then so is $T + B^*$.

Axiom (5): $\Diamond A \rhd A$. Suppose λ is a Π_1!-sentence provable in $T + A^*$. We need to show, arguing in $T + (\Diamond A)^*$, that then λ is true. Indeed, suppose $T + A^*$ is consistent. Then it cannot prove a false Π_1!-sentence (by Σ_1!-completeness), and hence λ must be true.

Axiom (M): $(A \rhd B) \to (A \wedge \Box C \rhd B \wedge \Box C)$. Suppose $T + B^*$ is Π_1-conservative over $T + A^*$ and λ is a Π_1!-sentence provable in $T + B^* \wedge (\Box C)^*$. Then $T + B^*$ proves $(\Box C)^* \to \lambda$. But the latter is a Π_1-sentence and therefore it is also proved by $T + A^*$. Hence, $T + A^* \wedge (\Box C)^* \vdash \lambda$.

The following proof of the ($\Longleftarrow$) part of the theorem is taken from Japaridze [1994b] and has considerable similarity to proofs given in Japaridze [1992,1993] and Zambella [1992]. Just as in Japaridze [1992,1993], the Solovay function is defined in terms of regular witnesses rather than provability in finite subtheories (as in

Berarducci [1990], Shavrukov [1988], Zambella [1992]). Disregarding this difference, the function is almost the same as the one given in Zambella [1992], for both proofs, unlike the ones in Berarducci [1990] and Shavrukov [1988], employ finite **ILM**-models rather than infinite Visser-models.

Suppose $\nvdash_{\mathbf{ILM}} A$. Then, by theorem 13.15, there is a finite **ILM**-model $\langle W, R, \{S_w\}_{w \in W}, \Vdash\rangle$ in which A is not valid. We may assume that $W = \{1, \ldots, l\}$, 1 is the root of the model in the sense that $1 R\, w$ for all $1 \neq w \in W$, and $1 \nVdash A$. We define a new frame $\langle W', R', \{S'_w\}_{w \in W'}\rangle$:

$$W' = W \cup \{0\},$$
$$R' = R \cup \{(0, w) \mid w \in W\}.$$
$$S'_0 = S_1 \cup \{(1, w) \mid w \in W\} \text{ and for each } w \in W,\ S'_w = S_w.$$

Observe that $\langle W', R', \{S'_w\}_{w \in W'}\rangle$ is a finite **ILM**-frame.

Just as in section 3, we are going to embed this frame into T by means of a Solovay style function $g : \omega \to W'$ and sentences Lim_w for $w \in W'$ which assert that w is the limit of g. This function will be defined in such a way that the following basic lemma holds:

14.3. Lemma.
(a) *T proves that g has a limit in W', i.e., $T \vdash \bigvee\{\mathrm{Lim}_r \mid r \in W'\}$,*
(b) *If $w \neq u$, then $T \vdash \neg(\mathrm{Lim}_w \wedge \mathrm{Lim}_u)$,*
(c) *If $w\, R'\, u$, then $T + \mathrm{Lim}_w$ proves that $T \nvdash \neg \mathrm{Lim}_u$,*
(d) *If $w \neq 0$ and not $w\, R'\, u$, then $T + \mathrm{Lim}_w$ proves that $T \vdash \neg \mathrm{Lim}_u$,*
(e) *If $u\, S'_w v$, then $T + \mathrm{Lim}_w$ proves that $T + \mathrm{Lim}_v$ is Π_1-conservative over $T + \mathrm{Lim}_u$,*
(f) *Suppose $w\, R'\, u$ and V is a subset of W' such that for no $v \in V$, $u\, S_w v$; then $T + \mathrm{Lim}_w$ proves that $T + \bigvee\{\mathrm{Lim}_v \mid v \in V\}$ is not Π_1-conservative over $T + \mathrm{Lim}_u$,*
(g) *Lim_0 is true,*
(h) *For each $i \in W'$, Lim_i is consistent with T.*

To deduce the main thesis from this lemma, we define a realization * by setting for each propositional letter p,

$$p^* = \bigvee\{\mathrm{Lim}_r \mid r \in W,\ r \Vdash p\}.$$

14.4. Lemma. *For any $w \in W$ and any* **ILM***-formula B,*
(a) *if $w \Vdash B$, then $T + \mathrm{Lim}_w \vdash B^*$;*
(b) *if $w \nVdash B$, then $T + \mathrm{Lim}_w \vdash \neg B^*$.*

Proof. By induction on the complexity of B. The cases when B is atomic or has the form $\Box C$ are handled just as in the proof of lemma 3.3, so we consider only the case when $B = C_1 \rhd C_2$.

Assume $w \in W$. Then we can always write $w\, R\, x$ and $x\, S_w\, y$ instead of $w\, R'\, x$ and $x\, S'_w\, y$. Let $\alpha_i = \{r \mid w\, R\, r,\ r \Vdash C_i\}$ $(i = 1, 2)$. First we establish that for both $i = 1, 2$,

$(*)$ $\quad T + \mathrm{Lim}_w$ proves that $T \vdash C_i^* \leftrightarrow \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_i\}$.

Indeed, argue in $T + \mathrm{Lim}_w$.

Since each $r \in \alpha_i$ forces C_i, we have by the induction hypothesis for clause (a) that for each such r, $T \vdash \mathrm{Lim}_r \to C_i^*$, whence $T \vdash \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_i\} \to C_i^*$. Next, according to lemma 14.3(a), $T \vdash \bigvee\{\mathrm{Lim}_r \mid r \in W'\}$ and, according to lemma 14.3(d), T disproves every Lim_r with *not* $wR\,r$; consequently, $T \vdash \bigvee\{\mathrm{Lim}_r \mid wR\,r\}$; at the same time, by the induction hypothesis for clause (b), C_i^* implies in T the negation of each Lim_r with $r \nVdash C_i$. We conclude that $T \vdash C_i^* \to \bigvee\{\mathrm{Lim}_r \mid wR\,r,\ r \Vdash C_i\}$, i.e., $T \vdash C_i^* \to \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_i\}$. Thus, $(*)$ is proved. Now continue:

(a) Suppose $w \Vdash C_1 \rhd C_2$. Argue in $T + \mathrm{Lim}_w$. By $(*)$, to prove that $T + C_2^*$ is Π_1-conservative over $T + C_1^*$, it is enough to show that $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_2\}$ is Π_1-conservative over $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_1\}$. Consider an arbitrary $u \in \alpha_1$ (the case with empty α_1 is trivial, for any theory is conservative over $T + \bot$). Since $w \Vdash C_1 \rhd C_2$, there is $v \in \alpha_2$ such that $u\,S_w\,v$. Then, by lemma 14.3(e), $T + \mathrm{Lim}_v$ is Π_1-conservative over $T + \mathrm{Lim}_u$. Then so is $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_2\}$ (which is weaker than $T + \mathrm{Lim}_v$). Thus, for each $u \in \alpha_1$, $\quad T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_2\}$ is Π_1-conservative over $T + \mathrm{Lim}_u$. Clearly this implies that $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_2\}$ is Π_1-conservative over $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_1\}$.

(b) Suppose $w \nVdash C_1 \rhd C_2$. Let us then fix an element u of α_1 such that $u\,S_w\,v$ for no $v \in \alpha_2$. Argue in $T + \mathrm{Lim}_w$.

By lemma 14.3(f), $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_2\}$ is not Π_1-conservative over $T + \mathrm{Lim}_u$. Then, neither is it Π_1-conservative over $T + \bigvee\{\mathrm{Lim}_r \mid r \in \alpha_1\}$ (which is weaker than $T + \mathrm{Lim}_u$). This means by $(*)$ that $T + C_2^*$ is not Π_1-conservative over $T + C_1^*$. $\dashv$

Now we can pass to the desired conclusion: since $1 \nVdash A$, lemma 14.4 gives $T \vdash \mathrm{Lim}_1 \to \neg A^*$, whence $T \nvdash \neg\,\mathrm{Lim}_1 \Rightarrow T \nvdash A^*$. But we do have $T \nvdash \neg\,\mathrm{Lim}_1$ according to lemma 14.3(h). This ends the proof of theorem 14.2. $\dashv$

Our remaining duty is to define the function g and to prove lemma 14.3. The recursion theorem enables us to define this function simultaneously with the sentences Lim_w (for each $w \in W'$), which, as we have mentioned already, assert that w is the limit of g, and the formulas $\Delta_{wu}(y)$ (for each pair (w, u) with $wR'u$), which we define by

$\Delta_{wu}(y) \equiv \exists\, t > y\,(g(t) = \bar{u} \wedge \forall z(y \leqslant z < t \to g(z) = \bar{w}))$.

14.5. Definition. (function g)
We define $g(0) = 0$.

Assume that $g(y)$ has already been defined for every $y \leqslant x$, and let $g(x) = w$. Then $g(x+1)$ is defined as follows:

(1) Suppose $wR'u$, $n \leqslant x$ and for all z with $n \leqslant z \leqslant x$ we have $g(z) = w$. Then, if $\vdash_x \mathrm{Lim}_u \to \neg\,\Delta_{wu}(\bar{n})$, we define $g(x+1) = u$.

(2) Else, suppose $m \leqslant x$, λ is a Π_1!-sentence and the following holds:

(a) λ has a regular counterwitness which is $\leqslant x$,
(b) $\vdash_m \mathrm{Lim}_u \to \lambda$,
(c) $w\, S_{g(m)}\, u$,
(d) m is the least number for which such λ and u exist, i.e., there are no $m' < m$, world u' and Π_1!-sentence λ' satisfying the conditions (a)–(c) with m', u' and λ' substituted for m, u and λ.
Then we define $g(x+1) = u$.

(3) In all remaining cases $g(x+1) = g(x)$.

It is not hard to see that g is primitive recursive. Before we start proving lemma 14.3, let us agree on some jargon and prove two auxiliary lemmas. When the transfer from $w = g(x)$ to $u = g(x+1)$ is determined by definition 14.5(1), we say that at the moment $x+1$ the function g makes (or we make) an R'-*move* from the world w to the world u. If this transfer is determined by definition 14.5(2), then we say that an S'-*transfer* takes place and call the number m from definition 14.5(2) the *rank* of this S'-transfer. Sometimes the S'-transfer leads to a new world, but 'mostly' it does not, i.e., $(u =) g(x+1) = g(x) (= w)$, and then it is not a move in the proper sense. Those S'-transfers which lead to a new world we call S'-*moves*. As for R'-transfers, they (by irreflexivity of R') always lead to a new world, so we always say "R'-move" instead of "R'-transfer".

In these terms, the formula $\Delta_{wu}(n)$ asserts that starting at or before the moment n and until some moment t, we stay at the world w without moving and then, at the moment t, we move directly to u.

Intuitively, we make an R'-move from w to some u with $wR'u$ in the following situation: since some moment n and up to the present we have been staying at world w, and just now we have reached evidence that $T + \mathrm{Lim}_u$ thinks that the first (proper) move which happens after passing moment n (and thus our next move) cannot lead directly to the world u; then, to spite this belief of $T + \mathrm{Lim}_u$, we immediately move to u.

And the conditions for an S'-transfer from w to u can be described as follows: we are staying at the world w and by the present moment we have reached evidence that $T + \mathrm{Lim}_u$ proves a false Π_1!-sentence λ. This evidence consists of two components: (1) a regular counterwitness, which indicates that λ is false, and (2) the rank m of the transfer, which indicates that $T + \mathrm{Lim}_u \vdash \lambda$. Then, as soon as $w\, S_{g(m)}\, u$, the next moment we must be at u (move to u, if $u \neq w$, and remain at w, if $u = w$); if there are several possibilities for such a transfer, we choose the one with the least rank. An additional necessary condition for an S'-transfer is that in the given situation an R'-move is impossible; R'-moves have priority over S'-moves.

Note that the condition for an R'-move here is weaker than for the function h defined in section 3: T only needs to prove $\mathrm{Lim}_u \to \neg\Delta_{wu}(\bar{n})$. This feature will play a crucial role in the verification of 14.3(f).

14.6. Lemma. $(T \vdash :)$ *For each natural number m and each $w \in W'$, $T + \mathrm{Lim}_w$ proves that no S'-transfer to w can have rank less than m.*

Proof. Indeed, "the rank of an S'-transfer is $< m$" means that $T + \mathrm{Lim}_w$ proves a false (i.e., one with a regular counterwitness) $\Pi_1!$-sentence λ and the code of this proof (i.e., of the T-proof of $\mathrm{Lim}_w \to \lambda$) is smaller than m. But the number of all $\Pi_1!$-sentences with such short proofs is finite, and as $T + \mathrm{Lim}_w$ proves each of them, it also proves that none of these sentences has a regular counterwitness (recall our assumptions about the formula $\mathrm{Regwit}(x, y)$ from section 12). ⊣

14.7. Lemma. $(T \vdash :)$ *If $g(x)R'w$, then for all $y \leqslant x$, $g(y)R'\,w$.*

Proof. Suppose $g(x)R'\,w$ and $y \leqslant x$. We proceed by induction on $n = x - y$. If $y = x$, we are done. Suppose now $g(y+1)R'\,w$. If $g(y) = g(y+1)$, we are done. If not, then at the moment $y+1$ the function makes either an R'-move or an S'-move. In the first case we have $g(y)R'\,g(y+1)$ and, by transitivity of R', $g(y)R'\,w$; in the second case we have $g(y)S'_v\,g(y+1)$ for some v, and the desired thesis then follows from property (iv) of **ILM**-frames (definition 13.4). ⊣

Proof of lemma 14.3. In each case below, except in (g) and (h), we reason in T.

(a): First observe that there exists some z such that for all $z' \geqslant z$, not $g(z')\,R'\,g(z'+1)$. Indeed, suppose this is not the case. Then, by lemma 14.7, for all z, there is z' with $g(z)R'\,g(z')$. This means that there is an infinite (or "sufficiently long") chain $w_1R'\,w_2R'\cdots$, which is impossible because W' is finite and R' is transitive and irreflexive.

So, let us fix this number z. Then we never make an R'-move after the moment z. We claim that S'-moves can also take place at most a finite number of times (whence it follows that g has a limit and this limit is, of course, one of the elements of W').

Indeed, let $x+1$ be an arbitrary moment after z at which we make an S'-move, and let m be the rank of this move. That is, for some $\Pi_1!$-sentence λ with a $\leqslant x$ regular counterwitness, we have $\vdash_m \mathrm{Lim}_u \to \lambda$ and $wS_{g(m)}u$, where $w = g(x)$ and $u = g(x+1)$. Suppose we make the next S'-move, with rank m', at some moment $x'+1$, $x' > x$, from the world u to a world v, $v \neq u$. Since $S_{g(m)}$ is reflexive, conditions (a)-(c) of definition 14.5(2) hold for x', u, u, λ, m in the roles of x, w, u, λ, m, respectively, and then, according to condition (d) of definition 14.5(2), the only reason for moving to v instead of u — instead of remaining at u, that is — could be that $m > m'$ (the case $m = m'$ is ruled out because $\mathrm{Lim}_u \neq \mathrm{Lim}_v$). Similarly, the rank m'' of the following S'-move will be less than m', etc. Thus, consecutive S'-moves without an R'-move between them have decreasing ranks. Therefore, S'-moves can take place at most m times after passing x.

(b): Clearly g cannot have two different limits w and u.

(c): Assume w is the limit of g and $w\,R'\,u$. Let n be such that for all $x \geqslant n$, $g(x) = w$. We need to show that $T \nvdash \neg\,\mathrm{Lim}_u$. Deny this. Then $T \vdash \mathrm{Lim}_u \to \neg\,\Delta_{wu}(\bar{n})$ and, since every provable formula has arbitrary long proofs, there is $x \geqslant n$ such that $\vdash_x \mathrm{Lim}_u \to \neg\,\Delta_{wu}(\bar{n})$; but then, according to definition 14.5(1), we must have $g(x+1) = u$, which, as $u \neq w$ (by irreflexivity of R'), is a contradiction.

(d): Assume $w \neq 0$, w is the limit of g and not $wR'\,u$. If $u = w$, then (since $w \neq 0$) there is x such that $g(x) = v \neq u$ and $g(x+1) = u$. This means that at the

moment $x+1$ we make either an R'-move or an S'-move. In the first case we have $T \vdash \mathrm{Lim}_u \to \neg\, \Delta_{vu}(\bar{n})$ for some n for which, as it is easy to see, the Σ_1!-sentence $\Delta_{vu}(\bar{n})$ is true, whence, by Σ_1!-completeness, $T \vdash \neg\, \mathrm{Lim}_u$. And if an S'-move is taken, then again $T \vdash \neg\, \mathrm{Lim}_u$ because $T + \mathrm{Lim}_u$ proves a false (with a $\leqslant x$ regular counterwitness) Π_1!-sentence.

Next, suppose $u \neq w$. Let us fix a number z with $g(z) = w$. Since g is primitive recursive, T proves that $g(z) = w$. Now argue in $T + \mathrm{Lim}_u$: since u is the limit of g and $g(z) = w \neq u$, there is a number x with $x \geqslant z$ such that $g(x) \neq u$ and $g(x+1) = u$. Since not $(w =)\, g(z) R'\, u$, we have by lemma 14.7 that

$(*)$ for each y with $z \leqslant y \leqslant x$, not $g(y) R'\, u$.

In particular, not $g(x) R'\, u$ and the transfer from $g(x)$ to $g(x+1)(= u)$ can have been determined only by definition 14.5(2). Then $(*)$ together with the property (i) of **IL**-frames and definition14.5(2c), implies that the rank of this S'-move is less than z, which, by lemma 14.6, is a contradiction. Thus, $T + \mathrm{Lim}_u$ is inconsistent, i.e., $T \vdash \neg\, \mathrm{Lim}_u$.

(e): Assume $u\, S'_w\, v \neq u$ (the case $v = u$ is trivial). Suppose w is the limit of g, λ is a Π_1-sentence and $T \vdash_z \mathrm{Lim}_v \to \lambda$. We may suppose that $\lambda \in \Pi_1$! and that z is sufficiently large, namely, $g(z) = w$. Fix this z. We need to show that $T + \mathrm{Lim}_u \vdash \lambda$.

Argue in $T + \mathrm{Lim}_u$. Suppose not λ. Then there is a regular counterwitness c for λ. Let us fix a number $x > z, c$ such that $g(x) = g(x+1) = u$ (as u is the limit of g, such a number exists). Then, according to definition 14.5, the only reason for $g(x+1) = u \neq v$ can be that we make an S'-transfer from u to u and the rank of this transfer is less than z, which, by lemma 14.6, is not the case. Conclusion: λ (is true).

(f): Assume w is the limit of g, $wR'\, u$, $V \subseteq W'$ and for each $v \in V$, not $u\, S'_w\, v$. Let n be such that for all $z \geqslant n$, $g(z) = w$. By primitive recursiveness of g, T proves that $g(n) = w$. By definition 14.5(1), $T + \mathrm{Lim}_u \nvdash \neg\, \Delta_{wu}(\bar{n})$. So, as $\neg\, \Delta_{wu}(\bar{n})$ is a Π_1-sentence, in order to prove that $T + \bigvee\{\mathrm{Lim}_v \mid v \in V\}$ is not Π_1-conservative over $T + \mathrm{Lim}_u$, it is enough to show that for each $v \in V$, $T + \mathrm{Lim}_v \vdash \neg\, \Delta_{wu}(\bar{n})$. Let us fix any $v \in V$. According to our assumption, not $u\, S'_w\, v$ and, by reflexivity of S'_w, $u \neq v$.

Argue in $T + \mathrm{Lim}_v$. Suppose, for a contradiction, that $\Delta_{wu}(n)$ holds, i.e., there is $t > n$ such that $g(t) = u$ and for all z with $n \leqslant z < t$, $g(z) = w$. As v is the limit of g and $v \neq u$, there is $t' > t$ such that $g(t'-1) \neq v$ and at the moment t' we arrive at v to stay there for ever. Let then $x_0 < \ldots < x_k$ be all the moments in the interval $[t, t']$ at which R'- or S'-moves take place, and let $u_0 = g(x_0), \ldots, u_k = g(x_k)$. Thus $t = x_0$, $t' = x_k$, $u = u_0$, $v = u_k$ and $u_0, \ldots, u_k$ is the route of g after departing from w (at the moment t).

Now let j be the least number among $1, \ldots, k$ such that for all $j \leqslant i \leqslant k$, not $u_0\, R'\, u_i$. Note that such a j does exist because at least $j = k$ satisfies the condition (otherwise, if $(u =)\, u_0\, R'\, u_k\, (= v)$, property (iv) of **ILM**-frames would imply $u\, S'_w\, v$). Note also that, for each i with $j \leqslant i \leqslant k$, the move to u_i cannot be an R'-move. Otherwise, we must have $u_{i-1} R'\, u_i$, whence, by lemma 14.7, $u_0\, R'\, u_i$, which is impossible for $i \geqslant j$.

Thus, from the moment x_j onwards, each move is an S'-move. Moreover, for each i with $j \leqslant i \leqslant k$, the rank of the S'-move to u_i is less than x_0. Indeed, suppose, for such an i, the rank of the S'-move to u_i is m for $m \geqslant x_0$. We have $g(m) = u_e$ for some e with $0 \leqslant e \leqslant k$ and, by definition 14.5(2c), we should have $u_{i-1} S'_{u_e} u_i$, and by property (i) of **IL**-frames, $u_e R' u_i$, whence, as above, lemma 14.7 gives that $u_0 \, R' \, u_i$, which is impossible for $i \geqslant j$. On the other hand, since consecutive S'-moves decrease the rank (as we noted in the proof of (a) above), and since the rank of the S'-move to u_k cannot be less than n (lemma 14.6), we conclude: for each i with $j \leqslant i \leqslant k$, the rank of the S'-move to u_i is in the interval $[n, x_0 - 1]$. But the value of g in this interval is w, and by definition 14.5(2c) this means that $u_{j-1} \, S'_w \, u_j \, S'_w \ldots S'_w \, u_k$. At the same time, we have either $u_0 = u_{j-1}$ or $u_0 \, R' \, u_{j-1}$. In both cases we then have $u_0 \, S'_w \, u_{j-1}$ (in the first case by reflexivity of S'_w and in the second case by property (iii) of **IL**-frames), whence, by transitivity of S'_w, $u_0 \, S'_w \, u_k$, i.e., $u \, S'_w \, v$, which is a contradiction. Conclusion: $T + \mathrm{Lim}_v \vdash \neg \, \Delta_{wu}(\bar{n})$.

(g): By (a), as T is sound, one of the Lim_w for $w \in W'$ is true. Since for no w do we have $wR'w$, (d) means that each Lim_w, except Lim_0, implies in T its own T-disprovability and therefore is false. Consequently, Lim_0 is true.

(h): As 3.2(f). $\dashv$

The proof of theorem 14.2 is complete. In de Jongh and Pianigiani [1998] this theorem and its extension to an interpretability logic with witness comparison formulas (Hájek and Montagna [1992]) was applied to solve a conjecture of Guaspari [1983]. This conjecture stated that those formulas of modal logic that under each arithmetic realization are interpreted as Σ_1-sentences are **L**-equivalent to disjunction of $\Box$-sentences (already proved in Visser [1995]), and those of modal logic extended with witness comparison formulas are **R**-equivalent to disjunctions which contain as their members conjunctions of witness comparison formulas and $\Box$-formulas. A companion paper is Beklemishev [1993a], in which it is shown that the realization of other formulas, i.e., the ones that are not always realized as Σ_1-sentences, cannot be restricted to any particular class in the arithmetic hierarchy, thereby improving Guaspari [1983]'s results as well.

Visser [1990] showed that **ILP** is the interpretability logic for all reasonable finitely axiomatizable theories that contain $\mathbf{I}\Delta_0 + \mathrm{SUPEXP}$. An open problem is the axiomatization of the logic of the principles valid for interpretability in all reasonable r.e. theories. Visser [1991] showed that this logic is not just the intersection of **ILM** and **ILP**.

15. Tolerance logic and other interpretability logics

15.1. The logics of cointerpretability and faithful interpretability

Unlike interpretability, no modal axiomatization for the logic of cointerpretability or faithful interpretability (over **PA** or any other reasonable theory) has been found so far. Even the question of decidability of these logics remains open.

However, the logics of weak interpretability and the more general relations of tolerance and cotolerance (see section 11) have been studied thoroughly. Here is a brief history of research in this field, which starts from some digression from the subject.

15.2. The logic of the arithmetic hierarchy

Japaridze [1990b,1994a] introduced a decidable propositional logic **HGL** with infinitely many unary modal operators: $\Box, \Sigma_1, \Sigma_1^+, \Sigma_2, \Sigma_2^+, \ldots$ and proved its soundness and completeness with respect of the arithmetic interpretation where $\Box A$ is understood as a formalization of "A^* is provable (in **PA**)", $\Sigma_n A$ as "A^* is (**PA**-equivalent to) a Σ_n-sentence" and $\Sigma_n^+ A$ as "A^* is (**PA**-equivalent to) a Boolean combination of Σ_n-sentences". The logic has a reasonable axiomatization and Kripke semantics.

15.3. The logic of tolerance and its fragments

Ignatiev [1990] (see Ignatiev [1993b]) strengthened the $(\Box, \Sigma_1)$-fragment of the logic of the arithmetic hierarchy by switching from the unary modal operator Σ_1 to the more general binary operator $\gg$, where $A \gg B$ is interpreted as "there is a Σ_1-sentence φ such that $\mathbf{PA} \vdash (A^* \to \varphi) \wedge (\varphi \to B^*)$" (for comparison: the interpretation of $\Sigma_1 A$ is nothing but "there is a Σ_1-sentence φ such that $\mathbf{PA} \vdash (A^* \to \varphi) \wedge (\varphi \to A^*)$"). He constructed a logic **ELH** in this language, called "the logic of Σ_1-interpolability", and proved its arithmetic completeness. Although the author of the logic of Σ_1-interpolability did not suspect this, he actually had found the logic of weak interpretability over **PA**, because, as it is now easy to see in view of corollary 12.8, the formula $\neg(A \gg \neg B)$ expresses that $\mathbf{PA} + B^*$ is weakly interpretable in $\mathbf{PA} + A^*$.

We know that weak interpretability is a special (binary) case of linear tolerance, and the latter is a special (linear) case of tolerance of a tree of theories. Japaridze [1992] gave an axiomatization of the logic **TOL** of linear tolerance over **PA**, and Japaridze [1993] did the same for the logic **TLR** of the most general relation of tolerance for trees.

All three logics **ELH**, **TOL** and **TLR** are decidable. Among them **TOL** has the most elegant language, axiomatization and Kripke semantics, and although **TOL** is just a fragment of **TLR**, here we are going to have a look only at this intermediate logic.

The language of **TOL** contains the single variable-arity modal operator $\diamondsuit$: for any n, if $A_1, \ldots, A_n$ are formulas, then so is $\diamondsuit(A_1, \ldots, A_n)$. This logic is defined as classical logic plus the rule $\neg A / \neg\diamondsuit(A)$ plus the following axiom schemata:

1. $\diamondsuit(\vec{C}, A, \vec{D}) \to \diamondsuit(\vec{C}, A \wedge \neg B, \vec{D}) \vee \diamondsuit(\vec{C}, B, \vec{D})$,

2. $\diamondsuit(A) \to \diamondsuit(A \wedge \neg\diamondsuit(A))$,

3. $\diamondsuit(\vec{C}, A, \vec{D}) \to \diamondsuit(\vec{C}, \vec{D})$,
4. $\diamondsuit(\vec{C}, A, \vec{D}) \to \diamondsuit(\vec{C}, A, A, \vec{D})$,
5. $\diamondsuit(A, \diamondsuit(\vec{C})) \to \diamondsuit(A \wedge \diamondsuit(\vec{C}))$,
6. $\diamondsuit(\vec{C}, \diamondsuit(\vec{D})) \to \diamondsuit(\vec{C}, \vec{D})$.

(Here $\vec{A}$ stands for $A_1, \ldots, A_n$ for an arbitrary $n \geqslant 0$, $\diamondsuit(\langle\rangle)$ is identified with $\top$.)

15.4. Definition. A *Visser-frame* (see Berarducci [1990]) is a triple $\langle W, R, S\rangle$, where $\langle W, R\rangle$ is a Kripke-frame for **L** and S is a transitive, reflexive relation on W such that $R \subseteq S$ and, for all $w, u, v \in W$, we have $wSuRv \Longrightarrow wRv$.

A **TOL**-*model* is a quadruple $\langle W, R, S, \Vdash\rangle$ with $\langle W, R, S\rangle$ a Visser-frame combined with a forcing relation $\Vdash$ with the clause

$w \Vdash \diamondsuit(A_1 \ldots, A_n)$ iff there are $u_1, \ldots, u_n$ with $u_1 S \ldots S u_n$ such that, for all i, wRu_i and $u_i \Vdash A_i$.

Such a model is said to be *finite*, if W is finite.

15.5. Theorem. (Japaridze [1992]) *For any* **TOL**-*formula* A, $\vdash_{\mathbf{TOL}} A$ *iff* A *is valid in every* **TOL**-*model; the same is true if we consider only finite* **TOL**-*models.*

15.6. Theorem. (Japaridze [1992]) *Let* T *be a sound superarithmetic theory, and let, for* * *an arithmetic realization,* $(\diamondsuit(A_1, \ldots, A_n))^*$ *be interpreted as a natural formalization of "the sequence* $T + A_1^*, \ldots, T + A_n^*$ *is tolerant". Then, for any* **TOL**-*formula* A, $\vdash_{\mathbf{TOL}} A$ *iff for every realization* *, $T \vdash A^*$.

With the arithmetic interpretation in mind, note that **L** is the fragment of **TOL** in which the arity of $\diamondsuit$ is restricted to 1. This is because consistency of A^* with T, expressed in **L** by $\diamondsuit A$, means nothing but tolerance of the one-element sequence $\langle T + A^*\rangle$ of theories, expressed in **TOL** by $\diamondsuit(A)$.

As for cotolerance, one can easily show, using theorems 12.7 and 12.13 ((i)$\Longleftrightarrow$(iii)), that a sequence of superarithmetic theories is cotolerant iff the sequence where the order of these theories is reversed is tolerant. Moreover, it was shown in Japaridze [1993] that cotolerance — though not tolerance — for trees can also be expressed in terms of linear tolerance. In particular, a tree of superarithmetic theories is cotolerant iff one of its topological sortings is. Hence, given a tree Tr of modal formulas, cotolerance of the corresponding tree of theories can be expressed in **TOL** by $\diamondsuit(\vec{A_1}) \vee \ldots \vee \diamondsuit(\vec{A_n})$, where $\vec{A_1}, \ldots, \vec{A_n}$ are all the reverse-order topological sortings of Tr. Thus **TOL**, being the logic of linear tolerance, can, at the same time, be viewed as the logic of (unrestricted) cotolerance over **PA**.

Just like tolerance, the notion of Γ-consistency (see definition 12.4) can be generalized to finite trees, including sequences as special cases of trees: a tree Tr of theories is Γ-*consistent* iff there are consistent extensions of these theories, of which each one is Γ-conservative over its predecessors in the tree.

Then the corollaries of theorems 12.7 and 12.13 generalize to the following:

15.7. Theorem. (Japaridze [1993], $\mathbf{PA} \vdash$) *For any finite tree Tr of superarithmetic theories,*

(a) *Tr is tolerant iff Tr is Π_1-consistent;*

(b) *Tr is cotolerant iff Tr is Σ_1-consistent.*

Just as in the case of **ILM**, in the arithmetic completeness theorems for **TOL** and **TLR**, the requirement of superarithmeticity (essential reflexivity) of T can be weakened to $\mathbf{I\Sigma_1} \subseteq T$ if we view these logics as logics of Π_1-consistency rather than tolerance.

15.8. Truth interpretability logics

We want to finish our discussion of propositional interpretability logics by noting that the closure under modus ponens of the set of theorems of **ILM**, or any other of the logics mentioned in this section, supplemented with the axiom $\Box A \to A$ or its equivalent, yields the logic (in case of **ILM** called $\mathbf{ILM}^\omega$) that describes all *true* principles expressible in the corresponding modal language, just as this was shown to be the case for **L** in section 3. The original sources usually contain proofs of both versions of the arithmetic completeness theorems for these logics.

Strannegård [1997] considers infinite r.e. sets of modal formulas of interpretability logic. He generalizes his theorem 5.3 for the specific case of interpretability over **PA** to the following theorem.

15.9. Theorem. *Let T be a well-specified r.e. set of formulas of interpretability logic. Then T is realistic iff it is consistent with $\mathbf{ILM}^\omega$.*

As in the case of **L** (corollary 5.2), a stronger version of this theorem implies as a corollary a uniform version of the arithmetic completeness of **ILM** with regard to **PA**. For a further consequence, let us first note that the existence of *Orey-sentences* in **PA**, i.e., arithmetic sentences A such that both $\mathbf{PA} + A$ and $\mathbf{PA} + \neg A$ are interpretable in **PA** (first obtained by Orey [1961]), follows immediately from the arithmetic completeness of **ILM** with regard to **PA**. In Strannegård's terminology this can be phrased as: Orey [1961] showed that the set $\{\top \rhd p, \top \rhd \neg p\}$ is realistic. Orey continued by asking what similar sets (such as $\{\top \rhd p,\ \top \rhd q,\ \top \rhd \neg(p \wedge q),$ $\neg(\top \rhd \neg p),\ \neg(\top \rhd \neg q),\ \neg(\top \rhd p \wedge q)\}$) are realistic. Let an *Orey set* be a set of modal formulas of the form $(\neg)(B \rhd C)$, where B and C are Boolean formulas. Strannegård can then give the following answer to Orey's question.

15.10. Theorem. *Let T be an r.e. Orey set. Then T is realistic iff it is consistent with $\mathbf{ILM}^\omega$.*

16. Predicate provability logics

16.1. The predicate modal language and its arithmetic interpretation

The language of predicate provability logic is that of first order logic (without identity or function symbols) together with the operator $\Box$. We assume that this language uses the same individual variables as the arithmetic language. Throughout this section T denotes a sound theory in the language of arithmetic containing **PA**. We also assume that T satisfies the Löb derivability conditions.

As in the previous sections, we want to regard each modal formula $A(P_1, ..., P_n)$ as a schema of arithmetic formulas arising from $A(P_1, ..., P_n)$ by substitution of arithmetic predicates $P_1^*, ..., P_n^*$ for the predicate letters $P_1, ..., P_n$ and replacing $\Box$ by $\mathrm{Pr}_T()$. However, some caution is necessary when we try to make this approach precise. In particular, we need to forbid for P_i^* to contain quantifiers that bind variables occurring in A.

16.2. Definition. A *realization* for a predicate modal formula A is a function * which assigns to each predicate symbol P of A an arithmetic formula $P^*(v_1, \ldots, v_n)$, whose bound variables do not occur in A and whose free variables are just the first n variables of the alphabetical list of the variables of the arithmetic language if n is the arity of P. For any realization * for A, we define A^* by the following induction on the complexity of A:

- in the atomic cases, $(P(x_1, \ldots, x_n))^* = P^*(x_1, \ldots, x_n)$,
- * commutes with quantifiers and Boolean connectives: $(\forall x B)^* = \forall x (B^*)$, $(B \to C)^* = B^* \to C^*$, etc.,
- $(\Box B)^* = \mathrm{Pr}_T[B^*]$.

For an explanation of the notation "[ ]" see notation 12.2. Observe from this that A^* always contains the same free variables as A. We say that an arithmetic formula φ is a *realizational instance* of a predicate modal formula A, if $\varphi = A^*$ for some realization * for A.

The main task is to investigate the set of predicate modal formulas which express valid principles of provability, i.e., all of whose realizational instances are provable, or true in the standard model.

16.3. The situation here is not as smooth as in the propositional case, ...

Having been encouraged by the impressive theorems of Solovay on the decidability of propositional provability logic, one might expect that the valid principles captured by the predicate modal language are also axiomatizable (decidability is ruled out of course). However, the situation here is not as smooth as in the propositional case. The first doubts about this were raised by Montagna [1984]. In fact, it turned out

afterwards that we have very strong negative results, one of which is the following theorem on nonarithmeticity of truth predicate logics of provability.

16.4. Theorem. (Artëmov [1985a]) *Suppose T is recursively enumerable. Then (for any choice of the provability predicate Pr_T) the set* Tr *of predicate modal formulas all of whose realizational instances are true, is not arithmetic.*

It was later shown by Vardanyan [1986], and also by Boolos and McGee [1987] that Tr is in fact Π_1-complete in the truth set of arithmetic.

Proof of theorem 16.4. We assume here that the arithmetic language contains one two-place predicate letter E and two three-place predicate letters A and M, and does not contain any other predicate, functional or individual letters. Thus, this language is a fragment of our predicate modal language. In the standard model $E(x,y)$, $A(x,y,z)$ and $M(x,y,z)$ are interpreted as the predicates $x=y$, $x+y=z$ and $x\times y=z$, respectively.

One variant of a well-known theorem of Tennenbaum (see e.g., Chapter 29 of Boolos and Jeffrey [1989]) asserts the existence of an arithmetic sentence β such that:

(1) β is true ("true" here always means "true in the standard model"),

(2) any model of β, with domain ω, E interpreted as the identity relation, and A and M as recursive predicates, is isomorphic to the standard model.

We assume that β conjunctively contains the axioms of Robinson's arithmetic **Q**, including the identity axioms. Therefore, using standard factorization, we can pass from any model D of β with domain ω and such that E, A and M are interpreted as recursive predicates, to a model D' which satisfies the conditions of (2) and which is elementarily equivalent to D. Thus, (2) can be changed to the following:

($2'$) any model D of β, with domain ω and E, A and M interpreted as recursive predicates, is elementarily equivalent to the standard model (i.e., $D\vDash\gamma$ iff γ is true, for all sentences γ).

Let C be the formula

$$\begin{array}{l}\forall x,y\,(\Box E(x,y)\vee\Box\neg E(x,y))\wedge\\ \quad\forall x,y,z\,(\Box A(x,y,z)\vee\Box\neg A(x,y,z))\wedge\\ \quad\forall x,y,z\,(\Box M(x,y,z)\vee\Box\neg M(x,y,z)).\end{array}$$

The following lemma yields the algorithmic reducibility of the set of all true arithmetic formulas (which, by Tarski's theorem, is nonarithmetic) to the set Tr, and this proves the theorem.

16.5. Lemma. *For any arithmetic formula φ, φ is true if and only if every realizational instance of $\beta \wedge C \to \varphi$ is true.*

Proof. $\Longrightarrow$: Suppose φ is true, * is a realization for $\beta \wedge C \to \varphi$ and $\beta^* \wedge C^*$ is true. We want to show that φ^* is also true. It is not hard to see that, since T is consistent and recursively enumerable (this condition is essential!), the truth of C^* means that the relations defined on ω in the standard model by the formulas E^*, A^* and M^* are recursive. Let us define a model D with domain ω such that, for all $k, m, n \in \omega$,

$D \vDash E(k,m)$ iff $E^*(k,m)$ is true,
$D \vDash A(k,m,n)$ iff $A^*(k,m,n)$ is true,
$D \vDash M(k,m,n)$ iff $M^*(k,m,n)$ is true.

Observe that for every arithmetic formula γ (for which the realization * is legal), we have $D \vDash \gamma$ iff γ^* is true. In particular $D \vDash \beta$, and thus D satisfies the conditions of $(2')$, i.e., D is elementarily equivalent to the standard model, whence (as φ is true) $D \vDash \varphi$, whence φ^* is true.
$\Longleftarrow$: Suppose φ is false. Let * be the trivial realization, i.e., such that $E^*(x,y) = E(x,y)$, $A^*(x,y,z) = A(x,y,z)$, $M^*(x,y,z) = M(x,y,z)$. Then $\beta^* = \beta$, $\varphi^* = \varphi$ and therefore it suffices to show that $\beta \wedge C^* \to \varphi$ is false, i.e., that $\beta \wedge C^*$ is true. But β is true by (1), and from the decidability in T of the relations $x = y$, $x + y = z$ and $x \times y = z$, it follows that C^* is also true. $\dashv$

Formalizing in arithmetic the idea employed in the above proof, Vardanyan [1986] also proved that if T is recursively enumerable, then the set of predicate modal formulas whose realizational instances are provable in T (or in **PA**) is not recursively enumerable and is in fact Π_2-complete.

There is one perhaps even more unpleasant result which should also be mentioned here. For recursively enumerable T, the answer to the question whether a predicate modal formula expresses a valid provability principle, turns out to be dependent on the choice of the formula Pr_T, that is, on the concrete way of formalization of the predicate "x is the code of an axiom of T", even if a set of axioms is fixed (Artëmov [1986]). Note that the proofs of Solovay's theorems for propositional provability logic are insensitive in this respect and actually the only requirement is that the three Löb-conditions must be satisfied.

16.6. . . . but still not completely desperate

Against this gloomy background one still can succeed in obtaining positive results in two directions. Firstly, although the predicate logic of provability in full generality is not (recursively) axiomatizable, some natural fragments of it can be so and may be stable with respect to the choice of the formula Pr_T.

And secondly, all the above-mentioned negative facts exclusively concern recursively enumerable theories, and the proofs hopelessly fail as soon as this condition is removed. There are however many examples of interesting and natural theories

which are not recursively enumerable (e.g., the theories induced by ω-provability or the other strong concepts of provability mentioned in section 8), and it well might be that the situation with their predicate provability logics is as nice as in the propositional case.

The main positive result we are going to consider is the following: the "arithmetic part" of Solovay's theorems, according to which the existence of a Kripke countermodel (with a transitive and converse well-founded accessibility relation) implies arithmetic nonvalidity of the formula, can be extended to the predicate level. This gives us a method of establishing nonvalidity for a quite considerable class of predicate modal formulas.

16.7. Kripke-models for the predicate modal language

A *Kripke-frame* for the predicate modal language is a system

$$M = \langle W, R, \{D_w\}_{w \in W} \rangle,$$

where $\langle W, R \rangle$ is a Kripke-frame in the sense of section 2, $\{D_w\}_{w \in W}$ are nonempty sets ("domains of individuals") indexed by elements of W such that if $w\,R\,u$, then $D_w \subseteq D_u$, and a *Kripke-model* is a Kripke-frame together with a forcing relation $\Vdash$, which is now a relation between worlds $w \in W$ and closed formulas with parameters in D_w; for the Boolean connectives and $\Box$, $\Vdash$ behaves as described in section 2, and we have only the following additional condition for the universal quantifier:

- $w \Vdash \forall x A(x)$ iff $w \Vdash A(a)$ for all $a \in D_w$,

and a similar one for the existential quantifier. A formula is said to be *valid* in a Kripke-model $\langle W, R, \{D_w\}_{w \in W}, \Vdash \rangle$, if A is forced at every world $w \in W$. Such a model is said to be finite if W as well as all D_w are finite.

16.8. The predicate version of Solovay's theorems

For every predicate modal formula A, let $\mathrm{REFL}(A)$ denote the universal closure of $\bigwedge\{\Box B \to B \mid \Box B \in \mathrm{Sb}\}$, where Sb is the set of the subformulas of A.

16.9. Theorem. (Artëmov and Japaridze [1987,1990]). *For any closed predicate modal formula A,*
(a) *if A is not valid in some finite Kripke-model with a transitive and converse well-founded accessibility relation, then there exists a realization * for A such that $T \nvdash A^*$,*
(b) *if $\mathrm{REFL}(A) \to A$ is not valid in such a model, then there exists a realization * for A such that A^* is false.*

Proof. We prove here only clause (a), leaving (b) as an exercise for the reader. Some details in this proof are in fact redundant if we want to prove only (a), but they are of assistance in passing to a proof of (b).

Assume that $\langle W, R, \{D_w\}_{w \in W}, \Vdash \rangle$ is a model with the above-mentioned properties in which A is not valid. As before, without loss of generality we may suppose that $W = \{1, \ldots, l\}$, 1 is the root and $1 \nVdash A$. We suppose also that $D_w \subseteq \omega$ and $0 \in D_w$ for each $w \in W$. Let us define a model $\langle W', R', \{D'_w\}_{w \in W'}, \Vdash' \rangle$ by setting

- $W' = W \cup \{0\}$,
- $R' = R \cup \{(0, w) \mid w \in W\}$,
- $D'_0 = D_1$ and, for all $w \in W$, $D'_w = D_w$,
- for any atomic formula P, $0 \Vdash' P$ iff $1 \Vdash P$ and, if $w \in W$, $w \Vdash' P$ iff $w \Vdash P$.

We accept the definitions of the Solovay function h and the sentences Lim_w from section 3 without any changes; the only additional step is the following:

For each a from $D = \bigcup\{D_w \mid x \in W\}$ we define an arithmetic formula $\gamma_a(x)$ with only x free by setting

$$\gamma_a(x) = \bigvee\{\exists t \leqslant x(\, h(t) = h(x) = w \wedge \neg\, \exists z < t (h(z) = w) \wedge x = t + a) \mid a \in D_w\}.$$

Thus, using the jargon from section 14, $\gamma_a(x)$ says that we have reached some world w such that $a \in D_w$, at the moment x we are still at w, and exactly a moments have passed since we moved to this world (we assume that the first "move", to the world 0, happened at the initial moment 0). We define the predicates γ'_a by

- for each $0 \neq a \in D$, $\gamma'_a(x) = \gamma_a(x)$, and
- $\gamma'_0(x) = \gamma_0(x) \vee \neg \bigvee\{\gamma_a(x) \mid a \in D \backslash \{0\}\}$.

(It is easy to check that the left disjunct of $\gamma'_0(x)$ is redundant; it implies the right disjunct.) Since we employ the same Solovay function h as in section 3, lemma 3.2 continues to hold. In addition, we need the following lemma:

16.10. Lemma.

(i) $T \vdash \neg\, (\gamma'_a(x) \wedge \gamma'_b(x))$ *for all* $a \neq b$,

(ii) $T \vdash \mathrm{Lim}_w \to \bigwedge\{\exists x\, \gamma'_a(x) \mid a \in D_w\}$ *for all* $w \in W'$,

(iii) $T \vdash \mathrm{Lim}_w \to \forall x(\bigvee\{\gamma'_a(x) \mid a \in D_w\})$ *for all* $w \in W'$.

Proof. (i): The formulas $\gamma_a(x)$ and $\gamma_b(x)$ for $a \neq b$ are defined so that each disjunct of $\gamma_a(x)$ is inconsistent with each disjunct of $\gamma_b(x)$. And the right disjunct of $\gamma'_0(x)$, by definition, is inconsistent with each $\gamma_a(x)$, $a \neq 0$.

(ii): Suppose $a \in D_w$ and argue in $T + \mathrm{Lim}_w$. Since w is the limit of h, there is a moment t at which we arrive in w, and stay there for ever (more formally: we have $\neg\, \exists y < t\, (h(y) = w)$ and $\forall y \geqslant t\, (h(y) = w)$). Then, by definition, $\gamma_a(t + a)$ holds, whence $\gamma'_a(t + a)$ holds, whence $\exists x\, \gamma'_a(x)$. And so for each $a \in D_w$.

(iii): Argue in $T + \mathrm{Lim}_w$. Consider an arbitrary number x. We must show that $\gamma'_a(x)$ holds for some $a \in D_w$. The definition of h implies that, either $h(x) R' w$, or $h(x) = w$; in both cases we then have $D_{h(x)} \subseteq D_w$. Let t be the least number such that $h(t) = h(x)$, and let $a = x - t$. By definition, if $a \in D_{h(x)}$ (and thus $a \in D_w$),

then $\gamma_a(x)$ holds, whence $\gamma'_a(x)$ holds and we are done; and if $a \notin D_{h(x)}$, then (the right disjunct of) $\gamma'_0(x)$ holds and we are also done, because $0 \in D_w$. $\dashv$

We now define a realization *. For each n-place predicate letter P, let P^* be

$$\bigvee\{\mathrm{Lim}_w \wedge \gamma'_{a_1}(v_1) \wedge \ldots \wedge \gamma'_{a_n}(v_n) \mid a_1, \ldots, a_n \in D_w,\ w \Vdash' P(a_1, \ldots, a_n)\}.$$

16.11. Lemma. *Let B be a predicate modal formula with precisely $x_1, \ldots, x_n$ free. Then, for each $w \in W$ and for all $a_1, \ldots, a_n \in D_w$,*

(a) *if $w \Vdash' (B(a_1, \ldots, a_n)$, then $T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to B^*$;*
(b) *if $w \nVdash' (B(a_1, \ldots, a_n)$, then $T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \neg B^*$.*

Proof. We proceed by induction on the complexity of B. Suppose $B(x_1, \ldots, x_n)$ is atomic. If $w \Vdash' B(a_1, \ldots, a_n)$, then $\mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)$ is one of the disjuncts of B^* and the desired result is obvious. If $w \nVdash' B(a_1, \ldots, a_n)$, then that formula is not a disjunct of B^* and, according to lemma 3.2(b) and 16.10(i), it implies in T the negations of all the disjuncts of B^*.

Next suppose that $B(x_1, \ldots, x_n)$ is $\forall y\, C(y, x_1, \ldots, x_n)$.
If $w \Vdash \forall y\, C(y, a_1, \ldots, a_n)$, then $w \Vdash C(b, a_1, \ldots, a_n)$ for all $b \in D_w$. Then, by the induction hypothesis, for all $b \in D_w$,

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_b(y) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to (C(y, x_1, \ldots, x_n))^*.$$

Therefore,

$$T \vdash \mathrm{Lim}_w \wedge \Big(\bigvee\{\gamma'_b(y) \mid b \in D_w\}\Big) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to (C(y, x_1, \ldots, x_n))^*.$$

Note that there is no free occurrence of y in either Lim_w or $\gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)$. Universal quantification over y gives

$$T \vdash \mathrm{Lim}_w \wedge \forall y\Big(\bigvee\{\gamma'_b(y) \mid b \in D_w\}\Big) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \forall y (C(y, x_1, \ldots, x_n))^*.$$

By lemma 16.10(iii), we can eliminate the conjunct $\forall y\Big(\bigvee\{\gamma'_b(y) \mid b \in D_w\}\Big)$ in the antecedent of the above formula and conclude that

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \forall y\, (C(y, x_1, \ldots, x_n))^*.$$

If on the other hand $w \nVdash \forall y\, C(y, a_1, \ldots, a_n)$, then there is $b \in D_w$ such that $w \nVdash C(b, a_1, \ldots, a_n)$. By the induction hypothesis,

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_b(y) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \neg\, (C(y, x_1, \ldots, x_n))^*.$$

Again, neither Lim_w nor $\gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)$ contains y free, and existential quantification over y gives

$$T \vdash \mathrm{Lim}_w \wedge \exists y\, \gamma'_b(y) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \exists y\, \neg (C(y, x_1, \ldots, x_n))^*.$$

According to lemma 16.10(ii), $T \vdash \mathrm{Lim}_w \to \exists y\, \gamma'_b(y)$. Therefore,

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \neg (\forall y\, C(y, x_1, \ldots, x_n))^*.$$

Finally, suppose that B is $\Box C$. If $w \Vdash \Box C(a_1, \ldots, a_n)$, then for each u such that $wR'u$, we have $u \Vdash C(a_1, \ldots, a_n)$ and, by the induction hypothesis,

$$T \vdash \mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to (C(x_1, \ldots, x_n))^*.$$

Therefore,

$$T \vdash \Big(\bigvee \{\mathrm{Lim}_u \mid wR'u\} \Big) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to (C(x_1, \ldots, x_n))^*,$$

and, by the first two Löb conditions,

$$T \vdash \mathrm{Pr}_T[\Big(\bigvee \{\mathrm{Lim}_u \mid wR'u\} \Big) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)] \to (\Box C(x_1, \ldots, x_n))^*.$$

Observe that the formulas $\gamma_a(x)$ are primitive recursive and we have that $T \vdash \gamma_a(x) \to \mathrm{Pr}_T[\gamma_a(x)]$; together with lemma 3.2(d) this means that

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \\ \mathrm{Pr}_T[\Big(\bigvee \{\mathrm{Lim}_u \mid wR'u\} \Big) \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)].$$

Thus, we get $T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to (\Box C(x_1, \ldots, x_n))^*$.

If $w \nVdash \Box C(a_1, \ldots, a_n)$, then there is u such that $wR'u$ and $u \nVdash C(a_1, \ldots, a_n)$. By the induction hypothesis,

$$T \vdash \mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \neg\, (C(x_1, \ldots, x_n))^*.$$

Therefore,

$$T \vdash (C(x_1, \ldots, x_n))^* \to \neg\, (\mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)),$$

$$T \vdash \mathrm{Pr}_T[(C(x_1, \ldots, x_n))^*] \to \mathrm{Pr}_T[\neg\, (\mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n))],$$

$$T \vdash \neg\, \mathrm{Pr}_T[\neg\, (\mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n))] \to \neg\, (\Box C(x_1, \ldots, x_n))^*.$$

On the other hand, we have

$$T \vdash \neg\, \mathrm{Pr}_T[\neg\, \mathrm{Lim}_u] \wedge \mathrm{Pr}_T[\gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)] \to \\ \neg\, \mathrm{Pr}_T[\neg\, (\mathrm{Lim}_u \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n))]$$

(this is a realizational instance of the principle $\Diamond p \wedge \Box q \to \Diamond(p \wedge q)$ which is provable in **K**). According to lemma 3.2(c), and since $T \vdash \gamma'_a(x) \to \mathrm{Pr}_T[\gamma'_a(x)]$, we have

$$T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \\ \neg\, \mathrm{Pr}_T[\neg\, \mathrm{Lim}_u] \wedge \mathrm{Pr}_T[\gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n)].$$

Therefore, $T \vdash \mathrm{Lim}_w \wedge \gamma'_{a_1}(x_1) \wedge \ldots \wedge \gamma'_{a_n}(x_n) \to \neg\, (\Box C(x_1, \ldots, x_n))^*$. ⊣

To finish the proof of theorem 16.9: since A is closed and $1 \nVdash A$, we have by lemma 16.11, $T \vdash \mathrm{Lim}_1 \to \neg\, A^*$. By lemma 3.2(f), Lim_1 is consistent with T, and consequently $T \nvdash A^*$. ⊣

16.12. Further positive results

One of the applications of theorem 16.9 is the following. Consider the fragment of our predicate modal language which arises by restricting the set of variables to one single variable x. In this case, without loss of generality, we may assume that every predicate letter is one-place. Since the variable x is fixed, it is convenient to omit it in the expressions $\forall x$, $P(x)$, $Q(x)$, ... and simply write $\forall$, p, q, ... In fact, we then have a bimodal propositional language with the modal operators $\Box$ and $\forall$. The modal logic **Lq**, introduced by Esakia [1988], is axiomatized by the following schemata:

1. all propositional tautologies in the bimodal language,
2. the axioms of **L** for $\Box$,
3. the axioms of **S5** for $\forall$, i.e.,
   - $\forall(A \to B) \to (\forall A \to \forall B)$,
   - $\forall A \to A$,
   - $\exists A \to \forall \exists A$ ($\exists$ abbreviates $\neg \forall \neg$),
4. $\Box \forall A \to \forall \Box A$,

together with the rules modus ponens, $A/\Box A$ and $A/\forall A$. For this logic (the language of which is understood as a fragment of the predicate modal language) we have the following modal completeness theorem:

16.13. Theorem. (Japaridze [1988a,1990a]) *For any* **Lq***-formula* A, $\vdash_{\mathbf{Lq}} A$ *iff* A *is valid in all finite predicate Kripke-models with a transitive and converse well-founded accessibility relation.*

In view of the evident arithmetic soundness of **Lq**, this modal completeness theorem together with the above predicate version of Solovay's first theorem implies the arithmetic completeness of **Lq**:

16.14. Corollary. *For any* **Lq***-formula* A, $\vdash_{\mathbf{Lq}} A$ *iff every realizational instance of* A *is provable in* T.

Japaridze [1988a,1990a] also introduced the bimodal version **Sq** of **S** and proved that $\vdash_{\mathbf{Sq}} A$ iff every realizational instance of A is true. The axioms of **Sq** are all theorems of **Lq** plus $\Box A \to A$, and the rules of inference are Modus Ponens and $A/\forall A$.

Taking into account that we deal with a predicate language, the requirement of finiteness of the models in theorem 16.9 is a very undesirable restriction however. In Japaridze [1990a] a stronger variant of this theorem was given with the condition of finiteness replaced by a weaker one. What we need instead of finiteness, is roughly the following:

(1) The relations $w \in W$, $w\,R\,u$, $a \in D_w$ must be binumerable in T (see definition 12.1), and this fact must be provable in T.

(2) The relation $\Vdash$ must be numerable in T and T must prove that fact. To be more precise, $\Vdash$ need not be defined for all worlds and all formulas, but only for those which are needed to falsify the formula A in the root of the model (i.e., in some cases we may have neither $w \Vdash B$ nor $w \Vdash \neg B$); T should just prove that $\Vdash$ behaves "properly", e.g., $w \Vdash B \implies w \nVdash \neg B$, $w \Vdash B \vee C \implies (w \Vdash B$ or $w \Vdash C)$, $w \Vdash \neg (B \vee C) \implies (w \Vdash \neg B$ and $w \Vdash \neg C)$, ...

(3) T also must "prove" that the relation R is transitive and converse well-founded. Of course, well-foundedness is not expressible in the first order language, and T should somehow simulate a proof of this property of R. This is the case if, e.g., T proves the scheme of R-induction for the elements of W, i.e.,

$$T \vdash \forall w \in W \Big(\forall u(wR\,u \to \varphi(u)) \to \varphi(w)\Big) \to \forall w \in W\, \varphi(w).$$

We want to end this section by mentioning one last positive result. Let **QL** be the logic which arises by adding to **L** (written in the predicate modal language) the axioms and rules of the classical predicate calculus. Similarly, let **QS** be the closure of **S** with respect to classical predicate logic.

16.15. Theorem. (Japaridze [1990a,1991]). *Suppose T is strong enough to prove all true Π_1-sentences, and A is a closed predicate modal formula which satisfies one of the following conditions:*

(i) *no occurrence of a quantifier is in the scope of some occurrence of $\Box$ in A, or*

(ii) *no occurrence of $\Box$ is in the scope of some other occurrence of $\Box$ in A, or*

(iii) *A has the form $\Box^n \bot \to B$.*

Then we have:

(a) $\vdash_{\mathbf{QL}} A$ *iff all realizational instances of A are provable in T,*

(b) $\vdash_{\mathbf{QS}} A$ *iff all realizational instances of A are true.*

(Of course, clause (b) is trivial in case (iii).) The proof for the (ii) and (iii)-fragments in Japaridze [1990a] is based on the above-mentioned strong variant of the predicate version of Solovay's theorems. Both Vardanyan's and Artëmov's theorems on nonenumerability and nonarithmeticity hold for the (i) and (ii)-fragments as well, but this is not in contradiction with theorem 16.15. The point is that the use of Tennenbaum's theorem in the proofs of these negative results is possible only on assumption of the recursive enumerability of T, whereas no consistent theory which proves all the true Π_1-sentences can be recursively enumerable. Thus there are no immediate objections against the optimistic conjecture that **QL** and **QS** are complete for such strong theories without any restriction on the language.

17. Acknowledgements

In the first place we are very grateful to Lev Beklemishev for providing us with a draft for the sections 6, 7 and 8 in a near perfect state. He also gave extensive

comments on other sections. Sergei Artëmov supported us with section 10, and in answering some questions for us. Albert Visser was very helpful with comments and discussions, answering questions, and pointing out mistakes. Giovanni Sambin provided valuable comments. Claes Strannegård shared his expertise with us. Joost Joosten, Rosalie Iemhoff and Eva Hoogland found quite a number of inaccuracies in the manuscript. Anne Troelstra stood by us with advice. Sam Buss was a very helpful editor and careful reader.

The first author was supported by N.W.O., the Dutch Foundation for Scientific Research, while working on this chapter in 1992-1993.

References

S. N. Artëmov

[1980] Arithmetically complete modal theories, *Semiotika i Informatika, VINITI, Moscow*, 14, pp. 115–133. In Russian, English translation in: *Amer. Math. Soc. Transl.* (2), 135: 39–54, 1987.

[1985a] Nonarithmeticity of truth predicate logics of provability, *Doklady Akademii Nauk SSSR*, 284, pp. 270–271. In Russian, English translation in *Soviet Math. Dokl.* 32 (1985), pp. 403-405.

[1985b] On modal logics axiomatizing provability, *Izvestiya Akad. Nauk SSSR, ser. mat.*, 49, pp. 1123–1154. In Russian, English translation in: *Math. USSR Izvestiya* 27(3).

[1986] Numerically correct logics of provability, *Doklady Akademii Nauk SSSR*, 290, pp. 1289–1292. In Russian.

[1994] Logic of proofs, *Annals of Pure and Applied Logic*, 67, pp. 29–59.

[1995] *Operational Modal Logic*, Tech. Rep. MSI 95-29, Cornell University.

S. N. Artëmov and G. K. Japaridze (Dzhaparidze)

[1987] On effective predicate logics of provability, *Doklady Akademii Nauk SSSR*, 297, pp. 521–523. In Russian, English translation in *Soviet Math. Dokl.* 36 (1987), pp. 478-480.

[1990] Finite Kripke models and predicate logics of provability, *Journal of Symbolic Logic*, 55, pp. 1090–1098.

A. Avron

[1984] On modal systems having arithmetical interpretations, *Journal of Symbolic Logic*, 49, pp. 935–942.

L. D. Beklemishev

[1989a] On the classification of propositional provability logics, *Izvestiya Akademii Nauk SSSR, ser. mat. D*, 53, pp. 915–943. In Russian, English translation in *Math. USSR Izvestiya* 35 (1990) 247–275.

[1989b] A provability logic without Craig'sprotect interpolation property, *Matematicheskie Zametkie*, 45, pp. 12–22. In Russian, English translation in *Math. Notes* 45 (1989).

[1991] Provability logics for natural Turing progressions of arithmetical theories, *Studia Logica*, pp. 107–128.

[1992] Independent numerations of theories and recursive progressions, *Sibirskii Matematichskii Zhurnal*, 33, pp. 22–46. In Russian, English translation in *Siberian Math. Journal*, 33 (1992).

[1993a] On the complexity of arithmetic applications of modal formulae, *Archive for Mathematical Logic*, 32, pp. 229–238.

[1993b] Review of de Jongh and Montagna [1988,1989], Carbone and Montagna [1989,1990], *Journal of Symbolic Logic*, 58, pp. 715–717.

[1994] On bimodal logics of provability, *Annals of Pure and Applied Logic*, 68, pp. 115–160.

[1996a] Bimodal logics for extensions of arithmetical theories, *Journal of Symbolic Logic*, 61, pp. 91–124.

[1996b] Remarks on Magari-algebras of PA and $I\Delta_0 + EXP$, in: *Logic and Algebra*, A. Ursini and P. Aglianò, eds., Marcel Dekker, Inc., New York, pp. 317–326.

A. Berarducci

[1990] The interpretability logic of Peano arithmetic, *Journal of Symbolic Logic*, 55, pp. 1059–1089.

A. Berarducci and R. Verbrugge

[1993] On the provability logic of bounded arithmetic, *Annals of Pure and Applied Logic*, 61, pp. 75–93.

C. Bernardi

[1976] The uniqueness of the fixed-point in every diagonalizable algebra, *Studia Logica*, 35, pp. 335–343.

G. Boolos

[1979] *The Unprovability of Consistency*, Cambridge University Press.

[1981] Provability, truth and modal logic, *Journal of Philosophic Logic*, 9, pp. 1–7.

[1982] Extremely undecidable sentences, *Journal of Symbolic Logic*, 47, pp. 191–196.

[1993a] The analytical completeness of Dzhaparidze's polymodal logics, *Annals of Pure and Applied Logic*, 61, pp. 95–111.

[1993b] *The Logic of Provability*, Cambridge University Press.

G. Boolos and R. C. Jeffrey

[1989] *Computability and Logic, 3rd ed.*, Cambridge University Press.

G. Boolos and V. McGee

[1987] The degree of the set of sentences of predicate provability logic that are true under every interpretation, *Journal of Symbolic Logic*, 52, pp. 165–171.

G. Boolos and G. Sambin

[1991] Provability: the emergence of a mathematical modality, *Studia Logica*, 50, pp. 1–23.

A. Carbone and F. Montagna

[1989] Rosser orderings in bimodal logics, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 35, pp. 343–358.

[1990] Much shorter proofs: a bimodal investigation, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 36, pp. 47–66.

T. Carlson

[1986] Modal logics with several operators and provability interpretations, *Israel Journal of Mathematics*, 54, pp. 14–24.

B. F. Chellas

[1980] *Modal Logic: An Introduction*, Cambridge University Press.

P. Clote and J. Krajíček

[1993] eds., *Arithmetic, Proof Theory and Computational Complexity*, Oxford University Press.

D. van Dalen

[1994] *Logic and Structure*, Springer Verlag, Berlin, Amsterdam.

L. L. Esakia

[1988] Provability logic with quantifier modalities, in: *Intensional Logics and the Logical Structure of Theories: Material from the fourth Soviet-Finnish Symposium on Logic, Telavi, May 20-24, 1985*, Metsniereba, Tbilisi, pp. 4–9. In Russian.

S. Feferman
[1960] Arithmetization of metamathematics in a general setting, *Archive for Mathematical Logic*, 6, pp. 52–63.
[1962] Transfinite recursive progressions of axiomatic theories, *Journal of Symbolic Logic*, 27, pp. 259–316.

S. Feferman, G. Kreisel, and S. Orey
[1960] 1-consistency and faithful interpretations, *Fundamenta Mathematicae*, 49, pp. 35–92.

Z. Gleit and W. Goldfarb
[1990] Characters and fixed points in provability logic, *Notre Dame Journal of Formal Logic*, 31, pp. 26–551.

K. Gödel
[1933] Eine Interpretation des intuitionistischen Aussagenkalkuls, *Ergebnisse Math. Colloq.*, Bd. 4, pp. 39–40.

D. Guaspari
[1979] Partially conservative extensions of arithmetic, *Transactions of the American Mathematical Society*, 254, pp. 47–68.
[1983] Sentences implying their own provability, *Journal of Symbolic Logic*, 48, pp. 777–789.

D. Guaspari and R. M. Solovay
[1979] Rosser sentences, *Annals of Mathematical Logic*, 16, pp. 81–99.

P. Hájek
[1971] On interpretability in set theories I, *Comm. Math. Univ. Carolinae*, 12, pp. 73–79.
[1972] On interpretability in set theories II, *Comm. Math. Univ. Carolinae*, 13, pp. 445–455.

P. Hájek and F. Montagna
[1990] The logic of Π_1-conservativity, *Archiv für Mathematische Logik und Grundlagenforschung*, 30, pp. 113–123.
[1992] The logic of Π_1-conservativity continued, *Archiv für Mathematische Logik und Grundlagenforschung*, 32, pp. 57–63.

P. Hájek, F. Montagna, and P. Pudlak
[1993] Abbreviating proofs using metamathematical rules, in: *Clote and Krajíček [1993]*, pp. 387–428.

D. Harel
[1984] Dynamic logic, in: *Handbook of Philosophic Logic, Volume II, Extensions of Classical Logic*, D. Gabbay and F. Guenthener, eds., Kluwer Academic Publishers, Dordrecht, Boston, pp. 497–604.

D. Hilbert and P. Bernays
[1939] *Grundlagen der Mathematik II*, Springer, Berlin.

G. E. Hughes and M. J. Cresswell
[1984] *A Companion to MODAL LOGIC*, Methuen, London, New York.

K. N. Ignatiev
[1990] *The logic of Σ_1-interpolability over Peano arithmetic.* Manuscript. In Russian.
[1993a] On strong provability predicates and the associated modal logics, *Journal of Symbolic Logic*, 58, pp. 249–290.
[1993b] The provability logic of Σ_1-interpolability, *Annals of Pure and Applied Logic*, 64, pp. 1–25.

G. K. Japaridze (Dzhaparidze)
[1986] *The Modal Logical Means of Investigation of Provability*, PhD thesis, Moscow State University. In Russian.

[1988a] The arithmetical completeness of the logic of provability with quantifier modalities, *Bull. Acad. Sci. Georgian SSR*, 132, pp. 265–268. In Russian.

[1988b] The polymodal logic of provability, in: *Intensional Logics and the Logical Structure of Theories: Material from the fourth Soviet-Finnish Symposium on Logic, Telavi, May 20-24, 1985*, Metsniereba, Tbilisi, pp. 16–48. In Russian.

[1990a] Decidable and enumerable predicate logics of provability, *Studia Logica*, 49, pp. 7–21.

[1990b] Provability logic with modalities for arithmetical complexities, *Bull. Acad. Sci. Georgian SSR*, 138, pp. 481–484.

[1991] Predicate provability logic with non-modalized quantifiers, *Studia Logica*, 50, pp. 149–160.

[1992] The logic of linear tolerance, *Studia Logica*, 51, pp. 249–277.

[1993] A generalized notion of weak interpretability and the corresponding logic, *Annals of Pure and Applied Logic*, 61, pp. 113–160.

[1994a] The logic of arithmetical hierarchy, *Annals of Pure and Applied Logic*, 66, pp. 89–112.

[1994b] A simple proof of arithmetical completeness for Π_1-conservativity logic, *Notre Dame Journal of Formal Logic*, 35, pp. 346–354.

D. H. J. de Jongh

[1987] A simplification of a completeness proof of Guaspari and Solovay, *Studia Logica*, 46, pp. 187–192.

D. H. J. de Jongh, M. Jumelet, and F. Montagna

[1991] On the proof of Solovay's theorem, *Studia Logica*, 50, pp. 51–70.

D. H. J. de Jongh and F. Montagna

[1988] Provable fixed points, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 34, pp. 229–250.

[1989] Much shorter proofs, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 35, pp. 247–260.

[1991] Rosser-orderings and free variables, *Studia Logica*, 50, pp. 71–80.

D. H. J. de Jongh and D. Pianigiani

[1998] Solution of a problem of David Guaspari, *Studia Logica*, 57. To appear.

D. H. J. de Jongh and F. Veltman

[1990] Provability logics for relative interpretability, in: *Petkov [1990]*, pp. 31–42.

D. H. J. de Jongh and A. Visser

[1991] Explicit fixed points in interpretability logic, *Studia Logica*, 50, pp. 39–50.

G. Kreisel and A. Lévy

[1968] Reflection principles and their use for establishing the complexity of axiomatic systems, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 14, pp. 97–142.

P. Lindström

[1984] On faithful interpretability, in: *Computation and Proof Theory*, M. M. Richter, E. Börger, W. Oberschelp, B. Schinzel, and W. Thomas, eds., Lecture Notes in Mathematics #1104, Springer Verlag, Berlin, Berlin, pp. 279–288.

[1994] *The Modal Logic of Parikh Provability*, Tech. Rep. Filosofiska Meddelanden, Gröna serien, No. 5, University of Göteborg.

J. C. C. McKinsey and A. Tarski

[1948] Some theorems about the calculi of Lewis and Heyting, *Journal of Symbolic Logic*, 13, pp. 1–15.

F. Montagna

[1979] On the diagonalizable algebra of Peano arithmetic, *Bulletino della Unione Matematica Italiana*, 5, 16B, pp. 795–812.

[1984] The predicate modal logic of provability, *Notre Dame Journal of Formal Logic*, 25, pp. 179–189.

[1987] Provability in finite subtheories of PA, *Journal of Symbolic Logic*, 52, pp. 494–511.

[1992] Polynomially and superexponentially shorter proofs in fragments of arithmetic, *Journal of Symbolic Logic*, 57, pp. 844–863.

S. Orey

[1961] Relative interpretations, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 7, pp. 146–153.

R. Parikh

[1971] Existence and feasibility, *Journal of Symbolic Logic*, 36, pp. 494–508.

P. P. Petkov

[1990] ed., *Mathematical Logic, Proceedings of the Heyting 1988 Summer School*, New York, Plenum Press.

M. de Rijke

[1992] Unary interpretability logic, *Notre Dame Journal of Formal Logic*, 33, pp. 249–272.

G. Sambin

[1976] An effective fixed-point theorem in intuitionistic diagonalizable algebras, *Studia Logica*, 35, pp. 345–361.

G. Sambin and S. Valentini

[1982] The modal logic of provability. The sequential approach., *Journal of Philosophical Logic*, 11, pp. 311–342.

[1983] The modal logic of provability: cut elimination., *Journal of Philosophical Logic*, 12, pp. 471–476.

D. S. Scott

[1962] Algebras of sets binumerable in complete extensions of arithmetic, in: *Recursive Function Theory*, American Mathematical Society, Providence, R.I., pp. 117–121.

V. Y. Shavrukov

[1988] *The Logic of Relative Interpretability over Peano Arithmetic*, Tech. Rep. Report No.5, Stekhlov Mathematical Institute, Moscow. (in Russian).

[1991] On Rosser's provability predicate, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 37, pp. 317–330.

[1993a] A note on the diagonalizable algebras of PA and ZF, *Annals of Pure and Applied Logic*, 61, pp. 161–173.

[1993b] Subalgebras of diagonalizable algebras of theories containing arithmetic, *Dissertationes mathematicae (Rozprawy matematycne)*, 323. Instytut Matematyczny, Polska Akademia Nauk, Warsaw.

[1994] A smart child of Peano's, *Notre Dame Journal of Formal Logic*, 35, pp. 161–185.

[1997] Undecidability in diagonalizable algebras, *Journal of Symbolic Logic*, 62, pp. 79–116.

C. Smoryński

[1977] The incompleteness theorems, in: *Handbook of Mathematical Logic*, J. Barwise, ed., vol. 4, North–Holland, Amsterdam, Amsterdam, pp. 821–865.

[1978] Beth's theorem and self–referential statements, in: *Computation and Proof Theory*, A. Macintyre, L. Pacholski, and J. B. Paris, eds., North–Holland, Amsterdam, Amsterdam, pp. 17–36.

[1985] *Self-reference and modal logic*, Springer-Verlag, Berlin.

C. Strannegård

[1997] *Arithmetical Realizations of Modal Formulas*, PhD thesis, University of Göteborg, Acta Philosophica Gothoburgensia 5.

A. Tarski, A. Mostowski, and R. M. Robinson
[1953] *Undecidable Theories*, North–Holland, Amsterdam, Amsterdam.

A. S. Troelstra and H. Schwichtenberg
[1996] *Basic Proof Theory*, Cambridge University Press.

A. Turing
[1939] System of logics based on ordinals, *Proceedings of the London Mathematical Society, Ser. 2*, 45, pp. 161–228.

V. A. Vardanyan
[1986] Arithmetic complexity of predicate logics of provability and their fragments, *Doklady Akademii Nauk SSSR*, 288, pp. 11–14. In Russian, English translation in *Soviet Math. Dokl.* 33 (1986), pp. 569-572.

R. Verbrugge
[1993a] *Efficient Metamathematics*, PhD thesis, Universiteit van Amsterdam, ILLC-disseration series 1993-3.
[1993b] Feasible interpretability, in: *Clote and Krajíček [1993]*, pp. 197–221.

A. Visser
[1980] Numerations, λ-calculus and arithmetic, in: *To H.B. Curry: Essays on Combinatory logic, lambda calculus and formalism*, J. P. Seldin and J. R. Hindley, eds., Academic Press, Inc., London, pp. 259–284.
[1981] *Aspects of Diagonalization and Provability*, PhD thesis, University of Utrecht, Utrecht, The Netherlands.
[1984] The provability logics of recursively enumerable theories extending Peano arithmetic at arbitrary theories extending Peano arithmetic, *Journal of Philosophical Logic*, 13, pp. 97–113.
[1985] *Evaluation, Provably Deductive Equivalence in Heyting's arithmetic of Substitution Instances of Propositional Formulas*, Tech. Rep. LGPS 4, Department of Philosophy, Utrecht University.
[1989] Peano's smart children. a provability logical study of systems with built-in consistency, *Notre Dame Journal of Formal Logic*, 30, pp. 161–196.
[1990] Interpretability logic, in: *Petkov [1990]*, pp. 175–209.
[1991] The formalization of interpretability, *Studia Logica*, 50, pp. 81–106.
[1994] *Propositional Combinations of Σ–Sentences in Heyting's Arithmetic*, Tech. Rep. LGPS 117, Department of Philosophy, Utrecht University. To appear in the Annals of Pure and Applied Logic.
[1995] A course in bimodal provability logic, *Annals of Pure and Applied Logic*, 73, pp. 109–142.
[1997] An overview of interpretability logic, in: *Advances in Modal Logic '96*, M. Kracht, M. de Rijke, and H. Wansing, eds., CSLI Publications, Stanford.

A. Visser, J. van Benthem, D. H. J. de Jongh, and G. R. Renardel de Lavalette
[1995] *NILL*, a study in intuitionistic propositional logic, in: *Modal Logic and Process Algebra, a Bisimulation Perspective*, A. Ponse, M. de Rijke, and Y. Venema, eds., CSLI Lecture Notes #53, CSLI Publications, Stanford, pp. 289–326.

F. Voorbraak
[1988] A simplification of the completeness proofs for Guaspari and Solovay's R, *Notre Dame Journal of Formal Logic*, 31, pp. 44–63.

D. Zambella
[1992] On the proofs of arithmetical completeness of interpretability logic, *Notre Dame Journal of Formal Logic*, 35, pp. 542–551.
[1994] Shavrukov's theorem on the subalgebras of diagonalizable algebras for theories containing $I\Delta_0 + \mathrm{EXP}$, *Notre Dame Journal of Formal Logic*, 35, pp. 147–157.

Name Index

Subject Index

Discard this page.