Notre Dame Journal of Formal Logic
Volume 34, Number 3, Summer 1993

Shavrukov's Theorem on the Subalgebras of Diagonalizable Algebras for Theories Containing $I\Delta_0 + \exp$

DOMENICO ZAMBELLA

Abstract Recently Shavrukov pioneered the study of subalgebras of diagonalizable algebras of theories of arithmetic. We show that his results extend to weaker theories (namely to theories containing $I\Delta_0 + \exp$).

1 Introduction A diagonalizable algebra (cf. Magari [4],[5], Bernardi [2], Bellisima [1], and Montagna [6]) is a Boolean algebra $(\mathcal{D},\to,\perp)$ with an additional operator $\Box$ which satisfies the axioms:

$$\forall x,y\, \Box(x \to y) \to (\Box x \to \Box y) = \top,$$
$$\forall x\, \Box(\Box x \to x) \to \Box x = \top,$$
$$\Box\top = \top$$

Let T be a sufficiently strong axiomatized theory in the language of arithmetic. The predicate of provability of T generates in a natural way an operator on the Lindenbaum algebra of T. The resulting diagonalizable algebra $\mathcal{D}_T$ is called the *diagonalizable algebra of* T. The subalgebras of $\mathcal{D}_T$ have been studied in Shavrukov [7], in particular the general problem of when a diagonalizable algebra $\mathcal{D}$ is embeddable in $\mathcal{D}_T$ was considered there. We intend to present a modification of Shavrukov's construction that allows us to prove the same results for a wider class of theories, namely all those containing $I\Delta_0 + \exp$.

We will translate this question about subalgebras into problems of provability logic. For this we need some notation. Let $\mathcal{L}$ be the set of modal formulas generated by the language $(\to,\Box,\perp,\{p_i\}_{i\in\omega})$. We write $B \models A$ if A can be derived using modus ponens and necessitation from the formula B and Löb's axioms (hence $\models A$ means that A is a theorem of Löb's logic and $B \models A$ means $\models \boxdot B \to A$, where $\boxdot B$ is $B \wedge \Box B$). We write $B \Vdash A$ iff $\models B \to A$. When $\mathcal{A}$ is a set of modal formulas in the

Received August 24, 1992; revised May, 1993

language $\mathcal{L}$ we write $\mathcal{A} \models A$, and $\mathcal{A} \Vdash A$ if for some conjunction B of formulas in $\mathcal{A}$, $B \models A$, resp. $B \Vdash A$. Given a set $\mathcal{A}$, consider the equivalence relation on $\mathcal{L}$: $A \approx_{\mathcal{A}} B$ iff $\mathcal{A} \models A \leftrightarrow B$, and let $\mathcal{L}/\mathcal{A}$ be the sets of $\approx_{\mathcal{A}}$-equivalence classes. The operator which maps the equivalence class of A to that of $\Box A$ is a well defined operator on $\mathcal{L}/\mathcal{A}$ which turns it into a diagonalizable algebra. For every (denumerable) diagonalizable algebra $\mathcal{D}$ there is a set $\mathcal{A}$ such that $\mathcal{D}$ is isomorphic to $\mathcal{L}/\mathcal{A}$.

Let T be an axiomatized theory in the language of arithmetic and let *Thm*(.) be the provability predicate of T. A *T-interpretation* is a map ι which maps formulas of $\mathcal{L}$ to sentences of the language of arithmetic such that T proves:

(1) $\iota(\Box A) \leftrightarrow \mathit{Thm}[\iota(A)]$;
(2) $\neg\iota(\bot)$;
(3) $\iota(A \to B) \leftrightarrow (\iota(A) \to \iota(B))$.

(In the following we shall simply say an *interpretation* since the theory T will be fixed.) If for every formula A in $\mathcal{L}$, $\mathcal{A} \models A$ iff $T \vdash \iota(A)$ we say that ι *interprets* $\mathcal{A}$ in T. We say that $\mathcal{A}$ is *interpretable* in T if there exists an interpretation which interprets $\mathcal{A}$ in T.

Given an interpretation of $\mathcal{A}$ in T one can construct in a natural way an embedding of $\mathcal{L}/\mathcal{A}$ in $\mathcal{D}_T$ and vice versa: from an embedding one can easily construct an interpretation. So for any given theory T, the problem of classifying the subalgebras of $\mathcal{D}_T$ reduces to classifying the sets of modal formulas $\mathcal{A}$ which are interpretable in T.

We write as usual $\Box^0\bot$ for $\bot$ and $\Box^{n+1}\bot$ for $\Box\Box^n\bot$; the minimal n such that $\mathcal{A} \models \Box^n\bot$ is called the *height* of $\mathcal{A}$. If such an n does not exist, we say that $\mathcal{A}$ has *infinite height*. We say that $\mathcal{A}$ has the *strong disjunction property* (s.d.p.) or, equivalently, that $\mathcal{A}$ is *strongly disjunctive* (s.d.) iff $\mathcal{A}$ is consistent and for all formulas A and B if $\mathcal{A} \models \Box A \vee \Box B$ then either $\mathcal{A} \models A$ or $\mathcal{A} \models B$. The same classification is, mutatis mutandis, applied to diagonalizable algebras. In the following T will be a fixed axiomatized theory (i.e., the theory is given along with a Kalmar elementary axiomatization of it). The language of T contains the language of the arithmetic and—only for the sake of convenience—a symbol for exponentiation. *Thm*(.) is the provability predicate of T. We write $\mathit{Thm}^0(\bot)$ for the sentence $0 \neq 0$ and $\mathit{Thm}^{n+1}(\bot)$ for $\mathit{Thm}(\mathit{Thm}^n(\bot))$ (in the following we shall always omit the Gödel number symbols $\ulcorner\urcorner$). The minimal n such that $T \vdash \mathit{Thm}^n(\bot)$ is called the *height* of T. If such an n does not exist we say that T has *infinite height*. The height of T is in fact the height of its diagonalizable algebra $\mathcal{D}_T$. If all Σ_1 sentences provable in T are true in the standard model, then T is Σ_1*-sound*, otherwise T is Σ_1*-ill*. Shavrukov proved that every r.e. set of modal formulas is interpretable in the diagonalizable algebra of every (sufficiently strong) Σ_1-ill theory provided it has the same height as the theory. Moreover an r.e. set of modal formulas is interpretable in the diagonalizable algebra of every (sufficiently strong) Σ_1-sound theory if and only if it is s.d. Recall that the Gödel numbering of arithmetical sentences gives a natural recursive enumeration of a set $\mathcal{A}$ such that $\mathcal{L}/\mathcal{A}$ is isomorphic to $\mathcal{D}_T$. So an interesting consequence is that diagonalizable algebras of Σ_1-sound theories are mutually embeddable. The same holds for Σ_1-ill theories of any fixed height.

The results mentioned above have been proved in [7] for theories which contain Σ_1 induction. In fact, the construction makes use of a Solovay function which ranges over a Kripke model. In the case of infinite height theories the models used have nonstandard height, so Σ_1 induction is needed to guarantee the existence of the limit. In Section 3 we show by Theorems 3.1 and 3.2 that the use of Σ_1 induction is inessential and the result is valid for all theories containing $I\Delta_0 + \exp$. (Actually Theorems 3.1 and 3.2 consider only theories of infinite height; in fact in the case of finite height the proof in [7] goes through for $I\Delta_0 + \exp$ with minor modifications.)

For Σ_1-ill theories a stronger result holds. In [7] a characterization was given of all (non necessarily r.e.) subalgebras of the diagonalizable algebra of a Σ_1-ill theory. Also this theorem holds for weaker theories than those considered in [7]. We shall not give a proof of this fact since it is easily derivable from Shavrukov's as follows. To embed $\mathcal{D}$ in the diagonalizable algebra of some "weak" theory T, first apply the result of [7] to embed $\mathcal{D}$ in the diagonalizable algebra of some sufficiently "strong" theory T^*. Finally, embed $\mathcal{D}_{T^*}$ in $\mathcal{D}_T$. Composing the two embeddings one obtains the desired subalgebra.

2 A lemma In this section we prove a lemma which will be used to characterize the r.e. sets of modal formulas interpretable in a theory $T \supseteq I\Delta_0 + \exp$. We assume the reader is familiar with the techniques introduced in Solovay [8].

A finite tree-like Kripke model k (in the sequel simply a *model*) is a triple $(W, R, \Vdash)$ where (W, R) is a finite tree with nodes $w \in W$ strictly ordered by the relation R, and $\Vdash$ is a finite subset of $W \times \omega$. We call W the *universe of* k and (W, R) the *frame of* k. We write $w \Vdash p_i$ if $(w, i) \in \Vdash$. The relation $w \Vdash A$ (w *forces* A) is then extended to all the formulas of $\mathcal{L}$ in the usual way. We say that $k' = (W', R', \Vdash')$ is a generated submodel (in the sequel simply a *submodel*) of $k = (W, R, \Vdash)$ if the universe of k' is $W' = \{w\} \cup \{u \mid wRu\}$ for some node w of k, and R' and $\Vdash'$ are the restrictions of R and $\Vdash$. We write $k \Vdash A$ (k *forces* A) iff the formula A is forced at the root of the model k, and we write $k \models A$ (k *is a model of* A) if every node of k forces A. Then we have that k is a model of A iff k forces $\boxdot A$. If $\mathcal{A}$ is a finite set of formulas we write $k \Vdash \mathcal{A}$ (resp. $k \models \mathcal{A}$) if for every $A \in \mathcal{A}$, $k \Vdash A$ (resp. $k \models A$). Then it is easy to check that, if $\mathcal{A}$ is finite, then $\mathcal{A} \models A$ iff every model of $\mathcal{A}$ is a model of A, and $\mathcal{A} \Vdash A$ iff every model which forces $\mathcal{A}$ forces A (if $\mathcal{A}$ is infinite this may not be the case).

In a first-order formula an occurrence of a quantifier is said to be bounded if it is of the form $\forall x < t$ or $\exists x < t$, where t is a term of the language of T. The Δ_0-formulas of T are the formulas provably equivalent to formulas with only bounded quantifiers (having assumed exponentiation as a primitive function of the language we should properly write $\Delta_0(\exp)$, but in the present paper there will be no risk of confusion). The Σ_1-formulas are those equivalent to a Δ_0-formula preceded by an existential quantifier. The theory whose axioms are those of Robinson arithmetic plus the characteristic axioms for exponentiation and the induction schema for Δ_0-formulas is called $I\Delta_0 + \exp$; the theory which contains also the schema of Σ_1 induction is called $I\Sigma_1$. We refer the reader to Hájeck and Pudlák [3] for more details on these theories.

We fix a natural coding of modal formulas and of models in arithmetic; we shall use the same symbol both for a formula (resp. model) and its code. We require that the coding assigns to proper submodels of k a smaller code than to k itself. Having exponentiation as a primitive function, we may require without loss of generality that $k \Vdash A$ and $k \models A$ translate into Δ_0-formulas. We also use in the following that the completeness theorem of Löb's logic with respect to (finite) models is formalizable in $I\Delta_0 + \exp$. Given an r.e. set $\mathcal{A}$ of modal formulas we may find, formalizing in the language of arithmetic the algorithm enumerating $\mathcal{A}$, a Δ_0-formula "$A \in \mathcal{A}_{,x}$" (here A and x are the free variables of the formula) such that for every $A \in \mathcal{L}$, $A \in \mathcal{A}$ iff $\exists n \in \omega\, T \vdash A \in \mathcal{A}_{,n}$. We also require that (provably in T) if $A \in \mathcal{A}_{,x}$ then $A < x$, i.e., the code of A is less than that of x. We call such a formula a *description* of $\mathcal{A}$ (in T). We may formalize in T also the notion of Löb's derivability so that we can use the expression $\mathcal{A}_{,n} \models A$ both when arguing in the real world and in the theory. Formalizing the proof of the completeness theorem for Löb's logic in $I\Delta_0 + \exp$ one can find a Δ_0-formula describing the relation $\mathcal{A}_{,n} \models A$. We shall also use the expression "$\mathcal{A} \models A$" when reasoning in T; this stands for $\exists x\, (\mathcal{A}_{,x} \models A)$.

Once we fix a description of $\mathcal{A}$, it makes perfect sense to say "T proves that $\mathcal{A}$ is s.d." This simply means:

$$T \vdash \neg(\mathcal{A} \models \bot) \wedge \forall A, B(\mathcal{A} \models \Box A \vee \Box B) \to (\mathcal{A} \models A \vee \mathcal{A} \models B).$$

Obviously, an r.e. set of formulas $\mathcal{A}$ may have different descriptions, and for one description the theory T may prove that $\mathcal{A}$ is s.d. whereas for another description it may not. Note also that possibly the "opinion" of T about $\mathcal{A}$ may be incorrect. In fact, when T is Σ_1-ill there are descriptions of $\mathcal{A}$ which do not satisfy $A \in \mathcal{A}$ iff $T \vdash \exists x (A \in \mathcal{A}_{,x})$. So it may happen T proves $\mathcal{A}$ is s.d. when this fails to reflect reality. We use essentially this fact in the next section; for the moment we keep the description fixed and assume T proves that $\mathcal{A}$ is s.d.

Lemma 2.1 *Let T be an axiomatized theory of infinite height containing $I\Delta_0 + \exp$ and $\mathcal{A}$ an r.e. set of modal formulas. If there is a description of $\mathcal{A}$ in T such that T proves that $\mathcal{A}$ is s.d. then $\mathcal{A}$ is interpretable in T.*

Proof: Let T be an axiomatized theory and "$A \in \mathcal{A}_{,n}$" be a description of an r.e. set of modal formulas as in the hypothesis of the lemma. We shall define a Solovay function $h(n)$ whose value is either 0 or the code of a model of $\mathcal{A}_{,m}$ for some $m \leq n$. We agree that $0 \Vdash A$ is some fixed provably false sentence (e.g., $0 \neq 0$), so the expression $h(n) \Vdash A$ will always have a meaning. The Solovay function is defined simultaneously with the sentences λ_0 and λ_A, by an arithmetical fixed point. The definition is the following.

Let λ_0 be the sentence $\forall n\, h(n) = 0$. We order the modal formulas by increasing code and let A_i be the i-th formula in this order (this enumeration of formulas is redundant, since here formulas are actually codes, but we introduce it for better readability). For every i and every string $\sigma \in 2^i$ define a formula:

$$A_\sigma := \bigwedge\{A_n \mid n < i \text{ and } \sigma(n) = 1\} \wedge \bigwedge\{\neg A_n \mid n < i \text{ and } \sigma(n) = 0\}.$$

The formula λ_A (with free variable A) is:

$$\begin{aligned}\lambda_A := \; & \exists \sigma \in 2^{i+1}[\sigma(i) \\ & = 1 \wedge \exists^\infty n\, h(n) \Vdash A_\sigma \;\wedge \forall \tau \in 2^{i+1}(\tau < \sigma \to \forall^\infty n\, h(n) \nVdash A_\tau)],\end{aligned}$$

where i is such that $A = A_i$ and $\tau < \sigma$ has to be read as τ precedes σ in the lexicographic order. $\exists^\infty n$ is an abbreviation of $\forall m \exists n > m$ and $\forall^\infty n$ of $\neg\exists^\infty n \neg$.

Let $h(0) = 0$. For $n+1$ if n codes a proof of $\lambda_0 \vee \lambda_A$ for some formula A, then:

(a) if $h(n) = 0$ and $\mathcal{A}_{,n} \not\models A$, then choose the minimal model k of $\mathcal{A}_{,n}$ which forces $\neg A$ and define $h(n+1) = k$.
(b) if $h(n) = h \neq 0$ and the root of some submodel of h forces $\neg A$ then let k be the minimal such submodel and define $h(n+1) = k$.
(c) in all other cases let $h(n+1) = h(n)$.

Note that (provably in T) the graph of h is Δ_0. A straightforward formalization of the completeness theorem for Löb's modal logic shows that $h(n)$ is (roughly) bounded by 2^{2^n} (h increases only if at stage n case (a) obtains; at that stage the code of $\neg A$ and of all the formulas in $\mathcal{A}_{,n}$ is bounded by n). So Δ_0 induction shows that h is a total function.

If the theory T is strong enough one is able to use for λ_A simply the sentence $\exists m \forall n > m\, h(n) \Vdash A$. Then $\lambda_0 \vee \lambda_A$ simply means that the limit of h is either 0 or a model which forces the formula A; in particular, if h moved to $h(n+1)$ because n codes a proof of $\lambda_0 \vee \lambda_A$, there will be a proof that $h(n+1)$ is not the limit of the function (in fact $h(n+1)$ is chosen so that $h(n+1) \Vdash \neg A$). But in $I\Delta_0 + \exp$ it we do not know how to prove that the limit of the Solovay function exists (one needs Σ_1 induction). It cannot be excluded that for some formula A both $h(n) \Vdash A$ and $h(n) \Vdash \neg A$ occurs for infinitely many n; thus one would not have as desired, $\lambda_{\neg A} \leftrightarrow \neg\lambda_A$. To help the reader's intuition we present the following semi-formal description of λ_A which should clarify the definition above. To each formula A we attach an infinite set $C(A)$ such that either $\forall n \in C(A)\, h(n) \Vdash A$ or $\forall n \in C(A)\, h(n) \Vdash \neg A$. The set $C(A)$ is defined in the following way. Let $C(A_0) = \{n \mid h(n) \Vdash \neg A_0\}$ if this is infinite, $C(A_0) = \{n \mid h(n) \Vdash A_0\}$ otherwise. Let $C(A_{i+1}) = \{n \in C(A_i) \mid h(n) \Vdash \neg A_{i+1}\}$ if this is infinite, $C(A_{i+1}) = \{n \in C(A_i) \mid h(n) \Vdash A_{i+1}\}$ otherwise. Finally, let λ_A be the sentence $\forall n \in C(A)\, h(n) \Vdash A$.

Claim 2.2 *T proves $\forall n\, [h(n) \neq 0 \to Thm[\exists m\, h(m)$ is a proper submodel of $h(\dot{n})]\,]$.*

Proof: In fact, if $h(n) \neq 0$, then at some stage $s < n$ for some formula A, s codes a proof of $\lambda_0 \vee \lambda_A$ and $h(s+1) = h(n) \Vdash \neg A$. By provable Σ_1 completeness $Thm[\neg\lambda_0]$. This together with $Thm[\lambda_0 \vee \lambda_A]$ yields $Thm[\lambda_A]$ and in particular $Thm[\exists^\infty n\, h(n) \Vdash A]$. From $h(n) \Vdash \neg A$ we get $Thm[h(\dot{n}) \Vdash \neg A]$ by provable Σ_1 completeness, and the claim follows.

Claim 2.3 *$\forall n \in \omega\, \exists m \in \omega$ such that T proves $h(n) \neq 0 \to Thm^m(\bot)$. (So, since T has infinite height, for every standard n, $h(n) = 0$.)*

Proof: This is an easy corollary of the previous claim.

To define $\iota(A)$ we need to assign "ad hoc" a model to 0. Following Shavrukov we will construct a formula $\mathcal{T}$ in such a way that for all standard formulas A and B the following properties are provable in T.

(1) $\neg\mathcal{T}(\bot)$
(2) $\mathcal{T}(A \to B) \leftrightarrow (\mathcal{T}(A) \to \mathcal{T}(B))$
(3) $\mathcal{A} \models A \to \mathcal{T}(A)$
(4) $\mathcal{T}(\Box A) \to \mathcal{A} \models A$.

(Roughly speaking the formula $\mathcal{T}(A)$ says that A belongs to some maximal consistent set $\mathcal{T}$ containing $\mathcal{A} \cup \{\neg\Box A \mid \mathcal{A} \not\models \Box A\}$. Such a set $\mathcal{T}$ exists (within T) since otherwise for some $A_0, \ldots, A_n$ such that $\mathcal{A} \not\models \Box A_0, \ldots, \mathcal{A} \not\models \Box A_n$ we would have $\mathcal{A} \models \Box A_0 \vee \ldots \vee \Box A_n$. This contradicts the provable s.d.p. of $\mathcal{A}$.) For the proof of the lemma only (1)–(4) are needed, so we prefer to postpone the definition of $\mathcal{T}$ and the proof of (1)–(4) until after the proof of the lemma.

We define τ_A to be the sentence $\lambda_0 \wedge \mathcal{T}(A)$, and finally define: $\iota(A) := \lambda_A \vee \tau_A$, i.e., $\lambda_A \vee [\lambda_0 \wedge \mathcal{T}(A)]$. We shall prove that ι is an interpretation (Claim 2.6) and that ι interprets $\mathcal{A}$ in T (Claim 2.7).

Claim 2.4 *For every $A \in \mathcal{L}$, T proves $\forall^\infty n\, h(n) \Vdash A \to \lambda_A$.*

Proof: Since A is standard we can replace in the defintion of λ_A the quantifications over strings by finite conjunctions and disjunctions. So the claim is trivial.

Claim 2.5 *For every $A \in \mathcal{L}$, T proves $\forall n[h(n) = 0 \wedge \mathcal{A}_{,n} \models A \to \iota(A)]$.*

Proof: Assume $h(n) = 0$ and $\mathcal{A}_{,n} \models A$. Reasoning in T we want to show $\lambda_A \vee \tau_A$. Since $h(n) = 0$ and $\mathcal{A}_{,n} \models A$, the function can leave 0 only to a model of A and eventually move to some submodel of it. So $\neg\lambda_0$ implies $\forall^\infty n\, h(n) \models A$. By the previous claim, this implies λ_A. On the other hand, by (3), we have $\mathcal{T}(A)$, so λ_0 implies τ_A.

Claim 2.6 *The function ι is an interpretation (i.e., properties (1)–(3) from Section 1 are provable in T).*

Proof: We have to prove that for every standard formula A properties (1)–(3) are provable in T, i.e., $\iota(\Box A) \leftrightarrow Thm[\iota(A)]$, $\neg\iota(\bot)$, and $\iota(A \to B) \leftrightarrow (\iota(A) \to \iota(B))$. The proof is more readable if we derive them both from $T + \lambda_0$ and from $T + \neg\lambda_0$. In fact, under the hypothesis λ_0, the sentence $\iota(A)$ is equivalent to $\mathcal{T}(A)$ (by our convention that $0 \not\Vdash A$), and under the hypothesis $\neg\lambda_0$, $\iota(A)$ is equivalent to λ_A.

$T + \lambda_0 \vdash \iota(\Box A) \to Thm[\iota(A)]$. Assume $\iota(\Box A)$ and λ_0 and reason in T. As we just remarked, under the assumption λ_0, $\iota(\Box A)$ reduces to $\mathcal{T}(\Box A)$. By (4) we obtain $\mathcal{A} \models A$, so for some n, $\mathcal{A}_{,n} \models A$. Since we assume λ_0, $h(n) = 0$. Both $\mathcal{A}_{,n} \models A$ and $h(n) = 0$ are Σ_1 formulas, so by provable Σ_1 completeness we have $Thm[\mathcal{A}_{,\dot{n}} \models A]$ and $Thm[h(\dot{n}) = 0]$. By Claim 2.5 we have $Thm[\iota(A)]$.

$T + \lambda_0 \vdash \iota(\Box A) \to \iota(\Box A)$. Assume $Thm[\lambda_A \vee \tau_A]$ and λ_0. It suffices to show, reasoning in T, that $\mathcal{T}(\Box A)$. Since $Thm[\lambda_A \vee \tau_A]$, a fortiori $Thm[\lambda_0 \vee \lambda_A]$. Let n be the code of a proof of $\lambda_0 \vee \lambda_A$. Since we assumed λ_0, $h(n) = 0$. Then $\mathcal{A}_{,n} \models A$, or else the function would leave 0 at stage $n + 1$, contradicting λ_0. Then $\mathcal{A} \models A$, and so by (3), $\mathcal{T}(\Box A)$.

$T + \lambda_0 \vdash \neg\iota(\bot)$. Immediate from (1).

$T + \lambda_0 \vdash \iota(A \to B) \leftrightarrow (\iota(A) \to \iota(B))$. Immediate from (2).

$T + \neg\lambda_0 \vdash \iota(\Box A) \to Thm[\iota(A)]$. Assume $\iota(\Box A)$ and $\neg\lambda_0$. It suffices to prove $Thm[\lambda_A]$ in T. By our assumption $\lambda_{\Box A}$ holds, in particular for some n, $h(n) \Vdash \Box A$. The latter is a Σ_1 formula so $Thm[h(\dot{n}) \Vdash \Box A]$. Since $h(n) \neq 0$, by Claim 2.2 we have Thm["$\exists m\, h(m)$ is a submodel of $h(\dot{n})$"], thus $Thm[\forall^\infty n\, h(n) \Vdash A]$. By Claim 2.4, $Thm[\lambda_A]$ follows.

$T + \neg\lambda_0 \vdash Thm[\iota(A)] \to \iota(\Box A)$. Assume $Thm[\lambda_A \vee \tau_A]$ and $\neg\lambda_0$. It suffices to derive $\lambda_{\Box A}$ reasoning in T. Since $Thm[\lambda_A \vee \tau_A]$, a fortiori $Thm[\lambda_0 \vee \lambda_A]$. Let n be the code of a proof of $\lambda_0 \vee \lambda_A$ which is large enough to have $h(n) \neq 0$. (Such an n

exists since we assumed $\neg\lambda_0$ and any provable sentence has arbitrary large proofs.) If $h(n) \Vdash \Box A$ then $h(n+1) = h(n)$, otherwise $h(n+1)$ will be the least submodel of $h(n)$ forcing $\neg A$. In both cases $h(n+1) \Vdash \Box A$ (recall that the code of a model is larger than the code of its proper submodels). Afterwards, h remains confined in a submodel of $h(n+1)$, so we can conclude that $\forall^\infty n\, h(n) \Vdash \Box A$. Thus $\lambda_{\Box A}$ follows by Claim 2.4.

$T + \neg\lambda_0 \vdash \neg\iota(\bot)$. Immediate.

$T + \neg\lambda_0 \vdash \iota(A \to B) \leftrightarrow (\iota(A) \to \iota(B))$. Proof is left to the reader.

This concludes the proof of Claim 2.6.

Claim 2.7 *For every $A \in \mathcal{L}$, $\mathcal{A} \models A$ iff $T \vdash \iota(A)$.*

Proof: ($\Longrightarrow$) Assume $\mathcal{A} \models A$, so for some $\mathcal{A}_{,n} \models A$. Since n is standard, $h(n) = 0$ and, by Σ_1 completeness, $T \vdash h(n) = 0 \wedge \mathcal{A}_{,n} \models A$. So $\iota(A)$ by Claim 2.7.

($\Longleftarrow$) Vice versa, if $T \vdash \iota(A)$ we have in particular that $T \vdash \lambda_0 \vee \lambda_A$. Assume for a contradiction that $\mathcal{A} \not\models A$ and let n be the code of the proof of $\lambda_0 \vee \lambda_A$. In particular we have that $\mathcal{A}_{,n} \not\models A$ then $h(n+1) = 0$. This n is a standard number, so this contradicts the fact that h will spend all its standard life in 0.

The proof of the lemma is complete but for the definition of the predicate $\mathcal{T}$. We introduce the formula $V(\sigma)$ which says roughly: A_σ is $\Box$-conservative over $\mathcal{A}$, i.e.,

$$V(\sigma) := \forall A[(\mathcal{A} \models A_\sigma \to \Box A) \to (\mathcal{A} \models \Box A)].$$

Assume strings have been coded into numbers in some natural way (e.g., choose $\Sigma_{\sigma(i)\downarrow=1} 2^i$ as the code for σ), so that on strings of equal length the relation "$<$" coincides with the relation "precedes lexicographically," or, when strings are thought of as nodes of a binary tree, "is on the left of." Let $U(\sigma)$ be the sentence which says that σ is the leftmost string satisfying $V(\sigma)$:

$$U(\sigma) := V(\sigma) \wedge \forall\tau \in 2^{i+1}\, (\tau < \sigma \to \neg V(\tau)).$$

If $A = A_i$ let $\mathcal{T}(A)$ hold if there is $\sigma \in 2^{i+1}$ such that $U(\sigma)$ and $\sigma(i) = 1$. We have to show that for every standard formula properties (1)–(4) of $\mathcal{T}$ are provable in T. First let us remark that for all standard i, T proves $\exists\sigma \in 2^{i+1}\, U(\sigma)$, i.e., there exists the leftmost string σ satisfying $V(\sigma)$. Reason in T. A string satisfying $V(\sigma)$ must exist or else for every $\sigma \in 2^{i+1}$ there would be a modal formula C_σ such that $\mathcal{A} \models A_\sigma \to \Box C_\sigma$ and $\mathcal{A} \not\models \Box C_\sigma$. Since $\bigvee_{\sigma \in 2^{i+1}} A_\sigma$ is a tautology, one would have $\mathcal{A} \models \bigvee_{\sigma \in 2^{i+1}} \Box C_\sigma$. By the s.d.p. of $\mathcal{A}$ (provable in T), $\mathcal{A} \models \Box C_\sigma$ for some σ, a contradiction. Now once we know that one string σ exists satisfying $V(\sigma)$, the existence of the minimal one is again a consequence of the standardness of i since the quantifiers over strings in 2^{i+1} may be transformed in finite conjunctions and disjunctions. This proves our remark. Now we check in turn that properties (1)–(4) which we required for $\mathcal{T}$ are provable in T.

(1) $\neg\mathcal{T}(\bot)$
(2) $\mathcal{T}(A \to B) \leftrightarrow (\mathcal{T}(A) \to \mathcal{T}(B))$
(3) $\mathcal{A} \models A \to \mathcal{T}(A)$
(4) $\mathcal{T}(\Box A) \to \mathcal{A} \models A$.

We reason in T. It is obvious that for no string σ such that $V(\sigma)$, $\sigma(\bot) = 1$, so (1) holds. (We write $\sigma(A)$ for $\sigma(i)$ where $A = A_i$.) To prove (2) assume first that $\mathcal{T}(A \to B)$ and $\mathcal{T}(A)$. Let σ be a sufficiently long string such that $U(\sigma)$ and $\sigma(A \to B) = \sigma(A) = 1$. Then $\sigma(B) = 1$ or else $A_\sigma \leftrightarrow \bot$ and surely could not satisfy $V(\sigma)$. The converse is similar. Property (3) is also a direct consequence of the existence of an arbitrary (standard) long string satisfying $U(\sigma)$. For such a string we must have $\sigma(A) = 1$ or else $\mathcal{A} \models A_\sigma \to \bot$ and, by the definition of $V(\sigma)$ we have that $\mathcal{A} \models \bot$. Lastly, to prove (4) assume that $\mathcal{T}(\Box A)$. Let σ be a sufficiently long string such that $U(\sigma)$ and $\sigma(\Box A) = 1$. Then $\mathcal{A} \models A_\sigma \to \Box A$, so, by the definition of $V(\sigma)$, we have that $\mathcal{A} \models \Box A$. By the s.d.p. of $\mathcal{A}$ we get $\mathcal{A} \models A$.

This completes the proof of Lemma 2.1.

3 The theorems We shall use Lemma 2.1 to prove the two theorems announced in the Introduction. They characterize the r.e. sets interpretable in a theory of infinite height.

Theorem 3.1 *If $\mathcal{A}$ is an r.e. set of modal formulas and T is a Σ_1-sound theory containing $I\Delta_0 + exp$, then $\mathcal{A}$ is interpretable in T iff $\mathcal{A}$ is s.d.*

Theorem 3.2 *If $\mathcal{A}$ is an r.e. set of modal formulas and T is a Σ_1-ill theory of infinite height containing $I\Delta_0 + exp$, then $\mathcal{A}$ is interpretable in T iff $\mathcal{A}$ has infinite height.*

The "only if" parts of the theorems are trivial. To prove the first theorem we show that, if $\mathcal{A}$ is an r.e. set with the s.d.p. and T is a Σ_1-sound theory, then we can find a description of $\mathcal{A}$ in T such that T proves the s.d.p. of $\mathcal{A}$. Analogously for the second theorem. For the sake of readiability we shall give these proofs in an informal style, i.e., we shall merely describe algorithms and take for granted their formalizability in the language of T.

Suppose $\mathcal{A}$ is an r.e. set of modal formulas and let $A \in \mathcal{A}_{,s}$ be any description of $\mathcal{A}$. With this description we associate in a natural way the algorithm $\{\mathcal{A}_{,s}\}_{s\in\omega}$ enumerating $\mathcal{A}$, i.e., an increasing recursive sequence of finite sets $\{\mathcal{A}_{,s}\}_{s\in\omega}$ such that $\mathcal{A} = \bigcup_{s\in\omega} \mathcal{A}_{,s}$. We shall construct a new algorithm $\{\mathcal{V}_{,s}\}_{s\in\omega}$ enumerating the same set $\mathcal{A}$ such that the canonical translation of $\{\mathcal{V}_{,s}\}_{s\in\omega}$ in the language of arithmetic yields a description with the desired properties.

The proofs of Theorems 3.1 and 3.2 need two modal lemmas, respectively Lemmas 3.3 and 3.4. These are the adaptations of some lemmas from [7]. We shall present them in a form which is easily formalized and proved in $I\Delta_0 + \exp$. Their proofs are moved to the end of this section.

A finite set $\mathcal{C}$ of formulas is said to be *adequate* if it is closed under subformulas and (up to provable equivalence) closed under boolean connectives; i.e., (i) $\bot \in \mathcal{C}$; (ii) all subformulas of any $B \in \mathcal{C}$ are in $\mathcal{C}$; and (iii) for every $B, C \in \mathcal{C}$ there exists $D \in \mathcal{C}$ such that $\Vdash D \leftrightarrow (B \to C)$.

Lemma 3.3 *Let $\mathcal{C}$ be a finite adequate set containing $\mathcal{A}$. The following are equivalent:*

(a) $\mathcal{A}$ is s.d.;

(b) $\mathcal{A} \not\models \bot$ and $\forall B, C \in \mathcal{C}\; \mathcal{A} \models \Box B \vee \Box C \Longrightarrow \mathcal{A} \models B$ or $\mathcal{A} \models C$.

Proof of Theorem 3.1: We are now ready to present the algorithm required to prove Theorem 3.1. We may code finite sets of formulas with natural numbers. The property

"s codes an adequate set" is Δ_0. With the same notation as the example given above, consider the following algorithm $\{\mathcal{V}_{,s}\}_{s\in\omega}$.

Stage 0. $\mathcal{V}_{,0} = \varnothing$.

Stage $s+1$. Let A be the minimal formula (if such exists) such that $A \in \mathcal{A}_{,s} - \mathcal{V}_{,s}$. If for some adequate set $\mathcal{C}$ of code less than s, $A \in \mathcal{C}$, $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,s} \cap \mathcal{C}$, and condition (b) of Lemma 3.3 holds for $\mathcal{A}_{,s} \cap \mathcal{C}$, then let $\mathcal{V}_{,s+1} = (\mathcal{A}_{,s} \cap \mathcal{C})$; otherwise let $\mathcal{V}_{,s+1} = \mathcal{V}_{,s}$.

We check by induction on the code of the (standard) formula A that $A \in \mathcal{A}$ iff $A \in \bigcup_{s\in\omega} \mathcal{V}_{,s}$. Since $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,s}$, only one implication needs to be proved. Suppose for a contradiction there is a formula such that $A \in \mathcal{A}_{,s} - \mathcal{V}_{,s}$ for all large enough $s \in \omega$. Fix A and s such that for all $r \geq s$, A is the least formula in $\mathcal{A}_{,r} - \mathcal{V}_{,r}$. Fix an adequate set $\mathcal{C}$ such that $\{A\} \cup \mathcal{V}_{,s} \subseteq \mathcal{C}$. Clearly $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,n} \cap \mathcal{C}$. Since $\mathcal{A}$ is s.d. and we assumed it closed under $\models$, condition (b) of Lemma 3.3 holds for $\mathcal{A}_{,n} \cap \mathcal{C}$. So $\mathcal{V}_{,n+1} = \mathcal{A}_{,n} \cap \mathcal{C}$, a contradiction. It remains to be checked that T proves the s.d.p. of $\bigcup_s \mathcal{V}_{,s}$. For this we need a formalized version of Lemma 3.3 in $I\Delta_0 + \exp$, and we invite the reader to check that all models used in the proof reported below are bounded by a few nested exponentiations of the code of the given adequate set $\mathcal{C}$. Consequently, the theorem holds in any model of $I\Delta_0 + \exp$. From Lemma 3.3 it follows that for all stages s the sets $\mathcal{V}_{,s}$ are s.d., which clearly suffices.

Lemma 3.4 *Let $\mathcal{C}$ be a finite adequate set containing $\mathcal{A}$. The following are equivalent:*

(1) $\mathcal{A}$ has infinite height;

(2) there exists $B \in \mathcal{C}$ such that B is s.d. and $B \models \bigwedge \mathcal{A}$.

Proof of Theorem 3.2: Given a Σ_1-ill theory T choose a Δ_0 formula $\sigma(x)$ such that $T \vdash \exists x\, \sigma(x)$ and $\omega \models \forall x\, \neg\sigma(x)$. In every model of T there is a Δ_0-definable number n, namely the minimal witness of $\exists x\, \sigma(x)$. The idea of the proof is the following: given any algorithm $\mathcal{A}_{,s}$ enumerating $\mathcal{A}$, we construct a new algorithm which simulates $\mathcal{A}_{,s}$ until the nonstandard stage n. Once this stage is reached we stop the simulation and enumerate some arbitrary s.d. set containing $\mathcal{A}_{,n}$. In the real world this stage n is never reached, so this new algorithm enumerates the same set as the old one. But in any model of T this algorithm enumerates a finite s.d. set. Lemma 3.4 is used to guarantee that some s.d. formula $B \models \mathcal{A}_{,s}$ always exists.

Stage 0. $\mathcal{V}_{,0} = \varnothing$.

Stage $s+1$. Let A be the minimal formula (if such exists) such that $A \in \mathcal{A}_{,s} - \mathcal{V}_{,s}$. If for some adequate set $\mathcal{C}$ of code less than s, $A \in \mathcal{C}$, $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,s} \cap \mathcal{C}$, for some $B \in \mathcal{C}$ condition (b) of Lemma 3.3 holds, and $B \models \mathcal{A}_{,s} \cap \mathcal{C}$, then:

Case 1: if $\forall x \leq s\, \neg\sigma(x)$ let $\mathcal{V}_{,s+1} = \mathcal{V}_{,s} \cup (\mathcal{A}_{,s} \cap \mathcal{C})$.

Case 2: if $\exists x < s\, \sigma(x)$ let $\mathcal{V}_{,s+1} = \mathcal{V}_{,s} \cup \{\mathcal{A}\}$.

Otherwise, let $\mathcal{V}_{,s+1} = \mathcal{V}_{,s}$.

We check by induction on the code of the formula A that $A \in \mathcal{A}$ iff $A \in \bigcup_{s\in\omega} \mathcal{V}_{,s}$. Since $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,s}$, only one implication needs to be proved. We need consider only standard stages (recall that a description of $\mathcal{A}$ should verify: $A \in \mathcal{A}$ iff $\exists s \in \omega\, T \vdash A \in \mathcal{V}_{,s}$), so Case 2 never obtains. Suppose for a contradiction that there is a formula such that $A \in \mathcal{A}_{,s} - \mathcal{V}_{,s}$ for all $s \in \omega$. Fix A and s such that for all $r \geq s$, A is the least formula in $\mathcal{A}_{,r} - \mathcal{V}_{,r}$. Fix an adequate set $\mathcal{C}$ such that $\{A\} \cup \mathcal{V}_{,s} \subseteq \mathcal{C}$ (such an adequate set exists since A is standard). Let $n > s$ be larger than the code of $\mathcal{C}$ and such that $\mathcal{A} \cap \mathcal{C} \subseteq \mathcal{A}_{,n} \cap \mathcal{C}$. Clearly $\mathcal{V}_{,s} \subseteq \mathcal{A}_{,n} \cap \mathcal{C}$, and since $\mathcal{A}$ has infinite height, so does $\mathcal{A}_{,n} \cap \mathcal{C}$. Thus, condition (2) of Lemma 3.4 holds for $\mathcal{A}_{,n} \cap \mathcal{C}$. We may conclude that $\mathcal{V}_{,n+1} = \mathcal{A}_{,n} \cap \mathcal{C}$, a contradiction. To check that T proves the s.d.p. of $\bigcup_s \mathcal{V}_{,s}$ recall that in every model of T, $\bigcup_s \mathcal{V}_{,s} = \bigcup_{s<n+1} \mathcal{V}_{,s}$, where n is the least number such that $\sigma(n)$ and $\bigcup_{s<n+1} \mathcal{V}_{,s}$ is equivalent to a single s.d. formula B.

Proof of Lemma 3.3: The direction (a) $\Longrightarrow$ (b) is trivial. For the converse assume (b). Fix a set $\mathcal{A}t \subseteq \mathcal{C}$ such that:

$$\mathcal{A}t := \{G \in \mathcal{C} \mid \forall C \in \mathcal{C} \text{ either } G \Vdash C \text{ or } G \Vdash \neg C\}.$$

The elements $\mathcal{A}t$ are called *atoms*; roughly, they are conjunctions of maximal consistent subsets of $\mathcal{C}$. By the adequateness of $\mathcal{C}$, for every $C \in \mathcal{C}$, if $\nVdash \neg C$ then there is some atom $G \Vdash C$. Also, $\Vdash \bigvee \mathcal{A}t$, or else for some atoms G, $G \Vdash \neg \bigvee \mathcal{A}t$ quod non. Let $\gamma = \{G \in \mathcal{A}t \mid A \not\models G\}$. From $\Vdash \bigvee \mathcal{A}t$ and $\mathcal{A} \not\models \bot$ we can conclude that $\gamma \neq \varnothing$. We claim that there is a model of $\mathcal{A} \cup \{\Diamond G \mid G \in \gamma\}$. In fact, if not then $\mathcal{A} \models \bigvee_{G \in \gamma} \Box\neg G$. By (b), there is $G \in \gamma$ such that $\mathcal{A} \models \neg G$ quod non. This proves the claim.

Suppose now that for some formulas B_1, B_2 both $\mathcal{A} \not\models B_1$ and $\mathcal{A} \not\models B_2$, so we may assume that there are two models k_1 and k_2 of $\mathcal{A}$ forcing respectively $\neg B_1$ and $\neg B_2$. We shall show that $\mathcal{A} \not\models \Box B_1 \vee \Box B_2$ by constructing a model k' of $\mathcal{A}$ which contains k_1 and k_2 as proper submodels. The s.d.p. of $\mathcal{A}$ will follow.

Let k be a model of $\mathcal{A} \cup \{\Diamond G \mid G \in \gamma\}$. Let r, r_1 and r_2 be the roots of respectively k, k_1, and k_2. Let R, R_1 and R_2 be the respective accessibility relations. Let k' be the model obtained by grafting k_1 and k_2 above the root of k. More precisely, the universe of k' is the disjoint union of the universes of k, k_1, and k_2, and the accessibility relation of k' is the transitive closure of the relation $R \cup R_1 \cup R_2 \cup \{(r, r_1), (r, r_2)\}$. The forcing relation of k' is the union of the forcing relations of k, k_1, and k_2.

We claim that k' is a model of $\mathcal{A}$ and $k' \Vdash \neg\Box B_1 \wedge \neg\Box B_2$. Obviously k' forces $\neg\Box B_1 \wedge \neg\Box B_2$ because k_1 and k_2 are submodels of k', forcing respectively B_1 and B_2. To show that k' is a model of $\mathcal{A}$, we prove by induction on the complexity of subformulas $C \in \mathcal{C}$ that $k' \Vdash C$ iff $k \Vdash C$. The basis step is trivial, as is the induction for Boolean connectives. We prove the induction step for $\Box$. Assume $k' \Vdash \neg\Box C$. Then for some proper submodel w' of k', $w' \Vdash \neg C$. The model w' is a submodel of k_1 or k_2 or is a proper submodel of k. If w' is a proper submodel of k, then $k \Vdash \neg\Box C$ follows. Otherwise, let G be the atom forced in w'; since $C \in \mathcal{C}$, by the definition of an atom either $G \Vdash C$ or $G \Vdash \neg C$. But $G \Vdash C$ leads immediately to contradiction, so $G \Vdash \neg C$. Since both k_1 and k_2 are models of $\mathcal{A}$, $G \in \gamma$. By our choice of k, $k \Vdash \bigwedge_{G \in \gamma} \Diamond G$, so there is a proper submodel w of k which forces G. Hence $w \Vdash \neg C$ and $k \Vdash \neg\Box C$. Vice versa, if $k \Vdash \neg\Box C$ then for some proper submodel w of k, $w \Vdash \neg C$. Since w is also a proper submodel of k', $k' \Vdash \neg\Box C$ follows. This completes the proof of Lemma 3.3.

Proof of Lemma 3.4: ($\Longleftarrow$) is immediate. ($\Longrightarrow$) List the formulas of $\mathcal{C} = \{C_1, \ldots, C_n\}$. Define $\mathcal{A}_0 := \mathcal{A}$ and for all $i \leq n$ let $\mathcal{A}_{i+1} := \mathcal{A}_i \cup \{C_i\}$ if this has infinite height, $\mathcal{A}_{i+1} := \mathcal{A}_i$ otherwise. Finally choose in $\mathcal{C}$ a formula B equivalent to $\bigwedge \mathcal{A}_{n+1}$. If $B \models \Box C_i \vee \Box C_j$ then $B \wedge C_i$ or $B \wedge C_j$ has infinite height. (For suppose for some n both $B \wedge C_i \models \Box^n \bot$ and $B \wedge C_j \models \Box^n \bot$ then $B \models \Box C_i \to \Box^{n+1}$ and $B \models \Box C_j \to \Box^{n+1}$. Thus $B \models \Box^{n+1} \bot$, quod non.) So, one of C_i and C_j, say C_i, has been enumerated in $\mathcal{A}_{n+1}$, so $B \models C_i$. By Lemma 3.3, B is s.d.

Acknowledgment I wish to thank Volodya Shavrukov for numerous suggestions and corrections. I owe very much also to the stimulating criticisms and friendly encouragements of Lev Beklemishev. Comments from Dick de Jongh and Alessandro Berarducci have helped to make this paper more readable.

REFERENCES

[1] Bellisima, F., "On the modal logic corresponding to diagonalizable algebra theory," *Bolletino dell' Unione Matematica Italiana*, vol. 15 (1978), pp. 915–930.

[2] Bernardi, C., "On the equational class of diagonalizable algebras," *Studia Logica*, vol. 34, pp. 321–331.

[3] Hájeck, P. and P. Pudlák, *Metamathematics of First Order Arithmetic*, Springer–Verlag, Berlin, 1993.

[4] Magari, R., "Representation and duality theory for diagonalizable algebras," *Bolletino dell' Unione Matematica Italiana*, vol. 12 (1975), pp. 117–125.

[5] Magari, R., "The diagonalizable algebras," *Studia Logica*, vol. 34 (1975), pp. 305–313.

[6] Montagna, F., "On the diagonalizable algebra of Peano arithmetic," *Bolletino dell' Unione Matematica Italiana*, vol. 16 (1979), pp. 795–812.

[7] Shavrukov, V. Y., "Subalgebras of diagonalizable algebras of theories containing arithmetic," *Dissertationes Mathematicae*, vol. 323 (1993), 82 pp.

[8] Solovay, R., "Provability interpretations of modal logic," *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287–304.

Department of Mathematics and Computer Science
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands