MAX-PLANCK-INSTITUT
FÜR
INFORMATIK

Reflection using the derivability conditions

Seán Matthews Alex K. Simpson

MPI–I–94–234 July 1994

Im Stadtwald
66123 Saarbrücken
Germany

Reflection using the derivability conditions

Seán Matthews Alex K. Simpson

MPI–I–94–234 July 1994

Authors' Addresses

Sean Matthews,
Max-Planck-Institut für Informatik,
Saarbrücken, Germany
<sean@mpi-sb.mpg.de>
Alex Simpson,
Dept. of Computer Science,
University of Edinburgh, Scotland
<als@dcs.ed.ac.uk>.

Publication Notes

To appear in the proceedings of the conference on logic and algebra in memory of Roberto Magari, Siena, May 1994.

Abstract

We extend arithmetic with a new predicate Pr, giving axioms for Pr based on first-order versions of Löb's derivability conditions. We hoped that the addition of a reflection schema mentioning Pr would then give a non-conservative extension of the original arithmetic theory. The paper investigates this possibility. It is shown that, under special conditions, the extension is indeed non-conservative. However, in general such extensions turn out to be conservative.

Keywords

Proof theory, reflection.

§ 1 Introduction

In any 1-consistent recursively enumerable theory of arithmetic, T, one can follow Gödel's construction to obtain a 'provability predicate', a Σ_1-formula $Bew_T(x)$, satisfying:

$$T \vdash Bew_T(\ulcorner A \urcorner) \qquad \textit{iff} \qquad T \vdash A,$$

where $\ulcorner A \urcorner$ is the Gödel number of the formula A. Moreover, if T is sufficiently strong then Bew_T satisfies the following predicate (or 'uniform') versions of Löb's derivability conditions [7]:

$(D1)$ $\qquad\qquad$ *if* $\;T \vdash \forall x A \;$ *then* $\;T \vdash \forall x Bew_T(\ulcorner A\langle x\rangle \urcorner)$,

$(D2)$ $T \vdash \forall x(Bew_T[\ulcorner (A \to B)\langle x\rangle \urcorner] \to (Bew_T(\ulcorner A\langle x\rangle \urcorner) \to Bew_T(\ulcorner B\langle x\rangle \urcorner)))$,

$(D3)$ $\qquad T \vdash \forall x(Bew_T(\ulcorner A\langle x\rangle \urcorner) \to Bew_T(\ulcorner Bew_T(\ulcorner A\langle x\rangle \urcorner)\langle x\rangle \urcorner))$,

where we write $\ulcorner A\langle x\rangle \urcorner$ for a term with a free variable x 'disquoting' any occurrence of x in A (see section 2). Solovay, [9], showed that the original propositional versions of the derivability conditions identify all the valid 'modal' schematic properties of Bew_T (the other modal axiom, the formalization of Löb's theorem, is derivable from $(D1)$–$(D3)$ using the diagonalization lemma). Although the first-order derivability conditions above do not capture all the valid first-order schematic properties of Bew_T (see [1]), they do isolate a natural class of 'modal' properties satisfied by Bew_T.

All the aforementioned work treats the derivability conditions as *descriptive* in that their purpose is to describe properties of the *Bew* predicate. In this paper we consider them in an alternative *prescriptive* rôle. We define a language, $\mathcal{L}'$, by adding a new unary predicate symbol, Pr, to the original language $\mathcal{L}$. Then we define an $\mathcal{L}'$-theory, T', as the least theory containing T that is closed under the following analogues of $(D1)$–$(D3)$:

$(C1)$ $\qquad\qquad$ *if* $\;T' \vdash \forall x A \;$ *then* $\;T' \vdash \forall x Pr(\ulcorner A\langle x\rangle \urcorner)$,

$(C2)$ $\qquad T' \vdash \forall x(Pr(\ulcorner (A \to B)\langle x\rangle \urcorner) \to Pr(\ulcorner A\langle x\rangle \urcorner) \to Pr(\ulcorner B\langle x\rangle \urcorner))$,

$(C3)$ $\qquad\qquad T' \vdash \forall x(Pr(\ulcorner A\langle x\rangle \urcorner) \to Pr(\ulcorner Pr(\ulcorner A\langle x\rangle \urcorner)\langle x\rangle \urcorner))$,

where we assume that Gödel numbering has been extended to $\mathcal{L}'$. It is natural to ask how much of the behaviour of $Bew_{T'}$ is forced upon Pr by the satisfaction of $(C1)$–$(C3)$.

As remarked in Boolos and Jeffrey [2, p. 185], there are many 'predicates' other than Bew_T that satisfy $(D1)$–$(D3)$; for example, the predicate

expressing the property of being (the Gödel number of) a well-formed formula. Therefore it does not hold that $T' \vdash \forall x(Pr(x) \rightarrow Bew_{T'}(x))$. We shall see below that the converse implication fails too.

However, it occurred to us to consider the effect of adjoining the following analogue of the uniform reflection schema to T':

$$(R) \qquad \forall x(Pr(\ulcorner A\langle x\rangle \urcorner) \rightarrow A).$$

The question we were interested in was whether $T' + R$ is a non-conservative extension of the original theory T.

The possibility that $T' + R$ might not be conservative over T is initially plausible for the following reason. There is an evident 'intended' interpretation of T' in T under which Pr is (modulo some mapping of Gödel numbers) translated as Bew_T. This interpretation can be used to prove that T' is a conservative extension of T. However, the proof cannot be extended to show that $T' + R$ is conservative over T, and moreover one can show that no other interpretation of Pr will do instead (see theorem 1).

On the other hand, the same interpretation can be used to establish that any $\mathcal{L}$-formula entailed by $T' + R$ is a theorem in the theory obtained by extending T with its uniform reflection schema:

$$(Rfn) \qquad \forall x(Bew_T(\ulcorner A\langle x\rangle \urcorner) \rightarrow A).$$

Now, by Gödel's second incompleteness theorem, $T' + R$ is a non-conservative extension of T. Further, if T proves only true statements of arithmetic then so does $T + Rfn$. Thus, by the translation, $T' + R$ proves only true statements of arithmetic. Indeed our initial hope was that $T' + R$ might be a (necessarily conservative) extension of $T + Rfn$.

This possibility is of practical interest. If $T' + R$ were an extension of $T + Rfn$, then the definition of $T' + R$ would provide a feasible way of extending the reasoning powers of T without having to go through the laborious construction of Gödel's Bew_T predicate (although admittedly the definition of $T' + R$ does still require a Gödel numbering of formulae). However, in theorem 2 we prove that, unfortunately, $T' + R$ is always conservative over T. (This shows that, as claimed above, $T' \nvdash \forall x(Bew_{T'}(x) \rightarrow Pr(x))$.) Thus our construction of $T' + R$ does not give the desired *general* method of achieving a non-conservative extension of T.

Nevertheless, a slight and natural modification of the construction of $T' + R$ does lead to a non-conservative extension in one notable case. Since $Pr(t)$ is intended to mimic $Bew_T(t)$ it ought to be treated as a Σ_1-formula.

So if T supports induction over Σ_1-formulae then it is reasonable to include induction over atomic formulae of the form $Pr(t)$ in T'. In this case $T' + R$ provides full induction over formulae of $\mathcal{L}'$ (theorem 3), and thus contains Peano Arithmetic. So for any T containing Σ_1-induction but not full induction, a non-conservative extension can be obtained by our method.

Unfortunately, the non-conservative effect does not extend beyond Peano Arithmetic. Since Peano Arithmetic supports induction over arbitrary formulae of $\mathcal{L}$ it is natural to allow induction over arbitrary formulae of $\mathcal{L}'$ in T'. However, even allowing such induction, if T is Peano Arithmetic then $T' + R$ is conservative over T (theorem 4).

The paper is structured as follows. In section 2 we give the technical background to our work. In section 3 we give a semantic proof that, in general, $T' + R$ is conservative over T. In section 4 we consider extending induction to the new language, proving the non-conservativity result for arithmetic with Σ_1-induction and the conservativity result for Peano Arithmetic. Finally, section 5 contains some concluding remarks.

§ 2 Preliminaries

Throughout the paper we work, for convenience, with the language, $\mathcal{L}$, of Primitive Recursive Arithmetic (PRA) [4]. Thus when we refer to Peano Arithmetic (PA) we mean a definitional extension in $\mathcal{L}$ of the usual Peano Arithmetic (which is in the language of elementary arithmetic). As in section 1, we define a language $\mathcal{L}'$ by adding a new unary predicate symbol, Pr, to $\mathcal{L}$.

A Gödel-numbering of $\mathcal{L}'$ is an injective mapping from $\mathcal{L}'$ into the natural numbers. We assume some such mapping. We denote the number standing for a formula A of $\mathcal{L}'$ by $\ulcorner A \urcorner$, and similarly for terms, etc. We assume that all the relevant operations and predicates on formulae/terms are primitive recursive. In particular there is a primitive recursive function $sub(\cdot, \cdot, \cdot)$, such that for any formula A (or term t), and number n:

$$sub(\ulcorner A \urcorner, \ulcorner x \urcorner, n) = \ulcorner A[\overline{n}/x] \urcorner$$

where $\overline{n}$ is the numeral $s^n(0)$. We define the abbreviation:

$$\ulcorner A\langle t \rangle \urcorner \;=\; sub(\ulcorner A \urcorner, \ulcorner x \urcorner, t).$$

The restriction of $\ulcorner \cdot \urcorner$ to $\mathcal{L}$ gives us also a Gödel-numbering of $\mathcal{L}$.

Let T be any theory in $\mathcal{L}$ that extends PRA (i.e. supports quantifier-free induction), is recursively enumerable, and entails no false arithmetical sentence. Let $Bew_T(x)$ be Gödel's provability predicate for T. Because T

extends *PRA*, the formula Bew_T does indeed satisfy the properties $(D1)$–$(D3)$ of section 1.

We define the $\mathcal{L}'$-theory T' as in section 2.

Proposition 1 *T' is a conservative extension of T.*

Proof. We define a translation $(\cdot)^*$ from formulae of $\mathcal{L}'$ to formulae of $\mathcal{L}$. By the second recursion theorem, there is a number r such that (writing $\{r\}$ for the r-th partial recursive function):

$$\{r\}(\ulcorner A \urcorner) \;=\; \ulcorner A^* \urcorner$$

where the $(\cdot)^*$ translation is defined primitively recursively by:

$$\begin{aligned} P(t_1,\ldots,t_n)^* &= P(t_1,\ldots,t_n) \quad (\text{where } P \not\equiv Pr) \\ Pr(t)^* &= \exists y(T(\overline{r},t,y) \wedge Bew_T(U(y))) \\ (A \circ B)^* &= A^* \circ B^* \\ (\mathsf{Q}xA)^* &= \mathsf{Q}x(A^*) \end{aligned}$$

(here $\circ$ and Q are any propositional connective and quantifier, and T and U are Kleene's primitive-recursive T predicate and result-extraction function). By definition $\{r\}$ is primitive recursive. So there is a function symbol, *trans*, such that, by the formalized recursion theorem and quantifier-free induction:

(1) $$T \vdash \forall x \exists y(T(\overline{r},x,y) \wedge U(y) = trans(x)),$$

(2) $$T \vdash \forall x(trans(\ulcorner A\langle x\rangle \urcorner) = \ulcorner A^*\langle x\rangle \urcorner).$$

We now show that for all $\mathcal{L}'$-formulae A, if $T' \vdash A$ then $T \vdash A^*$. And, since $A^* \equiv A$ for any $\mathcal{L}$-formula A, this establishes the desired conservativity result. The proof is a straightforward induction on the closure conditions of T':

$(C1)$ Assume that $T' \vdash \forall xA$. By the induction hypothesis we have that $T \vdash (\forall xA)^*$, and therefore that $T \vdash \forall xA^*$. We need to show that $T \vdash (\forall xPr(\ulcorner A\langle x\rangle \urcorner))^*$; i.e., that

$$T \vdash \forall x \exists y(T(\overline{r},\ulcorner A\langle x\rangle \urcorner,y) \wedge Bew_T(U(y))).$$

However, $T \vdash \forall xyzw.(T(x,y,z) \wedge T(x,y,w)) \rightarrow z = w$. Therefore, by (1) and (2), the above formula is equivalent to $T \vdash \forall x Bew_T(\ulcorner A^*\langle x\rangle \urcorner)$. And this, in turn, follows from $(D1)$ and the fact that $T \vdash \forall xA^*$.

($C2$) We have to show that

$$T \vdash (\forall x\ Pr(\ulcorner (A \to B)\langle x\rangle \urcorner) \to (Pr(\ulcorner A\langle x\rangle \urcorner) \to Pr(\ulcorner B\langle x\rangle \urcorner)))^*$$

which, in the same way as ($C1$) above, reduces to

$$T \vdash \forall x(Bew_T(\ulcorner (A \to B)^*\langle x\rangle \urcorner) \to Bew_T(\ulcorner A^*\langle x\rangle \urcorner) \to Bew_T(\ulcorner B^*\langle x\rangle \urcorner)),$$

an instance of ($D2$).

($C3$) Similar to ($C2$) only making use of ($D3$) instead. □

Proposition 2 *For any $\mathcal{L}$-formula A, if $T' + R \vdash A$ then $T + Rfn \vdash A$.*

Proof. Let $(\cdot)^*$ be the translation from $\mathcal{L}'$ to $\mathcal{L}$ defined in the last proof. We already know that if $T' \vdash A$ then $T \vdash A^*$ and hence $T + Rfn \vdash A^*$. So we need only show that $T + Rfn \vdash R^*$. However, as in the proof above, this translates to showing that:

$$T + Rfn \vdash \forall x(Bew_T(\ulcorner A^*\langle x\rangle \urcorner) \to A^*),$$

which is an instance of *Rfn*. □

The above translation cannot be used to prove the conservativity of $T' + R$ over T, because it is not in general the case that $T \vdash \forall x(Bew_T(\ulcorner A^*\langle x\rangle \urcorner) \to A^*)$. One might wonder whether there is a cleverer translation that works instead. We shall give a quite general proof that in fact there is none.

Definition 3 *A* retraction *of $\mathcal{L}'$ onto $\mathcal{L}$ is a translation, $(\cdot)^\dagger$, from $\mathcal{L}'$-formulae to $\mathcal{L}$-formulae in which the predicates and function symbols of $\mathcal{L}$ are interpreted by themselves and Pr is interpreted by a formula $H(x)$ of $\mathcal{L}$. (It is a retraction in the appropriate category of languages and translations.) The translation $(\cdot)^*$ used in the above proofs is an example of a retraction of $\mathcal{L}'$ onto $\mathcal{L}$ in which $H(x)$ is the formula $\exists y(T(\overline{r}, x, y) \wedge Bew_T(U(y)))$.*

Let S be any $\mathcal{L}$-theory and S' be any $\mathcal{L}'$-theory extending S. A *retraction* of S' onto S is a retraction, $(\cdot)^\dagger$, from $\mathcal{L}'$ to $\mathcal{L}$ such that, for any $\mathcal{L}'$-formula A, $S' \vdash A$ implies $S \vdash A^\dagger$. (It is a retraction in the appropriate category of theories and interpretations.) It is clear that the existence of a retraction from S' to S implies that S' is a conservative extension of S. Indeed the proof of proposition 1 worked by establishing that $(\cdot)^*$ is a retraction of T' onto T. The impossibility of obtaining a similar translational proof of the conservativity of $T' + R$ over T is given by:

Theorem 1 *There is no retraction of $T' + R$ onto T.*

Proof. Suppose, for contradiction, that $(\cdot)^\dagger$ is a retraction of $T' + R$ onto T in which Pr is translated to $H(x)$. By the diagonalization lemma, there is an $\mathcal{L}$-sentence A such that:

$$T \vdash A \leftrightarrow \neg H(\ulcorner A \urcorner). \tag{3}$$

However, we claim that:

$$\text{if} \quad T \vdash A \quad \text{then} \quad T \vdash H(\ulcorner A \urcorner), \tag{4}$$

$$T \vdash H(\ulcorner A \urcorner) \rightarrow A. \tag{5}$$

To see that (4) holds, suppose that $T \vdash A$. Then $T' \vdash A$. So, by $(C1)$, it follows that $T' \vdash Pr(\ulcorner A \urcorner)$. Therefore $T \vdash (Pr(\ulcorner A \urcorner))^\dagger$. So $T \vdash H(\ulcorner A \urcorner)$ as required. For (5), we have that $T' + R \vdash Pr(\ulcorner A \urcorner) \rightarrow A$. So $T \vdash (Pr(\ulcorner A \urcorner) \rightarrow A)^\dagger$. Thus indeed $T \vdash H(\ulcorner A \urcorner) \rightarrow A$.

But from (3)–(5) it is easy to derive that T is inconsistent, which is a contradiction. □

This proof is similar to Montague's proof of the inconsistency of syntactic interpretations of certain modal logics [8].

§ 3 The general conservativity proof

Theorem 1 gives hope that $T' + R$ might be non-conservative over T. Unfortunately, this turns out not to be the case. The main theorem of this section is:

Theorem 2 *$T' + R$ is a conservative extension of T.*

The proof of the theorem involves some analysis of properties of Gödel-numbering when formalized in T. Recall that all the relevant operations and predicates on Gödel-numbers have been assumed to be primitive recursive. More specifically, we require primitive recursive 'constructors' for all function symbols, predicate symbols, connectives and quantifiers, which can be used to assemble terms and formulas. As T supports quantifier-free induction, each constructor is provably injective. Furthermore it is provable in T that the Gödel-number of a compound term/formula has a unique decomposition into the components out of which it is built. We also require a primitive recursive function *free-in*$(\cdot, \cdot)$, such that *free-in*$(\ulcorner A \urcorner, \ulcorner x \urcorner)$ if and only if x is

free in A (and similarly for terms). Again, quantifier-free induction suffices to ensure that:

$(S1)$ $\quad T \vdash \ulcorner A[s(x)/x]\langle y\rangle\urcorner = \ulcorner A\langle s(y)\rangle\urcorner$,

$(S2)$ $\quad T \vdash \neg \mathit{free\text{-}in}(\ulcorner A\urcorner, \ulcorner x\urcorner) \rightarrow \ulcorner A\langle y\rangle\urcorner = \ulcorner A\langle z\rangle\urcorner$,

$(S3)$ $\quad T \vdash (\mathit{free\text{-}in}(\ulcorner A\urcorner, \ulcorner x\urcorner) \wedge \ulcorner A\langle y\rangle\urcorner = \ulcorner A\langle z\rangle\urcorner) \rightarrow y = z$,

$(S4)$ $\quad T \vdash (\mathit{free\text{-}in}(\ulcorner t\urcorner, \ulcorner x\urcorner) \wedge \ulcorner t\urcorner \neq \ulcorner x\urcorner \wedge y \leq z) \rightarrow \ulcorner x\langle y\rangle\urcorner \neq \ulcorner t\langle z\rangle\urcorner$.

The meanings of $(S1)$–$(S3)$ are clear. The more cumbersome $(S4)$ reflects the fact that if t is different from, but contains, x and $m \leq n$, then $\overline{m}$ is different from $t[\overline{n}/x]$ (since the former is a strict subterm of the latter).

A further important property of a reasonable Gödel numbering is that if x occurs free in A then it is provable in T that the function $\mathit{sub}(\ulcorner A\urcorner, \ulcorner x\urcorner, \cdot)$ tends to infinity. (This follows from $(S3)$ by Σ_2-induction, but more generally is provable using quantifier free induction because $\mathit{sub}(\ulcorner A\urcorner, \ulcorner x\urcorner, \cdot)$ has a primitive recursive 'inverse'.) Our use of this fact is model-theoretic: let $\mathfrak{M} = (D, \leq, 0, s, \ldots)$ be a model of T. If $d \in D$ is non-standard, then the denotation, $\ulcorner A\langle d\rangle\urcorner$, of the term $\ulcorner A\langle x\rangle\urcorner$ with d assigned to x is also non-standard. On the other hand, if x does not occur free in A then, by (S2), $\ulcorner A\langle d\rangle\urcorner$ is standard and equal to $\ulcorner A\urcorner$.

Theorem 2 will be proved semantically. Let $\mathfrak{M}$ be a model of T. We extend $\mathfrak{M}$ to a $\mathcal{L}'$-structure, $\mathfrak{M}'$, by defining, for $d \in D$:

$Pr(d)$ if there exists an $\mathcal{L}'$-formula A and an element $d' \in D$ such that $d = \ulcorner A\langle d'\rangle\urcorner$ and $T' \vdash \forall x A$.

We now prove a sequence of results aiming to show that $\mathfrak{M}'$ is a model of $T' + R$ (proposition 7). We write $A \equiv B$ ($t \equiv t'$) for syntactic identity between formulae (terms).

Lemma 4 *If $d \in D$ is non-standard, $d \leq d' \in D$ and $\ulcorner A\langle d\rangle\urcorner = \ulcorner B\langle d'\rangle\urcorner$ then there exists n such that $A \equiv B[s^n(x)/x]$.*

Proof. The proof is by induction on the structure of B. Suppose that $d \in D$ is non-standard and $d \leq d' \in D$.

We first show, by induction on the structure of terms t, that if $\ulcorner t'\langle d\rangle\urcorner = \ulcorner t\langle d'\rangle\urcorner$ then:

1. If x does not occur free in t, then $t' \equiv t$.

2. If x occurs free in t then there exists n such that $d' = s^n(d)$ and $t' \equiv t[s^n(x)/x]$.

Suppose the term is a variable, y, different from x, and $\ulcorner t'\langle d\rangle \urcorner = \ulcorner y\langle d'\rangle \urcorner$. Now x is not free in y so, by (S2), $\ulcorner y\langle d'\rangle \urcorner = \ulcorner y \urcorner$ and is standard. Thus $\ulcorner t'\langle d\rangle \urcorner$ is standard, which implies that x does not occur free in t'. So $\ulcorner t'\langle d\rangle \urcorner = \ulcorner t' \urcorner$. Therefore, by the injectivity of Gödel numbering, $t' \equiv y$ as required.

Suppose the term is x and $\ulcorner t'\langle d\rangle \urcorner = \ulcorner x\langle d'\rangle \urcorner$. We prove, by induction on the structure of t', that there exists n such that $d' = s^n(d)$ and $t' \equiv s^n(x)$. First, t' cannot be a variable y different from x because then $\ulcorner t'\langle d\rangle \urcorner$ would be standard whereas $\ulcorner x\langle d'\rangle \urcorner$ is non-standard. If t' is x then we are done with $n = 0$, as $d = d'$ by (S3). Lastly, suppose that t' is of the form $f'(t'_1, \ldots, t'_h)$ (with h possibly zero). Now $d' \in D$ is non-standard so it has a predecessor $d'' \in D$. Thus $\ulcorner x\langle d'\rangle \urcorner = \ulcorner x\langle s(d'')\rangle \urcorner = \ulcorner s(x)\langle d''\rangle \urcorner$, the last equality by (S1). But then $\ulcorner t'\langle d\rangle \urcorner = \ulcorner s(x)\langle d''\rangle \urcorner$. So, by the formalized injectivity of Gödel numbering, t' is of the form $s(t'')$ for some t'' such that $\ulcorner t''\langle d\rangle \urcorner = \ulcorner x\langle d''\rangle \urcorner$. Then, by the induction hypothesis, there exists n such that $d'' = s^n(d)$ and $t'' \equiv s^n(x)$. Thus $n + 1$ is the number required as $d' = s^{n+1}(d)$ and $t' \equiv s^{n+1}(x)$.

Suppose that the term is $f(t_1, \ldots, t_k)$ (where k is possibly zero) and $\ulcorner t'\langle d\rangle \urcorner = \ulcorner f(t_1, \ldots, t_k)\langle d'\rangle \urcorner$. Then t' cannot be a variable y different from x. If t' is x then x must occur free in some t_i (otherwise $\ulcorner f(t_1, \ldots, t_k)\langle d'\rangle \urcorner$ would be standard). However, $d \leq d'$ so, by (S4), $\ulcorner x\langle d\rangle \urcorner \neq \ulcorner f(t_1, \ldots, t_k)\langle d'\rangle \urcorner$, a contradiction. So t' must be of the form $f'(t'_1, \ldots, t'_h)$. But then, by formalized injectivity, we have that $f \equiv f'$. So $h = k$ and, for all i $(1 \leq i \leq k)$ $\ulcorner t'_i\langle d\rangle \urcorner = \ulcorner t_i\langle d'\rangle \urcorner$. If x does not occur free in any t_i then, by the induction hypothesis, $t'_i \equiv t_i$ for all i and thus $t' \equiv f(t_1, \ldots, t_k)$ as required. If x does occur free in some t_i then, by the induction hypothesis, there exists n such that $d' = s^n(d)$ and, for all i, $t'_i \equiv t_i[s^n(x)/x]$. So indeed $t' \equiv f(t_1, \ldots, t_k)[s^n(x)/x]$.

It remains only to extend the induction to formulae. One proves, by induction on the structure of B, that $\ulcorner A[d] \urcorner = \ulcorner B[d'] \urcorner$ implies that if x does not occur free in B then $A \equiv B$ and if x does occur free in B then there exists n such that $d' = s^n(d)$ and $A \equiv B[s^n(x)/x]$. The straightforward argument, similar to the case for $f'(t'_1, \ldots, t'_h)$ and $f(t_1, \ldots, t_k)$ above, is omitted. The result follows. □.

Lemma 5 *1. If $d \in D$ is standard then $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$ if and only if $T' \vdash A[\overline{d}/x]$.*

2. *If $d \in D$ is non-standard then $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$ if and only if there exists n such that $T' \vdash \forall x(A[s^n(x)/x])$.*

Proof.

1. Suppose $d \in D$ is standard and $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. Then $\ulcorner A[\overline{d}/x]\urcorner = \ulcorner A\langle d\rangle \urcorner = \ulcorner B\langle d'\rangle \urcorner$ for some $d' \in D$ and B such that $T' \vdash \forall x B$ (by the definition of the extension of Pr in $\mathfrak{M}'$). Now if d' is standard then $T' \vdash B[\overline{d'}/x]$ and $\ulcorner A[\overline{d}/x]\urcorner = \ulcorner B[\overline{d'}/x]\urcorner$ so $A[\overline{d}/x] \equiv B[\overline{d'}/x]$. Thus indeed $T' \vdash A[\overline{d}/x]$. If, however, d' is non-standard then x cannot occur free in B. Therefore $T' \vdash B$ and $\ulcorner A[\overline{d}/x]\urcorner = \ulcorner B\urcorner$ so $A[\overline{d}/x] \equiv B$. Thus again $T' \vdash A[\overline{d}/x]$ as required.

   Conversely, suppose that $T' \vdash A[\overline{d}/x]$. Then trivially $T' \vdash \forall x A[\overline{d}/x]$. It follows that $\mathfrak{M}' \models Pr(\ulcorner A[\overline{d}/x]\urcorner)$. Thus indeed $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$.

2. Suppose that $d \in D$ is non-standard, and that $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. Then $\ulcorner A\langle d\rangle \urcorner = \ulcorner B\langle d'\rangle \urcorner$ for some $d' \in D$ and B such that $T' \vdash \forall x B$. If $d \leq d'$ then, by lemma 4, $A \equiv B[s^m(x)/x]$ for some m. So clearly $T' \vdash \forall x A$, and the n we are required to find is zero. If $d' < d$ and d' is non-standard then, by lemma 4, $A[s^n(x)/x] \equiv B$ for some n. But then we have found an n such that $T' \vdash \forall x A[s^n(x)/x]$. Lastly, if d' is standard then $\ulcorner B\langle d'\rangle \urcorner$ is standard, so x cannot occur free in A. Thus $A \equiv B[\overline{d'}/x])$ and $T' \vdash B[\overline{d'}/x]$. Therefore $T' \vdash \forall x A$ and again n is zero.

   Conversely, suppose there exists n such that $T' \vdash \forall x A[s^n(x)/x]$. As d is non-standard, there exists $d' \in D$ such that $d = s^n(d')$. By the definition of the extension of Pr, $\mathfrak{M}' \models Pr(\ulcorner A[s^n(x)/x]\langle d'\rangle \urcorner)$. But, by (S1), $\ulcorner A\langle d\rangle \urcorner = \ulcorner A[s^n(x)/x]\langle d'\rangle \urcorner$. So indeed $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. □

Proposition 6 *$\mathfrak{M}'$ is a model of T'.*

Proof. We must show that $\mathfrak{M}'$ validates (C1)–(C3).

(C1) Suppose $T' \vdash \forall x A$ and $d \in D$. Then it is immediate from the definition of the extension of Pr in $\mathfrak{M}'$ that $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$ as required.

(C2) Suppose $d \in D$, $\mathfrak{M}' \models Pr(\ulcorner (A \to B)\langle d\rangle \urcorner)$ and $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. If d is standard then, by lemma 5(1), $T' \vdash (A \to B)[\overline{d}/x]$ and $T' \vdash A[\overline{d}/x]$. So $T' \vdash B[\overline{d}/x]$ whence, by lemma 5(1), $\mathfrak{M}' \models Pr(\ulcorner B\langle d\rangle \urcorner)$ as required. If d is non-standard then, by lemma 5(2), there exists m such that $T' \vdash \forall x(A \to B)[s^m(x)/x]$ and there exists m' such that

$T' \vdash \forall x A[s^{m'}(x)/x]$. Therefore $T' \vdash \forall x B[s^n(x)/x]$ where n is the maximum of m and m'. So, by lemma 5(2), $\mathfrak{M}' \models Pr(\ulcorner B\langle d\rangle \urcorner)$ as required.

(C3) Suppose that $d \in D$ and $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. We omit the easy argument if d is standard. If d is non-standard then, by lemma 5(2), there exists n such that $T' \vdash \forall x A[s^n(x)/x]$. Whence, by (C1), $T' \vdash \forall x Pr(\ulcorner A[s^n(x)/x]\langle x\rangle \urcorner)$. Now, by n applications of (S1), $T' \vdash \forall x(Pr(\ulcorner A\langle s^n(x)\rangle \urcorner))$. So, by lemma 5(2), it follows, as required, that $\mathfrak{M}' \models Pr(\ulcorner Pr(\ulcorner A\langle x\rangle \urcorner)\langle d\rangle \urcorner)$. □

We now have a second proof of proposition 1. We have shown that any model $\mathfrak{M}$ of T extends to a model $\mathfrak{M}'$ of T'. It follows that T' is a conservative extension of T.

Proposition 7 *$\mathfrak{M}'$ is a model of $T' + R$.*

Proof. We need only verify R. Suppose then that $d \in D$ and $\mathfrak{M}' \models Pr(\ulcorner A\langle d\rangle \urcorner)$. If d is standard then, by lemma 5(1), $T' \vdash A[\overline{d}/x]$. Thus, by proposition 6, $\mathfrak{M}' \models A[\overline{d}/x]$. Therefore $\mathfrak{M}' \models A[d]$ as required. If d is non-standard then, by lemma 5(2), there exists n such that $T' \vdash \forall x(A[s^n(x)/x])$. By proposition 6, $\mathfrak{M}' \models \forall x(A[s^n(x)/x])$. But d is non-standard, so there exists $d' \in D$ such that $d = s^n(d')$. Therefore $\mathfrak{M}' \models A[d]$ as required. □

We have shown that any model of T extends to a model of $T' + R$. This completes the proof of theorem 2.

§ 4 Extending induction to $\mathcal{L}'$

The conservativity result of the last section is very general, as the proof works for an arbitrary T extending *PRA*. However, one important possibility has been overlooked: that of extending induction to the language $\mathcal{L}'$. Such an extension of induction would render impossible any model-theoretic conservativity proof along the lines of that above.

However, the rules of how one ought to extend induction are not immediately clear. For example, if T is *PRA* then it only has induction over quantifier-free formulae. Given that we are thinking of Pr as a Σ_1-formula in disguise, it does not seem reasonable to give T' any instances of induction not already available in *PRA*. Thus although uniform reflection together with *PRA* gives *PA*, there is no analogous situation using T' and R.

Things becomes a good deal more interesting if we consider *PRA* together with Σ_1-induction as the initial theory. We shall refer to this theory as $I\Sigma_1$.

With $I\Sigma_1$ as the base theory it seems reasonable to give the extended theory induction over some appropriate analogue of Σ_1 in $\mathcal{L}'$. To this end, we extend the whole arithmetical hierarchy to $\mathcal{L}'$. We define sets Σ'_n, Π'_n $(1 \leq n)$ as the least sets closed under:

1. $\Sigma'_n \subseteq \Sigma'_{n+1}, \Pi'_{n+1}$ and $\Pi'_n \subseteq \Sigma'_{n+1}, \Pi'_{n+1}$
2. If P is not Pr then $P(t_1, \ldots, t_n) \in \Sigma'_1, \Pi'_1$.
3. $Pr(t) \in \Sigma'_1$.
4. If $A, B \in \Sigma'_n$ then $A \wedge B, \exists x A \in \Sigma'_n$ and $\neg A \in \Pi'_n$.
5. If $A, B \in \Pi'_n$ then $A \wedge B, \forall x A \in \Pi'_n$ and $\neg A \in \Sigma'_n$.

The motivation is that Pr is supposed to be emulating a Σ_1 (but not Π_1) formula.

We now give the extended theory, $I\Sigma'_1$, the evident definition. $I\Sigma'_1$ is the smallest $\mathcal{L}'$-theory containing $I\Sigma_1$ and Σ'_1-induction and closed under $(C1)$–$(C3)$. Again we consider adding the analogue of uniform reflection, R. This time we do get the desired non-conservativity.

Theorem 3 *$I\Sigma'_1 + R$ contains PA.*

Proof. Suppose that we have that $I\Sigma'_1 \vdash A[0/x]$ and $I\Sigma'_1 \vdash \forall x\ (A \rightarrow A[s(x)/x])$. By applying $(C1)$ we get that $I\Sigma'_1 \vdash Pr(\ulcorner A[0/x] \urcorner)$ and $I\Sigma'_1 \vdash \forall x Pr(\ulcorner (A \rightarrow A[s(x)/x])[x] \urcorner)$. The former gives immediately:

$$I\Sigma'_1 \vdash Pr(\ulcorner A\langle 0 \rangle \urcorner).$$

The latter gives, by (C2), $I\Sigma'_1 \vdash \forall x (Pr(\ulcorner A\langle x \rangle \urcorner) \rightarrow Pr(\ulcorner A[s(x)/x]\langle x \rangle \urcorner))$ whence, by (S1):

$$I\Sigma'_1 \vdash \forall x (Pr(\ulcorner A\langle x \rangle \urcorner) \rightarrow Pr(\ulcorner A\langle s(x) \rangle \urcorner)).$$

We can now apply Σ'_1-induction to derive $I\Sigma'_1 \vdash \forall x Pr(\ulcorner A\langle x \rangle \urcorner)$. Therefore, by one application of R, we have that $I\Sigma'_1 + R \vdash \forall x A$.

It is now easy to see that $I\Sigma'_1 + R$ derives induction for any $\mathcal{L}'$-formula, B. Just apply the above argument to the formula:

$$A \equiv (B[0/x] \wedge \forall y (B[y/x] \rightarrow B[s(y)/x])) \rightarrow B.$$

The result follows. □

The above argument can be translated back to give an elegant proof, using only $(D1)$, $(D2)$ and $(S1)$, that $I\Sigma_1 + Rfn$ is a theory as strong as PA. Note that condition $(C3)$ was not needed in the proof. Also R was used only as a rule. We conjecture that if any of $(C1)$, $(C2)$ and R are weakened to their propositional versions then the resulting extension of $I\Sigma_1$ is conservative.

It is a special case of Lemma 9 below that $I\Sigma_1' + R$ is actually conservative over PA.

We conclude by showing that the trick used to prove theorem 3 cannot be generalized to derive stronger principles than full induction. Define PA' to be the least $\mathcal{L}'$-theory containing PA and induction over every $\mathcal{L}'$-formula and closed under $(C1)$–$(C3)$.

Theorem 4 *$PA' + R$ is a conservative extension of PA.*

We write $I\Sigma_n$ for the $\mathcal{L}$-theory obtained by extending PRA with Σ_n-induction. Following the definition of $I\Sigma_1'$ above, define $I\Sigma_n'$ to be the least $\mathcal{L}'$-theory containing PRA and Σ_n'-induction and closed under $(C1)$–$(C3)$. The proof of theorem 4 uses the observation that:

$$PA' = \bigcup_n I\Sigma_n'. \tag{6}$$

The inclusion $\bigcup_n I\Sigma_n' \subseteq PA'$ is obvious. For the converse, it is easy to show that $\bigcup_n I\Sigma_n'$ contains PRA, contains induction for arbitrary $\mathcal{L}'$-formulae and is closed under $(C1)$–$(C3)$. Thus $\bigcup_n T_n'$ satisfies the closure conditions of PA'. Therefore it contains PA'.

Lemma 8 *For all n, the theory $I\Sigma_n'$ is a conservative extension of $I\Sigma_n$.*

Proof. Let n be fixed but arbitrary. Consider the translation $(\cdot)^*$ from formulae of $\mathcal{L}'$ to formulae of $\mathcal{L}$ defined in the proof of proposition 1 (where T is taken to be $I\Sigma_n$). We claim that for all $A \in \mathcal{L}'$, if $I\Sigma_n' \vdash A$ then $I\Sigma_n \vdash A^*$. The claim is shown by a straightforward modification of the proof of proposition 1. The only additional case is to show that if A is an instance of Σ_n'-induction then $I\Sigma_n \vdash A^*$. But this holds because $(\cdot)^*$ maps Σ_n'-formulae to Σ_n-formulae, so A^* is an instance of Σ_n-induction. □

Lemma 9 *For all n, the theory $I\Sigma_n' + R$ is a conservative extension of PA.*

Proof. By theorem 3, $I\Sigma_n' + R$ contains PA. Let $(\cdot)^*$ be the translation used in the last proof. We claim that $I\Sigma_n' + R \vdash A$ implies $PA \vdash A^*$. We

already know that if $I\Sigma'_n \vdash A$ then $I\Sigma_n \vdash A^*$ and hence $PA \vdash A^*$. So we need only show that $PA \vdash R^*$. However, as in the proof of proposition 2, this follows from the following fact about PA [6]:

$$\text{for all } n, \quad PA \vdash \forall x(Bew_{I\Sigma_n}(\ulcorner A\langle x\rangle \urcorner) \to A). \qquad \square$$

Theorem 4 is now easily proved. By (6), it is clear that $PA' + R = \bigcup_n (I\Sigma'_n + R)$. So it follows from Lemma 9 that $PA' + R$ is indeed conservative over PA.

§ 5 Conclusions

In this paper we have investigated the potential of using the derivability conditions to induce properties of a provability predicate without having to go to the effort of following Gödel's construction. In particular we have focused on the possibility of obtaining non-conservative extensions using an extra axiom that mimics the uniform reflection schema.

Unfortunately, our results are mainly negative. Although we have obtained a non-conservative extension in one notable case, the resulting theory, PA, can be obtained much more easily just by giving the full induction schema. Nevertheless, we believe that our results (both of non-conservativity and of conservativity) are interesting.

One natural question is whether a more general method of obtaining non-conservative extensions could be obtained by using more powerful axioms than $(C1)$–$(C3)$. It is clear that the proof of theorem 4 is general to apply to any T' generated by a collection of axioms based on arithmetically valid formulae of predicate provability logic [1]. Nevertheless, the possibility remains that a more general method could be obtained by going beyond what we get from predicate provability logic (for example, by replacing $(C2)$ and $(C3)$ with single axioms quantifying over the Gödel numbers of formulae). We believe it to be an interesting programme to investigate such generalizations.

There are other ways of adding a new predicate to the language to obtain non-conservative extensions. For example, one can axiomatize the property of being a *satisfaction class* like in the work of Robinson, Kotlarski and others (see [5, Ch. 15]). Also, Feferman has obtained non-conservative extensions by axiomatizing a partial truth predicate [3]. It is unclear how this work relates to the provability based approach of this paper.

Acknowledgements

We thank Alan Smaill for encouraging this work.

References

[1] G. Boolos. *The logic of provability.* Cambridge University Press, Cambridge, 1993.

[2] G. Boolos and R. Jeffrey. *Computability and logic (third edition).* Cambridge University Press, Cambridge, 1989.

[3] S. Feferman. Reflecting on completeness. *J. Symbolic Logic*, 56:1–49, 1991.

[4] J.-Y. Girard. *Proof Theory and Logical Complexity, volume 1, Studies in Proof Theory.* Bibliopolis, Naples, 1987.

[5] R. Kaye. *Models of Peano arithmetic.* Oxford University Press, Oxford, 1991.

[6] G. Kreisel and A. Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitshrift für Mathematische Logic und Grundlagen der Mathematik*, 14:97–142, 1968.

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

[8] R. Montague. Syntactical treatment of modality, with corollaries on reflexion principles and finite axiomatisability. *Acta Phil. Fennica*, 16:153–167, 1963.

[9] R. Solovay. Provability interpretations of modal logic. *Israel J. Math.*, 25:287–304, 1976.