Predicative Functionals and an Interpretation of $\widehat{ID}_{<\omega}$ [*]

Jeremy Avigad

December 22, 1997

Abstract

In 1958 Gödel published his *Dialectica* interpretation, which reduces classical arithmetic to a quantifier-free theory T axiomatizing the primitive recursive functionals of finite type. Here we extend Gödel's T to theories P_n of "predicative" functionals, which are defined using Martin-Löf's universes of transfinite types. We then extend Gödel's interpretation to the theories of arithmetic inductive definitions $\widehat{ID}_n$, so that each $\widehat{ID}_n$ is interpreted in the corresponding P_n. Since the strengths of the theories $\widehat{ID}_n$ are cofinal in the ordinal Γ_0, as a corollary this analysis provides an ordinal-free characterization of the $<\Gamma_0$-recursive functions.

1 Introduction

1.1 Background

In 1958, Gödel [18] published what is now known as the *Dialectica* interpretation of arithmetic, consisting of a quantifier-free theory T and interpretation of Peano Arithmetic (PA) in that theory. T allows for the definition of functionals of arbitrary finite type using a generalized form of primitive recursion, and at the end of the article Gödel suggested that extensions of T could be constructed using "transfinite" types of some sort.

The question we address here is as follows: are there interesting extensions of T that are appropriate for the interpretation of stronger predicative theories, that is, theories of proof-theoretic strength less than or equal to Γ_0? Martin-Löf's theories of constructive mathematics [9, 22, 23, 24, 12, 8, 34, 36] offer a natural means of extension, by incorporating a rich type-building structure based on "universes" of types. We show here that these universes are sufficient for the purpose just described.

First, we present theories $P_0, P_1, \ldots$, axiomatizing what we are calling the "predicative" functionals. The theories P_n are bare-bones versions of Martin-Lof's theories ML_n, and are ordinal-free in that no mention of ordinals is made

[*] The work described here comprises a part of the author's Ph.D. dissertation [5], and was supported in part by an Alfred E. Sloan Dissertation Year Fellowship.

in the defining schemata or axioms. We then show how, via a sequence of syntactical manipulations, each of the theories $\widehat{ID}_n$ can be interpreted in the corresponding P_n. $\widehat{ID}_0$ is just PA and P_0 is a logic-free variant of T, so in the initial case our interpretation is essentially the same as Gödel's.

The theories $\widehat{ID}_n$ are so named because they allow one to define sets of natural numbers using a weak form of arithmetic inductive definition. Each $\widehat{ID}_n$ has proof-theoretic ordinal ordinal γ_n, where $\langle \gamma_n \rangle$ forms a sequence that is cofinal in the Feferman-Schütte ordinal Γ_0. The union of these theories, $\widehat{ID}_{<\omega}$, therefore has strength Γ_0 itself.

In [6] we show that the theory ATR_0, which is of just the right strength to formalize some important mathematical arguments (see [28, 29, 16]), is conservative over $\widehat{ID}_{<\omega}$ for arithmetic formulas. As a result, the work here characterizes the provably total recursive functions of both these theories.

In recent years a number of other theories have been reduced to versions of Martin-Löf's. Most of these reductions are carried out in the "formulas-as-types" framework (see [12, 19]). For example, Aczel [2, 4] interprets constructive set theory, and Palmgren [25] interprets intuitionistic versions of the theories ID_n. But while the formulas-as-types framework (which bears a close kinship to Kleene's modified realizability; see [32]) provides an elegant setting for the interpretation of constructive mathematics, it does not seem to provide the means to interpret classical theories. In particular, modified realizability does not verify Markov's principle for primitive-recursive predicates, and so does not directly provide a characterization of a classical theory's provable functions.

At present, many reductions of classical theories to constructive ones pass through the ordinal analysis of the former. We believe that the direct reduction of classical theories to constructive ones is an illuminating way of extracting their "constructive content," thereby complementing the gains of an ordinal analysis. To that end, functional interpretation seems to offer a fruitful approach. We hope that the results reported here will spark further research along these lines, motivated by the question: "What computational principles are needed to interpret stronger classical theories?"

1.2 Overview

In Section 2 we introduce axiomatizations of first- and second-order logic that are amenable to our interpretations, and define the relevant theories of first-order and second-order arithmetic, including $\widehat{ID}_n$ and $\Sigma^1_1{-}AC$. In Section 3 we define the purely equational theories P_n. Under a natural identification of quantifier-free formulas φ in the language of arithmetic with assertions φ' in the language of P_0, our main theorem can be stated as follows:

Theorem 1.1 *Each theory $\widehat{ID}_n$ is interpreted (in the sense of [21]) in the corresponding P_n. In particular, if $\widehat{ID}_n$ proves $\forall x\ \exists y\ \varphi(x, y)$, where φ is a quantifier-free formula in the language of arithmetic, there is a term t such that P_n proves $\varphi'(x, t(x))$.*

This implies that each provably total recursive function of $\widehat{ID}_n$ is represented by a term of P_n. In Section 3.2 we point out that the converse is also true, so that the provably total recursive functions of $\widehat{ID}_n$ are exactly the ones so represented.

The interpretations are defined iteratively. In Section 4 we recast the Dialectica interpretation in our framework, as follows:

$$\begin{array}{rcl} PA & \rightsquigarrow & PA^i \\ & \rightsquigarrow & P_0. \end{array}$$

First PA is interpreted in a variant PA^i based on a fragment of intuitionistic logic, and this in turn is interpreted in P_0.

In Section 5, we describe a four-step interpretation of $\widehat{ID}_1$ in P_1:

$$\begin{array}{rcl} \widehat{ID}_1 & \rightsquigarrow & \Sigma^1_1\text{-}AC \\ & \rightsquigarrow & \Sigma^1_1\text{-}AC^i \\ & \rightsquigarrow & \textit{Frege-}PA^i \\ & \rightsquigarrow & P_1. \end{array}$$

First, we interpret the fixed-point constants of $\widehat{ID}_1$ by Σ^1_1 formulas in the theory Σ^1_1-AC; this step is due to Aczel. A "double-negation" interpretation reduces Σ^1_1-AC to a more contructive variant Σ^1_1-AC^i. Then we use a Dialectica-style interpretation to reduce Σ^1_1-AC^i to a theory *Frege-*PA^i which is halfway between Σ^1_1-AC^i and P_1, by "reflecting" arithmetic formulas of Σ^1_1-AC down to quantifier-free formulas of *Frege-*PA^i. The final interpretation of *Frege-*PA^i in P_1 builds on the interpretation of PA in P_0: because P_1 has a universe of small types, it can "internalize" the Dialectica interpretation and therefore handle the reflected arithmetic formulas of *Frege-*PA^i.

In Section 6, we sketch an interpretation of $\widehat{ID}_2$ in P_2, which relies on the interpretation of $\widehat{ID}_1$ in P_1. Iterating this process yields the main result.

In [6] we present an effective proof of the following

Theorem 1.2 *ATR_0 is conservative over $\widehat{ID}_{<\omega}$ for arithmetic sentences, though there is necessarily a superexponential "speedup" in the lengths of proofs.*

Letting $P_{<\omega}$ be the union of the theories P_n, Theorem 1.1 yields

Corollary 1.3 *Suppose ATR_0 or $\widehat{ID}_{<\omega}$ proves $\forall x\ \exists y\ \varphi(x, y)$ where φ is a quantifier-free formula in the language of arithmetic. Then there is a term t such that $P_{<\omega}$ proves $\varphi'(x, t(x))$.*

In other words, our interpretation characterizes the provably total recursive functions of both theories.

The work described here builds on not only the Dialectica interpretation, but also work due to Feferman [15] and Aczel regarding $\widehat{ID}_{<\omega}$; nonconstructive Dialectica-style interpretations due to Feferman (see [13]); and Martin-Löf's theories. Though familiarity with these results will provide some context for ones described here, we have tried to keep this account self-contained. A more detailed presentation can be found in [5].[1]

2 The Relevant Theories of Arithmetic

2.1 Predicate logic

Below we use the term "theory" quite broadly, applying it to any system of axioms and rules of derivation. So, for example, the theory Σ^1_1-AC consists of a certain set of axioms, combined with the usual rules for forming terms and formulas in the language of second-order arithmetic, and a set of axioms and rules of classical predicate logic. The theory Σ^1_1-AC^i is similar, except that it is based on a fragment of intuitionistic predicate logic. Finally, P_1 is an entirely different type of theory, with rules of derivation and term formation that interleave with one another. In each case we'll take the notation $T \vdash \varphi$ and the phrase "T proves φ" to mean that φ can be derived from the axioms and rules of T, and an *interpretation* of T_1 in T_2 translates derivations in T_1 to derivations in T_2.

We want to emphasize that the symbol "$\vdash$" doesn't imply any particular choice of underlying logic. For example, if a theory is built on classical logic, we will always specify this as part of the theory. In this section we present axiomatizations of first- and second-order classical predicate logic, denoted respectively by C_1 and C_2, and then pick out intuitionistically justified fragments I_1 and I_2. By "second-order logic" we really mean a two-sorted logic with variables ranging over objects and sets, and an $\in$ relation between the two: semantically, we don't assume that the second-order variables range over the entire power set of the universe of objects, and the logic itself doesn't include any set comprehension axioms.

We fix $\forall$, $\wedge$, $\rightarrow$, and $\bot$ (false) as our basic connectives, defining $\neg\varphi$ as $\varphi \rightarrow \bot$. Our reason for doing so is that classical and intuitionistic logic agree on these connectives, a convenient fact that will facilitate our interpretations. We define $\varphi \vee \psi$ as $\neg(\neg\varphi \wedge \neg\psi)$ and $\exists x\ \varphi$ as $\neg\forall x\ \neg\varphi$, but note that although these definitions are classically justified, they are *not* adequate definitions for the corresponding intuitionistic connectives. If $\varphi(x)$ is a term or formula with free variable x, we will write $\varphi(t)$ to represent the formula obtained by replacing free instances of x by t, changing the bound variables of $\varphi(x)$ to prevent clashes, if necessary.

Many axiomatizations of predicate logic include all propositional tautologies, but to verify our interpretations we need to choose a finite set of schemata. Troelstra [32] provides a number of suitable candidates, due to Gödel [18], Kleene, and Spector [30]. The axioms presented here are Spector's, with the axioms regarding $\vee$ and $\exists$ removed, and the law of the excluded middle expressed

[1] The account here has been modified slightly to make the reliance on a double-negation interpretation more explicit.

in the form $\neg\neg\varphi \to \varphi$.

Axioms and rules of propositional logic

1. $\varphi \to \varphi$
2. If ψ, then $\varphi \to \psi$
3. If φ and $\varphi \to \psi$ then ψ
4. If $\varphi \to \psi$ and $\psi \to \theta$ then $\varphi \to \theta$
5. $\varphi \wedge \psi \to \varphi$, $\varphi \wedge \psi \to \psi$
6. If $\varphi \to \psi$ and $\varphi \to \theta$ then $\varphi \to \psi \wedge \theta$
7. If $\varphi \wedge \psi \to \theta$ then $\varphi \to (\psi \to \theta)$
8. If $\varphi \to (\psi \to \theta)$ then $\varphi \wedge \psi \to \theta$
9. $\bot \to \varphi$
10. $\neg\neg\varphi \to \varphi$

The diligent reader can check that the remaining axioms on Spector's list can be derived from these, once $\vee$ and $\exists$ are defined as described above.

Quantifier axioms and rules

1. If $\varphi \to \psi(x)$, and x is not free in φ, then $\varphi \to \forall x\ \psi(x)$
2. $\forall x\ \varphi(x) \to \varphi(t)$, for any term t whose variables are not bound in $\varphi(t)$.

For second-order logic, which has variables and quantifiers ranging over two different sorts, these quantifier rules are duplicated for each sort. The language of second-order logic also has a fixed binary relation symbol $\in$ to denote membership between objects of the two sorts. Finally, both logics come equipped with an equality symbol for first-order terms, governed by the axioms below.

Equality axioms and rules

1. $x = x$
2. $x = y \to (\varphi(x) \to \varphi(y))$, for quantifier-free formulas φ

The symmetry and transitivity of equality can be derived from these. We do not include a separate equality symbol for second-order terms, but rather define $X = Y$ to mean

$$\forall x\ (x \in X \leftrightarrow x \in Y).$$

Of the propositional axioms given above, only axiom 10, the law of the excluded middle, is not intuitionistically valid. We therefore define I_1 and I_2 to consist of C_1 and C_2 respectively, with this axiom removed. We need to emphasize the fact that I_1 and I_2 represent only a *fragment* of the usual intuitionistic predicate logic, since they have nothing to say about the connectives $\vee$ and $\exists$. As it turns out, this fragment is sufficient for our interpretations, and ignoring the latter connectives shortens our task considerably.

In this setting, Gödel's "double negation" interpretation simply adds a double-negation before atomic formulas; that is, we define $\bot^N$ to be $\bot$, φ^N to be $\neg\neg\varphi$ for atomic φ, and otherwise let the $\cdot^N$ translation commute with $\to$, $\wedge$, and $\forall$. (If we had included $\vee$ and $\exists$ among the basic connectives, the $\cdot^N$ mapping would replace them with the classically equivalent definitions mentioned above.) We then have

Theorem 2.1 *Suppose a set of sentences S proves a formula φ in C_1 or C_2. Then S^N proves φ^N in I_1 or I_2, respectively.*

To prove this one only needs to show that the claim is true of the axioms of each C_j and is maintained under the rules of interence. All the axioms and rules of each C_j other than propositional axiom 10 reduce to their counterparts in I_j, and a straightforward induction on formula complexity shows that the translation of the law of the excluded middle

$$(\neg\neg\varphi \to \varphi)^N = \neg\neg\varphi^N \to \varphi^N$$

is also derivable. (One uses the fact that in general formulas of the form $\neg\neg\neg\psi \to \neg\psi$ are intuitionisitically valid and provable from the axioms of I_j.)

For a full axiomatization of intuitionistic logic, or for more information on its relationship to classical logic, see [8, 32, 33, 36].

2.2 The theories $\widehat{ID}_n$

The language of Peano Arithmetic is a first-order language containing a constant symbol 0, a unary function symbol S denoting the successor operation, and binary function symbols $+$ and $\times$. Peano Arithmetic consists of classical predicate logic C_1 together with the following axioms and rules, where x' abbreviates $S(x)$.

Axioms of Peano Arithmetic

1. $\neg x' = 0$
2. $x' = y' \to x = y$
3. $x + 0 = x$
4. $x + y' = (x + y)'$
5. $x \times 0 = 0$
6. $x \times y' = x \times y + x$

7. Induction: From $\varphi(0)$ and $\forall x\ (\varphi(x) \rightarrow \varphi(x'))$ conclude $\forall x\ \varphi(x)$

The usual axiomatic form of induction follows easily from the last rule.

$\widehat{ID}_0$ is another name for PA. The language of $\widehat{ID}_1$ extends that of PA by adding an additional predicate P_φ for each arithmetic formula $\varphi(n, X)$ in which a new predicate X occurs only positively. ("Occurs postively" means that if φ is expressed solely in terms of $\neg$, $\wedge$, and $\forall$, then X only occurs within the scope of an even number of negations.) Such an arithmetic formula defines a set function

$$\Gamma_\varphi : P(\omega) \rightarrow P(\omega)$$

given by

$$\Gamma_\varphi(X) = \{y | \varphi(y, X)\};$$

positivity insures this function is monotone, i.e. for any sets A and B, $A \subset B$ implies $\Gamma_\varphi(A) \subset \Gamma_\varphi(B)$. Classically such a function has a fixed point, and the predicate P_φ is intended to denote such a set.[2] Consequently, the axioms of $\widehat{ID}_1$ consist of the axioms of PA with induction extended to formulas involving the new predicates, and an axiom

$$\forall y\ (P_\varphi(y) \leftrightarrow \varphi(y, P_\varphi))$$

for each P_φ.

The move from $\widehat{ID}_0$ to $\widehat{ID}_1$ can be iterated, whereby each theory $\widehat{ID}_{n+1}$ adds new constants for positive arithmetic formulas in the language of $\widehat{ID}_n$ and the corresponding fixed-point axioms. Taking the union of all these theories yields $\widehat{ID}_{<\omega}$; see [15] for more details.[3]

2.3 The theories Σ^1_1-$AC(\widehat{ID}_n)$

The theories Σ^1_1-$AC(\widehat{ID}_n)$ are all based on second-order logic C_2. Σ^1_1-AC includes the axioms of PA, the induction rule extended to second-order formulas, and a comprehension schema

$$\exists X\ \forall y\ (y \in X \leftrightarrow \varphi(y)) \tag{ACA}$$

for arithmetic formulas φ, possibly with set parameters other than X. To state the remaining axiom of Σ^1_1-AC, we define $m \in X_n$ to mean $\langle n, m\rangle \in X$, so that X can be interpreted as coding a countable sequence of sets $\langle X_n \rangle$. The Σ^1_1 axiom of choice asserts

$$\forall y\ \exists X\ \varphi(y, X) \rightarrow \exists X\ \forall y\ \varphi(y, X_y), \tag{Σ^1_1-AC}$$

[2] In fact, every monotone set function has a *least* fixed point, and the stronger theory ID_1 has additional axioms which assert that P_φ is contained in any other arithmetically-defined fixed point. See [10, 1].

[3] Note that the presentation in [15] adds only one new constant at each stage, so that each $\widehat{ID}_n$ has only n new constants. This difference is inessential, since in any proof in our version of $\widehat{ID}_n$ one can "collapse" all the fixed-point constants of each iterative depth to a single one.

where φ is any Σ^1_1 formula. By coding pairs of sets as a single set we can assume that φ is in fact arithmetic, since we can always "absorb" an existentially quantified set in X.

More generally, Σ^1_1-$AC(\widehat{ID}_n)$ is defined the same way except that we start with the language and axioms of $\widehat{ID}_n$ instead of PA and add (ACA) and $(\Sigma^1_1$-$AC)$ to those, allowing the new fixed-point constants to appear in the induction, comprehension, and choice axioms.

3 Predicative Functionals

3.1 The theories P_n

In this section we describe the functional theories P_n. P_0 is just a logic-free version of Gödel's T, while the theories P_n gain added strength through the use of "universes," which allow one to contruct transfinite types. Each P_n is essentially a bare-bones version of Martin-Löf's ML_n, described in more detail in [9, 22, 23, 24, 12, 8, 34, 36].

We start with an infinite stock of variables $w, x, y, z, \ldots$. Terms are built up from variables and constants as described by the rules below. As in Martin-Löf's type theories, we allow for four different kinds of assertions, or *judgements*:

1. A type, asserting that the term A denotes a type

2. $A = B$ type, asserting that the terms A and B denote the same type

3. $a \in A$, asserting that the term a denotes an element of the type A

4. $a = b \in A$, asserting that the terms a and b denote equal elements of A

Note that we have just one stock of variables, rather than variables of each type. To each judgement we append a list of *assumptions*, indicating what types the variables are supposed to represent. For example, we will write

$$a \in A \ (x \in B, y \in C)$$

to assert that a is a term of type A, on the assumption that the variable x is of type B and y is of type C. Such a list of assumptions is often called a "context" and the corresponding relationship between contexts and type judgements is sometimes written

$$(x \in B, y \in C) \vdash a \in A.$$

Below, however, we'll stick with the first notation.

Variables occuring in terms can be free or bound. If we want to emphasize that the variable x can occur freely in the term a we'll write it as $a[x]$, and we'll use the notation $a[b/x]$ to denote the term obtained by replacing every free occurance of x in a by the term b. Although we won't spell out the details of when a variable is free or bound, the rules are as one would expect: for example,

x occurs freely in the term x, but becomes bound in the terms $\lambda x.a$, $\Pi_{x \in A} B[x]$, and $\Sigma_{x \in A} B[x]$.

In presenting the rules below, we omit type assumptions that are unchanged in the conclusion. We also often leave type assumptions on terms implicit, so that the rule expressing the symmetry of equality,

$$\frac{a = b \in A}{b = a \in A}$$

is more properly written

$$\frac{A \text{ type} \quad a \in A \quad b \in A \quad a = b \in A}{b \in A}.$$

We will sometimes even omit the type of an equality judgement, writing

$$\pi_0(\langle a, b \rangle) = a$$

to define the projection function and leaving the reader to fill in the necessary type assumptions. For a more complete discussion of issues such as these we refer the reader to [9, 34].

We've divided the rules into four groups, depending on the form of their conclusion: A type, $A = B$ type, $a \in A$, or $a = b \in A$. The exception is the group of rules allowing us to state induction in the theory. These appear last.

Types of $\mathbf{P_n}$

1. Natural numbers: $\mathbb{N}$ type

2. Universes: U_0 type, U_1 type, $\ldots$, U_{n-1} type

3. Product type: $\dfrac{A \text{ type} \qquad \overset{(x \in A)}{B[x] \text{ type}}}{\Pi_{x \in A} B[x] \text{ type}}$

4. Sum type: $\dfrac{A \text{ type} \qquad \overset{(x \in A)}{B[x] \text{ type}}}{\Sigma_{x \in A} B[x] \text{ type}}$

5. Reflection: $\dfrac{a \in U_i}{a \text{ type}}$

6. Instantiation of variables: $\dfrac{a \in A \qquad \overset{(x \in A)}{B[x] \text{ type}}}{B[a/x] \text{ type}}$

The product type $\Pi_{x \in A} B[x]$ is a generalization of Gödel's $A \to B$: it denotes the type of functions that take an object a of A to an object b of $B[a]$, so that the range is dependent on the argument. If x is not free in B, this just boils down to $A \to B$. Similarly, $\Sigma_{x \in A} B[x]$ is a generalization of $A \times B$, denoting the type of ordered pairs of objects $\langle a, b \rangle$ where a is from A and b is from $B[a]$. Once again, if x is not free in A this type can be written $A \times B$.

Notice that the reflection rule allows one to define types that depend on variables, so that the types of P_0, which has no universes, are just the finite types of Gödel's T.

Equality of Types in $\mathbf{P_n}$

1. Reflection: $\dfrac{a = b \in U_i}{a = b \text{ type}}$

Note that the first rule says that an equality that holds of objects in a universe is transfered to the corresponding types. One again, this rule is unavailable in P_0.

Terms of $\mathbf{P_n}$

1. Variables: $x \in A$ $(x \in A)$
2. Zero: $0 \in \mathbb{N}$
3. Successor: $S \in \mathbb{N} \to \mathbb{N}$
4. Explicit definition: $\dfrac{\begin{matrix}(x \in A)\\ b[x] \in B[x]\end{matrix}}{\lambda x.b \in \Pi_{x \in A} B}$
5. Application: $\dfrac{a \in A \quad b \in \Pi_{x \in A} B}{b(a) \in B[a/x]}$
6. Pairing: $\dfrac{a \in A \quad b \in B[a/x]}{\langle a, b \rangle \in \Sigma_{x \in A} B}$
7. Projection: $\dfrac{c \in \Sigma_{x \in A} B}{\pi_0(c) \in A}, \dfrac{c \in \Sigma_{x \in A} B}{\pi_1(c) \in B[\pi_0(c)/x]}$
8. Primitive recursion: $\dfrac{a \in A[0/x] \quad b \in \Pi_{x \in \mathbb{N}}(A[x] \to A[x'/x])}{R_{a,b} \in \Pi_{x \in \mathbb{N}} A}$
9. Substitution of equal types: $\dfrac{a \in A \quad A = B \text{ type}}{a \in B}$
10. Instantiation of variables: $\dfrac{a \in A \qquad \begin{matrix}(x \in A)\\ b[x] \in B[x]\end{matrix}}{b[a/x] \in B[a/x]}$

In the application rule, $b(a)$ is shorthand for a formal term $\mathit{Apply}(b, a)$. To increase readability, if a is a term of an appropriate type, we'll use $a(b, c)$ as an abbreviation for $(a(b))(c)$. It will also be convenient, for objects $c \in \Sigma_{x \in A} B$, to use the notations c_0 and c_1 for the projections $\pi_0(c)$ and $\pi_1(c)$ respectively. As usual, x' denotes $S(x)$.

We haven't yet described rules for building terms to denote elements of the universes. Though they really belong with the previous group, we present them separately in the following list.

Terms of universes of $\mathbf{P_n}$

1. Natural numbers: $\mathbb{N} \in U_i$
2. Universes: $U_0 \in U_i, U_1 \in U_i, \ldots, U_{i-1} \in U_i$
3. Product type: $\dfrac{A \in U_i \qquad \overset{(x \in A)}{B[x] \in U_i}}{\Pi_{x \in A} B \in U_i}$
4. Sum type: $\dfrac{A \in U_i \qquad \overset{(x \in A)}{B[x] \in U_i}}{\Sigma_{x \in A} B \in U_i}$

All the rules above are duplicated for each $i < n$. These rules "reflect" type-building operations down into the universes, and therefore look much the same as P_n's ordinary type-building rules.

Next we present the rules that allow one to conclude that two terms of P_n are equal.

Equality of terms in $\mathbf{P_n}$

1. Reflexivity: $\dfrac{a \in A}{a = a \in A}$
2. Symmetry: $\dfrac{a = b \in A}{b = a \in A}$
3. Transitivity: $\dfrac{a = b \in A \quad b = c \in A}{a = c \in A}$
4. Explicit definition: $(\lambda x.a)(b) = a[b/x]$
5. Projection: $\pi_0(\langle a, b\rangle) = a$, $\pi_1(\langle a, b\rangle) = b$
6. Primitive recursion: $R_{ab}(0) = a$, $R_{ab}(x') = b(x, R_{ab}(x))$ $(x \in \mathbb{N})$
7. Substitution of equal terms: $\dfrac{a = b \in A \qquad \overset{(x \in A)}{c[x] \in C[x]}}{c[a/x] = c[b/x] \in C[a/x]}$

8. Substitution of equal types: $\dfrac{a = b \in A \quad A = B \text{ type}}{a = b \in B}$

9. Instantiation of variables: $\dfrac{a \in A \qquad \begin{matrix}(x \in A)\\ c[x] = d[x] \in C[x]\end{matrix}}{c[a/x] = d[a/x] \in C[a/x]}$

Finally, we define terms that allow us to express implication between assertions of equality of type $\mathbb{N}$ objects, and induction.

Logical Axioms of $\mathbf{P_n}$

1. Implication term: $\mathit{implies} \in \mathbb{N} \times \mathbb{N} \to \mathbb{N}$

2. Implication rules: $\mathit{implies}(0, x) = x$, $\mathit{implies}(x', y) = 0$, and $\mathit{implies}(x, 0) = 0$

3. Equality term: $\mathit{equals} \in \mathbb{N} \times \mathbb{N} \to \mathbb{N}$

4. Equality rules: $\mathit{equals}(0, x) = x$, $\mathit{equals}(y, 0) = y$, $\mathit{equals}(x', y') = \mathit{equals}(x, y)$

5. Substitution: $\mathit{implies}(\mathit{equals}(a, b), \mathit{equals}(c[a/x], c[b/x])) = 0$

6. Induction rule: $\dfrac{a(0) = 0 \quad \mathit{implies}(a(x), a(x')) = 0 \ (x \in \mathbb{N})}{a(x) = 0 \ (x \in \mathbb{N})}$

7. Equality transfer: $\dfrac{\mathit{equals}(a, b) = 0}{a = b \in \mathbb{N}}$

The idea behind the function *implies* is that numerical variables can be interpreted as propositional variables, where 0 denotes “true” and any nonzero value denotes “false.” So, for example, using appropriate substitution and identity rules we can have the following

Lemma 3.1 *Modus ponens is a derived rule of P_n; that is, we can derive the rule*

$$\frac{a = 0 \quad \mathit{implies}(a, b) = 0}{b = 0}.$$

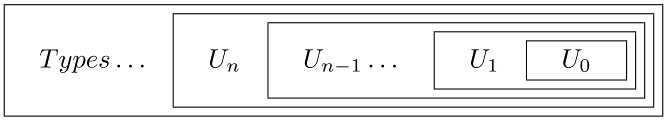
Similarly, $\mathit{equals}(a, b) = 0$ asserts that a and b represent the same number. We could have defined both *implies* and *equals* using primitive recursion, but then verifying that they satisfy the defining properties above would require some kind of induction rule, which is what we are using them to state.

The final rule allows us to transfer equality assertions of the form $\mathit{equals}(a, b) = 0$ to assertions of the form $a = b \in \mathbb{N}$. This allows us, for example, to use induction to derive the commutativity of the primitive recursively defined multiplication function. On the other hand, our interpretations will rely on the *equals* function rather than type $\mathbb{N}$ equality to interpret the equality symbol in the language of arithmetic. As a result, for our purposes, the transfer rule is unnecessary; we could just as well have omitted it.

We define the theory $P_{<\omega}$ to be the union of the theories P_n. In P_0 one can define primitive recursive functions such as addition, multiplication, the predecessor function, and truncated subtraction in the usual way. For each type A, we can also use primitive recursion and explicit definition to define a function $cases \in \mathbb{N} \times A \times A \to A$ by

$$\begin{aligned} cases(0, a, b) &= a \\ cases(x', a, b) &= b. \end{aligned}$$

When dealing with systems with multiple universes, it may be helpful to keep the following picture in mind:



By reflection anything in U_0 is in U_1, anything in U_1 is in U_2, and so on; and anything in any of the universes is a type. Note that the passage from P_n to P_{n+1} by adding a universe U_n actually allows us to define more terms in the smaller universes. For example, in P_0 we can do nothing more than define the finite types. Once we add the universe U_0, we can still define these types as elements of U_0, but we can do more: for example, we can use primitive recursion to define a function $T \in \mathbb{N} \to U_0$ by $T(0) = \mathbb{N}$ and $T(x') = T(x) \to T(x)$, and then conclude that $\Pi_{n \in \mathbb{N}} T(n)$ is an element of U_0, and hence a type.

3.2 The strength of the theories P_n

Our theories P_n are strongly derivative of Martin-Löf's theories ML_n, employing n universes and intensional equality, but lacking the W (well-foundedness) types. The differences between the theories stem from differences in underlying motivation.

Martin-Löf's theories provide a powerful framework for constructive mathematics in which mathematical statements φ are associated with types T_φ; the classical assertion that φ is true is associated with the constructive assertion that the type T_φ is inhabited by a term t witnessing φ's truth. This is known as the "formulas-as-types" interpretation, or the "Curry-Howard" isomorphism (see [12, 19]). To implement this framework, ML_n contains a rich assortment of type-forming operations. For example, for each type A and terms a and b of type A there is an associated equality type $I(a, b, A)$. In the intended interpretation, $I(a, b, A)$ is inhabited by a canonical element if and only if a and b represent the same object of type A.

Here we avoid the formulas-as-types framework, and so our systems lack many of the added formalisms of Martin-Löf's. The theories P_n can be viewed as minimal extensions of T, and as functional calculi they are conceptually simpler than the theories ML_n.

Given functional theories like ML_n and P_n, it is natural to ask what their models look like. In [32] there is a discussion of a number of different models of T, including the "full" set-theoretic model, models based on recursive functions, and term models. Many of these constructions extend to theories with universes as well (see [36]). In particular, Beeson [9] extends the "hereditarily effective operations" to form models of the theories ML_n, in which terms and types are represented by indices coding corresponding recursive objects. An adequate interpretation of the universe U_0 can be obtained in $\widehat{ID}_1$, and, furthermore, each theory $\widehat{ID}_n$ can appropriately define n such universes. As a result, each theory ML_n can be interpreted in $\widehat{ID}_n$, in such a way that terms of type $\mathbb{N} \to \mathbb{N}$ in ML_n correspond to provably total recursive functions in the $\widehat{ID}_n$. Since each P_n can be interpreted in the corresponding ML_n, we have the following converse to Theorem 1.1:

Theorem 3.2 *Each P_n is interpreted in $\widehat{ID}_n$. In particular, if P_n proves*

$$\varphi'(x, t(x))\ (x \in \mathbb{N})$$

for some term t in $N \to N$, there is an index e such that $\widehat{ID}_n$ proves

$$\forall x\, \exists y\, (\{e\}(x) \downarrow= y \wedge \varphi(x, y))$$

where $\{e\}$ denotes the eth recursive function.

Altogether, then, we have the chain of reductions

$$\widehat{ID}_n \rightsquigarrow P_n \rightsquigarrow ML_n \rightsquigarrow \widehat{ID}_n$$

indicating that the proof-theoretic strength of all these theories is the same.

4 Interpreting Peano Arithmetic in P_0

4.1 Interpreting PA in PA^i

The first part of the interpretation of PA in P_0 involves reducing the former theory to one that avoids the nonconstructive law of the excluded middle. To that end, we define the theory PA^i to consist of the same axioms and rules of PA given in Section 2.2, only this time based on the fragment of intuitionistic logic I_1.

Theorem 4.1 *Suppose PA proves φ. Then PA^i proves φ^N.*

Proof. If PA proves φ, then PA^N proves φ^N in I_1, so it suffices to show that PA^i proves these. But the doubly-negated defining equations for successor, plus, and times are easily provable in PA^i, and instances of induction in PA translate to instances of induction in PA^i. □

In fact, we can do a bit better. Since PA^i proves

$$\neg\neg x = y \to x = y$$

with a double induction on x and y, we could have allowed the $\cdot^N$-translation to leave these atomic formulas alone. For the sake of uniformity, we use the notation $\cdot^{N_0}$ to denote this "do nothing" translation.

4.2 Interpreting PA^i in P_0

In this section we show how to interpret PA^i in P_0. Recall that without universes the types of P_0 are obtained from $\mathbb{N}$ using the operations → and ×, and so P_0 is just a logic-free version of T. The interpretation we give here is essentially Gödel's, recast only slightly. As such, we only sketch the details and refer the reader to [35, 18, 30, 33, 32] for more information.

Since P_0 is logic-free, the first thing we have to do is embed the quantifier-free fragment of predicate logic, that is, the axioms and rules dealing with propositional logic and equality. The functions *implies* and *equals* were introduced for just this purpose, and allow us to build up logical combinations of assertions of type $\mathbb{N}$ equality. We define $false = 1$,

$$not(x) = implies(x, false),$$

and

$$and(x, y) = not(implies(x, not(y))).$$

We can then interpret any propositional combination φ of statements of type $\mathbb{N}$ equality in P_n with a single assertion $t_\varphi = 0$, by replacing $a = b \in \mathbb{N}$ by $equals(a, b) = 0$ and then using the logical functions just described. In the context of our functional theories, we will use greek letters φ to represent assertions of the form $t_\varphi = 0$. The notation $a \approx b$ will stand for the assertion $equals(a, b) = 0$, $\varphi \land \psi$ will stand for the assertion $and(t_\varphi, t_\psi) = 0$, and so on. With a little bit of work (see [5] for details), one can prove

Theorem 4.2 *In P_0 one can derive all the quantifier-free axioms and rules of predicate logic.*

We use $+$ and $\times$ to denote addition and multiplication in P_0, $x \dot{-} y$ to denote truncated subtraction, $x < y$ to denote

$$\neg(y \dot{-} x \approx 0),$$

and $x \leq y$ to denote

$$x < y \lor x \approx y.$$

We can then associate quantifier-free formulas φ of Peano Arithmetic with formulas φ' in the language of P_0. With an easy induction on formula complexity one can prove

Lemma 4.3 *Let φ be a closed quantifier-free formula in the language of PA, and φ' be the associated formula in P_0. Then*

$$(PA \vdash \varphi) \Leftrightarrow (P_0 \vdash \varphi') \Leftrightarrow (\mathbb{N} \models \varphi).$$

The idea of using functions to interpret type $\mathbb{N}$ equality and propositional connectives is well-known; see, for more information see also [30, 18, 35].

We now turn to the *Dialectica* interpretation proper. Our version of this interpretation maps every formula φ of PA^i to a formula φ^{D_0} of the form

$$\exists x \in A \, \forall y \in B \; f(x, y) = 0 \tag{1}$$

where A and B are types of P_0, and f is a term of type $A \times B \to \mathbb{N}$ whose free-variables are the same as those of φ. (For the moment, we allow for the possibility that either or both quantifiers are absent, in which case the type of f has to be modified accordingly.) The intepretation is defined in such a way that if PA^i proves φ, there is a term $a \in A$ of P_0 witnessing the existential quantifier in φ^{D_0}, in the sense that that P_0 proves

$$f(a, y) = 0 \; (y \in B).$$

We now define the mapping $\cdot^{D_0}$. Mapping terms t of PA to terms $t^{D_0} \in \mathbb{N}$ of P_0, we define $(t_1 = t_2)^{D_0}$ to be $t_1^{D_0} \approx t_2^{D_0}$. If φ^{D_0} is given by

$$\exists x \in A \, \forall y \in B \; \varphi_{D_0}(x, y)$$

and ψ^{D_0} is given by

$$\exists w \in C \, \forall z \in D \; \psi_{D_0}(w, z)$$

we define $(\varphi \to \psi)^{D_0}$ by

$$\begin{array}{rcl}
\varphi \to \psi & \rightsquigarrow & \exists x \, \forall y \; \varphi_{D_0}(x, y) \to \exists w \, \forall z \; \psi_{D_0}(w, z) \\
& \rightsquigarrow & \forall x \, \exists w \; (\forall y \; \varphi_{D_0}(x, y) \to \forall z \; \psi_{D_0}(w, z)) \\
& \rightsquigarrow & \forall x \, \exists w \, \forall z \, \exists y \; (\varphi_{D_0}(x, y) \to \psi_{D_0}(w, z)) \\
& \rightsquigarrow & \exists y, w \, \forall x, z \; (\varphi_{D_0}(x, y(x, z)) \to \psi_{D_0}(w(x), z)) \\
& \rightsquigarrow & \exists u \, \forall v \; \lambda u, v. t_{\varphi_{D_0}(v_0, u_0(v_0, v_1)) \to \psi_{D_0}(u_1(v_0), v_1))}(u, v) = 0.
\end{array}$$

Here only the last line is important (intermediate lines are presented for motivation) and we've suppressed the types for clarity: in the end, u is of type

$$(A \to C) \times (A \times D \to B)$$

and v is of type $A \times D$. In the final step we use the sum type and projection operations to combine types, and explicit definition and application so that the net result is of the form given by (1). To simplify the notation below we will omit this final step, but it should be taken as implicit.

Assuming φ^{D_0} an ψ^{D_0} are given as in the last paragraph, the translation of $\varphi \wedge \psi$ is unsurprising:

$$\begin{array}{rcl}
\varphi \wedge \psi & \rightsquigarrow & \exists x \, \forall y \; \varphi_{D_0}(x, y) \wedge \exists w \, \forall z \; \psi_{D_0}(w, z) \\
& \rightsquigarrow & \exists x, w \, \forall y, z \; (\varphi_{D_0}(x, y) \wedge \psi_{D_0}(w, z)).
\end{array}$$

$\bot^D$ is defined to be the formula $false = 0$. Finally, suppose the variable u is free in φ, and hence φ_{D_0}. In that case, $(\forall u\ \varphi)^{D_0}$ is given by the translation

$$\begin{aligned} \forall z\ \varphi &\rightsquigarrow \forall u\ \exists x\ \forall y\ \varphi_{D_0}(u, x, y) \\ &\rightsquigarrow \exists x\ \forall u, y\ \varphi_{D_0}(u, x(u), y), \end{aligned}$$

where the variable x is skolemized.

Note that if φ is a quantifier-free formula of PA^i then φ_{D_0} is essentially just φ'. Gödel's main theorem is as follows:

Theorem 4.4 *Suppose PA^i proves φ. Then P_0 proves φ^{D_0}. In other words, if φ^{D_0} is given by*

$$\exists x \in A\ \forall y \in B\ \varphi_{D_0}(x, y)$$

then there is a term $a \in A$ whose free variables correspond to those of φ such that P_0 proves

$$\varphi_{D_0}(a, y)\ (y \in B).$$

To prove this theorem one need only show that it holds true of the axioms of PA^i, and is maintained under rules of inference. The details are routine and well-known, and so here we only highlight two cases that will be of interest later on.

Assuming φ^{D_0} and ψ^{D_0} are as above, to handle the rule $\psi \Rightarrow \varphi \to \psi$ we can assume that there is an a such that P_0 proves $\psi_{D_0}(a, z)$, and we need terms b and c such that P_0 proves

$$\varphi_{D_0}(x, b(x, z)) \to \psi_{D_0}(c(x), z).$$

For this purpose we use the following

Lemma 4.5 (Canonical constants) *For each type A of P_0, there is a constant $canon_A \in A$.*

Proof. We can take

$$\begin{aligned} canon_{\mathbb{N}} &= 0 \\ canon_{A \to B} &= \lambda x.canon_B \\ canon_{A \times B} &= \langle canon_A, canon_B \rangle . \Box \end{aligned}$$

Using the lemma, define $c(x) = a$ and $b(x, z) = canon_B$.

To handle the rule $\varphi \to \psi, \varphi \to \theta \Rightarrow \varphi \to \psi \wedge \theta$, we are given terms a, a', b, and c such that

$$\varphi_{D_0}(x, a(x, z)) \to \psi_{D_0}(b(x), v) \qquad (2)$$

and

$$\varphi_{D_0}(x, a'(x, v)) \to \theta_{D_0}(c(x), v). \qquad (3)$$

We want terms e, f, and g such that

$$\varphi_{D_0}(x, e(x, v, z)) \to \psi_{D_0}(f(x), z) \wedge \theta_{D_0}(g(x), v). \tag{4}$$

If we define $f(x) = b(x)$, $g(x) = c(x)$, and

$$e(x, v, z) = cases(t_{\psi_{D_0}(f(x),z)}, a'(x, v), a(x, z))$$

we can show

$$\psi_{D_0}(f(x), z) \to (\varphi_{D_0}(x, e(x, v, z)) \leftrightarrow \varphi(v, a'(x, v)))$$

and

$$\neg\psi_{D_0}(f(x), z) \to (\varphi_{D_0}(x, e(x, v, z)) \leftrightarrow \varphi(x, a(x, z)).$$

By substitition we can replace $b(x)$ by $f(x)$ in (2) and $c(x)$ by $g(x)$ in (3). Equation (4) now follows using ordinary propositional logic.

The other axioms are readily taken care of; see, for example, [32, 30] for details. Note that since $\neg\varphi$ is defined as $\varphi \to \bot$, the translation of the axiom $\neg\neg\varphi \to \varphi$ is *not* in general verifiable in P_0, which explains our reliance on I_1.

Theorem 4.4 yields a characterization of the provably total recursive functions of PA:

Corollary 4.6 *Suppose PA proves $\forall x\ \exists y\ \varphi(x, y)$ where $\varphi(x, y)$ is quantifier free. Then there is a term t such that P_0 proves*

$$\varphi'(x, t(x))\ (x \in N).$$

In particular, if PA proves a quantifier-free φ, P_0 proves φ', and if PA proves $\bot$ then P_0 proves $1 = 0$.

The $\cdot^{D_0}$ mapping as we've defined it takes formulas of PA^i to formulas of the form (1), but allows for the possibility that either quantifier in the translation is absent. This nonuniformity will cause problems in Section 5.5, where we need to assume that every formula is mapped to something strictly in this form. We can achieve this by mapping atomic formulas $t_1 = t_2$ to

$$\exists u \in \mathbb{N}\ \forall v \in \mathbb{N}\ equals(t_1^{D_0}, t_2^{D_0}) = 0$$

where u and v are variables that don't appear in either term. Though this is inelegant, it does not harm the proof of Theorem 4.4 or Corollary 4.6. In fact, since u and v above are just dummy variables, we can prove the following

Lemma 4.7 *Let φ be a quantifier-free formula in the language of PA^i, and suppose φ^{D_0} is given by*

$$\exists w\ \forall z\ \varphi_{D_0}(w, z)$$

under the modified $\cdot^{D_0}$ translation. Then P_0 proves

$$\varphi_{D_0}(w, z) \leftrightarrow \varphi'.$$

In what follows, we'll use $\cdot^{D_0}$ to denote this modified form of the interpretation.

5 Interpreting $\widehat{ID}_1$ in P_1

5.1 Interpreting $\widehat{ID}_1$ in Σ^1_1-AC

Having fixed the interpretation of PA in P_0, we now turn to the interpretation of $\widehat{ID}_1$ in P_1. The first step is to reduce $\widehat{ID}_1$ to the theory Σ^1_1-AC by interpreting the fixed-point predicates by suitable Σ^1_1 formulas. The method given below is due to Aczel (see [15]) and hinges on the following

Lemma 5.1 *Let $\varphi(x, X)$ be an arithmetic formula in the language of Σ^1_1-AC, and assume X occurs only positively in φ. Then there is a Σ^1_1 formula $\psi(x)$ such that Σ^1_1-AC proves*

$$\forall x(\varphi(x, \{z \mid \psi(z)\}/X) \leftrightarrow \psi(x)).$$

Here $t \in \{z \mid \psi(z)\}$ should be interpreted as $\psi(t)$.

Proof (sketch). The proof is similar to the proof of Gödel's fixed-point lemma: we use a complete Σ^1_1 truth predicate and diagonalize. Let $Tr_\Sigma(\ulcorner\theta\urcorner, y, x)$ be such a truth predicate with the free variables shown, and consider the formula

$$\varphi(x, \{z \mid Tr_\Sigma(y, y, z)\}/X).$$

Since X occurs only positively in φ, Σ^1_1-AC proves this to be equivalent to a Σ^1_1 formula $\theta(y, x)$ (the axiom of choice is used to bring the second-order quantifiers out front), and so equivalent to $Tr_\Sigma(\ulcorner\theta\urcorner, y, x)$. Replacing y by $\ulcorner\theta\urcorner$ on both sides of the equivalence we get

$$\varphi(x, \{z \mid Tr_\Sigma(\ulcorner\theta\urcorner, \ulcorner\theta\urcorner, z)\}/X) \leftrightarrow Tr_\Sigma(\ulcorner\theta\urcorner, \ulcorner\theta\urcorner, x).$$

We can therefore take $\psi(x)$ to be the formula $Tr_\Sigma(\ulcorner\theta\urcorner, \ulcorner\theta\urcorner, x)$. □

This gives us the interpretation of $\widehat{ID}_1$ into Σ^1_1-AC, since we can interpret the fixed-point predicates by the corresponding formulas ψ given by the lemma. Induction in $\widehat{ID}_1$ reduces to induction in Σ^1_1-AC, and the other axioms of arithmetic are unchanged. Calling this interpretation $\varphi \mapsto \varphi^{\Sigma_1}$, we have

Theorem 5.2 *Suppose $\widehat{ID}_1$ proves φ. Then Σ^1_1-AC proves φ^{Σ_1}.*

By relativizing the truth predicate described above to constants in $\widehat{ID}_n$ we can, in the same way, interpret $\widehat{ID}_{n+1}$ in Σ^1_1-$AC(\widehat{ID}_n)$. Letting $\cdot^{\Sigma_{n+1}}$ denote the interpretation in which level $n+1$ fixed-point predicates are interpreted as Σ^1_1 formulas in $\Sigma^1_1(\widehat{ID}_n)$, we have

Theorem 5.3 *Suppose $\widehat{ID}_{n+1}$ proves φ. Then Σ^1_1-$AC(\widehat{ID}_n)$ proves $\varphi^{\Sigma_{n+1}}$.*

Note that formulas φ which don't contain any fixed-point predicates are unchanged by $\cdot^{\Sigma_n}$.

5.2 Interpreting Σ^1_1-AC in Σ^1_1-AC^i

Having reduced $\widehat{ID}_1$ to Σ^1_1-AC, our next step is to reduce Σ^1_1-AC to a more constructive variant, as we did with the interpretation of PA in PA^i. The theory Σ^1_1-AC^i is based on the intuitionistic fragment of second-order logic I_2, with the same arithmetic and induction axioms as Σ^1_1-AC. We then add arithmetic comprehension axioms (ACA^i) of the form

$$\exists X\ \forall y\ (y \in X \leftrightarrow \varphi(y))$$

and choice axioms (Σ^1_1-AC^i) of the form

$$\forall y\ \exists X\ \varphi(y, X) \to \exists X\ \forall y\ \exists u\ \varphi(y, X_{\langle u,y\rangle})$$

where φ is an arithmetic formula. (Remember that in both of these axioms, as well as the classical versions, we are taking existential quantification to be defined in terms of universal quantification and negation.) We are not claiming that (Σ^1_1-AC^i) is the "correct" intuitionistic translation of (Σ^1_1-AC); only that it is strong enough for our current interpretation and weak enough to be interpreted at the next stage.

The $\cdot^{N_0}$-translation extends to a new translation $\cdot^{N_1}$, obtained by adding the clause

$$(t \in X)^{N_1} = \neg\neg t \in X$$

to cover the new atomic formulas $t \in X$.

Theorem 5.4 *Suppose Σ^1_1-AC proves φ. Then Σ^1_1-AC^i proves φ^{N_1}.*

Proof. As in the case of PA, the interpretation verifies classical logic C_2, where the double-negation in front of formulas $t \in X$ guarantees that the translations of the law of the excluded middle can be derived in I_2. Translations of instances of (ACA) follow from instances of (ACA^i), using the fact that intuitionistic logic proves the equivalence of $\neg\neg\varphi^{N_1}$ (which is equal to $(\neg\neg\varphi)^{N_1}$) and φ^{N_1}.

Finally, over the other axioms of Σ^1_1-AC, instances of (Σ^1_1-AC) are equivalent to (Σ^1_1-AC^i). That is, given an X satisfying the conclusion of (Σ^1_1-AC^i), we can use arithmetic comprehension to define a set X' so that $X'_y = X_{\langle y,u_y\rangle}$ for each y, where u_y is the least value of u satisfying $\varphi(x, X_{\langle y,u\rangle})$; this X' then satisfies the conclusion of (Σ^1_1-AC). But instances of (Σ^1_1-AC^i) translate to instances of (Σ^1_1-AC^i), so we are done. □

In Section 6 we will need to extend the above interpretation to theories with fixed-point axioms. Define the theory $\widehat{ID}^i_1$ to be the theory $\widehat{ID}_1$ based on the intuitionistic fragment I_1, and extend the $\cdot^{N_0}$-translation to the language of this theory by adding the clause

$$(P_\varphi(x))^{N'_1} = P_{\varphi^{N_0}}(x).$$

The $\cdot^{N_1'}$-translation of a fixed-point axiom of $\widehat{ID}_1$ is now of the form

$$\forall x \; (P_{\varphi^{N_0}}(x) \leftrightarrow \varphi^{N_0}(x, P_{\varphi^{N_0}})) \tag{5}$$

which is a fixed-point axiom of $\widehat{ID}_1^i$. We only need to verify that the interpretation of the classical axiom $\neg\neg\psi \to \psi$ holds when ψ is an atomic formula of the form $P_\varphi(t)$; but this follows from (5) and the equivalence of $\neg\neg\varphi^{N_0}$ and φ^{N_0}.

Define $\cdot^{N_0'}$ to be another name for the N_0-translation, and inductively define $\cdot^{N_{n+1}'}$ by adding the clause

$$(P_\varphi(x))^{N_{n+1}'} = P_{\varphi^{N_n'}}(x),$$

to the definition of $\cdot^{N_n'}$. Also, let $\cdot^{N_{n+1}}$ be the extension of $\cdot^{N_n'}$ to second-order logic as above. Then just as in the case of PA and Σ_1^1-AC we can prove

Theorem 5.5 *If $\widehat{ID}_n$ proves a formula φ, then $\widehat{ID}_n^i$ proves $\varphi^{N_n'}$. If Σ_1^1-$AC(\widehat{ID}_n)$ proves a formula ψ, then Σ_1^1-$AC^i(\widehat{ID}_n^i)$ proves $\varphi^{N_{n+1}}$.*

5.3 The theory *Frege*-PA^i

This section introduces a theory called *Frege*-PA^i. Although this system is designed to be a stepping-stone for the interpretation of Σ_1^1-AC^i in P_1, it is possible that the underlying ideas may prove useful in other proof-theoretic contexts as well.

In his work on the foundations of mathematics, Frege identified formulas with mappings from their free variables to a truth values. He then interpreted quantifiers as higher-order mappings, taking formula-mappings to truth values, initiating a process that can be continued to even higher orders as well. We've named the theory *Frege*-PA^i after him because it makes these characterizations explicit: a formula $\varphi(x)$ gives rise to a map $\mathbb{N} \to \mathit{Prop}$, where elements of type Prop are either true or not, and quantifiers are interpreted as maps from $\mathbb{N} \to \mathit{Prop}$ to Prop. Once we've begun treating formulas as functions returning truth values, all of a sudden we will find ourselves with additional means for defining new formulas, using, say, composition and primitive recursion. Such possibilities form the basis of *Frege*-PA^i.[4]

Consider the following example. In the language of arithmetic, we can define a formula $\psi(x)$ that represents the empty set; for example, $0 = 1$ will do. Given a predicate variable X, we can also define a formula $\varphi(x, X)$ specifying what it means to be in the Turing jump of X; namely $\exists z \; T(x, 0, z, X)$, where $T(x, 0, z, X)$ is a form of Kleene's predicate asserting that z codes a halting computation of Turing machine x with oracle X, acting on input 0. Using ψ and φ we can then define succesive elements of the jump hierarchy, defining

$$J_0(x) = \psi(x)$$

[4] The name was suggested by Solomon Feferman, who noted the similarities with Aczel's notion of a "Frege structure" (see [8, 3]). Frege structures may help elucidate the semantics of *Frege*-PA^i; we have not explored this possibility.

and

$$J_{n+1}(x) = \varphi(x, \{x \mid J_n(x)\}/X).$$

If we could somehow define the J predicates *uniformly*, i.e. define a formula $H(n, x)$ equivalent to $J_n(x)$, we could then define the ωth jump of the empty set. Of course, this takes us outside the realm of PA, since $H(n, X)$ would provide us with an adequate truth definition for formulas of PA. But *Frege-*PA^i was designed for just this kind of task. In this theory, defining $H(n, X)$ involves a simple instance of primitive recursion.

Frege-PA^i is built on a finite type structure like that of T. In fact, for each ordinal notation $\alpha < \varepsilon_0$, *Frege-* PA^i allows us to define formulas representing αth Turing-jump of the empty set, just as in T we can define terms representing functions in the $<\varepsilon_0$ fast-growing hierarchy (see [26]).

In *Frege-*PA^i there are four types of judgements:

1. A type, asserting that A denotes a type
2. $a \in A$, asserting that the term a denotes an element of type A
3. $a = b \in A$, asserting that the terms a and b denote equal elements of A
4. For terms a of type $Prop$, the judgement $a\ True$, asserting that the proposition denoted by a is true

The term-forming rules of *Frege-*PA^i are more like those of T than the theories P_n, in that terms are terms and types are types and there is no confusion between the two. Using term-forming operations one can construct formulas (terms of type *Prop*) which are just like the formulas of ordinary PA^i, only more complex. And, as in ordinary PA^i, there are rules by which certain well-formed formulas can be shown to be true on the basis of the axioms.

Types of Frege-PA^i

1. Natural numbers: $\mathbb{N}$ type
2. Cross product: $\dfrac{X \text{ type} \quad Y \text{ type}}{X \times Y \text{ type}}$
3. Functions: $\dfrac{X \text{ type} \quad Y \text{ type}}{X \to Y \text{ type}}$
4. Propositions: *Prop* type

Terms of Frege-PA^i

1. Variables: $x \in A$ $(x \in A)$
2. Zero: $0 \in \mathbb{N}$

3. Successor: $S \in \mathbb{N} \to \mathbb{N}$

4. Explicit definition: $\dfrac{\begin{matrix}(x \in A)\\ b[x] \in B\end{matrix}}{\lambda x.b \in A \to B}$

5. Application: $\dfrac{a \in A \quad b \in A \to B}{b(a) \in B}$

6. Pairing: $\dfrac{a \in A \quad b \in B}{\langle a, b\rangle \in A \times B}$

7. Projection: $\dfrac{c \in A \times B}{\pi_0(c) \in A}$, $\dfrac{c \in A \times B}{\pi_1(c) \in B}$

8. Primitive recursion: $\dfrac{a \in A \quad b \in \mathbb{N} \to (A \to A)}{R_{a,b} \in \mathbb{N} \to A}$

9. Equality: $Equals \in \mathbb{N} \times \mathbb{N} \to Prop$

10. Implication: $Implies \in Prop \times Prop \to Prop$

11. Conjunction: $And \in Prop \times Prop \to Prop$

12. Universal Quantifier: $Forall \in (\mathbb{N} \to Prop) \to Prop$

13. Falsity: $False \in Prop$

The first few type- and term-building rules are just those of T. The meaning of the last few functionals on the previous list should be intuitively clear: $Equals$ takes two terms a and b and returns the proposition $a = b$, $Implies$ takes two propositions φ and ψ and returns the proposition $\varphi \to \psi$, and so on. The functional $Forall$ deserves further comment. If $\varphi(x)$ is an element of $Prop$ with free variable $x \in \mathbb{N}$, then the term $\lambda x.\varphi(x)$ is of type $\mathbb{N} \to Prop$, and the term $Forall(\lambda x.\varphi(x))$ is the element of $Prop$ that denotes, intuitively, the proposition $\forall x\ \varphi(x)$. In order to have our notation agree with more common usage, below we will write $Forall\ x\ \varphi(x)$ for $Forall(\lambda x.\varphi(x))$, $\varphi\ Implies\ \psi$ for $Implies(\varphi, \psi)$, and so on. We will also omit parentheses according to the usual conventions of predicate logic.

The rules for determining that two terms of *Frege-*PA^i are equal are no different from those of T. Note that the equality symbol of *Frege-*PA^i has nothing per se to do with the $Equals$ functional, though rules given later on will allow us to derive the assertion $Equals(a, b)\ True$ from $a = b$, for any terms a and b of type $\mathbb{N}$.

Equality of terms in Frege-PA^i

1. Reflexivity: $\dfrac{a \in A}{a = a \in A}$

2. Symmetry: $\dfrac{a = b \in A}{b = a \in A}$

3. Transitivity: $\dfrac{a = b \in A \quad b = c \in A}{a = c \in A}$

4. Explicit definition: $(\lambda x.a)(b) = a[b/x]$

5. Projection: $\pi_0(\langle a, b\rangle) = a$, $\pi_1(\langle a, b\rangle) = b$

6. Primitive recursion: $R_{ab}(0) = a$, $R_{ab}(x') = b(x, R_{ab}(x))$ $(x \in \mathbb{N})$

7. Substitution: $\dfrac{a = b \in A \qquad \begin{matrix}(x \in A)\\ c[x] \in C\end{matrix}}{c[a/x] = c[b/x] \in C}$

8. Instantiation of variables: $\dfrac{a \in A \qquad \begin{matrix}(x \in A)\\ c[x] = d[x] \in C\end{matrix}}{c[a/x] = d[a/x] \in C}$

The equality rules just presented have little to say about the logical functions *Implies*, *Forall*, and so on. The rules we've seen so far give us mechanisms to define complex elements of the type *Prop*, as well as ways of showing that certain elements of *Prop* are in a sense the same.

There is a natural map taking terms t in the language of arithmetic to terms $\widehat{t}$ of type $\mathbb{N}$ in *Frege-*PA^i. Similarly, we can map formulas φ in the language of arithmetic to elements $\widehat{\varphi}$ of type *Prop* in the language of *Frege-*PA^i, by taking atomic formulas $t_1 = t_2$ to $\widehat{t_1}$ *Equals* $\widehat{t_2}$, formulas $\varphi \to \psi$ to $\widehat{\varphi}$ *Implies* $\widehat{\psi}$, and so on. If $\Phi(\varphi_1, \ldots, \varphi_n)$ is a schema involving the formulas $\varphi_1(x_1, \ldots, x_k)$ to $\varphi_n(x_1, \ldots, x_k)$, we we define $\widehat{\Phi}$ by replacing each φ_i with a corresponding variable of type $\mathbb{N}^k \to Prop$; an instance of $\widehat{\Phi}$ is obtained by substituting appropriate terms for the variables. Similarly, an instance of a rule

$$\frac{\widehat{\Phi}_0\ True, \ldots, \widehat{\Phi}_m\ True}{\widehat{\Phi}\ True}$$

is also obtained by replacing variables by specific terms. We can now present the rules concerning

Truth of propositions in Frege-PA^i

1. If Φ an axiom or schema PA^i, then any instance of $\widehat{\Phi}\ True$

2. If $\Phi_0, \ldots, \Phi_m \Rightarrow \Phi$ is a rule of PA^i, then any instance of

$$\frac{\widehat{\Phi}_0\ True, \ldots, \widehat{\Phi}_m\ True}{\widehat{\Phi}\ True}$$

3. Substitution of equal terms: $\dfrac{a = b \in A \quad c[a/x]\ True}{c[b/x]\ True}$

4. Instantiation of variables: $\dfrac{a \in A \qquad \overset{(x \in A)}{b[x]\ True}}{b[a/x]\ True}$

More explicitly, clauses 1 and 2 include the propositional schemata, schemata involving quantifiers, equality axioms, the defining equations for succesor, plus, and times, and finally the induction rule. It should be clear that *Frege-*PA^i is an extension of PA^i. However, since *Frege-*PA^i is built on intuitionistic logic I_1, the law of the excluded middle $\neg\neg\varphi \to \varphi$ does *not* necessarily hold for arbitrary $\varphi \in Prop$.

Note that we have

$$x\ \mathit{Iff}\ x\ True\ (x \in Prop),$$

so that if *Frege-*PA^i proves that two terms a and b of type $Prop$ are equal, then by substitution it also proves $a\ \mathit{Iff}\ b\ True$.

5.4 Interpreting Σ^1_1-AC^i in *Frege-*PA^i

In this section we describe a Dialectica-style interpretation of Σ^1_1-AC^i in *Frege-*PA^i, in which the number variables of Σ^1_1-AC^i are mapped to number variables of *Frege-*PA^i and the set variables of Σ^1_1-AC^i are mapped to variables of type $\mathbb{N} \to Prop$. As in the Dialectica interpretation we associate to every formula φ of Σ^1_1-AC^i a formula φ^{F_1} of the form

$$\exists x \in A\ \forall y \in B\ \varphi_{F_1}(x, y)\ True \tag{6}$$

where φ_{F_1} is a term of *Frege-*PA^i of type $A \times B \to Prop$. Once again, we allow for the possibility that either quantifier is absent. We'll then show, as in Section 4.2, that from a proof of φ in Σ^1_1-AC we can extract a term witnessing the existential quantifier of (6). For arithmetic φ, φ^{F_1} will be just $\widehat{\varphi}\ True$, so that arithmetic formulas become "quantifier-free" in the translation.

The $\cdot^{F_1}$ mapping is very similar to the $\cdot^{D_1}$ mapping. One key difference is in the clause for implication. In Section 4.2 verifying the interpretation of the rule $\varphi \to \psi, \varphi \to \theta \Rightarrow \varphi \to \psi \wedge \theta$ required a "definition by cases" dependent on the truth value of a quantifier-free formula. In *Frege-*PA^i the "quantifier-free" formulas are arithmetic, and the theory does not provide the means to define a function dependent on the truth of a term in $Prop$. As a result, we have to use a trick due to Diller and Nahm [11] which avoids the need for such functions.

We start by defining $(t_1 = t_2)^{F_1}$ to be

$$\widehat{t_1}\ Equals\ \widehat{t_2}\ True$$

and once variables $\widehat{X} \in \mathbb{N} \to Prop$ are assigned to set variables X of Σ^1_1-AC, we can define $(t \in X)^{F_1}$ to be

$$\widehat{X}(\widehat{t})\ True.$$

Now assuming φ^{F_1} is given by

$$\exists x \in A\, \forall y \in B\ \varphi_{F_1}(x,y)\ True$$

and ψ^{F_1} is given by

$$\exists w \in C\, \forall z \in D\ \psi_{F_1}(w,z)\ True$$

we define $(\varphi \to \psi)^{F_1}$ by

$$\begin{array}{lll} \varphi \to \psi & \rightsquigarrow & \exists x\, \forall y\ \varphi_{F_1}(x,y)\ True \to \exists w\, \forall z\ \psi_{F_1}(w,z)\ True \\ & \rightsquigarrow & \exists x\, \forall y\ Forall\ u\ \varphi_{F_1}(x,y(u))\ True \to \exists w\, \forall z\ \psi_{F_1}(w,z)\ True \\ & \rightsquigarrow & \forall x\, \exists w\ (\forall y\ Forall\ u\ \varphi_{F_1}(x,y(u))\ True \to \forall z\ \psi_{F_1}(w,z)\ True) \\ & \rightsquigarrow & \forall x\, \exists w\, \forall z\, \exists y\ (Forall\ u\ \varphi_{F_1}(x,y(u))\ Implies\ \psi_{F_1}(w,z))\ True \\ & \rightsquigarrow & \exists y, w\, \forall x, z\ (Forall\ u\ \varphi_{F_1}(x,y(u,x,z))\ Implies\ \psi_{F_1}(w(x),z))\ True \end{array}$$

Note that in the second line we replace

$$\exists x\, \forall y\ \varphi_{F_1}(x,y)\ True$$

by

$$\exists x\, \forall y\ Forall\ u\ \varphi_{F_1}(x,y(u))\ True.$$

Intuitively, we're replacing elements y by sequences of elements $\lambda u.y(u)$. (In the Diller-Nahm interpretation one uses *finite* sequences of elements, which would work here also; for our purposes, using infinite sequences is sufficient and simplifies the notation.)

The interpretations of the other propositional connectives remain the same: $(\varphi \wedge \psi)^{F_1}$ is given by

$$\exists x, w\, \forall y, z\ (\varphi_{F_1}(x,y)\ And\ \psi_{F_1}(w,z))\ True,$$

and $\bot^{F_1}$ is given by $False\ True$. The translation of second-order quantifiers is also as before: $\forall Z\ \varphi(Z)$ is translated to

$$\begin{array}{lll} \forall Z\ \varphi(Z) & \rightsquigarrow & \forall z\, \exists x\, \forall y\ \varphi_{F_1}(z,x,y)\ True \\ & \rightsquigarrow & \exists x\, \forall z, y\ \varphi_{F_1}(z,x(z),y)\ True \end{array}$$

where z is of type $\mathbb{N} \to Prop$. More interesting is the interpretation of a first-order quantifier: $\forall z\ \varphi(z)$ is translated to

$$\begin{array}{lll} \forall z\ \varphi(z) & \rightsquigarrow & \forall z\, \exists x\, \forall y\ \varphi_{F_1}(z,x,y)\ True \\ & \rightsquigarrow & \exists x\, \forall y\ Forall\ z\ \varphi_{F_1}(z,x(z),y)\ True. \end{array}$$

In other words, the first-order universal quantifier is absorbed by the "quantifier-free" part of φ^{F_1}.

Theorem 5.6 *Suppose Σ^1_1-AC^i proves φ^i. Then* Frege-PA^i *proves φ^{F_1}. In other words, if φ^{F_1} is given by*

$$\exists x \in A\ \forall y \in B\ \varphi_{F_1}(x, y)\ True$$

then there is a term $a \in A$ whose free variables correspond to those of φ such that Frege-PA^i *proves*

$$\varphi_{F_1}(a, y)\ True\ (y \in B).$$

The verification of first-order axioms is routine, with induction and axioms regarding implication handled as in [11]. $(ACA^i)^{F_1}$ is easily taken care of: axioms (ACA^i) are of the form

$$\neg\forall X\ \neg\forall y\ (y \in X \leftrightarrow \varphi(y))$$

and their $\cdot^{F_1}$-translations are of the form

$$\exists x\ Not\ Forall\ u\ Not\ Forall\ y\ (x(u, y)\ \mathit{Iff}\ \ \varphi_{F_1}(y))\ True,$$

where x is of type $\mathbb{N} \to (\mathbb{N} \to Prop)$; note that the "$Forall\ u$" arises from the Diller-Nahm interpretation of a negation.[5] We need a term a to witness the existential quantifier; taking $a = \lambda u \lambda y.\varphi_{F_1}(y)$ suffices.

Axioms (Σ^1_1-AC^i) are of the form

$$\forall y\ \neg\forall X\ \neg\varphi(y, X)\ \to \neg\forall X\ \neg\forall y\ \neg\forall u\ \neg\varphi(y, X_{\langle u,y\rangle}).$$

The left-hand side translates to

$$\exists x\ Forall\ y\ Not\ Forall\ u\ Not\ \varphi_{F_1}(y, x(u, y))\ True$$

while the right-hand side translates to

$$\exists x'\ Not\ Forall\ v\ Not\ Forall\ y\ Not\ Forall\ u\ Not\ \varphi_{F_1}(y, x'(v)_{\langle u,y\rangle})\ True.$$

Given an a witnessing the left-hand side, we need to define an a' witnessing the right. Defining $a' = \lambda v \lambda w.a(w_{00}, w_{01}, w_1)$ yields $a'(v, \langle\langle u, y\rangle, t\rangle) = a(u, y, t)$ for any term t, and hence

$$a'(v)_{\langle u,y\rangle}(t)\ \mathit{Iff}\ \ a(u, y, t)\ True.$$

We can use this to show

$$\varphi_{F_1}(y, a(u, y))\ \mathit{Iff}\ \varphi_{F_1}(y, a'(v)_{\langle u,y\rangle})\ True,$$

which yields the necessary implication.

For arithmetic φ that don't involve set variables φ_{F_1} is just $\widehat{\varphi}$. As a result, we have

[5] This extra quantifier was omitted in [5]. We are grateful to Justus Diller for pointing out this error.

Corollary 5.7 *If Σ^1_1-AC^i proves φ, where φ is arithmetic, then* Frege-PA^i *proves $\widehat{\varphi}$ True. In particular, if Σ^1_1-AC^i proves $\perp$ then* Frege-PA^i *proves $False\ True$.*

As in Section 4.2, though we initially allowed for the possibility that either quantifier is absent in the definition of φ^{F_1}, we now want to rule out this possibility. We can bring this about by translating the atomic formula $t_1 = t_2$ to

$$\exists u \in N\ \forall v \in N\ \widehat{t_1}\ Equals\ \widehat{t_2}\ True$$

where u and v don't appear in $\widehat{t_1}$ or $\widehat{t_2}$, and similarly for atomic formas $X(t)$. This has the net effect that arithmetic formulas φ are translated to formulas $\exists u \in A\ \forall v \in B\ \bar{\varphi}\ True$ where u and v don't appear in $\bar{\varphi}$. Once again we can show that this doesn't harm Theorem 5.6 or Corollary 5.7, and so from now on we adopt this modified translation as $\cdot^{F_1}$.

5.5 Interpreting *Frege-*PA^i in P_1

Frege-PA^i looks a lot like T, except that formulas of PA^i are "reflected down" to quantifier-free formulas. Similarly, P_1 looks a lot like T, except that types of P_0 are "reflected down" to the universe U_0. Since PA^i is interpreted in P_0, it might seem plausible that *Frege-*PA^i can be interpreted in P_1. In this section we will show that this is indeed the case.

To interpret *Frege-*PA^i in P_1 we will translate types A of the former system to types A^{D_1} of the latter, and terms $a \in A$ to terms $a^{D_1} \in A^{D_1}$. To motivate the definition of $Prop^{D_1}$, remember that in the *Dialectica* interpretation formulas φ of PA^i are mapped to assertions φ^{D_0} of the form

$$\exists x \in A\ \forall y \in B\ f(x, y) = 0 \tag{7}$$

where A and B are types of P_0 and f is of a term of type $A \times B \to \mathbb{N}$. In *Frege-*$PA^i$, terms of type $Prop$ are represent generalized formulas of PA^i, so it makes sense to define

$$Prop^{D_1} = \Sigma_{X \in U_0, Y \in U_0} f \in X \times Y \to \mathbb{N}$$

of P_1. We can now interpret terms $\varphi \in Prop$ of *Frege-*PA^i by elements $\langle A, B, f\rangle$ of $Prop^{D_1}$, in such a way that if *Frege-*PA^i proves $\varphi\ True$ then P_1 proves (7); that is, there is a term $a \in A$ such that P_1 proves

$$f(a, y) = 0\ (y \in B).$$

This is done by "internalizing" the *Dialectica* interpretation, so that the propositional maps $Implies$, And, and $Forall$ of *Frege-*PA^i translate to the corresponding operations on elements of $Prop$.

A small technical complication arises with this approach. The verification of the rule $\varphi \Rightarrow \psi \to \varphi$ in Section 4 relied on the existence of "canonical elements" of each type, given by Lemma 4.5. Unfortunately, we cannot generate such elements uniformly; that is, there is no functional $Canon \in \Pi_{X \in U_0} X$ that provides an element $Canon(A) \in A$ for each type A of U_0. However, we *can* pick a canonical element of type $\mathbb{N}$ (e.g. 0), and given canonical elements for the types appearing in φ^{D_0} and ψ^{D_0} we can get canonical elements for the types appearing in $(\varphi \rightarrow \psi)^{D_0}$, $(\varphi \wedge \psi)^{D_0}$, and so on. So the solution is to leave room for these elements in $Prop^{D_1}$. In short, we set

$$Prop^{D_1} = \Sigma_{X \in U_0, Y \in U_0} (X \times Y \rightarrow \mathbb{N}) \times X \times Y$$

so that elements $\langle A, B, f, canon_A, canon_B \rangle$ of $Prop^{D_1}$ include the canonical constants as well as A, B, and f as above.

We now define $\mathbb{N}^{D_1}$ to be $\mathbb{N}$, $(A \times B)^{D_1}$ be to be given inductively by $A^{D_1} \times B^{D_1}$, and $(A \rightarrow B)^{D_1}$ to be given by $A^{D_1} \rightarrow B^{D_1}$. Similarly, to define the mapping on terms of *Frege-PA^i*, take $0^{D_1} = 0$, $S^{D_1} = S$, map variables $x \in A$ of *Frege- PA^i* to variables $x^{D_1} \in A^{D_1}$ in P_1, and let the map $\cdot^{D_1}$ commute with explicit definition, application, pairing, projection, and primitive recursion.

All that is left to do is define the action of $\cdot^{D_1}$ on the *Prop*-valued functions $Implies$, And, $False$, $Equals$, and $Forall$. To define

$$Implies^{D_1} \in Prop^{D_1} \times Prop^{D_1} \rightarrow Prop^{D_1},$$

recall that if φ^{D_0} is given by

$$\exists x \in A \; \forall y \in B \; f(x, y) = 0$$

and ψ^{D_0} is given by

$$\exists w \in C \; \forall z \in D \; g(w, z) = 0$$

then $(\varphi \rightarrow \psi)^{D_0}$ is given by

$$\begin{aligned} &\exists u \in (A \rightarrow C) \times (A \times D \rightarrow B) \; \forall v \in A \times D \\ &\qquad (\lambda u, v. t_{implies(f(v_0, u_0(v_0, v_1)), g(u_1(v_0), v_1)))})(u, v) = 0. \end{aligned}$$

We can then define

$$\begin{aligned} &Implies^{D_1}(\langle A, B, f, canon_A, canon_B \rangle, \langle C, D, g, canon_C, canon_D \rangle) = \\ &\qquad \langle (A \rightarrow C) \times (A \times D \rightarrow B), A \times D, \\ &\qquad \lambda u, v. t_{implies(f(v_0, u_0(v_0, v_1)), g(u_1(v_0), v_1))}, \\ &\qquad \langle \lambda x. canon_C, \lambda x. canon_B \rangle, \langle canon_A, canon_D \rangle \rangle. \end{aligned}$$

Similarly, we define

$$\begin{aligned} &And^{D_1}(\langle A, B, f, canon_A, canon_B \rangle, \langle C, D, g, canon_C, canon_D \rangle) = \\ &\qquad \langle A \times C, B \times D, \\ &\qquad \lambda u, v. t_{and(f(u_0, v_0), g(u_1, v_1))}, \\ &\qquad \langle canon_A, canon_C \rangle, \langle canon_B, canon_D \rangle \rangle; \end{aligned}$$

also

$$Equals^{D_1}(x,y) = \langle \mathbb{N}, \mathbb{N}, \lambda u, v.t_{equals(x,y)}, 0, 0\rangle$$

and

$$False^{D_1} = \langle \mathbb{N}, \mathbb{N}, \lambda u, v.1, 0, 0\rangle.$$

The definition of $Forall^{D_1}$ involves a slight twist. Recall that if $\varphi(z)^{D_1}$ is given by

$$\exists x \in A\ \forall y \in B\ f(z,x,y) = 0$$

then $(\forall z\ \varphi(z))^{D_1}$ is given by

$$\exists x \in \mathbb{N} \to A\ \forall \langle z, y\rangle \in \mathbb{N} \times B\ f(z, x(z), y) = 0.$$

Now we must consider the possibility that A and B also depend on z, so that in general $\varphi(z)^{D_1}$ is given by

$$\exists x \in A(z)\ \forall y \in B(z)\ f(z,x,y) = 0.$$

Instead of $\mathbb{N} \to A$ and $\mathbb{N} \times B$ we now use the dependent product and sum types, and define $(\forall z\ \varphi(z))^{D_1}$ by

$$\exists x \in \Pi_{z\in\mathbb{N}} A(z)\ \forall \langle z, y\rangle \in \Sigma_{z\in\mathbb{N}} B(z)\ f(z, x(z), y) = 0.$$

In short, we define

$$Forall^{D_1} \in (\mathbb{N} \to Prop^{D_1}) \to Prop^{D_1}$$

by

$$Forall^{D_1}(d) = \langle \Pi_{z\in N}.d(z)_0, \Sigma_{z\in N}.d(z)_1,$$
$$\lambda u, v.d(z)_2(v_0, u(v_0), v_1), \lambda z.(d(z)_3), \langle 0, d(0)_4\rangle\rangle.$$

This completes the definition of $\cdot^{D_1}$.

Theorem 5.8 *The mapping $\cdot^{D_1}$ has the following properties:*

1. *If A is a type of* Frege-PA^i*, then A^{D_1} is a type of P_1.*
2. *If* Frege-PA^i *proves $a \in A$ then P_1 proves $a^{D_1} \in A^{D_1}$.*
3. *If* Frege-PA^i *proves $a = b \in A$ then P_1 proves $a^{D_1} = b^{D_1} \in A^{D_1}$.*
4. *Let $d \in Prop$ in* Frege-PA^i*, and suppose* Frege-PA^i *proves $d\,True$. Then P_1 proves*

   $$\exists x \in d_0^{D_1}\ \forall y \in d_1^{D_1}\ d_2^{D_1}(x,y) = 0.$$

   In other words, there is a term $t \in d_0^{D_1}$ whose free variables correspond to those of d such that P_1 proves

   $$d_2^{D_1}(t,y) = 0\ (y \in d_1^{D_1}).$$

The details of the proof are routine, and can be found in [5]. For arithmetic φ, $\widehat{\varphi}^{D_1}$ is essentially φ^{D_0}. As a result, we have

Corollary 5.9 *Suppose φ is an arithmetic formula in the language of arithmetic such that* Frege-PA^i *proves $\widehat{\varphi}$ $True$. Then P_1 proves φ^{D_0}. In particular, if φ is the Π^0_2 assertion $\forall x\, \exists y\, \psi(x,y)$ where ψ is quantifier-free then there is a term t such that P_1 proves*

$$\psi'(x, t(x))\ (x \in \mathbb{N})$$

and if Frege-PA^i *proves $False\ True$ then P_1 proves $1 = 0$.*

5.6 Putting it all together

Now let's consider what happens when the $\cdot^{\Sigma_1}$, $\cdot^{N_1}$, $\cdot^{F_1}$, and $\cdot^{D_1}$ mappings are composed. The $\cdot^{\Sigma_1}$ mapping takes formulas φ in the language of $\widehat{ID}_1$ to formulas φ^{Σ_1} in the language of second-order arithmetic, and $\cdot^{N_1}$ adds double-negations before atomic formulas of the form $t \in X$. $\varphi^{\Sigma_1 N_1 F_1}$ is then a formula of the form

$$\exists x \in A\ \forall y \in B\ f(x,y)\ True, \tag{8}$$

where A and B are types of *Frege-PA^i* and F is a term of *Frege-PA^i* of type $A \times B \to Prop$. The $\cdot^{D_1}$ mapping is only defined on types and terms of *Frege- PA^i* but we can extend it in the obvious way to map the formula (8) to $\varphi^{\Sigma_1 N_1 F_1 P_1}$, given by

$$\exists x \in A^{D_1}\ \forall y \in B^{D_1}\ \exists w \in (f^{D_1}(x,y))_0\ \forall z \in (f^{D_1}(x,y))_1\ (f^{D_1}(x,y))_2(w,z) = 0.$$

Now, combining Theorems 5.2, 5.6, and 5.8, we have the following

Theorem 5.10 *Suppose $\widehat{ID}_1$ proves φ. Then P_1 proves $\varphi^{\Sigma_1 N_1 F_1 P_1}$. In other words, if $\varphi^{\Sigma_1 N_1 F_1 P_1}$ is given by*

$$\exists x \in A\ \forall y \in B\ \exists w \in (f(x,y))_0\ \forall z \in (f(x,y))_1\ (f(x,y))_2(w,z) = 0$$

then there are terms $t \in A$ and $s[y] \in (f(t,y))_0\ (y \in B)$ such that P_1 proves

$$(f(t,y))_2(s,z) = 0\ (y \in B, z \in (f(t,y))_1).$$

Since the $\cdot^{\Sigma_1 N_1 F_1 P_1}$ mapping agrees with the $\cdot^{N_0 D_0}$ mapping on arithmetic formulas, we get the following

Corollary 5.11 *Suppose $\widehat{ID}_1$ proves φ, where φ is does not involve any fixed-point constants. Then P_1 proves $\varphi^{N_0 D_0}$. In particular, if $\widehat{ID}_1$ proves $\forall x\, \exists y\, \varphi(x,y)$ where φ is quantifier-free then there is a term t such that P_1 proves*

$$\varphi'(x, t(x)) = 0\ (x \in N)$$

and if $\widehat{ID}_1$ proves $\perp$ then P_1 proves $1 = 0$.

6 Iterating the Interpretation

In Section 4 we showed how to interpret PA in P_0. In Section 5 we saw that this interpretation was "uniform" enough to allow for the interpretation of *Frege-*PA^i in P_1 and hence, ultimately, the interpretation of $\widehat{ID}_1$. This interpretation can in turn be lifted to an interpretation of *Frege-*$\widehat{ID}_1^i$ in P_2, yielding an interpretation of $\widehat{ID}_2$, and so on up the hierarchy.

The interpretation of PA in P_0 ran

$$\begin{array}{lll} PA & \rightsquigarrow^{N_0} & PA^i \\ & \rightsquigarrow^{D_0} & P_0. \end{array}$$

and the interpretation of $\widehat{ID}_1$ in P_1 ran

$$\begin{array}{lll} \widehat{ID}_1 & \rightsquigarrow^{\Sigma_1} & \Sigma^1_1\text{-}AC \\ & \rightsquigarrow^{N_1} & \Sigma^1_1\text{-}AC^i \\ & \rightsquigarrow^{F_1} & \textit{Frege-}PA^i \\ & \rightsquigarrow^{D_1} & P_1, \end{array}$$

where the last step relied on a uniform version of the interpretation of PA^i in P_0. Turning to $\widehat{ID}_2$, we'd like to interpret

$$\begin{array}{lll} \widehat{ID}_2 & \rightsquigarrow^{\Sigma_2} & \Sigma^1_1\text{-}AC(\widehat{ID}_1) \\ & \rightsquigarrow^{N_2} & \Sigma^1_1\text{-}AC^i(\widehat{ID}_1^i) \\ & \rightsquigarrow^{F_2} & \textit{Frege-}\widehat{ID}_1^i \\ & \rightsquigarrow^{D_2} & P_2, \end{array}$$

where the last step similarly makes use of a uniform version of the interpretation of $\widehat{ID}_1^i$ in P_1.

The theory Σ^1_1-$AC(\widehat{ID}_1)$ and the map $\cdot^{\Sigma_2}$ were defined in Section 5.1, and Σ^1_1-$AC^i(\widehat{ID}_1^i)$ and the $\cdot^{N_2}$ mapping were defined in Section 5.2. We now define *Frege-*$\widehat{ID}_1^i$ to consist of *Frege-*PA^i together with additional terms

$$Fixed_\varphi \in \mathbb{N} \to Prop$$

for each positive arithmetic φ, and axioms

$$Forall\ z\ (Fixed_\varphi(z)\ \mathit{Iff}\ \widehat{\varphi}(z, \lambda z.Fixed_\varphi(z)))\ True.$$

Just as in Section 5.4 we can define an interpretation $\cdot^{F_2}$ of Σ^1_1-$AC(\widehat{ID}_1^i)$ in *Frege-*$\widehat{ID}_1^i$. To finish off the interpretation of $\widehat{ID}_2$, we only need to define a map $\cdot^{D_2}$ which will interpret *Frege-*$\widehat{ID}_1^i$ in P_2.

The idea is as follows. In interpreting *Frege-PA^i* in P_1, we used the interpretation of PA^i in P_0 to define functions $Implies^{D_1}$, And^{D_1}, $False^{D_1}$, $Equals^{D_1}$, and $Forall^{D_1}$, returning values in

$$Prop^{D_1} = \Sigma_{X \in U_0, Y \in U_0}(X \times Y \to N) \times X \times Y.$$

This enabled us to translate terms $t \in Prop$ of *Frege-PA^i* to terms $t^{D_1} \in Prop^{D_1}$ of P_1 in such a way that axioms and rules of PA were preserved. We can repeat this maneuver to interpret *Frege-$\widehat{ID}_1^i$* in P_2. According to Theorem 5.10, the interpretation of $\widehat{ID}_1$ takes formulas φ of that theory to formulas $\varphi^{\Sigma_1 N_1 F_1 P_1}$ of the form

$$\exists x \in A \, \forall y \in B \, \exists w \in (f(x,y))_0 \, \forall z \in (f(x,y))_1 \, (f(x,y))_2(w,z) = 0 \qquad (9)$$

where A and B are types of P_1 and $f \in A \times B \to Prop^{D_1}$. Since the type building operations of P_1 can be mirrored by operations on the universe U_1 of P_2, we can define

$$Prop^{D_2} = \Sigma_{X \in U_1, Y \in U_1}(X \times Y \to Prop^{D_1}) \times X \times Y,$$

so that intuitively an element $\langle A, B, f, canon_A, canon_B \rangle$ of $Prop^{D_2}$ corresponds to the formula given by (9). We can then define functions $Implies^{D_2}$, And^{D_2}, $False^{D_2}$, $Equals^{D_2}$, $Forall^{D_2}$, and $Fixed_\varphi^{D_2}$ returning elements in $Prop^{D_2}$ corresponding to the interpretation of $\widehat{ID}_1$ in P_1. With these maps we can identify element $t \in Prop$ of *Frege-$\widehat{ID}_1^i$* with elements $t^{D_2} \in Prop^{D_2}$ of P_2, such that the axioms and rules of $\widehat{ID}_1^i$ are verified in the translation.

Iterating this idea gives the main result.

Theorem 6.1 *For each n $\widehat{ID}_n$ is interpreted in P_n.*

Here "interpreted" has to be construed in the appropriate sense. Since all the interpretations are essentially the *Dialectica* interpretation when restricted to on arithmetic formulas, we have

Corollary 6.2 *Suppose $\widehat{ID}_n$ proves a formula φ involving none of the fixed-point constants. Then P_n proves $\varphi^{N_0 D_0}$. In particular, if $\widehat{ID}_n$ proves $\forall x \, \exists y \, \psi(x,y)$ where ψ is quantifier free, there is a term t such that P_n proves*

$$\psi'(x, t(x)) = 0 \; (x \in \mathbb{N}),$$

and if $\widehat{ID}_n$ proves $\perp$ then P_n proves $1 = 0$.

But this is just our main result, Theorem 1.1.

7 Final Comments

When we began this project we were looking for a functional interpretation of ATR_0 in a theory like $P_{<\omega}$. As it turns out, the fact that $P_{<\omega}$ is "stratified" whereas ATR_0 is not, together with the speedup result given by Theorem 1.2, makes a direct interpretation unlikely. This, however, suggests an interesting question: is there a "second-order" version of $P_{<\omega}$ in which ATR_0 can be interpreted directly?

Can one prove that a reduction procedure for closed terms of P_n is strongly normalizing? If this could be carried out via an assignment of ordinals to terms as in [31, 20], the result, combined with the analysis in this paper, would provide an alternate route to the ordinal analysis of $\widehat{ID}_{<\omega}$.

We'd like to see this work extended to stronger systems. Proof-theoretic investigations by Spector [30] and Girard [17] have shown that bar recursion and an impredicative form of polymorphism suffice to interpret full second-order arithmetic. Can fragments and variants of these schemata be used to interpret other classical theories? We feel that such interpretations serve to illuminate the classical strength of computational schemata, as well as the "constructive content" of classical reasoning.

I'd like to thank Solomon Feferman for numerous discussions and suggestions regarding the work here, and my advisor, Jack Silver for his support. Finally, I'd like to thank the Sloan Foundation for supplying me with a dissertation-year fellowship while I carried out this work.

References

[1] Aczel, Peter, "An introduction to inductive definitions," in [7], pp. 739-782.

[2] Aczel, Peter, "The Type Theoretic Interpretation of Constructive Set Theory," in A. Macintyre et al. eds., *Logic Colloqium '77*, North Holland, 1978.

[3] Aczel, Peter, "Frege structures and the notion of proposition, truth, and set," in Jon Barwise et al. eds., *The Kleene Symposium,* North-Holland, 1980.

[4] Aczel, Peter, "The Type Theoretic Interpretation of Constructive Set Theory: Choice Principles," in A. S. Troelstra and D. van Dalen eds., *The L.E.J. Brouwer Centenary Symposium*, North Holland, 1982.

[5] Avigad, Jeremy, *Proof-Theoretic Investigations of Subsystems of Second-Order Arithmetic*, Ph.D. dissertation, University of California, Berkeley, 1995.

[6] Avigad, Jeremy, "On the Relationship Between ATR_0 and $\widehat{ID}_{<\omega}$," *Journal of Symbolic Logic* 61 (1996), 768-779.

[7] Barwise, Jon, ed., *The Handbook of Mathematical Logic,* North-Holland, 1977.

[8] Beeson, Michael J., *Foundations of Constructive Mathematics*, Springer, 1985.

[9] Beeson, Michael J., "Recursive Models for Constructive Set Theories," *Annals of Mathematical Logic* 23 (1982), 127-178.

[10] Buchholz, Wilfried, et. al., *Iterated Inductive Definitions and Subsystems of Analysis: Recent Proof-Theoretical Studies*, Lecture Notes in Mathematics, vol. 897, Springer, 1981.

[11] Diller, J. and W. Nahm, "Eine Variante zur Dialectica Interpretation der Heyting-Arithmetik endlicher Typen," Archive Mathematische Logik Grundlagenforschung 16 (1974), 49-66.

[12] Diller, J. and A.S. Troelstra, "Realizability and Intuitionistic Logic," Synthese 60 (1984), 253-282.

[13] Feferman, Solomon, "Theories of Finite Type Related to Mathematical Practice," in [7], pp. 913-972.

[14] Feferman, Solomon et al., eds., *Kurt Gödel: Collected Works,* vol. 2, Oxford University Press, 1990.

[15] Feferman, Solomon, "Iterated Inductive Fixed-Point Theories: Application to Hancock's Conjecture", in G. Metakides ed. *Patras Logic Symposium*, North-Holland, 1982.

[16] Gallier, J. H., "What's So Special About Kruskal's Theorem and the Ordinal Γ_0? A Survey of Some Results in Proof Theory," *Annals of Pure and Applied Logic* 53 (1991), 199-260.

[17] Girard, Jean-Yves, Yves Lafont, and Paul Taylor, *Proofs and Types*, Cambridge University Press, 1989.

[18] Gödel, Kurt, "Uber ein bisher noch nicht benutzte Erweiterung des finiten Standpunktes," Dialectica 12 (1958), 280-287; also appears, with English translation, in [14].

[19] Howard, W. A., "The Formulae-As-Types Notion of Construction," in J.P. Seldin and J.R. Hindley, eds., *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism,* Academic Press, 1980.

[20] Howard, W. A., "Assignment of ordinals to terms for primitive recursive functionals of finite type," in J. Myhill et al. eds., *Intuitionism and Proof Theory*, North-Holland, 1970.

[21] Kreisel, G., "On the Interpretation of Non-Finitist Proofs," I and II, Journal of Symbolic Logic 16 (1951) 241-267 and 17 (1952) 43-58.

[22] Martin-Löf, Per, "An Intuitionistic Theory of Types: Predicative Part," in H.E. Rose and J.C. Shepherdson, eds., *Logic Colloqium*, North-Holland, 1975.

[23] Martin-Löf, Per, "Constructive Mathematics and Computer Programming," in C.A.R. Hoare and J.C. Shepherdson, eds., *Mathematical Logic and Programming Languages,* Prentice-Hall, 1985.

[24] Martin-Löf, Per, *Intuitionistic Type Theory,* Bibliopolis, 1984.

[25] Palmgren, Erik, "Type-theoretic Interpretation of Iterated, Strictly Positive Inductive Definitions," *Archive For Mathematical Logic 32* (1992) 75-99.

[26] Schwichtenberg, Helmut, *Einige Anwendungen von unendlichen Termen und Wertfunktionalen*, Habilitationsschrift, Westfälische Wilhelms-Universität, Münster, 1973.

[27] Schwichtenberg, Helmut, "Some Applications of Cut-Elimination," in [7]

[28] Simpson, Stephen G., "Subsystems of Z_2 and Reverse Mathematics," appendix to Gaisi Takeuti, *Proof Theory* (second edition), North-Holland, 1987.

[29] Simpson, Stephen G., "Friedman's Research on Subsystems of Second Order Arithmetic," in Leo Harrington et al. eds, *Harvey Friedman's Research on the Foundations of Mathematics,* North-Holland, 1985.

[30] Spector, Clifford, "Provably Recursive Functionals of Analysis: A Consistency Proof of Analysis by an Extension of Principles Formulated in Current Intuitionistic Mathematics," in J. Dekker, ed., *Recursive Function Theory*, Proceedings of Symposia in Pure Mathematics Vol. 5, American Mathematics Society, 1962.

[31] Tait, William, "Infinitely Long Terms of Transfinite Type," in J. Crossley and M. Dummet eds., *Formal systems and recursive functions*, North-Holland, 1965.

[32] Troelstra, A.S., ed., *Metamathematical Investigation of Intuitionist Arithmetic and Analysis,* Lecture Notes in Mathematics 344, Springer, 1973.

[33] Troelstra, A.S., "Aspects of Constructive Mathematics" in [7].

[34] Troelstra, A.S., "On The Syntax of Martin-Löf's Type Theories," *Theoretical Computer Science 51* (1987) 1-26.

[35] Troelstra, A.S., Introductory notes to [18], in [14].

[36] Troelstra, A.S. and D. van Dalen, *Constructivism in Mathematics, An Introduction*, vols. I and II, North-Holland, 1988.