Interpretations between Theories

Mathematical Logic I
Fall 2023
Robert Rynasiewicz

December 5, 2023

Defined Symbols as Abbreviations

Suppose we're given an elementary language $\mathcal{L}$. There are two ways to treat defined symbols.

1. The symbol (in sequence with existing symbols) is an abbreviation for some expression in $\mathcal{L}$.
2. The symbol is added to $\mathcal{L}$ to form an expanded language together with a definition.

Examples of (1):

- $\exists x$ is short for $\neg\forall x\neg$.
- $(\varphi \vee \psi)$ is short for $\neg\varphi \to \psi$.
- $x \subseteq y$ is short for $\forall z(z \in x \to z \in y)$.

For an n-ary predicate symbol P, the general scheme is that $Px_1 \cdots x_n$ is short for $\varphi(x_1, \ldots, x_n)$ for some wff φ of $\mathcal{L}$ with exactly those free variables.

Defined Symbols as Added Vocabulary

Alternatively, in the case of a defined predicate or function symbol, we can expand $\mathcal{L}$ to a larger language $\mathcal{L}^+$ containing the new symbol. In order to capture its intended meaning, we provide a definition by means of a sentence δ in $\mathcal{L}^+$. For an n-ary predicate symbol P, δ takes the form

$$\forall x_1 \cdots \forall x_n (P x_1 \cdots x_n \leftrightarrow \varphi(x_1, \ldots, x_n)),$$

where $\varphi(x_1, \ldots, x_n)$ is a wff of $\mathcal{L}$ with free variables $x_1, \ldots, x_n$. For an n-place function symbol f, δ takes the form

$$\forall x_1 \cdots \forall x_n \forall x_{n+1} (f x_1 \cdots x_n = x_{n+1} \leftrightarrow \varphi(x_1, \ldots, x_n, x_{n+1})),$$

where, again, φ is a wff of $\mathcal{L}$ with free variables $x_1, \ldots, x_n, x_{n+1}$.

Conservative Extensions

Suppose T is a theory in $\mathcal{L}$ and we are given a definition δ of a new predicate symbol (respectively, function symbol). The definition generates a larger theory, $T^+ = \mathrm{Cn}\ T; \delta$, in $\mathcal{L}^+$. Obviously $T \subseteq T^+$. In fact T is a *proper* subset of T^+ (since $\delta \in T^+$ but $\delta \notin T$). However, although we get new theorems in the extended language $\mathcal{L}^+$, we don't want any new theorems in the original language $\mathcal{L}$. In other words, we want T^+ to be a *conservative extension* of T.

Defn. T^+ is a **conservative extension** of T iff there is no sentence σ of the original language $\mathcal{L}$ s.t. $\sigma \in T^+$ but $\sigma \notin T$.

As we will see below, we don't have to worry whether the addition of a definition of a predicate symbol leads to a conservative extension. The case with defined function symbols is different.

Example of Defining a New Predicate Symbol

We saw in connection with Peano arithmetic that we could define $<$ by

$$\forall x \forall y (x < y \leftrightarrow \exists z : x + Sz = y).$$

Let this sentence be δ and A the axioms of Peano arithmetic. Let $T = \text{Cn } A$ and $T^+ = \text{Cn } A; \delta$.

We claim that T^+ is a conservative extension. In fact, the situation is no different with any other predicate symbol.

Defined Predicates and Conservative Extensions

Theorem. If δ is the definition of a new predicate symbol, then $T^+ =_{df}$ Cn $T; \delta$ is a conservative extension of T.

Proof 1. Suppose $T; \delta \models \sigma$, where σ is a sentence of the narrower language $\mathcal{L}$. We need to show that T alone entails σ. So let $\mathfrak{A}$ be a model of T. We show it is also a model of σ. To do so, we expand $\mathfrak{A}$ to a structure $\mathfrak{A}^+$ for the expanded language $\mathcal{L}^+$ in accordance with the definition of the new predicate symbol P. Let n be the arity of P. The general form of the definition is

$$\forall x_1 \cdots \forall x_n (P x_1 \cdots x_n \leftrightarrow \varphi(x_1, \ldots, x_n)).$$

Define the interpretation of P in $\mathfrak{A}^+$ as the set of n-tuples:

$$P^{\mathfrak{A}^+} = \{\langle d_1, \ldots, d_n \rangle \in |\mathfrak{A}|^n \ : \ \models_{\mathfrak{A}} \sigma \ [\![d_1, \ldots, d_n]\!]\}.$$

Proof 1 (cont.) and Prologue for Proof 2

By hypothesis, $\mathfrak{A}$ satisfies T, and hence $\mathfrak{A}^+$ satisfies T since P does not occur in T. Moreover, $\mathfrak{A}^+$ was constructed so as to satisfy δ. Hence, $\mathfrak{A}^+$; satisfies $T; \gamma$. Thus, since $T; \gamma \models \sigma$, it follows that $\mathfrak{A}^+$ also satisfies σ. Hence, since σ does not contain P, $\mathfrak{A}$ satisfies σ. Therefore $T \models \sigma$. ■

Alternatively, if $T; \gamma \models \sigma$, then, by completeness, there is a Hilbert proof of σ from $T; \gamma$, which can be converted to a proof of σ from T alone. This requires a preliminary definition of the result of uniformly substituting in a wff $Pt_1 \cdots t_n$ with $\varphi(t_1, \ldots, t_n)$, where $t_1, \ldots, t_n$ are any terms of $\mathcal{L}$, and hence terms of $\mathcal{L}^+$.

A Preliminary Definition and Proof 2

The definition is by recursion. We adopt the notation $\alpha^\dagger$ for the result of the uniform substitution on α. Here's the definition.

$$
\begin{aligned}
(Pt_1 \cdots t_n)^\dagger &= \varphi(t_1, \ldots, t_n) \\
(Qt_t \cdots t_m)^\dagger &= Qt_t \cdots t_m \text{ for any other predicate } m\text{-ary symbol } Q \\
(\approx t_1 t_2)^\dagger &= \approx t_1 t_2 \\
(\neg\alpha)^\dagger &= \neg(\alpha^\dagger) \\
(\alpha \to \beta)^\dagger &= (\alpha^\dagger \to \beta^\dagger) \\
(\forall x\alpha)^\dagger &= \forall x(\alpha^\dagger)
\end{aligned}
$$

Proof 2. Suppose that $T; \gamma \models \sigma$. By completeness, there is a Hilbert proof $\langle \gamma_1, \ldots, \gamma_m \rangle$ of σ from $T; \delta$, where $\gamma_m = \sigma$. The proof proceeds by strong induction on $\mathbb{N}$. The inductive hypothesis is that, for given $k \leq m$, if $j < k$, then $\langle \gamma_1^\dagger, \ldots, \gamma_j^\dagger \rangle$ is a proof of $\gamma_j^\dagger$ from T.

Proof 2 (cont.)

We need to show that $\langle \gamma_1^\dagger, \ldots, \gamma_k^\dagger \rangle$ is then a proof of $\gamma_k^\dagger$ from T. This is done by separation of cases: (i) $\gamma_k \in T$, (ii) γ_k is a logical axiom, (iii) there exist $i, j < k$ such that γ_k follows from γ_i and γ_j by MP.

Case (i). $\gamma_k \in T$ entails $\gamma_k^\dagger = \gamma_k$, so trivially $\langle \gamma_1^\dagger, \ldots, \gamma_k^\dagger \rangle$ is a proof of $\gamma_k^\dagger$ from T.

Case (ii). One needs a subproof that, for any logical axiom λ, the wff $\lambda^\dagger$ is also a logical axiom. Given that subresult, the reasoning is the same as is case (i).

Case (iii). Suppose γ_k follows from γ_i and γ_j by MP. Then γ_i has the form $(\gamma_j \rightarrow \gamma_k)$. Since $(\gamma_j \rightarrow \gamma_k)^\dagger = (\gamma_j^\dagger \rightarrow \gamma_k^\dagger)$ it follows that $\gamma_k^\dagger$ follows from $\gamma_i^\dagger$ and $\gamma_j^\dagger$ by MP. Thus, $\langle \gamma_1^\dagger, \ldots, \gamma_k^\dagger \rangle$ is a proof of $\gamma_k^\dagger$ from T.

So, this holds for $\gamma_m^\dagger$, and since $\gamma_m = \sigma$, $\gamma_m^\dagger = \gamma_m$. ■

Defined Function Symbols

The case is more delicate for defining new function symbols. Again, the general form of a definition of a new function symbol f is

$$\forall x_1 \cdots \forall x_n \forall x_{n+1}(f x_1 \cdots x_n = x_{n+1} \leftrightarrow \varphi(x_1, \ldots, x_n, x_{n+1})),$$

Example. In ZF, the axiom of the power set reads:

$$\forall x \exists y \forall z(z \subseteq x \rightarrow z \in y).$$

This asserts, not the existence of a power set for each set, but rather at least one set that contains the power set. An application of instance of the Separation/Comprehension Schema, rids us of any elements of S_x not in the power set, so that the sentence

$$\forall x \exists y \forall z(z \subseteq x \leftrightarrow z \in y)$$

is a theorem of ZF. The definition of the power set symbol $\mathscr{P}$ is then

$$\forall x \forall y(\mathscr{P} x = y \leftrightarrow \forall z(z \subseteq x \leftrightarrow z \in y).$$

Existence and Uniqueness Requirements

In the example above, we were fortunate enough to have first secured the existence of a power set. It turns out that it is also a theorem of ZF that

$$\forall x \exists! y \forall z (z \subseteq x \leftrightarrow z \in y)$$

so that we have both existence and uniqueness. But existence and uniqueness do not come out of the blue. So, unlike the case of defined predicate symbols, which we can introduce any time without regard to a theory, in the case of defined function symbols, we must first specify of theory T in the original language that guarantees existence and uniqueness.

In other words, after fixing T, it must be the case that

$$T \models \forall x_1 \cdots x_n \exists! y \varphi \tag{1}$$

if φ is to be use as the *definiens* of a function symbol.

Cases of Failure

That we can't get away without first having a theory for which (1) above holds, is clear from examples (if not already clear!).

Example. Consider Peano arithmetic. Suppose we introduce a new function symbol f with the following definition.

$$\forall x \forall y (fx = y \leftrightarrow SS0 \cdot y < x).$$

(I'm assuming that $<$ has been defined as given earlier.)

Since there is no y s.t. $2y < 0$, the existence requirement for $f(0)$ fails.

For $f(1)$ we have both existence and uniqueness, viz., $f(1) = 0$.

For, say, $f(4)$, we have existence but lack uniqueness, since $2 \cdot 0 < 4$ and $2 \cdot 1 < 4$.

Evidently, PA$\not\models \forall x \exists! y : SS0 \cdot y < x$, and thus fails the necessary definitional condition.

The Necessary Condition is also Sufficient

"Sufficient for what?" you might ask. Answer: sufficient for Cn $T;\delta$ to be conservative, where δ is the proposed definition.

Lemma. Let φ be a proposed *definiens* for a new function symbol f and δ the definition. Then Cn $T;\delta$ is conservative iff $T \models \forall x_1 \cdots \forall x_n \exists! y \varphi(x_1, \ldots, x_n, y)$.

Proof. To avoid to much notational clutter, we show how the proof works for the case $n = 1$. That establishes the pattern for the general case.
($\Leftarrow$). Suppose that $T \models \forall x \exists! y \varphi(x, y)$. Further suppose that $T;\delta \models \sigma$, where σ is in the original language of T. We need to show that $T \models \sigma$. So suppose $\mathfrak{A}$ is a model of T. Expand $\mathfrak{A}$ to a structure $\mathfrak{A}^+$ for $\mathcal{L}^+$ that satisfies δ as well as T as follows. Take $f^{\mathfrak{A}^+}$ to be the operation on $|\mathfrak{A}|$ such that for each $d \in |\mathfrak{A}|$, it holds that $\models_{\mathfrak{A}^+} \varphi(x, y)\ [\![d, f^{\mathfrak{A}^+}(d)]\!]$. Since $\mathfrak{A}^+$ satisfies $T;\delta$, it satisfies σ as well. But since σ is a sentence of $\mathfrak{L}$, $\models_{\mathfrak{A}} \sigma$, and therefore $T \models \sigma$.

Remainder of the Proof; Intro to Interpretations

($\Rightarrow$). Assume that Cn $T;\delta$ is conservative. Let σ be the sentence $\forall x \exists! y \varphi(x,y)$. Since $\delta \models \sigma$, trivially $T;\delta \models \sigma$. Since Cn $T;\delta$ is conservative, it follows that $T \models \sigma$, i.e., $T \models \forall x \exists! y \varphi(x,y)$. ■

Introduction to Interpretations

If we have two theories T_1, T_2 in the same language $\mathcal{L}$ and $T_1 \subseteq T_2$, we say that T_1 is a *subtheory* of T_2 and T_2 an *extension* of T_1. The relation of set inclusion tells us which theory is logically stronger, viz., here T_2 is stronger than T_1 since whatever is the case according to T_1 is also the case according to T_2. This much is banal. But can we compare the logical strength of theories otherwise, say, neither is a subset of the other, or, more significantly, the theories are not even in the same language, for example, PA and ZF? We tend to think ZF the stronger, but on what grounds? To establish that it is requires interpreting the *language* of PA into ZF in such a way that PA becomes a subtheory of ZF.

Interpreting a Language into a Theory

Rather than taking this more challenging case head on, let's introduce definitions as motivated by some toy examples.

Let $\mathfrak{Z} =_{df} (\mathbb{Z}, +, \cdot)$ be the ring of integers and $T = \text{Th}\ \mathfrak{Z}$. Further, let $\mathcal{L} = \{0, S\}$. There are many ways to interpret $\mathcal{L}$ into T. However, one way is suggestive if we have in mind the structure $\mathfrak{N}_S =_{df} (\mathbb{N}, 0, S)$, which we'll call *successor arithmetic*. For we might be curious as to the relative strengths of Th $\mathfrak{N}_S$ and Th $\mathfrak{Z}$. We use π to denote an interpretation.

In this case, we want π to restrict universal quantification to $\mathbb{N}$ rather than over all of $\mathbb{Z}$. This is known as *relativization*. One way to do this is via Lagrange's 4-square theorem: every non-negative integer is the sum of four squares. So, π replaces a wff of the form $\forall x \varphi$ with the wff

$$\forall x \left(\exists y_1, y_2, y_3, y_4 : x = \sum_{i=1}^{4} y_i \cdot y_i \rightarrow \varphi \right).$$

π Applied to $\forall$

What's going on here is that π maps the universal quantifier to the wff

$$\exists y_1, y_2, y_3, y_4 : v_0 = \sum_{i=1}^{4} y_i \cdot y_i.$$

We'll write $\pi_\forall$ for $\pi(\forall)$. If we adopt the Enderton convention that, for a terms $t, t_1, t_2, \ldots, t_n$,

$$\begin{aligned} \varphi(t) &= \varphi_t^{v_0} \\ \varphi(t_1, \ldots, t_{n+1}) &= (\varphi(t_1, \ldots, t_n))_{t_{n+1}}^{v_n}, \end{aligned}$$

then the wff at the bottom of the last slide reads

$$\forall x(\pi_\forall(x) \to \varphi).$$

π Applied to the Elements of $\mathcal{L}$

For the vocabulary of $\mathcal{L}$, we know that the wff

$$v_0 + v_0 = v_0$$

defines the set $\{0\}$ in the standard model $\mathfrak{N}$ of arithmetic. So, let this be π_0.

Similarly, the successor relation (on $\mathbb{Z}$) can be defined

$$\forall x((x \cdot x = x \wedge x \neq x + x) \rightarrow v_0 + x = v_1).$$

So, take this to be π_S.

Then, e.g., the axiom $\forall x\, Sx \neq 0$ becomes

$$\forall x(\pi_\forall(x) \rightarrow \neg(\forall y(\pi_0(y) \rightarrow \forall z(\pi_S(x, z) \rightarrow y = z))).$$

This may seem like a lot of work for little benefit, but the rewards will come. We need a general definition of interpreting $\mathcal{L}$ into T.

Definition of Interpreting $\mathcal{L}$ into T

Defn. An **interpretation** of $\mathcal{L}$ into T is a function on $\mathcal{L}; \forall$ such that

1. $\pi(\forall)$ is a wff $\pi_\forall$ in which at most v_0 appears free and $T \models \exists v_0 \pi_\forall$.
   (The last requirement is so that $\pi_\forall$ defines a *non-empty* domain for any structure for $\mathcal{L}$.)
2. Let P be an n-ary predicate of $\mathcal{L}$. Then π_P is a wff of the language of T s.t. $FV(\pi_P) = \{v_i \mid 0 \leq i < n\}$.
3. Let f be an n-place function symbol of $\mathcal{L}$. Then π_f is a wff of the language of T s.t. $FV(\pi_f) = \{v_i \mid 0 \leq i \leq n\}$ and

$$T \models \forall v_0, \ldots, v_{n-1} \left(\bigwedge_{i=0}^{n-1} \pi_\forall(v_i) \rightarrow \exists x(\pi_\forall(x) \wedge \forall y(\pi_f(v_0, \ldots, v_n) \leftrightarrow v_n = x)) \right).$$

The Case of the Individual Constant

In the case of a 0-ary function, a.k.a., individual constant c, π_c is a wff with free variable v_0 and the last condition reads:

$$T \models \exists x(\pi_\forall(x) \wedge \forall y(\pi_c(y) \leftrightarrow y = x)).$$

Structures for $\mathcal{L}$ from Models of T

Under an interpretation π each model $\mathfrak{B}$ of T generates a structure ${}^{\pi}\mathfrak{B}$ for $\mathcal{L}$ as follows. For the domain of discourse, we have

$$|{}^{\pi}\mathfrak{B}| = \{d \in |\mathfrak{B}| \ : \ \models_{\mathfrak{B}} \pi_{\forall}(v_0)\ [\![d]\!]\}.$$

For the interpretation of an $(n+1)$-ary predicate symbol P:

$$P^{{}^{\pi}\mathfrak{B}} = \{\langle d_0, \ldots, d_n\rangle \in |{}^{\pi}\mathfrak{B}|^{n+1} \ : \ \models_{\mathfrak{B}} \pi_P(v_0, \ldots, v_n)\ [\![d_0, \ldots, d_n]\!]\}.$$

The interpretation of an n-ary function symbol looks similar:

$$f^{{}^{\pi}\mathfrak{B}} = \{\langle d_1, \ldots, d_n, e\rangle \in |{}^{\pi}\mathfrak{B}|^{n+1} \ : \ \models_{\mathfrak{B}} \pi_f(v_0, \ldots, v_n)\ [\![d_1, \ldots, d_n, e]\!]\},$$

only that for each $d_1, \ldots, d_n$ there exists a unique e s.t. $\langle d_1, \ldots, d_n, e\rangle \in f^{{}^{\pi}\mathfrak{B}}$ because of clause (3) in the definition of an interpretation.