SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

System T and the Dialectica Interpretation

av

Friðgeir Ingi Jónsson

2019 - No K36

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

System T and the Dialectica Interpretation

Friðgeir Ingi Jónsson

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Peter Lumsdaine

2019

Abstract

The Dialectica interpretation and system T were first introduced by Kurt Gödel in 1958. The original Dialectica interpretation was an interpretation of intuitionistic arithmetic into a typed functional theory called system T. In this thesis we present a modernized version of the Dialectica interpretation along with a proof of the soundness of the interpretation. We also look at two different ways to provide semantics for System T.

Acknowledgements

I would like to thank my supervisor Peter LeFanu Lumsdaine for introducing me to the deeply fascinating Dialectica interpretation as well as for his support and the invaluable advise he has given me.

System **T** and the Dialectica Interpretation

Friðgeir Ingi Jónsson

Contents

1 Introduction

In 1958 an article was published by Kurt Gödel in a special issue of the journal *Dialectica*, issued to commemorate the 70th birthday of the mathematician Paul Bernays. The article, titled 'Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes' which has been translated into English as 'On a hitherto unutilized extension of the finitary standpoint', described what is now called the *Dialectica interpretation*, an interpretation of intuitionistic arithmetic into a typed functional theory that Gödel called system **T**. The interpretation was meant by Gödel to provide a consistency proof for classical arithmetic by using it in conjunction with his double negation translation, which interprets classical theories into intuitionistic ones.

The Dialectica interpretation as well as system **T** have since then been shown to have a wide variety of applications within the studies of both mathematical logic and computer science. A survey of some of these applications can be found in Avigad and Feferman (1998).

In this paper we will present the original Dialectica interpretation in a modernized form and prove the soundness of it. This presentation is inspired by presentations of the interpretation in both Avigad and Feferman (1998) and Pédrot (2015). It is the writers opinion that this modernized form of the presentation makes it both clearer and more palatable for modern readers.

The paper also includes a short chapter on the semantics of system **T**. We show how to construct two very different models of system **T** and discuss some properties of these models.

2 The systems HA, T and HA+T

The Dialectica interpretation translates the formulas of Heyting arithmetic (**HA**) into formulas of a logical system we call **HA+T**, a first order theory of arithmetic taking its terms from a term rewriting system called system **T**. This section is dedicated to the presentation of these three systems.

2.1 Heyting Arithmetic

Heyting arithmetic (**HA**) is a first order theory of aritmethic identical in every sense to Peano arithmetic (**PA**) with the exception that the underlying deductive system of the theory is intuitionistic instead of classical.

2.1.1 Syntax of HA

The language of **HA** consists of the logical constants $\wedge, \vee, \rightarrow, \exists, \forall, \bot$; denumerably many variables $x, y, z, \dots$; an equality predicate symbol $=$; a symbol 0, denoting zero; a symbol $\mathbf{S}$, denoting the successor function; and symbols $+, \cdot$, denoting addition and multiplication respectively.

Definition 2.1. The *terms* of **HA** are defined inductively as follows

1. Every variable $x, y, z, \dots$ and the constant 0 are terms.
2. If t_1 and t_2 are terms, then $\mathbf{S}t_1$, $t_1 + t_2$ and $t_1 \cdot t_2$ are terms.

Definition 2.2. The *formulas* of **HA** are defined inductively as follows

1. If t_1 and t_2 are terms of **HA**, then $t_1 = t_2$ is a formula of **HA**.
2. $\bot$ is formula of **HA**.
3. If φ and ψ are formulas, then $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$ and $(\varphi \rightarrow \psi)$ are formulas as well.
4. If φ is a formula and x is a variable, then $\forall x\varphi$ and $\exists x\varphi$ are formulas as well.

We call the formulas of clauses 1 and 2 in this definition the *prime formulas* of **HA**.

Notation 2.3. We list some notational conventions:

- We let $\neg\varphi$ abbreviate $\varphi \rightarrow \bot$.
- We let $\varphi \leftrightarrow \psi$ abbreviate $(\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$.
- We stick to the convention of letting $\neg$ and the quantifiers bind stronger than $\wedge$ and $\vee$, who in turn bind stronger than $\rightarrow$ and $\leftrightarrow$, for example $\varphi \vee \neg\psi \rightarrow \theta \wedge \chi$ is equivalent to $(\varphi \vee (\neg\psi)) \rightarrow (\theta \wedge \chi)$.

We end the discussion of the syntax of **HA** by giving a formal definition of free and bound variables in **HA**.

Definition 2.4. We define the set of free variables of a formula φ of **HA**, denoted by $\mathbf{FV}(\varphi)$, inductively as follows.

- If φ is a prime formula and x a variable occurring in φ, then $x \in \mathbf{FV}(\varphi)$.
- If $\varphi = \psi \wedge \theta$, $\varphi = \psi \vee \theta$ or $\varphi = \psi \to \theta$ where ψ and θ are formulas, then $\mathbf{FV}(\varphi) = \mathbf{FV}(\psi) \cup \mathbf{FV}(\theta)$.
- If x is a variable, then $\mathbf{FV}(\forall x \varphi) = \mathbf{FV}(\varphi) - \{x\}$ and $\mathbf{FV}(\exists x \varphi) = \mathbf{FV}(\varphi) - \{x\}$.

2.1.2 Deductions in HA

In the historical presentations of the Dialectica Interpretation **HA** has been presented using Hilbert-style axioms and rules as its deduction system (see Gödel (1990), Troelstra (1973) and Avigad and Feferman (1998)). Rather than doing so we opt to present deduction in **HA** using a natural deduction style system.

Before presenting the deduction rules we give a few definitions.

Definition 2.5. A *capture free substitution* of a free variable x with a term s in a formula φ, denoted by $\varphi[x := s]$, is defined inductively as follows for variables x and y, terms t, s and r and formulas ψ_1 and ψ_2

1. $0[x := s] = 0$.
2. $x[x := s] = s$.
3. $y[x := s] = y$, if $y \neq x$.
4. $\mathbf{S}(t)[x := s] = \mathbf{S}(t[x := s])$.
5. $\bot[x := s] = \bot$.
6. $(t = r)[x := s] = (t[x := s] = r[x := s])$.
7. $(\psi_1 \wedge \psi_2)[x := s] = (\psi_1[x := s] \wedge \psi_2[x := s])$. The definition is the same for $(\psi_1 \vee \psi_2)[x := s]$ and $(\psi_1 \to \psi_2)[x := s]$.
8. $(\exists x \psi_1)[x := s] = \exists x \psi_1$ and $(\exists y \psi_1)[x := s] = \exists z(\psi_1[y := z][x := s])$ where z is chosen fresh for both φ and s, if $x \neq y$. The definition is the same for $(\forall x \psi_1)[x := s]$ and $(\forall y \psi_1)[x := s]$ where $x \neq y$.

Definition 2.6. An *environment* is a list of formulas, possibly empty. Environments can be defined inductively as follows

1. An empty list is an environment.
2. If Γ is an environment and φ a formula then Γ, φ is an environment.

Definition 2.7. A *sequent* is an expression of the form $\Gamma \vdash \varphi$ where Γ is an environment and φ a formula.

The intuitive meaning of a sequent $\Gamma \vdash \varphi$ should be clear, that from the formulas of the environment Γ one is able to deduce the formula φ. The way in which these deductions are done is of course governed by the rules of intuitionistic arithmetic, which we will list out now. We begin by giving the two so called *structural rules*:

$$\frac{}{\Gamma, \varphi \vdash \varphi}\text{Axiom} \qquad \frac{\Gamma \vdash \varphi}{\Gamma, \psi \vdash \varphi}\text{Weakening}$$

followed by the rules of intuitionistic propositional logic:

$$\frac{\Gamma \vdash \varphi \qquad \Gamma \vdash \psi}{\Gamma \vdash \varphi \wedge \psi}\wedge\text{I} \qquad \frac{\Gamma \vdash \varphi \wedge \psi}{\Gamma \vdash \varphi}\wedge\text{E}_1 \qquad \frac{\Gamma \vdash \varphi \wedge \psi}{\Gamma \vdash \psi}\wedge\text{E}_2$$

$$\frac{\Gamma \vdash \varphi}{\Gamma \vdash \varphi \vee \psi}\vee\text{I}_1 \qquad \frac{\Gamma \vdash \psi}{\Gamma \vdash \varphi \vee \psi}\vee\text{I}_2 \qquad \frac{\Gamma \vdash \varphi \vee \psi \qquad \Gamma, \varphi \vdash \theta \qquad \Gamma, \psi \vdash \theta}{\Gamma \vdash \theta}\vee\text{E}$$

$$\frac{\Gamma, \varphi \vdash \psi}{\Gamma \vdash \varphi \to \psi}\to\text{I} \qquad \frac{\Gamma \vdash \varphi \to \psi \qquad \Gamma \vdash \varphi}{\Gamma \vdash \psi}\to\text{E}$$

$$\frac{\Gamma \vdash \bot}{\Gamma \vdash \varphi}\bot\text{E}$$

By adding the following four rules for quantifiers to those above we get first order predicate logic:

$$\frac{\Gamma \vdash \varphi}{\Gamma \vdash \forall x \varphi}\forall\text{I} \qquad \frac{\Gamma \vdash \forall x \varphi}{\Gamma \vdash \varphi[x := t]}\forall\text{E}$$

$$\frac{\Gamma \vdash \varphi[x := t]}{\Gamma \vdash \exists x \varphi}\exists\text{I} \qquad \frac{\Gamma \vdash \exists x \varphi \qquad \Gamma, \varphi \vdash \psi}{\Gamma \vdash \psi}\exists\text{E}$$

In $\forall$I x cannot occur freely in any of the formulas in Γ and in $\exists$E x cannot occur freely in any of the formulas of Γ nor can it occur freely in ψ.

Taken together, these rules form the deduction system of first order intuitionistic logic. Should one want to expand this system to a classical one it suffices to add the law of excluded middle

$$\frac{}{\vdash \varphi \vee \neg\varphi}\text{LEM}$$

to the system.

To get **HA** we add to the system of first order intuitionistic logic the following rules. The rules for equality:

$$\frac{}{\Gamma \vdash n = n} \qquad \frac{\Gamma \vdash n = m \qquad \Gamma \vdash \varphi[x := n]}{\Gamma \vdash \varphi[x := m]}$$

the defining rules for **S** and 0:

$$\frac{}{\Gamma \vdash \neg(0 = \mathbf{S}n)} \qquad \frac{\Gamma \vdash \mathbf{S}n = \mathbf{S}n}{\Gamma \vdash n = n}$$

the defining rules for addition and multiplication:

$$\frac{}{\Gamma \vdash n + 0 = n} \qquad \frac{}{\Gamma \vdash n + \mathbf{S}m = \mathbf{S}(n + m)}$$

$$\frac{}{\Gamma \vdash n \cdot 0 = 0} \qquad \frac{}{\Gamma \vdash n \cdot \mathbf{S}m = n + (n \cdot m)}$$

and the induction rule:

$$\frac{\Gamma \vdash \varphi[x := 0] \qquad \Gamma \vdash \varphi[x := y] \to \varphi[x := \mathbf{S}y]}{\Gamma \vdash \varphi[x := n]}$$

2.2 System T

In Gödel's original article on the Dialectica translation the target language of the translation was a system he called system **T**. In that article system **T** was a full blown quantifier-free theory of arithmetic, a system including a typed term-rewriting system as well as a quantifier-free logic allowing one to reason about these terms.

What we call system **T** is basically the term-rewriting part of Gödel's system. In short our **T** is a typed λ-calculus with a few extra tools allowing us to encode arithmetic in it.

2.2.1 Types of T

Every term of system **T** is endowed with a type. We therefore begin our discussion of **T** by introducing the type-structure of **T**.

Definition 2.8. The *types* of **T** are defined inductively as follows

1. **N** is a type.

2. If τ and σ are types, then $\tau \to \sigma$ is a type.

The base-type **N** should be understood to be the type of the natural numbers and for each two types τ and σ the type $\tau \to \sigma$ should be understood to be the type of functions from elements of type τ to elements of type σ.

Notation 2.9. We stick to the convention of associating parentheses to the right, i.e.

$$\tau_1 \to \tau_2 \to \cdots \to \tau_{n-1} \to \tau_n$$

we read as

$$\tau_1 \to (\tau_2 \to \ldots (\tau_{n-1} \to \tau_n) \ldots).$$

2.2.2 Terms of T

Having defined the types of $\mathbf{T}$ we can now show how the terms of $\mathbf{T}$ are formed. Since each term has a type it is very important keep track of the type structure of the terms formed. We define the terms of $\mathbf{T}$ with this in mind.

Definition 2.10. The set of *terms* of $\mathbf{T}$ is defined inductively as follows, where $t : \tau$ reads as t is of the type τ.

1. The constant $0 : \mathbf{N}$ is a term.
2. For each type τ of $\mathbf{T}$ the variables $x^\tau : \tau, y^\tau : \tau, z^\tau : \tau, \ldots,$ are terms.
3. When $n : \mathbf{N}$ is a term, then $\mathbf{S}(n) : \mathbf{N}$ is a term.
4. When x^σ is a variable and $t : \tau$ is a term, then $(\lambda x^\sigma.t) : \sigma \to \tau$ is a term.
5. When $t : \sigma \to \tau$ and $s : \sigma$ are terms, then $t(s) : \tau$ is a term.
6. When $f : \tau$, $g : \mathbf{N} \to \tau \to \tau$ and $n : \mathbf{N}$ are terms, then $\mathbf{R}_\tau(f, g, n) : \tau$ is a term.

Terms of the form $\lambda x^\sigma.t : \sigma \to \tau$ are called λ-*abstractions.* They are the main tool we use construct functions from terms of type σ to terms of type τ. Terms of the form $t(s) : \tau$ where $s : \sigma$ and $t : \sigma \to \tau$ are called *applications.* An application should be considered to be the value of a function t applied to the term s.

The constants 0, $\mathbf{S}$ and $\mathbf{R}_\tau$ are used to code arithmetic. Just as in **HA** the constant 0 should be interpreted as zero and the constant $\mathbf{S}$ as the function that takes a natural number to its successor. The constant $\mathbf{R}_\tau$ is the so called recursor. It is used to define primitive recursive functions such as multiplication and addition.

Notation 2.11. Before moving on we introduce a few notational conventions designed to increase readability.

- Parentheses are omitted when there is no danger of confusion.
- Whenever it is clear from the context we suppress the type superscript of variables $x^\tau, y^\tau, \ldots$ and the type subscript of $\mathbf{R}_\tau$ and simply write $x, y, \ldots$ and $\mathbf{R}$.
- We write $\lambda x_1 x_2 \ldots x_n.t$ as a shorthand for $\lambda x_1.\lambda x_2.\ldots\lambda x_n.t$.
- When we have terms $t : \sigma_1 \to \cdots \to \sigma_n \to \tau, s_1 : \sigma_1, \ldots, s_n : \sigma_n$, we usually write $t(s_1, \ldots, s_n)$ instead of $t(s_1)\ldots(s_n)$.

As we will see here below when defining the reduction rules of $\mathbf{T}$, λ-abstractions are an incredibly simple and elegant tool to define functions. But before we can define the reduction rules we must have a clear notion of substitution. There are certain precautions that must be made when defining substitution in terms involving λ-abstractions to avoid syntactic mix-ups.

We will therefore state a few definitions and conventions regarding the naming of variables, substitutions and equivalences. These conventions ensure that our definitions of the reduction rules for **T**, which we give later in this section, are unproblematic.

In a λ-abstraction of the form $\lambda x.t$, the λ-symbol is said to *bind* any free occurrence of the variable x in the term t. We make this notion of *free* and *bound variables* precise in the following definition.

Definition 2.12. The set of free variables of a term t of **T**, denoted by $\mathbf{FV}(t)$, are defined inductively as follows. Let t, s and r be terms and x a variable. Then

1. $\mathbf{FV}(0) = \emptyset$,
2. $\mathbf{FV}(x) = \{x\}$,
3. $\mathbf{FV}(\mathbf{S}(t)) = \mathbf{FV}(t)$
4. $\mathbf{FV}(\lambda x.t) = \mathbf{FV}(t) - \{x\}$,
5. $\mathbf{FV}(t(s)) = \mathbf{FV}(t) \cup \mathbf{FV}(s)$, and
6. $\mathbf{FV}(\mathbf{R}(t, s, r)) = \mathbf{FV}(t) \cup \mathbf{FV}(s) \cup \mathbf{FV}(r)$.

A member of $\mathbf{FV}(t)$ is said to be *free* in t. Any variable occurring in t that is not free is said to be *bound* in t. If $\mathbf{FV}(t)$ is empty we say that t is *closed*.

There are two different notions of substitution in our system, the so-called capture-free substitution used for free variables and the so-called change of bound variables (sometimes also called α-conversion). We now define these two notions.

Definition 2.13. A *capture-free substitution* of a free variable x^τ with a term $s : \tau$ in a term t, denoted $t[x := s]$, is defined inductively as follows where the variables x and y, and the terms t, s, r and u are of the appropriate types:

1. $0[x := s] = 0$.
2. $x[x := s] = s$.
3. $y[x := s] = y$, if $y \neq x$.
4. $\mathbf{S}(t)[x := s] = \mathbf{S}(t[x := s])$
5. $t(r)[x := s] = t[x := s](r[x := s])$.
6. $\mathbf{R}(t, r, u)[x := s] = \mathbf{R}(t[x := s], r[x := s], u[x := s])$
7. $(\lambda x.t)[x := s] = \lambda x.t$.
8. $(\lambda y.t)[x := s] = \lambda z.(t[y := z][x := s])$ where z is chosen fresh for t and s, if $y \neq x$.

Definition 2.14. We call it a *change of a bound variable* in a term t (or α-*conversion* of a term) when some part of t of the form $\lambda x.s$ is swapped out for $\lambda y.(s[x := y])$ where the variable y is fresh in s.

When discussing the reduction rules of $\mathbf{T}$ we will note that a change of bound variables in a term does not constitute any change in the operational meaning of that term. Thus we define the following class of equivalences.

Definition 2.15. We say that two terms s and t are α-*equivalent*, written

$$s \equiv_\alpha t$$

if t can be obtained by a series of changes of bound variables in s.

Since all α-equivalent terms behave in the same way we will from here on out identify terms that are α-equivalent. But while we do not want to distinguish between α-equivalent terms we do want to distinguish between terms that differ only in the names of their free variables. For example, we would let $\lambda x.x = \lambda y.y$ while $\lambda x.yx \neq \lambda x.zx$.

Thus whenever we have two or more different terms where certain variables occur freely in some terms but bounded in others we change those bound variables to variables that do not occur in any of those terms. Following this convention ensure that there is no danger that substitution results in free variables unintentionally becoming bound.

Remark 2.16. While following the conventions we just introduced ensures that no problems arise when using substitution, we did omit a lot of technical details needed for a perfectly rigorous treatment of these issues. However such a treatment is really outside the scope of this presentation. For a detailed treatment of these issues the reader can look up chapter 2 and appendix C in Barendregt's *The Lambda Calculus, Its Syntax and Semantics* (1984) and appendix A1 of Hindley and Shelley's *Lambda-Calculus and Combinators, an Introduction* (2008).

As $\mathbf{T}$ is a term rewriting system it must have some rewriting rules. We introduce the reduction rules in $\mathbf{T}$.

Definition 2.17. If u and v are terms we define the relationship $u \triangleright v$ by the following directed equations, the so called *reduction rules* of $\mathbf{T}$:

1. $(\lambda x.t)s \triangleright t[x := s]$, where $x : \sigma$ is a variable and $t : \tau$ and $s : \sigma$ are terms.
2. $\mathbf{R}_\tau(f, g, 0) \triangleright f$, where $f : \tau$ and $g : \mathbf{N} \to \tau \to \tau$ are terms.
3. $\mathbf{R}_\tau(f, g, \mathbf{S}n) \triangleright g(n, \mathbf{R}(f, g, n))$, where $n : \mathbf{N}$, $f : \tau$ and $g : \mathbf{N} \to \tau \to \tau$ are terms.

The first reduction rule is usually called *β-reduction.* Note that all these equations are directed. This is because these are reduction rules, they describe how complex terms are reduced to simpler ones. These rules allow us to define the notion of *reduction.*

Definition 2.18. Let u and v be terms.

1. We say that u *reduces to v in one step*, written $u \to v$, if v can be obtained by replacing some subterm t of u by a term s such that $t \rhd s$.

2. We say that u *reduces* to v, written $u \to^* v$, if v can be obtained from u by a finite sequence of one step reductions.

3. We write $u \overset{*}{\leftrightarrow} v$, if $u \to^* v$ or $v \to^* u$.

Note that $\to^*$ is the reflexive and transitive closure of $\to$ and $\overset{*}{\leftrightarrow}$ the reflexive, symmetric and transitive closure of $\to$.

We close this section with a little demonstration of the expressive power of system $\mathbf{T}$.

Example 2.19. To illuminate the expressive power of the system we show how to define a few useful primitive recursive functions in $\mathbf{T}$.

$$\begin{aligned}
\mathrm{ADD} : \mathbf{N} \to \mathbf{N} \to \mathbf{N} &= \lambda xy.\mathbf{R}(x, \lambda pq.\mathbf{S}q, y) \\
\mathrm{MULT} : \mathbf{N} \to \mathbf{N} \to \mathbf{N} &= \lambda xy.\mathbf{R}(0, \lambda pq.\mathrm{ADD}(x, q), y) \\
\mathrm{SIGN} : \mathbf{N} \to \mathbf{N} &= \lambda x.\mathbf{R}(1, \lambda pq.0, x) \\
\mathrm{PRED} : \mathbf{N} \to \mathbf{N} &= \lambda x.\mathbf{R}(0, \lambda pq.p, x) \\
\mathrm{SUB} : \mathbf{N} \to \mathbf{N} \to \mathbf{N} &= \lambda xy.\mathbf{R}(x, \lambda pq.\mathrm{PRED}(q), y) \\
\mathrm{DIFF} : \mathbf{N} \to \mathbf{N} \to \mathbf{N} &= \lambda xy.\mathrm{ADD}(\mathrm{SUB}(x, y), \mathrm{SUB}(y, x)).
\end{aligned}$$

The terms $\mathrm{ADD}(x, y)$, $\mathrm{MULT}(x, y)$, $\mathrm{SUB}(x, y)$ and $\mathrm{DIFF}(x, y)$ are usually denoted by $x + y$, $x \cdot y$, $x \dot{-} y$ and $|x - y|$ respectively.

2.2.3 Sequences

For the Dialectica interpretation mere terms of $\mathbf{T}$ do not suffice. We often need sequences of terms as well. It is surprisingly simple to reason about sequences of terms in system $\mathbf{T}$. If one just follows a few simple notational convention one can treat sequences of terms of $\mathbf{T}$ almost as one would treat single terms.

The notational conventions presented here are taken from Pédrot (2015). Interestingly, while these notational convention are not made explicit, neither in the original, Gödel (1990), nor in the more recent presentations of the Dialectica by Troelstra (1973), (1990) and Avigad and Feferman (1998), they are usually at least implicitly followed. Making these explicit therefore adds a valuable clarity to the presentation.

Notation 2.20. A sequence $t_1, \ldots, t_n$ of terms will be denoted by $\vec{t}$ and a sequence $\tau_1, \ldots, \tau_n$ of types will be denoted by $\vec{\tau}$. A sequence, whether of terms or of types, can be a singleton, that is a single term t or a single type τ, or an empty sequence, denoted by $\emptyset$. For any sequence $\vec{x}$ we let $|\vec{x}|$ denote the length of the sequence and we let $\vec{x}, \vec{y}$ denote the concatenation of two sequences.

Now let $\vec{\tau} = \tau_1, \dots, \tau_n$ be a sequence of types and τ a type. Then we let $\tau \to \vec{\tau}$ denote a sequence of types:

$$\tau \to \vec{\tau} = \tau \to \tau_1, \dots, \tau \to \tau_n$$

and we let $\vec{\tau} \to \tau$ denote a type:

$$\vec{\tau} \to \tau = \tau_1 \to \dots \tau_n \to \tau.$$

Note that in this notation, given that $|\vec{\sigma}| = n$ and $|\vec{\tau}| = m$, we get

$$\vec{\sigma} \to \vec{\tau} = \sigma_1 \to \dots \to \sigma_n \to \tau_1, \dots, \sigma_1 \to \dots \to \sigma_n \to \tau_m$$

regardless of whether you expand $\vec{\sigma}$ or $\vec{\tau}$ first which showing us that this is not an ambiguous notation.

We deal with terms in a corresponding way. Given a sequence of terms $\vec{t} = t_1, \dots, t_n$ and a term t we let $\vec{t}(t)$ denote a sequence of terms:

$$\vec{t}(t) = t_1(t), \dots, t_n(t)$$

and we let $t(\vec{t})$ denote a term:

$$t(\vec{t}) = t(t_1, \dots, t_n).$$

The λ-abstractions must be dealt with in a way that matches the definition on applications so if we are additionally given a sequence of variables $\vec{x} = x_1, \dots, x_n$ and a variable x we let $\lambda x.\vec{t}$ denote a sequence of terms:

$$\lambda x.\vec{t} = \lambda x.t_1, \dots, \lambda x.t_n$$

and we let $\lambda \vec{x}.t$ denote a term:

$$\lambda \vec{x}.t = \lambda x_1 \dots x_n.t.$$

We also let $\vec{t} : \vec{\tau}$ denote

$$t_1 : \tau_1, \dots, t_n : \tau_n$$

given that $|\vec{t}| = |\vec{\tau}| = n$.

Since empty sequences are used a lot in the Dialectica interpretation we list out here below all the different types of situations in which empty sequences might occur in types and terms:

$$\emptyset \to \tau = \tau \qquad t\emptyset = t \qquad \lambda\emptyset.t = t$$

$$\tau \to \emptyset = \emptyset \qquad \emptyset t = \emptyset \qquad \lambda x.\emptyset = \emptyset$$

These are of course all special cases of notational conventions for sequences in general.

Given the right length of sequences the rules for the relationship : between terms and types, established in Definition 2.10, extend naturally to sequences and so does β-reduction, i.e. if $|\vec{x}| = |\vec{s}|$, then

$$(\lambda \vec{x}.\vec{t})\vec{s} \rhd \vec{t}[\vec{x} := \vec{s}]$$

where the substitution $[\vec{x} := \vec{s}]$ denotes the substitution of of each variable in $\vec{x}$ for the corresponding term in $\vec{s}$ and $\triangleright$ is taken to mean that the reduction rule holds for the terms corresponding to each other in the sequences on both sides of the symbol.

Thus we see that this notation allows us to extend **T** from single terms to sequences of terms with remarkable ease.

2.3 Higher type arithmetic

As we mentioned in the introductory remarks for section 2.2 Gödel intended system **T** to be the target language of the Dialectica translation. We also mentioned that Gödel's system **T** was endowed with an underlying logic. Since we have stripped **T** of all of its logical content, we must devise a logical system to reason about the terms of **T** and act as a target language for the Dialectica translation. We therefore introduce **HA+T**, a first-order theory of arithmetic for higher types.

2.3.1 HA+T

The logical theory **HA+T** is really just **HA** extended to allow reasoning about the terms of **T**. We will show below that **HA** is just a fragment of **HA+T**.

The terms of **HA+T** are simply the terms of **T** and the formulas of **HA+T** are defined as the formulas of **HA** are with two crucial changes: the prime formulas are restricted to equalities between terms of type **N** and variables in the scope of quantifiers are only allowed to range over terms of one type.

Definition 2.21. The formulas of **HA+T** are defined inductively as follows.

1. If $t : \mathbf{N}$ and $s : \mathbf{N}$ are terms of **HA+T**, then $t = s$ is a formula of **HA+T**.

2. $\perp$ is a formula of **HA+T**.

3. If φ and ψ are formulas of **HA+T**, then $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$ and $(\varphi \to \psi)$ are formulas of **HA+T** as well.

4. If φ is a formula of **HA+T** and $x : \tau$ a variable then $\forall x : \tau.\varphi$ and $\exists x : \tau.\varphi$ are formulas of **HA+T** as well.

As in **HA** we call the formulas consisting only of equalities as well as $\perp$ prime formulas. We let $\neg\varphi$ and $\varphi \leftrightarrow \psi$ denote the same formulas as we did in **HA**. We also stick to the same convention on how strongly logical connectives bind and the dropping of parentheses.

Notation 2.22. When we have a sequence $\vec{x} : \vec{\tau}$ of variables and a formula $\varphi(\vec{x})$ in which the variables of $\vec{x}$ possibly occur freely we use the following notation:

$$\exists\vec{x} : \vec{\tau}.\varphi(\vec{x}) = \exists x_1 : \tau_1. \dots \exists x_n : \tau_n.\varphi(\vec{x})$$

$$\forall\vec{x} : \vec{\tau}.\varphi(\vec{x}) = \forall x_1 : \tau_1. \dots \forall x_n : \tau_n.\varphi(\vec{x})$$

The rules of **HA**+**T** are very similar to the rules of **HA** but there are some key differences so we state them in their entirety for the sake of clarity. Definitions 2.5 and 2.6, of environments and sequents, can be used unchanged for **HA**+**T** and we define the rules of **HA**+**T** as follows. First we state the structural rules and the rules of propositional logic, all of which stay completely unchanged:

$$\frac{}{\Gamma, \varphi \vdash \varphi}\ \text{Axiom} \qquad \frac{\Gamma \vdash \varphi}{\Gamma, \psi \vdash \varphi}\ \text{Weakening}$$

$$\frac{\Gamma \vdash \varphi \qquad \Gamma \vdash \psi}{\Gamma \vdash \varphi \wedge \psi}\ \wedge\text{I} \qquad \frac{\Gamma \vdash \varphi \wedge \psi}{\Gamma \vdash \varphi}\ \wedge\text{E}_1 \qquad \frac{\Gamma \vdash \varphi \wedge \psi}{\Gamma \vdash \psi}\ \wedge\text{E}_2$$

$$\frac{\Gamma \vdash \varphi}{\Gamma \vdash \varphi \vee \psi}\ \vee\text{I}_1 \qquad \frac{\Gamma \vdash \psi}{\Gamma \vdash \varphi \vee \psi}\ \vee\text{I}_2 \qquad \frac{\Gamma \vdash \varphi \vee \psi \qquad \Gamma, \varphi \vdash \theta \qquad \Gamma, \psi \vdash \theta}{\Gamma \vdash \theta}\ \vee\text{E}$$

$$\frac{\Gamma, \varphi \vdash \psi}{\Gamma \vdash \varphi \to \psi}\ \to\text{I} \qquad \frac{\Gamma \vdash \varphi \to \psi \qquad \Gamma \vdash \varphi}{\Gamma \vdash \psi}\ \to\text{E}$$

$$\frac{\Gamma \vdash \bot}{\Gamma \vdash \varphi}\ \bot\text{E}$$

The rules for first-order logic on the other hand all have to be changed to make sure the type-structure of the terms of **T** is respected:

$$\frac{\Gamma \vdash \varphi \qquad x : \tau}{\Gamma \vdash \forall x : \tau . \varphi}\ \forall\text{I} \qquad \frac{\Gamma \vdash \forall x : \tau . \varphi \qquad t : \tau}{\Gamma \vdash \varphi[x := t]}\ \forall\text{E}$$

$$\frac{\Gamma \vdash \varphi[x := t] \qquad t : \tau}{\Gamma \vdash \exists x : \tau . \varphi}\ \exists\text{I} \qquad \frac{\Gamma \vdash \exists x : \tau . \varphi \qquad t : \tau \qquad \Gamma, \varphi \vdash \psi}{\Gamma \vdash \psi}\ \exists\text{E}$$

The restrictions of x not occurring freely in Γ in $\forall$I and x not occurring freely in Γ or ψ in $\exists$E are maintained here as usual.

There are also some changes made in the rules of arithmetic. The equality rules are changed to ensure that equality is only admissible between terms of the type of natural number:

$$\frac{n : \mathbf{N}}{\Gamma \vdash n = n} \qquad \frac{\Gamma \vdash n = m \qquad \Gamma \vdash \varphi[x := n]}{\Gamma \vdash \varphi[x := m]}$$

The defining rules for 0 and **S** remain unchanged but since + and $\cdot$ can be defined in **T** we remove the defining axioms for these and instead add a rule that allows for substitution between terms that are equivalent in terms of the rewriting rules:

$$\frac{}{\Gamma \vdash \neg(0 = \mathbf{S}n)} \qquad \frac{\Gamma \vdash \mathbf{S}n = \mathbf{S}n}{\Gamma \vdash n = n}$$

$$\frac{t \overset{*}{\leftrightarrow} s \qquad \Gamma \vdash \varphi[x := t]}{\Gamma \vdash \ \varphi[x := s]}$$

The induction rule is only changed to ensure that the terms in question are natural numbers:

$$\frac{y : \mathbf{N}, n : \mathbf{N} \qquad \Gamma \vdash \varphi[x := 0] \qquad \Gamma \vdash \varphi[x := y] \to \varphi[x := \mathbf{S}y]}{\Gamma \vdash \varphi[x := n]}$$

When it is necessary to make a distinction between derivations made in **HA** and derivations made in **HA+T** we write $\vdash_{\mathbf{HA}}$ to denote that a derivation was made in **HA** and $\vdash_{\mathbf{HA+T}}$ to denote that a derivation was made in **HA+T**.

It should be noted that any term of **HA** can easily be translated to a term of type **N** in **T**. This is done inductively: the base cases, 0 and the variables of variables of **HA** are translated to their obvious counterparts in **T**, the successor function of **HA** translates to its counterpart in **T** and the functions + and $\cdot$ translate to the function we denote by these symbols in example 2.19. Thus any term of **HA** has a counterpart in **T**.

Using this translation of the terms the formulas of **HA** can be translated by induction to formulas of **HA+T**: terms are translated in the way stated above, logical connectives of **HA** are translated to their corresponding logical connectives in **HA+T** and the quantifiers of **HA** are translated to quantifiers binding variables of type **N**. This way of translating terms and formulas of **HA** into **HA+T** will be used implicitly from here on.

Given this translation the following proposition holds.

Proposition 2.23. Given an environment Γ of **HA** and a formula φ of **HA**, if $\Gamma \vdash_{\mathbf{HA}} \varphi$, then $\Gamma \vdash_{\mathbf{HA+T}} \varphi$.

Proof. This is proved by induction on the rules of **HA**. Most of the rules of **HA** translate directly into almost identical rules in **HA+T**, the only difference being that in some cases there are restriction on the type of the terms in question. Since all of the terms of **HA** are of the type **N** this poses no problem and the proof follows immediately for these.

The only rules of **HA** that do not have a direct counterpart in **HA+T** are the defining rules for addition and multiplication. These rules are easily shown to correspond to the way in which these functions are defined in **T** and thus the substitution-rule for $\overset{*}{\leftrightarrow}$-equivalent terms provides the proof for these terms. □

This shows that **HA** is a fraction of **HA+T**, more precisely **HA** is just **HA+T** restricted to terms of type **N** and allowing only the functions + and $\cdot$.

3 The Dialectica Interpretation

The Dialectica interpretation provides for each well formed formula, φ, in the language of **HA** a *Dialectica translation*, φ^D, which is a **HA**+**T** formula of the form

$$\varphi^D = \exists \vec{x} : \mathcal{W}_\varphi . \forall \vec{y} : \mathcal{C}_\varphi . \varphi_D(\vec{x}, \vec{y})$$

where $\mathcal{W}_\varphi$ is a so called witness type, $\mathcal{C}_\varphi$ is a so called counter type and $\varphi_D(x, y)$ is a quantifier-free formula in the language of **HA**+**T**.

An intuitive way to understand φ^D is to think of it in the terms of game semantics. On this reading the formula $\varphi_D(\vec{x}, \vec{y})$ is a 'game' or a set of 'rules' and $\vec{x} : \mathcal{W}_\varphi$ represents the possible 'moves' one player can make while $\vec{y} : \mathcal{C}_\varphi$ represents the possible 'moves' of his opponent.

Then φ^D reads as the statement that there exists some sequence of terms $\vec{t} : \mathcal{W}_\varphi$ that represents moves that will defeat any possible choice of moves $y : \mathcal{C}_\varphi$ at the game $\varphi_D(\vec{x}, \vec{y})$. The hope is therefore that **T** provides such a witness for any theorem of **HA**. As we shall see, the beauty of Gödel's Dialectica interpretation is that **T** does in fact do this.

3.1 The witness and counter types

To begin our presentation of the interpretation we give a definition of the witness and counter types of each formula of **HA**.

Definition 3.1. The *witness types* and *counter types* of the formulas of **HA**, denoted by $\mathcal{W}_\varphi$ and $\mathcal{C}_\varphi$ respectively, where φ is a formula are sequences of types defined as follows

1. If φ is prime, then $\mathcal{W}_\varphi = \mathcal{C}_\varphi = \emptyset$.

Otherwise assume that ψ_1 and ψ_2 are formulas. Then

2. $\mathcal{W}_{\psi_1 \wedge \psi_2} = \mathcal{W}_{\psi_1}, \mathcal{W}_{\psi_2}$ and $\mathcal{C}_{\psi_1 \wedge \psi_2} = \mathcal{C}_{\psi_1}, \mathcal{C}_{\psi_2}$
3. $\mathcal{W}_{\psi_1 \vee \psi_2} = \mathbf{N}, \mathcal{W}_{\psi_1}, \mathcal{W}_{\psi_2}$ and $\mathcal{C}_{\psi_1 \vee \psi_2} = \mathcal{C}_{\psi_1}, \mathcal{C}_{\psi_2}$
4. $\mathcal{W}_{\forall z \psi_1} = \mathbf{N} \to \mathcal{W}_{\psi_1}$ and $\mathcal{C}_{\forall z \psi_1} = \mathbf{N}, \mathcal{C}_{\psi_1}$
5. $\mathcal{W}_{\exists z \psi_1} = \mathbf{N}, \mathcal{W}_{\psi_1}$ and $\mathcal{C}_{\exists z \psi_1} = \mathcal{C}_{\psi_1}$
6. $\mathcal{W}_{\psi_1 \to \psi_2} = \mathcal{W}_{\psi_1} \to \mathcal{W}_{\psi_2}, \mathcal{W}_{\psi_1} \to \mathcal{C}_{\psi_2} \to \mathcal{C}_{\psi_1}$ and $\mathcal{C}_{\psi_1 \to \psi_2} = \mathcal{W}_{\psi_1}, \mathcal{C}_{\psi_2}$.

If $\mathcal{W}_\varphi = \emptyset$ or $\mathcal{C}_\varphi = \emptyset$ we say that these types are *empty*.

Notation 3.2. Given a sequence $\Gamma = \varphi_1, \ldots, \varphi_n$ of **HA** formulas we let

$$\mathcal{W}_\Gamma = \mathcal{W}_{\varphi_1}, \ldots, \mathcal{W}_{\varphi_n}$$

and

$$\mathcal{C}_\Gamma = \mathcal{C}_{\varphi_1}, \ldots, \mathcal{C}_{\varphi_n}.$$

We end this discussion of the witness and counter types by presenting an important lemma involving them.

Lemma 3.3. For any formula φ of **HA** and any variable x, if t is a term of **HA** then

$$\mathcal{W}_{\varphi[x:=t]} = \mathcal{W}_{\varphi}$$

and

$$\mathcal{C}_{\varphi[x:=t]} = \mathcal{C}_{\varphi}.$$

Proof. By induction on the length of φ. □

This lemma shows that the witnesses and counters are invariant of which free variables or terms occur in the formulas of **HA**, so the witness and counter types are only determined by the logical structure of their formulas.

3.2 Translating HA

The crux of the Dialectica interpretation is the Dialectica translation itself.

Definition 3.4. For any formula φ in the language of **HA** its *Dialectica translation*, φ^D, is a **HA**+**T** formula of the form

$$\exists \vec{x} : \mathcal{W}_{\varphi}.\forall \vec{y} : \mathcal{C}_{\varphi}.\varphi_D(\vec{x}, \vec{y})$$

where φ_D is a quantifier-free formula in the language of **HA**+**T**. The formulas φ^D and φ_D are defined inductively as follows:

1. If φ is a prime formula then $\mathcal{W}_{\varphi}$ and $\mathcal{C}_{\varphi}$ are empty, so $\varphi^D = \varphi_D = \varphi$.

Otherwise assume

$$\psi_1^D = \exists \vec{x} : \mathcal{W}_{\psi_1}.\forall \vec{y} : \mathcal{C}_{\psi_1}.\psi_{1D}(\vec{x}, \vec{y}) \text{ and } \psi_2^D = \exists \vec{u} : \mathcal{W}_{\psi_2}.\forall \vec{v} : \mathcal{C}_{\psi_2}.\psi_{2D}(\vec{u}, \vec{v}).$$

Then

2. $(\psi_1 \wedge \psi_2)^D = \exists \vec{x}, \vec{u} : \mathcal{W}_{\psi_1 \wedge \psi_2}.\forall \vec{y}, \vec{v} : \mathcal{C}_{\psi_1 \wedge \psi_2}.(\psi_{1D}(\vec{x}, \vec{y}) \wedge \psi_{2D}(\vec{u}, \vec{v}))$
3. $(\psi_1 \vee \psi_2)^D = \exists z, \vec{x}, \vec{u} : \mathcal{W}_{\psi_1 \vee \psi_2}.\forall \vec{y}, \vec{v} : \mathcal{C}_{\psi_1 \vee \psi_2}.((z = 0 \wedge \psi_{1D}(\vec{x}, \vec{y}))$
   $\vee (z = 1 \wedge \psi_{2D}(\vec{u}, \vec{v})))$
4. $(\forall z \psi_1)^D = \exists \vec{X} : \mathcal{W}_{\forall z \psi_1}.\forall z, \vec{y} : \mathcal{C}_{\forall z \psi_1}.\psi_{1D}(\vec{X}(z), \vec{y})$
5. $(\exists z \psi_1)^D = \exists z, \vec{x} : \mathcal{W}_{\exists z \psi_1}.\forall \vec{u} : \mathcal{C}_{\exists z \psi_1}.\psi_{1D}(z, \vec{x}, \vec{y})$
6. $(\psi_1 \to \psi_2)^D = \exists \vec{U}, \vec{Y} : \mathcal{W}_{\psi_1 \to \psi_2}.\forall \vec{x}, \vec{v} : \mathcal{C}_{\psi_1 \to \psi_2}.(\psi_{1D}(\vec{x}, \vec{Y}(\vec{x}, \vec{v})) \to \psi_{1D}(\vec{U}(\vec{x}), \vec{v}))$.

From our definition of $\neg\varphi$ one can add the following to the list above.

7. $(\neg\psi_1)^D = (\psi_1 \to \bot)^D = \exists \vec{Y} : \mathcal{W}_{\psi_1} \to \mathcal{C}_{\psi_1}.\forall \vec{x} : \mathcal{W}_{\psi_1}.(\psi_{1D}(\vec{x}, \vec{Y}(\vec{x})) \to \bot)$

The Dialectica translation is structured in such a way that the free variables of φ_D are always either free in φ or components of $\vec{x} : \mathcal{W}_\varphi$ or $\vec{y} : \mathcal{C}_\varphi$. The only cases where there is any threat that variables bound in a formula φ become free in φ_D is when φ is a formula including quantifiers, that is in clauses 4 and 5 in the definition above. In both cases it is avoided by integrating the previously bounded z into the witness sequence.

This is makes the following proposition possible.

Proposition 3.5. If φ is a formula of **HA** and z is a variable that is free in φ but does not occur in $\vec{x} : \mathcal{W}_\varphi$ or $\vec{y} : \mathcal{C}_\varphi$, then for any term t

$$\varphi[z := t]_D(\vec{x}, \vec{y}) = \varphi_D(\vec{x}, \vec{y})[z := t].$$

Proof. By induction on the length of φ. □

Example 3.6. At first encounter the Dialectica translation can seem very confusing. In order to disperse some of that confusion we give examples of the Dialectica translations of two formulas of **HA**.

1. Let us first look at the formula $\forall x(x = 0 \vee \exists y(x = \mathbf{S}y))$, a theorem of **HA**. We let the formula be denoted by φ. A translation of a formula relies on the translation of each of its subformulas so we should begin by looking at the translations of the smallest subformulas. These are the prime formulas $x = 0$ and $x = \mathbf{S}y$ which have empty witness and counter types and trivial translations. The second smallest subformula is $\exists y(x = \mathbf{S}y)$ which has the witness type $\mathbf{N}$, an empty counter type and translates as

$$(\exists y(x = \mathbf{S}y))^D = \exists y : \mathbf{N}.(x = \mathbf{S}y)$$

The next subformula is $x = 0 \vee \exists y(x = \mathbf{S}y)$. This formula has the witness type $\mathbf{N}, \mathbf{N}$ an empty counter type and the translation

$$(x = 0 \vee \exists y(x = \mathbf{S}y))^D = \\ \exists y, z : \mathbf{N}, \mathbf{N}.((z = 0 \wedge x = 0) \vee (z = 1 \wedge x = \mathbf{S}y)).$$

We can then translate φ. The formula has the witness and counter types

$$\mathcal{W}_\varphi = \mathbf{N} \to \mathbf{N}, \mathbf{N} \to \mathbf{N} \text{ and } \mathcal{C}_\varphi = \mathbf{N}$$

and its Dialectica translation is

$$\varphi^D = \exists Y, Z : \mathcal{W}_\varphi.\forall x : \mathcal{C}_\varphi.((Z(x) = 0 \wedge x = 0) \vee (Z(x) = 1 \wedge x = \mathbf{S}(Y(x)))).$$

It is not difficult to construct effective witnesses for φ. We let

$$t_1 = \lambda x.\mathbf{R}(0, \lambda pq.1, x) \text{ and } t_2 = \text{PRED}$$

and then it is easy to see that

$$\vdash_{\mathbf{HA+T}} (t_1(x) = 0 \wedge x = 0) \vee (t_1(x) = 1 \wedge x = \mathbf{S}(t_2(x)))$$

regardless of how x is chosen.

2. Define $x \leq y$ as the formula $\exists z(x + z = y)$. We take a look at the formula $\forall x(\mathbf{S}(0) \leq x) \rightarrow \forall y(0 \leq y)$. This implication is undeniably a theorem of **HA** and although it may seem like a silly example since the hypothesis is clearly not provable and the conclusion is provable independently of the hypothesis, its translation provides an illuminating demonstration of how the Dialectica interpretation treats implications.

   Let φ denote the whole formula and let ψ_1 and ψ_2 denote $\forall x(\mathbf{S}(0) \leq x)$ and $\forall x(0 \leq x)$ respectively. By following the same method as in the the previous example we get

$$\mathcal{W}_{\psi_1} = \mathcal{W}_{\psi_2} = \mathbf{N} \rightarrow \mathbf{N}$$
$$\mathcal{C}_{\psi_1} = \mathcal{C}_{\psi_2} = \mathbf{N}$$

   and

$$\psi_1^D = \exists Z : \mathcal{W}_{\psi_1}.\forall x : \mathcal{C}_{\psi_1}.(\mathbf{S}(0) + Z(x) = x)$$
$$\psi_2^D = \exists U : \mathcal{W}_{\psi_2}.\forall y : \mathcal{C}_{\psi_2}.(0 + U(y) = y)$$

   Thus for φ we get the following witness and counter types

$$\mathcal{W}_\varphi = (\mathbf{N} \rightarrow \mathbf{N}) \rightarrow (\mathbf{N} \rightarrow \mathbf{N}), (\mathbf{N} \rightarrow \mathbf{N}) \rightarrow \mathbf{N} \rightarrow \mathbf{N}$$
$$\mathcal{C}_\varphi = \mathbf{N} \rightarrow \mathbf{N}, \mathbf{N}$$

   and φ translates as

$$\varphi^D = \exists U', X : \mathcal{W}_\varphi.\forall Z, y : \mathcal{C}_\varphi.$$
$$((\mathbf{S}(0) + Z(X(Z, y)) = X(Z, y)) \rightarrow (0 + U'(Z, y) = y)).$$

   There are several different ways of constructing witnesses t_1 and t_2 for φ such that

$$\vdash_{\mathbf{HA}+\mathbf{T}} (\mathbf{S}(0) + Z(t_1(Z, y)) = t_1(Z, y)) \rightarrow (0 + t_2(Z, y) = y)$$

   for all Z, y. For example we might let $t_1 = \lambda pq.0$. Then the hypothesis leads to absurdity, proving the translation regardless of how t_2 is chosen. Similarly we could let $t_2 = \lambda pq.q$ which would prove the translation regardless of the choice of t_1. These two choices of witnesses represent the proofs of φ in **HA** consisting of either showing the absurdity of ψ_1 or proving ψ_2 independently of ψ_1.

   Interestingly enough **T** offers another way of constructing witnesses for φ. Let

$$t_1 = \lambda pq.q \text{ and } t_2 = \lambda pq.\mathbf{S}(p(q)).$$

   Then we get

$$\mathbf{S}(0) + Z(y) = y \rightarrow 0 + \mathbf{S}(Z(y)) = y.$$

   Not only do these two terms effectively witness the translation but they are also very much in the spirit of the Dialectica translation of implications,

> t_1 effectively converts witnesses of the hypothesis into witnesses of the conclusion and t_2 converts counters of the conclusion into counters of the hypothesis.

The main result of Gödel's original article is that system **T** effectively provides sequences of terms witnessing the Dialectica translation of every theorem of **HA**. This is proved by induction over the length of deductions in **HA**. We have mentioned that in the historical presentations of the Dialectica interpretation, deductions in **HA** are usually presented in Hilbert style systems. We however chose to present deductions in a natural deduction style system. In Theorem 3.7, we will present a version of Gödel's main result modified to suit the parameters of our presentation of **HA**. For us to be able to present and prove such a result we need a way to extend the Dialectica translation to sequents.

Luckily for us there is a natural way to do this. Let $\Gamma = \varphi_1, \dots, \varphi_n$ be an environment. We remind the reader of the following well known metatheorem:

$$\Gamma \vdash \psi \text{ if and only if } \vdash \varphi_1 \wedge \dots \wedge \varphi_n \to \psi.$$

A Dialectica translation of the sequent on the left hand side should therefore be the same as stating that the Dialectica translation of the formula on the right hand side is provable in **HA+T**. In accordance with this we define the witness and counter types for sequents as follows:

$$\begin{array}{rcl} \mathcal{W}_{\varphi_1,\dots,\varphi_n \vdash \psi} & = & \mathcal{W}_{\varphi_1 \wedge \dots \wedge \varphi_n \to \psi} \\ & = & \mathcal{W}_{\varphi_1} \to \dots \to \mathcal{W}_{\varphi_n} \to \mathcal{W}_{\psi}, \\ & & \mathcal{W}_{\varphi_1} \to \dots \to \mathcal{W}_{\varphi_n} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\varphi_1}, \\ & & \dots, \\ & & \mathcal{W}_{\varphi_1} \to \dots \to \mathcal{W}_{\varphi_n} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\varphi_n} \\ \\ \mathcal{C}_{\varphi_1,\dots,\varphi_n \vdash \psi} & = & \mathcal{C}_{\varphi_1 \wedge \dots \wedge \varphi_n \to \psi} \\ & = & \mathcal{W}_{\varphi_1}, \dots, \mathcal{W}_{\varphi_n}, \mathcal{C}_{\psi} \end{array}$$

If we were to define the Dialectica translation of the sequent in the same manner we would get the following:

$$\begin{aligned} (\Gamma \vdash \psi)^D &= (\varphi_1 \wedge \dots \wedge \varphi_n \to \psi)^D \\ &= \exists \vec{U}, \vec{Y} : \mathcal{W}_{\Gamma \vdash \psi}. \forall \vec{x}, \vec{v} : \mathcal{C}_{\Gamma \vdash \psi}. \\ &\quad (\varphi_{1D}(\vec{x}_1, \vec{Y}_1(\vec{x}, \vec{v})) \wedge \dots \wedge \varphi_{nD}(\vec{x}_n, \vec{Y}_n(\vec{x}, \vec{v})) \to \psi_D(\vec{U}(\vec{x}), \vec{v})) \end{aligned}$$

where $\vec{x} = \vec{x}_1, \dots, \vec{x}_n$ and $\vec{Y} = \vec{Y}_1, \dots, \vec{Y}_n$.

However this is a very cumbersome formula so we would like some simpler way of stating a Dialectica translation of a sequent. Note that the Dialectica translation of a sequent is equivalent with the meta-statement that there exists a sequence of terms $\vec{U}, \vec{Y}_1, \dots, \vec{Y}_n : \mathcal{W}_{\Gamma \vdash \psi}$ such that

$$\varphi_{1D}(\vec{x}_1, \vec{Y}_1(\vec{x}, \vec{v})), \dots, \varphi_{nD}(\vec{x}_n, \vec{Y}_n(\vec{x}, \vec{v})) \vdash_{\mathbf{HA+T}} \psi_D(\vec{U}(\vec{x}), \vec{v})$$

for any choice of counters $\vec{x}_1, \ldots, \vec{x}_n, \vec{v} : \mathcal{C}_{\Gamma \vdash \psi}$. By letting $\vec{x}$ and $\vec{Y}$ keep the same meaning as above we can use the following method to abbreviate environments in **HA+T**:

$$\Gamma_D(\vec{x}, \vec{Y}(\vec{x}, \vec{v})) = \varphi_{1D}(\vec{x}_1, \vec{Y}_1(\vec{x}, \vec{v})), \ldots, \varphi_{nD}(\vec{x}_n, \vec{Y}_n(\vec{x}, \vec{v})).$$

The Dialectica translation of a sequent can then be stated in relatively compact way as:

There exists a sequence $\vec{U}, \vec{Y} : \mathcal{W}_{\Gamma \vdash \psi}$ such that

$$\Gamma_D(\vec{x}, \vec{Y}(\vec{x}, \vec{v})) \vdash_{\mathbf{HA+T}} \psi_D(\vec{U}(\vec{x}), \vec{v})$$

for any choice of $\vec{x}, \vec{v} : \mathcal{C}_{\Gamma \vdash \psi}$.

This last statement of the translation captures the meaning of the translation perfectly. We will therefore use it as our Dialectica translation of sequents from here on.

We finish this section by taking a little deeper look into the inner workings of the Dialectica translation. Let us think of what it means to prove a formula of the form $\varphi^D = \exists \vec{x} : \mathcal{W}_\varphi . \forall \vec{u} : \mathcal{C}_\varphi . \varphi_D(\vec{x}, \vec{y})$ in a constructive manner. To do so one would construct a sequence of terms $\vec{t} : \mathcal{W}_\varphi$ such that $\vdash_{\mathbf{HA+T}} \varphi_D(\vec{t}, \vec{y})$ given any possible choice of a counter $\vec{y} : \mathcal{C}_\varphi$. In terms of the game semantics reading we gave above, one has to construct a witness capable of defeating any counters at the game φ_D.

Let us now look at the Dialectica translation with this in mind. The definition of the translations of prime formulas, conjunctions and formulas involving existential quantification does not need much explanation. The translation does not need much explanation for disjunctions either, to prove a translated disjunction we simply have to construct witnesses for each disjunct and a term $z : \mathbf{N}$ containing information pointing at a witness that defeats its counters.

The translation of formulas involving universal quantification are proved by constructing a function that maps each natural number $z : \mathbf{N}$ to a witness of φ. This makes sense since a constructive proof of a formula of the form $\forall x \varphi$ should indeed consist of constructing a function mapping each element of the domain of discourse to a proof of φ.

To prove the translation of implications is perhaps the most confusing part. To do so one has to construct two functions, one from the witnesses of the hypothesis to the witnesses of the conclusion and another one from the witnesses of the hypothesis to a function from the counters of the conclusion to counters of the hypothesis. The basic idea is that if an implication is provable, then from any witness of the hypothesis one should be able to find both a witness of the hypothesis and a function transforming the counters of the conclusion into counters of the hypothesis.

There is another more technical way to justify the Dialectica interpretation, by showing through induction that $\varphi \leftrightarrow \varphi^D$. We will see that this is only possible by allowing the use of a few non-intuitionistic principles.

It is clear that if φ is prime formula, then $\varphi \leftrightarrow \varphi^D$. It is also easy to see that the equivalences $(\varphi \wedge \psi)^D \leftrightarrow (\varphi^D \wedge \psi^D)$, $(\varphi \vee \psi)^D \leftrightarrow (\varphi^D \vee \psi^D)$ and $(\exists z\varphi(z))^D \leftrightarrow \exists z(\varphi(z)^D)$ are all justified intuitionistically. However the equivalence $(\forall z\varphi(z))^D \leftrightarrow \forall z(\varphi(z)^D)$ can only be justified by an application of the axiom of choice

$$\forall x\exists y\varphi(x,y) \to \exists Y\forall x\varphi(x,Y(x)), \tag{AC}$$

a principle not generally accepted to be constructive.

The equivalence $(\varphi \to \psi)^D \leftrightarrow (\varphi^D \to \psi^D)$ requires a bit more work than the rest. It is justified by stepwise applying the following equivalences:

$$\begin{array}{ll}
\exists\vec{x}\forall\vec{y}\varphi_D(\vec{x},\vec{y}) \to \exists\vec{u}\forall\vec{v}\psi_D(\vec{u},\vec{v}) & \leftrightarrow \text{(i)} \\
\forall\vec{x}(\forall\vec{y}\varphi_D(\vec{x},\vec{y}) \to \exists\vec{u}\forall\vec{v}\psi_D(\vec{u},\vec{v})) & \leftrightarrow \text{(ii)} \\
\forall\vec{x}\exists\vec{u}(\forall\vec{y}\varphi_D(\vec{x},\vec{y}) \to \forall\vec{v}\psi_D(\vec{u},\vec{v})) & \leftrightarrow \text{(iii)} \\
\forall\vec{x}\exists\vec{u}\forall\vec{v}(\forall\vec{y}\varphi_D(\vec{x},\vec{y}) \to \psi_D(\vec{u},\vec{v})) & \leftrightarrow \text{(iv)} \\
\forall\vec{x}\exists\vec{u}\forall\vec{v}\exists\vec{y}(\varphi_D(\vec{x},\vec{y}) \to \psi_D(\vec{u},\vec{v})) & \leftrightarrow \text{(v)} \\
\exists\vec{U}\vec{Y}\forall\vec{x}\vec{v}(\varphi_D(\vec{x},\vec{Y}(\vec{x},\vec{v})) \to \psi_D(\vec{U}(\vec{x}),\vec{v})). &
\end{array}$$

Of these equivalences only (i) and (iii) are intuitionistically acceptable. Equivalence (v) is a double application of the axiom of choice. Equivalence (ii) can be justified as a special case the classically acceptable independence principle

$$(\varphi \to \exists x\psi) \to \exists x(\varphi \to \psi) \tag{IP}$$

and equivalence (iv) can be justified using a special case of Markov's principle

$$\neg\forall x\theta \to \exists x\neg\theta \tag{MP'}$$

where θ is quantifier-free. Neither (IP) nor (MP') are generally accepted as constructive principles. Interestingly enough however, the Dialectica interpretaion verifies the three non-intuitionistic principles (AC), (IP) and (MP') making it an interpretation of slightly more than just pure intuitionistic arithmetic. We will not discuss this any further but interested readers can look up section 3.1 in Avigad and Feferman (1998) as well as sections 7.4 and 7.6 of Pédrot (2015).

3.3 The soundness of the Dialectica interpretation

We now want to show the soundness of the Dialectica interpretation, that is to say that for any sequent it is possible to deduce in **HA**, there exists a sequence of terms in **T** witnessing the sequents Dialectica translation. We state this as a theorem.

Theorem 3.7. Let $\Gamma = \varphi_1, \ldots, \varphi_n$ be an environment in **HA** and ψ be a formula of **HA** and assume that

$$\Gamma \vdash_{\mathbf{HA}} \psi.$$

Then there exist sequences of terms

$$\begin{aligned}&\vec{p}^{+}_{\psi} : \mathcal{W}_{\Gamma} \to \mathcal{W}_{\psi},\\ &\vec{p}^{-}_{\varphi_1} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\varphi_1},\\ &\dots,\\ &\vec{p}^{-}_{\varphi_n} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\varphi_n},\end{aligned}$$

of system **T** terms such that

$$\varphi_{1D}(\vec{x}_1, \vec{p}^{-}_{\varphi_1}(\vec{x}, \vec{v})), \dots, \varphi_{nD}(\vec{x}_n, \vec{p}^{-}_{\varphi_n}(\vec{x}, \vec{v})) \vdash_{\mathbf{HA}+\mathbf{T}} \psi_D(\vec{p}^{+}_{\psi}(\vec{x}), \vec{v})$$

for any choice of sequences $\vec{x} : \mathcal{W}_{\Gamma} = \vec{x}_1 : \mathcal{W}_{\varphi_1}, \dots, \vec{x}_n : \mathcal{W}_{\varphi_n}$ and $\vec{v} : \mathcal{C}_{\psi}$.

Note that $\vec{p}^{+}_{\psi}, \vec{p}^{-}_{\varphi_1}, \dots, \vec{p}^{-}_{\varphi_n}$ are precisely the components of a sequence of terms of type $\mathcal{W}_{\Gamma \vdash \psi}$. When we translate sequents in the proof of this theorem we will use the superscript $^{+}$ to mark components that are of the witness to witness part of a witness of the sequent while the superscript $^{-}$ will mark components of the witness to counter to counter part. In particular we let $\vec{p}^{-}_{\Gamma} = \vec{p}^{-}_{\varphi_1}, \dots, \vec{p}^{-}_{\varphi_n}$ denote the sequence of terms producing the counters for an environment $\Gamma = \varphi_1, \dots, \varphi_n$.

For the proof of Theorem 3.7 we need a few definitions and lemmas.

Definition 3.8. For each type τ of **T** we define its *dummy term* $\varnothing_{\tau}$ inductively as follows:

- $\varnothing_{\mathbf{N}} = 0$,
- $\varnothing_{\sigma \to \tau} = \lambda x.\varnothing_{\tau}$ for some fresh variable x.

The dummy terms ensure that each witness and counter type is inhabited. They are useful when the structure of witness types of **HA** formulas requires an inactive placeholder term.

Notation 3.9. Dummy terms can naturally be extended to sequences:

$$\varnothing_{\vec{\tau}} = \varnothing_{\tau_1}, \dots, \varnothing_{\tau_n}.$$

Definition 3.10. For each formula φ of **HA** define the function

$$\mathbf{Decide}_{\varphi} : \mathcal{W}_{\varphi} \to \mathcal{C}_{\varphi} \to \mathbf{N}$$

as follows:

- $\mathbf{Decide}_{\perp} = \mathbf{S}(0)$
- $\mathbf{Decide}_{n=m} = \lambda \vec{x}\vec{y}.|n - m|$
- $\mathbf{Decide}_{\varphi \wedge \psi} = \lambda \vec{x}\vec{u}\vec{y}\vec{v}.(\mathbf{Decide}_{\varphi}(\vec{x}, \vec{y}) + \mathbf{Decide}_{\psi}(\vec{u}, \vec{v}))$
- $\mathbf{Decide}_{\varphi \vee \psi} = \lambda z\vec{x}\vec{u}\vec{y}\vec{v}.\mathbf{R}(\mathbf{Decide}_{\varphi}(\vec{x}, \vec{y}), \lambda pq.\mathbf{Decide}_{\psi}(\vec{u}, \vec{v}), z)$

- $\mathbf{Decide}_{\varphi\to\psi} = \lambda \vec{U}\vec{Y}\vec{x}\vec{v}.\,\mathrm{SIGN}(\mathbf{Decide}_\varphi(\vec{x}, \vec{Y}(\vec{x}, \vec{v}))) \cdot \mathbf{Decide}_\psi(\vec{U}(\vec{x}), \vec{v}))$
- $\mathbf{Decide}_{\forall z\varphi} = \lambda \vec{X}z\vec{y}.\mathbf{Decide}_\varphi(\vec{X}(z), \vec{y})$
- $\mathbf{Decide}_{\exists z\varphi} = \lambda z\vec{x}\vec{y}.\mathbf{Decide}_\varphi((z, \vec{x}), \vec{y})$.

The definitions of the functions SIGN and $|n - m|$ can be found in Example 2.19.

Lemma 3.11. If φ is a formula of **HA** and $\vec{x} : \mathcal{W}_\varphi, \vec{y} : \mathcal{C}_\varphi$, then

$$\vdash_{\mathbf{HA+T}} (\mathbf{Decide}_\varphi(\vec{x}, \vec{y}) = 0 \wedge \varphi_D(\vec{x}, \vec{y})) \vee (\mathbf{Decide}_\varphi(\vec{x}, \vec{y}) \neq 0 \wedge \neg\varphi_D(\vec{x}, \vec{y})).$$

Proof. This is proved by induction on the length of φ and by checking on case by case basis when needed. The details are both tedious and obvious so we skip writing up the whole proof. □

This lemma shows us that for any formula φ of **HA**, the formula φ_D is decidable. We can therefore use classical logic to reason about φ_D. More importantly it allows us to define the following very important function.

Definition 3.12. For each formula φ of **HA** define the function

$$\mathbf{Merge}_\varphi : \mathcal{C}_\varphi \to \mathcal{C}_\varphi \to \mathcal{W}_\varphi \to \mathcal{C}_\varphi$$

as follows:

Let $\vec{y}_1 = y_{1,1}, \ldots, y_{1,n}$ and $\vec{y}_2 = y_{2,1}, \ldots, y_{2,n}$ be sequences of the type $\mathcal{C}_\varphi$. We first define $\mathbf{Merge}^k_\varphi$ for $k = 1, \ldots, n$ as

$$\mathbf{Merge}^k_\varphi = \lambda \vec{y}_1\vec{y}_2\vec{x}.\mathbf{R}(y_{2,k}, \lambda pq.y_{1,k}, \mathbf{Decide}_\varphi(\vec{x}, \vec{y}_1)).$$

Then we define $\mathbf{Merge}_\varphi$ as

$$\mathbf{Merge}_\varphi = \lambda \vec{y}_1\vec{y}_2\vec{x}.\mathbf{Merge}^1_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}), \ldots, \mathbf{Merge}^n_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}).$$

While it might seem unnecessarily complex to do the definition like this it is in fact necessary to get around the restriction that **R** takes terms as arguments but not sequences of terms. While the definition might not be particularly transparent a little inspection reveals that

$$\mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}) = \begin{cases} \vec{y}_1 \text{ if } \mathbf{Decide}_\varphi(x, \vec{y}_1) \neq 0 \\ \vec{y}_2 \text{ if } \mathbf{Decide}_\varphi(x, \vec{y}_1) = 0. \end{cases}$$

In other words $\mathbf{Merge}_\varphi$ has the value $\vec{y}_2$ if $\vdash \varphi_D(\vec{x}, \vec{y}_1)$ and the value $\vec{y}_1$ otherwise. The purpose of the function is make it possible to create one counter by merging two counters. This is summed up in following lemma.

Lemma 3.13. If φ is a formula of **HA** and $\vec{x} : \mathcal{W}_\varphi, \vec{y}_1 : \mathcal{C}_\varphi, \vec{y}_2 : \mathcal{C}_\varphi$, then

$$\vdash_{\mathbf{HA+T}} \varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x})) \leftrightarrow \varphi_D(\vec{x}, \vec{y}_1) \wedge \varphi_D(\vec{x}, \vec{y}_2).$$

Proof. We first prove $\vdash \varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x})) \to \varphi_D(\vec{x}, \vec{y}_1) \wedge \varphi_D(\vec{x}, \vec{y}_2)$. We already know from Lemma 3.11 that φ_D is decidable so we can assume that either $\vdash \varphi_D(\vec{x}, \vec{y}_1)$ or $\vdash \neg\varphi_D(\vec{x}, \vec{y}_1)$. If $\vdash \varphi_D(\vec{x}, \vec{y}_1)$, then $\mathbf{Decide}_\varphi(\vec{x}, \vec{y}_1) = 0$ and $\mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}) = \vec{y}_2$ so $\varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x})) \equiv \varphi_D(\vec{x}, \vec{y}_2)$. Then both conjuncts of the conclusion are true whenever the hypothesis is true, proving the implication. If $\vdash \neg\varphi_D(\vec{x}, \vec{y}_1)$, then $\mathbf{Decide}_\varphi(\vec{x}, \vec{y}_1) \neq 0$ and $\mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}) = \vec{y}_1$ giving us $\varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x})) \equiv \varphi_D(\vec{x}, \vec{y}_1)$. Then the hypothesis becomes false and this proves the implication.

We then prove $\vdash \varphi_D(\vec{x}, \vec{y}_1) \wedge \varphi_D(\vec{x}, \vec{y}_2) \to \varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x}))$. We already know that if $\vdash \varphi_D(\vec{x}, \vec{y}_1)$ then $\varphi_D(\vec{x}, \mathbf{Merge}_\varphi(\vec{y}_1, \vec{y}_2, \vec{x})) \equiv \varphi_D(\vec{x}, \vec{y}_2)$. So it is clear that if we can prove the hypothesis, the conclusion follows and this concludes the proof. □

This makes it clear that if a witness is to beat a merger of two counters it must be able beat both of the counters.

Notation 3.14. It is possible to extend the definition of **Merge** to a sequence $\Gamma = \varphi_1, \ldots, \varphi_n$ of formulas as follows:

$$\mathbf{Merge}_\Gamma : \mathcal{C}_\Gamma \to \mathcal{C}_\Gamma \to \mathcal{W}_\Gamma \to \mathcal{C}_\Gamma$$
$$\mathbf{Merge}_\Gamma(\vec{y}_1, \vec{y}_2, \vec{x}) = \mathbf{Merge}_{\varphi_1}(\vec{y}_{1,1}, \vec{y}_{2,1}, \vec{x}_1), \ldots, \mathbf{Merge}_{\varphi_n}(\vec{y}_{1,n}, \vec{y}_{2,n}, \vec{x}_n).$$

We are now ready to prove Theorem 3.7.

Proof of Theorem 3.7. This theorem is proved by induction on the length of deductions in **HA**, that is we assume as an induction hypothesis that it has been shown to hold for the premises of each rule and show that in that case it also holds for the conclusion. We will use the letter p for the witnesses we construct for conclusions and the letters q, r and s for witnesses we extract from the translation of premises.

We begin with the structural rules.

1. $\dfrac{}{\Gamma, \varphi \vdash \varphi}$ Axiom

This rule has no premises so we simply have to extract the witnesses directly from the Dialectica translation of the conclusion. This is very simple for $\vec{p}^{\,+}_\varphi$:

$$\vec{p}^{\,+}_\varphi : \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\varphi$$
$$\vec{p}^{\,+}_\varphi = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{u} : \mathcal{W}_\varphi).\vec{u}$$

as well as for $\vec{p}^{\,-}_\varphi$:

$$\vec{p}^{\,-}_\varphi : \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\varphi \to \mathcal{C}_\varphi$$
$$\vec{p}^{\,-}_\varphi = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{u} : \mathcal{W}_\varphi)(\vec{v} : \mathcal{C}_\varphi).\vec{v}.$$

It is slightly more complicated to construct the sequence $\vec{p}_{\Gamma}^{-}$ since there is no obvious way of extracting it from the translation of the sequent. Here the dummy terms come in handy:

$$\begin{aligned}\vec{p}_{\Gamma}^{-} &: \mathcal{W}_{\Gamma} \to \mathcal{W}_{\varphi} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma}\\ \vec{p}_{\Gamma}^{-} &= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{u} : \mathcal{W}_{\varphi})(\vec{v} : \mathcal{C}_{\varphi}).\varnothing_{\mathcal{C}_{\Gamma}}.\end{aligned}$$

We now have to show that these terms actually witness the Dialectica-translation of the sequent in **HA+T**, that is:

$$\Gamma_D(\vec{x}, \vec{p}_{\Gamma}^{-}(\vec{x}, \vec{u}, \vec{v})), \varphi_D(\vec{u}, \vec{p}_{\varphi}^{-}(\vec{x}, \vec{u}, \vec{v})) \vdash \varphi_D(\vec{p}_{\varphi}^{+}(\vec{x}, \vec{u}), \vec{v}).$$

A simple unfolding of the definitions of each of our witnesses gives us

$$\Gamma_D(\vec{x}, \varnothing_{\mathcal{C}_{\Gamma}}), \varphi_D(\vec{u}, \vec{v}) \vdash \varphi_D(\vec{u}, \vec{v})$$

which clearly holds in **HA+T**.

2. $\dfrac{\Gamma \vdash \varphi}{\Gamma, \psi \vdash \varphi}$ Weakening

A translation of the premise gives us witnesses of the following types:

$$\begin{aligned}\vec{q}_{\varphi}^{+} &: \mathcal{W}_{\Gamma} \to \mathcal{W}_{\varphi}\\ \vec{q}_{\Gamma}^{-} &: \mathcal{W}_{\Gamma} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma}\end{aligned}$$

and an induction hypothesis:

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}_{\varphi}^{+}(\vec{x}), \vec{v}).$$

We simply construct $\vec{p}_{\varphi}^{+}$ and $\vec{p}_{\Gamma}^{-}$ using the witnesses extracted from the premise:

$$\begin{aligned}\vec{p}_{\varphi}^{+} &: \mathcal{W}_{\Gamma} \to \mathcal{W}_{\psi} \to \mathcal{W}_{\varphi}\\ \vec{p}_{\varphi}^{+} &= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{w} : \mathcal{W}_{\psi}).\vec{q}_{\varphi}^{+}(\vec{x})\\ \vec{p}_{\Gamma}^{-} &: \mathcal{W}_{\Gamma} \to \mathcal{W}_{\psi} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma}.\\ \vec{p}_{\varphi}^{+} &= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{w} : \mathcal{W}_{\psi})(\vec{v} : \mathcal{C}_{\varphi}).\vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v}).\end{aligned}$$

We then use a dummy term to construct the remaining witness:

$$\begin{aligned}\vec{p}_{\psi}^{-} &: \mathcal{W}_{\Gamma} \to \mathcal{W}_{\psi} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\psi}\\ \vec{p}_{\Gamma}^{-} &= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{u} : \mathcal{W}_{\psi})(\vec{v} : \mathcal{C}_{\varphi}).\varnothing_{\mathcal{C}_{\psi}}.\end{aligned}$$

The rest is easy since

$$\Gamma_D(\vec{x}, \vec{p}_{\Gamma}^{-}(\vec{x}, \vec{u}, \vec{v})), \psi_D(\vec{w}, \vec{p}_{\psi}^{-}(\vec{x}, \vec{w}, \vec{v})) \vdash \varphi_D(\vec{p}_{\varphi}^{+}(\vec{x}, \vec{w}), \vec{v}).$$

unfolds to

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v})), \psi_D(\vec{w}, \varnothing_{\mathcal{C}_{\psi}}) \vdash \varphi_D(\vec{q}_{\varphi}^{+}(\vec{x}), \vec{v}).$$

and this clearly follows from the induction hypothesis and the weakening rule of **HA+T**.

Next are the rules of propositional logic

3. $\dfrac{\Gamma \vdash \varphi \qquad \Gamma \vdash \psi}{\Gamma \vdash \varphi \wedge \psi} \wedge\mathrm{I}$

This is the first rule in which we have two premises. There are certain things that we must keep in mind any time we deal with rules with multiple premises. We begin as usual though, by unfolding the witness types of the premises:

$$\begin{aligned}\vec{q}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi\\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma\\ \vec{r}^{\,+}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi\\ \vec{r}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma.\end{aligned}$$

We then get the following two induction hypotheses:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{v})$$

$$\Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w})) \vdash \psi_D(\vec{r}^{\,+}_{\psi}(\vec{x}), \vec{w})$$

It is obvious how these witnesses are used to define $\vec{p}^{\,+}_{\varphi\wedge\psi}$:

$$\begin{aligned}&\vec{p}^{\,+}_{\varphi\wedge\psi} : \mathcal{W}_\Gamma \to \mathcal{W}_\varphi, \mathcal{W}_\Gamma \to \mathcal{W}_\psi\\ &\vec{p}^{\,+}_{\varphi\wedge\psi} = \vec{q}^{\,+}_{\varphi}, \vec{r}^{\,+}_{\psi}.\end{aligned}$$

It is a bit more difficult to define $\vec{p}^{\,-}_{\varphi\wedge\psi}$. We begin by unfolding its type:

$$\begin{aligned}\vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_{\varphi\wedge\psi} \to \mathcal{C}_\Gamma\\ &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma.\end{aligned}$$

There are two ways of extracting a sequence of this type from the witnesses of the premises, one from $\vec{q}^{\,-}_{\Gamma}$ and one from $\vec{r}^{\,-}_{\Gamma}$. The problem is that we will need both of them to get access to both of the induction hypotheses. Here the **Merge** function comes in handy. We use it to merge the counters produced by each premise:

$$\vec{p}^{\,-}_{\Gamma} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{v} : \mathcal{C}_\varphi)(\vec{w} : \mathcal{C}_\psi).\mathbf{Merge}_\Gamma(\vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v}), \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w}), \vec{x}).$$

We have to show that

$$\Gamma_D(\vec{x}, \vec{p}^{\,-}_{\Gamma}(\vec{x}, \vec{v}, \vec{w})) \vdash \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{v}) \wedge \psi_D(\vec{r}^{\,+}_{\varphi}(\vec{x}), \vec{w}).$$

This unfolds to

$$\Gamma_D(\vec{x}, \mathbf{Merge}_\Gamma(\vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v}), \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w}), \vec{x})) \vdash \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{v}) \wedge \psi_D(\vec{r}^{\,+}_{\varphi}(\vec{x}), \vec{w})$$

and Lemma 3.13 tells us that

$$\Gamma_D(\vec{x}, \mathbf{Merge}_\Gamma(\vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v}), \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w}), \vec{x})) \leftrightarrow \Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v})) \wedge \Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w}))$$

where $\Gamma_D \wedge \Gamma_D$ denotes the pointwise conjuction of the terms of each sequence. It therefore all boils down to showing that

$$\Gamma_D(\vec{x}, \vec{q}_\Gamma^-(\vec{x}, \vec{v})), \Gamma_D(\vec{x}, \vec{r}_\Gamma^-(\vec{x}, \vec{w})) \vdash \varphi_D(\vec{q}_\varphi^+(\vec{x}), \vec{v}) \wedge \psi_D(\vec{r}_\varphi^+(\vec{x}), \vec{w})$$

which clearly follows from the two induction hypotheses and the $\wedge$I-rule in **HA+T**. We will see the **Merge** function used like this every time we need to show the soundness of rules with more than one premise.

4. $$\frac{\Gamma \vdash \varphi \wedge \psi}{\Gamma \vdash \varphi} \wedge \mathrm{E}_1$$

We give the proof for this rule and skip the analogous proof for $\wedge$E$_2$. We begin by unfolding the types of the witnesses for the premise:

$$\begin{aligned} \vec{q}^{\,+}_{\varphi\wedge\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi, \mathcal{W}_\Gamma \to \mathcal{W}_\psi \\ \vec{q}_\Gamma^- &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \end{aligned}$$

Let us call the first and second component of $\vec{q}^{\,+}_{\varphi\wedge\psi}$, $\vec{q}_\varphi^+$ and $\vec{q}_\psi^+$ respectively. Then the induction hypothesis is:

$$\Gamma_D(\vec{x}, \vec{q}_\Gamma^-(\vec{x}, \vec{v}, \vec{w}) \vdash \varphi_D(\vec{q}_\varphi^+(\vec{x}), \vec{v}) \wedge \psi_D(\vec{q}_\psi^+(\vec{x}), \vec{w})$$

Defining $\vec{p}_\varphi^+$ is simple:

$$\begin{aligned} \vec{p}_\varphi^+ &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{p}_\varphi^+ &= \vec{q}_\varphi^+. \end{aligned}$$

The term $\vec{q}_\Gamma^-$ is almost of the right type for $\vec{p}_\Gamma^-$ the only difference is that the term takes a $\mathcal{C}_\psi$ term as an argument. We get around this by using a dummy term:

$$\begin{aligned} \vec{p}_\Gamma^- &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \\ \vec{p}_\Gamma^- &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{v} : \mathcal{C}_\varphi).\vec{q}_\varphi^+(\vec{x}, \vec{v}, \varnothing_{\mathcal{C}_\psi}). \end{aligned}$$

Then we just have to prove that

$$\Gamma_D(\vec{x}, \vec{p}_\Gamma^-(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{p}_\varphi^+(\vec{x}), \vec{v})$$

which boils down to proving

$$\Gamma_D(\vec{x}, \vec{q}_\Gamma^-(\vec{x}, \vec{v}, \varnothing_{\mathcal{C}_\psi})) \vdash \varphi_D(\vec{q}_\varphi^+(\vec{x}), \vec{v})$$

but this follows from the induction hypothesis and the the $\wedge$I$_1$-rule in **HA+T**.

5. $$\frac{\Gamma \vdash \varphi}{\Gamma \vdash \varphi \vee \psi} \vee \mathrm{I}_1$$

As with the elimination rules for conjunction we only give the proof for this rule and skip the analogous one for $\vee\mathrm{I}_2$. The witness types of the premise are the following:

$$\begin{aligned} \vec{q}_{\varphi}^{\,+} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{q}_{\Gamma}^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \end{aligned}$$

and this gives the induction hypothesis:

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{\,-}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}_{\varphi}^{\,+}(\vec{x}), \vec{v})$$

The type of $\vec{p}_{\varphi\vee\psi}^{\,+}$ is

$$\begin{aligned} \vec{p}_{\varphi\vee\psi}^{\,+} &: \mathcal{W}_\Gamma \to \mathbf{N}, \mathcal{W}_\varphi, \mathcal{W}_\psi \\ &: \mathcal{W}_\Gamma \to \mathbf{N}, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi, \mathcal{W}_\Gamma \to \mathcal{W}_\psi. \end{aligned}$$

The three components of $\vec{p}_{\varphi\vee\psi}^{\,+}$ we denote by $\vec{p}_{z}^{\,+}$, $\vec{p}_{\varphi}^{\,+}$ and $\vec{p}_{\psi}^{\,+}$ respectively. Since we can extract an effective witness of first disjunct, φ, from the induction hypothesis the natural number is supposed to be 0. For the witness of ψ we use a dummy term. So we have:

$$\begin{aligned} \vec{p}_{z}^{\,+} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).0 \\ \vec{p}_{\varphi}^{\,+} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).\vec{q}_{\varphi}^{\,+}(\vec{x}) \\ \vec{p}_{\psi}^{\,+} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).\varnothing_{\mathcal{W}_\psi}. \end{aligned}$$

The $\vec{p}_{\Gamma}^{\,-}$ part its easy to extract from $\vec{q}_{\Gamma}^{\,-}$:

$$\begin{aligned} \vec{p}_{\Gamma}^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \\ \vec{p}_{\Gamma}^{\,-} &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{v} : \mathcal{C}_\varphi)(\vec{w} : \mathcal{C}_\psi).\vec{q}_{\Gamma}^{\,-}(\vec{x}, \vec{v}). \end{aligned}$$

We then show that the following holds

$$\begin{aligned} &\Gamma_D(\vec{x}, \vec{p}_{\Gamma}^{\,-}(\vec{x}, \vec{v}, \vec{w})) \\ &\qquad \vdash ((\lambda\vec{x}.0)\vec{x} = 0 \wedge \varphi_D(\vec{q}_{\varphi}^{\,+}(\vec{x})), \vec{v}) \vee ((\lambda\vec{x}.0)\vec{x} = 1 \wedge \psi_D((\lambda\vec{x}.\varnothing_{\mathcal{W}_\psi})\vec{x}, \vec{w})) \end{aligned}$$

this can of course be reduced to

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{\,-}(\vec{x}, \vec{v})) \vdash (0 = 0 \wedge \varphi_D(\vec{q}_{\varphi}^{\,+}(\vec{x}), \vec{v}) \vee (0 = 1 \wedge \psi_D(\varnothing_{\mathcal{W}_\psi}, \vec{w})).$$

Since $0 = 0$ always holds this clearly follows from the induction hypothesis and the $\vee\mathrm{I}_1$-rule in **HA+T**.

6. $$\frac{\Gamma \vdash \varphi \vee \psi \qquad \Gamma, \varphi \vdash \theta \qquad \Gamma, \psi \vdash \theta}{\Gamma \vdash \theta}\vee\mathrm{E}$$

This one is quite difficult. The first of the premises give us the following witness types:

$$\begin{aligned}
\vec{q}^{\,+}_{\varphi\vee\psi} &: \mathcal{W}_\Gamma \to \mathbf{N}, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi, \mathcal{W}_\Gamma \to \mathcal{W}_\psi \\
\vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma
\end{aligned}$$

Let the respective parts of $\vec{q}^{\,+}_{\varphi\vee\psi}$ be called $\vec{q}^{\,+}_{z}$, $\vec{q}^{\,+}_{\varphi}$ and $\vec{q}^{\,+}_{\psi}$. Then the first induction hypothesis is:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{v}, \vec{w})) \vdash (\vec{q}^{\,+}_{z}(\vec{x}) = 0 \wedge \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{v})) \vee (\vec{q}^{\,+}_{z}(\vec{x}) = 1 \wedge \psi_D(\vec{q}^{\,+}_{\psi}(\vec{x}), \vec{w})).$$

The other two premises give us witnesses with the following types:

$$\begin{aligned}
\vec{r}^{\,+}_{\theta} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\theta \\
\vec{r}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
\vec{r}^{\,-}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\theta \to \mathcal{C}_\varphi \\
\vec{s}^{\,+}_{\theta} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi \to \mathcal{W}_\theta \\
\vec{s}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
\vec{s}^{\,-}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi \to \mathcal{C}_\theta \to \mathcal{C}_\psi
\end{aligned}$$

and the following two induction hypotheses:

$$\Gamma_D(\vec{x}, \vec{s}^{\,-}_{\Gamma}(\vec{x}, \vec{u}, \vec{t})), \varphi_D(\vec{u}, \vec{s}^{\,-}_{\varphi}(\vec{x}, \vec{u}, \vec{t})) \vdash \theta_D(\vec{s}^{\,+}_{\theta}(\vec{x}, \vec{u}), \vec{t})$$

$$\Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{y}, \vec{t})), \psi_D(\vec{y}, \vec{r}^{\,-}_{\psi}(\vec{x}, \vec{y}, \vec{t})) \vdash \theta_D(\vec{r}^{\,+}_{\theta}(\vec{x}, \vec{y}), \vec{t}).$$

The types of the witnesses we are seeking are:

$$\begin{aligned}
\vec{p}^{\,+}_{\theta} &: \mathcal{W}_\Gamma \to \mathcal{W}_\theta \\
\vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma.
\end{aligned}$$

We begin by constructing $\vec{p}^{\,+}_{\theta}$ as follows

$$\vec{p}^{\,+}_{\theta} = \lambda\vec{x} : \mathcal{W}_\Gamma.\mathbf{R}(\vec{r}^{\,+}_{\theta}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x})), \lambda cd.\vec{s}^{\,+}_{\theta}(\vec{x}, \vec{q}^{\,+}_{\psi}(\vec{x})), \vec{q}^{\,+}_{z}(\vec{x})).$$

This is of course a blatant abuse of notation, since **R** only takes three arguments, not three sequences of arguments. Such a function could of course be defined using similar tricks as we used to define **Merge**. We however skip showing the details of such a definition to keep an already long proof from becoming any longer.

For the construction of $\vec{p}^{\,-}_{\Gamma}$ we have three different ways to get to a sequence of the right type, one from each premise:

$$\begin{aligned}
&\vec{p}_q^{\,-} : \mathcal{W}_\Gamma \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
&\vec{p}_q^{\,-} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{t} : \mathcal{C}_\theta).\vec{q}_\Gamma^{\,-}(\vec{x}, \vec{r}_\varphi^{\,-}(\vec{x}, \vec{q}_\varphi^{\,+}(\vec{x}), \vec{t}), \vec{s}_\psi^{\,-}(\vec{x}, \vec{q}_\psi^{\,+}(\vec{x}), \vec{t})) \\
&\vec{p}_r^{\,-} : \mathcal{W}_\Gamma \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
&\vec{p}_r^{\,-} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{t} : \mathcal{C}_\theta).\vec{r}_\Gamma^{\,-}(\vec{x}, \vec{q}_\varphi^{\,+}(\vec{x}), \vec{t}) \\
&\vec{p}_s^{\,-} : \mathcal{W}_\Gamma \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
&\vec{p}_s^{\,-} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{t} : \mathcal{C}_\theta).\vec{s}_\Gamma^{\,-}(\vec{x}, \vec{q}_\psi^{\,+}(\vec{x}), \vec{t}).
\end{aligned}$$

We then have to merge all of these counters:

$$\begin{aligned}
&\vec{p}_\Gamma^{\,-} : \mathcal{W}_\Gamma \to \mathcal{C}_\theta \to \mathcal{C}_\Gamma \\
&\vec{p}_\Gamma^{\,-} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{t} : \mathcal{C}_\theta).\mathbf{Merge}_\Gamma[\vec{p}_q^{\,-}(\vec{x}, \vec{t}), \mathbf{R}(\vec{p}_r^{\,-}(\vec{x}, \vec{t}), \lambda cd.\vec{p}_s^{\,-}(\vec{x}, \vec{t}), \vec{q}_z^{\,+}(\vec{x})), \vec{x}].
\end{aligned}$$

We have here again the same abuse of the notation $\mathbf{R}$ as before. This function merges the counter extracted from the first premise and one of the counters extracted from the other premises, depending on which one of the disjuncts is proved by the first premise. Now we have to use all of this to show that

$$\Gamma_D(\vec{x}, \vec{p}_\Gamma^{\,-}(\vec{x}, \vec{t})) \vdash \theta_D(\vec{p}_\varphi^{\,+}(\vec{x}), \vec{t}).$$

We begin by noting that from the induction hypotheses we can easily see that either $\vec{q}_z^{\,+}(\vec{x}) = 0$ or $\vec{q}_z^{\,+}(\vec{x}) = 1$. If $\vec{q}_z^{\,+}(\vec{x}) = 0$, then an unfolding of $\vec{p}_\Gamma^{\,-}(\vec{x}, \vec{t})$ and $\vec{p}_\varphi^{\,+}(\vec{x})$ gives

$$\Gamma_D(\vec{x}, \mathbf{Merge}_\Gamma(\vec{p}_q^{\,-}(\vec{x}, \vec{t}), \vec{p}_r^{\,-}(\vec{x}, \vec{t}), \vec{x})) \vdash \theta_D(\vec{r}_\theta^{\,+}(\vec{x}, \vec{q}_\varphi^{\,+}(\vec{x})), \vec{t}).$$

As we already pointed out Lemma 3.13 shows that this boils down to

$$\Gamma_D(\vec{x}, \vec{p}_q^{\,-}(\vec{x}, \vec{t})), \Gamma_D(\vec{x}, \vec{p}_r^{\,-}(\vec{x}, \vec{t})) \vdash \theta_D(\vec{r}_\theta^{\,+}(\vec{x}, \vec{q}_\varphi^{\,+}(\vec{x})), \vec{t}). \qquad (*)$$

From the first induction hypothesis, the fact that $\vec{p}_q^{\,-}(\vec{x}, \vec{t})$ is of the form $\vec{q}_\Gamma^{\,-}(\vec{x}, \vec{v}, \vec{w})$ and the assumption $\vec{q}_z^{\,+}(\vec{x}) = 0$ we can deduce that

$$\Gamma_D(\vec{x}, \vec{p}_q^{\,-}(\vec{x}, \vec{t})) \vdash \varphi_D(\vec{q}_\varphi^{\,+}(\vec{x}), \vec{v})$$

and the second induction hypothesis reads:

$$\Gamma_D(\vec{x}, \vec{s}_\Gamma^{\,-}(\vec{x}, \vec{u}, \vec{t})), \varphi_D(\vec{u}, \vec{s}_\varphi^{\,-}(\vec{x}, \vec{u}, \vec{t})) \vdash \theta_D(\vec{s}_\theta^{\,+}(\vec{x}, \vec{u}), \vec{t}).$$

From these two facts it is easy to show that $(*)$ holds.

The proof is analogous if $\vec{q}_z^{\,+}(\vec{x}) = 1$.

7. $\dfrac{\Gamma, \varphi \vdash \psi}{\Gamma \vdash \varphi \to \psi} \to\mathrm{I}$

We unfold the witness types of the premise:

$$\begin{aligned}\vec{q}^{\,+}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\psi \\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \\ \vec{q}^{\,-}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\varphi\end{aligned}$$

and the induction hypothesis:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{u}, \vec{w})), \varphi_D(\vec{x}, \vec{q}^{\,-}_{\varphi}(\vec{x}, \vec{u}, \vec{w})) \vdash \psi_D(\vec{q}^{\,+}_{\psi}(\vec{x}, \vec{u}), \vec{w}).$$

The witness types of the conclusion are as follows:

$$\begin{aligned}\vec{p}^{\,+}_{\varphi\to\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\psi, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\varphi \\ \vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma.\end{aligned}$$

These have the same types as the witnesses of the premise, so we simply put:

$$\begin{aligned}\vec{p}^{\,+}_{\varphi\to\psi} &= \vec{q}^{\,+}_{\psi}, \vec{q}^{\,-}_{\varphi} \\ \vec{p}^{\,-}_{\Gamma} &= \vec{q}^{\,-}_{\Gamma}.\end{aligned}$$

We then have to show that

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{u}, \vec{w}) \vdash \varphi_D(\vec{x}, \vec{q}^{\,-}_{\varphi}(\vec{x}, \vec{u}, \vec{w})) \to \psi_D(\vec{q}^{\,+}_{\psi}(\vec{x}, \vec{u}), \vec{w})$$

but this follows directly from the induction hypothesis and the →I-rule.

8. $\dfrac{\Gamma \vdash \varphi \to \psi \qquad \Gamma \vdash \varphi}{\Gamma \vdash \psi} \to\mathrm{E}$

We begin as usual. From the premises we get the following witnesses:

$$\begin{aligned}\vec{q}^{\,+}_{\varphi\to\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\psi, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\varphi \\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \\ \vec{r}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{r}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma.\end{aligned}$$

The first and second component of $\vec{q}^{\,+}_{\varphi\to\psi}$ we denote by $\vec{q}^{\,+}_{\psi}$ and $\vec{q}^{\,-}_{\varphi}$ respectively. Then we get the two induction hypotheses:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{u}, \vec{w})) \vdash \varphi_D(\vec{x}, \vec{q}^{\,-}_{\varphi}(\vec{x}, \vec{u}, \vec{w})) \to \psi_D(\vec{q}^{\,+}_{\psi}(\vec{x}, \vec{u}), \vec{w})$$

$$\Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{w})) \vdash \varphi_D(\vec{r}^{\,+}_{\varphi}(\vec{x}), \vec{w})$$

The witnesses of the conclusion have the following types

$$\begin{aligned}\vec{p}^{\,+}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi \\ \vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma.\end{aligned}$$

It is easy to construct $\vec{p}_\psi^+$:

$$\vec{p}_\psi^+ = \lambda(\vec{x} : \mathcal{W}_\Gamma)\vec{q}_\psi^+(\vec{x}, \vec{r}_\varphi^+(\vec{x})).$$

We have two premises so as usual, for $\vec{p}_\Gamma^-$ we have to construct two sequences of the right type, one from each premise, and then merge them. We construct these two sequences as follows:

$$\begin{aligned}
\vec{p}_q^- &: \mathcal{W}_\Gamma \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \\
\vec{p}_q^- &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{w} : \mathcal{C}_\Gamma).\vec{q}_\Gamma^-(\vec{x}, \vec{r}_\varphi^+(\vec{x}), \vec{w}) \\
\vec{p}_r^- &: \mathcal{W}_\Gamma \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma \\
\vec{p}_r^- &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{w} : \mathcal{C}_\Gamma).\vec{r}_\Gamma^-(\vec{x}, \vec{q}_\varphi^-(\vec{x}, \vec{w}), \vec{w})
\end{aligned}$$

and then we merge them:

$$\vec{p}_\Gamma^- = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{w} : \mathcal{C}_\Gamma).\mathbf{Merge}_\Gamma(\vec{p}_q^-(\vec{x}, \vec{w}), \vec{p}_r^-(\vec{x}, \vec{w}), \vec{x}).$$

We must show that

$$\Gamma_D(\vec{x}, \vec{p}_\Gamma^-(\vec{x}, \vec{w})) \vdash \psi_D(\vec{p}_\psi^+(\vec{x}), \vec{w}).$$

This unfolds to

$$\Gamma_D(\vec{x}, \mathbf{Merge}_\Gamma(\vec{p}_q^-(\vec{x}, \vec{w}), \vec{p}_r^-(\vec{x}, \vec{w}), \vec{x})) \vdash \psi_D(\vec{q}_\psi^+(\vec{x}, \vec{r}_\varphi^+(\vec{x})), \vec{w})$$

which by Lemma 3.13 is equivalent to

$$\Gamma_D(\vec{x}, \vec{p}_q^-(\vec{x}, \vec{w})), \Gamma_D(\vec{x}, \vec{p}_r^-(\vec{x}, \vec{w})) \vdash \psi_D(\vec{q}_\psi^+(\vec{x}, \vec{r}_\varphi^+(\vec{x})), \vec{w}).$$

We recall that $\vec{p}_q^-(\vec{x}, \vec{w}) = \vec{q}_\Gamma^-(\vec{x}, \vec{r}_\varphi^+(\vec{x}), \vec{w})$ and $\vec{p}_r^-(\vec{x}, \vec{w}) = \vec{r}_\Gamma^-(\vec{x}, \vec{q}_\varphi^-(\vec{x}, \vec{w}), \vec{w})$, so we can use the two induction hypotheses and the →E-rule in **HA+T** to show that this holds.

9. $\dfrac{\Gamma \vdash \bot}{\Gamma \vdash \varphi}\ \bot\mathrm{E}$

A translation of the witness types of the premise gives us

$$\begin{aligned}
\vec{q}_\bot^+ &: \mathcal{W}_\Gamma \to \mathcal{W}_\bot \\
&: \emptyset \\
\vec{q}_\Gamma^- &: \mathcal{W}_\Gamma \to \mathcal{C}_\bot \to \mathcal{C}_\Gamma \\
&: \mathcal{W}_\Gamma \to \mathcal{C}_\Gamma
\end{aligned}$$

and the induction hypothesis is simply:

$$\Gamma_D(\vec{x}, \vec{q}_\Gamma^-(\vec{x})) \vdash \bot.$$

The rest is very simple. We construct $\vec{p}_{\varphi}^{+}$ and $\vec{p}_{\Gamma}^{-}$ as follows:

$$\begin{aligned}
&\vec{p}_{\varphi}^{+} : \mathcal{W}_{\Gamma} \to \mathcal{W}_{\varphi} \\
&\vec{p}_{\varphi}^{+} = \lambda(\vec{x} : \mathcal{W}_{\Gamma}).\varnothing_{\mathcal{W}_{\varphi}} \\
&\vec{p}_{\Gamma}^{-} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma} \\
&\vec{p}_{\Gamma}^{-} = \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{v} : \mathcal{C}_{\varphi}).\vec{q}_{\Gamma}^{-}(\vec{x}).
\end{aligned}$$

We then have to show that

$$\Gamma_D(\vec{x}, \vec{p}_{\Gamma}^{-}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{p}_{\varphi}^{+}(\vec{x}), \vec{v})$$

which unfolds to

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{-}(\vec{x})) \vdash \varphi_D(\varnothing_{\mathcal{W}_{\varphi}}, \vec{v})$$

which follows from the induction hypothesis and $\perp$E in **HA**+**T**.

Next up are the rules of first-order logic.

10. $\dfrac{\Gamma \vdash \varphi}{\Gamma \vdash \forall z \varphi}\ \forall\text{I}$, where z does not occur freely in Γ.

We unfold the witness types of the premise:

$$\begin{aligned}
&\vec{q}_{\varphi}^{+} : \mathcal{W}_{\Gamma} \to \mathcal{W}_{\varphi} \\
&\vec{q}_{\Gamma}^{-} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma}.
\end{aligned}$$

The induction hypothesis then says that

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}_{\varphi}^{+}(\vec{x}), \vec{v}).$$

We then unfold the witness types of the conclusion:

$$\begin{aligned}
\vec{p}_{\forall z\varphi}^{+} &: \mathcal{W}_{\Gamma} \to \mathbf{N} \to \mathcal{W}_{\varphi} \\
\vec{p}_{\Gamma}^{-} &: \mathcal{W}_{\Gamma} \to \mathbf{N} \to \mathcal{C}_{\varphi} \to \mathcal{C}_{\Gamma}.
\end{aligned}$$

These are easy to construct these using the witnesses from the premise:

$$\begin{aligned}
\vec{p}_{\forall z\varphi}^{+} &:= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(z : \mathbf{N}).\vec{q}_{\varphi}^{+}(\vec{x}) \\
\vec{p}_{\Gamma}^{-} &= \lambda(\vec{x} : \mathcal{W}_{\Gamma})(z : \mathbf{N})(\vec{v} : \mathcal{C}_{\varphi}).\vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v}).
\end{aligned}$$

We then have to show that

$$\Gamma_D(\vec{x}, \vec{p}_{\Gamma}^{-}(\vec{x}, z, \vec{v})) \vdash \varphi_D(\vec{p}_{\forall z\varphi}^{+}(\vec{x}, z), \vec{v})$$

but unfolding this thus just gives us the induction hypothesis:

$$\Gamma_D(\vec{x}, \vec{q}_{\Gamma}^{-}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}_{\varphi}^{+}(\vec{x}), \vec{v})$$

where the z occcurring freely in $\vec{q}_{\Gamma}^{-}$, $\vec{q}_{\varphi}$ and φ_D have been absorbed by $\vec{p}_{\Gamma}^{-}$ and $\vec{p}_{\forall z\varphi}^{+}$, so we have the desired result.

11. $\dfrac{\Gamma \vdash \forall z \varphi}{\Gamma \vdash \varphi[z := t]} \forall\text{E}$

Unfolding the witnesses of the premise gives us:

$$\begin{aligned} \vec{q}^{\,+}_{\forall z\varphi} &: \mathcal{W}_\Gamma \to \mathbf{N} \to \mathcal{W}_\varphi \\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathbf{N} \to \mathcal{C}_\Gamma \end{aligned}$$

and the induction hypothesis

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_\Gamma(\vec{x}, z, \vec{v})) \vdash \varphi_D(\vec{q}^{\,+}_{\forall z\varphi}(\vec{x}, z), \vec{v}).$$

So to construct the witnesses of the conclusion, with the following types:

$$\begin{aligned} \vec{p}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \end{aligned}$$

we only need some natural number to occupy that position for a number in the witnesses from the premises. Note that we have a term $t : \mathbf{N}$ that can be extracted from the conclusion of the rule. We use this term $t : \mathbf{N}$ to construct the witnesses as follows:

$$\begin{aligned} \vec{p}^{\,+}_{\varphi} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).\vec{q}^{\,+}_{\forall z\varphi}(\vec{x}, t) \\ \vec{q}^{\,-}_{\Gamma} &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{v} : \mathcal{C}_\varphi).\vec{q}^{\,-}_{\Gamma}(\vec{x}, t, \vec{v}). \end{aligned}$$

Then we have to show that

$$\Gamma_D(\vec{x}, \vec{p}^{\,-}_\Gamma(\vec{x}, \vec{v})) \vdash \varphi[z := t]_D(\vec{p}^{\,+}_{\varphi}(\vec{x}), \vec{v})$$

which unfolds to

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_\Gamma(\vec{x}, t, \vec{v})) \vdash \varphi[z := t]_D(\vec{q}^{\,+}_{\forall z\varphi}(\vec{x}, t), \vec{v})$$

which is just the induction hypothesis where all occurrences of z in φ_D, $\vec{q}^{\,-}_\Gamma$ and $\vec{q}^{\,+}_{\forall z\varphi}$ have been substituted for t, thus showing the desired result.

12. $\dfrac{\Gamma \vdash \varphi[z := t]}{\Gamma \vdash \exists z \varphi} \exists\text{I}$

We begin by unfolding the types of the witnesses for the premise:

$$\begin{aligned} \vec{q}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \end{aligned}$$

and the induction hypothesis

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_\Gamma(\vec{x}, \vec{v})) \vdash \varphi[z := t]_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{v}).$$

The witnesses of the conclusion have the following types:

$$\begin{aligned}\vec{p}^{+}_{\exists z\varphi} &: \mathcal{W}_\Gamma \to \mathbf{N}, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{p}^{-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathbf{N} \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma\end{aligned}$$

For $\vec{p}^{+}_{\exists z\varphi}$ we will use $\vec{p}^{+}_{z}$ and $\vec{p}^{+}_{\varphi}$ to denote the first and second component respectively. We construct these terms as follows:

$$\begin{aligned}\vec{p}^{+}_{z} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).t \\ \vec{p}^{+}_{\varphi} &= \lambda(\vec{x} : \mathcal{W}_\Gamma).\vec{q}^{+}_{\varphi}(\vec{x}) \\ \vec{p}^{-}_{\Gamma} &= \lambda(\vec{x} : \mathcal{W}_\Gamma)(z : \mathbf{N})(\vec{v} : \mathcal{C}_\varphi).\vec{q}^{-}_{\Gamma}(\vec{x}, \vec{v})\end{aligned}$$

where t comes from the premise. Then we have to show that

$$\Gamma_D(\vec{x}, \vec{p}^{-}_{\Gamma}(\vec{x}, z, \vec{v})) \vdash \varphi_D(\vec{p}^{+}_{z}(\vec{x}), \vec{p}^{+}_{\varphi}(\vec{x}), \vec{v})$$

which unfolds to

$$\Gamma_D(\vec{x}, \vec{q}^{-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi_D(t, \vec{q}^{+}_{\varphi}(\vec{x}), \vec{v})$$

which by Proposition 3.5 is equivalent to the induction hypothesis

$$\Gamma_D(\vec{x}, \vec{q}^{-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi[z := t]_D(\vec{q}^{+}_{\varphi}(\vec{x}), \vec{v})$$

giving us the desired result.

13. $\dfrac{\Gamma \vdash \exists z\varphi \qquad \Gamma, \varphi \vdash \psi}{\Gamma \vdash \psi}\ \exists\mathrm{E}$, where z does not occur freely in Γ nor in ψ.

We unfold the witness types of the premises:

$$\begin{aligned}\vec{q}^{+}_{\exists z\varphi} &: \mathcal{W}_\Gamma \to \mathbf{N}, \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{q}^{-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \\ \vec{r}^{+}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\psi \\ \vec{r}^{-}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\varphi \\ \vec{r}^{-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma.\end{aligned}$$

As usual the two components of $\vec{q}^{+}_{\exists z\varphi}$, will be called $\vec{q}^{+}_{z}$ and $\vec{q}^{+}_{\varphi}$. Then the two induction hypotheses are:

$$\Gamma_D(\vec{x}, \vec{q}^{-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi_D(\vec{q}^{+}_{z}(\vec{x}), \vec{q}^{+}_{\varphi}(\vec{x}), \vec{v})$$

$$\Gamma_D(\vec{x}, \vec{r}^{-}_{\Gamma}(\vec{x}, \vec{u}, \vec{w})), \varphi_D(\vec{u}, \vec{r}^{-}_{\varphi}(\vec{x}, \vec{u}, \vec{w})) \vdash \psi_D(\vec{r}^{+}_{\psi}(\vec{x}, \vec{u}), \vec{w}).$$

The types of the witnesses for the conclusion are

$$\begin{aligned}\vec{p}^{+}_{\psi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\psi \\ \vec{p}^{-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\psi \to \mathcal{C}_\Gamma.\end{aligned}$$

We construct $\vec{p}^{\,+}_{\psi}$ in a way that may seem slightly odd at first glance:

$$\vec{p}^{\,+}_{\psi} = \lambda(\vec{x} : \mathcal{W}_{\Gamma}).(\lambda z.\vec{r}^{\,+}_{\psi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x})))\vec{q}^{\,+}_{z}(\vec{x}).$$

We will explain this added substitution of z for $\vec{q}^{\,+}_{z}(\vec{x})$ later. To construct $\vec{p}^{\,-}_{\Gamma}$ we do the usual work of constructing two different terms of the right type, one extracted from each premise, and then merging them:

$$\begin{aligned}
&\vec{p}^{\,-}_{q} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\Gamma}\\
&\vec{p}^{\,-}_{q} = \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{w} : \mathcal{C}_{\psi}).\vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))\\
&\vec{p}^{\,-}_{r} : \mathcal{W}_{\Gamma} \to \mathcal{C}_{\psi} \to \mathcal{C}_{\Gamma}\\
&\vec{p}^{\,-}_{r} = \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{w} : \mathcal{C}_{\psi}).(\lambda z.\vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))\vec{q}^{\,+}_{z}(\vec{x})\\
&\vec{p}^{\,-}_{\Gamma} = \lambda(\vec{x} : \mathcal{W}_{\Gamma})(\vec{w} : \mathcal{C}_{\psi}).\mathbf{Merge}_{\Gamma}(\vec{p}^{\,-}_{q}(\vec{x}, \vec{w}), \vec{p}^{\,-}_{r}(\vec{x}, \vec{w}), \vec{x}).
\end{aligned}$$

Note that we added an odd substitution of z for $\vec{q}^{\,+}_{z}(\vec{x})$ in our definition of $\vec{p}^{\,-}_{r}$ as well. We then have to show that

$$\Gamma_D(\vec{x}, \vec{p}^{\,-}_{\Gamma}(\vec{x}, \vec{w})) \vdash \psi_D(\vec{p}^{\,+}_{\psi}(\vec{x}), \vec{w}).$$

We have seen how **Merge** functions a few times now so we know that this equivalent to showing that

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))), \Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w})) \vdash \psi_D(\vec{r}^{\,+}_{\psi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x})), \vec{w})$$

where any occurrence of z in $\vec{r}^{\,-}_{\Gamma}$ and $\vec{r}^{\,+}_{\psi}$ has been substituted with $\vec{q}^{\,+}_{z}(\vec{x})$. The first induction hypothesis gives us:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))) \vdash \varphi_D(\vec{q}^{\,+}_{z}(\vec{x}), \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))$$

which is equivalent to:

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x}, \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))) \vdash \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))[z := \vec{q}^{\,+}_{z}(\vec{x})].$$

The second induction hypothesis along with the →I-rule in **HA+T** gives us:

$$\begin{aligned}
&\Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w}))\\
&\qquad\qquad\vdash \varphi_D(\vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{r}^{\,-}_{\varphi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x}), \vec{w})) \to \psi_D(\vec{r}^{\,+}_{\psi}(\vec{x}, \vec{q}^{\,+}_{\varphi}(\vec{x})), \vec{w})
\end{aligned}$$

Note that z does not occur freely in Γ by assumption and we have substituted any occurrence of it in $\vec{r}^{\,-}_{\Gamma}$ and $\vec{r}^{\,+}_{\psi}$ for $\vec{q}^{\,+}_{z}$. Hence we can freely substitute any free occurrence of z left in φ_D for $\vec{q}^{\,+}_{z}$. This means that the two induction hypotheses along with the →E rule gives us the desired result.

Lastly we do the rules of arithmetic. The establishing rules for $0, \mathbf{S}, +$ and $\cdot$ as well as the first equality rule only deal with deductions of prime formulas and to show that the theorem holds for them we do not really rely on the witnesses for them but rather on the rules in **HA+T** corresponding these rules. We show how to prove the result for the first equality axiom and then explain how to show the result for the rest of these rules in a similar way.

14. $\overline{\Gamma \vdash n = n}$

We have no premises here so there is no induction hypothesis. We therefore begin directly with the unfolding of the types of the witnesses of the conclusion:

$$\begin{aligned} \vec{p}^{\,+}_{n=n} &: \mathcal{W}_\Gamma \to \emptyset \\ &: \emptyset \\ \vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \emptyset \to \mathcal{C}_\Gamma \\ &: \mathcal{W}_\Gamma \to \mathcal{C}_\Gamma. \end{aligned}$$

So there is no need to construct $\vec{p}^{\,+}_{n=n}$ since it is empty and $\vec{p}^{\,-}_{\Gamma}$ is constructed with a dummy term:

$$\vec{p}^{\,-}_{\Gamma} = \lambda(\vec{x} : \mathcal{W}_\Gamma).\varnothing_{\mathcal{C}_\Gamma}.$$

Then all that is left is to prove that

$$\Gamma_D(\vec{x}, \vec{p}^{\,-}_{\Gamma}(\vec{x})) \vdash n = n.$$

This is just the first equality axiom of **HA+T** and thus we get the desired result.

The proofs for the establishing rules for $0, \mathbf{S}, +$ and $\cdot$ are analogous to this one. The only difference is that for the establishing rules for $+$ and $\cdot$ we rely on the fact the substitution rule for $\overset{*}{\leftrightarrow}$-equivalent terms and the way in which these functions are defined in **T** instead of relying on any establishing rules.

15. $\dfrac{\Gamma \vdash n = m \qquad \Gamma \vdash \varphi[z := n]}{\Gamma \vdash \varphi[z := m]}$

We unfold the witness types of the premises:

$$\begin{aligned} \vec{q}^{\,+}_{n=m} &: \emptyset \\ \vec{q}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\Gamma \\ \vec{r}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{r}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \end{aligned}$$

and the two induction hypotheses are

$$\Gamma_D(\vec{x}, \vec{q}^{\,-}_{\Gamma}(\vec{x})) \vdash n = m$$

$$\Gamma_D(\vec{x}, \vec{r}^{\,-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi[z := n]_D(\vec{r}^{\,+}_{\varphi}(\vec{x}), \vec{v}).$$

The witnesses for the conclusion have the following type:

$$\begin{aligned} \vec{p}^{\,+}_{\varphi} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\ \vec{p}^{\,-}_{\Gamma} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma. \end{aligned}$$

We can construct $\vec{p}_\varphi^{\,+}$ in an obvious way:

$$\vec{p}_\varphi^{\,+} = \lambda(\vec{x} : \mathcal{W}_\Gamma).\vec{r}_\varphi^{\,+}(\vec{x})$$

and for $\vec{p}_\Gamma^{\,-}$ we use **Merge** as usual when we are dealing with two premises: The witnesses for the conclusion have the following type:

$$\vec{p}_\Gamma^{\,-} = \lambda(\vec{x} : \mathcal{W}_\Gamma)(\vec{v} : \mathcal{C}_\Gamma).\mathbf{Merge}_\Gamma(\vec{r}_\Gamma^{\,-}(\vec{x}, \vec{v}), \vec{q}_\Gamma^{\,-}(\vec{x}), \vec{x}).$$

We then have to show that

$$\Gamma_D(\vec{x}, \vec{p}_\Gamma^{\,-}(\vec{x}, \vec{v})) \vdash \varphi[z := m]_D(\vec{p}_\varphi^{\,+}(\vec{x}), \vec{v}).$$

which is equivalent to

$$\Gamma_D(\vec{x}, \vec{r}_\Gamma^{\,-}(\vec{x}, \vec{v})), \Gamma_D(\vec{x}, \vec{q}_\Gamma^{\,-}(\vec{x})) \vdash \varphi[z := m]_D(\vec{r}_\varphi^{\,+}(\vec{x}), \vec{v}).$$

But by the induction hypotheses and the second equality rule of **HA+T** this holds.

15. $$\frac{\Gamma \vdash \varphi[z := 0] \qquad \Gamma \vdash \varphi[z := y] \to \varphi[z := \mathbf{S}y]}{\Gamma \vdash \varphi[z := n]}$$

This is the last and by far the most complex part of the proof. We must of course begin as usual by unfolding the types of the witnesses of the premises:

$$\begin{aligned}
\vec{q}_\varphi^{\,+} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\
\vec{q}_\Gamma^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma \\
\vec{r}_\varphi^{\,+} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{W}_\varphi \\
\vec{r}_\varphi^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\varphi \to \mathcal{C}_\varphi \\
\vec{r}_\Gamma^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma.
\end{aligned}$$

These give us two induction hypotheses:

$$\Gamma_D(\vec{x}, \vec{q}_\Gamma^{\,-}(\vec{x}, \vec{v})) \vdash \varphi[z := 0]_D(\vec{q}_\varphi^{\,+}(\vec{x}), \vec{v})$$

$$\Gamma_D(\vec{x}, \vec{r}_\Gamma^{\,-}(\vec{x}, \vec{u}, \vec{v})) \vdash \varphi[z := y]_D(\vec{u}, \vec{r}_\varphi^{\,-}(\vec{x}, \vec{u}, \vec{v})) \to \varphi[z := \mathbf{S}y]_D(\vec{r}_\varphi^{\,+}(\vec{x}, \vec{u}), \vec{v})$$

The types of the witnesses of the conclusion are the following:

$$\begin{aligned}
\vec{p}_\varphi^{\,+} &: \mathcal{W}_\Gamma \to \mathcal{W}_\varphi \\
\vec{p}_\Gamma^{\,-} &: \mathcal{W}_\Gamma \to \mathcal{C}_\varphi \to \mathcal{C}_\Gamma.
\end{aligned}$$

Part of the difficulty of the proof lies in the fact that the witnesses of the conclusion must all be defined at the same time because they rely on one another:

$$\begin{aligned}
\vec{p}_\varphi^{\,+}, \vec{p}_\Gamma^{\,-} :=& \lambda(\vec{x} : \mathcal{W}_\varphi).\mathbf{R}((\vec{q}_\varphi^{\,+}(\vec{x}), \vec{q}_\Gamma^{\,-}(\vec{x})), \\
&\lambda z\vec{y}\vec{f}.(\vec{r}_\varphi^{\,+}(\vec{x}, \vec{y}), \lambda(\vec{v} : \mathcal{C}_\varphi).\mathbf{Merge}_\Gamma(\vec{f}(\vec{r}_\varphi^{\,-}(\vec{x}, \vec{y}, \vec{v})), \vec{r}_\Gamma^{\,-}(\vec{x}, \vec{y}, \vec{v}), \vec{x}), n).
\end{aligned}$$

Now let $\vec{p}^{+}_{n}$ and $\vec{p}^{-}_{n}$ stand for $\vec{p}^{+}_{\varphi}$ and $\vec{p}^{-}_{\Gamma}$ respectively where n represents the n in $\vec{p}^{+}_{\varphi}, \vec{p}^{-}_{\Gamma}$. Then the following equivalences hold:

$$\begin{aligned}
\vec{p}^{+}_{0}(\vec{x}) &\overset{*}{\leftrightarrow} \vec{q}^{+}_{\varphi}(\vec{x}) \\
\vec{p}^{+}_{\mathbf{S}n}(\vec{x}) &\overset{*}{\leftrightarrow} \vec{r}^{+}_{\varphi}(\vec{x}, \vec{p}^{+}_{n}(\vec{x})) \\
\vec{p}^{-}_{0}(\vec{x}, \vec{v}) &\overset{*}{\leftrightarrow} \vec{q}^{-}_{\Gamma}(\vec{x}, \vec{v}) \\
\vec{p}^{+}_{\mathbf{S}n}(\vec{x}, \vec{v}) &\overset{*}{\leftrightarrow} \mathbf{Merge}_{\Gamma}(\vec{p}^{-}_{n}(\vec{x}, \vec{r}^{-}_{\varphi}(\vec{x}, \vec{p}^{+}_{n}(\vec{x}), \vec{v})), \vec{r}^{-}_{\Gamma}(\vec{x}, \vec{p}^{+}_{n}(\vec{x}), \vec{v}), \vec{x}).
\end{aligned}$$

Now we must prove that

$$\Gamma_{D}(\vec{x}, \vec{p}^{-}_{\Gamma}(\vec{x}, \vec{v})) \vdash \varphi[z := n]_{D}(\vec{p}^{+}_{\varphi}(\vec{x}), \vec{v}).$$

We do this by induction. First assume that $n = 0$. Then we want to show that

$$\Gamma_{D}(\vec{x}, \vec{p}^{-}_{0}(\vec{x}, \vec{v})) \vdash \varphi[z := 0]_{D}(\vec{p}^{+}_{0}(\vec{x}), \vec{v})$$

which is equivalent with the first induction hypothesis:

$$\Gamma_{D}(\vec{x}, \vec{q}^{-}_{\varphi}(\vec{x}, \vec{v})) \vdash \varphi[z := 0]_{D}(\vec{q}^{+}_{\varphi}(\vec{x}), \vec{v})$$

giving us the desired result.

Now assume that we have already showed that the result holds for $n = m$, using $\vec{p}^{-}_{m}$ and $\vec{p}^{+}_{m}$ as witnesses. We want to show that it holds for $n = \mathbf{S}m$. Then we have to show that

$$\Gamma_{D}(\vec{x}, \vec{p}^{-}_{\mathbf{S}m}(\vec{x}, \vec{v})) \vdash \varphi[z := \mathbf{S}m]_{D}(\vec{p}^{+}_{\mathbf{S}m}(\vec{x}), \vec{v}).$$

We know that this leads to the following by Lemma 3.13:

$$\begin{aligned}
&\Gamma_{D}(\vec{x}, \vec{p}^{-}_{m}(\vec{x}, \vec{r}^{-}_{\varphi}(\vec{x}, \vec{p}^{+}_{n}(\vec{x}), \vec{v}))), \Gamma_{D}(\vec{x}, \vec{r}^{-}_{\Gamma}(\vec{x}, \vec{p}^{+}_{n}(\vec{x}), \vec{v})) \\
&\qquad\qquad\qquad\qquad \vdash \varphi[z := \mathbf{S}m]_{D}(\vec{r}^{+}_{\varphi}(\vec{x}, \vec{p}^{-}_{m}(\vec{x})), \vec{v}).
\end{aligned}$$

By assumption we have

$$\Gamma_{D}(\vec{x}, \vec{p}^{-}_{m}(\vec{x}, \vec{r}^{-}_{\varphi}(\vec{x}, \vec{p}^{+}_{n}(\vec{x}), \vec{v}))) \vdash \varphi[z := m]_{D}(\vec{r}^{+}_{\varphi}(\vec{x}, \vec{p}^{-}_{m}(\vec{x})), \vec{v})$$

and from the second induction hypothesis we have

$$\begin{aligned}
&\Gamma_{D}(\vec{x}, \vec{r}^{-}_{\Gamma}(\vec{x}, \vec{p}^{+}_{m}(\vec{x}), \vec{v})) \\
&\qquad \vdash \varphi[z := m]_{D}(\vec{u}, \vec{r}^{-}_{\varphi}(\vec{x}, \vec{p}^{+}_{m}(\vec{x}), \vec{v})) \to \varphi[z := \mathbf{S}m]_{D}(\vec{r}^{+}_{\varphi}(\vec{x}, \vec{p}^{+}_{m}(\vec{x})), \vec{v})
\end{aligned}$$

and from these two results along with the →E-rule in **HA+T** we get the desired result.

This concludes the proof. □

4 Semantics of T

This paper has up until now been dedicated entirely to explaining the Dialectica interpretation and therefore we have only looked at system **T** in light of its role as an interpreter of **HA**. However **T** is an interesting system in its own right. To give the reader a little more insight into the inner workings of **T** this last section will be dedicated to showing some ways in which semantics can be given for **T**. We look in particular at two very different models, a purely syntactical term model and a purely denotational model based on mathematical structures called coherence spaces.

4.1 A term model for T

In section 2.2.2 we introduced the reduction rules of **T**. What kind of meaning should be ascribed to these reduction rules? They are a finite set of rules that determine a process through which certain terms can be reduced to other terms. This certainly has a computational flavour to it. Let us now take a look at a few concept that can be be defined using the reduction rules.

Definition 4.1.

1. A subterm s of a term t is called a *redex* if it is possible to apply a reduction rule to it, that is if there exists a term r such that $s \rhd r$.
2. If a term has a redex it is said to be *reducible.*
3. If a term is not reducible it is said to be in *normal form* or *irreducible.*
4. A term is said to be *normalizable* if it can be reduced to a term in normal form. If every term of a system is normalizable, then the system is said to be *normalizing.*
5. A term is said to be *strongly normalizing* if no infinite sequence of reductions begins with it, that is any sequence of reductions beginning with the term ends with a term in normal form. If every term of a system is strongly normalizable, then the system is said to be *strongly normalizing.*
6. A system is said to be *confluent* or have the *Church-Rosser property* if for every term s of the system, when $s \to^* u$ and $s \to^* v$ then, there exists a term t such that $u \to^* t$ and $v \to^* t$.

It is easy to see that if a system is confluent every normalizable term will have only one normal form. Moreover, if the system is also normalizing, then every term will have a unique normal form. This suggests a very simple way to construct a model for systems posessing both of these qualities, namely a so called *term model* in which each term of the system is identified with its normal form.

It is possible to show that **T** is both confluent and normalizing. In fact system **T** is also strongly normalizing. Proofs of these facts can be found in appendices A2 and A3 of Hindley and Seldin's *Lambda-Calculus and Combinators, an Introduction* (2008).

The fact that **T** is strongly normalizing means that if a term of **T** has more than one redex, it does not matter in which order the reduction rules are applied to the redexes of the term, any chain of reductions starting with the term will terminate and in fact, since **T** is confluent, result in the same term.

So we see that these properties allow us to create a term model for **T**. By taking the computational flavour of the reduction rules even more seriously one might think of the closed terms of **T** as programs that compute terms in normal form when they are applied to other terms in normal form. With this reading **T** is nothing more than a programming language.

There are certain interesting aspects to this interpretation. For example it is obvious that each closed term in normal form of type **N** is a numeral. It is therefore possible to identify these with the natural numbers. Then the fact that **T** is confluent and normalizing makes it impossible to show that $0 = 1$ and thus gives us a way of showing the consistency of arithmetic.

This account of the term model for system **T** is based on sections 4.2 and 4.3 of Avigad and Feferman (1998). Readers interested in term models will find more information on them there.

4.2 Denotational semantics for T

While the term model does provide an adequate model for **T** it is a very naive model, in the sense that it doesn't really interpret the terms of **T** as anything other than other terms of **T**. One might therefore wish to find a denotational model for **T**, a model that does not in any way involve the syntax of **T**.

There is of course an obvious way of constructing a denotational model using set theory. In such a model the terms of type **N** would be identified with the natural numbers and then each type $\sigma \to \tau$ would represent the set of functions from from the type σ to the type τ.

Another way of providing a model for **T** is given by Jean-Yves Girard in chapters 8 and 9 of his book *Proofs and Types* (1989). There Girard uses ideas developed from domain theory to construct a model for **T**. At the heart of this model are structures called coherence spaces.

Definition 4.2. A *coherence space* is a set of sets A satisfying the following two conditions:

1. If $a \in A$ and $a' \subset a$, then $a' \in A$.
2. If $X \subseteq A$ and for all $a_1, a_2 \in X$ it holds that $a_1 \cup a_2 \in A$, then $\bigcup\limits_{x \in X} x \in A$.

The members of a coherence space A are called the *points* of A and the set $|A| = \{\alpha : \{\alpha\} \in A\}$, the union of all the members of A, is called the *web* of A. The elements of $|A|$ are called the *tokens* of A.

For the readers familiar with domain theory it is possible to think of coherence spaces as domains where the objects are ordered by inclusion. Then the minimal member of every coherence space is the empty set, $\emptyset$.

Another possible way of looking at coherence spaces is to think of them as a graph. We define the following relation on the members of $|A|$.

Definition 4.3. For any two tokens $\alpha_1, \alpha_2 \in |A|$, we say that α_1 *is coherent with* α_2 *modulo* A, written

$$\alpha_1 \coh \alpha_2 \pmod{A}$$

if and only if

$$\{\alpha_1, \alpha_2\} \in A.$$

Since the relationship $\coh$ is clearly both reflexive and symmetric it is clear that a coherence space A must define an undirected graph with its web $|A|$ as the set of nodes and the relationship $\coh$ defining the edges. The points of A are simply the complete subgraphs of this graph.

In fact any undirected graph defines a coherence space. This is easily seen from the equivalence:

$$a \in A \leftrightarrow a \subseteq |A| \wedge \forall \alpha_1, \alpha_2 \in a(\alpha_1 \coh \alpha_2 \pmod{A}).$$

Example 4.4. We give few some simple examples to demonstrate how coherence spaces work.

- The simplest possible coherence space is the one consisting only of $\emptyset$.
- Coherence spaces consisting only of singleton sets along with the empty set are called *flat* spaces. A particularly important flat space is the set we shall call **Nat**, consisting of the singletons $\{0\}, \{1\}, \{2\}, \ldots$ as well as $\emptyset$. The graphs representing flat spaces are the discrete graphs, that is the graphs that have no edges.
- Let us look the set

  $$A = \{\{\}, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}.$$

  This is not a coherence space since $\{\{1\}, \{2\}, \{3\}\} \subset A$ and the union of each pair of points in the subset is also a member of A and yet the union of all the members of the subset, $\{1,2,3\}$ is not in A. This clearly violates the second condition of the definition of coherence spaces. However the set

  $$A \cup \{1,2,3\}$$

  is a coherence space.

Before moving on we define a few useful concepts.

Definition 4.5.

1. A family X of points of a coherence space A is said to be *directed* if for every pair $x_1, x_2 \in X$ there exists a $y \in X$ such that $x_1 \cup x_2 \subseteq y$.
2. An element a of a coherence space A is said to be *maximal* [*minimal*] if for all $a' \in A$, $a \subseteq a'$ $[a' \subseteq a]$ implies that $a = a'$.
3. The subset of a coherence space A consisting of all the finite members of A is denoted by $A_{\mathbf{fin}}$

Our goal is to interpret each type of **T** as a coherence space. But for us to be able to do so we must have some method of interpreting types of the form $\sigma \to \tau$. We will now define a class of functions that can be used to construct spaces suitable for the interpretation of these types.

Definition 4.6. Let A and B be coherence spaces. A function $F : A \to B$ is said to be *stable* if it satisfies the following conditions:

1. If $a_1 \subset a_2 \in A$, then $F(a_1) \subset F(a_2)$.
2. If X is a directed family of points of A, then $\bigcup_{x \in X} F(x) = F(\bigcup_{x \in X} x)$.
3. If $a_1 \cup a_2 \in A$, then $F(a_1 \cap a_2) = F(a_1) \cap F(a_2)$.

The first two conditions should be familiar to those who are familiar with domain theory. The first condition says that stable functions are monotone and the second one says that stable functions are continuous in a domain theoretical sense, that is to say they preserve least upper bounds of directed families of points. The third condition says that the function has a property called stability. This property allows us to prove the following lemma.

Lemma 4.7. Let F be a stable function from a coherence space A to a coherence space B and let $a \in A$ and $\beta \in |B|$. Then

1. If $\beta \in F(a)$, then there exists a finite $a_0 \subseteq a$ such that $\beta \in F(a_0)$.
2. For each β there exists a unique minimal solution a_0 to the first part of the lemma.

Proof.

1. Let X_a denote the set of finite subsets of a. Then it is clear that $a = \bigcup_{x \in X} x$ and thus $\bigcup_{x \in X} F(x) = F(\bigcup_{x \in X} x) = F(a)$. Thus if $\beta \in F(a)$ there must exist some $a_0 \in X$ such that $\beta \in F(a_0)$.
2. Let a_0 be a minimal solution to the first part of the lemma. Take some finite a' such that $a' \subseteq a$ and $\beta \in F(a')$. Then it is clear that $a_0 \cup a' \subseteq a$ so $a_0 \cup a' \in A$ and thus, $\beta \in F(a_0) \cap F(a') = F(a_0 \cap a')$. But since a_0 is minimal we must have $a_0 \subseteq a_0 \cap a'$ making it clear that $a_0 \subseteq a'$. This does indeed show that a_0 is unique. □

This lemma allows us to make the following definition.

Definition 4.8. Let F be a stable function. Then the *trace* of F denoted by $\mathbf{Tr}(F)$ is the set of pairs (a_0, β) such that $a_0 \in A$ is finite, $\beta \in F(a_0)$ and for any $a' \subseteq a_0$ such that $\beta \in f(a')$, $a = a_0$.

The following lemma is then an immediate corollary of Lemma 4.7.

Lemma 4.9. The trace of every stable function F determines F completely through the following equation:

$$F(a) = \{\beta : \exists a_0(a_0 \subseteq a \wedge (a_0, \beta) \in \mathbf{Tr}(F))\}.$$

We are now ready to show how to construct a coherence space defined by stable functions.

Definition 4.10. Let A and B be coherence spaces. The *function space* of functions from A to B denoted by $A \to B$ is defined as follows:

- $|A \to B| = A_{\mathbf{fin}} \times |B|$
- $(a_1, \beta_1) \mathrel{\overset{\frown}{\smile}} (a_2, \beta_2) \, (\text{mod } A \to B)$ if and only if:
  1. if $a_1 \cup a_2 \in A$, then $\beta_1 \mathrel{\overset{\frown}{\smile}} \beta_2 \, (\text{mod } B)$, and
  2. if $a_1 \cup a_2 \in A$ and $\beta_1 = \beta_2$, then $a_1 = a_2$

On first glance this definition seems rather arbitrary. But the following theorem gives meaning to it.

Theorem 4.11. The members of $A \to B$ are the traces $\mathbf{Tr}(F)$, where F ranges over the stable functions from A to B.

Proof. We begin by showing that if F is a stable function, then $\mathbf{Tr}(F) \in A \to B$. It is easy to see that $\mathbf{Tr}(F) \in |A \to B|$. Let

$$(a_1, \beta_1), (a_2, \beta_2) \in \mathbf{Tr}(F).$$

Assume that $a_1 \cup a_2 \in A$. We know that $\beta_1 \in F(a_1)$ and $\beta_2 \in F(a_2)$. Thus by the monotonicity of F we have $\{\beta_1, \beta_2\} \subseteq F(a_1 \cup a_2)$ which clearly implies that $\beta_1 \mathrel{\overset{\frown}{\smile}} \beta_2 \, (\text{mod } B)$.

Now assume that $a_1 \cup a_2 \in A$ and $\beta_1 = \beta_2$. Clearly $a_1 \cup a_2 \in A$ implies that $\beta_1, \beta_2 \in F(a_1 \cup a_2)$. By the definition of $\mathbf{Tr}(F)$, a_1 and a_2 must be the minimal subsets of $a_1 \cup a_2$ such that $\beta_1 \in F(a_1)$ and $\beta_2 \in F(a_2)$. But according to Lemma 4.7 a_1 and a_2 are unique, so since $\beta_1 = \beta_2$ we must have $a_1 = a_2$. This shows that $\mathbf{Tr}(F) \in A \to B$.

We now want to show that if $f \in A \to B$, then there is some stable function $F : A \to B$ such that $f = \mathbf{Tr}(F)$. We define a function $F : A \to B$ by the following equation:

$$F(a) = \{\beta : \exists a_0(a_0 \subset a \wedge (a_0, \beta) \in f)\}.$$

We first show that F is in fact a function from A to B. It is clear that given any $a \in A$, $F(a) \subseteq B$. Thus we only have to show that the members of $F(a)$ are coherent modulo B. Let $\beta_1, \beta_2 \in F(a)$. Then by the definition of F there exist some $a_1, a_2 \subseteq a$ such that $(a_1, \beta_1), (a_2, \beta_2) \in f$. But this implies $a_1 \cup a_2 \subseteq a \in A$ and then by the coherence of f we have $\beta_1 \overset{\frown}{\smile} \beta_2 \pmod{B}$.

We then show that F is stable. The monotonicity of F is immediate from the definition. To prove the continuity of F assume that X is a directed family of points of A and let $a = \bigcup_{x \in X} x$. Then it is easy to see that monotonicity forces $\bigcup_{x \in X} F(x) \subseteq F(a)$. Now let $\beta \in F(a)$. Then there is some finite $a_0 \subseteq a$ such that $\beta \in F(a_0)$. This also means that $a_0 \subseteq \bigcup_{x \in X} x$ and hence $a_0 \subseteq x'$ for some $x' \in X$. But then it clearly holds that $\beta \in \bigcup_{x \in X} F(x)$ showing that $f(a) \subseteq \bigcup_{x \in X} F(x)$, proving the continuity of F.

Then we prove the stability of F. Assume that $a_1 \cup a_2 \in A$. It is clear from the monotonicity of F that $F(a_1 \cap a_2) \subseteq F(a_1) \cap F(a_2)$. Now let $\beta \in F(a_1) \cap F(a_2)$. Then there are some $a_1' \subseteq a_1, a_2' \subseteq a_2$ such that $(a_1', \beta), (a_2', \beta) \in f$. Thus (a_1', β) and (a_2', β) are coherent and since $a_1' \cup a_2' \subseteq a_1 \cup a_2 \in A$ we must have $a_1' = a_2'$. It is clear that $a_1' \subseteq a_1 \cap a_2$ so $\beta \in F(a_1 \cap a_2)$ showing that $F(a_1) \cap F(a_2) \subseteq F(a_1 \cap a_2)$. Thus F is a stable function.

Then the only thing left is to show that $f = \mathbf{Tr}(F)$. The inclusion $\mathbf{Tr}(F) \subseteq f$ follows from Lemma 4.9. Assume that $(a_0, \beta) \in f$. By the construction of $|A \to B|$ we have $a_0 \in A_{\mathbf{fin}}$ and $\beta \in |B|$ and by the construction of F we have $\beta \in F(a_0)$. We then only have to show that for any $a' \subseteq a_0$, if $\beta \in F(a')$, then $a' = a_0$. It is clear by the construction of F that if $\beta \in F(a')$ there is some $a'' \subseteq a'$ such that $(a'', \beta) \in f$. But by the coherence of f and the fact that $a'' \cup a_0 \in A$ we get $a'' = a_0$ which clearly forces $a' = a_0$. □

We now have defined all the concepts we need for our interpretation of $\mathbf{T}$. In what follows we shall use $[\![\cdot]\!]$ to denote the interpretation function that maps the objects of $\mathbf{T}$ to their interpretations.

The first order of business is to find a way to interpret $\mathbf{N}$ as a coherence space. One might be tempted to try to interpret $\mathbf{N}$ as the flat coherence space $\mathbf{Nat}$, mentioned in Example 4.4 by the following obvious interpretation:

$$[\![0]\!] = \{0\} \quad \text{and} \quad [\![\mathbf{S}n]\!] = \mathcal{S}([\![n]\!])$$

where $\mathcal{S}$ is defined by the following equations:

$$\mathcal{S}(\{n\}) = \{n+1\} \quad \text{and} \quad \mathcal{S}(\emptyset) = \emptyset.$$

However it turns out that this interpretation does not work. Here below we will show how to interpret terms of the form $\mathbf{R}(f, g, n)$ with stable functions defined on $[\![\mathbf{N}]\!]$. Now assume we have some terms t and s of $\mathbf{T}$ such that

$$\mathbf{R}(t, s, 0) = n \quad \text{and} \quad \mathbf{R}(t, s, \mathbf{S}(x)) = m$$

for some integers n and m. Then we must have some stable function F which interprets the function $x \mapsto \mathbf{R}(t, s, x)$. It is clear that we would have

$$F(\{0\}) = \{n\} \quad \text{and} \quad F(\mathcal{S}(x)) = \{m\}.$$

In particular we get $F(\mathcal{S}(\emptyset)) = m$. But $\mathcal{S}(\emptyset) = \emptyset \subseteq \{0\}$ while $F(\mathcal{S}(\emptyset)) \nsubseteq F(\{0\})$ which clearly contradicts the fact that F should be stable.

The problem here lies in the fact that we interpret $\mathcal{S}(\emptyset)$ as $\emptyset$, something lacking any information, while we do in fact have some information, we know we have a successor. Hence we need a different way to interpret $\mathbf{N}$.

So we construct a new coherence space in search of a way in which we can make $\mathcal{S}(\emptyset)$ have the desired meaning. We call this coherence space $\mathbf{Nat}^+$. We let $|\mathbf{Nat}^+| = \{0, 0^+, 1, 1^+, 2, 2^+, \dots\}$. Now assume that the variables n, m range over$\{0, 1, 2, \dots\}$ while the variables n^+, m^+ range over $\{0^+, 1^+, 2^+, \dots\}$. Then we define the coherence relationship modulo $\mathbf{Nat}^+$ as follows:

$$\begin{aligned} &n \frown\!\!\!\!\smile m \,(\text{mod } \mathbf{Nat}^+) \text{ iff } n = m\\ &n^+ \frown\!\!\!\!\smile m \,(\text{mod } \mathbf{Nat}^+) \text{ iff } n^+ < m\\ &n^+ \frown\!\!\!\!\smile m^+ \,(\text{mod } \mathbf{Nat}^+) \text{ for all } n^+, m^+.\end{aligned}$$

Let us take a look at the maximal points of this space. There are two different types of maximal points in $\mathbf{Nat}^+$.

- If $a \in \mathbf{Nat}^+$ is maximal and there is some $n \in a$, then
  $a = \{0^+, \dots, (n-1)^+, n\}$.
- If a is maximal and there is no $n \in a$, then $a = \{0^+, 1^+, 2^+, \dots\}$.

We then interpret the elements of $\mathbf{N}$ as follows:

$$[\![0]\!] = \{0\} \quad \text{and} \quad [\![\mathbf{S}n]\!] = \mathcal{S}([\![n]\!])$$

where $\mathcal{S}$ is the function defined by the following equation:

$$\mathcal{S}(a) = \{0^+\} \cup \{n+1 : n \in a\} \cup \{(n+1)^+ : n^+ \in a\}.$$

Under this interpretation we get $[\![n]\!] = \{0^+, \dots, (n-1)^+, n\}$, if n denotes the n-th successor of 0. This notation also gives a meaning to the intuition that applying the successor function to $\emptyset$ should convey more information than just $\emptyset$ since we get $\mathcal{S}^k(\emptyset) = \{0^+, \dots, k^+\}$.

Having found a suitable interpretation of the terms of type $\mathbf{N}$ we would like to extend this interpretation to the type $\mathbf{N}$ itself. So we let

$$[\![\mathbf{N}]\!] = \mathbf{Int}^+.$$

We then simply interpret the rest of the types of $\mathbf{T}$ as follows

$$[\![\sigma \to \tau]\!] = [\![\sigma]\!] \to [\![\tau]\!].$$

We now want to find a method to interpret the rest of the terms of $\mathbf{T}$. We are not interested in finding a good interpretation of all the terms of $\mathbf{T}$ though, we are only really interested in interpreting the closed terms of $\mathbf{T}$. However since we will interpret each term componentwise, that is the interpretation of each term relies on the interpretation of its subterms, we must have some method that we can use to deal with free variables since they will inevitably occur in the subterms of many closed terms. The following definition gives us a tool to solve this problem.

Definition 4.12. A *variable assignment* is a function that assigns to each variable x^τ of $\mathbf{T}$ a unique element of $[\![\tau]\!]$. If π is a variable assignment, x^τ is a variable and $t \in [\![\tau]\!]$, then we let $\pi[x \leftarrow t]$ denote the variable assignment identical to π with the exception that $\pi[x \leftarrow t](x) = t$.

We now let $[\![\cdot]\!]_\pi$ denote the interpretation function relative to the variable assignment π. Now assume that t is any term of $\mathbf{T}$ not of the form 0 or $\mathbf{S}(n)$. We define $[\![t]\!]_\pi$ as follows:

- If $t = x$ where x is a variable, we let $[\![t]\!]_\pi = \pi(x)$.
- Assume $t = u(v)$ where $u : \sigma \to \tau$ and $v : \sigma$. Then
$$[\![t]\!]_\pi = \{t' : \exists v'(v' \subseteq [\![v]\!]_\pi \wedge (v', t') \in [\![u]\!]_\pi)\}$$
- Assume $t = \lambda x.u$ where $x : \sigma$ and $u : \tau$. Then let $F(a) = [\![u]\!]_{\pi[x \leftarrow a]}$. Then
$$[\![t]\!]_\pi = \mathbf{Tr}(F).$$
- Assume $t = \mathbf{R}_\tau(f, g, n)$. Then $[\![t]\!]_\pi = F([\![n]\!])$ where F is a function from $\mathbf{Nat}^+$ to $[\![\tau]\!]$ defined by the following equations:
$$F(\{0\}) = [\![f]\!]_\pi \quad F(\mathbf{S}(a)) = [\![g]\!]_\pi(a)(G(a)) \quad F(a) = \emptyset \text{ if } 0, 0^+ \notin a.$$

For $[\![\cdot]\!]$ to be well defined the functions denoted by F must be stable. It is not particularly difficult to prove this, neither in the case for λ-abstraction nor in the case for $\mathbf{R}$. However as these are very tedious proofs we omit them.

It is easy to see that if t is a closed term, then for any two variable assignments π_1 and π_2 we have $[\![t]\!]_{\pi_1} = [\![t]\!]_{\pi_2}$. Thus for t closed the meaning of $[\![t]\!]$ is unambiguous. When t is not closed it is just as clear that this equality will not hold for every choice of π_1 and π_2. But as we already mentioned we are not particularly interested in the interpretation of open terms so this is of no particular concern for us.

We conclude this discussion of the denotational semantics of $\mathbf{T}$ by showing that our interpretation satifies $\mathbf{T}$, in the sense that for all terms t and s, if $t \to^* s$, then $[\![t]\!] = [\![s]\!]$. It suffices to show that $t \rhd s$ implies $[\![t]\!] = [\![s]\!]$, that is to check that this equality holds for the reduction rules of $\mathbf{T}$. In the case of β-reduction we want to show that $[\![(\lambda x.t)s]\!]_\pi = [\![t[x := s]]\!]_\pi$:

$$\begin{aligned}[\![(\lambda x.t)s]\!]_\pi &= \{\alpha : \exists s'[s' \subseteq [\![s]\!]_\pi \wedge (s', \alpha) \in \mathbf{Tr}(a \mapsto [\![t]\!]_{\pi[x \leftarrow a]})]\} \\ &= [\![t]\!]_{\pi[x \leftarrow [\![s]\!]_\pi]}.\end{aligned}$$

It is obvious that $[\![t[x := s]]\!]_\pi = [\![t]\!]_{\pi[x \leftarrow [\![s]\!]_\pi]}$ which gives us the desired equality. The result for the other two reduction rules follows immediately from the definition of $[\![\mathbf{R}(f, g, n)]\!]_\pi$. So we see that our interpretation satisfies $\mathbf{T}$.

5 Conclusion

Let us now look back at our results, summarize them and see if we can draw any conclusions and make some final comments. We have shown how the Dialectica interpretation translates **HA** into our higher type theory of arithmetic, **HA**+**T**. Each formula φ of **HA** was given a translation

$$\varphi^D = \exists \vec{x} : \mathcal{W}_\varphi . \forall \vec{y} : \mathcal{C}_\varphi . \varphi_D(\vec{x}, \vec{y}).$$

We then showed how this translation could be extended to sequents allowing us to prove the soundness of the interpretation.

The soundness proof has some interesting features. It consists of showing that for each sequent $\Gamma \vdash \varphi$ that can be deduced in **HA**, sequences $\vec{p}_\Gamma^-, \vec{p}_\varphi^+$ of terms of the correct witness type $\mathcal{W}_{\Gamma \vdash \varphi}$ can be constructed such that these sequences satisfy

$$\Gamma_D(\vec{x}, \vec{p}_\Gamma^-(\vec{x}, \vec{v})) \vdash_{\mathbf{HA}+\mathbf{T}} \varphi_D(\vec{p}_\varphi^+(\vec{x}), \vec{v})$$

where each formula of the environment Γ_D and the formula φ_D are decidable. This shows that the real work of proving translations of **HA** sequents in **HA**+**T** consists of constructing sequences of effective witnesses, as the fact that Γ_D, φ_D are decidable shows that for each sequence of the type $\mathcal{W}_{\Gamma \vdash \varphi}$ we can test algorithmically whether it constitutes an effective witness of $\Gamma \vdash \varphi$ or not.

If we now connect this observation to the discussions of the computational aspects of **T** in section 4.1 we can make an interesting observation. The Dialectica interpretation reduces the proofs of the theorems of **HA** to a computational process. If we interpret the closed terms of **T** as programs in the programming language **T**, as we did in section 4.1, the act of proving a Dialectica translation of a theorem φ of **HA** boils down to writing a set of programs in **T** and then testing them in the metaprogram φ_D to see whether they produce the desired results.

This observation is very much in line with Gödel's original ideas about the Dialectica translation. In Gödel (1990) he devotes a chunk of the text to the discussion of the idea that the Dialectica interpretation exposes the computational nature of intuitionistic arithmetic. I believe that in this discussion we have corroborated this idea.

Before concluding this summary I want to make a few comments on the presentation of the Dialectica interpretation that can be found in this thesis. The presentation of the translation as well as the soundness proof of the interpretation are in large part based on the presentation of the interpretation that can be found in Pédrot (2015). There are some differences between the presentations, I defined certain concepts in a different manner (the counter type $\mathcal{C}_{\exists z \varphi}$, the definition of the Dialectica translation of sequents (which Pédrot treats as an abuse of notation) and the functions **Decide** and **Merge** for example), I stated the soundness theorem in a different way and I took different paths in the proof of certain parts of it. However this thesis would never have taken the form it has had it not been for the work of Pédrot and I am deeply grateful for having been able to use it as a point of reference during the writing of this thesis.

References

Avigad, J., Feferman, S. (1998) 'Gödel's functional ('Dialectica') interpretation' in: Buss, S. R. (editor), *Handbook of Proof Theory. Studies in Logic and the Foundations of Mathematics.* Vol. 137. Elsevier, Amsterdam, pp. 337–405.

Barendregt, H. P. (1984) *The Lambda Calculus Its Syntax and Semantics. Revised.* Vol. 103. Elsevier, Amsterdam.

Girard, J.-Y., Lafont, Y. and Taylor P. (1989) *Proofs and Types.* Cambridge University Press, Camebridge.

Gödel, K., *Collected Work, Vol. 2*, S. Feferman et al. (editors). (1990) Oxford University Press, New York.

Gödel, K. 'On a hitherto unutilized extension of the finitary standpoint' in Gödel (1990), pp. 241–251.

Gödel, K. 'On an extension of finitary methods which has not yet been used', in: Gödel (1990), pp. 271–280.

Hindley, J. R., Seldin, J. P. (2008) *Lambda-Calculus and Combinators, an Introduction.* Camebridge University Press, Camebridge.

Pédrot, P.-M. (2015). *A Materialist Dialectica.* Diderot, Paris.

Troelstra, A. S. (1973) 'Realizability and functional interpretations' in: Troelstra A.S. (editor) *Metamathematical Investigation of Intuitionistic Arithmetic and Analysis. Lecture Notes in Mathematics, vol 344.* Springer, Berlin, Heidelberg, pp. 175-274.

Troelstra, A. S. (1990) 'Introductory note to 1958 and 1972', in: Gödel (1990), pp. 217–241.