Philosophy of Arithmetic and Incompleteness

Daya Singh, Based off Lectures by Professor Tim Button

June 2024

Table of Contents

Introduction

This document is written for the UCL course Philosophy of Arithmetic and Incompleteness, run by Professor Tim Button. It is split into two major parts; the first of which is Arithmetic, and the second is Incompleteness. These notes are an attempt to arrange semantic notions, that are usually defined in plain text in a formal manner similar to that of common mathematics notes; as I understand it, this has been done (successfully) for other modules. I do not make the argument that these notes are superior to books, or extracts, but I certainly hope that they are.

The obvious difference in philosophy and mathematics notes are definitions. Mathematical definitions vary very little from author to author (and are not subject to a reader's interpretation), whereas English definitions are broad, and are usually open to the interpretation to the reader. To ensure clarity as to which definitions are what, I shall define things with '**Definition** $(\cdot)$'. '**(formal)**' indicates a definition is exact, both in these notes, and as taught in the course, '**(informal)**' implies that it's open to interpretation and '**(attempted)**' means that the definition is attempted to be made more precise, but the precision only applies within this document (and should be assumed when referred, should something be defined twice, as informal and attempted).

Furthermore, solutions to problems in philosophy cannot be chosen arbitrarily as they can in mathematics. For example there are several proofs to Pythagoras' Theorem, and so long as it's sufficiently simple, there is little value gained in documenting both proofs; however many solutions to philosophy have value and may be incomplete in some areas that others cover. As a result, in conjunction with the usual '*theorem x.y.z*', '*lemma x.y.z*' etc, I will present suggestions to a question as '*idea*'; these will be more subjective.

It is worth noting that the latter part of this document will likely resemble a set of mathematics lecture notes than the first.

Finally, whilst I make reference to the weekly reading, I will not, unless where necessary, be making notes on them.

Part 1: Arithmetic

1 Characterising arithmetic

Arithmetic is notoriously difficult to characterise, and has been the debate of mathematicians and philosophers alike. In this chapter, we will attempt to define the parameters that we require a characterisation to satisfy, and some of the main propositions which attempt to satisfy these parameters. We will first need to define arithmetic informally, to differentiate it from mathematics.

Definition (informal) 1.1 (Arithmetic)**.** The language concerned with the natural numbers (and '0'), and the operations + and ×.

1.1 The parameters to motivate a characterisation

In order to motivate arithmetic, we need to define parameters (these parameters are motivated by what generally fits the properties of arithmetic as we *generally* view it).

- **Definition (formal) 1.2** (Infinitary)**.** Allows for arithmetic to have infinite (i.e., non-finite) concepts.
- **Definition (informal) 1.3** (Apodictic)**.** Is self-evidently or demonstrably true, and would be considered absurd otherwise.
- **Definition (informal) 1.4** (A priori)**.** Truths are arrived at by reflection alone; without the need for sensory observation.

  Definition (attempted) 1.5 (A priori)**.** All truths are either apodictic or are a syntactic consequence of another.
- **Definition (informal) 1.6** (Necessary)**.** There is no alternative or hypothetical in which arithmetic is different (i.e., what is in arithmetic cannot have been otherwise).

  Definition (attempted) 1.7 (Necessary)**.** If a statement of arithmetic syntactically implies another in any interpretation, it applies to all others, including contradictory interpretations.
- **Definition (informal) 1.8** (Universally Applicable)**.** It is motivated that arithmetic applies to all areas (e.g chemistry and physics).

  Definition (attempted) 1.9 (Universally Applicable)**.** Any interpretation of arithmetic is valid, and motivated.
- **Definition (informal) 1.10** (Indispensable)**.** We cannot do away with arithmetic; it is needed.
- **Definition (formal) 1.11** (Inter-subjectively Robust)**.** The validity of mathematical claims is not dependent on the person assessing it.

Note that we have decided on these parameters to avoid making the characterisation of arithmetic non-epistemic; that is, without the need for us to have to characterise that it means to *know* something.

Before I introduce possible answers to satisfy these parameters, I would like to define one more term, which will be useful in this module:

Definition (informal) 1.12 (A Posteriori)**.** Dependent on our experience and observations of the physical world.

Remark 1.12.1 (wrt 'a priori')*.* It is not possible for an a posteriori statement to also be priori. This is not unanimously true; Potter [2002] argues it's possible.

1.2 Ideas for a characterisation

Below are 4 possible characterisations on arithmetic:

- **Idea 1.1** (Mill's Empiricism)**.** Arithmetic is defined by the our sensory experience; it is a posteriori. This was advanced by John Stuart Mill.

    - Infinitary: No, as we cannot experience or observe the infinite.
    - A priori: No, as it is a posteriori.
    - Apodictic: Yes, because it would be absurd to look at two apples, and two oranges and say one represents 3 and the other, 2.
    - Necessary: No, because we can only imagine alternate worlds, and are unable to experience them.
    - Universally applicable: Yes, because we are motivated to impose the rules of the world as we observe them onto whatever else we do.
    - Indispensable: Yes, as we certainly cannot reject our own senses.
    - Inter-subjectively Robust: Yes, so long as we can agree that our experiences of the world are not unique.

- **Idea 1.2** (Leibniz's formalism)**.** Arithmetic is a system of definitions, and we evaluate facts from those definitions.

    - Infinitary: Yes, as we can define an infinite limit and/or induction
    - A priori: Yes, as we deduce things from apodictic definitions
    - Apodictic: Yes, so long as we pick sensible definitions
    - Necessary: Yes, because contrary definitions lead to absurd results.
    - Universally applicable: No, as we have no motivation (on this reasoning alone) to assume arithmetic applies to all areas
    - Indispensable: No, because there is nothing to suggest that these rules are of any specific value
    - Inter-subjectively Robust: Yes, since so long as we agree on the definitions, which is assumed by the idea, we can evaluate whether something is valid.

- **Idea 1.3** (Thomae's Game formalism)**.** Arithmetic is a system of rules, posed upon whichever interpretation we put on it

  - Infinitary: Yes, as we can define a rule that extends forever.
  - A priori: Yes, by definition.
  - Apodictic: Yes, so long as we pick sensible rules
  - Necessary: Yes, because contrary rules lead to absurd results.
  - Universally applicable: No, as we have no motivation (on this reasoning alone) to assume all areas adhere to the rules we define.
  - Indispensable: No, because there is nothing to suggest that these rules are of any specific value
  - Inter-subjectively Robust: Yes, since so long as we agree on the rules, which is assumed by the idea, we can evaluate whether something is valid.

  It is worth reflecting on this by understanding it as similar to the rules of chess. These rules are arbitrary, and cannot be agreed upon by everyone; nor can you apply the rules of chess to physics. However it has the benefit of you dimissing the properties of the pieces themselves, as you only care about the abstract notion of the rules associated with them.

- **Idea 1.4** (Intutitionism)**.** Arithmetic is a part of language, and that the language defines the intuition.

  - Infinitary: Yes, as we can can describe infinity
  - Apodictic: No, as our language has no self-evident truths to it.
  - A priori: Yes, as our language motivates the rules of formal logic.
  - Necessary: Yes, because we have a concept of absurdity described by language.
  - Universally applicable: Yes, because all interpretations are governed by language
  - Indispensable: No, because there is nothing to suggest that arithmetic is of any special value
  - Inter-subjectively Robust: No, since our language is unique to us; based on our experiences that we use to try to convey to each other (but not objectively).

  It's worth noting that as we describe 'a priori', an a priori statement stems from an apodictic one. But language is circular in that we use language to describe language.

Below is a summary table of all 4 ideas:

	Mill's Empiricism	Leibniz's formalism	Thomae's Game Formalism	Intuitionism
Infinitary	No	Yes	Yes	Yes
A Priori	No	Yes	Yes	Yes
Apodictic	No	Yes	Yes	No
Necessary	No	Yes	Yes	Yes
Universally Applicable	Yes	No	No	Yes
Indispensable	Yes	No	No	No
Inter-subjectively Robust	Yes	Yes	Yes	No

Unfortunately, none of these characterisation satisfy all parameters individually.

1.3 Benacerraf's Dilemma

To aid in the formulation of Benacerraf's Dilemma and its consequences, I have defined a term below:

Definition (formal) 1.13 (epistemology)**.** The study of knowledge, and what defines something to be 'known'.

From there we can express Benacerraf's Dilemma

Idea 1.5 (Benacerraf's Dilemma)**.** Epistemically speaking, something is known if we have a causal link between that which we already know. However, there is no causal link between mathematical facts and the way in which we express them.

Consider the three statements below:

- There are 2 Tube Stops between Brixton and Victoria
- There are two prime numbers between 1 and 4
- $\exists x \exists y (x \neq y \wedge F(x) \wedge F(y) \wedge R(a, x, b) \wedge R(a, y, b))$ (with the obvious semantics)

Each of these are related in to arithmetic in some way; and we are able to know (note the intentional epistemic link) what each statement conveys without there being a direct link to the numbers applied to them. For example, we refer to 'Tube Stops' in the first statement and assign it the number '2'.

However '2', under the relevant perspectives on mathematics (e.g **Idea 1.2**), is an abstract notion (or under some perspectives, an abstract object); how do we know:

- *How* do our justify the link/our arithmetic beliefs (side note: what constitutes a satisfactory justification)?
- How can we even *have* arithmetic beliefs?
- How do we even *express* them?

Note that, despite Benacerraf's original interpretation, this isn't entirely epistemic, and can also be viewed in the context of the link between the formal language.

1.4 Robinson's Arithmetic

Recall the definition of a theory.

Definition (formal) 1.14 (Theory)**.** A set of sentences in first order logic

In arithmetic we can express some further parameters for our characterisation:

- Axiomatisable
- Structural
- Incomplete

We will formally define these in later sections. For now, we will broadly accept the below definition:

Definition (informal) 1.15 (Axiomatisable)**.** The theory of mathematics can be broken down into a (not-necessarily finite) number of statements which we will call 'axioms'.

Robinson proposed the below set of axioms for arithmetic:

Definition (formal) 1.16 (Robinson Arithmetic)**.** Robinson arithmetic, called **Q** is defined by the axioms below:

- Axioms for successor

  Q1 $\forall x \forall y (x' = y' \rightarrow x = y)$

  Q2 $\forall x (0 \neq x')$

  Q3 $\forall x (x = 0 \vee \exists y (x = y'))$

- Axioms for Addition

  Q4 $\forall x (x + 0 = x)$

  Q5 $\forall x \forall y (x + y' = (x + y)')$

- Axioms for multiplication

  Q6 $\forall x(x \times 0 = 0)$
  Q7 $\forall x \forall y(x \times y'(x \times y) + x)$

- Debatably a definition, but we have the axioms for inequality

  Q8 $\forall x \forall y(x < y \iff \exists z(z' + x = y))$

An obvious observation is the lack of digits to resemble numbers. Instead, we denote an everywhere-defined function $\square'$ to denote numbers as successors (+1) of 0. I.e we can intepret 3 to be $0'''$. Upon further inspection, Robinson's Arithmetic is incredibly weak, and cannot, for example, prove commutativity of addition ($\forall x \forall y(x + y = y + x)$). However, it is adequate to prove the result of Gödel's Incompleteness Theorem, which we will do later in the module. This was instead proven with Peano Arithmetic.

1.5 Peano Arithmetic and induction

We can characterise Peano Arithmetic Below:

Definition (formal) 1.17 (Peano Arithmetic)**.** Peano arithmetic is Q and (or the union of) every instance of the induction scheme

$$[\phi(0) \land \forall x(\phi(x) \to \phi(x'))] \to \forall x(\phi(x))$$

We denote Peano Arithmetic by **PA**.

Remark 1.17.1 (Size of **PA**). Note that since the induction scheme is not a sentence, and instead a formula, **PA** is infinitely large, as we have to include *every instance* of ϕ.

Consider a bunch of statements numbered starting from 0. Induction proves that if any statement can be proven from the statement before it, and you have proven the first, you can prove them all. i.e if I prove statement 0, I know statement 1 is proven from statement 0, and that statement 2 is proven from statement 1 and so on and so forth. We can apply **PA** below.

Example 1.17.1. From **PA**, we can prove $\forall x(0 + x = x)$:

Proof.

$$\begin{aligned}
& \phi(x) :\equiv x + 0 = x && \\
& 0 + 0 = 0 && \text{(Q4)} \\
& \therefore\ \phi(0) && \\
\text{Suppose: } & 0 + x = x && \text{(i.e } \phi(x)) \\
& 0 + x' = (0 + x)' && \text{(Q5)} \\
& \quad = x' && \text{(by assumption)} \\
& \therefore\ \phi(x') && \\
& \therefore\ \forall x(\phi(x) \to \phi(x)) && \text{(As shown)} \\
& \therefore\ 0 + x = x && \text{(Using induction schema)}
\end{aligned}$$

□

One can attempt to prove induction using the theory of deduction:

Proof. Suppose $\phi(0)$ and $\forall x(\phi(x) \to \phi(x'))$. for any choice of n:

$$\begin{array}{lr}
\forall x(\phi(x) \to \phi(x')) & \text{(given)} \\
\phi(0) & \text{(given)} \\
\phi(0) \to \phi(1) & (\forall\text{E}) \\
\phi(1) & (\to\text{E}) \\
\phi(1) \to \phi(2) & (\forall\text{E}) \\
\phi(2) & (\to\text{E}) \\
\vdots & \\
\phi(n-1) \to \phi(n) & (\forall\text{E}) \\
\phi(n) & (\to\text{E})
\end{array}$$

□

This proof (minus the obvious aesthetic changes) is obviously valid, however it has a couple points worth considering:

- One could argue we have used induction by the repetition of a step (note the structure of our proof on the right is periodic), We could make the counterargument that we have instead reduced the argument of induction into a weaker statement about how we structure proofs.
- This does not agree with Mill's empiricism because it's infinitary and we therefore will not experience every proof.
- This is not necessarily a formal proof and instead a blueprint for proofs. But to characterise that if we don't treat this scheme as valid, we can construct a proof that breaks our rules, you need a stronger language (and I just did so in this sentence with English).

We can instead rewrite Peano Arithmetic in *Second-order logic*:

Definition (informal) 1.18 (Second-order logic)**.** First Order logic but we can also quantify over functions.

Example 1.18.1. Consider the statement "I am scared, and you are scared hence there is a feeling common amongst us". With the obvious semantics, we can write this as:

$$S(x) \wedge S(y) \vdash \exists F(F(x) \wedge F(y))$$

Definition (formal) 1.19 (Peano Arithmetic (alternative))**.** **PA** in second-order logic, is denoted $\mathbf{PA}_2$ and restates **PA** with the induction schema with a single axiom:

$$\forall Y[Y(0) \wedge \forall x(Y(x) \to Y(x'))]$$

2 Frege's Approach

Let us revisit the earlier definitions for the parameters we need to define mathematics:

- Infinitary
- Apodictic
- A priori
- Necessary
- Universally Applicable
- Indispensable
- Inter-subjectively robust

One apparent thing is that almost all of these are properties of logic. However it is not infinitary. Whilst it may have some infinite properties, in that domains need not be finite, it does not necessarily *need* to be. However, since we're almost there, Frege proposes that arithmetic is just an extension on logic. In order to deal with the inability to justify an infinitary characterisation Frege has to make the assertion that logic is infinitary.

Idea 2.1 (Frege's Logicism)**.** Arithmetic is an extension of logic, with the added benefit that logic is asserted to be infinitary

2.1 The underlying assumptions of Frege's Logicism

We will describe Frege's Logicism using our (naive) understanding of Second Order Logic. It is important to note that although we can be more rigorous in our characterisation of Second Order logic, out current understanding is sufficient in understanding the main issues associated to Logicism (i.e the ones we care about). We first define Frege's idea of a concept:

Definition (formal) 2.1 (concept)**.** A concept is defined to be a function that output True or False values. We know this, in the modern day, as a 1 place predication but in order to fit with Frege's initial ideas we denote these predications to be a property or characteristic of the variable that is applied to them.

Note 2.1. To supplement my personal and gain a stronger formal and symbolic understanding, I have deferred to "https://plato.stanford.edu/entries/frege-theorem/" to help pull Frege's ideas into a more tangible modern language,

Example 2.1.1. Fx can be interpreted as x belongs to the concept of "is a part of negative feeling" which we can denote as F.

The immediate issue here is that if we do have Fx, we are asserting that x belongs to a concept without specifying how, or even that it could exists. In a way, we are *defining* x to exist. One attempt to discredit this reasoning is to discredit as a premise of the arguments of logic; we are claiming this is a truth of logic, and we are attacking this claim. However, this is a weak argument, as there is no reason not to apply this to any other law of logic, such as Modus Ponens. There is an alternate method:

Idea 2.2 (Motives for Logic)**.** We could claim that logic itself is meant to be completely fundamental, and all encompassing, but by applying additional constraints on it, you lose this property. I.e anything that is given can also fail to exist. However we can fight back and claim that the necessity of arithmetic means that there is no such possibility where numbers don't exist.

2.2 Constructing Numbers and Frege's Theorem

Logicism defines number very similarly to how sets are compared. Sets are said to have the same size iff there is a one to one correspondence between their elements (you should note that there are many parallels between the "falling under a concept" and "belonging to a set"). Frege applies Hume's principle:

Definition (formal) 2.2 (Hume's Principle)**.** The number of Fs= the number of Gs iff there are as many Fs as Gs. I.e we comment about the object 'number' (we denote by #) by considering relations. This is formally defined in SoL:

$$\#F = \#G \leftrightarrow F \approx G,$$

where,

$$F \approx G := \exists R \forall x((F(x) \to \exists! y(G(y) \wedge R(x,y))) \wedge (G(x) \to \exists! y((F(y) \wedge R(y,x))))$$

Example 2.2.1. If we take F to mean a waiter's plates, and G to mean to a napkin, a waiter knows that there are an equal number of napkins and plates if they place them next to each other (denoted by the relation R).

Let us define two terms below to assess Hume's laws

Definition (informal) 2.3 (Analytic)**.** The truth of analytic propositions are determined solely by the definitions of the concepts that they include

Definition (informal) 2.4 (Synthetic)**.** Not Analytic

It may be tempting to view Hume's law as analytic, however the problem is that it doesn't in fact define a number, but only implicitly describes them. It shows that they are at least comparable. So it is debatable whether we wish to call Hume's law analytic. We can now begin to construct out numbers, starting with the characterisation of $\#F$:

Proposition 2.3. *Given some F,*

$$\exists x(x = \#F)$$

Proof.

$$\begin{aligned} F &\approx F && \text{(Using } R\text{=Id)} \\ \#F &= \#F && \text{(Using HP)} \\ \exists x(x &= \#F) && (\exists\text{I}) \end{aligned}$$

□

We can now define the numbers as below:

- **Definition (formal) 2.5** (Zero).

$$0 := \#H_0$$
$$H_0(x) :\equiv x \neq x$$

- **Definition (formal) 2.6** (One).

$$1 := \#H_1$$
$$H_1(x) :\equiv x = 0$$

- **Definition (formal) 2.7** (Two).

$$2 := \#H_2$$
$$H_2(x) :\equiv (x = 0) \vee (x = 1)$$

  Note we have used HP and our existence proposition to generate the definition. Concerningly, we will have to describe induction to prove this for all numbers.

We can offer an equivalent definition for 0 below:

Lemma 2.4.

$$0 = \#F \leftrightarrow \neg\exists x(F(x))$$

Proof. (→) For the sake of contradiction, suppose $\exists x(F(x))$ and $0 = \#F$.

$$\begin{aligned}
&F(a) && \text{(Where } a = x\text{)}\\
&a = a && (=\text{I})\\
&\exists x(F(x) \wedge x = x) && (\exists\text{I } \wedge\text{I})\\
&F \not\approx H_0 && \text{(by defintion)}\\
&\#F \neq \#H_0 && \text{(HP)}\\
&\#F \neq 0 && \text{(Transitivity of =)}\\
&\bot &&
\end{aligned}$$

(←) I want to prove, given the RHS, $\#F = \#H_0$; the LHS follows as shown

$$\begin{aligned}
&\neg\exists x(x \neq x) && \text{(Reflectivity)}\\
&\neg\exists x(H_0(x)) && \text{(By definition)}\\
&H_0 \approx F && \text{(By definition)}\\
&\#H_0 = \#F && \text{(HP)}\\
&0 = \#F && \text{(Transitivity of =)}
\end{aligned}$$

□

We can attempt to prove parts of **PA** by defining the successor

Definition (formal) 2.8 (Successor)**.**

$$\begin{aligned} F_a(x) &:\equiv (F(x) \land x \neq a) \\ m' = n &:\equiv \exists F \exists x(n = \#F \land F(x) \land (m = \#F_x)) \end{aligned}$$

We require a successor to have a concept that is associated with it, and that there is some x that falls under the concept (i.e non-0), and that the number that the successor succeeds corresponds to the number of the successor without the additional x.

Theorem 2.5 (Frege's Theorem for Q1 in **PA**)**.**

$$\forall x \forall y(x' = y' \to x = y)$$

Proof.

$$\begin{aligned} x' = y' &\equiv \exists F \exists a(y' = \#F \land F(a) \land (x = \#F_a)) \\ y' = x' &\equiv \exists G \exists b(x' = \#G \land G(b) \land (y = \#G_b)) \\ \#F = \#G & && (\exists \text{ E \& transitivity of } =) \\ F \approx G & && (\text{HP}) \end{aligned}$$

Let's call the one to one relation from G to F, S. Hence:

$$\exists \overline{a}(S(a, \overline{a}))$$
$$\exists \overline{b}(S(b, \overline{b}))$$

We can now define a new relation T such that $T = S$ except:

$$T(\overline{b}, \overline{a})$$

T works to show:

$$\begin{aligned} F_a &\approx G_b && (\text{Through } T) \\ \#F_a &= \#G_b && (\text{HP}) \\ x &= y && (\text{Transitivity of } =) \end{aligned}$$

□

Frege further proved that all PA axioms follow through from his definition, therefore satisfying the extra parameter, that mathematics is axiomatisable.

2.3 Flaws of the theory; The Julius Caesar Paradox

Kant critiqued Frege's work by pointing out that he assumed the existence of a function in Hume's principle. Hume's principle seemingly evokes the function out of nowhere. This wasn't the issue Frege had; he wanted a characterisation that proved the uniqueness of the function. From the definition alone, we can assign multiple definitions to 0; is 0 Julius Caesar for example? Or is 0 a Rhino?

The concern here isn't that we are unsure about whether Julius Caesar is a number; the concern is that we *know* that Julius Caesar is not a number, and therefore need a characterisation that rules that out. HP doesn't do this, hence HP does not give us everything we know about arithmetic. In his attempt, he defined an 'extension of a concept' as follows:

Definition (informal) 2.9 (Extension of a concept, F)**.**

$$F = \{x : F(x)\}$$

Note that this definition is informal as he hasn't defined sets extensions particularly well. He then explicitly characterised $\#F$ as follows:

Definition (informal) 2.10.

$$\#F = \{G : F \equiv G\}$$

The problem is, with his vagueness, it is unclear whether the Caesar Problem has been solved. Below is his (failed) attempt to resolve this:

Idea 2.6 (Basic Law V)**.**

$$\S F = \S G \leftrightarrow \forall x(F(x) \leftrightarrow G(x))$$

The idea of extensions is very similar to sets; this states that two extensions were equivalent if and only if the same things fell under their extensions. This idea fails in an irrecoverable way due to Russell's Paradox.

Theorem 2.7 (Russell's Paradox)**.** *Basic Law V leads to an inconsistency*

Proof. Frege stated $a \in b :\equiv (b = \S G \wedge G(a))$. Note that, by using Basic Law V, we can show $\forall F \exists \S F$, much like we did for HP. Russel suggested $R(x) := x \neq x . \S R \in \S R$ if and only if $R(\S R)$ by definition, but this is only true iff $\S R \notin \S R$. □

3 Dedekind Structuralism

3.1 Benacerraf's Identification Problem

Structuralism is a term with multiple definitions, so before we begin to discuss it, we need to define it. However, even before then we need to understand Benacerraf's Identification Problem. It is based on two approaches to defining natural numbers in set theory; assuming that set theory is logically motivated.

Idea 3.1 (Von-Neumann Ordinals)**.** Von-Neumann defined numbers as below

$$\begin{aligned} 0 &:= \emptyset \\ 1 &:= \{0\} \\ 2 &:= \{1, 2\} \\ &\vdots \\ n' &:= \{1, 2 \ldots, n\} \end{aligned}$$

Alternatively we have this definition:

Idea 3.2 (Zermello Ordinals)**.** Von-Neumann defined numbers as below

$$
\begin{aligned}
0 &:= \emptyset \\
1 &:= \{0\} \\
2 &:= \{2\} \\
&\vdots \\
n' &:= \{n\}
\end{aligned}
$$

Benacerraf acknowledged that this permitted there to exist an infinite number of ways to define natural numbers. That presents a problem:

Idea 3.3 (Benacerraf's Identification Problem)**.** Note the two disagree in a substantial way; for example, $0 \in 2$ takes different truth values depending on our definition. Benacerraf proposes this line of reasoning:

Premise 1 For both to be right is absurd, therefore at most one is.

Premise 2 There is no philosophical or mathematical reason to choose one over the other

Argument 1 Hence it would unreasonable to pick one over the other

Argument 2 Hence they are both wrong.

Conclusion 1 What matters in Arithmetic is the structure between numbers.

Conclusion 2 There cannot be an object-oriented definition of numbers; numbers are not objects.

There are some issues one may take with this argumentation:

- Just because we do not have justification for one, that doesn't necessarily mean that the answer isn't one or the other; it just states a limit on our ability to argue. (e.g Theseus' Ship). So Premise 2 doesn't necessarily entail argument 1.
- We have seen a characterisations of arithmetic, Logicism, that disproves this reasoning by refuting premise 2 (note that it failed for *other* reasons).
- He assumes that the ZFC framework holds. This is not too much of a problem, since we are disproving that such a framework generates numbers as objects; however since we have little knowledge on what specifically motivates ZFC, we ignore the possibility that the motivation specifies a true definition of numbers.

Definition (informal) 3.1 (Structuralism)**.** The believe in conclusion 1 and/or conclusion 2. Most believe in conclusion 1, but conclusion 2 is not as popular.

3.2 Dedekind's characterisation of the infinite

Dedekind maintains, much like Frege, that arithmetic is a consequence of logic, and that it governed the laws of thought. So, much like Frege, his aim was to tackle the infinite. In order to do this he needs to create a precise formulation. We can begin to understand his thinking by taking a look at a thought experiment:

Idea 3.4 (Hilbert's Hotel)**.** Hilbert proposed a Hotel with an infinite number of rooms, indexed by the natural numbers. Suppose someone new wants to book a room. Even if the hotel was fully booked he'd have no problem. All he would have to do is move everyone in room n to room $n+1$. Everyone still gets a room, with the added bonus of room 1 being free for the guest. Many people take to calling this a paradox; however if it were, it's easily resolved by the fact that no such hotel existss. Instead, it's moreso an illustration of the successor function, which mapped $\mathbb{N}_0$ to $\mathbb{N}_0 - \{1\}$.

Dedekind begins with a system, which is, in a way, as set. He doesn't particularly care what the system is (he's a structuralist); but he wishes to define a structure on it. He defines an injection on this system as follows:

Definition (formal) 3.2 (Injection)**.** An injection is a function, f, such that in a system,

$$\forall x \forall y (f(x) = f(y) \rightarrow x = y)$$

This is a fairly standard definition for an injection; in fact given we are working in FoL, we may be tempted to call a system the domain of interpretation, but this is quite anachronistic, and gets complicated when we go forward. If we call our system A, Dedekind defined it as infinite as follows:

Definition (formal) 3.3 (Dedekind-Infinite)**.** A set/system, A is Dedekind-infinite if there exists an injection f that maps from A such that there is some o such that $\forall x f(x) \neq o$

Note we quantified over a function, putting us in the domain of second order logic; making it harder to define a system as a domain of interpretation. Note that there are several notions of an infinte set, hence why we are specific when calling a set *Dedekind*-infinite. By accepting **PA** (Q1-Q3), and by using the successor function (as we did in Hilbert's Hotel), we have therefore stated that $\mathbb{N}_0$ is Dedekind-infinite. Now to get induction going, Dedekind considered 'minimal' infinite systems.

Definition (formal) 3.4 (closed sets)**.** For any function, f, a set B is closed if $\forall x \in B, f(x) \in B$. A set is f-closed iff the codomain of f is in the set.

We further define:

Definition (formal) 3.5 (closure function)**.** The closure function is intersection of all f-closed subsets of A:

$$\text{clo}_f(o) := \bigcap \{B \subseteq A : B \ f\text{-closed and } o \in A\}$$

This all leads to a Dedekind-algebra.

3.3 Dedekind's algebra

Definition (formal) 3.6 (Dedekind Algebra)**.** A Dedekind algebra is a structure with three components:

- A system/set A
- An object, $o \in A$
- An injection f such that $A = \text{clo}_f(o)$ and o is not in the range of f

This set of descriptions gives way to two theorems

Theorem 3.5. *All Dedekind infinite systems have a subsystem/subset that has a Dedekind Algebra structure.*

Theorem 3.6. *All Dedekind algebras satisfy the* ***PA*** *axioms.*

We will assume these without proof (although it is provable). There are two simple questions left to answer:

- Ontological: Does there exist *a* Dedekind-infinite system?
- Uniqueness: Is there only *one* Dedekind-infinite system?

The answer to the second is easy:

Proposition 3.7. *Any Dedekind-infinite system entails the existence of another.*

Proof. We can prove this by swapping o with some arbitrary n. More rigorously we can define a new function:

$$g(x) = \begin{cases} f(x) & \text{if } x \neq 0 \text{ and } x \neq n \\ f(o) & \text{if } x = n \\ f(n) & \text{if } x = 0 \text{ and } f(n) \text{ was defined} \end{cases}$$

This yields a new algebra in g. □

Now we may think that, as we have a more general version of Benaceraff's dilemma, very much like the Caesar problem. However, we must remember that Dedekind was a structuralist, and therefore didn't care about an explicit definition of the objects. He argued that making an explicit definition of numbers assigned them properties we didn't care about. For an explicit example, we could take 2 pencils to define the number 2, but that allows us to say '2 can be sharpened'. However, one could argue that this just avoids the question, and be viewed as a restatement that there is no mathematical or philosophical motivation for a specific or canonical definition which has flaws stated prior.

As stated, he only really cares about the structure of the system, not the objects. So what he wants is a structure that is preserved between systems:

Theorem 3.8. *All Dedekind algebras are isomorphic. That is given two Dedekind Algebras* (A_1, o_1, f_1) *and* (A_2, o_2, f_2)*, there is a bijection* $g : A_1 \to A_2$ *such that:*

$$g(o_1) = o_2$$
$$g(f_1(x))) = f_2(g(x))$$

This theorem is given without proof. To answer the ontological question, Dedekind proposes this system:

Idea 3.9 (Dedekind's first system)**.** The total number of things I can think of is infinite; we can call this S. Take any thought in S. The successor of S can be the thought x', which proves that the system is infinite

There are some flaws to this:

- We don't have a good characterisation on what a thought is; it might not even be abstract.
- S has only been described, but has never actually been constructed.
- This contradicts the physical constraints of reality, which tells us that there are only a finite (but incredibly large) number of states the brain can take.

4 The Frege-Hilbert Correspondence

We have explored the justification of arithmetic through Logic. However, Hilbert makes a striking proposition that consistency of a system (we return to our normal understanding of 'system') implies it's truth and/or existence. We won't necessarily explore what these mean, but we can develop out understanding from our vague idea.

4.1 Geometry

Euclid's elements are one of the most popular old examples of axiomatisation. In it, Euclid lays out 5 postulates and 5 axioms for geometry. The postulates are as follows:

1. You can draw a straight line between two points.
2. You can produce a finite straight line within a straight line.
3. You can describe a circle with any centre and distance
4. All right angles are equivalent
5. Two straight lines that intersect one another cannot be parallel to the same straight line

The 5th has been a point contention due to its open-endedness. The reason is because given two non-parallel lines, we would have to follow along one of them an arbitrary distance to find an intersection. Whilst discussing geometry, Hilbert and Frege discuss what an axiom is.

4.2 Frege's Axioms

Frege adopts this idea:

Idea 4.1 (Frege's characterisation of axioms)**.** An axiom represents a fact of intuition. It is a truth we assume to be correct that we cannot prove. The difference between a definition and an axiom is that a definition just lays out a language with which a claim can be made, but does not make an assertion on its truth. I.e I can define a unicorn to have a rainbow tail, but that doesn't mean that unicorns exist (although interestingly by this definition a chameleon is a unicorn). We must also require that all expressions used in writing an axiom must be completely understood.

Using Frege's axioms, we can express geometry.

Example 4.0.1 (Frege's Geometry)**.**

- We know what points and lines are, so we can take it as an *axiom* that two points determine a straight line.
- We can *define* a right angle to be the angle formed by two straight lines where the angles formed are equal in magnitude.
- The sum of interior angles of a triangle being the sum of two right angles is a fact with a correct yes/no answer.

In **PA**:

Example 4.0.2 (Frege's Arithmetic)**.** Ignoring his logicism.

- We know what expressions such as 0 and + mean. Hence we can interpret the **PA** axioms.
- We can make definitions such as:

$$x|y :\equiv \exists m(m \neq 0 \vee x \times m = y)$$
$$d = \gcd(x, y) := d|x \wedge d|y \wedge \neg\exists(d' > d)(d|x \wedge d|y)$$

- We can then make facts such as:

$$\{c|(a \times b), \gcd(a, b) = 1\} \models c|a \vee c|b$$

There are a couple issues with Frege's approach:

- Our own intuition can sometimes betray us; Frege's 'axiom 5' leads to Russel's paradox. However a Fregian would just argue that's a limitation of us, not the characterisation.
- Not all axioms are obvious, such as the axiom of choice.

4.3 Hilbert's axioms

Hilbert proposes this alternative:

Idea 4.2 (Hilbert's characterisation of axioms)**.** Hilbert argues that an axiom is an arbitrary selection of formulae; any collection will do. The symbolism comes after the fact in response to a need. He describes axioms as an implicit definition on the expressions that they contain and that each axiom changes the definition of the expression by recontextualising it.

For geometry

Example 4.0.3.

- Words such as 'point', 'line' or 'plane' have no particular meaning, and you can lay any axioms about them as you like.
- Your axioms define the expressions that they entail, but you can add further explicit definitions if you like
- You can always ask whether some sentence is entailed from your axioms, but they may or may not, and neither may their negation.

Frege suggests that Hilbert's idea doesn't decide whether his pocket watch is a point, but that is his point; every axiom is just a premise for the expressions they define.

4.4 Hilbertian Scaffolding

We can try and gain a better idea of what it means to have an implicit definition through **Toy**, which is a simple theory (recall a theory is a set of sentences which we will now refer to as axioms):

$$\mathbf{Toy} := \{R(a,b), R(b,c), R(c,a)\}$$

Let's illustrate **Toy** with the model (i.e the structures/interpretations), $\models$**Toy**:

$$\begin{aligned} D_{\mathcal{M}} &= \{\text{Daya}\} \\ |a|_{\mathcal{M}} &= |b|_{\mathcal{M}} = |c|_{\mathcal{M}} \\ |R|_{\mathcal{M}} &= \{(\text{Daya}, \text{Daya})\} \end{aligned}$$

We can instead illustrate this with a second order intuition of scaffolding, which allows us to quantify over variables and functions but also to form relations between them. so using Φ as an example:

$$\begin{aligned} \Phi(a,b,c,R) &: \equiv \bigwedge \mathbf{Toy} \\ &: \equiv (R(a,b) \wedge R(b,c) \wedge R(c,a)) \end{aligned}$$

Frege and Hilbert disagree on the relationship between truth/existence and consistency. Frege believes the former entails the latter, whereas Hilbert believes the converse. Given Frege assumes some axioms are true, he is able to prove that any entailed inconsistency negates the truth of at least one of those axioms. Consider that we interpret R in **Toy** to be 'strictly shorter than' and a, b, c can all be some three persons. This is obviously false in the Fregian sense, so the axioms cannot be true simultaneously. Hilbert says that the contradiction just mean that the axioms do not define truth. To prove consistency, we can approach this by using a higher order theory. For example, we just need to find entities witnessing **Toy**:

$$\exists R \exists a \exists b \exists c \left(\bigwedge \textbf{Toy} \right)$$

But we need a background theory from which to find these values; what resources do we have at hand, and what assumptions can we make. For **Toy**, this isn't particularly serious, but we have big issues if we look at more complicated theories in mathematics. The background theory must also be consistent.

1. Suppose we have two theories **U** and **T** such that **U**⊢"**T** is consistent"
2. If **U** is inconsistent it proves everything (by explosion [though not everyone agrees])
3. So to ensure **T** really *is* consistent, we need to prove that **U** is consistent

To terminate this link, we may think that we have to Fregean about some theories, however Hilbert may have another way to prove consistency beyond a parent theory.

5 Hilbert's programme

5.1 Aims for consistency

As we saw in the last section, Hilbert wanted to prove consistency in a system without a parent system, otherwise the burden of consistency just falls onto that system. The problem Hilbert wants to tackle is infinity; he believes that infinity is nowhere to be found in reality; often science rids itself of infinities, and for rational thought, paradoxes arise such as Russell's paradox. A major issue for mathematicians is calculus:

$$\text{Gradient of } f \text{ at 2: } f'(2) = \frac{f(2+h) - f(2)}{h}$$

where h is small. However, this is an approximation depending on how small $|h|$ is. The question follows whether h is a number, and if so, is it 0? Hilbert wasn't particularly worried about this; it was resolved by Cauchy some times later. However, this sort of issue led Hilbert to split mathematics in two:

- **Definition (informal) 5.1** (Finitary Mathematics)**.** "Ordinary elementary number theory" that does not depend on "actual infinity"; this is seen as consistent
- **Definition (informal) 5.2** (Ideal Mathematics)**.** Everything else; which means its consistency is in doubt.

It is Hilbert's view that, for example in Cantor's Set Theory, $\mathbb{N}_0$ is an actual infinity, as it is the totality of numbers; a holistic viewpoint. This would fall into Ideal Mathematics. Finite mathematics only handles the potentially infinite; the idea that you there is always a step afterwards after a finite start (i.e counting numbers). Hilbert is a game-formalist for ideal mathematics; but he is a 'term-formalist' for finitary mathematics:

Definition (informal) 5.3 (Term-Formalism)**.** Mathematics is about the manipulation of symbols, and so claims in mathematics express truths about manipulations.

5.2 Hilbert's method

Unlike in game-formalism, Hilbert believes that this allows us to express truths for finitary mathematics. Remember that game-formalism believes arithmetic is the arbitrary manipulation of symbols, but term-formalism believes that we have truths about *how* we are allowed to manipulate these symbols; this offers a much needed justification for the universal applicability that was missing when considering Thomae. To put it another way, we want Ideal mathematics to work like natural deduction for First Order Logic. For example, if we know $A \wedge B$ is true, and A is false, we can either use the truth table, or more quickly, natural deduction, to prove B. Similarly we have a similar scenario:

$$\begin{tikzcd}
\text{Ideal} \arrow[r] & \text{Ideal} \arrow[d] \\
\text{Finitary} \arrow[u] \arrow[r] & \text{Finitary}
\end{tikzcd}$$

Of course we need to know if Ideal mathematics is 'trustworthy'. Hilbert makes this more precise:

Definition (formal) 5.4 (Conservative)**.** $\mathbf{S}$ is conservative over $\mathbf{T}$ when, for any ϕ in the language of $\mathbf{T}$, $\mathbf{S} \vdash \phi \to \mathbf{T} \vdash \phi$

Hilbert defines finitary arithmetic as the theory **PRA**, 'Primitive Recursive Arithmetic'. He assumes that **PRA** is complete:

Definition (formal) 5.5 (Completeness)**.** For a statement, φ in $\mathbf{T}$, $\mathbf{T} \models \phi \to \mathbf{T} \vdash \phi$. Note this is equivalent to proving that a complete theory proves either a statement or its negation.

Definition (formal) 5.6 (Consistency)**.** A theory, $\mathbf{T}$, is consistent iff $\mathbf{T} \vdash \phi \to \neg(\mathbf{T} \vdash \neg\phi)$

Definition (formal) 5.7 (Extension)**.** A theory, $\mathbf{S}$, extends a theory $\mathbf{T}$ iff $\mathbf{T} \subset \mathbf{S}$

Proposition 5.1. *If* $\boldsymbol{T}$ *is consistent and* $\boldsymbol{S}$ *is conservative over* $\boldsymbol{T}$*, then* $\boldsymbol{S}$ *is consistent.*

Proof. Assume $\mathbf{S}$ is conservative over $\mathbf{T}$:

$$\begin{aligned}
\mathbf{S} \vdash \phi &\to \mathbf{T} \vdash \phi && (\mathbf{S}\text{ is conservative}) \\
&\to \neg(\mathbf{T} \vdash \neg\phi) && (\text{by consistency of }\mathbf{T}) \\
&\to \neg(\mathbf{S} \vdash \neg\phi) && (\mathbf{S}\text{ is conservative [contrapositive]}) \\
&\therefore\ \mathbf{S}\text{ is consistent}
\end{aligned}$$

□

Proposition 5.2. *If $\boldsymbol{T}$ is complete and $\boldsymbol{S}$ is consistent, then $\boldsymbol{S}$ is conservative over $\boldsymbol{T}$, assuming $\boldsymbol{S}$ extends $\boldsymbol{T}$.*

Proof. Assume for the sake of contradiction that **S** is not conservative over **T**; i.e that there is some contradictory statement ϕ that **S** proves but **T** does not

$$\begin{aligned}(\mathbf{S} \vdash \phi) \wedge \neg(\mathbf{T} \vdash \phi) & \\ \mathbf{T} \vdash \neg\phi & \qquad \text{(Completeness of } \mathbf{T}) \\ \mathbf{S} \vdash \neg\phi & \qquad (\mathbf{S} \text{ extends } \mathbf{T}) \\ \bot & \qquad (\mathbf{S} \text{ is consistent})\end{aligned}$$

□

We can use this to show that if **PRA** is consistent and complete, we can infer that Ideal mathematics is complete. All we need to do is construct a complete and consistent **PRA**.

5.3 Defining PRA

Hilbert decided to start with Stroke Notation:

- $| := 1$
- $|| := 2$
- $\underbrace{||...||}_{n \text{ times}} := n$

The numeral is what we call a token:

Definition (informal) 5.8 (Token)**.** A token represents a specific item representing a specific instance of the real thing

Definition (informal) 5.9 (Type)**.** A token represents the idea of something

Example 5.9.1. 'The canary is becoming common' vs 'The canary is in the cage'. In the former 'the canary' is a type, and in the latter, a token.

In stroke notation, we denote $\mathfrak{a} + \mathfrak{b}$ to be $\mathfrak{a}$ followed by $\mathfrak{b}$. We also define $\mathfrak{a} \times \mathfrak{b}$ to denote every stroke in a copy of $\mathfrak{a}$ with a copy of $\mathfrak{b}$. Given tokens (i.e writing these out ourselves) we can see this is true, which (possibly?) allows us to learn about the types. However we hit a problem when we hit quantifiers. Consider the below statement:

$$\text{For a prime } n, \exists p(p > n)$$

This statement is unbounded because we are implicitly quantifying p over every number more than n, which is infinite. This goes against, **PRA**, hence we are not allowed to quantify over infinite numbers. **PRA** would, however, allow the proof of this alternative statement:

$$\text{For a prime } n, \exists p(n < p < n! + 1)$$

Euclid's proof (kind of). If $1 < m \leq n$ then m does not divide $n! + 1$ so either $n! + 1$ is prime or there is some prime strictly between n and $n! + 1$ □

So what about claims such as $\mathfrak{a} + \mathfrak{b} = \mathfrak{b} + \mathfrak{a}$? We cannot view this as $\forall a \forall b (a + b = b + a)$. It functions similarly, but the use of fraktur letters are to signify meta-variables (i.e beyond FoL); these represent every instance of a specific statement that can be substituted in. So to say $\mathfrak{a} + \mathfrak{b} = \mathfrak{b} + \mathfrak{a}$, we mean to say that $1 + 3 = 3 + 1$ and $1 + 2 = 2 + 1$ and $5 + 2 = 2 + 5$ and so and so forth. Note that we cannot negate the statement since, should we (mistakenly) view it as an universal quantifier, we apply this rule: $\forall x(\phi(x)) \equiv \neg \exists x(\neg \phi(x))$, which we cannot do. Further note that $\mathfrak{a} + \mathfrak{b} \neq \mathfrak{b} + \mathfrak{a}$ can (intuitionistically) be viewed as $\forall x(\neg \phi(x))$. Hilbert tried by induction:

Proof. Assume WLOG $\mathfrak{b} > \mathfrak{a}$, so $\mathfrak{b} = \mathfrak{a} + \mathfrak{c}$ So the statement is equivalent to $\mathfrak{a} + \mathfrak{a} + \mathfrak{c} = \mathfrak{a} + \mathfrak{c} + \mathfrak{a}$ which just boils down to proving $\mathfrak{a} + \mathfrak{c} = \mathfrak{a} + \mathfrak{c}$. This holds by an assumed induction hypothesis. □

So despite Hilbert's restrictions he has assumed (strong) induction:

$$\forall y((\forall x < y)\phi(x) \rightarrow \phi(y)) \rightarrow \forall y \phi(y)$$

So Hilbert uses a schematic inference rule, a rule on schemas:

$$[\phi(0) \wedge (\phi(\mathfrak{a}) \rightarrow \phi(\mathfrak{a}'))] \rightarrow \phi(\mathfrak{b})$$

Hilbert now has the issue of justifying this rule.

Part 2: Gödel's Incompleteness Theorem

In this part, we will be using formal language a lot more, though our proofs will not be formalised as we have done before.

6 Fundamentals

6.1 Theories and Axioms

Definition (formal) 6.1 (The language of arithmetic)**.** Denoted, $\mathscr{L}_{\mathbf{A}}$, the Language of arithmetic has the primitives 0, $+$, $\times$, $<$, and the successor, $'$. Note that formally, x' is denoted as $'(x)$ and $x \times y$ as $\times(x, y)$.

One way of interpreting this language is with the standard model of arithmetic, $\mathcal{N}$:

Definition (formal) 6.2 (The Standard Model of Arithmetic)**.** Denoted $\mathcal{N}$, it is the interpretation over the language of arithmetic defined by $|\mathcal{N}| = \mathbb{N}_0$ (remember this is domain of objects, not cardinality), $0^{\mathcal{N}} = 0$, $+^{\mathcal{N}} = +$, $\times^{\mathcal{N}} = \cdot$, $<^{\mathcal{N}}=<$ where the defined terms match their obvious counterparts (note we interpret $\cdot$ as the arithmetic product).

Now we have a language and a model to interpret it, we now need a theory.

Definition (formal) 6.3 (Theory)**.** A theory is a set of sentences closed under entailment. That is to say that for a theory, $\mathbf{T}$, $A \in \mathbf{T} \to \mathbf{T} \models A$

Note 6.1. The module tends to use Γ as an arbitrary theory notationally speaking; to keep in touch with the earlier proofs, I will use $\mathbf{S}$ and $\mathbf{T}$.

Example 6.3.1 (**TA**)**.** An important example in this course is **TA**, and is defined as $\{A : \mathcal{N} \models A\}$. This can be confusing at first. Note that the model $\mathcal{N}$ isn't explicitly defined by a set of axioms or as a set, however, it implicitly has a concept of truth through our normal intuition on what the symbols are (e.g $3 \times 2 = 6$)

We can make this more optimal by introducing a set of axioms:

Definition (formal) 6.4 (Axiomatisation)**.** A theory $\mathbf{T}$ is axiomatised by a (sub)set $\mathbf{T}_0$ iff $A \in \mathbf{T} \to \mathbf{T}_0 \models A$

Example 6.4.1. Every theory axiomatises itself.

Example 6.4.2. The (pretty boring) finite theory axiomatised by $\{A, B, C\}$ is also axiomatised by $\{A \wedge B, C\}$

Note we have encountered a theory before, **Q** being Robinson's Arithmetic, axiomatised by the set of axioms we described. Similarly we have encountered Peano Arithmetic, **PA**, which is **Q** with the addition of the induction schema (so note that **PA** is infinitely large!). Also note that since we are using first order logic, $\models$ and $\vdash$ are interchangeable by the soundness and completeness theorems. For the sake of getting everything together, we will redefine completeness and consistency (though the latter through its negation):

Definition (formal) 6.5 (Completeness)**.** A theory $\mathbf{T}$ is complete iff $\mathbf{T} \models \neg A$ or $\mathbf{T} \models A$ for any sentence A in its language.

Definition (formal) 6.6 (Inconsistency)**.** A theory $\mathbf{T}$ is inconsistent iff $\mathbf{T} \models \neg A$ and $\mathbf{T} \models A$ for some sentence A in its language.

Note the similarities between the logical shape of these definition. We know that $\mathcal{N} \models \mathbf{PA}$, i.e that $\mathbf{PA} \subset \mathbf{TA}$. So we can attempt to express a condition on their equivalence

Theorem 6.1. *$\boldsymbol{PA} = \boldsymbol{TA}$ iff $\boldsymbol{PA}$ is complete*

Proof. ($\rightarrow$) Note that since $\mathcal{N}$ models $\mathscr{L}_{\mathbf{A}}$, it is complete in the language, so $\mathbf{PA}$ is complete by equivalence.
($\leftarrow$) We know $\mathbf{PA} \subset \mathbf{TA}$ so it suffices to prove $\mathbf{TA} \subset \mathbf{PA}$. We can do this by noting that if $A \in \mathbf{TA}$, then $\neg A \notin \mathbf{TA}$ by completeness and so $\neg A \notin \mathbf{PA}$ by subsethood. But by assuming $\mathbf{PA}$ is complete, we have $A \in \mathbf{PA}$, as required. □

6.2 Decidability, and Computational Enumerability

We will define an idea of computation later; for this section we will just assume that a computational procedure is a finite list of explicit steps that outputs something after a given input. We first begin by defining decidability

Definition (formal) 6.7 (Decidability)**.** A set X is decidable iff there is some computational procedure that determines whether an input x is in X (by outputting 1) or not (by outputting 0).

Example 6.7.1. All finite sets are decidable. Just assign the set an arbitrary ranking, and for any input, compare it along the list. This comparison will end in a finite number of steps.

Example 6.7.2. All inconsistent theories are decidable because explosion means that it proves everything. So a procedure that outputs 1 always works.

Naturally the first instinct is to define any theory as axiomatisable if it axiomatised by some set, however this isn't necessarily useful (as any theory can axiomatise themself). Hence we reserve this term as below

Definition (formal) 6.8 (Axiomatisability)**.** A theory is axiomatisable iff it is axiomatised by a *decidable* set of axioms.

Example 6.8.1. Every finite theory axiomatises itself, and as all finite sets are decidable, it is therefore axiomatisable. For example, $\mathbf{Q}$.

Example 6.8.2. Infinite sets of axioms that can be written as an axiom schema (such as the axiom schema of indction) are also decidable since, for any input we have a finite list of axioms to compare it with and a finite list of schemata to check if it matches the shape of. Hence $\mathbf{PA}$ is axiomatisable.

We can also go for a weaker definition related to countability. Note that a set X is countable if there is an injection from $\mathbb{N}_0$ to X or if X is empty. We call this injection an enumeration.

Definition (formal) 6.9 (Computably Enumerable)**.** A set is computably enumerable (also described with c.e) iff there is computational procedure that is an enumeration.

Theorem 6.2. *If **T** is axiomatisable, it is computationally enumerable.*

Proof. Suppose **T** is axiomatised by some decidable set $\mathbf{T}_0$. We therefore have some algorithm, Φ, that tells us that whether a subset of $\mathbf{T}_0$ make up all the undischarged assumptions of an input proof. Since all proofs in the language can be ordered, we can define a ranking of all the groups (e.g alphabetisation orders words so vanity comes before varied in the dictionary). If we go through them and apply Φ, we can use it to enumerate the sentence in $\mathbf{T}_0$ it proves. □

Corollary 6.2.1. *If **T** is axiomatisable and complete, it is decidable*

Proof. If **T** is inconsistent, it proves everything (by explosion), hence everything is an element of **T**, making it decidable. If it is, we can, by **T** being axiomatisable enumerate everything in **T**. By consistency we know, for any A, either $\neg A$ or A will be found in finite time by counting along the enumeration. □

We use $\overline{n}$ to represent the nth canonical number; i.e $\overline{3} = 0'''$ in $\mathscr{L}_{\mathbf{A}}$. Recall that a formula is an expression that allows free variable, unlike a sentence. We can chacractise a representation of it in three ways

- **Definition (formal) 6.10** (Formula of a function)**.** A formula $A(x_1, \dots, x_k, y)$ *represents* a function $f : \mathbb{N}_0^k \to \mathbb{N}_0$ in **T** iff where $f(n_1, \dots, n_k) = m$

  1. $\mathbf{T} \vdash A(\overline{n_1}, \dots, \overline{n_k}, \overline{m})$

  2. $\mathbf{T} \vdash \forall y(A(\overline{n_1}, \dots, \overline{n_k} m y) \to y = \overline{m}$

- **Definition (formal) 6.11** (Formula of a relation)**.** A formula $A(x_1, \dots, x_k)$ *represents* a relation $R \subset \mathbb{N}_0^k$ iff

  1. If $R(\overline{n_1}, \dots, \overline{n_k})$, then $\mathbf{T} \vdash A(\overline{n_1}, \dots, \overline{n_k})$

  2. If not $R(\overline{n_1}, \dots, \overline{n_k})$, then $\mathbf{T} \vdash \neg A(\overline{n_1}, \dots, \overline{n_k})$

- **Definition (formal) 6.12** (Formula of set)**.** This is the same as the above definition but where a set represents a one-place relation.

We are now in a position to describe the theorem we aim to prove in this section:

Theorem 6.3 (Gödel's First Incompleteness Theorem: Partly)**.** *If **T** is a decidable and axiomatisable theory in $\mathscr{L}_{\boldsymbol{A}}$ which represents all decidable sets of naturals; if it is complete, it is not consistent.*

Proof. Suppose **T** is as described, and is complete, the set of formulas of sets in **T** is computably enumerable (as it is axiomatisable); we can index each such formula as $A_n(x)$. We can further define this set:

$$D = \{n \in \mathbb{N}_0 : (\neg A_n(\overline{n})) \in \mathbf{T}\}$$

We know D is decidable since we can computably enumerate the elements in **T**, and by completeness, it will either eventually print $A_n(\overline{n})$ or $\neg A_n(\overline{n})$. By decidability of **T** and D, there must be some formula, $A_m(x)$ such that:

- If $n \in D$, then $A_n(\overline{n}) \in \mathbf{T}$
- If $n \notin D$, then $(\neg A_n(\overline{n})) \in \mathbf{T}$

Suppose for the sake of contradiction, that $m \notin D$, then by definition $\neg((\neg A_m(\overline{m})) \in \mathbf{T})$ but this is a contradiction by our definition of $A_m(x)$ so $m \in D$ proving inconsistency as $A_m(\overline{m}), \neg A_m(\overline{m}) \in \mathbf{T}$. □

To make this proof more precise, we need to better define a theory of computation.

7 A Theory of Computation

7.1 Primitive Recursion

For the sake of proving the previous theorems (and the full Incompleteness Theorems) more precisely, we need to better define an algorithm. We can do this through primitive recursion. First, we need to define some primitive functions:

- $\text{zero}(x) = 0$
- $\text{succ}(x) = x + 1$
- $\mathrm{P}_i^n(x_0, \ldots, x_{n-1}) = x_i$ for $0 \leq i \leq n$. Note that when $n = 1$ we have the identity!

Note that some of these functions are fairly redundant in $\mathcal{N}$; for example we already have a successor operator. However, we can use these to create a lot more interesting functions.

Note 7.1. For ease of reading, I will denote $x_1, x_2, \ldots, x_n$ with $\mathbf{x}$

- **Definition (formal) 7.1** (Functions by Composition)**.** For functions f and g, we can define h to be the composition of f and g. For example, $a(x) = \text{succ}(\text{succ}(x))$. We can define composition for an n-placed function f and k-placed functions $g_1,\ g_2, \ldots, g_n$. We can define $h(\mathbf{x}) = f(g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_k(\mathbf{x}))$
- **Definition (formal) 7.2** (Functions by Recursion)**.** A function, h, is defined by *primitive recursion* from a k-placed function f and $k+2$-placed function g when:

  $$\begin{aligned} h(\mathbf{x}, 0) &:= f(\mathbf{x}) \\ h(\mathbf{x}, y+1) &:= g(\mathbf{x}, y, h(\mathbf{x}, y)) \end{aligned}$$

  We can see this is well defined due to the way $\mathbb{N}_0$ is structured.

  Example 7.2.1.

  $$\begin{aligned} \text{add}(x, 0) &:= x \\ \text{add}(x, y+1) &:= \text{succ}(\text{add}(x, y)) \end{aligned}$$

  Example 7.2.2.

  $$\begin{aligned} \text{mult}(x, 0) &:= 0 \\ \text{mult}(x, y+1) &:= \text{add}(\text{mult}(x, y), y) \end{aligned}$$

Example 7.2.3.

$$\begin{aligned}\text{power}(x,0) &:= 1\\ \text{power}(x,y+1) &:= \text{mult}(\text{power}(x,y),y)\end{aligned}$$

- **Definition (formal) 7.3** (Primitive Recursive Functions)**.** zero, succ P_i^n are all primitive recursive functions. A composition of primitive functions, and function defined by primitive recursion of primitive recursive functions are also primitive recursive.

Note that our examples of powers, multiplication and addition are all primitive recursive functions; explicit numbers work by composing successor on 0 (or more interestingly with multiplication as well; minimising this is known as integer complexity); also note that we can ignore inputs by use of a projection function (for example when defining something by primitive recursion we don't need to use *all* inputs of a function). Note that very obviously this shows that functions can have multiple definitions. This is fine! Two functions are equal if they agree on all outputs. To extend out definitions to relations:

Definition (formal) 7.4 (Characteristic Functions)**.** A relation, R's, characteristic function χ_R is defined:

$$\chi_R(\mathbf{x}) := \begin{cases}1 & \text{if } R(\mathbf{x})\\ 0 & \text{if } \neg R(\mathbf{x})\end{cases}$$

We call a relation primitive recursive iff its characteristic function is

We can further apply some logic to play about with it by mapping truth to 1 and falsity to 0.

Proposition 7.1. *Any expression in predicate logic (i.e using* $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$*) using primitive recursive relations is primitive recursive.*

Proof. It is known (and not particularly difficult to prove) that we can generate any of the symbols in predicate logic with a nor gate, denoted $\uparrow$, in which $(A \uparrow B) = (\neg(A \vee B))$. We can do that by defining, for relations P and Q, the relation below

$$\chi_{(P\uparrow Q)}(\mathbf{x}) := \text{power}(0, \text{add}(\chi_P(\mathbf{x}), \chi_Q(\mathbf{x})))$$

□

We can also define a piece-wise function with p.r outputs contingent on p.r relation conditions.

Proposition 7.2. *Given p.r functions* f_0*,* f_1*, …,* f_n *and* R_0*,* R_1*, …,* R_n *be primitive relations,* h*, defined below is also a primitive relation:*

$$h(\mathbf{x}) = \begin{cases} f_0(\mathbf{x}) & \text{if } R_0(\mathbf{x})\\ f_1(\mathbf{x}) & \text{if } R_1(\mathbf{x})\\ \dots & \dots\\ f_{n-1}(\mathbf{x}) & \text{if } R_{n-1}(\mathbf{x})\\ f_n(\mathbf{x}) & \text{Otherwise}\end{cases}$$

Proof. We can write it as an explicit composition of p.r functions:

$$\begin{aligned} h(\mathbf{x}) =& f_0(\mathbf{x}) \cdot \chi_{R_0}(\mathbf{x}) \\ &+ f_1(\mathbf{x}) \cdot \chi_{\neg R_0 \wedge R_1}(\mathbf{x}) \\ &+ \dots \\ &+ f_{n-1}(\mathbf{x}) \cdot \chi_{\neg R_0 \wedge \neg R_1 \wedge \dots \wedge R_{n-1}}(\mathbf{x}) \\ &+ f_n(\mathbf{x}) \cdot \chi_{\neg R_0 \wedge \neg R_1 \wedge \dots \wedge \neg R_{n-1}}(\mathbf{x}) \end{aligned}$$

□

Recall Hilbert's trouble with unbounded quantification. We can make this idea more precise:

Definition (formal) 7.5 (Bounded quantification)**.** An occurrence of $\forall \nu$ is bounded iff its scope is a formula of the form $(\nu < \tau \to A)$ where τ does not contain ν. Additionally, an occurrence of $\exists \nu$ is bounded iff its scope is a formula of the form $(\nu < \tau \wedge A)$ where τ does not contain ν

We abbreviate this by quantifying in the form $(Q\nu < \tau)A$ where Q is out quantifier.

Proposition 7.3. *For a p.r relation, $R(\boldsymbol{x}, z)$, $(\forall(z < \lambda)R$ and $\exists(z < \lambda)R$ are both also p.r*

Proof. It suffices to prove the universal case as we know that we can express the existential in terms of negation and the universal. Define the function below:

$$\begin{aligned} r(\mathbf{x}, 0) &:= 1 \\ r(\mathbf{x}, y+1) &:= r(\mathbf{x}, y) \cdot \chi_R(\mathbf{x}, y) \end{aligned}$$

$r(x, \lambda)$ is precisely a characteristic function for the relation of $\forall(z < \lambda)R$. We can observe this by example. Consider $r(\mathbf{x}, 3)$. The only way this could be 1 is if $r(\mathbf{x}, 2) \cdot \chi_R(\mathbf{x}, 2) = 1$ and this is true only if $r(\mathbf{x}, 1) \cdot \chi_R(\mathbf{x}, 1) = 1$ and this is true only if $\chi_R(\mathbf{x}, 0) = 1$. □

7.2 Bounded Searches

We define a special function below:

Definition (formal) 7.6 (Bounded Minimisation)**.** The bounded minimisation of a relation $R(\mathbf{x}, y)$ is as follows:

$$m_R(\mathbf{x}, y) = \begin{cases} z & \text{if } z < y \text{ is the least number such that } R(\mathbf{x}, z) \\ y & \text{otherwise} \end{cases}$$

It basically states the smallest value that satisfies the relation and if it doesn't, then it just restates the bound.

Proposition 7.4. *The Bounded minimisation of a relation $R(\mathbf{x}, y)$ is p.r*

Proof. Re-express it as follows:

$$m_R(\mathbf{x}, y) = \begin{cases} z & \text{if } (\forall c < z) \neg R(\mathbf{x}, c) \wedge R(\mathbf{x}, z) \wedge (x < y) \\ y & \text{otherwise} \end{cases}$$

Each condition is p.r, each output is p.r, and so this piece-wise function is p.r □

After a lot of work, we can now express more complicated expressions as p.r:

1. $x \mid y$ can be defined as $(\exists c \leq y)(x \cdot c = y)$.
2. The one-place relation prime(x) can be defined as $(\forall c \leq x)(c \mid x \to (c = 1) \wedge (c = x))$.
3. The function nextprime(x) allows us to find the next prime after x by use of Euclid's proof and bounded minimisation.
4. The function $\mathrm{p}(x)$ for the $x - 1$st prime can be defined as:

   $$\mathrm{p}(0) := 2$$
   $$\mathrm{p}(y+1) := \mathrm{nextprime}(y)$$

   $(\forall c \leq x)(c \mid x \to (c = 1) \wedge (c = x))$.
5. The power of the xth prime $\pi_x(y) := p_x^{y+1}$ (not to be confused with the prime counting function, which also is denoted using π).

Proposition 7.5. *A sequence can be coded as a single unique number through this expression:*

$$\langle a_0, \ldots, a_n \rangle \approx \pi_0(a_0) \cdot \pi_1(a_1) \cdot \cdots \cdot \pi_n(a_n)$$

Proof. By the fundamental theorem of arithmetic, not proven here, this is an injection into the naturals (and in actual fact a bijection if you vary n). □

Below are some useful p.r functions on sequences:

- len(s) returns the length of s
- append(s, n) adds n to the end of s
- element(s, i) returns the ith element of s or outputs 0 given a faulty choice of i.
- $s \frown t$ is the concatenation of the two sequences.

There are various other things we can do. For example since we can encode sequences of numbers as a single number, we can encode a sequence of sequences; remember we always know our encoding, so numbers can, if we're vague, be a number, a sequence and a sequence of sequences. A more interesting example is a tree.

Definition (formal) 7.7 (A Tree)**.** A tree is a sequence of the form $\langle k, t_1, t_2, \ldots, t_k, l \rangle$ where each t_i is also a (different) tree. This may seem a little bit of a circular definition, however recall we have no infinite notion, hence we must realise this definition is recursive.

We can potentially make this definition more well rounded by saying a tree cannot form loops, but if we use a well-founded (roughly, sets cannot contain themselves) set theory this is a given. We can view a tree as follows:

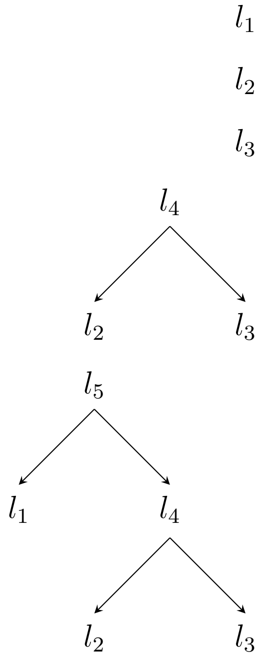
$$t_1 = \langle 0, l_1 \rangle$$

$$t_2 = \langle 0, l_1 \rangle$$

$$t_3 = \langle 0, l_1 \rangle$$

$$t_4 = \langle 2, t_2, t_3, l_4 \rangle$$

$$t_5 = \langle 2, t_1, t_4, l_5 \rangle$$



An interesting observation is that we don't actually need the first term as we can we just find len$(t) - 1$ to find the value. However the last term is very important. By allowing us to label each node we can actually classify proofs. The explicit description of how we can achieve this is omitted.

7.3 Defining Compatibility

Computability was described semantically; we don't have a rigorous definition for it. In this subsection we will finally define what it means. To do so, we shall 'prove' a couple things about primitive recursive functions.

Definition (informal) 7.8 (Bounded computability)**.** A boundedly computable function is a function that can be computed without an unbounded search.

Idea 7.6. Every initial p.r function is boundedly computable.

Proof. zero, and succ is obvious. The projection function eliminates data from a finite set, making it pretty obviously computably enumerable □

Idea 7.7. Functions defined from composition of boundedly computable functions are themselves boundedly computable

Proof. Suppose h is formed from composition of $g_1, g_2, \ldots, g_n$ on f; i.e we have that $h(\mathbf{x}) = f(g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_n(\mathbf{x}))$. We can compute g_i in a bounded way, and w can repeat this a finite number of times. Once we have all our g's together we can then evaluate f, which is boundedly computable. Since we did all this in a bounded manner, we can call h boundedly computable. □

Idea 7.8. Functions defined from primitive recursion on boundedly computable functions are themselves boundedly computable

Proof. For any input of such a function $h(\mathbf{x}, n)$ we can evaluate each instance of $h(\mathbf{x}, i)$ for i ranging from 0 to n. This is a bounded computation of bounded computable functions, so this must be boundedly computable. □

Corollary 7.8.1. *Every primitive recursive function is recursively enumerable.*

Proof. This is obvious as every p.r function is defined inductively from the initial functions through primitive recursion or composition; hence we can apply ideas 7.6 to 7.8 inductively too. □

By our motivated definition of p.r functions, that being that they are all c.e, we can prove the following result

Theorem 7.9. *Not all computable functions are p.r*

Proof. We shall use a diagonalisation argument. We know all p.r functions are c.e, so we can generate a list $f_0, f_1, \dots$; define this function:

$$D(x) := f_x(x) + 1$$

D is certainly computable as it is just 1 more than a c.e function. For the sake of contradiction, suppose it was p.r. That means that is f_k for some k. However that leads to this contradiction:

$$D(k) = f_k(k) = f_k(k) + 1$$

Hence D cannot be p.r. □

Corollary 7.9.1. *For any set of c.e functions, there is always a computable function not in that set. i.e the set of computable functions are not c.e.*

Proof. Apply the same argument as above, but instead of forming $f_0, f_1, \dots$ with p.r functions, form it with the set in question. □

We can extend p.r functions a bit more, to generalise computability a bit more. Define the functions below:

Definition (formal) 7.9 (Regular Function)**.** A function, $f(x, \mathbf{z})$ is a regular function iff

$$\forall \mathbf{z} \exists x (f(x, \mathbf{z}) = 0)$$

Definition (formal) 7.10 (Regular Minimisation)**.** For a regular function, $f(x, \mathbf{z})$, its (regular) *minimisation* is the function defined

$$\mu_f(\mathbf{z}) := \text{ The least } x \text{ s.t } f(x, \mathbf{z}) = 0$$

Note 7.2. There are two things to note:

1. Notationally, the course uses $\mu x[f(x, \mathbf{z})]$ but I have changed it to make familiarity easier.

2. This uses an unbounded search which is guaranteed to terminate by regularity of f.

We can extend to this to a relation by use of its characteristic function

Definition (formal) 7.11. Given a relation $R(x, \mathbf{z})$, if $\chi_{\neg R}$ is regular, then $\mu_R(\mathbf{z}) := \mu_{\chi_{\neg R}}$

We can now define general recursive functions as follows:

Definition (formal) 7.12 (General Recursive Functions)**.** A general recursive (g.r) function are all p.r functions and any function defined by minimisation on a *regular* g.r function, and nothing else.

Theorem 7.10 (Church Turing Thesis)**.** *The Church Turing Thesis is a mathematical-cum-philosophical argument that shows that all computable functions are, and only are, general recursive.*

To complete our theory we need to now create an idea of representability:

Definition (formal) 7.13 (Representability)**.** A function, $f : \mathbb{N}_0^k \to \mathbb{N}_0$ is representable in $\mathbf{T}$ iff there is a formula in $\mathbf{T}$'s language, $A(\mathbf{x}, y)$ such that for all numbers $\mathbf{n}$:

$$\mathbf{T} \vdash \forall y(A(\overline{\mathbf{n}}, y) \leftrightarrow y = \overline{f(\mathbf{n})})$$

We say A represents f in $\mathbf{T}$.

We will state the following theorem without proof:

Theorem 7.11. *A function is representable in* $\boldsymbol{Q}$ *iff it is g.r*

8 Arithmetisation

8.1 Codes and Gödel Numbers

Now that we have a theory for computation, we now want to be able to handle more interesting objects like we did with sequences and trees. In order to do this we will introduce what is known as a Gödel numbering. Before we do, we need a numbering system. Let's examine the following specification for symbols in $\mathscr{L}_\mathbf{A}$:

- Quantifiers: $\forall, \exists$
- Connectives: $\wedge, \vee, \rightarrow, \neg, \bot$
- Logical constants: $=$
- Improper symbols: brackets and commas
- Arithmetic symbols (0, $'$, $+$, $\times$, $<$)
- First order variables of which there is an infinite number

In order to make computations using these we need to assign these a number with which to operate with. Any will do, so long as we stick with it. Using sequences as natural numbers, we can code as follows:

- $c_\bot := \langle 0, 0\rangle$
- $c_\neg := \langle 0, 1\rangle$
- $c_\vee := \langle 0, 2\rangle$
- $c_\wedge := \langle 0, 3\rangle$
- $c_\to := \langle 0, 4\rangle$
- $c_\forall := \langle 0, 5\rangle$
- $c_\exists := \langle 0, 6\rangle$
- $c_= := \langle 0, 7\rangle$
- $c_(:= \langle 0, 8\rangle$
- $c_) := \langle 0, 9\rangle$
- $c_, := \langle 0, 10\rangle$

What we have done is encoding each symbol as a number represented by $c_\square$ for some symbol $\square$. Note we can enumerate symbols that represent variables, constants. predicates etc:

- The ith variable has the number associated with $\langle 1, i\rangle$
- The ith constant has the number associated with $\langle 2, i\rangle$
- The ith n-place function symbol has the number associated with $\langle 3, n, i\rangle$
- The ith n-place predicate symbol has the number associated with $\langle 4, n, i\rangle$

Definition (formal) 8.1 (Gödel Number)**.** *We* (other systems exist) call the Gödel number (g.n) of a sequence of symbols $s = s_1 s_2 \dots s_k$ the natural number associated with the finite sequence according to the aforementioned encoding (proposition 7.5). It is denoted by ${}^{\#}s^{\#}$

Example 8.1.1. Consider the sentence $\forall x_1(x_1 = x_1 + 0)$. The Gödel number of that sentence, ${}^{\#}\forall x_1(x_1 = x_1 + 0)^{\#}$ is given as follows:

$$
{}^{\#}\forall x_1(x_1 = x_1 + 0)^{\#} = \pi_0(c_\forall) \times \pi_1(c_{x_1}) \times \pi_2(c_() \times \pi_3(c_{x_1}) \times \pi_4(c_=) \times \pi_5(c_{x_1}) \times \pi_6(c_+) \times \pi_7(c_0) \times \pi_7(c_))
$$

And due to the 'I have a life theorem', I leave the calculation of this number as an excercise for the reader. To extend our vocabulary, we group together variables and constants under an umbrella term (see what I did there?):

8.2 Terms and Further Relations

Definition (formal) 8.2 ($\mathscr{L}$-Term)**.** A $\mathscr{L}$-term is a variable, constant, functions of variables and constants and nothing else.

Example 8.2.1. $\overline{1}$, $\overline{2}$, $\overline{2} + \overline{5}$, 0 and x_1 are all terms.

Definition (formal) 8.3 (Closed-Term)**.** A closed term is a term with no free variables.

We can now define a series of relations.

- **Definition (formal) 8.4** (Term relation)**.** Term(x) iff x is the g.n of an $\mathscr{L}_\mathbf{A}$-term

- **Definition (formal) 8.5** (Closed Term relation). ClTerm(x) iff x is the g.n of a closed $\mathscr{L}_{\mathbf{A}}$-term
- **Definition (formal) 8.6** (Numeral function). num(x):=$^{\#}\overline{x}^{\#}$
- **Definition (formal) 8.7** (Atom relation). Atom(x) iff x is the g.n of an atomic formula
- **Definition (formal) 8.8** (Formula relation). Frm(x) iff x is the g.n of an $\mathscr{L}_{\mathbf{A}}$-formula
- **Definition (formal) 8.9** (Sentence relation). Sent(x) iff x is the g.n of a sentence.
- **Definition (formal) 8.10** (Substitution relation). Subst($^{\#}A^{\#}$,$^{\#}t^{\#}$,$^{\#}u^{\#}$)= $^{\#}A[t/u]^{\#}$ iff x is the g.n of a sentence. where we denote $A[t/u]$ to be the result of replacing all free occurrences of u with t in formula A

Proposition 8.1. *All the above relations and functions are p.r.*

Proof sketches. Consider the n-long string $\perp\perp\ldots\perp$. Its gödel number obviously is the smallest possible gödel number of length n. So by choosing n such that the upper bound is larger than our input we can generate every possible "blank" length n or less using the rules of being a "blank" and calculate their gödel number which doesn't require an unbounded search. This reasoning applies to Term, ClTerm, Atom , Frm, and Sent. num and Subst are also p.r by very composition and a finite search & replace respectively. □

Since we can regard natural deductions as trees, with labels being sentences in our language, we c an regard natural deductions as numbers. More specifically, we can talk about the Gödel number of a *derivation*.

- **Definition (formal) 8.11** (Derivation function). Deriv(x) iff x is the g.n of a valid derivation.
- **Definition (formal) 8.12** (Proof relation). $\mathrm{Prf}_{\mathbf{T}}(x,y)$ iff x is the g.n. of a proof from $\mathbf{T}$ of the sentence with g.n. y

Proposition 8.2. *The above two relations are also p.r*

Sketch of why. By using the same idea as before, we can program a machine using only bounded searches to determine whether a line of reasoning is valid. □

9 The First Incompleteness Theorem

The plan for proving the first incompleteness theorem is to try and find a sentence, $G_{\mathbf{T}}$ that says 'I am not provable in $\mathbf{T}$'. The important thing to note is that there is no self reference to it, as we will see later. The contradiction can therefore, in some sense, be considered much richer than something akin to the liar's paradox, the sentence, "This Statement is false". For concision, we will use the following notation:

Definition (formal) 9.1. $\ulcorner A \urcorner$ We define $\ulcorner A \urcorner$ to be The numeral of a Gödel expression of A. i.e $\ulcorner A \urcorner = \overline{^{\#}A^{\#}}$. This allows us to keep to the language of $\mathscr{L}_{\mathbf{A}}$

9.1 The Diagonalisation Lemma

As we did in our mini-proof (Theorem 6.8), we will create a concept of diagonalisation.

Definition (formal) 9.2 (Diagonalisation)**.** The diagonalisation of any formula with a single free variable, $A(x)$, is the formula $A(\ulcorner A \urcorner)$. Note that $A(x)$ is a notational convention to tell us that there is a single free variable. However, in our coding, A would be what our encoding will take in.

We can now try and represent this as a number:

Definition (formal) 9.3 (Diagonalisation formula)**.** Let $\mathrm{di} : \mathbb{N}_0 \to \mathbb{N}_0$ be the function that maps the g.n. of a formula to the g.n. of its diagonalisation (with a 0 for a dummy output where the number doesn't represent a formula). This means that $\mathrm{di}({}^{\#}A^{\#}) = {}^{\#}A(\ulcorner A \urcorner)^{\#}$. It's easy to see that di is p.r. meaning by representation there is a formula $\mathsf{dia}(x, y)$ that represents it.

We can now state the diagonalisation lemma:

Lemma 9.1 (The Diagonalisation Lemma)**.** *For each $\mathscr{L}_{\mathbf{A}}$ formula, $B(x)$, there is an $\mathscr{L}_{\mathbf{A}}$ sentence, A such that $\boldsymbol{Q} \vdash B(\ulcorner A \urcorner) \leftrightarrow A$*

This is, more or less, the crux of the proof of Gödel. It is incredibly powerful.

Example 9.3.1. Let $T(x)$ be any one-place $\mathscr{L}_{\mathbf{A}}$-formula with x as its only free variable. For any $\mathscr{L}_{\mathbf{A}}$, sentence, X, there is a $\mathscr{L}_{\mathbf{A}}$ sentence, Y, such that:

$$\mathbf{Q} \vdash X \leftrightarrow T(\ulcorner Y \urcorner) \leftrightarrow Y$$

We can prove this by applying the diagonalisation lemma to $\mathscr{L}_{\mathbf{A}}$-formula $B(x) := X \leftrightarrow T(x)$.

Proof Sketch. Let's first prove the lemma for a theory $\mathbf{T}$, which enriches $\mathbf{Q}$ with a new function symbol, δ, which has the following property:

$$\mathbf{T} \vdash \delta(\overline{n}) = \overline{\mathrm{di}(n)}$$

Define the following:

- $E(x) := B(\delta(x))$
- $A := E(\ulcorner E \urcorner)$

Note, we have described A as the diagonalisation of E. Using our assumptions we can make the following steps:

$$\begin{aligned}
\mathbf{T} \vdash \delta(\ulcorner E \urcorner) &= \delta(\overline{{}^{\#}E^{\#}}) \\
&= \overline{\mathrm{di}({}^{\#}E^{\#})} \\
&= \overline{{}^{\#}E(\ulcorner E \urcorner)^{\#}} \\
&= \ulcorner A \urcorner
\end{aligned}$$

Hence $\mathbf{T} \vdash A \leftrightarrow E(\ulcorner E \urcorner) \leftrightarrow B(\delta(\ulcorner E \urcorner)) \leftrightarrow B(\ulcorner A \urcorner)$ □

The problem with this proof is that there is no such formula δ in $\mathbf{Q}$. However we do have Dia; it makes the proof a little more complicated, but we can use it in place of δ:

Proof. Let us define A as before, but define E using Dia:

$$E(x) := \exists y(\mathsf{Dia}(x, y) \land B(y))$$

Recall, by the definition of representability:

$$\mathbf{T} \vdash \forall y(\mathsf{Dia}(\ulcorner E \urcorner, y) \leftrightarrow y = \overline{\mathrm{di}({}^{\#}E^{\#})})$$

From the definition of di we have:

$$\mathbf{T} \vdash \forall y(\mathsf{Dia}(\ulcorner E \urcorner, y) \leftrightarrow y = \overline{{}^{\#}A^{\#}})$$

We can rewrite this as:

$$\mathbf{T} \vdash \forall y(\mathsf{Dia}(\ulcorner E \urcorner, y) \leftrightarrow y = \ulcorner A \urcorner)$$

In order to prove $\mathbf{T} \vdash A \leftrightarrow B(\ulcorner A \urcorner)$ we will work from one direction to the other.

($\rightarrow$): $A := E(\ulcorner E \urcorner)$ by definition. Now all that's left is to figure out what y is in the definition of E. However, we have by above that $\mathsf{Dia}(\ulcorner E \urcorner, y) \leftrightarrow y = \ulcorner A \urcorner$, giving us $B(\ulcorner A \urcorner)$

($\leftarrow$): Take $\ulcorner A \urcorner$ in above. By conjoining it to $B(\ulcorner A \urcorner)$ with $\land$ and generalising with $\exists$, we have $E(\ulcorner E \urcorner) = A$.

□

Note that we haven't done any self reference here. All these are formulae are only to do with numbers.

9.2 Gödel's First Incompleteness Theorem

Let $\mathbf{T} \supseteq Q$ be axiomatisable, meaning it represents all g.r functions and $\mathrm{Prf}_{\mathbf{T}}(x, y)$ is computable. Let $\mathsf{Prf}_{\mathbf{T}}$ represent $\mathrm{Prf}_{\mathbf{T}}$ in $\mathbf{T}$.

Definition (formal) 9.4 (Provability function)**.** Let $\mathsf{Prov}_{\mathbf{T}}(y) := \exists x \mathsf{Prf}_{\mathbf{T}}(x, y)$. This states that there is a $\mathbf{T}$-proof of the sentence with a g.n. of y.

Definition (formal) 9.5 (The Gödel Sentence)**.** The Gödel Sentence, $G_{\mathbf{T}}$, of a theory $\mathbf{T} \supseteq Q$ is the sentenced formed from diagonalising $\neg\mathsf{Prov}_{\mathbf{T}}$, giving it the following property

$$\mathbf{T} \vdash \neg\mathsf{Prov}_{\mathbf{T}}(G_{\mathbf{T}}) \leftrightarrow G_{\mathbf{T}}$$

The main takeaway of $G_{\mathbf{T}}$ is that if it is true, then there is no proof of it. More precisely, there is no natural number that is an encoding of its proof.

Lemma 9.2. *If $\boldsymbol{T} \supseteq \boldsymbol{Q}$ is computable and axiomatisable then $\boldsymbol{T} \nvdash G_{\boldsymbol{T}}$.*

Proof. Consider, for contradiction, that $\mathbf{T} \vdash G_{\mathbf{T}}$. From the properties of $G_{\mathbf{T}}$, we can infer that $\mathbf{T} \vdash \neg\mathsf{Prov}_{\mathbf{T}}(G_{\mathbf{T}})$. Breaking this apart from definitions we get that $\neg\exists x \mathsf{Prf}(x, \ulcorner G_{\mathbf{T}} \urcorner)$. However we know that there is a proof of $G_{\mathbf{T}}$ in $\mathbf{T}$ from our original assumption, and hence there must be a corresponding number encoding it. This is a contradiction. □

For this course, we will only be proving Gödel's incompleteness with ω-consistency. It is defined below:

Definition (formal) 9.6 (ω-consistency)**.** $\mathbf{T}$ is ω-consistent iff, for each $\mathscr{L}_{\mathbf{A}}$ formula A, if $\mathbf{T} \vdash \neg A(\overline{n})$ for each $\overline{n}$ then $\mathbf{T} \nvdash \exists x(A(x))$

Note that we are quantifying outside and within in the definition.

Remark 9.6.1. A few remarks:

- ω-consistency entails consistency by explosion.
- A theory can be consistent but not ω-consistent.
- No ω-inconsistent theory is true of the standard model, $\mathcal{N}$.

Lemma 9.3. *If $\boldsymbol{T} \supseteq \boldsymbol{Q}$ is ω-consistent and axiomatisable then $\boldsymbol{T} \nvdash \neg G_{\boldsymbol{T}}$.*

Proof. For contradiction, let's assume $\mathbf{T} \vdash \neg G_{\mathbf{T}}$. By consistency, we know that $\mathbf{T} \nvdash G_{\mathbf{T}}$, hence there is no number, $\overline{n}$ which is the g.n. of a proof of $G_{\mathbf{T}}$; for every possible $n \in \mathbb{N}_0$, $\mathbf{T} \vdash \neg\mathsf{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner G_{\mathbf{T}} \urcorner)$. By ω-consistency we have $\mathbf{T} \nvdash \exists x \mathsf{Prf}_{\mathbf{T}}(x, \ulcorner G_{\mathbf{T}} \urcorner)$ hence, $\mathbf{T} \nvdash \mathsf{Prov}_{\mathbf{T}}(\ulcorner G_{\mathbf{T}} \urcorner)$ and so $\mathbf{T} \nvdash \neg G_{\mathbf{T}}$ by the definition of $G_{\mathbf{T}}$. □

Theorem 9.4 (Gödel's First Incomplentess Theorem (GIT1))**.** *If $\boldsymbol{T} \supseteq \boldsymbol{Q}$ is ω-consistent and axiomatisable, then $G_{\boldsymbol{T}}$ is independent to $\boldsymbol{T}$.*

The inclusion of ω-consistency is annoying, and there are a few things GIT1 does not tell us. Assuming $\mathbf{T}$ is sound, $\mathcal{N} \models \mathbf{T}$ (hence $G_{\mathbf{T}}$ is true $[\mathcal{N} \models G_{\mathbf{T}}]$), we know $\mathbf{T} + G_{\mathbf{T}}$ is sound, axiomatisable and extends $\mathbf{Q}$, meaning it too is incomplete. However we cannot say the same for $\mathbf{T} + G_{\mathbf{T}}$ as it may be unsound. The full theorem is as follows:

Theorem 9.5 (Gödel-Rosser)**.** *No theory, $\boldsymbol{T}$ is such that*

- $\boldsymbol{T} \supseteq \boldsymbol{Q}$
- *$\boldsymbol{T}$ is axiomatisable*
- *$\boldsymbol{T}$ is complete*
- *$\boldsymbol{T}$ is consistent*

Corollary 9.5.1. ***TA** is not axiomatisable*

10 The Second Incompleteness Theorem

Gödelian reasoning led to us encoding information about $\mathbf{T}$ inside of $\mathbf{T}$ itself. To further this, we need to examine more about $\mathsf{Prov}_{\mathbf{T}}$.

Definition (formal) 10.1 (The modal operator)**.** To abbreviate $\mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner)$, we write $\Box_{\mathbf{T}} A$, and additionally, as we can infer the theory from context, write $\Box A$.

This is a specific instance of $\Box$ in the wider context of logic, and has general properties, even when not used specifically to this definition.

10.1 Modal Reasoning

Recall that we call $\mathbf{T}$ inconsistent iff $\mathbf{T} \vdash \bot$. Hence we can use out $\Box$ notation as follows:

Definition (formal) 10.2 (Consistency)**.** We can denote inconsistency of a theory as $\Box\bot$, and the consistency as $\neg\Box\bot$. We sometimes write $\mathsf{Con}_{\mathbf{T}}$ for the latter.

Recall that $\mathsf{Con}_{\mathbf{T}}$ is still an $\mathscr{L}_{\mathbf{A}}$ sentence, chosen arbitrarily by our encoding. We are still talking about $\mathbf{T}$ within $\mathbf{T}$. In order to prove things using $\Box$, we will need to list some of their properties.

Definition (formal) 10.3 (Löb Derivability Conditions)**.** Below are some properties of $\Box$:

Löb 1 If $\mathbf{T} \vdash A$, $\mathbf{T} \vdash \Box A$

Löb 2 $\mathbf{T} \vdash \Box(A \to B) \to (\Box A \to \Box B)$

Löb 3 $\mathbf{T} \vdash \Box A \to \Box\Box A$

Below is a useful condition derived from the first 3.

Proposition 10.1 (Löb 4)**.** *If $\boldsymbol{T} \vdash A \to B$, and $\boldsymbol{T} \vdash \Box A$ then $\boldsymbol{T} \vdash \Box B$.*

Proof. Apply Löb 1 to $\mathbf{T} \vdash A \to B$ to get $\mathbf{T} \vdash \Box(A \to B)$. Use Löb 2 to get $\mathbf{T} \vdash \Box A \to \Box B$. Modus ponens the second assumption to get the result. □

We will now introduce a theory between $\mathbf{Q}$ and $\mathbf{PA}$. In order to do so, let us define a class of formulae:

Definition (formal) 10.4 (Σ^1)**.** Formulae of the form $\exists \mathbf{x}(\phi(x))$ where $\exists$ is bounded are called Σ^1

We can now define a strengthening of $\mathbf{Q}$ as follows:

Definition (formal) 10.5 ($I\Sigma_1$)**.** We define this theory as $\mathbf{Q}$ along with an induction principle on all Σ_1 formulae.

Remark 10.5.1*.* $I\Sigma_1$ is axiomatisable

We can now state a criteria for the derivability conditions to be satisfied:

Theorem 10.2. *If $\boldsymbol{T} \supseteq I\Sigma_1$ is c.e.-axiomatisable, then $\Box_{\boldsymbol{T}}$ obeys the Löb derivability conditions.*

The proof of this has been omitted from this document.

10.2 Gödel's Second Incompleteness Theorem

The main point of Gödel's Second Incompleteness Theorem is that we want to show that **T** cannot prove its own consistency. That is, **T** proves that if it can prove its own consistency, it can prove $\bot$. The way we can do this is as follows:

Lemma 10.3 (Löb's Lemma)**.** *Suppose $\boldsymbol{T} \supseteq I\Sigma_1$ is c.e.-axiomatisable, then:*

- $\boldsymbol{T} \vdash \mathsf{Con}_{\boldsymbol{T}} \to G_{\boldsymbol{T}}$
- $\boldsymbol{T} \vdash \mathsf{Con}_{\boldsymbol{T}} \to \neg\Box\mathsf{Con}_{\boldsymbol{T}}$

Proof. Follow the reasoning:

$$\begin{aligned}
&\mathbf{T} \vdash G_{\mathbf{T}} \leftrightarrow \neg\Box G_{\mathbf{T}} && \text{(1. Property of } G_{\mathbf{T}}\text{)}\\
&\mathbf{T} \vdash G_{\mathbf{T}} \to (\Box G_{\mathbf{T}} \to \bot) && \text{(2. Proof by Logic)}\\
&\mathbf{T} \vdash \Box G_{\mathbf{T}} \to \Box(\Box G_{\mathbf{T}} \to \bot) && \text{(3. Löb 4)}\\
&\mathbf{T} \vdash \Box(\Box G_{\mathbf{T}} \to \bot) \to (\Box\Box G_{\mathbf{T}} \to \Box\bot) && \text{(4. Löb 2)}\\
&\mathbf{T} \vdash \Box G_{\mathbf{T}} \to \Box\Box G_{\mathbf{T}} && \text{(5. Löb 3)}\\
&\mathbf{T} \vdash \Box G_{\mathbf{T}} \to \Box\bot && \text{(6. Logically deduced from 3-5)}\\
&\mathbf{T} \vdash \neg\Box\bot \to G_{\mathbf{T}} && \text{(7. Logically deduced from 1,6)}\\
&\mathbf{T} \vdash \neg\Box\bot \to G_{\mathbf{T}} && \text{(8. Löb 4)}\\
&\mathbf{T} \vdash \neg\Box\bot \to \neg\Box\neg\Box\bot && \text{(9. Logically deduced from 6,8)}
\end{aligned}$$

The results follow from restating 7 and 9. □

We are now able to state Gödel's Second Incompleteness Theorem in full.

Theorem 10.4 (Gödel's Second Incompleteness Theorem (GIT2))**.** *No theory $\boldsymbol{T}$ is such that:*

- $\boldsymbol{T} \subseteq I\Sigma_1$
- $\boldsymbol{T}$ *is c.e.-axiomatisable*
- $\boldsymbol{T} \vdash \mathsf{Con}_{\boldsymbol{T}}$
- $\boldsymbol{T}$ *is consistent*

Proof. Assume all four for contradiction. By Löb's Lemma,, we have that $\mathbf{T} \vdash \mathsf{Con}_{\mathbf{T}} \to G_{\mathbf{T}}$. By the third point, we have $\mathbf{T} \vdash G_{\mathbf{T}}$, meaning **T** must be inconsistent. □

Corollary 10.4.1. *If $\boldsymbol{PA}$ is consistent, then $\boldsymbol{PA} \nvdash \mathsf{Con}_{\boldsymbol{PA}}$*

10.3 Consequences to Hilbert's Program

Let $\mathbf{I}$ be part of the ideal theory in Hilbert's arithmetic. Likely, Hilbertian's will want:

- $\mathbf{I}$ to be consistent
- $\mathbf{I}$ to be axiomatisable
- $\mathbf{I}$ to extend $I\Sigma_1$

However this means two things:

- $\mathbf{I} \nvdash \mathsf{Con}_\mathbf{I}$ by GIT2
- $\mathbf{PRA} \nvdash \mathsf{Con}_\mathbf{I}$ as $\mathbf{I} \supseteq \mathbf{PRA} \supseteq I\Sigma_1$

Therefore, to try and keep Hilbert's program afloat, we will need new ideas.

Idea 10.5 (Include infinitary rules)**.** Let us enrich our natural deduction with this rule, which we will denote the ω-rule:

$\phi(0), \phi(1), \phi(2)...$ entails $\forall x(\phi(x))$

Let us enrich $\mathbf{Q}$ with it, and call it $\mathbf{Q}^\omega$. This is extremely powerful; we can now prove $\mathbf{Q}^\omega = \mathbf{TA}$.

Proof sketch. Note that $\mathbf{Q}^\omega \subseteq \mathbf{TA}$. Firstly, we know that the axioms of $\mathbf{Q}$ are satisfied by $\mathcal{N}$, and we know the ω-rule is sound for $\mathcal{N}$ since, if $\mathcal{N} \models \phi(\overline{n})$ for each n, then $\mathcal{N} \models \forall x(\phi(x))$. To prove that $\mathbf{TA} \subseteq \mathbf{Q}^\omega$ you can induct over a 'prenex normal form' (standard expression of formulae with quantifiers on the outside). Where ϕ has no quantifiers, we know the answer. Now consider the induction case where we add a quantifier. Suppose $\mathcal{N} \models \forall x(\phi(x))$. Then $\mathcal{N} \models \phi(\overline{n})$ for each n. So, by the induction hypothesis, $\mathbf{Q}^\omega \vdash \phi(\overline{n})$ for each n. So $\mathbf{Q}^\omega \vdash \forall x(\phi(x))$ by the ω-rule. □

However, since the ω-rule is infinitary, we have gone beyond g.r computability, and this may contend with our existing perspectives on arithmetic.

Another idea we could have:

Idea 10.6 (Find an alternative sentence for $\mathbf{I}$)**.** We can do this by inducting over $\mathbf{PA}$ to generate a new theory, $\mathbf{PA}^*$. Iterating over the nth putative proof:

- If the proof results in a sentence, A with a negation proven for smaller n, it is not in $\mathbf{PA}^*$, Otherwise, it is.

There is no proof of a contradiction in $\mathbf{PA}^*$; it is arithmetisable meaning that $\mathbf{PA} \vdash \mathsf{Con}^*_\mathbf{PA}$, and if $\mathbf{PA}$ is consistent, $\mathbf{PA} = \mathbf{PA}^*$. However, we need to rely on $\mathbf{PA}$ being consistent, and therefore must have lost the Löb Derivability Conditions.

Additional Notes

In general, I haven't included notes on the optional slides which include the proof of **Q**'s representability, The Halting Problem, Gödel-Rosser and of the Löb Derivability conditions. If I ever have the time, I'll be sure to add them in. However, I feel as though, in order to better understand some of the topics, I should include some notes based off questions I have posed to Professor Button in person and via email and general independent research.

11 Model-Theoretic understanding

In order to better understand the relationship between $\mathcal{N}$ and **TA**, let us consider what it truly means to model something. The below definitions are a bit more in-depth and omit some structure with regards to examples:

Definition (formal) 11.1 (Language)**.** A language is merely a set of symbols; n-place symbols and terms. For example, we have $+(\cdot,\cdot)$ inside of the language of arithmetic, $\mathscr{L}_{\mathbf{A}}$, and constants such as $0, a, b, c \dots$

Definition (formal) 11.2 (Model)**.** A model is an assignment of meaning on the language. Consider $\mathscr{L}_{\mathbf{A}}$ on its own; it doesn't tell us that $3 + 2 = 5$; the model *does.* A model, could state that $+ := \{(2, 2, 4), (3, 2, 5) \dots\}$ (I am being a bit liberal with the distinction between $0'$ and 1 for example). $\mathcal{N}$ is the 'correct' version of mathematics. We can try and describe it with some other language but the bottom-line is we have to 'know' what we are aiming for.

Definition (formal) 11.3 (Theory)**.** We have already explained it, but a theory is a set of sentences in the language. Note that this is different to a model. $(3, 2, 5) \in +$ arises from a model, but the sentence $3 + 2 = 5$ does not (though of course we can observe some relation). This is a sentence in a theory.

Now that we have described a theory and a model, we can now ask what the relation between the two.

Definition (formal) 11.4 (Modelling)**.** We say $\mathcal{M} \models \mathbf{T}$ for a model $\mathcal{M}$ and theory $\mathbf{T}$ if the assignment of the model matches the sentences of the theory. We also want some recursion clauses to be satisfied, such as:

- $\mathcal{M}$ models $P\&Q$ iff $\mathcal{M}$ models P and $\mathcal{M}$ models Q
- $\mathcal{M}$ models $\neg P$ iff $\mathcal{M}$ does not model P
- $\mathcal{M}$ models $\exists x P(x)$ iff there is a constant c, such that $\mathcal{M}$ models $P(c)$.

Let us consider what **TA** is in context of the model of $\mathcal{N}$. We know that $\mathbf{TA} := \{A : \mathcal{N} \models A\}$, but we can generalise this idea:

Definition (formal) 11.5 (Complete Theory)**.** The complete theory, of a model, $\mathcal{M}$ is $\text{Th}(\mathcal{M}) := \{A : \mathcal{M} \models A\}$. **TA**, for example is $\text{Th}(\mathcal{M})$.

To prove **TA** is consistent, we will prove something more general.

Proposition 11.1. *Th($\mathcal{M}$) is consistent for all models $\mathcal{M}$.*

Proof. This is obviously true, because if it was inconsistent, it means Th($\mathcal{M}$) models $\perp$. But that must mean that the model models some P and $\neg P$ which we refuted in the recursion clause proves. □

As a result, we know **TA** is consistent, and so must be **PA**. The only question left is *if*, $\mathcal{N}$ exists. However, this is exactly the debate that we talked about in Part 1. In fact, Hilbert was concerned about the assumption of the infinitary side of $\mathcal{N}$, spurring the debate with respect to consistency. One big use of model theory is that, although it is anachronistic, we are better able to characterise Frege and Hilbert's ideas (here I work off Dummet):

Idea 11.2 (Frege)**.** A theory is true only if a model exists for it. Even if the theory is consistent, there must be a model.

Idea 11.3 (Hilbert)**.** A theory is true if it is consistent.

12 Second Order Logic

In general, we have been fairly ambiguous about how Second Order Logic Works. The reason is because there are many second order logics out there.

Definition (informal) 12.1 (Deductive Second Order Logic)**.** Most mathematicians and some philosophers use this type of second order logic. This logic allows one to quantify over predicates. For example, I can say something such as:

$$\forall X \exists Y \forall z (\neg X(z) \iff Y(z))$$

This states that every predicate has a negation. However, we must also entrich our deduction techniques (such as tertium non datur, modus ponens etc) by axioms such as the axiom schema of comprehension:

$$\exists Z \forall n (X(n) \leftrightarrow \mathfrak{A})$$

where $\mathfrak{A}$ is any formula not containing n.

Note 12.1. In the above logic, we can still arithmetise the variable predicates, and we can still encode proofs, allowing us to still apply GIT1.

Definition (informal) 12.2 (The Full Semantics)**.** Most philosophers in the past used this. The Full semantics are the same as above, however there is the added caveat that when quantifying, you are quantifying over every possibility, as opposed to first order, where you quantify over a domain.

Note 12.2. For above, GIT1 does not apply. Note that Dedekind proved that all Dedekind algebras are isomorphic and prove **PA**. Suppose some full model **PA2** $\not\models A$, for some $\mathscr{L}_{\mathbf{A}}$ sentence, A, hence no model of **PA2** $\models A$. This must mean all full models of **PA2** $\models \neg A$

13 A Small Bit of History

In general, the history is a bit convoluted, and only a few bits need mentioning. I have listed historical points that have come up in bullet-proof format:

- Why was Hilbert's Program promising if it never proved the consistency of anything interesting?
  The answer is simply because it did! First order real analysis is real, decidable and complete! There's also evidence to suggest Hilbert and his team were getting close or were mistakenly closer than they thought they were.

- Was Dedkind aware of Cantor's work on uncountable sets?
  Yes he was. In fact, they were incredibly good friends. A useful theorem you might want to know is Cantor's theorem, that a set, X is always smaller than its powerset, $\mathscr{P}(X)$. This is even true for infinite sets.

  Proof. Consider a surjective map, $f : X \to \mathscr{P}(x)$. Let $B := \{a \in X \mid a \notin f(x)\}$. We know that there must be some ξ such that $f(\xi) = B$. If $\xi \in B$ then $\xi \notin B$ and if $\xi \notin B$ then $\xi \in B$, a contradiction. □

14 My Essay

Below is my essay. I have included it as I aimed to be more technical with my topic, which is less likely to crop up amongst example essays (which themselves are uncommon).

To what extent is Frege's Logicism, and approach to axioms, reconcilable with Hilbert's Program and approach?

Daya Nidhan Singh

June 28, 2024

Introduction

Against criteria to follow, this essay shows that Hilbert and Frege's views on arithmetic cannot be reconciled to provide a description of it that can utilise Hilbert's program in a fulfilling capacity. Our method is to justify a process that picks parts of their views and *then* attempt to justify consistency between them, or failing that, make adjustments to resolve shortcomings. The idea follows:

- Frege's Theorem proves that some Second-order Logic (SoL) with Hume's Principle (HP) embeds Peano Arithmetic (**PA**). Here, we take Frege's perspective of "True ⇒ Consistent" (Frege and Gabriel, 1980) to prove the consistency of **PA**.
- Hilbert's Term Formalism resolves the Julius Caesar Problem (Linnebo, 1981).
- With SoL+HP being true and consistent, we can use it in Hilbert's Programme, with a 'Consistent ⇒ True' view, over Elementary Finitary Arithmetic (**EFA**).

The upside is that we can justify **PA** as apodictic, a priori, infinitary etc, but also as consistent without needing it to prove its own consistency (which is impossible by Gödel's Second Incompleteness Theorem (G2IT)). We will weigh this plan against the following criteria:

Necessity Verify if SoL+HP changes the outcome of G2IT, hence benefiting reconcilability (as I may need to call upon fewer arguments from Frege).

Agreement How much do the ideas that I call upon from Hilbert contradict with those from Frege? This would undermine the reconciliation.

Credibility To what extent do issues or resolutions thereof regarding *Agreement* undermine the utility and/or credibility or the described theory? E.G. If neither view is able to settle some issue, invoking an entirely external view would undermine the attempt of reconciliation.

The Process

The Justification

Before describing the process, we must be particular about our choice of their views. To narrow possibilities, let us first examine how they fail:

- Hilbert described a system where arithmetic could prove more interesting facts about mathematics. However, he failed at proving the consistency of the base theory, **PA**.
- Frege described a potential model for **PA**, which entails consistency. However, he was not able to explicitly describe it.

Whilst their views may not have a chronology to them, the processes are disjoint, hence the natural reconciliation is to solely use views that justify the above processes, and then solve their shortcomings.

G2IT

Note that, even when accepting the consistency of **PA**, we know that $\mathbf{PA} \vdash \mathsf{Con}_{\mathbf{T}} \Rightarrow \mathbf{PA} \vdash \mathsf{Con}_{\mathbf{PA}}$ for some $\mathbf{T} \supseteq \mathbf{PA}$, which is contradictory, leading to question why this may be wanted. The reason is that we can use the true theory $\{\mathsf{Con}_{\mathbf{PA}}\} \cup \mathbf{PA}$ and the theory $\{\mathsf{Con}_{\{\mathsf{Con}_{\mathbf{PA}}\} \cup \mathbf{PA}}, \mathsf{Con}_{\mathbf{PA}}\} \cup \mathbf{PA}$ and so on. This increases the range of provable theories.

The Description

To specify the SoL we will be using, we will state HP:

$$\#F = \#G \leftrightarrow F \approx G,$$

where,

$$F \approx G := \exists R \forall x((Fx \to \exists! y(Gy \wedge Rxy)) \wedge (Gx \to \exists! y((Fy \wedge Ryx)))$$

This is done using Zalta's definition (Zalta, 2024), where SoL is described as a deductive enrichment of First-order Logic (FoL). However, this description has SoL+HP use countably many objects, making it arithmetisable. Significantly, as it is (decidably) axiomatisable, proofs in that language are arithmetisable. Hence Gödel's First Incompleteness Theorem (G1IT) applies. This necessitates some justification that **PA** is consistent. However, the benefit is that this precision makes it easier to accept SoL+HP as true. As a result, we have fully realised *Necessity.*

We know that Frege believes that "True ⇒ Consistent", hence, using Frege's reasoning that SoL+HP is true (as a part of our way of thinking) (Frege, 1884) we can infer it is consistent. Here we have our first switch from Frege's views to Hilbert's. Frege's concern about SoL+HP was that it did not assign an explicit definition of a number

(Macbride, 2003). However, if we switch to Hilbert's term formalist view, we don't require one.

Term Formalism is the idea that all semantic notions can be reduced down to purely syntactic ones (Linnebo, 1981). HP is a syntactic notion used to describe the properties of numbers; we can argue that Frege's definition of the numerals *is* such a reduction. Hence, we have some theory that embeds **PA**, and therefore Robinson's Arithmetic (**Q**), meaning we can start talking about consistency of more advanced systems (through representability and arithmetisation); we can invoke Hilbert's view that 'Consistency $\Rightarrow$ Truth'.

Term Formalism vs Frege

We may find an issue first when considering Frege was largely against Term Formalism (Linnebo, 1981), however we are not accepting *all* of Frege's perspective. We need only compare the arguments for and against Term formalism with those already used (to invoke HP).

Invoking HP

Frege (Frege, 1884) confirms HP stands by means of its utility; it describes the properties of number that we want it to. This is Frege *confirming* the definition. However, under Frege's distinction between a definition and an axiom, we need to invoke HP by means of proving it as true. Frege does this by claiming it describes the properties of numbers as objects. He believes:

1. HP describes the properties of the objects which we call 'numbers'.
2. HP must therefore be true.
3. We can use HP to define the objects as 'numbers'.

Frege eventually failed on the last step in search of an explicit definition, raising the question if we even *can* define numbers as specific objects. He also emphasises that HP has a unique position over the general idea of equinumerosity (for example, parallel lines) because it allows for there to be an equality between two (albeit undefined) objects which we regard as numbers, consistent with the SoL notion of equality.

Term Formalism

As term formalists, numbers as objects is not considered a necessity, hence we do not assert that they are. This lack of assertion conflicts with step 1 when invoking HP via Frege's perspective. Furthermore, although we have 2 different objects on either side, the sentence $2 \times 2 = 4$ is regarded as true under Term Formalism (Linnebo, 1981); but this contradicts Frege's motivation for equinumerosity applying to HP.

Summary of issues; Is there a way out?

Unpacking the arguments made, we are left with two areas of contention:

- HP is motivated as it maintains consistency of $=$ in the numerical sense; $2 \times 2 = 4$ but Term formalism does not maintain this consistency.
- HP, by design, is motivated by defining numbers as objects whereas a term formalist, although not denying it, is by definition, ignorant of it.

The question remains if we can either reshuffle the views around, or if we will have to invoke some external view to get these perspectives to be compatible. Certainly, we can assume the existential of a number, and accept any representation such as $\#H_0$ or 2×2 as a token for each number; or more formally instead of $2 \times 2 = 4$ we have that 2×2 and 4 fall under the same concept. That is, we claim that there is a number, but any concrete statement '$\mathfrak{a}$ is a number' is false for any object in place of $\mathfrak{a}$. However, there is no immediate obvious reason to motivate this beyond an attempt to avoid contradiction between the two viewpoints, thereby weakening our claim for this system. Furthermore, we also know that Frege was against this notion of universal quantification (Linnebo, 1981). In the end we lose out more with respect to *Credibility.*

Truth Vs Consistency

Another issue that may arise is that we invoke "Truth $\Rightarrow$ Consistency" when we want to verify the consistency of SoL+HP but "Consistency $\Rightarrow$ Truth" when we want to employ Hilbert's program. The obvious marriage between them is "Consistency $\iff$ Truth". When examining much of the correspondence between Hilbert and Frege we notice that both not only insist that their implication is right, but the other is wrong; moreso in Frege's case than Hilbert's (Dummett, 1996). The reason, of course, is that "Truth $\Rightarrow$ Consistency" is deeply embedded in logic; for example proof by contradiction follows this idea. The issue therefore lies in whether Frege's assertion that "Consistency $\not\Rightarrow$ Truth" is a requirement for any other ideas we accept from Frege.

Truth

To develop a usable idea of truth:

1. Frege (anachronistically) requires a defined model for a theory to be true (Dummett, 1996).
2. Hilbert believed consistency implies a model's existence (and hence truth).
3. Therefore, proving a model's existence is a middle-ground requirement for truth.

A Contradiction

Under our reconcilation of SoL+HP with Term Formalism we assume the existence of objects invokes the existence of the model i.e it is true; but as the model is fixed, we encounter an issue when assuming "Truth $\iff$ Consistency". Consider the two consistent theories $\mathbf{PA} \cup \{\neg G_{\mathbf{PA}}\}$ and $\mathbf{PA} \cup \{G_{\mathbf{PA}}\}$. Both are consistent, hence we call them true, but crucially they are part of the same model. However, they are together inconsistent, which shouldn't be true. This wouldn't be an issue if not for our assumption that numbers are (albeit unidentifiable) objects. It is by this assumption that we cannot have contradicting truths about the way they behave. One way we can fix this is by our interpretation of axioms.

Which Axioms?

To see how the two perspectives may come come into play, we want to define axioms. Naturally, our first set of choices are from Hilbert or Frege, as those are banks of opinions we are calling upon. Frege's axioms require that they stem from intuition; the axioms of SoL are taken to be motivated by intuition, therefore we are in a position to do so. This also applies to HP, as although defined in response to a need (that is to define the numbers), it is still assumed to be true by an intuitive argument provided by Frege.

However, we cannot simultaneously accept Hilbert's view on axioms. The reason is that the views are completely opposing. This is because Hilbert states that axioms are *inherently* arbitrary, yet Frege states that they have meaning; in this scenario, that the axioms are derived from intuition and are therefore assumed to be true. To resolve the contradiction, we can limit ourselves in the current fashion:

- Truth $\Rightarrow$ Consistency
- Consistent & Intuitive $\Rightarrow$ True

This employs Frege's perspective on axioms but is an issue with Hilbert's Program as it requires that we not only restrict ourselves to theories that are provably consistent by HP+SoL but those that are also intuitive (i.e. can be accepted as axioms in a Fregian sense). However, this may not be too problematic as we can only hope to capture *some* part of mathematics using Hilbert's Program, and most of mainstream mathematics can be embedded by ZFC axioms, which one can attempt to argue are justified by intuition in the Fregian sense (and therefore any consistent subset thereof must be true).

We could also undermine the issue with the contradiction by noting it isn't necessarily the case that $\text{SoL} \cup \{\text{HP}\} \vdash \mathsf{Con}_{\mathbf{PA} \cup G_{\mathbf{PA}}}$ and $\text{SoL} \cup \{\text{HP}\} \vdash \mathsf{Con}_{\mathbf{PA} \cup \neg G_{\mathbf{PA}}}$. So a more appealing idea would be as follows:

- Truth $\Rightarrow$ Consistency
- Provably consistent from Truths $\Rightarrow$ True

We still run the risk of proving similar contradictions but this doesn't explicitly invoke one. However, this may be a point of contention to be accepted (especially given this is neither a Hilbertian (to some extent) nor Fregean point of view). This undermines *Credibility*.

Conclusion

In summary, we have developed a theory of SoL+HP that treats numbers as unidentifiable objects. This has allowed us to motivate a term formalist approach in the manipulation of numbers. However, in treating numbers as objects, we have undermined the ability of SoL+HP to reach more advanced mathematical ideas. To do so, we need to add caveats to the Hilbertian view on consistency and truth. We have, in most regards, seemed to have mostly satisfied *Necessity* and also *Agreement*, but at the cost from straying from both Frege and Hilbert. However, with regards to *Credibility*, the idea has mostly been disappointing. As a result, it is evident that Frege and Hilbert's views cannot be reconciled to a high degree.

References

Dummett, Michael (1996). *Frege and Other Philosophers*. Oxford Scholarship Online. ISBN: 9780198236283. DOI: `10.1093/019823628X.001.0001`. URL: `https://doi.org/10.1093/019823628X.001.0001`.

Frege, Gottlob (1884). "Grundlagen der Arithmetik. (German) [The Foundations of Arithmetic]". In: pp. 130–159.

Frege, Gottlob and Gottfried Gabriel (1980). "Philosophical and mathematical correspondence". In: pp. 31–52.

Linnebo, Øystein (1981). *Philosophy of Mathematics*. Princeton University Press. Chap. 3–4. DOI: `"https://doi.org/10.2307/j.ctt216687n.7"`.

Macbride, Fraser (2003). "Speaking with Shadows: A Study of Neo-Logicism". In: *Oxford University Press* 54.1, pp. 103–163.

Zalta, Edward N. (2024). "Frege's Theorem and Foundations for Arithmetic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2024. Metaphysics Research Lab, Stanford University.