MATHEMATICAL LOGIC FOR COMPUTER SCIENCE
(A.A. 20/21)

HANDOUT N. 15

ABSTRACT. Minimal Arithmetic. Representability. Incompleteness and undecidability of formal arithmetic through inseparable sets. Gödel's First Incompleteness Theorem.

1. Algorithmic consequences of the Completeness Theorem

The Completeness Theorem, i.e., the equivalence

$$T \vdash E \text{ if and only if } T \models E,$$

has some immediate algorithmic consequences.

We observed that the set of sentences provable from the axioms of the Predicate Calculus is computably enumerable. Since by Completeness $\{E :\models E\} = \{E :\vdash E\}$ we infer that the set of logically valid sentences is also computably enumerable..

The fact that $\{E :\vdash E\}$ is computably enumerable follows from the notion of formal proof and the fact that the axioms of the Predicate Calculus are a computable set i.e., there is an algorithm that decides whether a string is an (instance of an) axiom (scheme) of the Predicate Calculus. Therefore, by the same reasoning, for any decidable theory T, the set $\{E : T \vdash E\}$ is computably enumerable. We call this set $Theor(T)$, the set of "theorems of T". Note that by the Completeness Theorem this set coincides with the set of logical consequences of T, i.e., $\{E : T \models E\}$.

Let us call a theory T *formal* if $Theor(T)$ is c.e. By the above observation, if T is decidable then T is formal. It is an interesting exercise to prove that T is formal also if T is only semi-decidable, i.e., computably enumerable.

We are interested in conditions ensuring that $Theor(T)$ is also decidable, rather than just semi-decidable. If T is a formal theory and T is also (syntactically) complete, then it is easy to see that there is an algorithm to decide membership in $Theor(T)$: just enumerate all theorems of T and see whether E or $\neg E$ appears (first). Note that this works also if T is not consistent: then all sentences are theorems of T. Inconsistent theories are pretty much useless and therefore the above observation is only interesting for theories T with the following three properties:

(1) T is formal, i.e. $Theor(T)$ is c.e.,
(2) T is consistent,
(3) T is complete.

Proposition 1.1. *If T is a formal, consistent and complete theory then $Thm(T)$ is decidable.*

Recall, on the other hand, that we already proved that $Th(\mathbf{N}) = \{E : \mathbf{N} \models E\}$ (the theory of $\mathbf{N}$, or "True Arithmetic") is algorithmically undecidable. By putting together the above observations we reach immediately a rough form of the famous Gödel's First Incompleteness Theorem.

Theorem 1.2. *There is no theory T such that T is formal, consistent and complete and such that $Theor(T) = Th(\mathbf{N})$.*

We will investigate below a more refined and informative version of this important Theorem. In particular, we will isolate a "minimal theory" of arithmetic for which the incompleteness phenomenon holds.

2. Minimal Arithmetic

We fix a basic theory that we call Minimal Arithmetic and we denote with **MA**. The idea of **MA** is that it contains axioms that are so elementary that they are indispensable in any reasonable axiomatization of our informal ideas about the natural numbers.

MA axioms are the following axioms in the language $\{0, 1, +, \times\}$. We denote with **MA** also the conjunction of the universal closure of the axioms of **MA**. For every $n \in \mathbf{N}$ we denote ambiguously with n the term obtained by summing the constant 1 to itself n times. With $x \neq y$ we abbreviate $\neg(x = y)$, with $x \leq y$ we abbreviate $(\exists z)(x + z = y)$, and with $x < y$ abbreviate $x \leq y \wedge x \neq y$.

(Ax 1) $0 + 1 = 1$
(Ax 2) $\forall x(x + 1 \neq 0)$
(Ax 3) $\forall x(x \neq 0 \to \exists z(z + 1 = x))$
(Ax 4) $\forall x \forall y(x + 1 = y + 1 \to x = y)$
(Ax 5) $\forall x(x + 0 = x)$
(Ax 6) $\forall x \forall y(x + (y + 1) = (x + y) + 1$
(Ax 7) $\forall x(x \times 0 = 0)$
(Ax 8) $\forall x \forall y(x \times (y + 1) = (x \times y) + x)$
(Ax 9) $\forall x \forall y(x < y \vee x = y \vee y < x)$

We we need the following propositions, that collect some basic facts about what is provable in **MA**.

Proposition 2.1. *For every $n \in \mathbf{N}$, $\mathbf{MA} \vdash x < n + 1 \to x \leq n$.*

Proof. Recall that $x < n + 1$ abbreviates $\exists z(x + z = n + 1 \wedge x \neq (n + 1)$. We reason in **MA** by cases. If $z = 0$ then by (Ax 5) $x = n + 1$, which contradicts $x \neq n + 1$. Therefore $z \neq 0$. From (Ax 3) it follows that $\exists w(x + (w + 1) = n + 1)$. Therefore, by Ax (6) and (Ax 4), $\exists w(x + w = n)$, which by convention is abbreviated as $x \leq n$. □

Proposition 2.2. *For every $m, n, p \in \mathbf{N}$*

(1) $\mathbf{MA} \vdash x \leq n \to x = 0 \vee \cdots \vee x = n$.
(2) *If $m + n = p$ then* $\mathbf{MA} \vdash m + n = p$.
(3) *If $m \cdot n = p$ then* $\mathbf{MA} \vdash m \times n = p$.
(4) *If $m \neq n$ then* $\mathbf{MA} \vdash \neg(m = n)$.

Proof. We prove point (1) by induction on n. For the case $n = 0$ we must prove $\mathbf{MA} \vdash x \leq 0 \to x = 0$. $x \leq 0$ abbreviates $\exists z(x + z = 0)$. We reason in **MA**, and suppose $\exists z(x + z = 0)$. Reason by cases. If $z \neq 0$, then by (Ax 3) $\exists w(x + (w + 1) = 0)$ and from this, by (Ax 6) and by the assumption $\exists z(x + z = 0)$, it follows that $\exists w((x + w) + 1 = 0)$. But this contradicts (Ax 2). Therefore $z = 0$, and then $x + 0 = 0$. By (Ax 5) it follows $x = 0$. For the case $n + 1$, the induction hypothesis is that $\mathbf{MA} \vdash x \leq n \to x = 0 \vee \cdots \vee x = n$. We reason in **MA**, and suppose $x \leq n + 1$. Then $x < n + 1 \vee x = n + 1$. Then also $x \leq n \vee x = n + 1$, by the previous Proposition. By inductive hypothesis, from $x \leq n$ it follows $x = 0 \vee \cdots \vee x = n$ therefore from $x \leq n \vee x = n + 1$ it follows that $x = 0 \vee \cdots \vee x = n \vee x = n + 1$.

We prove point (2) by induction on n. For the case $n = 0$ we must prove that $\mathbf{MA} \vdash m + 0 = m$. This follows from Axiom (5). For the case $n + 1$, assume $m + (n + 1) = q$. Then $m + n = p$, where p is $q - 1$. By inductive hypothesis, $\mathbf{MA} \vdash m + n = p$. Then $\mathbf{MA} \vdash (m + n) + 1 = p + 1$. By (Ax 6) This implies $\mathbf{MA} \vdash m + (n + 1) = q$.

We prove point (3) by induction on n. For the case $n = 0$ we must prove that $\mathbf{MA} \vdash m \times 0 = 0$. This follows from (Ax 7). For the case $n + 1$, suppose $m \cdot (n + 1) = q$. Then $m \cdot n = p$, where $q = p + m$. By inductive hypothesis $\mathbf{MA} \vdash m \times n = p$. Then $\mathbf{MA} \vdash (m \times n) + m = p + m$. By (Ax 8) we have $\mathbf{MA} \vdash m \times (n+1) = p + m$. From point (2) it follows $\mathbf{MA} \vdash p + m = q$ (N.B. the closed term $p + m$ does not coincide syntactically with the closed term q). □

3. Representability

Although **MA** might look as a weak theory, we will show that it has a surprising expressive power. In particular we show that **MA** (and any theory extending **MA**) is strong enough to reflect/represent the input/output behavior of all the computable functions.

We need the following concept, that is the analogue, for the formal theory **MA**, of the concept of definability in $\mathbf{N}$.

Definition 3.1 (Representability of a function in a theory)**.** A function $\varphi : \mathbf{N}^k \to \mathbf{N}$ is representable in a theory T if there is a formula $F_\varphi(x_1, \ldots, x_k, z)$ in the language of T such that

$$T \vdash (\forall x_1) \ldots (\forall x_k)(\forall y)(\forall z)[F_\varphi(x_1, \ldots, x_k, y) \wedge F_\varphi(x_1, \ldots, x_k, z) \to y = z],$$

and for every $n_1, \ldots, n_k, m$ in $\mathbf{N}$

$$\text{If } \varphi(n_1, \ldots, n_k) = m \text{ then } T \vdash F_\varphi(\bar{n_1}, \ldots, \bar{n_k}, \bar{m}).$$

(if the last implication reverses we talk about strong representability).

The most important fact about representability in **MA** is the Theorem below, due to Gödel. The Theorem shows that all computable functions are representable in **MA**.

The proof essentially consists in verifying that what the proof of the Characterization Theorem (about definability of computable functions in the model $\mathcal{N}$) can be carried through formally in **MA**, replacing "$\mathcal{N} \models$" with "**MA** $\vdash$". The reverse implication is also true but we won't make use of it in our presentation.

Theorem 3.2 (Representability Theorem)**.** *The computable functions are representable in* **MA***.*

Proof. We prove the theorem in three steps:

- Basic functions are representable.
- Representable functions are closed under composition.
- Representable functions are closed under minimization.

Claim 3.3 (Representability of Basic Functions)**.** *Basic functions are representable in* **MA***.*

Proof. Addition is represented by the formula $x + y = z$, multiplication by the formula $x \times y = z$, the projections by the formulas $x_i = z$, equality by the formula $(x = y \wedge z = 1) \vee (x \neq y \wedge z = 0)$. □

Claim 3.4 (Closure under composition)**.** *Representable functions in* **MA** *are closed under composition.*

Proof. Let $G_i(\vec{x}, y_i)$ be the formula representing function θ_i, $i \in [1, m]$, let $H(y_1, \ldots, y_m, z)$ be the formula representing function ψ. Then

$$\exists y_1 \ldots \exists y_m (G_1(\vec{x}, y_1) \wedge \cdots \wedge G_m(\vec{x}, y_m) \wedge H(y_1, \ldots, y_m, z)$$

represents $\varphi(\vec{x}) = \psi(\theta_1(\vec{x}), \ldots, \theta_m(\vec{x}))$.

If $\varphi(\vec{k}) = p$ then there are $q_1, \ldots, q_m \in \mathbf{N}$ such that $\theta_i(\vec{k}) = q_i$ and $\psi(q_1, \ldots, q_m) = p$. Therefore **MA** $\vdash G_i(\vec{k}, q_i)$ and **MA** $\vdash H(q_1, \ldots, q_m, p)$ by hypotheses on G_i and H. □

Claim 3.5 (Closure under minimization)**.** *Representable functions in* **MA** *are closed under minimization.*

Proof. Let $G(\vec{x}, z, y)$ be a formula representing function ψ. Then we claim that

$$F(\vec{x}, z) \equiv \forall w (w \leq z \to \exists y (G(\vec{x}, w, y) \wedge (y = 0 \leftrightarrow w = z)$$

represents $\varphi(\vec{x}) = \min z(\psi(\vec{x}, z) = 0)$.

We first prove the uniqueness of $F(\vec{x}, z)$. We observe that **MA** $\vdash F(\vec{x}, z) \to G(\vec{x}, z, 0)$ and that by definition of F follows

$$\textbf{MA} \vdash (w < z \wedge F(\vec{x}, z)) \to \neg G(\vec{x}, w, 0).$$

therefore

$$\mathsf{MA} \vdash (w < z \wedge F(\vec{x}, z)) \rightarrow \neg F(\vec{x}, w).$$

from the Axiom of totality (Ax 9) of the order $<$ follows

$$\mathsf{MA} \vdash F(\vec{x}, z) \wedge F(\vec{x}, w) \rightarrow z = w.$$

In fact, if $w < z$, given that $F(\vec{x}, z)$, it follows $\neg G(\vec{x}, w, 0)$. On the other hand, from $F(\vec{x}, w)$ follows $G(\vec{x}, w, 0)$. Therefore $\neg(w < z)$. If instead a $z < w$, given that $F(\vec{x}, w)$, follows $\neg G(\vec{x}, z, 0)$, but given that $F(\vec{x}, z)$, follows $G(\vec{x}, z, 0)$. Therefore $\neg(z < w)$. By Ax (9) the only remaining possibility is $z = w$.

We now prove that $F(\vec{x}, z)$ correctly reflects in **MA** the input-output behavior of $\varphi(\vec{x})$. Let $k_1, \ldots, k_n, p$ be in $\mathbf{N}$ such that $\varphi(k_1, \ldots, k_n) = p$. Then, by the definition of φ, we have:

- $\psi(\vec{k}, \bar{p}) = 0$ and
- for each $q < p$ exist $m \neq 0$ such that $\psi(\vec{k}, q) = m$.

By hypotheses on ψ, we have

$$\mathsf{MA} \vdash G(\vec{k}, \bar{p}, 0),$$

and, for every $q < p$ there exists a $m \neq 0$ such that

$$\mathsf{MA} \vdash G(\vec{k}, \bar{q}, \bar{m}).$$

Also, for such an m we have

$$\mathsf{MA} \vdash \bar{m} \neq 0,$$

because $m \neq 0$.

By functionality of G (provable in **MA**), for every $q < p$,

$$\mathsf{MA} \vdash \neg G(\vec{k}, \bar{q}, 0).$$

We want to prove that $F(\vec{k}, \bar{p})$. We have already seen that $\mathsf{MA} \vdash G(\vec{k}, \bar{p}, 0)$ and therefore for $w = p$ the implication in F is true. Consider now the values $w < p$. Reason by cases.

If $p = 0$ then $w < p$ contradicts the axioms. Therefore

$$\forall w(w < 0 \rightarrow \neg G(\vec{k}, w, 0)).$$

If $p = s + 1$ then (by Proposition 3.2 part (1) we have

$$\mathsf{MA} \vdash w < \bar{p} \rightarrow w = 0 \vee \cdots \vee w = \bar{s}.$$

therefore also

$$\mathsf{MA} \vdash \forall w(w < \bar{p} \rightarrow \neg G(\vec{k}, \bar{p}, 0)).$$

We can then conclude that

$$\mathsf{MA} \vdash F(\vec{k}, \bar{p}).$$

□

The proof of the Theorem is thus completed. □

4. Undecidability and Incompleteness

Let us fix a programming system (model of computation) P. Let P_u be the universal function for our programming system P, i.e., the computable function such that

$$P_u(i, j) = P_i(j).$$

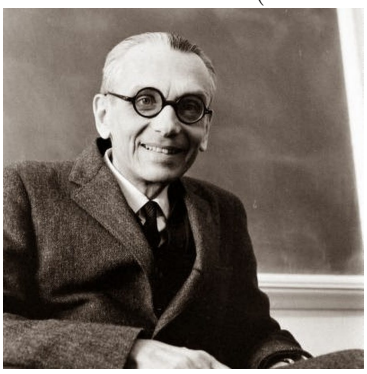
Let $U(x, y, z)$ be the formula that represents P_u in **MA**. The intuitive interpretation of $U(x, x, 0)$ is that $P_x(x) = 0$. Let T be a theory extending **MA**.

The basic idea is that if $P_i(i) = 0$ then **MA** is strong enough to prove $U(i, i, 0)$, and if $P_i(i) = 1$, then **MA** is strong enough to prove $\neg U(i, i, 0)$, i.e., to refute $U(i, i, 0)$. Thus the formal theory **MA** is enough for distinguishing between the two inseparable sets: $S_0 = \{i : P_i(i) = 0\}$ and $S_1 = \{i : P_i(i) = 1\}$.

Let T be any theory extending **MA**.

We prove the following implications, which are used to define a reduction of the theorems of T and of the unsatisfiable sentences to the inseparable pair S_0 and S_1.

FIGURE 1. Kurt Gödel (1906-1978)



(1) If $P_i(i) = 0$ then $\mathsf{MA} \wedge U(i,i,0)$ is **provable** in T.
(2) If $P_i(i) = 1$ then $\mathsf{MA} \wedge U(i,i,0)$ is **unsatisfiable**.

The first implication is a direct consequence of the Representation Theorem and of the choice of $U(x,x,0)$. The second implication is proved as follows.

Trivially

$$\mathsf{MA} \wedge U(i,i,0) \vdash U(i,i,0)$$

By the Representation Theorem $\mathsf{MA} \vdash U(i,i,1)$, and a fortiori we have that

$$\mathsf{MA} \wedge U(i,i,0) \vdash U(i,i,1).$$

Therefore

$$\mathsf{MA} \wedge U(i,i,0) \vdash U(i,i,0) \wedge U(i,i,1).$$

By the Representation Theorem and the univocity of U we have that

$$\mathsf{MA} \vdash U(i,i,0) \wedge U(i,i,1) \to 0 = 1.$$

Therefore

$$\mathsf{MA} \wedge U(i,i,0) \vdash 0 = 1.$$

On the other hand

$$\mathsf{MA} \vdash \neg(0 = 1),$$

because $0 \neq 1$.

Therefore

$$\mathsf{MA} \wedge U(i,i,0) \vdash 0 = 1 \wedge \neg(0 = 1),$$

Therefore $\mathsf{MA} \wedge U(i,i,0)$ is unsatisfiable, by the Completeness Theorem.

Note that in principle we don't know whether $Theor(T)$ and UNSAT are disjoint or not. Asking that $Theor(T) \cap \mathsf{UNSAT} = \emptyset$ is equivalent to asking that T is consistent. In fact, if T is inconsistent then T proves everything, so in particular it proves all sentences in UNSAT. On the other hand, if T is consistent, there can be no $E \in Theor(T) \cap \mathsf{UNSAT}$. Existence of such a T implies that T has no model, since $T \vdash E$ implies $T \models E$, and E has no model. T has no model implies T is inconsistent.

We just proved the following Main Theorem.

Theorem 4.1 (Main Theorem)**.** *Let T be a consistent theory extending MA. Then the theorems of T and the unsatisfiable sentences are computably inseparable.*

Proof. The map $i \mapsto \mathsf{MA} \wedge U(i,i,0)$ is an effective map that reduces the inseparable pair (S_0, S_1) to the pair $(Theor(T), \mathsf{UNSAT})$. Therefore the latter pair is also computably inseparable. □

We immediately obtain the following corollary. Note that it holds for theories that are not necessarily formal. In particular it holds for $Th(\mathcal{N})$ (i.e., True Arithmetic).

Corollary 4.2 (Undecidability of Consistent Arithmetic Theories)**.** *If $T \supseteq \mathsf{MA}$ is consistent, then the set of theorems of T is undecidable.*

Proof. If T is consistent, then no UNSAT sentence is provable in T, i.e.

$$\overline{Theor(T)} \subseteq \mathsf{UNSAT},$$

where $\overline{Theor(T)}$ denotes the complement of $Thm(T)$, i.e., $\{E : T \nvdash E\}$. From the Main Theorem we have that $(Theor(T), \mathsf{UNSAT})$ is inseparable and therefore the pair $(Theor(T); \overline{Theor(T)})$ is also an inseparable pair. Therefore $Theor(T)$ is undecidable □

The above corollary says that there is no way to amend for the undecidability of the theorems of **MA** by adding axioms, as long as we don't fall into an inconsistent (useless) theory. This is often abbreviated saying that **MA** is essentially undecidable.

We immediately obtain the following Theorem as a Corollary.

Theorem 4.3 (First Gödel's Incompleteness Theorem)**.** *If $T \supseteq \mathsf{MA}$ is consistent T is formal (i.e. $Theor(T)$ is c.e.), then T is incomplete, i.e. there exists a sentence G such that:*

(1) *$T \nvdash G$ (T does not prove G), and*
(2) *$T \nvdash \neg G$ (T does not disprove, or refute, G).*

Proof. If T were complete then there would exists a decision procedure for the set of Theorems of T (enumerate all theorems of T and look whether E or $\neg E$ appears in the list), contrary to the previous Corollary. □

This result is a milestone of modern Mathematical Logic. It shows that there is no way of writing a reasonable theory that captures all truths about the structure $\mathcal{N}$. Here "reasonable" means: containing **MA**, being formal and consistent. All such theories are doomed to be incomplete.

Note that the sentence G or its negation $\neg G$ is true in the standard model $\mathbf{N}$, therefore Gödel's First Theorem also shows that, for any consistent formal extentsion T of **MA**, there is *a true sentence* (i.e. a sentence true in the usual structure of $\mathbf{N}$) that is neither provable nor refutable from T (a sentence of this kind is sometimes called *independent* from T).

If we apply this result to **MA**, it gives us a sentence that is neither provable nor refutable from **MA** but is nevertheless true in the structure $\mathbf{N}$. This might not be surprising, considering that **MA** is axiomatized by rather weak axioms, containing a "minimum" amount of what we call arithmetic.

We can then think that we might extend **MA** by stronger axioms, allowing us to prove the truths that escape **MA**.

For example, the following theory has been widely considered a good candidate axiomatization of arithmetic, in that it contains all the usual assumptions we make about $\mathbf{N}$. Let Peano Arithmetic, **PA** be the theory containgin **MA** plus all the following Induction Axioms, one for each formula F in the language:

$$F(0) \wedge \forall x(F(x) \to F(x+1)) \to \forall x F(x).$$

This theory is much stronger than **MA** and one could think that it can avoid the incompleteness problem. But the theorem we proved above can be applied to **PA**: if **PA** is consistent, then it is incomplete. There will be another sentence that is true in $\mathbf{N}$ but is independent of **PA**. There is no way to escape the incompleteness phenomenon by extending **MA** with new axioms, as long as we preserve consistency and formality. This is often abbreviated by saysing that **MA** is essentially incomplete.

We can in fact prove more. By recalling the following proposition we proved about S_0, S_1 being *effectively computably inseparable*, we obtain an even stronger form of Gödels' First Incompleteness Theorem.

Proposition 4.4 (Effective Computable Inseparability)**.** *There exists an algorithm α such that, for all $a, b \in \mathbf{N}$, if the following conditions hold then $\alpha(a, b)$ is defined and $a \notin deom(P_a) \cup dom(P_b)$.*

(1) $dom(P_a) \cap dom(P_b) = \varnothing$.
(2) $S_0 \subseteq dom(P_a)$.
(3) $S_1 \subseteq dom(P_b)$.

Theorem 4.5 (Uniform Gödel's First Incompleteness Theorem)**.** *There is an algorithm A such that, given any formal theory T, if T if consistent and complete, then A computes a sentence E such that $T \nvdash E$ and $T \nvdash \neg E$.*

Proof. Let $U(x, y, z)$ the formula used in the proof of the Main Theorem. Note that this formula is picked as a formula that represents in T the universal computable function. Therefore it depends on T, and might not be unique. To claim uniformity we should acknowledge that there is a uniform effective way of obtaining the formula U representing the universal function in the theory T. That this is the case can be checked by verifying that the proof of the Representation Theorem gives an inductive algorithm to compute such a formula, based on a program to decide the axioms of T.

Then we have

$$P_i(i) = 0 \Rightarrow T \vdash U(i, i, 0)$$

and

$$P_i(i) = 1 \Rightarrow T \vdash \neg U(i, i, 0).$$

Therefore we have

$$S_0 = \{i \ : \ P_i(i) = 0\} \subseteq A_0 = \{i \mid T \vdash U(i, i, 0)\}$$

and

$$S_1 = \{i \ : \ P_i(i) = 1\} \subseteq A_1 = \{i \mid T \vdash \neg U(i, i, 0)\}.$$

Moreover the sets A_0 and A_1 are c.e., by hypothesis on T being formal. They are also disjoint by our hypothesis on T being consistent.

Thus, by our knowledge of S_0, S_1, there exists an $e \notin A_0 \cup A_1$. Therefore there exists an e such that

$$T \nvdash U(e, e, 0), \quad T \nvdash \neg U(e, e, 0).$$

□

In its uniform version, Gödels' First Incompleteness Theorem says that there is an algorithm that, given any formal consistent theory T of arithmetic extending **MA**, computes a witness of the incompleteness of T.

In the above proof, the witness also has a specific form: $U(e, e, 0)$, i.e. a sentence stating that program number e on input number e halts and outputs 0.