2 Gödel's Incompleteness Theorems

2.1 Hilbert's Programme

In the 1920s David Hilbert (1862–1943) formulated a programme for the further development of mathematics. He proposed to axiomatise various branches of mathematics in first-order logic and to reduce mathematical reasoning to formal derivations that can be processed automatically. This should be possible by constructing effective procedures (algorithms) to establish the truth of mathematical statements in a mathematical theory. A more global idea was to prove the consistency of mathematics.

This programme was realised only partially. For important areas of mathematics, axiom systems were developed, in particular for Peano arithmetic (PA) and for ZFC, which led to a formalisation of mathematics in set theory. Furthermore, precise notions of proofs were defined. Appropriate formal proof systems are Hilbert-Frege systems, the method of resolution, and sequent calculi.

In 1931, Gödel proved the completeness theorem for first-order logic. It says that a formula ψ follows from a set of formulae Φ if and only if it φ can be derived from Φ in sequent calculus, i.e. $\Phi \models \psi \Leftrightarrow \Phi \vdash \psi$. It follows that the set of all valid first-order sentences $\mathrm{val}(\mathrm{FO}) = \{\psi \in \mathrm{FO} : \ \vdash \psi\}$ is recursively enumerable.

Finally, algorithms for deciding satisfiability and validity for certain fragments of FO and other logics were found.

However, fundamental results from the 1930s showed that Hilbert's programme was due to fail.

Theorem 2.1 (Gödel's First Incompleteness Theorem)**.** Every sufficiently powerful recursively axiomatisable theory is incomplete.

We shall make the notion of *sufficiently powerful theory* precise later. Examples of such theories are ZFC and PA.

Theorem 2.2 (Church, Turing)**.** Satisfiability and validity of FO are undecidable.

Theorem 2.3 (Gödel's Second Incompleteness Theorem)**.** Let Φ be a decidable sufficiently powerful axiom system. Then the consistency of Φ is not provable in Φ ($\Phi \not\vdash \mathrm{Cons}_\Phi$).

In particular, this holds for $\Phi = \mathrm{ZFC}$, so the consistency of mathematics is provable if and only if it is inconsistent.

2.2 Theories

Definition 2.4. A *theory* $T \subseteq \mathrm{FO}(\tau)$ is a satisfiable set of τ-sentences which is closed under $\models$, i.e. $T \models \psi$ implies $\psi \in T$. We say that T is *complete* if for each sentence $\psi \in \mathrm{FO}(\tau)$ either $\psi \in T$ or $\neg\psi \in T$. We say that T is *recursively axiomatisable* if there is a decidable set $\Phi \subseteq \mathrm{FO}(\tau)$ of axioms such that $\Phi^{\models} := \{\psi \in \mathrm{FO}(\tau) : \Phi \models \psi\} = T$.

Theorem 2.5. Let T be a complete theory over a signature τ. Then the following statements are equivalent.

(1) T is recursively axiomatisable.

(2) There exist a recursively enumerable axiom system Φ such that $T = \Phi^{\models}$.

(3) T is recursively enumerable.

(4) T is decidable.

Proof. (1) $\Rightarrow$ (2) This case is trivial.

(2) $\Rightarrow$ (3) If the set Φ is recursively enumerable, then so is the set of all finite $\Phi_0 \subseteq \Phi$. Hence we can systematically generate all derivable sequents $\Phi_0 \Rightarrow \psi$ (with $\Phi_0 \subseteq \Phi$), hence we can also derive all ψ with $\Phi \vdash \psi$, but $\{\psi : \Phi \vdash \psi\} = T$. Hence T is recursively enumerable.

(3) $\Rightarrow$ (4) T is complete. Hence $\psi \notin T$ if and only if $\neg\psi \in T$. Hence also $\mathrm{FO} \setminus T$ is recursively enumerable. It follows that T is decidable.

(4) $\Rightarrow$ (1) Put $\Phi = T$. Q.E.D.

Consider the structure $\mathfrak{N} = (\mathbb{N}, +, \cdot, 0, 1)$, the arithmetic of natural numbers. The theory $\mathrm{TA} = \mathrm{Th}(\mathfrak{N}) = \{\varphi : \mathfrak{N} \models \varphi\}$ is the *true arithmetic*. As a theory of a structure, it is complete.

Another way to define arithmetic on natural numbers is Peano's axiom system, where well-known properties of arithmetic on natural numbers are given explicitly. However, we shall see later that the two axiom systems are not equivalent.

In monadic second-order logic (MSO) one can write the "axiom of induction" as

$$\forall X\big(X0 \land \forall y(Xy \to X(y+1))\big) \to \forall z Xz,$$

where X is a second-order variable and stands for the set of elements having some property. In FO we explicitly name each definable property, which leads to an infinite (but enumerable) set of axioms.

Let τ_{ar} be the signature of arithmetic: $\tau_{\mathrm{ar}} = \{+, \cdot, 0, 1\}$. The axiom system of Peano arithmetic Φ_{PA} consists of the following axioms:

(1) $\forall x \neg(x + 1 = 0)$,
(2) $\forall x \forall y((x + 1 = y + 1) \to (x = y))$,
(3) $\forall x(x + 0 = x)$,
(4) $\forall x \forall y(x + (y + 1) = (x + y) + 1)$,
(5) $\forall x(x \cdot 0 = 0)$,
(6) $\forall x \forall y(x \cdot (y + 1) = (x \cdot y) + x)$,

and the scheme of induction axioms:

(7) $\forall \bar{y}(\varphi(0, \bar{y}) \land \forall x(\varphi(x, \bar{y}) \to \varphi(x + 1, \bar{y})) \to \forall x \varphi(x, \bar{y}))$
for every formula $\varphi(x, y_1, \ldots, y_n) \in \mathrm{FO}(\tau_{\mathrm{ar}})$.

Remark 2.6. For every formula $\varphi(x, \bar{y})$ and every tuple $\bar{b} \in \mathbb{N}^k$, $\varphi(x, \bar{b})$ defines a set $\varphi^{\mathfrak{N}, \bar{b}} = \{a \in \mathbb{N} : \mathfrak{N} \models \varphi(a, \bar{b})\}$.

The theory $\mathrm{PA} := \Phi_{\mathrm{PA}}^{\models}$ is called *Peano arithmetic*. It gives us induction for all sets definable in that sense. Notice that PA is recursively axiomatisable and therefore recursively enumerable.

We introduce the notion of a representative axiom system, which formalises the above-mentioned notion of a sufficiently powerful axiom system.

Definition 2.7. An axiom system Φ *is representative* (or *permits coding*) if one can construct, for each $n \in \mathbb{N}$, a term t_n such that

(1) $\Phi \models \neg(t_n = t_m)$ for all $m \neq n$, and
(2) for every total computable function $f : \mathbb{N}^k \to \mathbb{N}$ there is a formula $\varphi_f(\bar{x}, y)$ such that for all $n_1, \dots, n_k \in \mathbb{N}$ and for all $m \in \mathbb{N}$
   (1) $\Phi \vdash \exists^{=1} y \varphi_f(t_{n_1}, \dots, t_{n_k}, y)$,
   (2) if $f(n_1, \dots, n_k) = m$ then $\Phi \vdash \varphi_f(t_{n_1}, \dots, t_{n_k}, t_m)$, and
   (3) if $f(n_1, \dots, n_k) \neq m$ then $\Phi \vdash \neg\varphi_f(t_{n_1}, \dots, t_{n_k}, t_m)$.

Remark 2.8. If Φ is representative, then each decidable relation $R \subseteq \mathbb{N}^k$ is represented by a formula $\varphi_R(x_1, \dots, x_k)$ such that

- $(n_1, \dots, n_k) \in R$ implies $\Phi \vdash \varphi_R(t_{n_1}, \dots, t_{n_k})$, and
- $(n_1, \dots, n_k) \notin R$ implies $\Phi \vdash \neg\varphi_R(t_{n_1}, \dots, t_{n_k})$

because we can put $\varphi_R(x_1, \dots, x_k) = \varphi_f(x_1, \dots, x_k, t_1)$ where $f : \mathbb{N}^k \to \mathbb{N}$ is the characteristic function of R.

Our next goal is to show that TA, Φ_{PA} and ZFC are representative. For this purpose we encode tuples of fixed length by numbers.

Definition 2.9. The function $[\cdot,\cdot] : \mathbb{N}^2 \to \mathbb{N}$ is defined as $[x, y] = \frac{1}{2}(x + y)(x + y + 1) + x$.

Lemma 2.10. $[\cdot,\cdot]$ is a bijection.

Proof. Enumerate $\mathbb{N}^2$ as depicted in Figure 2.1. There are $\frac{(x+y)(x+y+1)}{2}$ elements on the diagonals before the diagonal containing the element (x, y). On every diagonal $\{(x, y) : x + y = k\}$ there are $k + 1$ elements. Thus the pair (x, y) gets the number $(\sum_{0 \leq n < x+y} n + 1) + x = (\sum_{1 \leq n \leq x+y} n) + x = \frac{1}{2}(x + y)(x + y + 1) + x = [x, y]$. Q.E.D.

Now define $[a_0, \dots, a_{n-1}] := [a_0, [a_1, \dots, a_{n-1}]]$ for $n > 2$. Thus we have definable bijections $\mathbb{N}^k \to \mathbb{N}$ for every fixed k. To describe arbitrary computable functions, hence computations of arbitrary finite

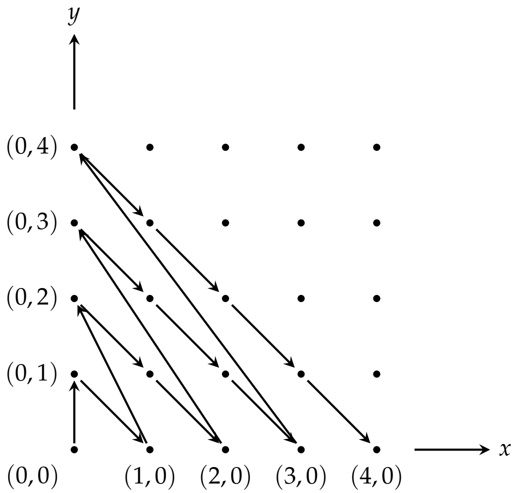

Figure 2.1. Enumeration of $\mathbb{N} \times \mathbb{N}$

length (for example of Turing Machines) by formulae of arithmetic, we need coding sequences of natural numbers of bounded length.

Theorem 2.11 (Chinese Remainder Theorem)**.** Let $q_1, \dots, q_{n-1} \in \mathbb{N}$ be pairwise relatively prime and let $q := \prod_{i<n} q_i$. Then the function $F : \mathbb{Z}/q\mathbb{Z} \to \mathbb{Z}/q_0\mathbb{Z} \times \cdots \times \mathbb{Z}/q_{n-1}\mathbb{Z}$ with $a \mapsto (a_0, \dots, a_{n-1})$ where $a \equiv a_i \pmod{q_i}$ is a bijection.

Proof. Since $\mathbb{Z}/q\mathbb{Z}$ and $\mathbb{Z}/q_0\mathbb{Z} \times \cdots \times \mathbb{Z}/q_{n-1}\mathbb{Z}$ are finite and have the same number of elements, it suffices to show that F is injective. Let $a, a' \in \mathbb{Z}/q\mathbb{Z}$ be such that $a \equiv a' \pmod{a_j}$ for all $j < n$. Then $a - a'$ is divisible by all q_j, hence (since the q_j are relatively prime) also by the product q. It follows that $a \equiv a' \pmod{q}$. Q.E.D.

Lemma 2.12 (β-Lemma by Gödel)**.** There is a total computable function $\beta : \mathbb{N}^3 \to \mathbb{N}$ such that for each finite sequence $(a_0, \dots, a_{n-1})$ on $\mathbb{N}$, there exist $a, b \in \mathbb{N}$ with $\beta(a, b, j) = a_j$ for all $j < n$.

Proof. Put $\beta(x,y,z) := x \pmod{1+y(z+1)}$. Obviously, β is definable (in TA, PA) by

$$\varphi_\beta(x,y,z,v) := v < 1+y(z+1) \land \exists u(x = u + uy(z+1) + v) .$$

It remains to show that for all n and all $a_0, \dots, a_{n-1}$ we can find appropriate a, b such that $a \equiv a_j \pmod{1+b(j+1)}$ for all $j < n$. Let $b := m!$ for $m = \max(n, a_0, \dots, a_{n-1})$.

Claim 2.13. For $0 \leq i < j \leq n$, the numbers $1+(i+1)b$ and $1+(j+1)b$ are relatively prime.

Otherwise there is a prime p with $p \mid 1+(i+1)b$ and $p \mid 1+(j+1)b$, and hence $p \mid (i-j)b$. But $p \nmid b$ (otherwise $p \nmid 1+(i+1)b$), hence $p \mid (i-j)$ and hence $p < n$. But all $p < n$ divide b, a contradiction. This proves the claim.

We can apply Chinese Remainder theorem and conclude that there is an $a < \prod_{j=i}^{n-1}(1+b(j+1))$ such that $a \equiv a_j \pmod{1+b(j+1)}$ for all $j < n$.

A sequence $(a_0, \dots, a_{n-1})$ on $\mathbb{N}$ can be coded by $\langle a_0, \dots, a_{n-1} \rangle := [a, b, n]$ so that $\beta(a,b,j) = a_j$ for all $j < n$ with $b = \max(n, a_0, \dots, a_{n-1})!$

Q.E.D.

Now we can define $\ln(\langle a_0, \dots, a_{n-1} \rangle) := n$, $\pi_i(\langle a_0, \dots, a_{n-1} \rangle) = a_i$. It is clear that $[\cdot,\cdot], \beta, \ln, \pi_i$ are definable in TA,PA and ZFC.

2.2.1 Coding Turing Machines

Let $M = (Q, \Sigma, \delta, q_0, F)$ be a deterministic Turing Machine. A configuration of M is a tuple $c = \langle q, w, p \rangle \in Q \times \Sigma^* \times \mathbb{N} \subseteq \mathbb{N} \times \mathbb{N}^* \times \mathbb{N}$ where q is the state of M in c, w is the tape inscript and p is the head position. A computation is a sequence $\langle c_0, c_1 \dots, c_m \rangle$ of configurations with $c_i \vdash_M c_{i+1}$.

Lemma 2.14. Let $\Phi \in \{\mathrm{TA}, \Phi_{\mathrm{PA}}, \mathrm{ZFC}\}$, and let M be a Turing Machine. Then there exist formulae $\mathrm{Conf}_M(x)$, $\mathrm{Start}_M(x,y)$, $\mathrm{End}_M(x,y)$ and $\mathrm{Run}_M(x)$ such that

- $\Phi \vdash \mathrm{Conf}_M(x)$ if and only if x represents a configuration of M,

- $\Phi \vdash \mathrm{Start}_M(x, y)$ if and only if x encodes the input configuration of M on input y,
- $\Phi \vdash \mathrm{End}_M(x, y)$ if and only if x encodes a final configuration of M with output y,
- $\Phi \vdash \mathrm{Run}_M(x)$ if and only if x encodes a computation of M.

It follows from Lemma 2.14 that TA, Φ_{PA}, ZFC are representative.

Corollary 2.15 (Tarski)**.** TA is undecidable.

Likewise, PA and ZFC$^{\models}$ are undecidable. For TA and PA we have the following theorem. Since TA is complete, it would be decidable if there were a decidable axiom system for TA. PA is recursively axiomatised; if it were complete, it would be decidable.

Theorem 2.16 (Gödel)**.**

(1) There is no decidable axiom system for TA.
(2) PA is incomplete.

There are many possible ways to encode terms and formulae with tuples of natural numbers. We consider the following Gödelisation of terms:

- $[x_i] := \langle 0, i \rangle \in \mathbb{N}$,
- $[0] := \langle 1, 0 \rangle \in \mathbb{N}$,
- $[1] := \langle 1, 1 \rangle \in \mathbb{N}$,
- $[t_0 + t_1] := \langle 2, [t_0], [t_1] \rangle \in \mathbb{N}$,
- $[t_0 \cdot t_1] := \langle 3, [t_0], [t_1] \rangle \in \mathbb{N}$,

and the following Gödelisation of formulae:

- $[t_0 = t_1] := \langle 4, [t_0], [t_1] \rangle \in \mathbb{N}$,
- $[\neg \varphi] := \langle 5, [\varphi] \rangle \in \mathbb{N}$,
- $[\varphi \wedge \psi] := \langle 6, [\varphi], [\psi] \rangle \in \mathbb{N}$,
- $\exists x_i \varphi := \langle 7, i, [\varphi] \rangle$.

For a formula $\vartheta(x)$ and $k \in \omega$ we write $\vartheta(k)$ for $\vartheta[x/k]$ where k is $1 + 1 + \cdots + 1$ (k times), i.e. the formula which we get substituting k times $1 + \cdots + 1$ for x in $\vartheta(x)$.

Theorem 2.17 (Fixed Point Theorem)**.** Let Φ be representative. For each formula $\psi \in \mathrm{FO}(\{+,\cdot,0,1\})$ there is a sentence φ such that $\Phi \vdash \varphi \leftrightarrow \psi([\varphi])$. In other words, the function $g_\psi : \varphi \mapsto \psi([\varphi])$ has a fixed point (up to logical equivalence).

Proof. Let $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ the function with $f([\vartheta(x)], k) := [\vartheta(k)]$ and $f(n,k) = 0$ if n is not the Gödelisation of a formula $\vartheta(x)$. Obviously, f is computable. Hence there is a formula $\alpha(x,y,z)$ such that $\Phi \vdash \alpha(n,k,m)$ if and only if $m = f(n,k)$ (because Φ is representative). For a given formula $\psi(x)$, set $\vartheta(x) := \forall z(\alpha(x,y,z) \to \psi(z))$ and $\varphi := \vartheta([\vartheta])$.

We show that $\Phi \vdash \varphi \leftrightarrow \psi([\varphi])$ holds, which proves the theorem. We have $f([\vartheta],[\vartheta]) = [\varphi]$, and hence

$$\Phi \vdash \alpha([\vartheta][\vartheta],[\varphi])\ . \tag{$*$}$$

- We show that $\Phi \vdash \varphi \to \psi([\varphi])$ holds. We have

  $$\Phi \vdash (\varphi \wedge \alpha([\vartheta],[\vartheta],[\varphi])) \to \psi([\varphi])$$

  because $\Phi \vdash \varphi$ if and only if $\Phi \vdash \forall z(f([\vartheta],[\vartheta]) = z \to \psi(z))$ and $\Phi \vdash \alpha([\vartheta],[\vartheta],[\varphi])$ if and only if $f([\vartheta],[\vartheta]) = [\varphi]$. Hence with $(*)$ we obtain $\Phi \vdash \varphi \to \psi([\varphi])$.
- It remains to show that $\Phi \vdash \psi([\varphi]) \to \varphi$. We have $\Phi \vdash \exists^{=1}z\,\alpha([\vartheta],[\vartheta],z)$ because α represents a function. By $(*)$, we obtain $\Phi \vdash \forall z\ \alpha([\vartheta],[\vartheta],z) \to z = [\varphi]$. Then $\Phi \vdash \psi([\varphi]) \to \Big(\forall z\ \alpha([\vartheta],[\vartheta],z) \to \psi(z)\Big)$ (in other words, $\forall z\ \alpha \ldots$ implies $z = [\varphi]$, so $\psi(\varphi)$ implies $\psi(z)$). By definitions of φ and ϑ, the expression in the larger brackets of the preceding formula is equal to φ, and hence we get $\Phi \vdash \psi([\varphi]) \to \varphi$. Q.E.D.

Using the Fixed Point Theorem and a diagonalisation argument, one can prove that in sufficiently powerful theories the set of all true (or all false) sentences is not definable in the theory.

Theorem 2.18. Let $\mathfrak{A}$ be a structure that extends $\mathfrak{N}$ so that $\mathrm{Th}(\mathfrak{A})$ permits coding. Then there is no first-order formula $\mathrm{True}_{\mathfrak{A}}(x)$ such that

$$\mathfrak{A} \models \psi \;\Leftrightarrow\; \mathfrak{A} \models \mathrm{True}_{\mathfrak{A}}([\psi]).$$

Proof. Suppose that such a formula $\mathrm{True}_{\mathfrak{A}}(x)$ exists. By the Fixed Point Theorem applied to $\neg\mathrm{True}_{\mathfrak{A}}(x)$, there exists a fixed point φ with $\mathrm{Th}(\mathfrak{A}) \vdash \varphi \leftrightarrow \neg\mathrm{True}_{\mathfrak{A}}([\varphi])$ (which can informally be interpreted as "φ claims that φ is false"). Hence $\mathfrak{A} \models \varphi \Leftrightarrow \mathfrak{A} \models \neg\mathrm{True}_{\mathfrak{A}}([\varphi]) \Leftrightarrow \mathfrak{A} \not\models \varphi$, which is a contradiction. (The last equivalence is due to the definition of $\mathrm{True}_{\mathfrak{A}}(x)$.) Q.E.D.

Corollary 2.19 (Tarski). The set of true sentences in $\mathfrak{N}$ is not definable in $\mathfrak{N}$.

We give still another formulation of the preceding result.

Theorem 2.20. Let T be a representative and complete theory. Then T is not definable in T, i.e. there is no formula $\mathrm{True}_T(x)$ with $\psi \in T \Leftrightarrow \mathrm{True}_T([\psi]) \in T$.

We get a different proof of Gödel's First Incompleteness Theorem: If T is recursively axiomatisable and representative, then T is incomplete. (Otherwise T would be decidable. Since T is representative, the decidable set $\{[\psi] : \psi \in T\}$ would be definable in T.)

Let Φ be decidable and representative. We select an appropriate coding of derivations in the sequent calculus and consider the relation $B \subseteq \mathbb{N} \times \mathbb{N}$ such that $(n, m) \in B$ if and only if n encodes a derivation of a sequent $\Phi_0 \Rightarrow \psi$ so that $\Phi_0 \subseteq \Phi$ and $m = [\psi]$. Since B is decidable, there is a formula $\mathrm{Proof}_{\Phi}(x, y)$ so that $\Phi \vdash \mathrm{Proof}_{\Phi}(n, m)$ if and only if $(n, m) \in B$. Put $\mathrm{Provable}_{\Phi}(x) := \exists y\, \mathrm{Proof}_{\Phi}(y, x)$ and $\mathrm{Cons}_{\Phi} := \neg\mathrm{Provable}([0 \neq 0])$. Hereby Cons_{Φ} expresses consistency of Φ, and $[0 \neq 0]$ is the Gödel number of a false sentence.

Theorem 2.21 (Gödel's Second Incompleteness Theorem). If $\Phi \supseteq \Phi_{\mathrm{PA}}$ is decidable and consistent, then $\Phi \not\vdash \mathrm{Cons}_{\Phi}$.

Proof. According to the Fixed Point Theorem, there is a sentence φ with

$$\Phi \vdash \varphi \leftrightarrow \neg\mathrm{Provable}_{\Phi}([\varphi]) \qquad (+)$$

(i.e. φ expresses its own non-provability).

First, we show that this implies $\Phi \nvdash \varphi$. Indeed, assume $\Phi \vdash \varphi$. Then there is a proof that $\Phi_0 \models \varphi$ for some finite subset Φ_0 of Φ. Hence we have $\Phi \vdash \text{Provable}_\Phi([\varphi])$ and $\Phi \vdash \neg\varphi$, which contradicts the assumption that $\Phi \vdash \varphi$ because Φ is consistent.

Thus the consistency of Φ implies the non-provability of φ, as a formula: $\text{Cons}_\Phi \to \neg\text{Provable}_\Phi([\varphi])$. One can formulate this proof in $\Phi \supseteq \Phi_{\text{PA}}$ and show that $\Phi \vdash \text{Cons}_\Phi \to \neg\text{Provable}_\Phi([\varphi])$. If $\Phi \vdash \text{Cons}_\Phi$ then $\Phi \vdash \neg\text{Provable}_\Phi([\varphi])$, and hence $\Phi \vdash \varphi$ by $(+)$. However, we have already shown that this is impossible, and hence $\Phi \nvdash \text{Cons}_\Phi$. Q.E.D.