teorema
Vol. XXII/3 2003, pp.5-15

A Short Guide to Gödel's Second Incompleteness Theorem

Joan Bagaria

RESUMEN

La demostración habitual del Segundo Teorema de Incompletud de Gödel a partir de teorías débiles como $I\Sigma_1$ es larga y técnicamente intrincada. Raramente se dan todos los detalles y en muchos casos se omiten completamente apelando a la capacidad de lector para completarlos. En la primera parte de este artículo presentamos una guía de los principales puntos técnicos de la demostración habitual del Segundo Teorema de Incompletud de Gödel a partir de teorías débiles. En la segunda parte presentamos una demostración distinta y más simple para la Teoría de Conjuntos de Zermelo-Fraenkel debida a T. Jech [Jech (1994)], y observamos que puede ser extendida de forma que englobe teorías débiles, evitando así muchas de las complicaciones técnicas que requieren las demostraciones habituales.

ABSTRACT

The usual proof of Gödel's second incompleteness theorem for weak theories like $I\Sigma_1$ is long and technically cumbersome. The details are rarely given in full and in most cases they are skipped altogether with dismissing vague sentences alluding to the reader's ability to fill them in. In the first part of this note we provide a guide through the main technical points of the usual proof of Gödel's theorem for weak theories. In the second part we present a different and simpler proof of the theorem for Zermelo-Fraenkel Set Theory, due to T. Jech [Jech (1994)], and we observe that it can be stretched to encompass weak theories, while avoiding many of the technicalities that are required in the usual proofs.

I. INTRODUCTION

The importance of Gödel's incompleteness theorems [Gödel (1967)] for both Logic and the Foundations of Mathematics can hardly be overstated. They not only represented a heavy blow on Hilbert's Program it its original form, but they also changed forever the role of Logic in Mathematics, not to mention the endless discussions to our day about their philosophical significance.

The theorems can be informally stated as follows:

THEOREM (*Gödel's first incompleteness theorem*) Let T be an axiomatizable theory that contains (a small fragment of) arithmetic. Then there is a sentence θ such that if T is consistent, then T does not prove θ, and if T satisfies certain additional consistency hypothesis, then T does not prove the negation of θ either.

THEOREM (*Gödel's second incompleteness theorem*) Let T be an axiomatizable theory that contains (a small fragment of) arithmetic. If T is consistent, then T does not prove that T is consistent.

For the usual first-order theories of Arithmetic and Set Theory, the first theorem is an easy corollary of the second. In this note we shall concentrate on the second incompleteness theorem. In the first section we provide a guide through the main technical points in the proof of the theorem for weak theories, aiming at its (almost) optimal form. Next, we present a short proof of the theorem, due to T. Jech [Jech (1994)], for Zermelo-Fraenkel Set Theory (*ZF*). Finally, we show how Jech's proof can be stretched to encompass weaker theories, like Peano's Arithmetic (*PA*) and Σ_1-Induction ($I\Sigma_1$), which is essentially the weakest theory for which the Theorem holds. These proofs avoid many of the technicalities that are required in the usual proofs of the Theorem as outlined in the first section of this note.

II. THE MAIN INGREDIENTS IN THE PROOF OF THE INCOMPLETENESS THEOREMS

The main ingredients in the usual proof of Gödel's second incompleteness theorem for a given theory T (in a language that contains the language of arithmetic) are the following:

- Recursive arithmetization of the syntax of the language of T.
- Σ_1-definability of the recursively enumerable predicates.
- Provability in T of true Σ_1 sentences of the language of arithmetic.
- Diagonalization.
- Provability in T of some of the properties of the *Provability* predicate Bew_T.

The first four ingredients are also present in the first incompleteness theorem. The fifth, which is harder to prove, is the crucial step that yields the stronger second incompleteness theorem.

II.1 *Recursive arithmetization of the syntax of the language of T*

Given a countable formal language $\mathcal{L}$, we can identify the symbols with natural numbers and we can code, in a (primitive) recursive way, the syntax of $\mathcal{L}$. This is possible because the syntactic notions, like formulas and proofs are defined recursively. The way the coding is done is quite irrelevant, as long as it is recursive, so that if the set of symbols is a recursive set of natural numbers, then so are the sets of codes of terms, formulas and proofs. We also need that the ternary relation *Sb* consisting of all $\langle x,y,z\rangle$ such that z is (the code of) the result of substituting the only free variable of the formula (coded by) x by the term (coded by) y, is recursive.

The main point is that if T is a recursively enumerable set of formulas of $\mathcal{L}$ (i.e., the set of codes of formulas of T is recursively enumerable), then the provability predicate Bew_T, consisting of the codes of all theorems of T is also a recursively enumerable set.

II.2 Σ_1-*definability of the Recursively Enumerable Predicates*

The language of arithmetic consists of two binary function symbols + and $\cdot$, one unary function symbol S and one constant symbol 0. The ordering relation $\leq$ is defined as $x \leq y$ iff $\exists z(x + z = y)$.

A Σ_1 formula (in the language of aritmetic) is a formula of the form $\exists x\varphi(x,y_1,\ldots,y_k)$, where $\varphi(x,y_1,\ldots,y_k)$ is a *bounded formula*, i.e., a formula whose quantifiers are all bounded, namely, they are of the form $\exists y \leq z$ or $\forall y \leq z$.

Every recursively enumerable set of natural numbers is Σ_1 definable in the standard model $\langle \mathbb{N}, +, \cdot, S, 0\rangle$. In fact, this is if and only if. In particular, there are Σ_1 formulas $Sb(x,y,z)$ and $Bew_T(x)$ that define the substitution relation *Sb* and the provability predicate Bew_T.

II.3 *Provability in T of true* Σ_1 *sentences of the language of arithmetic.*

We write $\overline{n}$ instead of the term

$$\underbrace{SSSS\ldots 0}_{n-times}$$

The terms $\overline{n}$ are called *numerals*.

The following fragment of arithmetic is called R_0. It is given by four infinite groups of axioms:

1. $\overline{n} + \overline{m} = \overline{p}$, for all $m,n,p \in \mathbb{N}$ such that $n + m = p$.
2. $\overline{n} \cdot \overline{m} = \overline{p}$, for all $m,n,p \in \mathbb{N}$ such that $n \cdot m = p$.
3. $\overline{n} \neq \overline{m}$, for all $m,n \in \mathbb{N}$ such that $n \neq m$.
4. And the universal closure of the formulas that are of the form:

$$x \leq \overline{n} \to (x = \overline{0} \vee x = \overline{1} \vee \ldots \vee x = \overline{n}),$$

for all $n \in \mathbb{N}$.

R_0 has the following important feature:

Every true Σ_1 sentence in the language of arithmetic is provable in R_0.

The reason is that every model M of the first three groups of axioms satisfies the diagram of $\langle \mathbb{N}, +, \cdot, S, 0\rangle$, hence all sentences without quantifiers are absolute between M and $\mathbb{N}$. The fourth group of axioms ensures that every sentence in the language of arithmetic with only bounded quantifiers is R_0-equivalent to a sentence without quantifiers. Hence, all Σ_1 sentences that hold in $\mathbb{N}$ also hold in M. It can be easily checked that R_0 is the weakest fragment of arithmetic that has this property.

The following property of the provability predicate Bew_T will play a crucial role in the proof of the incompleteness theorem: Since every true Σ_1 sentence is provable in T, we have that for all formulas φ,

D0. $T \vdash \varphi$ implies $T \vdash Bew_T(\ulcorner\varphi\urcorner)$

where $\ulcorner\varphi\urcorner$ is the Gödel notation for the numeral corresponding to the code of φ, i.e., if n codes φ, $\ulcorner\varphi\urcorner = \overline{n}$.

II.4 *Diagonalization*

If a is a term or a formula, let $[a]$ denote the code of a. If $n = [a]$ we write, following Gödel, $\ulcorner a\urcorner$ instead of $\overline{n}$.

THEOREM (*Gödel's diagonalization theorem*) Let T be a theory that contains R_0. Then for every formula $\varphi(x)$, where x is the only free variable, there is a sentence θ such that

$$T \vdash (\theta \leftrightarrow \varphi(\ulcorner\theta\urcorner)).$$

The proof of the diagonalization theorem hinges on the following:

(**) *Since $Sb(x,y,z)$ is Σ_1, by the provability in R_0 of the true Σ_1 sentences, if $Sb(m,n,p)$, then $T \vdash Sb(\overline{m},\overline{n},\overline{p})$, for all $m,n,p \in \mathbb{N}$.*

The proof of the diagonalization theorem then goes as follows:
Let

$$n = [\forall z(Sb(x, \ulcorner x\urcorner, z) \rightarrow \varphi(z))].$$

Let θ be the sentence

$$\forall z(Sb(\overline{n}, \ulcorner\overline{n}\urcorner, z) \rightarrow \varphi(z)).$$

Note that $Sb(n,[\overline{n}],[\theta])$ holds.

By (∗),

$$T \vdash Sb(\overline{n}, \ulcorner\overline{n}\urcorner, \ulcorner\theta\urcorner).$$

Clearly,

$$\vdash (\theta \to (Sb(\overline{n}, \ulcorner\overline{n}\urcorner, \ulcorner\theta\urcorner) \to \varphi(\ulcorner\theta\urcorner))).$$

Hence,

$$T \vdash (\theta \to \varphi(\ulcorner\theta\urcorner))$$

On the other hand, since $T \vdash Sb(\overline{n}, \ulcorner\overline{n}\urcorner, \ulcorner\theta\urcorner)$, we have

$$T \vdash (\varphi(\ulcorner\theta\urcorner) \to \forall z(Sb(\overline{n}, \ulcorner\overline{n}\urcorner, z) \to \varphi(z)))$$

But the consequent is precisely θ.

For the proof of both incompleteness theorems, we need the diagonalization theorem for only one particular instance, namely, the formula $\neg Bew_T(\ulcorner x\urcorner)$,

II.5 *Provability in T of some of the properties of the Provability predicate.*

Among the properties of the provability predicate Bew_T, the following two are relevant for the proof of the second incompleteness theorem:

For all formulas φ and ψ,

D1. $(Bew_T(\ulcorner\varphi\urcorner) \wedge Bew_T(\ulcorner\varphi \to \psi\urcorner)) \to Bew_T(\ulcorner\psi\urcorner))$
D2. $Bew_T(\ulcorner\varphi\urcorner) \to Bew_T(\ulcorner Bew_T(\ulcorner\varphi\urcorner)\urcorner)$

D1 is true as long as we have Modus Ponens as a deduction rule. As for D2, it is true as long as Σ_1 true sentences are provable in *T*.

For the proof of the second incompleteness theorem we need that both D1 and D2 are provable in *T*. This is not immediate, since the complexity of D1 and D2 is greater than Σ_1 (it is Δ_2, i.e., both Σ_2 and Π_2).

Up to this point, any recursive theory *T* that contains R_0 suffices. But to prove D1 and D2 in *T*, R_0 is not enough. What we need is the fragment of Peano's Arithmetic (*PA*) known as Σ_1-Induction ($I\Sigma_1$). This is *PA* with the induction schema restricted to Σ_1 formulae. To show that $I\Sigma_1$ proves D1 and D2 above requires a considerable amount of work.

Let T be $I\Sigma_1$ and consider the binary relation B_T:

x is a proof from T of y

If T is recursive, then this is a relation defined by primitive recursion from recursive relations and functions. Using Gödel's β function, which is primitive recursive and allows to code finite sequences, one can show that every relation definable by primitive recursion from recursive relations and total recursive functions is recursive, hence it has a definition by a Σ_1 formula. Let $B_T(x,y)$ be the Σ_1 formula that defines B_T.

It follows from the primitive recursive definition of B_T that for all formulae φ and ψ,

$$B_T(x,\ulcorner\varphi\urcorner) \wedge B_T(y,\ulcorner\varphi \to \psi\urcorner) \to B_T(x*y*\ulcorner\psi\urcorner,\ulcorner\psi\urcorner)$$

where $x*y*\ulcorner\psi\urcorner$ is the code of the proof obtained by concatenating the proof coded by x followed by the proof coded by y, and followed by ψ.

We need to see that the formula above is provable in T. So, let M be a model of T. The formula above will hold in M provided the binary relation defined in M by the formula $B_T(x,y)$ satisfies the same definition by primitive recursion it satisfied in $\mathbb{N}$. This will be the case provided the β function, as defined in M, has the same properties it has in $\mathbb{N}$, namely, it codes finite sequences. The crucial point is to show that the β function has the property that for every $a\in M$ and every sequence f of length a, there are $c,z\in M$ such that $f(i) = \beta(c,z,i)$, for all $i < a$. This is certainly not immediate, since a may be non-standard and so sequences of length a may be infinite. Fortunately, we need only to consider sequences f that are Σ_1-definable in M, and so $I\Sigma_1$ is enough. This is a delicate point, for we need to develop a bit of arithmetic within $I\Sigma_1$: the least number principle for Σ_1 formulae, the existence of the least common divisor of any two elements of M, the Chinese Remainder Theorem, etc.

All this granted, then we can show that the property of $B_T(x,y)$ displayed above holds in M. Hence, we have:

D1. $\quad T \vdash Bew_T(\ulcorner\varphi\urcorner) \wedge Bew_T(\ulcorner\varphi \to \psi\urcorner)) \to Bew_T(\ulcorner\psi\urcorner)$

To prove D2 in T, first we can see that there is a (Σ_1) formula, $Tr_1(x)$, such that for every Σ_1 sentence ψ,

(∗) $\quad T \vdash (Tr_1(\ulcorner\psi\urcorner) \leftrightarrow \psi)$

$Tr_1(x)$ is a truth definition for Σ_1 sentences and, although long, it can be easily written by going trough the usual recursive definitions of denotation of terms and satisfaction of formulae.

Moreover, T proves the completeness theorem for Σ_1 sentences, namely, for every Σ_1 sentence ψ,

$$(**) \quad T \vdash (Tr_1(\ulcorner\psi\urcorner) \to Bew_T(\ulcorner\psi\urcorner)).$$

Indeed, applying Σ_1-Induction to the Σ_1 formula (in the variable ψ)

$$\psi \text{ is a bounded sentence} \wedge Tr_0(\psi) \to Bew_T(\psi),$$

where $Tr_0(x)$ is a definition of truth for bounded sentences, we obtain

$$\forall\psi(\psi \text{ is a bounded sentence} \wedge Tr_0(\psi) \to Bew_T(\psi)).$$

Hence,

$$\forall\varphi(\varphi \text{ is a } \Sigma_1 \text{ sentence} \wedge Tr_1(\varphi) \to Bew_T(\varphi))$$

for if M is a model of T and $M \vDash \varphi \equiv \exists x\psi(x)$, $\psi(x)$ is bounded and $Tr_1(\varphi)$, then $M \vDash \varphi$ and, therefore, $M \vDash \psi(a)$ for some $a \in M$. Extend the language by adding a new constant symbol $\overline{a}$. Then, we have $M \vDash Bew_T(\psi(\overline{a}))$, and so $M \vDash Bew_T(\varphi)$.

Thus, $(*)$ and $(**)$ above yield

$$\text{D2.} \quad T \vdash (Bew_T(\ulcorner\varphi\urcorner) \to Bew_T(\ulcorner Bew_T(\ulcorner\varphi\urcorner)\urcorner))$$

Now all the elements are in place and we can prove Gödel's second incompleteness theorem.

Proof: By D2,

$$T \vdash Bew_T(\ulcorner\theta\urcorner) \to Bew_T(\ulcorner Bew_T(\ulcorner\theta\urcorner)\urcorner)$$

By diagonalization, let θ be such that $T \vdash (\neg Bew_T(\ulcorner\theta\urcorner) \to \theta)$. By D0, $T \vdash Bew_T(\ulcorner Bew_T(\ulcorner\theta\urcorner) \to \neg\theta\urcorner))$. Hence, by D1,

$$T \vdash Bew_T(\ulcorner\theta\urcorner) \to Bew_T(\ulcorner\neg\theta\urcorner)$$

By D0, $T \vdash Bew_T(\ulcorner\theta \to (\neg\theta \to (\theta\wedge\neg\theta))\urcorner)$, from which it follows, by D1,

$$T \vdash Bew_T(\ulcorner\theta\urcorner) \to Bew_T(\ulcorner\theta\wedge\neg\theta\urcorner)$$

By D0, $T \vdash Bew_T(\ulcorner(\theta\wedge\neg\theta) \to \bot\urcorner)$, where $\bot$ is any false sentence, e.g., $0\neq 0$. Hence, by D1,

$$T \vdash Bew_T(\ulcorner\theta\urcorner) \to Bew_T(\ulcorner\bot\urcorner)$$

Let $CON(T)$ be the sentence $\neg Bew_T(\ulcorner\bot\urcorner)$. Thus, $CON(T)$ says, via coding, that T is consistent. We have shown:

$$T \vdash CON(T) \to \theta.$$

We conclude that $T \nvdash CON(T)$, for if $T \vdash CON(T)$, then $T \vdash \theta$ and therefore, by D0, $T \vdash Bew_T(\ulcorner\theta\urcorner)$, that is, $T \vdash \neg\theta$, and so T is inconsistent.

Thus, we have proved:

THEOREM (*Gödel's second incompleteness theorem*) Let T be a recursive theory that contains $I\Sigma_1$. If T is consistent, then $T \nvdash CON(T)$.

The theorem is also true for recursive theories T in any language, not necessarily containing the language of arithmetic. What is required is that $I\Sigma_1$ be *interpretable* in T. This means, roughly, that there are formulas in the language of T that define, in T, a model of $I\Sigma_1$. For then we may add to the language of T the symbols of the language of arithmetic and we may add to T the defining formulas for these symbols, so that the new T in the extended language, call it T', satisfies all the axioms of $I\Sigma_1$. It follows that $T' \nvdash CON(T')$. But since all the new symbols are definable in T, this implies that $T \nvdash CON(T)$.

An important example is Zermelo-Fraenkel Set Theory (*ZF*) and its extensions. In *ZF* we may define the model which has as its universe the finite ordinals, $+$ and $\cdot$ are the usual sum and product of finite ordinals, S is the function that sends each finite ordinal α to $\alpha\cup\{\alpha\}$, and 0 is the empty set. *ZF* proves that this is a model of *PA*. So, if T is a recursive theory that contains *ZF*, we have that $T \nvdash CON(T)$.

III. SHORT PROOFS

We will next present a short proof of Gödel's second incompleteness theorem for Zermelo-Fraenkel set theory. The proof is due to T. Jech [Jech (1994)]. Notice that in this proof there is no need of arithmetizing the syntax.

THEOREM If ZF is consistent, then $ZF \nvdash CON(ZF)$.

To prove the theorem, notice that ZF proves the completeness theorem for first-order logic, hence,

$$ZF \vdash (CON(ZF) \rightarrow ZF \text{ has a model})$$

So, to prove the theorem, and towards a contradiction, suppose that ZF proves that ZF has a model.

Let S be a finite set of axioms of ZF which is enough to define the notions of *model* and *satisfaction*, contains a single instance of the *Comprehension* axiom that will be needed in II below, and proves that ZF has a model.

If $\langle M, E^M\rangle$ and $\langle N, E^N\rangle$ are models of S, we define: $M < N$ iff there is some $\langle m, E^m\rangle \in N$, such that $E^M = (E^m)^N \coloneqq \{\langle x,y\rangle : N \vDash xE^m y\}$. i.e., M is what N thinks that m is.

Notice that if $M < N$, then for every sentence σ of the language of Set Theory,

$$M \vDash \sigma \text{ iff } N \vDash (m \vDash \sigma)$$

Notice also that if $M < N$, then $M \subseteq N$.

I. *If $N \vDash S$, then there is $M < N$*. For suppose $N \vDash S$. Then there is $\langle m,E^m\rangle \in N$ such that $N \vDash (m \vDash ZF)$. Let $M = m$ and let $E^M = (E^m)^N$. Then $M < N$. Notice that $M \vDash S$.

II. *$<$ is a transitive relation*. For suppose m_1 witnesses $M_1 < M_2$ and m_2 witnesses $M_2 < M_3$. Since $E^{m_1}, E^{m_2} \in M_3$, and M_3 satisfies some *Comprehension*, there is $E \in M_3$ such that

$$M_3 \vDash \forall xy(xEy \rightarrow (xE^{m_1} y \wedge \langle x,y\rangle E^{m_2} E^{m_1}))$$

It can be easily seen that $\langle m_1,E\rangle$ witnesses that $M_1 < M_3$.

If $\varphi(x)$ is a formula with x as the only free variable, let $C_{\varphi(x)}$ be the set of natural numbers defined by $\varphi(x)$. Let

$$D = \{\varphi(x): \exists M(M \vDash S \wedge M \vDash \varphi(x) \notin C_{\varphi(x)})\}$$

Let $\theta(x)$ be the formula $\exists M(M \vDash S \wedge M \vDash x \notin C_x)$. So, $C_{\theta(x)} = D$. Then

$$S \vdash (\theta(x) \in D \text{ iff } \exists M(M \vDash S \wedge M \vDash \theta(x) \notin D)).$$

The sentence "$\theta(x) \in D$" plays the role of the sentence θ in the diagonalization theorem. So, let us call it also θ.

III. *If $N \vDash \theta$, then there is $M < N$ such that $M \vDash \neg\theta$*: If $N \vDash \theta$, then there is $m \in N$ such that $N \vDash (m \vDash \neg\theta)$. Let $M = m$ and let $E^M = (E^m)^N$. Then $M < N$ and this is witnessed by M, and so $M \vDash \neg\theta$.

IV. *If $N \vDash \neg\theta$ and $M < N$, then $M \vDash \theta$*: For let m witness $M < N$. If $M \vDash \neg\theta$, then $N \vDash (m \vDash \neg\theta)$. Hence, $N \vDash \theta$. A contradiction.

Now suppose $M_1 \vDash S$. If $M_1 \vDash \theta$, by III there is $M_2 < M_1$ such that $M_2 \vDash \neg\theta$. Otherwise, let $M_2 = M_1$. By I, let $M_3 < M_2$. By IV, $M_3 \vDash \theta$. By III, let $M_4 < M_3$ be such that $M_4 \vDash \neg\theta$. But by II, $M_4 < M_2$, which contradicts IV.

III. Short proofs for weak theories

The argument above can also be used to prove Gödel's second incompleteness theorem for weaker theories, like *PA*, or even $I\Sigma_1$. Let T be *PA*.

Suppose T^* is an extension of T such that:

1. $T^* \vdash$ "CON(T) $\rightarrow$ CON(T^*)".
2. $T^* \vdash$ "CON(T^*) $\rightarrow$ T^* has a model".
3. T^* proves *Comprehension* for bounded formulas.

Now suppose $T \vdash$ CON(T). Then, $T^* \vdash$ CON(T^*), and so $T^* \vdash$ "T^* has a model". We can now proceed as in the proof above and reach a contradiction.

Such a theory T^* exists, for instance, the weak form of second-order arithmetic known as ACA_0 (Arithmetical Comprehension Axiom. See [Hájek and Pudlák (1993)] or [Simpson (1993)]) construed as a first-order theory by the addition of new predicates *Number* and *Set* [Hájek and Pudlák (1993), 1.15].

If T is $I\Sigma_1$, then there also exist theories T^* satisfying (1)-(3) above. For instance, WKL_0 (Weak König's Lemma. See [Simpson (1993)]).

Institució Catalana de Recerca i Estudis Avançats (ICREA), and Departament de Lògica, Història i Filosofia de la Ciència. Universitat de Barcelona
C/ Baldiri Reixac, s/n. 08028 Barcelona
E-mail: bagaria@ub.es

REFERENCES

GÖDEL, K. (1967) 'On formally undecidable propositions of Principia Mathematica and related systems I' in van Heijenoort, J. (ed.), *From Frege to Gödel. A Source Book in Mathematical Logic, 1879-1931*, Harvard University Press, Cambridge, Massachussetts, pp. 596-616. First published as 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik* 38, pp. 173-198.

JECH, T. (1994) 'On Gödel's second incompleteness theorem', *Proceedings of the American Mathematical Society*, 121 (1), pp. 311-313.

HÁJEK, P. AND PUDLÁK, P. (1993). *Metamathematics of First-Order Arithmetic*, Springer-Verlag, Berlin.

SIMPSON, S. G. (1999) *Subsystems of Second Order Arithmetic*, Springer-Verlag, Berlin.