0. INTRODUCTION

Let T be a sufficiently strong theory formalized in the language L_A of (first order) arithmetic. Following Gödel, we want to show that there is a sentence φ of L_A which is true (of the natural numbers) but not provable in T. Gödel's idea was to achieve this by constructing φ in such a way that

(*) φ is true if and only if φ is not provable in T.

Then, assuming (for simplicity) that all theorems of T are true, we are done. For, suppose φ is provable in T. Then, by (*), φ is not true and so, by hypothesis, φ is not provable in T. Thus, φ is not provable in T. But then, by (*), φ is true.

One way to achieve (*) is to find a sentence φ which, in some sense, "says" of itself that it is not provable in T. There are then three major difficulties. First of all, the sentences of L_A deal with natural numbers, they do not deal with syntactical objects such as sentences (of a formal language), proofs, etc. Secondly, even if some of the sentences of L_A can, somehow, be understood as dealing with syntactical objects, it is not clear that it is possible to "say" anything about provability (in T) using only the means of expression available in L_A. And, finally, even if this is possible, there may be no sentence which "says" of *itself* that it isn't provable.

Gödel, however, was able to overcome these difficulties. The first problem is solved by assigning natural numbers to syntactical expressions in a certain systematic way. This is sometimes called a Gödel numbering, and the number assigned to an expression, the Gödel number of that expression. Thus, the numeral of the number assigned to an expression can be regarded as a name of that expression and certain number theoretic statements can be regarded as statements about syntactical objects. (In what follows "φ is a formula", "p is a proof", etc. are short for "φ is the Gödel number of a formula", "p is the Gödel number of a proof", etc.)

To overcome the second difficulty Gödel (re)invented the primitive recursive functions (sets, relations). He showed that a number of crucial properties of (Gödel numbers of) expressions, such as that of being a (well–formed) formula, are primitive recursive. In particular, Gödel showed that, if the set of axioms of T is primitive recursive, this is also true of the relation $PRF_T(\varphi,p)$: p is a proof of the sentence φ in T. φ is provable in T, $PR_T(\varphi)$, if and only if $\exists p PRF_T(\varphi,p)$. This property, however, is not (primitive) recursive.

Gödel then went on to prove that all primitive recursive functions (sets, relations) are definable in L_A. Thus, in particular, there is a formula $Prf_T(x,y)$ of L_A defining $PRF_T(k,m)$. But then $Pr_T(x) := \exists y Prf_T(x,y)$ defines $PR_T(k)$. (In what follows we write $T\vdash \varphi$ for $PR_T(\varphi)$.)

Gödel, however, proved more and this is crucial: for every sentence φ, $T\vdash \varphi$ if and only if $T\vdash Pr_T(\varphi)$. (This is the first time we use the assumption that T is sufficiently strong; but, of course, if T isn't, T is incomplete for that reason.)

This takes care of the second difficulty. So now there remains only the problem of finding a formal sentence which "says" of *itself* that it is not provable in T. Gödel solved this problem in the following way.

Consider the substitution function SBST(k,m) defined by:

SBST(k,m) = the formula obtained from the formula m by replacing the free variable "x" by the numeral of the number k, if m is a formula,

= 0 otherwise.

This function is primitive recursive and so is defined in T by a formula Subst(x,y,z) in the sense that for any formula $\delta(x)$ and any number k,

$T \vdash \forall z(\text{Subst}(k,\delta,z) \leftrightarrow z = \delta(k))$;

in other words, for all k, $\delta(x)$, T proves that: "z satisfies Subst(k,δ,z) if and only if z is the formula $\delta(k)$". Now consider the formula

$\exists y(\text{Subst}(x,x,y) \wedge \neg \text{Pr}_T(y))$,

call it $\gamma(x)$. Let θ be $\gamma(\gamma)$. Intuitively, θ "says" that the result of replacing the variable "x" by the numeral for the number γ in the formula $\gamma(x)$ is not provable in T. But this result, $\gamma(\gamma)$, is θ itself. Thus, θ "says" that θ is not provable in T.

Formalizing this argument we obtain:

(**) $T \vdash \theta \leftrightarrow \neg \text{Pr}_T(\theta)$.

(This is an instance of the very important fixed point lemma; Lemma 1, Chapter 1.) And now Gödel's proof can be completed as follows. First we show that

(***) $T \nvdash \theta$.

Suppose $T \vdash \theta$. Then $T \vdash \text{Pr}_T(\theta)$. But then, by (**), $T \vdash \neg\theta$ and so T is inconsistent (whence $T \vdash \bot$, where $\bot := \neg 0 = 0$), contrary to assumption.

Thus, (***) holds. But this is exactly what $\neg \text{Pr}_T(\theta)$ "says". So $\neg \text{Pr}_T(\theta)$ is true and consequently, by (**), θ is true.

Let Con_T be the sentence $\neg \text{Pr}_T(\bot)$. Con_T is then a natural formalization of "T is consistent". In proving (***) we actually proved that if $T \vdash \theta$ then $T \vdash \bot$ and so that if T is consistent ($T \nvdash \bot$), then $T \nvdash \theta$. It turns out that this proof can be formalized in T provided that T is sufficiently strong. Thus, $T \vdash \text{Con}_T \rightarrow \neg \text{Pr}_T(\theta)$ and so, by (**),

$T \vdash \text{Con}_T \rightarrow \theta$.

But then, since $T \nvdash \theta$, it follows that $T \nvdash \text{Con}_T$; in other words, T cannot prove its own consistency. This is Gödel's second incompleteness theorem.

This, in brief, is what Gödel accomplished (restricted to theories in L_A; generalizations to other theories containing arithmetic, for exampel set theory, are straightforward). In Gödel's original proofs it is assumed that the set of axioms is primitive recursive. Subsequently, when the (general) recursive functions had been defined, it turned out, however, that this assumption could, without altering the structure of the proofs, be replaced by the weaker assumption that the set of axioms is recursive. In fact, it became clear that Gödel's first incompleteness theorem holds for all formal systems, in the most general sense, and is actually a result belonging to recursion theory: the set of true (Π_1) sentences of L_A is not recursively enumerable.

In (the above sketch of) Gödel's proof, and in virtually all proofs in the following chapters, the method of arithmetizing metamathematics, i.e. translating metamathematical concepts, statements, etc. into arithmetic, plays a central role. This method is based on a large number of definitions and preliminary results. In Chapter 1 we introduce the basic notation and terminology and state a number of Facts concerning these notions. These Facts will not be proved but references will be given; some of them are proved in almost every exposition of Gödel's theorems, others require quite extensive proofs that would be out of harmony with the rest of the book. In Chapter 1 we also prove the fixed point lemma (Lemma 1.1), the essential undecidability of Robinson's Arithmetic, Q, a very weak finite subtheory of PA, (Theorem 1.2), and the nonexistence of truth–definitions (Theorem 1.3).

In Chapter 2 we present the first and most important results of the subject, Gödel's incompleteness theorem and his (second) theorem on the unprovability of consistency (Theorems 2.1 and 2.4). Gödel's results were subsequently improved in various respects and we present some of these improvements.

The main result of Chapter 3 is that, assuming that T contains a minimum of arithmetic (Q), every recursively enumerable set is numerated by a (Σ_1) formula in T (even if not all Σ_1 sentences provable in T are true) (Theorem 3.1). We also prove some refinements of this result.

Given that the set of axioms of a theory T is infinite, it is natural to ask if these axioms can be replaced by a finite set of axioms. In Chapter 4 we apply the so called reflection principles to prove some negative results concerning this problem (Theorems 4.1, 4.2). For example, neither Peano Arithmetic, PA, nor any one of its consistent extensions (in L_A) is finitely axiomatizable. On the other hand, every extension of PA has an axiomatization which is irredundant in the sense that none of the axioms can be derived from the other axioms (Theorem 4.6). We also prove the existence of not irredundantly axiomatizable theories (Theorem 4.7).

Let Γ be a set of sentences, for example Σ_n or Π_n. A sentence φ is Γ–conservative over T if every sentence in Γ provable in T + φ is provable already in T. Partial conservativity is studied in Chapter 5, where the basic existence theorems are proved (Theorems 5.2, 5.3, 5.4); it plays an important role in Chapters 6 and 7.

An interpretation of a theory S in a theory T is, roughly speaking, a function t on the set of formulas of S into the set of formulas of T such that t preserves logical form and $T \vdash t(\varphi)$ whenever $S \vdash \varphi$. S is interpretable in T, $S \leq T$, if there is an interpretation of S in T. These concepts were introduced by Tarski. If, in addition, $S \vdash \varphi$ whenever $T \vdash t(\varphi)$, we shall say that t is faithful and that S is faithfully interpretable in T.

Interpretability was originally used as a tool in proofs of (relative) consistency and undecidability. Interpretability (in arithmetical theories) is studied for its own sake in Chapter 6. The key result is that if T is an extension of PA and Con_S is a sentence which (in a suitable sense) in T "says" that S is consistent, then $S \leq T + Con_S$ (the arithmetization of Gödel's completeness theorem; Theorem 6.4). From this it

follows that $S \leq T$ if and only if for every finite subtheory S' of S, $T \vdash Con_{S'}$ (Lemma 6.2) and that if S, too, is an extension of PA, then $S \leq T$ if and only if every Π_1 sentence provable in S is provable in T (Theorem 6.6). We also prove similar characterizations of faithful interpretability (Theorems 6.13, 6.14).

Mutual interpretability is an equivalence relation; its equivalence classes will be called degrees of interpretability. Let T be a consistent extension of PA. The degrees of extensions of T, partially ordered by the relation induced by $\leq$, form a distributive lattice (Theorem 7.1). This lattice is studied in Chapter 7 both from a purely algebraic point of view and in terms of the nature of the theories belonging to a given degree.

It is quite often true in the following pages that a result stated for extensions of PA actually holds for all extensions of some (much) weaker, sometimes finitely axiomatizable, subtheory of PA. We shall, however, pay little attention to facts of this type: what we are mainly interested in here are the properties shared by all theories containing a sufficient amount of arithmetic. But, if a result is (proved to be) true of Q (and its extensions), this will be explicitly noted.

Almost all the results presented in this book hold in a very general setting. In spite of this we shall in Capters 1 – 7, for reasons of simplicity and readability, formulate (and prove) these results for theories formalized in L_A only. We partly make up for this lack of generality in Chapter 8, which is devoted to generalizations, usually straightforward, to theories formalized in other languages; the most important of these is the language of (first order) set theory.