Back to Basics:
Revisiting the Incompleteness Theorems

Peter Smith
Faculty of Philosophy, University of Cambridge

Contents

Preface

These notes were written to accompany three presentations given at a Cameleon workshop weekend in Cambridge in March 2009,[1] and to give proofs of most of the formal theorems that I mentioned.

As will quickly become all too obvious, I am not a research mathematician bringing back news from frontier territory. I'm a philosopher with a relatively amateur interest in logic, and – at least officially – I'm really more concerned with the conceptual implications of major results than with technical niceties. But, a couple of years ago I did write an introductory textbook on the incompleteness theorems,[2] which ended up more mathematically detailed that I originally intended. Even so, it remains at about the same level as other familiar introductory texts such as Boolos and Jeffrey (to mention what is probably the most familiar one),[3] though my book does explore some supplementary themes not well covered elsewhere. Here, I'm taking the opportunity to step back a bit from those hundreds of pages of details, and explain *one* speedy approach to Gödel's Theorems, in fact by a somewhat different route from that in my book. I assume a basic familiarity with some of the background, though I give quick reminders of some key ideas.

[1] http://www.dpmms.cam.ac.uk/∼tf/cameleon.html

[2] Peter Smith, *An Introduction to Gödel's Theorems*, Cambridge University Press, 2007, reprinted with corrections 2008. Anyone who knows my book be warned, however: some jargon here is different – e.g. my usage of the term 'nice theory'.

[3] Now G. Boolos, J. Burgess, and R. Jeffrey, *Computability and Logic*, Cambridge University Press, 5th edn. 2008.

1

The First Theorem revisited

This first chapter highlights two points. (1) We should be careful to distinguish the simple but profound proof of the First Incompleteness Theorem – which says that theories satisfying certain natural conditions will be negation-incomplete – from the further (mostly tedious) work needed to check that this or that theory in fact satisfies the relevant conditions. (2) We need to be careful – a lot more careful than is sometimes the case – in talking about a Gödel sentence's being "true because it truly says that it is unprovable".

1.1 Notational preliminaries

It is crucial for our purposes to be absolutely clear when we are doing ordinary informal arithmetic and when we are looking at formalized claims made inside a formal theory T. To this end, we adopt a now quite widely used convention: *italic* symbols belong to our informal mathematics, sans serif symbols belong to formal T-wffs.

We'll take it that the language of any theory which we are interested in will incorporate – perhaps by definitional extension – *the first-order language of arithmetic* $\mathcal{L}_A$, which has a term '$\mathsf{0}$', the relational expression $\leq$, and function expressions 'S', '$+$', '$\times$'. The *standard numerals* are formed by iterating applications of the successor function, so that e.g. '$\mathsf{SSSS0}$' is the standard numeral for four. To avoid a rash of overlinings, I simply use 'n' for the theory's standard numeral for n (late-alphabet letters like 'x' serve as variables, of course). We use '$\vec{n}$' for a k-tuple of natural numbers (k to be understood from context); '$\vec{\mathsf{n}}$' stands for a corresponding tuple of formal numerals; $\vec{\mathsf{x}}$ is a tuple of variables.

$\mathcal{L}_A$ also includes the usual first-order logical apparatus, in particular quantifiers to run over the numbers (or over proxies for them). Those quantifiers can either be primitive, or – e.g. in the case where T is a rich set theory – they will be defined by restricting the theory's native quantifiers. Either way, quantifier notation in the formal sentences below should always be read as indicating such numerical quantifiers. The *standard interpretation* of $\mathcal{L}_A$ is of course the one which gives the built-in non-logical expressions the obvious arithmetical interpretation, and takes the domain for the numerical quantifiers to be the naturals. To say an $\mathcal{L}_A$-sentence is *true* is just to say that it is true on the standard interpretation.

We'll assume some familiarity with various standard theories of arithmetic built in $\mathcal{L}_A$. Q is Robinson Arithmetic, i.e. it has the usual axioms for successor, and the recursion equations for addition and multiplication, together with the axiom $\forall \mathsf{x}(\mathsf{x} = 0 \lor \exists \mathsf{y}(\mathsf{x} = \mathsf{Sy}))$ which says that every number other than zero is a successor number. PA is first-order Peano Arithmetic, i.e. Q plus all $\mathcal{L}_A$-instances of the first-order induction axiom schema.

The idea of a sensible system of Gödel-numbering for systematically associating expressions with code numbers for them will be taken as well understood. Assume some such system as fixed throughout. If e is an expression, or a sequence of expressions, then in informal contexts $\ulcorner e \urcorner$ is the Gödel number for e, and in formal contexts $\ulcorner e \urcorner$, i.e. the *same* expression[1], stands in for the standard numeral for that number. Context always disambiguates.

We'll also take the idea of a primitive recursive function to be familiar. The main thing you need to recall about such functions is that they are (i) effectively computable, and (ii) computable *without* any open-ended searches such as might be implemented by 'do until' loops, i.e. 'for' loops are enough. Henceforth, 'p.r.' abbreviates 'primitive recursive' (not 'partial recursive').

1.2 Definitional preliminaries

We begin with a handful of definitions (the ideas should all be entirely familiar: these are just reminders, fixing terminology).

Definition 1.2.1 *A theory T is* numeral-aware *if, for all naturals m, n, if $m \neq n$, then $T \vdash \mathsf{m} \neq \mathsf{n}$.*

[1] Rather than the more conventional but rebarbative $\overline{\ulcorner e \urcorner}$.

Definition 1.2.2 *A theory T* captures *(most say* represents*) the k-place function $f(\vec{x})$ iff there is an $(k+1)$-place open wff* $\mathsf{F}(\vec{\mathsf{x}}, \mathsf{y})$ *of T's language such that for all k-tuples $\vec{m}$ of natural numbers, and for all n, if $f(\vec{m}) = n$, then $T \vdash \forall \mathsf{y}(\mathsf{F}(\vec{\mathsf{m}}, \mathsf{y}) \leftrightarrow \mathsf{y} = \mathsf{n})$.*

Definition 1.2.3 *A theory T is* p.r. adequate *iff it is numeral-aware and it captures all p.r. functions.*

It's natural to want a formal arithmetic T to be p.r. adequate: for this is just to say that if we can informally compute that the value of the p.r. function f and show $f(m) = n$, then we can replicate the computation inside the formal theory T and prove a corresponding sentence $\mathsf{F}(\mathsf{m}, \mathsf{n})$. In other words, going formal doesn't stop us doing the p.r. computations that we could do before.

We immediately have a mini-theorem about p.r. adequate theories, which shows that they can 'capture' p.r. relations as well as functions. We'll state it for the two-place case, for that's all we need (and the generalization is obvious).

Definition 1.2.4 *A theory T* captures *the 2-place relation R there is a two-place wff* $\mathsf{R}(\mathsf{x}, \mathsf{y})$ *of T's language such that for all m, n,*

> *if $R(m, n)$, then $T \vdash \mathsf{R}(\mathsf{m}, \mathsf{n})$,*
> *if not-$R(m, n)$, then $T \vdash \neg\mathsf{R}(\mathsf{m}, \mathsf{n})$.*

Theorem 1.2.5 *Suppose T is p.r. adequate. If R is a two-place p.r. relation – i.e. has a p.r. characteristic function – then there is a two-place wff* $\mathsf{R}(\mathsf{x}, \mathsf{y})$ *of T's language which captures R.*

Proof Suppose f is R's characteristic function. So we have $f(m, n) = 1$ if $R(m, n)$, and $f(m, n) = 0$ otherwise. Let F capture f. Then we can put $\mathsf{R}(\mathsf{x}, \mathsf{y}) =_{\text{def}} \mathsf{F}(\mathsf{x}, \mathsf{y}, \mathsf{1})$. Since T is numeral-aware and hence proves $\mathsf{0} \neq \mathsf{1}$, the result immediately follows. □

Without significant loss of generality, we'll take it that T-proofs can be treated as linear sequences of wffs. Then,

Definition 1.2.6 *The relation $Prf_T(m, n)$ holds when m is the Gödel number of a sequence of wffs that is a T-proof, and n is the Gödel number of the wff proved.*

It follows then that $T \vdash \varphi$ iff $\exists x Prf_T(x, \ulcorner\varphi\urcorner)$.

Now, a key construction in what follows involves putting the Gödel number for an open formula – or more accurately, the numeral for that Gödel number – into free variable places in that formula. If we think of the Gödel number as indexing the formula in an enumeration of open formulae, this is like the familiar diagonalization construction, where we put the enumerating index for a function as the argument for the function. Hence we'll say

Definition 1.2.7 *The* diagonalization *of a one-place open wff $\varphi(\xi)$ is $\varphi(\ulcorner\varphi\urcorner)$.*

Definition 1.2.8 *$diag_T(m) = n$ if m Gödel-numbers some one-place open wff $\varphi(\xi)$ of T's language and n Gödel-numbers its diagonalization $\varphi(\ulcorner\varphi\urcorner)$, or else $diag_T(m) = m$.*

Definition 1.2.9 *T is a* p.r. formalized *theory iff $diag_T$ and Prf_T are p.r.*

Definition 1.2.10 *T is a* nice *theory iff it is p.r. adequate, p.r. formalized, and contains first-order logic.*[2]

Again niceness, in this sense, is a desideratum for any formal theory of arithmetic. As we noted before, p.r. adequacy just means it can prove what we can mechanically calculate about p.r. functions. And being p.r. formalized is really just matter of some formal features of the theory being mechanically checkable (in bounded time, without open-ended searches) – and we plainly want formal features to be mechanically checkable, or else what would be the point of going formal in the first place. More about that in a moment.

A nice theory will have in its language two-place expressions which capture $diag_T$ and Prf_T – in fact, it will have many such expressions. We will use notation as follows:

Definition 1.2.11 $\mathsf{Diag}_T(\mathsf{x}, \mathsf{y})$ *stands for any T-wff that captures $diag_T$, and* $\mathsf{Prf}_T(\mathsf{x}, \mathsf{y})$ *stands for any T-wff that captures Prf_T. Further, we put* $\mathsf{Prov}_T(\mathsf{y}) =_{\mathrm{def}} \exists \mathsf{x}\, \mathsf{Prf}_T(\mathsf{x}, \mathsf{y})$*: we'll call such an expression* a generic provability predicate *for T.*

[2] Intuitionistic or even minimal logic would do for almost every result below.

Note however that we are (so far) using a generic notion of capturing, which puts *no* structural constraints on *how* a function or relation is captured. And we mustn't expect that, in particular, Prov_T – on the intended interpretation of T – will be true to the property of numbering a T-theorem (i.e. there's no guarantee that $\mathsf{Prov}_T(\ulcorner\varphi\urcorner) =_{\mathrm{def}} \exists \mathsf{x}\, \mathsf{Prf}_T(\mathsf{x}, \ulcorner\varphi\urcorner)$ will be true on that interpretation just when $T \vdash \varphi$, i.e. just when $\exists x Prf_T(x, \ulcorner\varphi\urcorner)$): we'll soon be finding some counterexamples to that. But as we'll see, our proof of the First Theorem, which deploys a construction using Prov_T, does *not* depend on that predicate having the 'right' semantic interpretation. It just depends on T being able to prove the right things about the embedded Prf_T, whatever it means.

Just for aesthetic reasons, we'll from now on largely suppress subscript T's, restoring them below only when clarity demands.

Next, by a slight abuse of jargon, we'll say that

Definition 1.2.12 *If $\varphi(\mathsf{x})$ is a wff with one free variable, then γ is a* fixed point *for $\varphi(\mathsf{x})$ in T iff $T \vdash \gamma \leftrightarrow \varphi(\ulcorner\gamma\urcorner)$.*

Finally, recall that we say

Definition 1.2.13 *A theory T is ω-*inconsistent *iff, for some $\varphi(\mathsf{x})$, $T \vdash \exists \mathsf{x}\, \varphi(\mathsf{x})$, yet for every m, $T \vdash \neg\varphi(\mathsf{m})$.*

Evidently, an ω-inconsistent theory can't have a sound standard interpretation in the numbers (for the theorem $\exists \mathsf{x}\, \varphi(\mathsf{x})$ can be standardly true if each of $\neg\varphi(\mathsf{m})$ is true too): so again, we want formal theories of arithmetic to be ω-consistent. And since ω-consistency is a matter of not being able to prove a certain combination, ω-consistency entails plain consistency.

1.3 A general version of Gödel's First Theorem

We can now prove a simple theorem about fixed points G for the negation of a generic provability-predicate Prov:

Theorem 1.3.1 *Suppose T is nice and $T \vdash \mathsf{G} \leftrightarrow \neg\mathsf{Prov}(\ulcorner\mathsf{G}\urcorner)$. Then (i) if T is consistent, $T \nvdash \mathsf{G}$. And (ii) if T is ω-consistent, $T \nvdash \neg\mathsf{G}$.*

Proof (i) Suppose $T \vdash \mathsf{G}$. Then $T \vdash \neg\mathsf{Prov}(\ulcorner\mathsf{G}\urcorner)$. But if there *is* a proof of G, then for some m, $Prf(m, \ulcorner\mathsf{G}\urcorner)$, so $T \vdash \mathsf{Prf}(\mathsf{m}, \ulcorner\mathsf{G}\urcorner)$, since T captures

Prf by Prf. Hence $T \vdash \exists \mathsf{x}\,\mathsf{Prf}(\mathsf{x}, \ulcorner \mathsf{G} \urcorner)$, i.e. we also have $T \vdash \mathsf{Prov}(\ulcorner \mathsf{G} \urcorner)$, making T inconsistent. So if T is consistent, $T \nvdash \mathsf{G}$.

(ii) Suppose $T \vdash \neg \mathsf{G}$. Then $T \vdash \mathsf{Prov}(\ulcorner \mathsf{G} \urcorner)$, i.e. $T \vdash \exists \mathsf{x}\,\mathsf{Prf}(\mathsf{x}, \ulcorner \mathsf{G} \urcorner)$. But given T is consistent, there is no proof of G, i.e. for every m, not-*Prf*$(m, \ulcorner \mathsf{G} \urcorner)$, whence for every m, $T \vdash \neg \mathsf{Prf}(\mathsf{m}, \ulcorner \mathsf{G} \urcorner)$. So we have a φ such that T proves $\exists \mathsf{x}\varphi(\mathsf{x})$ while it refutes each instance $\varphi(\mathsf{m})$, which makes T ω-inconsistent. So if T is ω-consistent, $T \nvdash \neg \mathsf{G}$. □

But now a little trick easily proves the *Diagonalization Lemma* which asserts the existence of such fixed points for $\neg\mathsf{Prov}$, and indeed for any open sentence φ:

Theorem 1.3.2 *If T is nice and φ is a one-place open sentence of T's language, then there is sentence γ such that $T \vdash \gamma \leftrightarrow \varphi(\ulcorner \gamma \urcorner)$.*

Proof Put $\delta =_{\mathrm{def}} \forall \mathsf{y}(\mathsf{Diag}(\mathsf{x}, \mathsf{y}) \to \varphi(\mathsf{y}))$, and $\gamma =_{\mathrm{def}} \delta(\ulcorner \delta \urcorner)$. Since diagonalizing δ yields γ, we have $diag(\ulcorner \delta \urcorner) = \ulcorner \gamma \urcorner$, and hence $T \vdash \forall \mathsf{y}(\mathsf{Diag}(\ulcorner \delta \urcorner, \mathsf{y}) \leftrightarrow \mathsf{y} = \ulcorner \gamma \urcorner)$ since T is p.r. adequate. But just from the definition of γ, $T \vdash \gamma \leftrightarrow \forall \mathsf{y}(\mathsf{Diag}(\ulcorner \delta \urcorner, \mathsf{y}) \to \varphi(\mathsf{y}))$. Hence, substituting provable equivalents, we have $T \vdash \gamma \leftrightarrow \forall \mathsf{y}(\mathsf{y} = \ulcorner \gamma \urcorner \to \varphi(\mathsf{y}))$, and therefore $T \vdash \gamma \leftrightarrow \varphi(\ulcorner \gamma \urcorner)$. □

For future reference, note that it would have done just as well to form the wff $\delta' =_{\mathrm{def}} \exists \mathsf{y}(\mathsf{Diag}(\mathsf{x}, \mathsf{y}) \wedge \varphi(\mathsf{y}))$, and put $\gamma =_{\mathrm{def}} \delta'(\ulcorner \delta' \urcorner)$, with the rest of the proof as before.

Theorem 1.3.2 tells us in particular that, for any generic provability predicate, $\neg\mathsf{Prov}(\mathsf{y})$ does indeed have at least one fixed point G. Putting that together with Theorem 1.3.1 immediately yields a general form of the First Incompleteness Theorem:

Theorem 1.3.3 *If T is nice, then there is a sentence G such that (i) if T is consistent, $T \nvdash \mathsf{G}$, and (ii) if T is ω-consistent, $T \nvdash \neg\mathsf{G}$.*

1.4 Giving the First Theorem bite

The proof in the last section was remarkably speedy! But of course, we've got to an incompleteness theorem so very quickly by just assuming that we are dealing with a nice theory, whatever one of those might look like, and by sidestepping the rather tedious work needed to establish that this or that particular theory *is* nice.

How do we establish, then, that a familiar theory of arithmetic like

PA is indeed nice? (1) We have to confirm that the function *diag* and the relation *Prf* for PA are primitive recursive. And (2) we need to show that PA is p.r. adequate.

We said we wanted formal features of the theory being mechanically checkable. Here's a quick and dirty argument that this holds in the right kind of way for the relevant properties of PA, making the two components of (1) true. First, to evaluate $diag(m)$ we need to check whether m is the Gödel code of a wff $\varphi(\xi)$ with one free variable. If it is, we form the wff $\varphi(\mathsf{m})$, and work out its Gödel number: otherwise we set the output to m. That can evidently be done without open-ended searches, But recall that a p.r. function is one that can be computed using only bounded loops: so *diag* is indeed primitive recursive.

Likewise, it only takes a simple bounded computation to confirm whether m numbers a sequence of PA-wffs (assuming, as always, a sensible style of Gödel coding). Another such computation will check whether the sequence confirms to the rules for being a proof in your favoured version of PA – here we are assuming e.g. that the axioms are given in a normal kind of way, so that it is evident that we can check in a bounded way what's an axiom, etc. And then another bounded computation will check whether n numbers this proof's conclusion. So overall, a bounded computation without open-ended searches is again enough to decide whether $Prf(m, n)$. Hence *Prf* for PA is primitive recursive.

If that doesn't convince, and you insist on an explicit p.r. definition for *diag* and (the characteristic function of) *Prf*, then you'll have go through the sort of palaver that Gödel went through in his 1931 paper to show that the cut-down type-theory P which he is discussing there is indeed p.r. axiomatized. This is, in effect, a laborious but routine programming exercise, filling in the details of the procedure we've just outlined. Do it for one standard theory, and you'll be readily convinced that it can be done for any other normally presented axiomatized theory. We'll say no more about this here.

Showing (2), that PA is p.r. adequate, is a lot more interesting. For the proof yields a neat specification for a very large swathe of the theories to which the First Theorem applies (the 'normal' ones, as I'll call them). We also get key information about just how logically simple the undecidable sentences proved to exist by Theorem 1.3.3 can be.

It outline terms, we need to show that (i) PA can capture the so-called 'initial' functions (the zero function, projection functions, and successor). Then we show that (ii) if PA can capture g and h, it can capture their composition $f = g \circ h$, and (iii) if PA can capture g and h, it can capture the function f defined from g and h by primitive recursion – i.e., the function such that $f(\vec{x}, 0) = g(\vec{x})$, and $f(\vec{x}, Sy) = h(\vec{x}, y, f(\vec{x}, y))$. Then, since any p.r. function is definable from the initial functions via repeated definitions by composition and/or primitive recursion, it follows by induction on the length of the definitional chain that PA can capture every p.r. function.

The tricky step in the proof is showing (iii). Take the very simplest sort of case, where we have $f(0) = k$, and $f(Sy) = h(f(x))$. Then $f(m) = n$ just if there is a sequence of numbers $a_0, a_1, a_2, \ldots a_m$, where $a_0 = k$, $h(a_i) = h(a_{i+1})$, and $a_m = n$. So we might hope to capture f in PA if we can in effect talk not just about numbers but about *sequences* of numbers from inside PA. As is familiar, the β-function trick is the way to pull this off – for a β-function in effect takes a code number and spits out a finite sequence. And things are set up so that there's a code for every finite sequence, and generalizing over codes enables us to say, in effect, 'there is a sequence of numbers such that ...'. Once we've got hold of that neat trick, the proof goes through fairly readily for the simplest case, and then the generalization is trivial.

We won't go into further details here. But if we dig inside the adequacy proof we find two things. First, inspection reveals that we don't need the full strength of PA to get adequacy: in fact, the induction-free Robinson Arithmetic Q has all we need. Suppose, then, we add another definition:

Definition 1.4.1 *A* normal *theory is one that p.r. formalized, and which includes Robinson Arithmetic* Q.

Then we have

Theorem 1.4.2 *Any normal theory is nice.*

But we can also extract additional information from adequacy proof, this time about the kinds of expressions we need to do the capturing job. To put the point snappily we need another couple of definitions, by way of reminder.

Definition 1.4.3 *In $\mathcal{L}_A$, we define* bounded quantifiers *contextually as follows:* $(\forall\xi \leq \tau)\varphi =_{\mathrm{def}} \forall\xi(\xi \leq \tau \to \varphi)$*,* $(\exists\xi \leq \tau)\varphi =_{\mathrm{def}} \exists\xi(\xi \leq \tau \wedge \varphi)$ *where τ holds the place for a numeral or a variable.*

Definition 1.4.4 *An $\mathcal{L}_A$ expression is Δ_0 if its only quantifiers can be rewritten as bounded quantifiers. An expression is Σ_1 (resp. Π_1) if* *it is logically equivalent to a Δ_0 expression preceded by zero or more existential (resp. universal) numerical quantifiers.*

Then inspecting the workings of the adequacy proof (perhaps after some small massaging) reveals that

Theorem 1.4.5 Q*, and hence any normal theory, can in fact capture any p.r. function by a* Σ_1 *formula.*

For we only need a Δ_0 wff to capture the initial functions and, it turns out, Gödel's β-function. So as we build up a wff F to reflect f's p.r. definition, we can do all the construction in such a way that the only unbounded quantifiers we need are initial existential ones. Those are introduced when we compose functions (in effect saying there's a number which is the output of one and input to the other), and when we mirror definitions by primitive recursion (in effect saying there's a number for which the β-function spits out a certain sequence).

Note an immediate corollary of this. A theory T containing Q can capture, in particular, $diag_T$ and Prf_T by Σ_1 wffs Diag and Prf. So then the corresponding $\neg\mathsf{Prov}$ will be Π_1, and by inspection the fixed point for $\neg\mathsf{Prov}$ constructed as in the proof of Theorem 1.3.2 will also be Π_1.

Putting everything together, then, from Theorems 1.3.3, 1.4.2 and our last remarks, we get a much sharper version of the First Incompleteness Theorem:

Theorem 1.4.6 *If T is normal, then there is a Π_1 sentence G such that (i) if T is consistent, $T \nvdash \mathsf{G}$, and (ii) if T is ω-consistent, $T \nvdash \neg\mathsf{G}$.*

In summary form, no matter how fancy T is, if it is normal and ω-consistent then it has undecidable purely arithmetic sentences with a merely Π_1 degree of quantifier complexity.

1.5 Generic Gödel sentences and arithmetic truth

The situation, then, is this. We proved the general version of the First Incompleteness Theorem, Theorem 1.3.3, *very* quickly and easily.[3] And we should keep a clear conceptual distinction between what it takes to prove that general result and the further work needed to check that the theorem applies to familiar theories like PA and to confirm that there will

[3] It was, by the way, philosophically important to Gödel that the key proof ideas were simple and immediate.

always be undecidable purely arithmetical sentences of a very low degree of complexity (measuring complexity by quantificational structure), so giving us Theorem 1.4.6

The proofs that I have sketched plainly have their roots in Gödel's original 1931 (though it wasn't until twenty years later that it was shown that containing Q suffices for the First Theorem to apply to a sensibly axiomatized theory). Since then, variant routes to the same incompleteness result have been discovered, for example the familiar one via the unsolvability of the Halting Problem, or the less familiar one due to Kleene which is given in the Appendix. One important thing to note, though, about all those sorts of proofs is that they proceed without deploying semantic ideas. In particular, note that thus far we have *not* needed to say – and consequently haven't said – anything about the truth-values of the undecidable sentences we've given a reciple for constructing.

It is tempting to say – people often *do* say – that our undecidable sentences are constructed so as to say of themselves that they are unprovable, and since they *are* unprovable (if the relevant theory T is consistent), they must be true. But as we'll now see, this is actually wrong.

Let's first start, however, by noting:

Theorem 1.5.1 *If φ is a true Σ_1 sentence, then $\mathsf{Q} \vdash \varphi$. In a word,* Q *is* Σ_1-complete.

Proof It is easy to check that Q can prove any true Δ_0 sentence (as such a sentence is provably equivalent in Q to a propositional combination of *unquantified* correct equations and inequations, which Q is purpose-designed to be able to prove). But the initial existential quantifier(s) in a true Σ_1 sentence must be witnessed by some corresponding terms in a true Δ_0 sentence. So Q can prove any true Σ_1 sentence by existentially quantifying some relevant true Δ_0 sentence which it can prove. □

From this it is immediate that

Theorem 1.5.2 *If φ is a Π_1 sentence of a normal theory T such that $T \nvdash \neg\varphi$, then φ is true.*

Proof Suppose φ is Π_1 but false. Then $\neg\varphi$ is Σ_1 and true. Hence Q can prove $\neg\varphi$, and therefore so can any normal theory T. So if T *can't* prove $\neg\varphi$, $\neg\varphi$ isn't true. So φ *is* true. □

Suppose we now define terms as follows:

Definition 1.5.3 *A* generic Gödel sentence *for a theory T is any fixed point for the negation of any generic provability predicate* Prov_T.

Then, since the generic Gödel sentences of normal ω-consistent theories are unprovable by Theorem 1.3.1, it is immediate that

Theorem 1.5.4 *Any Π_1 generic Gödel sentence for an ω-consistent normal theory T is true.*

Note moreover that *this* result about truth is established without giving any particular interpretation to generic provability predicates, and without claiming that Gödel sentences 'say' that they are unprovable-in-T (even in some stretched sense of 'say').

And that's good, because generic Gödel sentences sentences do *not* in general 'say' that. For we also have:

Theorem 1.5.5 *There are normal (and ω-consistent) theories with undecidable generic Gödel sentences which are* false *$\mathcal{L}_A$ sentences.*

Proof For a given T, there are $\mathcal{L}_A$-wffs (indeed, Σ_1-wffs) $\mathsf{Diag(x, y)}$ and $\mathsf{Prf(x, y)}$ which capture $diag_T$ and Prf_T respectively. The corresponding $\mathsf{Prov(y)}$ is of course an $\mathcal{L}_A$-wff too.

If $diag_T(m) = n$, then by definition $T \vdash \forall \mathsf{y}(\mathsf{Diag(m, y)} \leftrightarrow \mathsf{y} = \mathsf{n})$. Let Θ be *any* closed $\mathcal{L}_A$-wff which is a theorem of T. Plainly it follows that if $diag_T(m) = n$, then $T \vdash \forall \mathsf{y}((\mathsf{Diag(m, y)} \wedge \Theta) \leftrightarrow \mathsf{y} = \mathsf{n})$. In other words, the $\mathcal{L}_A$-wff $\mathsf{Diag'(x, y)} =_{\text{def}} (\mathsf{Diag(x, y)} \wedge \Theta)$ also captures $diag_T$.

So by the remark following the proof of Theorem 1.3.2, the $\mathcal{L}_A$-wff $\mathsf{G}_\Theta =_{\text{def}} \exists \mathsf{y}(\mathsf{Diag'(d, y)} \wedge \neg\mathsf{Prov(y)})$ is a fixed point for $\neg\mathsf{Prov}$, where d is the numeral for a certain Gödel number). Since $\mathsf{Diag'(x, y)}$ contains Θ as a conjunct, when Θ is false so is G_Θ.

Suppose then that T is arithmetically unsound, i.e. entails some *false* sentence of $\mathcal{L}_A$. And suppose Θ in the argument above is such a false theorem. Then, assuming T is normal and ω-consistent, it will have a Gödel sentence G_Θ which is also a false $\mathcal{L}_A$ sentence. □

And since such a Gödel sentence is false it plainly cannot in any sense 'say' what is true, i.e. that it is unprovable.

Note too that if $\mathsf{Prf(x, y)}$ captures Prf then, by the same trick as in the proof, so does $\mathsf{Prf(x, y)}' =_{\text{def}} \mathsf{Prf(x, y)} \wedge \Theta$ for a false theorem Θ, and the corresponding Prov' predicate can thus also be arithmetically false of all numbers, whether they number a theorem or not.

As we'll see in the next section, there is a *sub-class* of Gödel sentences – the 'standard' ones – whose members *are* always true $\mathcal{L}_A$-sentences (assuming T is normal and consistent). But the point to emphasize is that you *can't* just take that result about standard Gödel sentences and apply it to Gödel sentences of the generic kind. Yet is the latter which characteristically feature in modern treatments of the First Incompleteness Theorem.

It is worth really stressing this point, for it certainly isn't unknown for authors to forget that they've officially introduced Gödel sentences in the generic way, but then give an informal gloss about what's going on which only applies to standard Gödel sentences. To take just one example – selected to show that this can happen even in the best books – Mendelson in his classic text in effect characterizes a Gödel sentence G for T as one where $T \vdash \mathsf{G} \leftrightarrow \neg\mathsf{Prov}(\ulcorner\mathsf{G}\urcorner)$, with Prov a generic proof predicate, defined as we have defined it. Changing his notation, he then writes

> G is equivalent in T to asserting that there is no proof in T of G. Hence, G is equivalent in T to an assertion that G is unprovable in T. In other words, G says 'I am not provable in T.' ... Under the standard interpretation, G asserts its own unprovability in T. Therefore, G is true for the standard interpretation.[4]

But not so. For a start, that forgets that if T is unsound, the biconditional theorem $\mathsf{G} \leftrightarrow \neg\mathsf{Prov}(\ulcorner\mathsf{G}\urcorner)$ could be false.

1.6 Canonical and standard Gödel sentences

Consider again the proof that we gestured towards in Section 1.4 to show that Q, and hence any normal theory, is p.r. adequate. This works by telling us how to build up an $\mathcal{L}_A$ expression F which captures a given p.r. function f.

More carefully, you start by taking one way in which we can build up a canonical p.r. definition of f, starting from initial functions and constructing functions by composition and primitive recursion until you arrive at f.

[4] Elliott Mendelson, *Introduction to Mathematical Logic*, Chapman and Hall/CRC, 4th edn. 1997: pp. 204–206. For similar examples from other well-known texts, see the Appendix to Peter Milne, 'On Gödel sentences and what they say', *Philosophia Mathematica* 15 (2007), 193–226.

You then build up F step-by-step to describe that construction, starting from obvious simple ways of capturing the initial functions. Further, you *only* put into F what is necessary to describe the construction of f. We might say that such an F *explicitly recapitulates* a full p.r. definition of f. And using this notion, we'll stipulate

Definition 1.6.1 *An expression* F canonically captures *the p.r. function f if it captures f by explicitly recapitulating a canonical p.r. definition of the function.*

Of course, such an expression won't be unique. For a start, we'll typically be able to give different p.r. definitions for f, hence there are different p.r. definitions for a formal wff to track. And even fixing on one such definition, we can shift around the order of clauses, chose different variables to use in a formal representation, and so on.

Still, whatever the variations, since a wff that canonically captures f clause-by-clause recapitulates some p.r. definition of f, and no more, by construction it is immediate that

Theorem 1.6.2 *If* F *canonically captures the p.r. function f, then* F *is* true to *f, meaning that* $\mathsf{F}(\mathsf{m},\mathsf{n})$ *is arithmetically true iff $f(m) = n$. Similarly, if* R *canonically captures the p.r. relation R, then* R *is* true to *R, meaning that* $\mathsf{R}(\mathsf{m},\mathsf{n})$ *is arithmetically true iff $R(m,n)$.*

Compare the point that we made before. If F captures f in T, and Θ is some closed T-theorem or other, then $\mathsf{F}'(\mathsf{x},\mathsf{y}) =_{\mathrm{def}} \mathsf{F}(\mathsf{x},\mathsf{y}) \wedge \Theta$ also captures f. And if T is unsound, and Θ is some false theorem, then F' is *not* true to f, for $\mathsf{F}(\mathsf{m},\mathsf{n}) \wedge \Theta$ is always false, whether or not $f(m) = n$. Cases where a wff canonically captures f aren't like that: they don't carry along with them free-loading and always-false baggage.

Let's now use superscript 'c' to indicate wffs which canonically capture the relevant p.r. functions. Thus Diag^c_T and Prf^c_T are Σ_1 wffs which canonically capture $diag_T$ and Prf_T respectively, and we define a canonical proof predicate $\mathsf{Prov}^c_T(\mathsf{y}) =_{\mathrm{def}} \exists\mathsf{x}\,\mathsf{Prf}^c_T(\mathsf{x},\mathsf{y})$. We then have

Theorem 1.6.3 *A canonical proof predicate* is faithful *– i.e. is true to the property of numbering a T-theorem, so that* $\mathsf{Prov}^c_T(\ulcorner\varphi\urcorner)$ *is true just in case $T \vdash \varphi$.*

The proof of Theorem 1.3.2 tells us how construct a fixed point for $\neg\mathsf{Prov}^c$ in T using Diag^c (dropping subscripts again). In fact, given the comment after the proof, we have two similar constructions, one using the universal quantifier, one the existential quantifier. Call those indifferently the *diagonal* constructions for a fixed point. Then

Definition 1.6.4 *A* canonical Gödel sentence *for a theory is a fixed point* G^c *for some canonical* $\neg\mathsf{Prov}^c$*, formed by a diagonal construction using some canonical* Diag^c*.*

We already know from Theorem 1.5.4 that if T is normal and ω-consistent, then its canonical Gödel sentences are indeed arithmetically true, since canonical sentences are Π_1 by construction. But we can now strengthen the result just a bit:

Theorem 1.6.5 *If T is normal and (plain) consistent, its canonical Gödel sentences are true.*

Proof We'll take the construction via a universal quantifier (the proof for the existential case runs similarly). Then the canonical Gödel sentence G^c has the form $\forall\mathsf{y}(\mathsf{Diag}^c(\mathsf{d},\mathsf{y}) \to \neg\mathsf{Prov}^\mathsf{c}(\mathsf{y}))$, where d is the Gödel number for $\mathsf{D}^c =_{\mathrm{def}} \forall\mathsf{y}(\mathsf{Diag}^c(\mathsf{x},\mathsf{y}) \to \neg\mathsf{Prov}^\mathsf{c}(\mathsf{y}))$.

On the standard interpretation the quantifier in G^c runs over the numbers, so G^c is true just so long as $(\mathsf{Diag}^c(\mathsf{d},\mathsf{n}) \to \neg\mathsf{Prov}^\mathsf{c}(\mathsf{n}))$ is true for any n. So for G^c to be false, we'd need both (i) $\mathsf{Diag}^c(\mathsf{d},\mathsf{n})$ and (ii) $\mathsf{Prov}^c(\mathsf{n})$ to be true for some n.

That's impossible. Since Diag^c is true to *diag*, for (i) $\mathsf{Diag}^c(\mathsf{d},\mathsf{n})$ to be true requires n to Gödel-number the diagonalization of the wff with number d. But the wff with number d is D^c, and its diagonalization is G^c, so n will have to number G^c.

But that means we can't also have (ii) $\mathsf{Prov}^c(\mathsf{n})$ true. For given T is normal and consistent, T doesn't prove G^c. So we don't have $Prf(m,n)$ for any m. Since Prf^c is true to Prf, that means $\mathsf{Prf}^c(\mathsf{m},\mathsf{n})$ is always false. Whence $\exists\mathsf{x}\,\mathsf{Prf}^c(\mathsf{x},\mathsf{n})$, i.e. $\mathsf{Prov}^c(\mathsf{n})$, is false. □

The reason for spelling this out a bit laboriously is that it reveals that what essentially matters to the proof is that we are dealing with versions of the wffs Diag and Prf that are respectively true to *diag* and *Prf* (and so Prov is true to the property of numbering a provable wff). Constructing those wffs in a canonical way is just one way of forcing them to be true to function/relation that they capture. We can therefore generalize. First, say

Definition 1.6.6 *A* standard *Gödel sentence for T is one constructed by a diagonal construction from $\mathcal{L}_A$-wffs* Diag *and* Prf *that are respectively true to $diag_T$ and $Prov_T$.*

Then what we've shown is that

Theorem 1.6.7 *If T is normal and (plain) consistent, its standard Gödel sentences are true.*

And the point we were making in the last section was, in a slogan, don't confuse the generic with the standard (even among the Gödel sentences which are arithmetic).

There's more to be said about the truth of various subclasses of generic Gödel sentence: but that's enough to be going on with.

2
The Second Theorem revisited

This chapter reviews the familiar route to the Second Incompleteness Theorem, trying to be a bit careful about what it takes to show the unprovability, inside a theory, of standard consistency sentences for that theory. We then see how the Second Theorem can be used to improve the First Theorem – it shows us how to replace the assumption of ω-consistency for half the First Theorem by a significantly weaker condition which is rarely remarked on.

2.1 Definitional preliminaries

Any nice theory T proves $\mathsf{0} \neq \mathsf{1}$. Hence,

Definition 2.1.1 *The absurdity constant* $\perp$ *can be treated as abbreviating* $\mathsf{0} = \mathsf{1}$.

Thus, T is consistent just if it doesn't prove $\perp$.

Definition 2.1.2 $\Box_T\varphi$ *abbreviates* $\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$.

However, we'll in fact – as is conventional – restrict the use of this notation to the case when Prov_T satisfies some so-called 'derivability conditions', to be explained in a moment.

Note that our box actually does a double job: it further abbreviates the long open sentence already abbreviated by Prov_T, and it also absorbs the corner-quotes in a formal wff that turn an expression into the standard numeral for that expression's Gödel number. If you are logically pernickety, then you might be quite upset about introducing a notation which in this way disguises the complex logical character of what is going on.[1] But we'll take abbreviatory convenience to trump notational perfectionism. And to emphasize that $\Box$ is just an abbreviatory device for us, we'll cheerfully call that a provability predicate too.

As we saw, a canonical provability-predicate is true to the property of numbering a T-theorem. So a corresponding $\neg\Box_T\bot$, defined in terms of a canonical provability-predicate, is true just when T is consistent. This motivates a further bit of notation:

Definition 2.1.3 Con_T *abbreviates* $\neg\Box_T\bot$.

Note that a sentence Con_T built from a canonical provability-predicate – a *canonical consistency sentence* for short – is the negation of a Σ_1 sentence, so in that sort of case Con_T is Π_1.

That's the crispest definition of a consistency sentence for T. There are alternatives. Here's another natural one. Suppose $\mathsf{Contr(x,y)}$ captures the p.r. relation which holds between two numbers when one codes for some sentence φ and the other for $\neg\varphi$. Then we could put

Definition 2.1.4 $\mathsf{Con}'_T =_{\mathrm{def}} \neg\exists\mathsf{x}\exists\mathsf{y}(\mathsf{Prov}_T(\mathsf{x}) \wedge \mathsf{Prov}_T(\mathsf{y}) \wedge \mathsf{Contr}(\mathsf{x},\mathsf{y}))$.

But, on modest assumptions, this sort of definition and its variants are equivalent: so we'll stick to the crisp one.

From now on, we will again usually drop subscripts if confusion won't ensue.

2.2 Towards Gödel's Second Theorem

We start with an informal outline sketch for motivation. The First Theorem tells us that, for nice enough T,

(1) If T is consistent then G is not provable in T.

[1] The beef is this. $\Box\varphi$ *looks* as if it ought to be a complex wff embedding φ, so that as φ increases in logical complexity, so does $\Box\varphi$. But not so. However complex φ is, $\ulcorner\varphi\urcorner$ is just a numeral and $\Box\varphi$, i.e. $\mathsf{Prov}_T(\ulcorner\varphi\urcorner)$, stays resolutely at the same degree of complexity – a Σ_1 wff if Prov is canonical.

The logically kosher approach is not to regard the box as an abbreviation, but to introduce a *new* modal language, and then explore a mapping relation that links modal sentences to arithmetical 'realizations' via a $\Box/\mathsf{Prov}$ link. For a properly careful development of this approach, see George Boolos, *The Logic of Provability*, Cambridge University Press, 2003.

Now, at least if Con is canonical, it is true just when T is consistent; and the corresponding $\neg\Box\mathsf{G}$ built from the same canonical provability predicate is true just so long as G is not provable in T. So in that case we can faithfully express (1), half of the First Theorem, inside T by

(2) $\mathsf{Con} \to \neg\Box\mathsf{G}$.

But now reflect that the informal reasoning for the First Theorem is in fact rather elementary. So we might expect that if T is strong enough, it should itself be able to replicate that elementary reasoning, so we have

(3) $T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$.

Call such a result the *Formalized First Theorem* for the relevant provability predicate.

Now, G is defined as a fixed point of T, so that

(4) $T \vdash \mathsf{G} \leftrightarrow \neg\Box\mathsf{G}$.

And we know from the First Theorem that,

(5) If T is consistent, $T \nvdash \mathsf{G}$.

So the Formalized First Theorem immediately yields the unprovability of the relevant consistency sentence. Restoring the subscript, for strong enough T, we'll have

(6) If T is consistent, $T \nvdash \mathsf{Con}_T$.

Which is the Second Incompleteness Theorem: roughly, for the right kind of theories T and the right kind of consistency sentences Con_T, T can't prove its own consistency sentences.

2.3 A general version of Gödel's Second Theorem

So much, then, for an outline sketch. Now let's try to get to the same sort of conclusion more carefully. We start by saying:

Definition 2.3.1 *The* derivability conditions *hold for a provability predicate* Prov *for T just when, using $\Box\varphi$ to abbreviate* $\mathsf{Prov}(\ulcorner\varphi\urcorner)$*,*

(C1) If $T \vdash \varphi$, then $T \vdash \Box\varphi$,
(C2) $T \vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$,
(C3) $T \vdash \Box\varphi \to \Box\Box\varphi$.

When these derivability conditions obtain, we will say that Prov *is a* Bernays *provability predicate.*

The latter label is chosen to give the credit where the main credit is apparently due. As announced, we'll only use the box notation for Bernays provability predicates. In fact, many reserve the label 'provability predicate' for Bernays cases.[2]

Now, rather than blankly taking the derivability conditions as unexplained givens, it is worth pausing to consider *why* they are a *very* natural set of conditions to propose if we want to be able to derive versions of the Formalized First Theorem and hence the Second Theorem. So think again about our quick proof of Theorem 1.3.1, and in particular about how we showed that, if G is a fixed point for $\neg\mathsf{Prov}$ in T, then T can't prove G (assuming T is consistent).

In our new notation, we started from (half) the characterization of G as a fixed point,

(D) $T \vdash\ \mathsf{G} \to \neg\Box\mathsf{G}$,

and we then proved an instance of (C1), i.e.

(C) if $T \vdash \mathsf{G}$ then $T \vdash \Box\mathsf{G}$,

Putting those together shows that, assuming consistency, $T \nvdash \mathsf{G}$.

Now, given that (C) and (D) imply half the First Theorem, we might reasonably expect to be able to argue from the claims that *T 'knows' that (C) holds* and *T 'knows' that (D) holds* to the conclusion *T 'knows' that half the First Theorem holds.*

More carefully, the thought that T 'knows' that (D) holds is tantamount to

$$T \vdash \Box(\mathsf{G} \to \neg\Box\mathsf{G}),$$

And that follows from (D) by (C1). The thought that T 'knows' that (C) holds is captured by (C3). The hoped-for inference is therefore from (D) and instances of (C1) and (C3) to the Formalized First Theorem. That should go through so long as T is able to cope with the idea that if $\varphi \to \psi$ and φ are both provable, then so is their consequence ψ. Which is just what (C2) provides.

[2] And some call *any* $\Box$ satisfying (C1) to (C3) a provability predicate, whether it is not built from a predicate capturing *Prf*. On this generous usage, even the $\Box'\varphi =_{\text{def}} \varphi$ and $\Box''\varphi =_{\text{def}} (\ulcorner\varphi\urcorner = \ulcorner\varphi\urcorner)$ count as provability predicates in the stretched sense.

So, putting everything together, the three derivability conditions – together with the characterization of G as a fixed point – look as if they *ought* to give us the Formalized First Theorem.

And they do.

Theorem 2.3.2 *Suppose T is nice, $\Box$ is a Bernays provability predicate for T, G is any fixed point for $\neg\Box$, and $\mathsf{Con} =_{\mathrm{def}} \neg\Box\bot$. Then $T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$.*

Proof T's niceness means that it contains enough logic to ensure that $T \vdash \neg\varphi \leftrightarrow (\varphi \to \bot)$. Using (C1) and (C2) it follows that $\Box\neg\varphi$ and $\Box(\varphi \to \bot)$ are provably equivalent in T. Call that principle (A). Then we argue as follows, relying in T's propositional logic:

1.	$T \vdash \mathsf{G} \to \neg\Box\mathsf{G}$	Given
2.	$T \vdash \Box(\mathsf{G} \to \neg\Box\mathsf{G})$	From 1, by (C1)
3.	$T \vdash \Box\mathsf{G} \to \Box\neg\Box\mathsf{G}$	From 2, by (C2)
4.	$T \vdash \Box\neg\Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	Using (A)
5.	$T \vdash \Box\mathsf{G} \to \Box(\Box\mathsf{G} \to \bot)$	From 3 and 4
6.	$T \vdash \Box\mathsf{G} \to (\Box\Box\mathsf{G} \to \Box\bot)$	From 5, by (C2)
7.	$T \vdash \Box\mathsf{G} \to \Box\Box\mathsf{G}$	Instance of (C3)
8.	$T \vdash \Box\mathsf{G} \to \Box\bot$	From 6 and 7
9.	$T \vdash \neg\Box\bot \to \neg\Box\mathsf{G}$	Contraposing
10.	$T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$	Definition of Con $\Box$

And then the same argument as given in the last section immediately yields the following general version of Gödel's Second Theorem:

Theorem 2.3.3 *Suppose T is a nice consistent theory, $\Box$ is a Bernays provability predicate for T, and $\mathsf{Con}_T =_{\mathrm{def}} \neg\Box\bot$. Then $T \nvdash \mathsf{Con}_T$.*

Careful though. We *aren't* saying that every proof of an instance of the Second Theorem will go via the derivability conditions and the Formalized First Theorem. For example, for a sensible choice for $\mathsf{Con}_{\mathsf{Q}}$, $\mathsf{Q} \nvdash \mathsf{Con}_{\mathsf{Q}}$. But Robinson arithmetic is too weak to prove derivability conditions. But the route via the derivability conditions is, so to speak, the generic one.

2.4 Giving the Second Theorem bite

As with proving the general version the First Theorem, proving the general version of the Second Theorem was pretty simple. The hard work comes next. We have an abstract characterization of sufficient conditions for the unprovability of certain kinds of consistency sentence from inside a theory. But we now need to see which generic provability predicates for which theories are Bernays, so we can apply the general Theorem to particular cases.

Evidently, the interesting cases of the second theorem are going to be those where the unprovable sentence Con genuinely merits the label 'consistency sentence' – i.e. where Con is true (on the intended interpretation for T) if and only if T really *is* consistent. One way of ensuring this is to concentrate on cases where Con is built from a provability predicate $\Box$ which is both *faithful* (in the sense of Theorem 1.6.3) and simultaneously *Bernays* – if such a combination is possible. By definition, if $\Box$ is faithful, $\Box\varphi$ will be arithmetically true just if φ is provable; so in particular Con will be an arithmetically true sentence just if $\perp$ is not provable and hence T is consistent. If $\Box$ is Bernays as well, the general Second Theorem holds and Con will be unprovable.

So the question is: are there theories with provability predicates which are both faithful and Bernays?

Indeed there are. Let's say

Definition 2.4.1 *A theory T is Σ-normal if it is normal and has induction for all Σ_1 predicates.*

Obviously PA, and any stronger theory, is Σ-normal. And for future reference,

Definition 2.4.2 $\mathsf{I}\Sigma_1$ *is the weakest Σ-normal theory, i.e. it is* Q *plus induction for all Σ_1 predicates.*

Now we have the key result:

Theorem 2.4.3 *If T is Σ-normal, and* Prov^c *is canonical for T, then* Prov^c *satisfies the derivability conditions – so is both faithful and Bernays.*

Proof We'll just sketch, with vigorous hand-waving, how to show that a canonical – and hence Σ_1 – provability predicate gets to satisfy the third derivability condition (as this is the tricky case).

It is enough to show that, if T is Σ-normal and ψ is Σ_1, then $T \vdash \psi \to \mathsf{Prov}^c(\ulcorner\psi\urcorner)$. For if this is true in general, it will hold in the special case where ψ is the Σ_1 wff $\mathsf{Prov}^c(\ulcorner\varphi\urcorner)$, which gives us the third condition.

Now, Theorem 1.5.1 tells us than any theory containing Q is Σ_1-complete: so if T is normal, then if ψ is a true Σ_1 sentence, $T \vdash \psi$. So what we need to show is that if T can do induction for Σ_1 wffs, then T 'knows', at least case by case, that it is Σ_1-complete. And it will 'know' that if the argument for T's Σ_1-completeness can be formalized inside T. Which, needless to say, is a great deal easier said than done!

If you spell out the argument for Q's Σ_1-completeness, it works by an informal induction on the complexity of the relevant wffs. In other words, you first show that Q can prove the simplest true Σ_1-wffs; then you show that Q can prove more complex Σ_1 truths built up using the connectives and bounded quantifiers, given that it can prove all simpler Σ_1 truths.

We now need to run the same kind of induction again. We first show that T can prove $\psi \to \mathsf{Prov}^c(\ulcorner\psi\urcorner)$ for the simplest Σ_1-wffs ψ. Then we show that T can prove the same for more complex Σ_1-wffs so long as it can prove all the instances for simpler Σ_1-wffs. This is messy work but not particularly difficult: again, we need to dig inside the provability predicate – i.e. inside $\exists\mathsf{v}\mathsf{Prf}^c(\mathsf{v},\mathsf{x})$ – and rely on the fact that Prf^c recapitulates the definition of the p.r. relation *Prf*. By hypothesis, T can handle induction for proving facts about the relevant Σ_1 wffs which compose Prf: and it can also facts about the Σ_1 wffs which capture the p.r. functions which handle syntactic relations among wffs, e.g. when representing one as the negation of another. So, it all comes out in the wash! □

So putting together Theorems 2.3.3 and 2.4.3, we have

Theorem 2.4.4 *If T is Σ-normal and consistent, then for any consistency sentence Con_T defined from a canonical provability predicate, Con_T is true but $T \nvdash \mathsf{Con}_T$.*

2.5 Comparisons

Find any provability predicate Prov for T, do the diagonal construction trick, and you have an unprovable Gödel sentence. But it *isn't* the case that, find any provability predicate, form $\neg\mathsf{Prov}(\ulcorner\bot\urcorner)$, and you have an unprovable consistency sentence. You need to have constructed the right sort of consistency sentence, e.g. from a provability predicate which is Bernays. That's the key contrast to make between the sort of incompleteness the First Theorem tells us about, and the unprovability of consistency.

The natural way of constructing a Bernays provability predicate is via a canonical capturing of Prf_T. This crucially keeps everything Σ_1. Now, since, given the relevant Gödel coding, we can read off from a canonical predicate the p.r. function/relation it captures, we might say that the predicate 'has the right meaning': so it is sometimes said that dealing with canonical provability predicates injects a distinctively intensional, semantic, dimension into the Second Theorem as against with the First Theorem. But that doesn't seem a particular helpful way of describing the First/Second constrast. After all, our tightest version so far of the First Theorem, i.e. Theorem 1.4.6, traded on canonical provability predicates to ensure there's always a Π_1 undecidable sentence.

Note by the way that actually doing the construction of either a Gödel sentence or a consistency sentence requires more than knowing that T *is* recursively axiomatizable, somehow or other. We need to know a presentation of T *as* a (primitive) recursively axiomatized theory. (This matters when we come of consider general constructions where we form new theories by iterating some processes for extending a theory, e.g. by adding consistency sentences: but that's a story for another time.)

2.6 Further results about provability predicates

Let's return from the specific to the more abstract. Here's a portmanteau theorem bringing together four easy results with a fifth and much more intriguing one:

Theorem 2.6.1 *Suppose T is nice, $\Box$ is a Bernays provability predicate for T, Con is $\neg\Box\bot$, and G is a fixed point for $\neg\Box$. Then*

(i) *For any φ, $T \vdash \neg\Box\varphi \to \mathsf{Con}$.*
(ii) $T \vdash \neg\Box\mathsf{G} \to \mathsf{Con}$.
(iii) $T \vdash \mathsf{G} \leftrightarrow \mathsf{Con}$.
(iv) $T \vdash \mathsf{Con} \leftrightarrow \neg\Box\mathsf{Con}$.
(v) *For any φ, if $T \vdash \Box\varphi \to \varphi$, then $T \vdash \varphi$.*

Proof (i) By logic $T \vdash \bot \to \varphi$. So by (C1) and (C2), $T \vdash \Box\bot \to \Box\varphi$. Contraposing gives the result.

(ii) This is a special case of (i).

(iii) Follows from (ii), Theorem 2.3.3 and the definition of G as a fixed point for $\neg\Box$.

(iv) One direction follows from (i). For the other, we have already shown $T \vdash \mathsf{Con} \to \mathsf{G}$. So by (C1) and (C2), $T \vdash \Box\mathsf{Con} \to \Box\mathsf{G}$. But the Formalized First Theorem tells us that $T \vdash \mathsf{Con} \to \neg\Box\mathsf{G}$. Therefore $T \vdash \mathsf{Con} \to \Box\mathsf{Con}$.

(v) This is *Löb's Theorem.* Assume that, for a given φ,

1. $T \vdash \Box\varphi \to \varphi$, i.e. $T \vdash \mathsf{Prov}(\ulcorner\varphi\urcorner) \to \varphi$.

Now consider the open wff $(\mathsf{Prov}(\mathsf{x}) \to \varphi)$. By hypothesis, T is nice, so we can invoke the general Diagonalization Lemma and apply it to this wff. Hence for some γ, T proves $\gamma \leftrightarrow (\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \varphi)$. Or, in the box notation,

2.	$T \vdash \gamma \leftrightarrow (\Box\gamma \to \varphi)$	By the Lemma
3.	$T \vdash \gamma \to (\Box\gamma \to \varphi)$	From 2
4.	$T \vdash \Box(\gamma \to (\Box\gamma \to \varphi))$	From 3, by (C1)
5.	$T \vdash \Box\gamma \to \Box(\Box\gamma \to \varphi)$	From 4, by (C2)
6.	$T \vdash \Box\gamma \to (\Box\Box\gamma \to \Box\varphi)$	From 5, by (C2)
7.	$T \vdash \Box\gamma \to \Box\Box\gamma$	By (C3)
8.	$T \vdash \Box\gamma \to \Box\varphi$	From 6 and 7
9.	$T \vdash \Box\gamma \to \varphi$	From 1 and 8
10.	$T \vdash \gamma$	From 2 and 9
11.	$T \vdash \Box\gamma$	From 10, by (C1)
12.	$T \vdash \varphi$	From 9 and 11

Hence, if $T \vdash \Box\varphi \to \varphi$, then $T \vdash \varphi$. □

Some commentary. Since T can't prove Con, (i) means that T can't prove any $\neg\Box\varphi$, including the true ones. So although by the first derivability condition T is assumed to 'know' about what it *can* prove, it 'knows' nothing about what it *can't* prove. Next, (iii) implies that, since G is undecidable in a nice ω-consistent theory, so is Con. And (iv) reinforces the message that a fixed point for a negated provability predicate (even a canonical one) – a Gödel sentence in the generic sense – needn't in any sense 'say of itself that it is unprovable': consistency sentences don't do that. Finally, and perhaps surprisingly, (v) tells us that, T *can't* in general prove that if $\Box\varphi$ then φ, however otherwise well-behaved $\Box$ is.

To lead up to our next theorem, suppose that S is some nice theory, and $T = S + \alpha$ is the theory you get by adding the sentence α as an additional axiom to S. Trivially, if T entails no absurdity, S doesn't either; in other words, if the more encompassing theory T is consistent, so is the weaker theory S. That's a simple matter of logic. How might that logical fact be reflected in the relationship between consistency sentences Con_T and Con_S?

Theorem 2.6.2 *If* Con_S *is a canonical consistency sentence for* S*, then there is a canonical consistency sentence* Con_T *for* $T =_{\text{def}} S + \alpha$ *such that* $\vdash \mathsf{Con}_T \to \mathsf{Con}_S$*.*

Proof $\neg\mathsf{Con}_S$ is of the form $\exists \mathsf{x}\mathsf{Prf}_S(\mathsf{x}, \ulcorner\bot\urcorner)$, with Prf_S canonically capturing the Prf_S relation for S. This means it will spell out a condition which is satisfied by a number if it codes for a sequence where every member is an axiom or follows from previous members in the sequence by the rules of inference. This canonical Prf will therefore contain a clause like $\mathsf{Axiom(z)}$ which tells us what numbers code for S axioms. Replace that with $\mathsf{Axiom'(z)} =_{\text{def}} \mathsf{Axiom'(z)} \lor \mathsf{z} = \ulcorner\alpha\urcorner$. Then we get a new predicate Prf_T which canonically captures the Prf_T relation for T. A little reflection shows, moreover, that $\mathsf{Axiom(z)}$ occurs positively in Prf_S; so its disjunctive replacement means that $\exists \mathsf{x}\mathsf{Prf}_S(\mathsf{x}, \ulcorner\bot\urcorner)$ logically entails $\exists \mathsf{x}\mathsf{Prf}_T(\mathsf{x}, \ulcorner\bot\urcorner)$. Which is to say that the negation of the given canonical Con_S entails the negation of a corresponding canonical Con_T. Contrapose and we are done. □

I do not know how much that last result can be improved. But it is enough for our purposes, and enables us to show:

Theorem 2.6.3 *If* S *is* Σ*-normal and consistent, then it has an extension* T *which is consistent but* ω*-inconsistent, where* $T \vdash \neg\mathsf{Con}_T$ *for some canonical consistency sentence.*

Proof Put $T =_{\text{def}} S + \neg\mathsf{Con}_S$, for a canonical consistency sentence $\mathsf{Con}_S = \neg\exists \mathsf{x}\, \mathsf{Prf}(\mathsf{x}, \ulcorner\bot\urcorner)$. Then:

(i) That can't be inconsistent, since $S \nvdash \mathsf{Con}_S$.

(ii) By hypothesis S is consistent, so for every m, it is false that $Prf_S(m, \ulcorner\bot\urcorner)$. So for every m, $S \vdash \neg\mathsf{Prf}_S(\mathsf{m}, \ulcorner\bot\urcorner)$. Hence for every m, $T \vdash \neg\mathsf{Prf}_S(\mathsf{m}, \ulcorner\bot\urcorner)$. But trivially, $T \vdash \exists \mathsf{x}\, \mathsf{Prf}_S(\mathsf{x}, \ulcorner\bot\urcorner)$. So the augmented theory is ω-inconsistent.

(iii) Since $T \vdash \neg\mathsf{Con}_S$, the previous theorem gives $T \vdash \neg\mathsf{Con}_T$ for some canonical consistency sentence. □

How can a consistent theory 'prove its own inconsistency'? By speaking falsely. Which ω-inconsistent theories always do – for such theories cannot have a sound interpretation in arithmetic.

2.7 Back to the First Theorem

It's an immediate from Theorem 2.6.1 (iii), and Theorem 2.6.3 that

Theorem 2.7.1 *There are nice consistent but ω-inconsistent theories T such that, for some Gödel sentence* G*,* $T \vdash \neg\mathsf{G}$.

That shows that the First Theorem *has* to have an asymmetric form, at least while we are still constructing undecidable sentences by Gödel's method of diagonalizing the negation of a provability predicate: T's consistency is enough to ensure $T \nvdash \mathsf{G}$, but not enough to ensure $T \vdash \neg\mathsf{G}$. Can we, however, improve the First Theorem by weakening the current requirement of ω-consistency in the second limb of the theorem (while still constructing undecidable sentences in the same way)?

It was Kreisel who seems first to have remarked – a surprisingly long time after Gödel's paper – that the general requirement of ω-consistency can for normal theories be weakened to ω-consistency-for-Σ_1-wffs, or *1-consistency* for short. Equivalently, we can use the notion of Σ_1-soundness, where

Definition 2.7.2 *T is* Σ_1-sound *iff, for any Σ_1 sentence φ such that $T \vdash \varphi$, φ is true.*

For suppose T is normal. Then we can construct a canonical and hence true Gödel sentence G. Being canonical, G is Π_1. Its negation $\neg\mathsf{G}$ is therefore a false Σ_1 sentence. So if T is Σ_1-sound it can't prove $\neg\mathsf{G}$. Hence, we have another version of the First Theorem:

Theorem 2.7.3 *If T is normal, then there is a Π_1 sentence G such that (i) if T is consistent, $T \nvdash G$, and (ii) if T is Σ_1-sound, $T \nvdash \neg G$.*

Now, so long as T is just a bit stronger than Q, e.g. is Σ-normal, then T can prove the equivalence of – for example – $\exists \mathsf{x}_1 \exists \mathsf{x}_2 \exists \mathsf{x}_3 \varphi$ with $\exists \mathsf{x}(\exists \mathsf{x}_1 \leq \mathsf{x})(\exists \mathsf{x}_2 \leq \mathsf{x})(\exists \mathsf{x}_3 \leq \mathsf{x})\varphi$. Generalizing and applying, T can prove the equivalence of a many-unbounded-quantifier Σ_1 wff with a one-unbounded-quantifier Σ_1 wff. Suppose then that such a T is *not* Σ_1-sound. That means it proves a false theorem of the one-unbounded-quantifier form $\exists \mathsf{x}\, \varphi(\mathsf{x})$, where φ is Δ_0. Since the theorem is false, $\neg\varphi(\mathsf{m})$ must be true for each m. Since T is normal, it can prove all true Δ_0 wffs. So T proves $\neg\varphi(\mathsf{m})$ for each m. Which makes T ω-inconsistent. Contraposing, if T is ω-consistent, it is Σ_1-sound. Which makes Theorem 2.7.3 generally at least as strong as Theorem 1.4.6.

In fact it's strictly stronger, as there are Σ_1-sound theories which aren't ω-consistent – though giving an example would take us too far afield. And in any case, we'll now see how to *further* weaken the assumption of Σ_1-soundness in proving the undecidability of Gödel sentences. For we have

Theorem 2.7.4 *Suppose, as before, T is nice, $\Box$ is a Bernays provability predicate for T, Con is $\neg\Box\bot$, and G is a fixed point for $\neg\Box$. Then (i) if T is consistent, $T \nvdash \mathsf{G}$, and (ii) if $T + \mathsf{Con}$ is consistent, then $T \nvdash \neg\mathsf{G}$.*

Proof The new part is (ii). Suppose $T \vdash \neg\mathsf{G}$. Then by Theorem 2.6.1 (iii), $T \vdash \neg\mathsf{Con}$, and $T + \mathsf{Con}$ is inconsistent. So if $T + \mathsf{Con}$ is consistent, then $T \nvdash \neg\mathsf{G}$. $\Box$

We can't do better than this, for we already know that if $T + \mathsf{Con}$ is not consistent, i.e. $T \vdash \neg\mathsf{Con}$, then $T \vdash \neg\mathsf{G}$ by Theorem 2.6.1 (iii) again. While on the other hand,

Theorem 2.7.5 *There are nice theories T such that (i) $T + \mathsf{Con}$ is consistent but (ii) T is not Σ_1-sound (and hence not ω-consistent).*

Proof Start with a normal sound theory, R (take PA, to fix ideas). Let S be the theory $R + \mathsf{Con}_R$. Let T be the theory $R + \neg\mathsf{Con}_S$. (Take the consistency sentences here to be suitably chosen canonical ones so we can apply Theorem 2.6.2.)

Then (ii) T is not Σ_1-sound. For since R is sound, so is S and the Π_1 sentence Con_S is true, so trivially T entails the false Σ_1 sentence $\neg\mathsf{Con}_S$.

1.	$T + \mathsf{Con}_T$ is inconsistent	Assumption
2.	$R + \neg\mathsf{Con}_S \vdash \neg\mathsf{Con}_T$	From 1, by defn.

Now plainly, by our definitions, if $T \vdash \bot$, then $R \vdash \mathsf{Con}_\mathsf{S}$. And with a bit of effort, as in the proof of Theorem 2.6.2, we can get that logical truism reflected formally in the behaviour of provability predicates. So

3.	$R + \neg\mathsf{Con}_S \vdash \Box_R\mathsf{Con}_\mathsf{S}$	From 2
4.	$R \vdash \neg\mathsf{Con}_S \to \Box_R\mathsf{Con}_\mathsf{S}$	From 3
5.	$R \vdash \mathsf{Con}_R \to \neg\Box_R\mathsf{Con}_R$	Thm 2.6.1 (iv)
6.	$R \vdash \mathsf{Con}_S \to \mathsf{Con}_R$	Thm 2.6.2
7.	$R \vdash \Box_R\mathsf{Con}_S \to \Box_R\mathsf{Con}_R$	From 6, by (C1) and (C2)
8.	$R \vdash \mathsf{Con}_R \to \mathsf{Con}_S$	From 4, 5, 7
9.	$R + \mathsf{Con}_R \vdash \mathsf{Con}_S$	From 8

10. $S \vdash \mathsf{Con}_S$ By defn. □

But since S is sound and hence consistent, (10) is ruled out by the Second Theorem. Hence $T + \mathsf{Con}_T$ is consistent. □

We can now make our general Theorem 2.7.4 more specific in the obvious way:

Theorem 2.7.6 *Suppose, T is Σ-normal, and* G *and* Con *are canonical and framed from the same provability predicate. Then (i) if T is consistent, $T \nvdash \mathsf{G}$, and (ii) if $T + \mathsf{Con}$ is consistent, then $T \nvdash \neg\mathsf{G}$.*

When Con is canonical and so Π_1, Σ_1 soundness means $T \nvdash \neg\mathsf{Con}$ and hence that $T + \mathsf{Con}$ is consistent. We've seen the converse doesn't hold, however. Which gives an interesting twist to the usual story. We might naturally think of the Second Theorem as a pendant to the First, and then take 'consistency extensions' of theories as something to be explored later on, if at all. But now it turns out that at least the first consistency extension of a theory comes into the story earlier, in giving the Gödelian version of the First Theorem with weakest sort of consistency assumption.

2.8 Introducing Rosserized provability predicates

I said 'Gödelian version' to remind us that if we depart from the Gödelian mode of construction and use Rosser's construction instead we can – as is very familiar – at last get a fully symmetric form of the incompleteness theorem.

Take Prf for a given theory T as before, and let $\overline{\mathsf{Prf}}(\mathsf{m}, \mathsf{n})$ similarly capture the p.r. relation which holds when m numbers the proof of the *negation* of the wff with number n. Then we defined Rosser's provability predicate as follows:

Definition 2.8.1 $\mathsf{RProv}(\mathsf{x}) =_{\mathrm{def}} \exists\mathsf{v}(\mathsf{Prf}(\mathsf{v}, \mathsf{x}) \land (\forall\mathsf{w} \leq \mathsf{v})\neg\overline{\mathsf{Prf}}(\mathsf{w}, \mathsf{x}))$.

Then, echoing Theorem 1.3.1, we have – via a distinctly messier but still elementary proof –

Theorem 2.8.2 *Suppose T is nice and $T \vdash \mathsf{R} \leftrightarrow \neg\mathsf{RProv}(\ulcorner\mathsf{R}\urcorner)$. Then if T is consistent, $T \nvdash \mathsf{R}$ and $T \nvdash \neg\mathsf{R}$.*

Note next that by various bits of trickery we can so construct a canonical RProv for a normal theory so that it comes out Σ_1 and the Gödel-style construction of a fixed point for $\neg\mathsf{RProv}$ will again yield a Π_1 wff. Putting everything together we therefore get

Theorem 2.8.3 *If T is normal and consistent, then there is a Π_1 sentence R such that $T \nvdash \mathsf{R}$, and $T \nvdash \neg\mathsf{R}$.*

We won't pursue Rosserized incompleteness theorems any further here. But the general idea of Rosserized proof-predicates, which build in some consistency constraints into their very definition, is interesting for another reason: we'll take it up in the third lecture.

3
'The best and most general version'

After the glory days of the 1930s, Gödel's comments on the details of his incompleteness theorems were few and far between. However, he did add a brief footnote to the 1967 translation of a much earlier piece on 'Completeness and consistency'. And Gödel thought that his brisk remarks in that footnote were sufficiently important to repeat them in a short paper in 1972, in a section entitled 'The best and most general version of the unprovability of consistency in the same system'.[1]

Gödel makes two main points. This last chapter – a rather lighter-weight afterword – aims to explain them both.

3.1 Consistency minded proofs

Canonical provability predicates satisfy the derivability conditions and yield unprovable canonical consistency sentences. But can there be a *non-canonical* sentence CCon_T – a sentence built from a non-canonical proof predicate – which in some fairly natural way still 'says' that T is consistent but which *is* provable in T itself?

The answer is 'yes'. As Gödel puts it in his note:

> ... the consistency (in the sense of non-demonstrability of both a proposition and its negation), even of very strong systems S, *may* be provable in S.

There are a number of possibilities here, using various non-canonical consistency sentences. We'll concentrate on the construction that – at least at the outset – looks least like a cheap trick.

Start with the following thought. When trying to establish an as-yet-unproved conjecture, mathematicians will use any tools to hand, bringing to bear any background assumptions that they are prepared

[1] Kurt Gödel, *Collected Works Vol. II*, Oxford University Press, 1990, p. 305.

to accept in the context. The more improvisatory the approach, the less well-attested the assumptions, then the greater the risk of lurking inconsistencies emerging, requiring our working assumptions to be revised. A check needs to be kept that apparent new results cohere with secure background knowledge. Only a derivation which passes the coherence test has a chance of being accepted as a *proof.*

So let's consider a possible way of capturing something of the idea that a genuine proof should involve checking consistency with what's gone before. Say that φ and ψ are simple contradictories when one is the negation of the other. Then,

Definition 3.1.1 *There is a* consistency-minded proof *of φ in the formal system T iff (i) there is an ordinary T-derivation of φ with Gödel number m, while (ii) there isn't already a T-derivation of ψ with a Gödel number less than m, where φ and ψ are simple contradictories.*

Now, take the numerical relation *Contr* defined as follows:

Definition 3.1.2 *$Contr(m,n)$ holds if m and n are numbers for a pair of simple contradictories.*

Evidently, *Contr* is primitive recursive. And we can use it to define another relation

Definition 3.1.3 $CPrf(x,y) =_{\text{def}} Prf(x,y) \wedge$
$\neg(\exists u \leq x)(\exists w \leq u)(Prf(u,w) \wedge Contr(y,w))$

where, to give a nice bound to the second quantifier, we rely on the fact that, on a sensible coding scheme, if u codes for a proof of the wff with the Gödel number w, then $w < u$. By construction, $CPrf_T(m,n)$ holds when m codes for a consistency-minded proof of the sentence numbered n. And since the new quantifiers are bounded, it is primitive recursive.

The p.r. relation *Contr* can, of course, be captured in a nice theory by a corresponding wff $\mathsf{Contr}(\mathsf{x},\mathsf{y})$. And if we define

Definition 3.1.4 $\mathsf{CPrf}(\mathsf{x},\mathsf{y}) =_{\text{def}} \mathsf{Prf}(\mathsf{x},\mathsf{y}) \wedge$
$\neg(\exists \mathsf{u} \leq \mathsf{x})(\exists \mathsf{w} \leq \mathsf{u})(\mathsf{Prf}(\mathsf{u},\mathsf{w}) \wedge \mathsf{Contr}(\mathsf{y},\mathsf{w}))$.

this wff will in fact capture the p.r. relation *CPrf*. And now, just as we define the provability predicate Prov_T in terms of Prf_T, we can define the consistency-minded provability predicate

Definition 3.1.5 $\mathsf{CProv(y)} =_{\mathrm{def}} \exists \mathsf{x}\, \mathsf{CPrf(x, y)}$.

$\mathsf{CProv}(\ulcorner\varphi\urcorner)$ will therefore be true just so long as there is a consistency-minded proof of φ.

Two remarks about our definitions. First, CProv is, of course, a *very* close cousin of RProv, the Rosser provability predicate which we met in Section 2.8. But Rosser provability is, so to speak, lopsidedly consistency-minded: φ is only Rosser-provable so long as there is no 'smaller' proof of $\neg\varphi$ (but $\neg\varphi$ can be Rosser-provable even if there is a 'smaller' proof of plain φ). By contrast, the property expressed by CProv is *symmetrically* consistency-minded.

Second, and much more importantly we have the following. Suppose we are dealing with a theory T which actually *is* consistent: then every T-proof is de facto consistency-minded. In this case, the relation *CPrf* is extensionally the same relation as *Prf*, so CPrf (non-canonically) captures *Prf* too. Moreover, $\mathsf{CProv(n)}$ will be true just when $\mathsf{Prov(n)}$ is true. So, for a consistent theory, CProv is in fact faithful to the same familiar provability property as Prov.

3.2 A provable consistency sentence

So, at least at first sight, there seems to be some interest in the idea of consistency-minded proofs and consistency-minded proof-predicates like CProv_T. What if we build a consistency sentence from such a predicate?

We'll take a consistency sentence in the second style that we met in Defn. 2.1.4. Let's now put

Definition 3.2.1 $\mathsf{CCon}_T =_{\mathrm{def}}$
$$\neg\exists \mathsf{x}\exists \mathsf{y}(\mathsf{CProv}_T(\mathsf{x}) \land \mathsf{CProv}_T(\mathsf{y}) \land \mathsf{Contr(x, y)}).$$

This might seem to be a natural enough way of defining a consistency sentence in terms of consistency-minded provability. Note again that, if the theory we are dealing with *is* consistent, then since CProv_T is faithful to the property of plain provability, then we might say that CCon_T obliquely (via the Gödel coding) expresses the claim that the theory is consistent.

But now it is quite straightforward to show that

Theorem 3.2.2 *For any T, $\mathsf{I\Sigma_1} \vdash \mathsf{CCon}_T$; hence for any Σ-normal T, $T \vdash \mathsf{CCon}_T$.*

Proof Argue inside $\mathsf{I\Sigma_1}$ and let's drop subscript T's for readability. We'll suppose $\neg\mathsf{CCon}$, that is to say $\exists\mathsf{x}\exists\mathsf{y}(\mathsf{CProv}(\mathsf{x}) \wedge \mathsf{CProv}(\mathsf{y}) \wedge \mathsf{Contr}(\mathsf{x},\mathsf{y}))$, and derive a contradiction.

For some c and d, therefore, we have $\exists\mathsf{v}\,\mathsf{CPrf}(\mathsf{v},\mathsf{c}) \wedge \exists\mathsf{v}\,\mathsf{CPrf}(\mathsf{v},\mathsf{d}) \wedge \mathsf{Contr}(\mathsf{c},\mathsf{d})$. Hence, for some a, b, $\mathsf{CPrf}(\mathsf{a},\mathsf{c}) \wedge \mathsf{CPrf}(\mathsf{b},\mathsf{d}) \wedge \mathsf{Contr}(\mathsf{c},\mathsf{d})$.

Σ_1-induction is (more than) enough to prove $\forall\mathsf{x}(\mathsf{x} \leq \mathsf{y} \vee \mathsf{y} \leq \mathsf{x})$. Hence we can prove that $(\mathsf{a} \leq \mathsf{b} \vee \mathsf{b} \leq \mathsf{a})$.

Now argue by cases. Let $\mathsf{a} \leq \mathsf{b}$. By hypothesis, we have $\mathsf{CPrf}(\mathsf{a},\mathsf{c})$; which by definition implies $\mathsf{Prf}(\mathsf{a},\mathsf{c})$. But – assuming a standard kind of Gödel-numbering, where codes for proofs are larger than codes for the wff proved – that in turn implies $\mathsf{c} \leq \mathsf{a}$.

Put together $\mathsf{a} \leq \mathsf{b}$, and $\mathsf{c} \leq \mathsf{a}$ with $\mathsf{Prf}(\mathsf{a},\mathsf{c})$ and $\mathsf{Contr}(\mathsf{c},\mathsf{d})$ and we can derive $(\exists\mathsf{u} \leq \mathsf{b})(\exists\mathsf{w} \leq \mathsf{u})(\mathsf{Prf}(\mathsf{u},\mathsf{w}) \wedge \mathsf{Contr}(\mathsf{w},\mathsf{d}))$. But by definition $\mathsf{CPrf}(\mathsf{b},\mathsf{d})$ implies the negation of that. Contradiction.

In the same way, $\mathsf{b} \leq \mathsf{a}$ also leads to contradiction. So we are done: our initial supposition leads through to contradictions so the first part of the theorem is proved. The second part of the theorem then follows trivially. □

So here we've found a not-too-bizarrely-constructed consistency sentence CCon_T, one that 'says' of a consistent T that it *is* consistent, yet is such that T proves it.

3.3 What does that show?

So Gödel is right. Suppose T is Σ-normal and consistent. Then, being consistent, all T's proofs are consistency-minded. And it *can*, in particular, prove a sentence which claims, in Gödel's words, the 'non-demonstrability of both a proposition and its negation' in that system itself by using a provability-predicate which is faithful to the property of numbering a T-theorem.

But now note that, since $\mathsf{I\Sigma_1}$ is sound, and entails CCon_T for any T, CCon_T is true for any T – so tells us nothing discriminating about it.

For example, it is perfectly possible that there are consistency-minded proofs in an inconsistent T of φ, $\varphi \to \psi$, and $\neg\psi$. That's because the smallest-numbered putative T-derivation of ψ – e.g. the result of chaining a proof of φ after a proof of $\varphi \to \psi$ and then using modus ponens – could have a g.n. greater than a proof of $\neg\psi$.[2]

Which goes to show that the truth of CCon_T leaves wide open the

[2] Since $T \vdash \mathsf{CCon}_T$, we know that the derivability conditions can't all hold for

possibility that T's set of consistency-minded theorems is *semantically* inconsistent (i.e. there is no interpretation respecting the meaning of the logical apparatus which makes those consistency-minded theorems all true). Hence, to take an extreme example, showing that $\mathsf{I\Sigma_1} \vdash \mathsf{CCon_{ZFC}}$ and – therefore that $\mathsf{CCon_{ZFC}}$ is true – goes no way at all towards deciding whether the standard set theory ZFC is semantically and hence syntactically consistent in the ordinary sense.

So: even though the initial motivating idea was not implausible, it has turned out that 'symmetrically Rosserized' consistency sentences are in themselves useless constructions. But our exploration does reveal something important, that how we put together consistency sentences out of which sort of provability predicate matters crucially for their provability or unprovability. There's much more to be said about this, but not here.

3.4 Introducing reflection principles

Our second topic concerns what happens if you add instances of a reflection schema to a theory.

We start with some definitions.

Definition 3.4.1 $\Box_T\varphi \to \varphi$ *is a* reflection schema *for a theory T, where φ holds the place for any T-sentence, and $\Box$ is a provability predicate for T.*

Definition 3.4.2 *If the permitted instances of $\Box\varphi \to \varphi$ are restricted to cases where φ is Π_1, then by a slight abuse we'll call this a* Π_1 reflection schema.

Definition 3.4.3 *A theory S* proves Π_1 reflection for T *(for a given $\Box$) iff, for every Π_1 sentence φ belonging to T's language, $S \vdash \Box\varphi \to \varphi$.*

A theory T is Π_1-sound, of course, when for every Π_1 sentence φ, if $T \vdash \varphi$, then φ is true. It is immediate that a theory T is Π_1-sound then every instance of a Π_1 reflection schema is true.

Now we add two more specific definitions, concerning ways to extend a Σ-normal theory T. Fix on a canonical provability predicate for T:

the consistency-minded provability predicate CProv_T (else we'd be able to show $T \nvdash \mathsf{CCon}_T$). We now see that, for a start, the condition (C2) can't hold.

Definition 3.4.4 *The theory $T\mathsf{C}$ is the theory you get by adding T's canonical consistency sentence $\mathsf{Con} = \neg\Box\bot$ as a new axiom.*

Definition 3.4.5 *The theory $T\sqcap$ is the theory you get by adding to T all instances of its Π_1 reflection schema in $\Box$ as additional axioms.*

It turns out, as we'll show in a moment, that these are equivalent theories.

3.5 Limitations on reflection

But first, the key result – trivial given the First Theorem! – is that

Theorem 3.5.1 *If T is normal and consistent, then T cannot prove all instances of any given Π_1 reflection schema for T.*

Proof Suppose G_T is a Π_1 fixed point for the provability predicate in the reflection schema. Then if T proved all instances of its Π_1 reflection schema, then $T \vdash \Box\mathsf{G} \to \mathsf{G}$. But, being normal, $T \vdash \mathsf{G} \leftrightarrow \neg\Box\mathsf{G}$. From which it would follow that $T \vdash \mathsf{G}$ contrary to the first incompleteness theorem. □

Of course, if T and the relevant $\Box$ satisfy the conditions for Theorem 2.6.1, (v) to apply, then we get a more general failure. Take any Π_1 wff φ which isn't already a theorem of T. Then by Löb's Theorem, $T \nvdash \Box\varphi \to \varphi$.

3.6 An equivalence

Take it throughout this section that we are talking of Σ-normal theories, and the provability predicates we are talking about are canonical and hence the derivability conditions hold.

Suppose we think T is *sound.* Then we'll believe that whatever it proves is true. So we'll think that any instance of a reflection schema for T is in fact also *true.* In particular the instances of a Π_1 reflection schema are all true. So let's consider the extended theory $T\sqcap$ which we get by adopting all those instances as further axioms.

That's still a p.r. axiomatized theory, since we can test whether a wff is one of the new axioms without an unbounded search. Given that T is sound, $T\sqcap$ is also normal and sound, so must be consistent.

It is immediate that $T\sqcap$ proves the relevant Con_T. Why? Trivially,

$\bot$ is Π_1. So since $T\mathsf{\Pi}$ proves all instances of T's Π_1 reflection schema, $T\mathsf{\Pi} \vdash \Box_T\bot \to \bot$, hence $T\mathsf{\Pi} \vdash \neg\Box_T\bot$. It follows that $T\mathsf{\Pi}$ proves Gödel sentences which are fixed points for $\Box_T$.

We also have the converse: the theory $T\mathsf{C}$ which has Con_T as an additional axiom proves all instances of the relevant Π_1 reflection schema for T. For suppose φ is some Π_1 sentence. Then $\neg\varphi$ is Σ_1, so we can invoke the principle that T 'knows' that it is Σ_1 complete, that we sketched a proof of in demonstrating Theorem 2.4.3, so $T \vdash \neg\varphi \to \Box\neg\varphi$. Whence

1.	$T\mathsf{C} \vdash \neg\varphi \to \Box\neg\varphi$	Weakening
2.	$T\mathsf{C} \vdash \neg\varphi \to (\varphi \to \bot)$	Logic
3.	$T\mathsf{C} \vdash \Box(\neg\varphi \to (\varphi \to \bot))$	From 2, by C1
4.	$T\mathsf{C} \vdash \Box\neg\varphi \to \Box(\varphi \to \bot)$	From 3, using C2
5.	$T\mathsf{C} \vdash \Box(\varphi \to \bot) \to (\Box\varphi \to \Box\bot)$	By C2
6.	$T\mathsf{C} \vdash \neg\varphi \to (\Box\varphi \to \Box\bot)$	From 1, 4, 5
7.	$T\mathsf{C} \vdash \neg\Box\bot$	Con is an axiom!
8.	$T\mathsf{C} \vdash \Box\varphi \to \varphi$	From 6, 7

Since φ was an arbitrary Π_1 sentence, that shows that $T\mathsf{C}$ proves any instance of Π_1 reflection for T. So, in sum,

Theorem 3.6.1 *The theories $T\mathsf{\Pi}$ and $T\mathsf{C}$ are equivalent.*

What's the interest in these reflections about reflection? Well, let's ask: what routes are there that can lead us to accepting a canonical Gödel sentence G_T as true? *One* kind of route goes via an explicit judgement that T is consistent – and there is a considerable variety in the reasons we might have for forming that judgement, depending on T. But now let's ask: do we *have* to go via explicit reflections about consistency? Can we perhaps go instead via thoughts about Π_1 reflection and rely on results such as that $T\mathsf{\Pi} \vdash \mathsf{G}$?

The idea would be: cheerfully immersed in the theory T, I accept what it proves (at least about Π_1 arithmetic propositions). Reflecting on that, I recognize that prepared I'm also prepared to endorse any instance I meet of the Π_1 reflection schema, and so I'll advance to endorsing $T\mathsf{\Pi}$ as well. And that's enough reflection to get me to accept T's canonical Gödel sentences – without, so to speak, surveying T's output all-at-once to make a consistency judgement.[3]

[3] For more about whether this sort of consideration can be made to fly, see my book, §27.8.

3.7 'Inner' and 'outer' consistency

Suppose that, like Gödel, we are interested in Hilbert's project of trying to use relatively uncontentious reasoning to show that the wilder reaches of infinitistic mathematics are still 'safe'. What does 'safety' consist in?

Well, an inconsistent theory is of course not 'safe', as it allows us to 'prove' anything. But a consistent theory that allows us to 'prove' e.g. sentences of elementary arithmetic that are simply false is just as bad. Indeed, the second way of going wrong is arguably more problematic. After all, if a theory T is *internally* inconsistent, then we can in principle always come to recognize this using only elementary logic, for we can be presented with some finite derivation of both φ and $\neg\varphi$ from T. On the other hand, suppose T is *inconsistent-with-elementary-arithmetic*, meaning that $T \vdash \varphi$ where φ is a false sentence of elementary arithmetic. Then if $\neg\varphi$ can't be proved in e.g. PA (and we know that there are arithmetic truths which are independent of PA), then we stand in need of some other, non-infinitary, way of recognizing that T is going wrong.

And how far does 'elementary arithmetic' extend – meaning the unproblematically contentful stuff that isn't already infected with problematic infinistic ideas? Well, Hilbert and his school thought: at least as far as the Π_1 truths of arithmetic. (And if the finitist *isn't* allowed to endorse general truths like $\forall\mathsf{x}\forall\mathsf{y}(\mathsf{x}+\mathsf{y}=\mathsf{y}+\mathsf{x})$ as contentful then his position is hobbled indeed!) So, more specifically, we might say in Hilbertian spirit: a safe theory T should not entail any false Π_1 theorems. In other worlds, we want: if $T \vdash \varphi$ and φ is Π_1, then φ is true.

So: being Π_1-sound looks to be a entirely natural minimal condition for being a 'safe' theory. Gödel calls in effect this same condition 'outer consistency' (though the label hasn't ever really caught on). And he remarks that 'for the usual systems [outer consistency] is trivially equivalent with consistency'. Why so?

Theorem 3.7.1 *If T is normal, T is consistent iff it is Π_1-sound.*

Proof Suppose T proves a *false* Π_1 sentence φ. $\neg\varphi$ will then be a *true* Σ_1 sentence. But in that case, since T extends Q and so is Σ_1-complete, T will prove $\neg\varphi$, making T inconsistent. Contraposing, if T is consistent, it proves no false Π_1 sentence, so is Π_1-sound.

The converse is trivial, since if T is inconsistent, we can derive anything in T, including false Π_1 sentences and so T isn't Π_1-sound. □

Note that this is – so to speak – a proof from outside T. (Reality check: what's the relation between this result and Theorem 3.6.1?)

3.8 'The best and most general version'

To establish formally that a theory T is Π_1-sound and hence to that extent 'safe' is a matter of proving Π_1 reflection for T. But Theorem 3.5.1 tells us that a normal T can't prove Π_1 reflection for itself, and it follows that an arithmetic like PA certainly can't prove Π_1 reflection for a *stronger*, more infinitary, theory T either, unless T is inconsistent and proves everything. *In sum, we can't use relatively modest arithmetical reasoning to prove the 'safeness' of a (consistent) infinitary theory T.*

And this is already enough to undermine the Hilbertian project. Yet note that the unprovability of 'outer consistency', as Gödel calls it, is an easier result than the unprovability of 'inner consistency', i.e. the unprovability of Con_T. For the former result holds given that we can show that there are Π undecidable Gödel sentences. So, to show the unprovability of *outer* consistency/Π_1-soundness for T, we *don't* have to do the hard work of showing e.g. that the derivability conditions hold. By contrast, showing the unprovability of *inner* consistency is tougher.

Which is why the result that a theory T can't prove Π_1 reflection for T (if the conditions for the First Theorem apply to T) might be said to be 'the best and most general version of the unprovability of consistency in the same system'.

And that is almost, but not quite, what Gödel says in his 1967/1972 note. What he actually asserts to be the best result is this (with a trivial change):

> [W]hat can be shown to be unprovable in T is the fact that the rules of the equational calculus applied to equations demonstrable in T between primitive recursive terms yield only correct numerical equations (provided only that T possesses the property which is asserted to be unprovable).

Here Gödel's T is a properly axiomatized theory which includes enough arithmetic to be p.r. adequate. But 'equations between primitive recursive terms' are expressions of the form $\forall\mathsf{x}(\mathsf{fx} = \mathsf{gx})$, where fx and gx express p.r. functions, so $\mathsf{fx} = \mathsf{gx}$ expresses a p.r. property. And, assuming T contains enough arithmetic to know about p.r. functions, such expressions will be provably equivalent to Π_1 sentences.

So what Gödel is saying is in effect that T can't prove Π_1 reflection for T. And he claims that the condition under which this holds is that

Π_1 reflection is actually true for T, i.e. T is Π_1-sound, which – given T is normal – is equivalent to T's being consistent. But T's being consistent is just the key to the First Theorem holding for T (given that T is properly axiomatized and contains enough arithmetic). Hence Gödel's remark seems to be making a version of our point above: so long as the conditions for the First Theorem applying to T hold, T can't prove Π_1 reflection for T.

Commentators on Gödel have sometimes attributed him a much more complex line of argument here. But I hesitate to go beyond the simple reading of his remarks that I've given. And those remarks are surely interesting enough even on the simple reading.

Appendix: Kleene's proof of incompleteness

The incompleteness result, especially in that last summary form, can of course be proved in other ways that don't go via the Diagonalization Lemma. For example, one familiar proof is via the unsolvability of the Halting Problem: and it is perhaps not so surprising that we can relate claims about limitations of formal machines with claims about limitations of formal proof gadgets. However, just because it is so very pretty but is also relatively unknown, we'll here briefly explore a different route to incompleteness due to Kleene.

First we need a quick reminder about the Normal Form Theorem. Here's a slightly cut-down version:

Theorem A 1 *There is a three-place p.r. function C and a one-place p.r. function U such that any one-place partial recursive function can be given in the standard form*

$$f_e(n) =_{\mathrm{def}} U(\mu z[C(e, n, z) = 0])$$

for some value of e.

Since U and C are primitive recursive, and so can be computed without any open-ended searches, this means that any partial recursive function at all can be computed with a just a single open-ended search (here triggered by the operator μz which tells us: 'for given e and n, search through $z = 0, 1, 2, \ldots$ until $C(e, n, z) = 0$, and return the first value of z, if any, where the C-function defaults to zero').

A proof of the Normal Form Theorem can be found in any suitable text.[4] But perhaps we can usefully here give a rough-and-ready sketch of *one* way of showing that this is true. So recall that a partial recursive

[4] Standard notation has the T function rather than C: but let's keep 'T' for theories.

function f_e can be computed by a Turing machine M_e, the e-th machine in some standard enumeration. And now consider the state-of-play function $\mathit{code}(e, n, z)$ which takes the machine index number e and the input n, and then for step-number z returns a number giving a Gödel-like code describing the state of the Turing machine tape at step z of M_e's computation for input n. Suppose we set things up so that, if and only the machine halts at step z, the code then defaults to zero for larger values of z. Then using a decoding function to extract information from the tape, we have

$$f_e(n) = \mathit{decode}(\mathit{code}(e, n, \mu z[\mathit{code}(e, n, Sz) = 0]))$$

since $\mu z[c(e, n, Sz) = 0]$ gives the step number when the machine halts. But sensible *code* and *decode* functions will be primitive recursive (because with a sensible scheme, coding and decoding don't involve open-ended searches). Which more of less gives us the Normal Form Theorem – and a bit of not-quite-trivial massaging gets us the rest of the way.

We now can show

Theorem A 2 *Kleene's Normal Form Theorem entails that any normal ω-consistent theory has an undecidable Π_1 sentence.*

This seems to link a not-obviously-limitative theorem with Gödelian incompleteness in a slightly more surprising way.

Proof Suppose that there is a normal system of arithmetic T which is ω-consistent (and hence consistent), and is negation complete for Π_1 sentences: in other words, for every Π_1 sentence φ either $T \vdash \varphi$ or $T \vdash \neg\varphi$.

Since T is normal, there will be a four-place Σ_1-wff C which captures the p.r. function C that appears in Kleene's theorem (so note, for use below, that $\neg\mathsf{C}$ is therefore Π_1). Now consider the following definition,

$$\overline{f}_e(n) = \begin{cases} U(\mu z[C(e, n, z) = 0]) & \text{if } \exists z[C(e, n, z) = 0] \\ 0 & \text{if } T \vdash \forall \mathsf{z}\neg\mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{z}, \mathsf{0}) \end{cases}$$

We'll first show that, given our assumptions about T, this well-defines an effectively computable total function for any e.

Take this claim in stages. First, we need to show that the two conditions in our definition are exclusive and exhaustive:

(i) The two conditions are mutually exclusive (so the double-barrelled definition is consistent). For assume that both (a) $C(e, n, k) = 0$

for some number k, and also (b) $T \vdash \forall \mathsf{z} \neg \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{z}, \mathsf{0})$. Since the formal predicate C captures C, (a) implies $T \vdash \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{k}, \mathsf{0})$. Which contradicts (b), given that T is consistent.

(ii) The two conditions are exhaustive. Suppose the first of them doesn't hold. Then for every k, it isn't the case that $C(e, n, k) = 0$. So for every k, $T \vdash \neg \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{k}, \mathsf{0})$. But by hypothesis T is ω-consistent, so $T \nvdash \exists \mathsf{z}\, \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{z}, \mathsf{0})$. Hence $T \vdash \forall \mathsf{z}\, \neg \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{z}, \mathsf{0})$, or else T wouldn't be negation complete for Π_1 wffs.

Which proves that, given our initial assumptions, our conditions well-define a total function $\overline{f}_e$.

Now we check that, for given e, and with our our initial assumptions still in place, $\overline{f}_e(n)$ is effectively computable. Start searching through numbers $k = 0, 1, 2, \ldots$. Keep going until we find the first k such that either $C(e, n, k) = 0$ (and then put $\overline{f}_e(n) = U(\mu z[C(e, n, z) = 0])$, as instructed), or else k Gödel-numbers a proof in T of $\forall \mathsf{z} \neg \mathsf{C}(\mathsf{e}, \mathsf{n}, \mathsf{z}, \mathsf{0})$ (and then put $\overline{f}_e(n) = 0$). Each of those conditions can be effectively tested for – in the second case because T is p.r. formalized, so we can effectively check whether k codes for a sequence of expressions which is indeed an T-proof of the given wff. And it follows from what we've just shown that eventually one of the conditions must hold.

Two more observations (still with our original assumptions in play):

(iii) Suppose f_e is a total recursive function, then of course $f_e(n) = U(\mu z[C(e, n, z) = 0])$ with the condition $\exists z(C(e, n, z) = 0)$ obtaining for every n. And so for that index $f_e = \overline{f}_e$. Hence an enumeration of the $\overline{f}_e$ by their index e will certainly include all the total recursive functions.

(iv) However, since we now know how to effectively compute a value of the function $\overline{f}_e$ for any input, the diagonal function $d(n) =_{\text{def}} \overline{f}_n(n) + 1$ is *also* total and effectively computable.

So, by Church's Thesis in labour-saving mode, $d(n)$ must be a total recursive function. But by construction, d is distinct from all the $\overline{f}_e$, hence distinct from any total recursive function. Contradiction!

Hence, it follows from Kleene's Theorem that, if T is a normal ω-consistent theory, it can't also be negation complete for Π_1 sentences. □

Which surely *is* rather pretty!