

### 21.3 The Gödel-Rosser Theorem again

(a) Back to the technicalities! The next major task is to fulfil our promise to give a proof of the Gödel-Rosser Theorem, i.e. Theorem 19.6: we'll show that nice theories are incomplete *without* using the extra assumption of  $\omega$ -consistency (or weaker versions of that assumption, like 1-consistency).

As we noted in Section 19.3, Rosser's basic trick is to construct a sentence which 'says' *If I'm provable, there's already a proof of my negation*. Here's a natural way of developing that idea.

Consider the relation  $\overline{Prf}_T(m, n)$  which holds when  $m$  numbers a  $T$ -proof of the *negation* of the wff with number  $n$ . This relation is obviously p.r. given that  $Prf_T$  is; so assuming  $T$  is nice it will be captured by a wff  $\overline{Prf}_T(x, y)$ .<sup>4</sup> So let's consider *the Rosser provability predicate* defined as follows:

$$RProv_T(x) =_{\text{def}} \exists v(Prf_T(v, x) \wedge (\forall w \leq v) \neg \overline{Prf}_T(w, x)).$$

Then a sentence is Rosser-provable in  $T$  – its g.n. satisfies the Rosser provability predicate – if it has a proof (in the ordinary sense) and there's no 'smaller' proof of its negation.

(b) Now we apply the Diagonalization Lemma, not to the negation of a regular provability predicate (which is what we just did to get Gödel's First Theorem again), but to the negation of the Rosser provability predicate. The Lemma then tells us that there's a sentence  $R_T$  which is a fixed point for  $\neg RProv_T(x)$ . That is to say, assuming  $T$  is nice,

$$T \vdash R_T \leftrightarrow \neg RProv_T(\ulcorner R_T \urcorner).$$

Hence if  $T$  is sound and its theorems are true, then  $R_T$  will indeed be true just so long as it isn't Rosser-provable. In other words,  $R_T$  is true just if, if it's provable, there is already a proof of its negation. So by the argument of Section 19.3, if  $T$  is sound,  $T \not\vdash R_T$  and  $T \not\vdash \neg R_T$ .

(c) But now we want to show that we don't need the assumption of soundness: consistency is enough. To show this, we first prove the following general result, which is the analogue of Theorem 21.1:

**Theorem 21.2** *Let  $T$  be a nice theory, and let  $\gamma$  be any fixed point for  $\neg RProv_T(x)$ . Then  $T \not\vdash \gamma$  and  $T \not\vdash \neg\gamma$ .*

showing me a pre-publication version of Milne (2007), in which he gives chapter and verse on the sins of various textbooks!

<sup>4</sup>Consider the p.r. function defined by  $neg(x) =_{\text{def}} \ulcorner \neg \urcorner \star x$ , where ' $\star$ ' is the concatenation function from Section 15.6. We have  $neg(\ulcorner \varphi \urcorner) = \ulcorner \neg \urcorner \star \ulcorner \varphi \urcorner = \ulcorner \neg\varphi \urcorner$ . So  $neg$  takes the g.n. of a wff and returns the g.n. of its negation.

Now let's suppose we can introduce a function-symbol into  $T$ 's language to capture this function (see the end of Section 12.2). We'll use the symbol ' $\dot{\neg}$ ' to do the job. (So now ' $\neg$ ' has a double use: 'undotted' and attached to a wff, it is a truth-functional operator; 'dotted' and attached to a term, it expresses a corresponding numerical function. The dotting convention will stop us getting confused.) With this neat new notation, we can put  $\overline{Prf}_T(x, y) =_{\text{def}} Prf_T(x, \dot{\neg}y)$ .

*Proof for first half* Suppose  $\gamma$  is any theorem. Then – dropping subscripts for readability – for some  $m$ ,  $\text{Prf}(m, \ulcorner \gamma \urcorner)$ . Since  $\text{Prf}$  captures  $\text{Prf}$ ,  $T \vdash \text{Prf}(m, \ulcorner \gamma \urcorner)$ .

Also, since  $T$  is consistent,  $\neg\gamma$  is unprovable, so for all  $n$ ,  $\text{not-Prf}(n, \ulcorner \gamma \urcorner)$ . Since  $\overline{\text{Prf}}$  captures  $\overline{\text{Prf}}$ , then for each  $n \leq m$  in particular,  $T \vdash \neg\overline{\text{Prf}}(n, \ulcorner \gamma \urcorner)$ . Using the result (O4) of Section 9.4, that shows  $T \vdash (\forall w \leq m) \neg\overline{\text{Prf}}(w, \ulcorner \gamma \urcorner)$ .

Putting these results together,  $T \vdash \text{Prf}(m, \ulcorner \gamma \urcorner) \wedge (\forall w \leq m) \neg\overline{\text{Prf}}(w, \ulcorner \gamma \urcorner)$ . So existentially quantifying,  $T \vdash \text{RProv}(\ulcorner \gamma \urcorner)$ .

But now suppose that  $\gamma$  is indeed a fixed point for  $\neg\text{RProv}(x)$ , i.e.  $T \vdash \gamma \leftrightarrow \neg\text{RProv}(\ulcorner \gamma \urcorner)$ . Then if  $\gamma$  is provable, we'd also have  $T \vdash \neg\text{RProv}(\ulcorner \gamma \urcorner)$ . Contradiction. So a fixed point  $\gamma$  is not provable:  $T \not\vdash \gamma$ .  $\square$

*Proof for second half* Now suppose  $\neg\gamma$  is a theorem, for some  $\gamma$ . Then for some  $m$ ,  $\overline{\text{Prf}}(m, \ulcorner \gamma \urcorner)$ , so  $T \vdash \overline{\text{Prf}}(m, \ulcorner \gamma \urcorner)$ .

Also, since  $T$  is consistent,  $\gamma$  is unprovable, so for all  $n$ ,  $\text{not-Prf}(n, \ulcorner \gamma \urcorner)$ . Hence, by a parallel argument to before,  $T \vdash (\forall v \leq m) \neg\text{Prf}(v, \ulcorner \gamma \urcorner)$ . Elementary manipulation gives  $T \vdash \forall v (\text{Prf}(v, \ulcorner \gamma \urcorner) \rightarrow \neg v \leq m)$ . Now appeal to (O8) of Section 9.4, and that gives  $T \vdash \forall v (\text{Prf}(v, \ulcorner \gamma \urcorner) \rightarrow m \leq v)$ .

Combining these two results, it immediately follows that  $T \vdash \forall v (\text{Prf}(v, \ulcorner \gamma \urcorner) \rightarrow (m \leq v \wedge \overline{\text{Prf}}(m, \ulcorner \gamma \urcorner)))$ . Quantifying,  $T \vdash \forall v (\text{Prf}(v, \ulcorner \gamma \urcorner) \rightarrow (\exists w \leq v) \overline{\text{Prf}}(w, \ulcorner \gamma \urcorner))$ . So given our definition,  $T \vdash \neg\text{RProv}(\ulcorner \gamma \urcorner)$ .

Suppose again that  $\gamma$  is a fixed point for  $\neg\text{RProv}(x)$ , i.e.  $T \vdash \gamma \leftrightarrow \neg\text{RProv}(\ulcorner \gamma \urcorner)$ . Then if  $\neg\gamma$  is provable, we'd also have  $T \vdash \text{RProv}(\ulcorner \gamma \urcorner)$ . Contradiction. So if  $\gamma$  is a fixed point,  $\neg\gamma$  is not provable:  $T \not\vdash \neg\gamma$ .  $\square$

(d) So we now know that any fixed point for  $\neg\text{RProv}_T$  must be formally undecidable in  $T$ . But the Diagonalization Lemma has already told us that there has to be such a fixed point  $R_T$ . Hence  $R_T$  is formally undecidable, assuming no more than  $T$ 's niceness.

Which is almost what we wanted to show. But not quite. For recall our official statement of the Gödel-Rosser Theorem:

**Theorem 19.6** *If  $T$  is a nice theory, then there is an  $L_A$ -sentence  $\varphi$  of Goldbach type such that neither  $T \vdash \varphi$  nor  $T \vdash \neg\varphi$ .*

This says not just that a nice theory  $T$  has an undecidable sentence, but that it has a  $\Pi_1$  undecidable sentence. And how do we show *that*?

This time it isn't enough simply to appeal to the corollary of Theorem 20.4, i.e. to the principle that  $\Pi_1$  predicates have  $\Pi_1$  fixed points. For  $\neg\text{RProv}(x)$  isn't – or at least, isn't obviously –  $\Pi_1$ . So we are going to have to do a bit more work to demonstrate the full-strength theorem:

*Proof* Let's look at the proof of the previous theorem again, and generalize the leading idea.

Suppose, then, that instead of using the two-place predicates  $\text{Prf}$  and  $\overline{\text{Prf}}$  we use any other pair of two-place predicates  $P$  and  $\overline{P}$  which respectively “enumerate” the positive and negative  $T$ -theorems, i.e. satisfy the following conditions:

1. if  $T \vdash \gamma$ , then for some  $m$ ,  $T \vdash P(m, \ulcorner \gamma \urcorner)$ .
2. if  $T \not\vdash \gamma$ , then for all  $n$ ,  $T \vdash \neg P(n, \ulcorner \gamma \urcorner)$ .
3. if  $T \vdash \neg \gamma$ , then for some  $m$ ,  $T \vdash \bar{P}(m, \ulcorner \gamma \urcorner)$ .
4. if  $T \not\vdash \neg \gamma$ , then for all  $n$ ,  $T \vdash \neg \bar{P}(n, \ulcorner \gamma \urcorner)$ .

Now define  $RP_T(x) =_{\text{def}} \exists v(P(v, x) \wedge (\forall w \leq v) \neg \bar{P}(w, x))$ . This gives us another Rosser-style predicate, and the argument will go through *exactly* as before: for a nice theory  $T$ , any fixed point of  $\neg RP_T(x)$  will be undecidable.

So this tells what we need to look for. Suppose we can find predicates  $P$  and  $\bar{P}$  which satisfy our four “enumeration” conditions, but which are  $\Delta_0$  (i.e. lack unbounded quantifiers). Then the corresponding  $RP_T(x)$  will evidently be  $\Sigma_1$ , its negation  $\neg RP_T(x)$  will be  $\Pi_1$ , and it *will* have  $\Pi_1$  undecidable fixed points.

It just remains, then, to find a suitable pair of  $\Delta_0$  predicates  $P$  and  $\bar{P}$ . Well, consider the  $\Sigma_1$  formula  $\text{Prov}_T(x) =_{\text{def}} \exists v \text{Prf}(v, x)$ . That expresses the property of numbering a  $T$ -theorem (see Section 20.1). Since it is  $\Sigma_1$ ,  $\text{Prov}_T(x)$  is logically equivalent to a wff with a bunch of initial existential quantifiers followed by a  $\Delta_0$  wff. And we can now apply the same trick we invoked in proving Theorem 10.1 to get a wff that expresses the same property but which starts with just a *single* existential quantifier, i.e. has the form  $\exists u P(u, x)$  where  $P$  is  $\Delta_0$ .

But note that when  $\gamma$  is a theorem,  $\exists u P(u, \ulcorner \gamma \urcorner)$  is true, so for some  $m$ ,  $P(m, \ulcorner \gamma \urcorner)$  is true. So, being nice and hence  $\Delta_0$  complete,  $T$  proves that last wff. And if  $\gamma$  isn't a theorem,  $\exists u P(u, \ulcorner \gamma \urcorner)$  is false, so for every  $n$ ,  $P(n, \ulcorner \gamma \urcorner)$  is false, so each  $\neg P(n, \ulcorner \gamma \urcorner)$  is true. Being  $\Delta_0$  complete,  $T$  proves all those latter wffs too.

So  $P$  is  $\Delta_0$  and satisfies the “enumerating” conditions (1) and (2). We can similarly construct  $\bar{P}$  from  $\exists v \bar{\text{Prf}}(v, x)$ . So we are done.  $\square$

## 21.4 Capturing provability?

Having backtracked to the Diagonalization Lemma, we have re-established the First Theorem and have now proved the stronger Gödel-Rosser Theorem. Now we strike out to get a range of new theorems.

First, consider again the  $T$ -wff  $\text{Prov}_T(x)$  which *expresses* the property  $\text{Prov}_T$  of being the g.n. of a  $T$ -theorem. The obvious next question to ask is: does this wff also case-by-case *capture* that property?

No, it doesn't, because in fact

**Theorem 21.3** *No open wff in a nice theory  $T$  can capture the numerical property  $\text{Prov}_T$ .*

*Proof* Suppose for reductio that  $\text{Pr}(x)$  abbreviates an open wff – not necessarily identical to  $\text{Prov}_T(x)$  – which captures  $\text{Prov}_T$ . By the Diagonalization Lemma applied to  $\neg \text{Pr}(z)$ , there is some wff  $\gamma$  such that