(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 $\neg R Prov_T(x)$. Then $T \not\vdash \gamma$ and $T \not\vdash \neg \gamma$.

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

Also, since $T$ is consistent, $\neg \gamma$ is unprovable, so for all $n$, not-$Prf(m, \pull\gamma)$. Since $Prf$ captures $Prf$, then for each $n \leq m$ in particular, $T \vdash \neg Prf(n, \pull\gamma)$.

Using the result (O4) of Section 9.4, that shows $T \vdash (\forall w \leq m) \neg Prf(w, \pull\gamma)$.

Putting these results together, $T \vdash Prf(m, \pull\gamma) \land (\forall w \leq m) \neg Prf(w, \pull\gamma)$. So, existentially quantifying, $T \vdash R Prov(\pull\gamma)$.

But now suppose that $\gamma$ is indeed a fixed point for $\neg R Prov(x)$, i.e. $T \vdash \gamma \leftrightarrow \neg R Prov(\pull\gamma)$. Then if $\gamma$ is provable, we’d also have $T \vdash \neg R Prov(\pull\gamma)$. Contradiction. So a fixed point $\gamma$ is not provable: $T \not\vdash \gamma$.

**Proof for second half** Now suppose $\neg \gamma$ is a theorem, for some $\gamma$. Then for some $m$, $Prf(m, \pull\gamma)$, so $T \vdash Prf(m, \pull\gamma)$.

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

Combining these two results, it immediately follows that $T \vdash \forall v(Prf(v, \pull\gamma) \rightarrow (m \leq v \land Prf(m, \pull\gamma)))$. That implies $T \vdash \forall v(Prf(v, \pull\gamma) \rightarrow (\exists w \leq v) Prf(w, \pull\gamma))$. So given our definition, $T \vdash \neg R Prov(\pull\gamma)$.

Suppose again that $\gamma$ is a fixed point for $\neg R Prov(x)$, i.e. $T \vdash \gamma \leftrightarrow \neg R Prov(\pull\gamma)$. Then if $\neg \gamma$ is provable, we’d also have $T \vdash R Prov(\pull\gamma)$. Contradiction. So if $\gamma$ is a fixed point, $\neg \gamma$ is not provable: $T \not\vdash \neg \gamma$.

(d) So we now know that any fixed point for $\neg R Prov_T$ 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, assuming no more than $T$’s niceness, it follows that $T$ is negation-incomplete.

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 undecided sentence, but that it has a $\Pi_1$ undecided 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 $\neg R Prov(x)_T$
isn’t $\Pi_1$ (or at least, not evidently so),\(^5\) so we can’t conclude that its fixed point $R_T$ is $\Pi_1$. Hence we are going to have to do a bit more work to demonstrate the full-strength Gödel-Rosser 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 $Prf$ and $\neg Prf$ we use any other pair of two-place predicates $P$ and $\neg P$ 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 $m$, $T \vdash \neg P(m, \ulcorner \gamma \urcorner)$.
4. if $T \not\vdash \neg \gamma$, then for all $n$, $T \vdash P(n, \ulcorner \gamma \urcorner)$.

Now define $RP_T(x) =_{\text{def}} \exists v(P(v, x) \land (\forall w \leq v) \neg \neg P(w, x))$. 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.

This tells us what we need to look for. Suppose we can find predicates $P$ and $\neg P$ 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 $\Sigma_1$: so its negation $\neg RP_T(x)$ will be $\Pi_1$ and will indeed have $\Pi_1$ undecidable fixed points.

It just remains, then, to find a suitable pair of $\Delta_0$ predicates $P$ and $\neg P$. Well, consider the $\Sigma_1$ formula $Prf_T(x) =_{\text{def}} \exists v Prf(v, x)$. That expresses the property $Prov_T$, i.e. the property of G"odel-numbering a $T$-theorem (see Section 20.1). Since it is $\Sigma_1$, $Prf_T(x)$ 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 $Prov_T$ but which starts with just a *single* existential quantifier, i.e. has the form $\exists v P(u, x)$ 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.

Hence $P$ is $\Delta_0$ and satisfies the “enumerating” conditions (1) and (2). We can similarly construct a $\Delta_0$ wff $\neg P$ from $\exists v \neg P(v, x)$. So we are done. \(\Box\)

*Phew!*

---

\(^5\) *Why? Well, $RProv_T(x)$ is defined as $\exists v (Prf_T(v, x) \land (\forall w \leq v) \neg \neg P(w, x))$, and its component wff $\neg \neg P(w, x)$ is $\Pi_1$. So, after the initial existential quantifier, $RProv_T(x)$ in effect has an unbounded universal quantifier buried inside. Hence $RProv_T(x)$ isn’t strictly $\Sigma_1$: and it isn’t evidently logically equivalent to a strictly $\Sigma_1$ wff either. So it’s negation isn’t evidently $\Pi_1$.*