(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 γ be any fixed point for $\neg\mathsf{RProv}_T(\mathsf{x})$. Then $T \nvdash \gamma$ and $T \nvdash \neg\gamma$.*

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

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

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

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

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

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

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

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

(d) So we now know that any fixed point for $\neg\mathsf{RProv}_T$ 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, 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 φ 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 Π_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 Π_1 predicates have Π_1 fixed points. For $\neg\mathsf{RProv}(\mathsf{x})_T$

isn't Π_1 (or at least, not evidently so),[5] so we can't conclude that its fixed point R_T is Π_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 $\overline{\mathsf{Prf}}$ we use any other pair of two-place predicates P and $\overline{\mathsf{P}}$ which respectively "enumerate" the positive and negative T-theorems, i.e. satisfy the following conditions:

1. if $T \vdash \gamma$, then for some m, $T \vdash \mathsf{P}(\overline{\mathsf{m}}, \ulcorner\gamma\urcorner)$.
2. if $T \nvdash \gamma$, then for all n, $T \vdash \neg\mathsf{P}(\overline{\mathsf{n}}, \ulcorner\gamma\urcorner)$.
3. if $T \vdash \neg\gamma$, then for some m, $T \vdash \overline{\mathsf{P}}(\overline{\mathsf{m}}, \ulcorner\gamma\urcorner)$.
4. if $T \nvdash \neg\gamma$, then for all n, $T \vdash \neg\overline{\mathsf{P}}(\overline{\mathsf{n}}, \ulcorner\gamma\urcorner)$.

Now define $\mathsf{RP}_T(\mathsf{x}) =_{\text{def}} \exists\mathsf{v}(\mathsf{P}(\mathsf{v}, \mathsf{x}) \wedge (\forall\mathsf{w} \leq \mathsf{v})\neg\overline{\mathsf{P}}(\mathsf{w}, \mathsf{x}))$. 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 $\neg\mathsf{RP}_T(\mathsf{x})$ will be undecidable.

This tells what we need to look for. Suppose we can find predicates P and $\overline{\mathsf{P}}$ which satisfy our four "enumeration" conditions, but which are Δ_0 (i.e. lack unbounded quantifiers). Then the corresponding $\mathsf{RP}_T(\mathsf{x})$ will evidently be Σ_1: so its negation $\neg\mathsf{RP}_T(\mathsf{x})$ *will* be Π_1 and will indeed have Π_1 undecidable fixed points.

It just remains, then, to find a suitable pair of Δ_0 predicates P and $\overline{\mathsf{P}}$. Well, consider the Σ_1 formula $\mathsf{Prov}_T(\mathsf{x}) =_{\text{def}} \exists\mathsf{v}\mathsf{Prf}(\mathsf{v}, \mathsf{x})$. That expresses the property $Prov_T$, i.e. the property of Gödel-numbering a T-theorem (see Section 20.1). Since it is Σ_1, $\mathsf{Prov}_T(\mathsf{x})$ is logically equivalent to a wff with a bunch of initial existential quantifiers followed by a Δ_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 $\exists\mathsf{u}\mathsf{P}(\mathsf{u}, \mathsf{x})$ where P is Δ_0.

But note that when γ is a theorem, $\exists\mathsf{u}\mathsf{P}(\mathsf{u}, \ulcorner\gamma\urcorner)$ is true, so for some m, $\mathsf{P}(\overline{\mathsf{m}}, \ulcorner\gamma\urcorner)$ is true. So, being nice and hence Δ_0-complete, T proves that last wff. And if γ isn't a theorem, $\exists\mathsf{u}\mathsf{P}(\mathsf{u}, \ulcorner\gamma\urcorner)$ is false, so for every n, $\mathsf{P}(\overline{\mathsf{n}}, \ulcorner\gamma\urcorner)$ is false, so each $\neg\mathsf{P}(\overline{\mathsf{n}}, \ulcorner\gamma\urcorner)$ is true. Being Δ_0-complete, T proves all those latter wffs too.

Hence P is Δ_0 and satisfies the "enumerating" conditions (1) and (2). We can similarly construct a Δ_0 wff $\overline{\mathsf{P}}$ from $\exists\mathsf{v}\overline{\mathsf{Prf}}(\mathsf{v}, \mathsf{x})$. So we are done. ⊠

Phew!

[5] Why? Well, $\mathsf{RProv}(\mathsf{x})_T$ is defined as $\exists\mathsf{v}(\mathsf{Prf}_T(\mathsf{v}, \mathsf{x}) \wedge (\forall\mathsf{w} \leq \mathsf{v})\neg\overline{\mathsf{Prf}}_T(\mathsf{w}, \mathsf{x}))$, and its component wff $\neg\overline{\mathsf{Prf}}_T$ is Π_1. So, after the initial existential quantifier, $\mathsf{RProv}(\mathsf{x})_T$ in effect has an unbounded *universal* quantifier buried inside. Hence $\mathsf{RProv}(\mathsf{x})_T$ isn't strictly Σ_1: and it isn't evidently logically equivalent to a strictly Σ_1 wff either. So it's negation isn't evidently Π_1.