On Rosser sentences and proof predicates

Rasmus Blanck

Department of Philosophy
University of Göteborg
2006

On Rosser sentences and proof predicates[*]

Rasmus Blanck

25th August 2006

Abstract

It is a well known fact that the Gödel sentences γ of a theory T are all provably equivalent to the consistency statement of T, Con_T. This result is independent from choice of proof predicate. It has been proved by Guaspari and Solovay [4] that this is not the case for Rosser sentences of T. There are proof predicates whose Rosser sentences are all provably equivalent and also proof predicates whose Rosser sentences are not all provably equivalent. This paper is an attempt to investigate the matter and explicitly define proof predicates of both kinds.

1 Background

We suppose the reader is familiar with the standard logical notation. Some acquaintance with Gödels incompleteness results might be useful as well. *PA* is Peano arithmetic, formulated in your favourite first order logic. Every theory T is assumed to be a sufficiently strong, consistent extension of some fragment of *PA*. We use $\varphi, \psi, \chi, \ldots$ for formulas and $\overline{\varphi}$ for the term denoting the Gödel number of φ. If $\varphi(x)$ is a formula with one free variable, $\overline{\varphi(\dot{x})}$ is the term denoting the Gödel number of $\varphi(x)$ with *x still free in* φ.

A proof predicate $Prf(x, y)$ is a binumeration of the relation "y is a proof of x in T", and a provability predicate $Pr(x)$ is defined as the formula $\exists y Prf(x, y)$, which is an enumeration of the theory of T in T.

$Th(T)$ is the set of theorems of T, i.e. the set of all sentences provable from T.

Con_T is the consistency statement of T, stating that the theory T does not prove any contradictions, i.e. $0 = \overline{1}$.

A theory T is *ω-consistent* iff for every formula $\varphi(x)$, if

$$T \vdash \neg\varphi(k), \text{ for every } k,$$

then

$$T \nvdash \exists x \varphi(x).$$

[*] This thesis was submitted for the degree of Master of Arts at the Department of Philosophy, University of Göteborg. Supervisor: Christian Bennet.

In 1931, Kurt Gödel [5] proved the existence of a true arithmetic sentence that is neither provable nor — if the theory in question is ω-consistent — refutable. The technique applied to construct such a sentence was a general one, using fixed points — a method we will see a few examples of. Given a formula $\xi(x)$ and a theory T, φ is a fixed point of $\xi(x)$ in T if $T \vdash \varphi \leftrightarrow \xi(\overline{\varphi})$.

The construction Gödel used was

$$T \vdash \gamma \leftrightarrow \neg Pr(\overline{\gamma}).$$

By this, γ asserts its own unprovability, is evidently not provable in T, but true in the standard interpretation.

By accident, the theory Gödel used — a mathematical framework presented in Russel & Whiteheads *Principia Mathematica* — is ω-consistent, so γ is neither provable nor refutable in T.

However, ω-consistency is a somewhat artificial property, and in 1936, J. Barkley Rosser [10] presented a method to construct fixed points possessing the desired properties, presupposing only consistency. The sentence used the fixed point

$$T \vdash \rho \leftrightarrow \forall y(Prf(\overline{\rho}, x) \rightarrow \exists z{\leq}y Prf(\overline{\neg\rho}, z)).$$

Notably, all of these fixed points are true.

Leon Henkin [6] raised the question concerning wether sentences asserting their own provability are provable. Consider any sentence η satisfying

$$T \vdash \eta \leftrightarrow Pr(\overline{\eta}).$$

It is not intuitively clear wether η is true in the the standard interpretation, nor is it evident wether it is provable or refutable in T. By a theorem of M.H. Löb [8], these fixed points are indeed provable.

In his work, Löb applied the fixed point theorem to the formula $Pr(x) \rightarrow \psi$, for some sentence ψ of T, to obtain a sentence λ such that

$$T \vdash \lambda \leftrightarrow (Pr(\overline{\lambda}) \rightarrow \psi).$$

From this follows that if $T \vdash Pr(\overline{\psi}) \rightarrow \psi$ then $T \vdash \psi$, which gives the answer to Henkins question.

Strangely enough, the Gödel sentence γ, the Henkin sentence η and the Löb sentence λ are all explicitely definable. It is provable that:

I) Since γ asserts the unprovability of something, it immediately implies consistency, so $T \vdash \gamma \leftrightarrow Con_T$,

II) η is provable in T, so η is provably equivalent to e.g. $0 = 0$ or *any* provable sentence, and

III) $T \vdash \lambda \leftrightarrow (Pr(\overline{\psi}) \rightarrow \psi)$.

These observations all follow from a more general result, emerging from the study of modalised provability logic. It is possible to interpret the provability predicate Pr as the necessity operator $\Box$ in some suitable modal logic, and much work on modal fixed points was done in the seventies by C. Bernardi, D. de Jongh and G. Sambin. It was proven independently by the three that modal fixed points are unique, and de Jongh and Sambin also presented proofs for explicit definability of these fixed points. See for example Smoryński [11] for a thorough treatment. The sentences used in Rossers construction, however, are not modally expressible and as such seem to require some other way of investigating the desired properties.

In 1979, D. Guaspari and R. M. Solovay proved that the answer to the question concering uniqueness of the Rosser fixed points depends on the actual choice of proof predicate, which leads us onto the technical part.

2 Preliminaries

Definition 1. We need a symbol for witness comparison.

$$\exists x\varphi(x) \prec \exists x\psi(x) := \exists x(\varphi(x) \wedge \forall y{\leq}x\neg\psi(y))$$

A *Rosser sentence* is a sentence χ for which $\chi \leftrightarrow Pr(\overline{\neg\chi}) \prec Pr(\overline{\chi})$ is provable in T. Note that this is the dual of Rossers original notion — this one suits our purposes better. Additionaly, none of these Rosser sentence of T are true, nor provable or refutable in T.

Remark 1. If

$$T \vdash \varphi \leftrightarrow \neg(Pr(\overline{\varphi}) \prec Pr(\overline{\neg\varphi}))$$

then

$$T \vdash \neg\varphi \leftrightarrow Pr(\overline{\varphi}) \prec Pr(\overline{\neg\varphi}) \tag{1}$$

is not necessarily equivalent to

$$T \vdash \neg\varphi \leftrightarrow Pr(\overline{\neg\neg\varphi}) \prec Pr(\overline{\neg\varphi}) \tag{2}$$

which would make $\neg\varphi$ a Rosser sentence. Considering this situation from a point *inside* the theory T, $\neg\neg\varphi$ might have a shorter proof than φ, in which case (2) will not say the same thing as (1).

Remark 2. Let $Pr^R(\overline{\varphi})$ be the modified proof predicate:

$$\exists y(Prf(\overline{\varphi}, y) \wedge \forall z{<}y\neg Prf(\overline{\neg\varphi}, z)).$$

The Rosser sentence is actually the Gödel sentence for Pr^R:

$$T \vdash \rho \leftrightarrow \neg Pr^R(\overline{\rho}).$$

Definition 2. With a slight abuse of the arithmetical language, we use the dotted negation sign as a function for getting (the numeral of) the Gödel number of a negated formula.

$$\dot{\neg}\overline{\varphi} = \overline{\neg\varphi}$$

Definition 3. We use the dotted minus sign as a function that removes a negation sign from a formula, if possible.

$$\dot{-}\overline{\varphi} = \begin{cases} \overline{\psi} & \text{if } \overline{\varphi} = \overline{\neg\psi} \text{ for some } \psi \\ \overline{\varphi} & \text{otherwise} \end{cases}$$

Definition 4. A provability predicate is *standard* if it satisfies the first two of the following *Löb derivability conditions*:

L1) $T \vdash (Pr(\overline{\varphi \to \psi}) \land Pr(\overline{\varphi})) \to Pr(\overline{\psi})$

L2) $T \vdash Pr(\overline{\varphi}) \to Pr(\overline{Pr(\overline{\varphi})})$

L3) $T \vdash \varphi \Rightarrow T \vdash Pr(\overline{\varphi})$

for all φ, ψ.

Finally, we need a special case of the fixed point theorem.

Theorem 2.1 (Ehrenfeucht & Feferman). *For any Δ_0 formulas $\gamma_0(x, y)$ and $\gamma_1(x, y)$, we can effectively find Δ_0 sentences φ_0 and φ_1 s.t.*

$T \vdash \varphi_0 \leftrightarrow \gamma_0(\overline{\varphi_0}, \overline{\varphi_1})$

$T \vdash \varphi_1 \leftrightarrow \gamma_1(\overline{\varphi_0}, \overline{\varphi_1})$

See for example Lindström [7] for a proof.

3 All Rosser sentences can be equivalent

Theorem 3.1 (Guaspari & Solovay). *There is a standard proof predicate, all of whose Rosser sentences are provably equivalent.*

This theorem was first proven in terms of a recursive function, enumerating the theorems of T and having the desired properties concerning Rosser sentences. Here, however, we actually construct a formula that defines this proof predicate in FOL.

Let $Pr(x)$ be any standard provability predicate. We define some formulas simultaneously, such that PA proves:

I) $\rho(r, y) \leftrightarrow y = \overline{\dot{r} \leftrightarrow Pr'(\neg\dot{r}) \prec Pr'(\dot{r})}$

II) $\pi(r, y) \leftrightarrow \exists x{\le}y(\rho(r, x) \land Prf(x, y)) \land$
$\forall z{<}y \forall u{<}z \neg((\rho(\dot{\neg}r, z) \land Prf(r, u)) \lor (\rho(\dot{-}r, z) \land Prf(r, u)))$

III) $\lambda(x,y) \leftrightarrow \exists z{\leq}y\,\pi(x,z)$

IV) $\beta_0(x,y) \leftrightarrow \mathit{Prf}(x,y) \wedge \lambda(x,y) \wedge \forall z{<}y\neg\exists u{<}y(\mathit{Prf}(u,z) \wedge \lambda(u,z))$

V) $\beta_1(x,y) \leftrightarrow \mathit{Prf}(\dot{\neg}x,z) \wedge \lambda(x,z) \wedge \forall z{<}y\neg\exists u{<}y(\mathit{Prf}(\dot{\neg}u,z) \wedge \lambda(\dot{\neg}u,z))$

VI) $\mathit{Prf}'(x,y) \leftrightarrow$

$$\begin{aligned}
&\big(\neg\lambda(x,y) \wedge \mathit{Prf}(x,y) \wedge \forall y'{<}y\forall x'{<}y\neg\beta_0(x',y') \wedge \\
&\qquad \forall y'{<}y\forall x'{<}y\neg\beta_1(x',y')\big)\vee \\
&\big(\exists y'{<}y\exists x'{<}y\beta_0(x',y') \wedge \forall z{\leq}y'(\lambda(x,y') \wedge \pi(x,z) \leftrightarrow y=y') \wedge \\
&\qquad \forall z{\leq}y'(\lambda(\dot{\neg}x,y') \wedge \pi(\dot{\neg}x,z) \leftrightarrow y=2y') \wedge (\neg\lambda(x,y) \leftrightarrow y=0)\big) \vee \\
&\big(\exists y'{<}y\exists x'{<}y\beta_1(x',y') \wedge \forall z{\leq}y'(\lambda(x,y') \wedge \pi(x,z) \leftrightarrow y=2y') \wedge \\
&\qquad \forall z{\leq}y'(\lambda(\dot{\neg}x,y') \wedge \pi(\dot{\neg}x,z) \leftrightarrow y=y') \wedge (\neg\lambda(x,y) \leftrightarrow y=0)\big)
\end{aligned}$$

Of these new relations, *Prf′* is defined in terms of λ, β and π — β uses λ, which in turn uses π and ρ. Finally, π is defined in terms of ρ only, but ρ depends on *Pr′* and so, *Prf′*. All quantification is bounded, so these relations are primitive recursive, and we apply Theorem 2.1 to makes sure that the simultaneous construction of *Prf’* and ρ goes through.

The intended interpretation of these relations are:

I) $\rho(r,y)$: y is the statement that r is a *Pr′*-Rosser sentence,

II) $\pi(r,y)$: r is put on a certain *Rosser list* at time y whenever $\rho(r,x)$ holds for some $x{\leq}y$, y is a proof of x, and neither the negation of r, nor the sentence with one less negation than r, is on the list,

III) $\lambda(x,y)$: x is on the Rosser list at time y,

IV) $\beta_0(x,y)$: we encountered a proof y of some Rosser sentence x that is on the Rosser list, and this was indeed the first such,

V) $\beta_1(x,y)$: we encountered a proof y of the negation of some Rosser sentence x that is on the Rosser list, and again this is the first of its kind,

VI) $\mathit{Prf}'(x,y)$: y is a proof of x in the new meaning if either of the following holds:

  a) x is not on the Rosser list at time y, y is a *Prf*-proof of x, and we have not yet encountered any x' and y' for which $\beta_i(x',y')$ holds, for $i=0,1$.

  b) $\beta_0(x',y')$ holds for some x' and y', and all proofs of negated Rosser sentences are greater than the proofs of their positive counterparts. Besides, any sentence not on the Rosser list has the trivial proof $y=0$.

  c) $\beta_1(x',y')$ holds for some x' and y', and all proofs of positive Rosser sentences are greater than the proof of their negated counterparts. Again, any other sentence has the proof $y=0$.

Lemma 3.2. *If φ is* Pr*–Rosser, φ is eventually put on the list.*

Proof. Since no Rosser sentence is provable, neither of $\beta_0(x,y)$ or $\beta_1(x,y)$ are true for any x and y, so the only thing that could keep φ off the list is another Rosser sentence ψ such that $\varphi = \neg\psi$ or $\psi = \neg\varphi$. Any of the two cases would contradict the fact that all Rosser sentences are false. □

Lemma 3.3. PA *proves that if $\beta_0(x,y)$ is true for any x and y,* Th(T) *is inconsistent.*

Proof. Suppose $\beta_0(\overline{\varphi},\overline{k})$ for some φ and k. Then φ is provable, and additionally, by construction, $Pr'(\overline{\neg\varphi}) \prec Pr'(\overline{\varphi})$. Either $Prf'(\overline{\neg\varphi},\overline{i})$ for some $i<k$, in which case $Th(T)$ clearly is inconsistent. Otherwise, $Prf'(\overline{\neg\varphi},\overline{i})$ for no $i<k$, which yields
$Pr'(\overline{\varphi}) \prec Pr'(\overline{\neg\varphi})$. Thus the following sentence, call it ψ, is a Pr–theorem.

$$\exists x(Pr'(\dot{\neg}x) \prec Pr'(x) \wedge Pr'(x) \prec Pr'(\dot{\neg}x))$$

But PA proves $\neg\psi$, and since Pr is standard, PA also proves $Pr(\overline{\neg\psi})$ and $Th(T)$ is inconsistent.

□

The same proof applies, *mutatis mutandis*, in the case where $\beta_1(x,y)$ holds for some x and y.

Lemma 3.4. PA $\vdash$ Pr$(x) \leftrightarrow$ Pr$'(x)$.

Proof. If neither of $\beta_0(x,y)$ or $\beta_1(x,y)$ is true for any x and y, then Pr' obviously proves exactly what Pr does. In the other case, $Th(T)$ is inconsistent by the previous lemma, and so is $Th'(T)$, by construction of Prf'. □

Theorem 3.5 (Proof of Theorem 3.1, concluded).

Proof. Let ρ_0 and ρ_1 be Pr–Rosser. Then neither is provable, and at some stage k both are on the Rosser list. By construction,

$$\rho_0 \leftrightarrow Pr'(\overline{\neg\rho_0}) \prec Pr'(\overline{\rho_0}) \leftrightarrow Pr'(\overline{\neg\rho_1}) \prec Pr'(\overline{\rho_1}) \leftrightarrow \rho_1.$$

□

4 Some Rosser sentences are unequivalent

Theorem 4.1. *There is a standard proof predicate, not all of whose Rosser sentences are provably equivalent.*

Proof. Following the results in the preceding section, I) to V) are as before, and we continue by defining two new formulas simultaneously. Again, we use Theorem 2.1 to make sure the simultaneous construction of $\rho(r,y)$ and $Prf''(x,y)$ is admissible. Let $\beta'(x,y)$ and $Prf''(x,y)$ be formulas such that PA proves:

VI) $\beta'(x,y) \leftrightarrow \exists z{<}y\big(\pi(x,z) \wedge \exists z'{<}y(z \neq z' \wedge \exists x'{<}z'\pi(x',z'))\big)$

VII) $\mathit{Prf}''(x,y) \leftrightarrow$

$$\begin{aligned}
&\big(\neg\lambda(x,y) \wedge \mathit{Prf}(x,y) \wedge \forall y'{<}y\forall x'{<}y\neg\beta_0(x',y') \wedge \\
&\quad \forall y'{<}y\forall x'{<}y\neg\beta'_1(x',y')\big) \vee \\
&\big(\exists x'{<}y\exists y'{<}y\beta_0(x',y') \wedge \\
&\qquad (\lambda(x,y') \wedge \beta'(x,y') \leftrightarrow y = 1 \wedge \mathit{Prf}''(\dot{\neg}x,2)) \wedge \\
&\qquad (\lambda(x,y') \wedge \neg\beta'(x,y') \leftrightarrow y = 2y'+1) \wedge \\
&\qquad (\lambda(\dot{\neg}x,y') \wedge \beta'(x,y') \leftrightarrow y = 2 \wedge \mathit{Prf}''(x,1)) \wedge \\
&\qquad (\lambda(\dot{\neg}x,y') \wedge \neg\beta'(x,y') \leftrightarrow y = 2y') \wedge \\
&\qquad (\neg\lambda(x,y') \leftrightarrow y = 0)\big) \vee \\
&\big(\exists x'{<}y\exists y'{<}y\beta_1(x',y') \wedge \\
&\qquad (\lambda(x,y') \wedge \beta'(x,y') \leftrightarrow y = 2 \wedge \mathit{Prf}''(\dot{\neg}x,1)) \wedge \\
&\qquad (\lambda(x,y') \wedge \neg\beta'(x,y') \leftrightarrow y = 2y') \wedge \\
&\qquad (\lambda(\dot{\neg}x,y') \wedge \beta'(x,y') \leftrightarrow y = 1 \wedge \mathit{Prf}''(x,2)) \wedge \\
&\qquad (\lambda(\dot{\neg}x,y') \wedge \neg\beta'(x,y') \leftrightarrow y = 2y'+1) \wedge \\
&\qquad (\neg\lambda(x,y') \leftrightarrow y = 0)\big)
\end{aligned}$$

$\beta'(x,y)$ states that there are at least two syntactically distinct sentences on the Rosser list. y being a *Prf*$''$-proof of x now means that either of the following conditions a)–c) are satisfied:

a) Neither of $\beta_0(x',y')$ or $\beta_1(x',y')$ holds for any $x'{<}y$ and $y'{<}y$, x is not on the Rosser list at time y, and y is a *Prf*-proof of x.

The following are parts of the second disjunct of the formula:

b1) $\beta_0(x',y')$ holds for some $x'{<}y$ and $y'{<}y$. x is on the Rosser list, and $\beta'(x,y')$ holds. Now $y = 1$ and $\mathit{Prf}''(\dot{\neg}x,2)$. All other sentences has the trivial proof $y = 0$.

b2) $\beta_0(x',y')$ holds for some $x'{<}y$ and $y'{<}y$. x is on the Rosser list, but $\beta'(x,y')$ is false. $y = 2y'+1$, and all other sentences has proof $y = 0$.

b3) $\beta_0(x',y')$ holds for some $x'{<}y$ and $y'{<}y$. $\dot{\neg}x$ is on the Rosser list, and $\beta'(x,y')$ holds. Now $y = 2$ and $\mathit{Prf}''(\dot{\neg}x,1)$. All other sentences has the trivial proof $y = 0$.

b4) $\beta_0(x',y')$ holds for some $x'{<}y$ and $y'{<}y$. $\dot{\neg}x$ is on the Rosser list, but $\beta'(x,y')$ is false. $y = 2y'$, and all other sentences has proof $y = 0$.

c1–c4) The cases for $\beta_1(x',y')$ are similar to $\beta_0(x',y')$ and can be worked out by the interested reader.

Lemmata 3.2 – 3.4 holds for *Prf*$''$ as well. Finally, *Prf*$''$ orders the proofs of Rosser sentences in the following way

$$\rho_0, \neg\rho_0, \neg\rho_1, \rho_1, \neg\rho_2, \rho_2, \ldots,$$

and ρ_0 is not provably equivalent to ρ_1.

□

5 Concluding remarks and questions

Remarks

As pointed out by Smoryński, the derivability conditions L1-L3 together with Löb's theorem seem to tell the whole story of Pr. Indeed, the result on possible non-uniqueness of Rosser sentences is the first requiring more than these conditions, together with "the usual" ordering of proofs, for a settlement.

It is also clear that "the usual" ordering and "the usual" proof predicate is highly arbitrary. A change in the coding of finite sequences is likely to change the order of proofs, as is a transition between different proof systems, and even two different Gödel numberings of formulas.

Standardness of proof predicates does not provide any clues towards a solution — as we have seen there are standard proof predicates with as well equivalent as with non-equivalent Rosser sentences.

The fact that all Rosser sentences are false does not seem to have anything to do with this. By construction, their negations are not provable either, although they are true.

Questions

As none of $\beta_0(x, y)$ and $\beta_1(x, y)$ are ever true for any standard numbers, what Prf' and Prf'' actually does is rearranging the proofs of Rosser sentences in non-standard models to PA. Can this be used to clarify matters?

There seems to be three parts of the concept of Rosser sentences. The Gödel numbering of formulas and sequences, the proof predicate and the fixed point construction. We know that choice of proof predicate does matter, and also that Gödel numbering *should* matter when it comes to ordering proofs. Is it possible that technicalities on the fixed point theorem holds the answer?

In light of this, together with Theorem 3.1 and 4.1 (and maybe their counterparts concering definability) one can ask how interesting the question of equivalence of Rosser sentences really is. The problem concering "the usual" proof predicate, is still open, and as Guaspari & Solovay stated in their 1979 article, the answer seems to be very hard to find.

References

[1] C. Bennet, *Provability predicates and witness comparison*, 1984.

[2] G. Boolos, **The logic of provability**, Cambridge University Press, Cambridge, 1993.

[3] A. Ehrenfeucht & S. Feferman, *Representability of recrusively enumerable sets in formal theories* in **Arch. Math. Logik Grundlagenforsch., vol. 5**, 1959, pp. 37-41.

[4] D. Guaspari & R. M. Solovay, *Rosser sentences* in **Annals of mathematical logic, vol. 16**, North-Holland Publishing Company, Amsterdam, 1979, pp. 81-99.

[5] K. Gödel, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme 1* in **Monatsh. Math. Physik, vol. 38**, 1931, pp. 173-198.

[6] L. Henkin, *A problem concerning provability* in **Journal of Symbolic Logic, vol. 17**, 1952, p. 160.

[7] P. Lindström, **Aspects of incompleteness**, second edition, A. K. Peters, Natick, 2003.

[8] M. H. Löb, *Solution of a problem of Leon Henkin* in **Journal of Symbolic Logic, vol. 20**, 1955, pp. 115-118.

[9] E. Mendelson, **Introduction to Mathematical Logic**, fourth edition, Chapman & Hall/CRC, Boca Raton, 2001.

[10] J. B. Rosser, *Extensions of some theorems of Gödel and Church* in **Journal of Symbolic Logic, vol. 1**, 1936, pp. 87-91.

[11] C. Smoryński, **Self-reference and modal logic**, Springer-Verlag, New York, 1985.