On the conservation results for local reflection principles

Kogure, Haruka

Kurahashi, Taishi

On the conservation results for local reflection principles

Haruka Kogure[*†] and Taishi Kurahashi[‡§]

Abstract

For a class Γ of formulas, Γ local reflection principle $\mathrm{Rfn}_\Gamma(T)$ for a theory T of arithmetic is a scheme formalizing the Γ-soundness of T. Beklemishev [2] proved that for every $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$, the full local reflection principle $\mathrm{Rfn}(T)$ is Γ-conservative over $T + \mathrm{Rfn}_\Gamma(T)$. We firstly generalize the conservation theorem to nonstandard provability predicates: we prove that the second condition **D2** of the derivability conditions is a sufficient condition for the conservation theorem to hold. We secondly investigate the conservation theorem in terms of Rosser provability predicates. We construct Rosser predicates for which the conservation theorem holds and Rosser predicates for which the theorem does not hold.

Keywords: Local reflection principles, Provability predicates, Rosser provability predicates, Conservation theorem.

1 Introduction

Let T be any recursively axiomatized consistent extension of Peano Arithmetic PA. For a class Γ of formulas, the Γ local reflection principle $\mathrm{Rfn}_\Gamma(T)$ for T is the scheme $\{\mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \varphi \mid \varphi \text{ is a } \Gamma \text{ sentence}\}$ which is a formalization of the Γ-soundness of T. Here, $\mathrm{Prov}_T(x)$ is a canonical provability predicate of T. Local reflection principles have been extensively studied by many authors (cf. [3, 11, 17]). Among other things, in the present paper, we focus on the following conservation theorem by Beklemishev:

Theorem (Beklemishev [3, Theorem 1])**.** *For each* $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$*, the full local reflection principle* $\mathrm{Rfn}(T)$ *for* T *is* Γ*-conservative over* $T + \mathrm{Rfn}_\Gamma(T)$*.*

Goryachev [6] studied local reflection principles $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ for Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T. He proved that $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is equivalent to the

*Email: kogure1987@stu.kanazawa-u.ac.jp

†College of Science and Engineering, School of Mathematics and Physics, Kanazawa University, Kakuma, Kanazawa 920-1192, Japan

‡Email: kurahashi@people.kobe-u.ac.jp

§Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan.

usual one $\mathrm{Rfn}(T)$ over T if and only if $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$. Goryachev then provided a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is equivalent to $\mathrm{Rfn}(T)$ over T. Kurahashi [12] continued the work of Goryachev and extensively studied Rosser-type local reflection principles. In particular, the existence of a Rosser provability predicate whose local reflection principle is not equivalent to the usual one was proved. Then, the following problem was proposed:

Problem ([12, Problem 7.1])**.** *Let* $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$*. Is* $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ Γ-*conservative over the theory* $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ *for any Rosser provability predicate* $\mathrm{Pr}_T^{\mathrm{R}}(x)$ *of* T*?*

The present paper studies conservation property with respect to local reflection principles, focusing on Beklemishev's conservation theorem and this problem. Among other things, we generalize Beklemishev's theorem to non-standard provability predicates and provide a counterexample of the above problem.

In Section 3, we firstly prove that for any provability predicate $\mathrm{Pr}_T(x)$ of T, if $\mathrm{Pr}_T(x)$ satisfies the following condition **D2**, then the conservation theorem holds for local reflection principles based on $\mathrm{Pr}_T(x)$:

D2 : $T \vdash \mathrm{Pr}_T(\ulcorner \varphi \to \psi \urcorner) \to (\mathrm{Pr}_T(\ulcorner \varphi \urcorner) \to \mathrm{Pr}_T(\ulcorner \psi \urcorner))$ for any φ, ψ.

While Beklemishev's proof of his conservation theorem used the techniques of the modal logic GL of provability, our proof is simple without any detour to modal logic. In Section 3, we secondly investigate Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$ for which the conservation property holds by distinguishing whether $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ proves Con_T or not.

In Section 4, we prove the existence of Rosser provability predicates lacking the Π_1-conservation property, and this gives counterexamples of the above problem. We prove that for each $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, there exits a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$ but $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}}) \nvdash \mathrm{Con}_T$. Furthermore, we then prove the existence of a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that for any $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is not Π_1-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.

In the last section, we investigate the connection between the Σ_1-conservation property of Rosser provability predicates and Σ_1-soundness. We prove that T is Σ_1-sound if and only if for any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$, there exists $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$ such that $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is Σ_1-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.

2 Preliminaries and background

Throughout this paper, let T denote a recursively axiomatized consistent extension of Peano Arithmetic PA in the language $\mathcal{L}_A$ of first-order arithmetic. Let ω be the set of all natural numbers. For each $n \in \omega$, $\overline{n}$ denotes the numeral for n. For each formula φ, let $\ulcorner \varphi \urcorner$ denote the numeral of the Gödel number of φ.

We inductively define the classes Σ_n and Π_n of $\mathcal{L}_A$-formulas for each $n \geq 0$. Let $\Sigma_0 = \Pi_0$ be the set of all formulas whose every quantifier is bounded. The classes Σ_{n+1} and Π_{n+1} are inductively defined as the smallest classes satisfying the following conditions:

1. $\Sigma_n \cup \Pi_n \subseteq \Sigma_{n+1} \cap \Pi_{n+1}$.
2. Σ_{n+1} (resp. Π_{n+1}) is closed under conjunction, disjunction and existential (resp. universal) quantification.
3. If φ is in Σ_{n+1} (resp. Π_{n+1}), then $\neg\varphi$ is in Π_{n+1} (resp. Σ_{n+1}).
4. If φ is in Σ_{n+1} (resp. Π_{n+1}) and ψ is in Π_{n+1} (resp. Σ_{n+1}), then $\varphi \to \psi$ is in Π_{n+1} (resp. Σ_{n+1}).

If Γ is Σ_n (resp. Π_n), let Γ^d be Π_n (resp. Σ_n). The classes Σ_n and Π_n are primitive recursive, and then they are not closed under taking logically equivalent formulas. A formula φ is said to be $\Delta_1(\mathsf{PA})$ if φ is Σ_1 and is PA-provably equivalent to some Π_1 formula. Let $\mathrm{True}_{\Sigma_1}(x)$ be a Σ_1 formula naturally expressing that "x is a true Σ_1 sentence". We may assume that $\mathrm{True}_{\Sigma_1}(x)$ is of the form $\exists y\, \delta(x, y)$ for some Δ_0 formula $\delta(x, y)$. It is known that such a formula $\mathrm{True}_{\Sigma_1}(x)$ exists and that for any Σ_1 sentence φ, $\mathsf{PA} \vdash \varphi \leftrightarrow \mathrm{True}_{\Sigma_1}(\ulcorner \varphi \urcorner)$ holds (see [8, 9]).

The present paper heavily use the following witness comparison notation (see [7]). For any existential formulas $\exists x\varphi(x)$ and $\exists x\psi(x)$, we introduce two connectives $\prec$ and $\preccurlyeq$ as the following abbreviations:

- $\exists x\varphi(x) \prec \exists x\psi(x) :\equiv \exists x(\varphi(x) \land \forall y \leq x\, \neg\psi(y))$.
- $\exists x\varphi(x) \preccurlyeq \exists x\psi(x) :\equiv \exists x(\varphi(x) \land \forall y < x\, \neg\psi(y))$.

We can apply the witness comparison notation to formulas of the form $\exists x\varphi(x) \lor \exists x\psi(x)$ by considering the formula $\exists x(\varphi(x) \lor \psi(x))$. The following proposition is easily verified.

Proposition 2.1. *For any existential formulas φ and ψ, PA proves the following sentences:*

1. $\varphi \prec \psi \to \varphi \preccurlyeq \psi$.
2. $\neg(\varphi \prec \psi) \lor \neg(\psi \preccurlyeq \varphi)$.
3. $\varphi \lor \psi \to (\varphi \preccurlyeq \psi) \lor (\psi \prec \varphi)$.

2.1 Provability predicates

A $\Delta_1(\mathsf{PA})$ formula $\mathrm{Prf}_T(x, y)$ is said to be a *proof predicate* of T if it satisfies the following conditions:

- For any $\mathcal{L}_A$-formula φ, $T \vdash \varphi$ if and only if $\mathbb{N} \models \exists y\, \mathrm{Prf}_T(\ulcorner \varphi \urcorner, y)$.
- $\mathsf{PA} \vdash \forall y\big(\exists x\, \mathrm{Prf}_T(x, y) \to \exists! x\, \mathrm{Prf}_T(x, y)\big)$.

The second clause says that our proof predicates are single conclusion ones. For a proof predicate $\mathrm{Prf}_T(x, y)$, the Σ_1 formula $\exists y\, \mathrm{Prf}_T(x, y)$ is called a *provability predicate* of T. In his proof of the incompleteness theorems, Gödel constructed a natural proof predicate $\mathrm{Proof}_T(x, y)$ of T saying that "y is a T-proof of x". Let $\mathrm{Prov}_T(x)$ denote the canonical provability predicate $\exists y\, \mathrm{Proof}_T(x, y)$ of T. It is known that $\mathrm{Prov}_T(x)$ satisfies the following Hilbert–Bernays–Löb's derivability conditions and Löb's theorem:

Fact 2.2. *For any $\mathcal{L}_A$-formulas φ and ψ,*

1. $T \vdash \mathrm{Prov}_T(\ulcorner \varphi \to \psi \urcorner) \to \big(\mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \mathrm{Prov}_T(\ulcorner \psi \urcorner)\big)$. **(D2)**
2. $T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \mathrm{Prov}_T(\ulcorner \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \urcorner)$. **(D3)**
3. *If φ is a Σ_1 sentence, then $T \vdash \varphi \to \mathrm{Prov}_T(\ulcorner \varphi \urcorner)$.* (Formalized Σ_1-completeness)
4. *If $T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \varphi$, then $T \vdash \varphi$.* (Löb's theorem)

Let Con_T be the Π_1 sentence $\neg \mathrm{Prov}_T(\ulcorner 0 = 1 \urcorner)$ which expresses the consistency of T. We inductively define the sequence $\{\mathrm{Con}_T^n\}_{n \in \omega}$ of Π_1 sentences as follows:

- $\mathrm{Con}_T^0 :\equiv 0 = 0$;
- $\mathrm{Con}_T^{n+1} :\equiv \neg \mathrm{Prov}_T(\ulcorner \neg \mathrm{Con}_T^n \urcorner)$.

A formula $\mathrm{Pr}_T^{\mathrm{R}}(x)$ is said to be a *Rosser provability predicate* of T if $\mathrm{Pr}_T^{\mathrm{R}}(x)$ is of the form $\mathrm{Pr}_T(x) \prec \mathrm{Pr}_T(\dot{\neg} x)$ for some proof predicate $\mathrm{Prf}_T(x, y)$ of T. Here, $\dot{\neg} x$ is a term corresponding to a primitive recursive function computing the Gödel number of $\neg \varphi$ from that of φ. Rosser provability predicates are essentially introduced by Rosser [16] to improve the first incompleteness theorem. It is known that the second incompleteness theorem does not hold for Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$, that is, $\mathsf{PA} \vdash \neg \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner 0 = 1 \urcorner)$ holds.

2.2 Local reflection principles

For each $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, the Γ *local reflection principle* $\mathrm{Rfn}_\Gamma(T)$ for T is the set $\{\mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \varphi \mid \varphi \text{ is a } \Gamma \text{ sentence}\}$ which expresses the Γ-soundness of T. The *local reflection principle* $\mathrm{Rfn}(T)$ for T is the set $\bigcup_{n \geq 1} \mathrm{Rfn}_{\Sigma_n}(T)$. Similarly, for any provability predicate $\mathrm{Pr}_T(x)$ of T, let $\mathrm{Rfn}_\Gamma(\mathrm{Pr}_T) := \{\mathrm{Pr}_T(\ulcorner \varphi \urcorner) \to \varphi \mid \varphi \text{ is a } \Gamma \text{ sentence}\}$ and $\mathrm{Rfn}(\mathrm{Pr}_T) := \bigcup_{n \geq 1} \mathrm{Rfn}_{\Sigma_n}(\mathrm{Pr}_T)$. Let $\mathcal{B}(\Sigma_n)$ denote the class of all Boolean combinations of Σ_n formulas. Beklemishev proved the following conservation theorem by using the modal logic GL of provability.

Theorem 2.3 (Beklemishev [3, Theorem 1])**.** *For each $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$, the full local reflection principle $\mathrm{Rfn}(T)$ for T is Γ-conservative over $T + \mathrm{Rfn}_\Gamma(T)$. Moreover, $\mathrm{Rfn}(T)$ is $\mathcal{B}(\Sigma_1)$-conservative over $T + \mathrm{Rfn}_{\Sigma_1}(T)$.*

A pioneering work for Rosser-type reflection principle $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ was done by Goryachev [6]. The following proposition is a stratified version of Goryachev's characterization result:

Fact 2.4 ((Essentially) Goryachev [6])**.** *For any $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$ and any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T, the following are equivalent:*

1. *$T + \mathrm{Rfn}_\Gamma(T)$ and $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ are deductively equivalent.*
2. *$T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$.*

In [12], the second author continued the work of Goryachev and extensively studied Rosser-type local reflection principles. The following theorem is a refinement of Goryachev's theorem on the existence of a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T such that $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$.

Fact 2.5 (Kurahashi [12, Theorem 6.8 and Corollary 6.14])**.** *There exists a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T such that $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$ and $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$. Consequently, $T + \mathrm{Rfn}_\Gamma(T)$ and $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ are deductively equivalent for all $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$.*

On the other hand, the following result shows that whether $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ proves Con_T or not is dependent on the choice of $\mathrm{Pr}_T^{\mathrm{R}}(x)$.

Fact 2.6 (Kurahashi [12, Corollary 5.3])**.** *There exists a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T such that $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \nvdash \mathrm{Con}_T$. Consequently, $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ and $T + \mathrm{Rfn}(\mathrm{Pr}_T)$ are not deductively equivalent.*

For $\mathrm{Pr}_T^{\mathrm{R}}(x)$ as in Fact 2.5, by Beklemishev's theorem, $T + \mathrm{Rfn}_\Gamma(T)$ is Γ-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ for all $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$. On the other hand, for $\mathrm{Pr}_T^{\mathrm{R}}(x)$ as in Fact 2.6, it is not clear whether it has the conservation property or not. This situation raised the following problem.

Problem 2.7 ([12, Problem 7.1])**.** *Let $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$. Is $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ Γ-conservative over the theory $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ for any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$?*

3 Provability predicates for which the conservation theorem holds

In this section, we study provability predicates for which Beklemishev's conservation theorem holds. This section consists of three subsections. In the first subsection, we generalize Beklemishev's conservation theorem to non-standard provability predicates. The second subsection is devoted to investigating the conservation property of Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ proves Con_T. In the last subsection, we show the existence of a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ for which the conservation theorem holds but $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ does not prove Con_T.

3.1 A generalization of the conservation theorem

We prove that the conservation theorem generally holds for provability predicates satisfying **D2**. Beklemishev's original proof of the conservation theorem presented in [2] uses the modal logic **GL** of provability, while our proof is simple without detouring modal logic.

For each class Γ of formulas, let $\Gamma(T) := \{\varphi \mid T \vdash \varphi \leftrightarrow \psi \text{ for some } \psi \in \Gamma\}$. If $\mathrm{Pr}_T(x)$ satisfies **D2**, then it is easily shown that the theories $T + \mathrm{Rfn}_{\Gamma(T)}(\mathrm{Pr}_T)$ and $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T)$ are deductively equivalent. We state our theorem in a slightly general form. Basically, we intend $\Theta = \Sigma_1$ and $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$, but cases such as $\Theta = \Sigma_2$ and $\Gamma = \mathcal{B}(\Sigma_2)$ are also in the scope of our theorem.

Theorem 3.1. *Let Θ and Γ be any classes of formulas. Suppose that $\mathrm{Pr}_T(x)$ satisfies* **D2** *and is in $\Theta(T)$. If $\Theta(T) \subseteq \Gamma(T)$ and $\Gamma(T)$ is closed under taking disjunction, then $T + \mathrm{Rfn}(\mathrm{Pr}_T)$ is Γ-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T)$.*

Proof. Suppose that $\mathrm{Pr}_T(x)$ satisfies **D2**, $\mathrm{Pr}_T \in \Theta(T)$, $\Theta(T) \subseteq \Gamma(T)$, and $\Gamma(T)$ is closed under taking disjunction. Let γ be any Γ sentence such that $T + \mathrm{Rfn}(\mathrm{Pr}_T) \vdash \gamma$. We would like to prove $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T) \vdash \gamma$. There exist $k \in \omega$ and $\varphi_0, \ldots, \varphi_{k-1}$ such that

$$T \vdash \bigwedge_{i<k} (\mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \to \varphi_i) \to \gamma. \tag{1}$$

Let $[k] = \{0, 1, \ldots, k-1\}$.

We prove that for any $X \subseteq [k]$, we have

$$T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T) \vdash \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma$$

by induction on the cardinality $|X|$ of X. For $X = \emptyset$, we have $T \vdash \bigwedge_{i<k} \neg \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \to \gamma$ by (1). That is, $T \vdash \bigvee_{i \in [k] \setminus \emptyset} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma$.

Suppose that the statement holds for l and that $|X| = l+1$. For each $j \in X$, we have $|X \setminus \{j\}| = l$, and so by the induction hypothesis,

$$T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T) \vdash \bigvee_{i \in [k] \setminus (X \setminus \{j\})} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma,$$

and hence,

$$T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T) \vdash \mathrm{Pr}_T(\ulcorner \varphi_j \urcorner) \vee \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma.$$

Thus, we obtain

$$T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T) \vdash \bigwedge_{j \in X} \mathrm{Pr}_T(\ulcorner \varphi_j \urcorner) \vee \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma. \tag{2}$$

On the other hand, since $\neg\mathrm{Pr}_T(\ulcorner \varphi_i \urcorner)$ and φ_j respectively imply $\mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \to \varphi_i$ and $\mathrm{Pr}_T(\ulcorner \varphi_j \urcorner) \to \varphi_j$, by (1), we have

$$T \vdash \bigwedge_{i \in [k] \setminus X} \neg\mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \wedge \bigwedge_{j \in X} \varphi_j \to \gamma,$$

and hence,

$$T \vdash \bigwedge_{j \in X} \varphi_j \to \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma.$$

By **D2**, we have

$$T \vdash \bigwedge_{j \in X} \mathrm{Pr}_T(\ulcorner \varphi_j \urcorner) \to \mathrm{Pr}_T\left(\ulcorner \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma \urcorner\right).$$

By combining this with (2),

$$T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T) \vdash \mathrm{Pr}_T\left(\ulcorner \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma \urcorner\right) \vee \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma.$$

By the supposition, $\bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma$ is a $\Gamma(T)$ sentence. So, we have

$$T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T) \vdash \mathrm{Pr}_T\left(\ulcorner \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma \urcorner\right) \to \left(\bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma\right).$$

Therefore, we obtain

$$T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T) \vdash \bigvee_{i \in [k] \setminus X} \mathrm{Pr}_T(\ulcorner \varphi_i \urcorner) \vee \gamma.$$

This shows that the statement holds for X.

At last, we have that the statement holds for $X = [k]$. We then conclude $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T) \vdash \gamma$. □

3.2 Rosser-type reflection principles proving Con_T

We investigate the conservation property of Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$ proves Con_T. Beklemishev's conservation theorem is applicable to such Rosser provability predicates as follows.

Proposition 3.2. *Let $\Gamma \in \{\Sigma_{n+1}, \Pi_{n+1} \mid n \geq 1\}$ and suppose $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$. The following are equivalent:*

1. *$T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is Γ-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.*
2. *$T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ is Π_1-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.*

3. $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$.

Proof. ($1 \Rightarrow 2$): This is trivial because $\Gamma \supseteq \Pi_1$.

($2 \Rightarrow 3$): This is because Con_T is a Π_1 sentence.

($3 \Rightarrow 1$): Let γ be any Γ sentence such that $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \gamma$. Since $T + \mathrm{Rfn}(\mathrm{Prov}_T) \vdash \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \gamma$, we obtain $T + \mathrm{Rfn}_\Gamma(\mathrm{Prov}_T) \vdash \gamma$ by Beklemishev's conservation theorem. Since $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$, we have $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Rfn}_\Gamma(\mathrm{Prov}_T) \vdash \gamma$ by Goryachev's theorem (Fact 2.4). ☐

For $\Gamma = \Sigma_1$, the following proposition is proved in the same way as in the proof of Proposition 3.2.

Proposition 3.3. *Suppose* $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$. *The following are equivalent:*

1. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* $\mathcal{B}(\Sigma_1)$*-conservative over* $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.
2. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* Π_1*-conservative over* $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.
3. $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$.

For $\Gamma = \Pi_1$, we have the following:

Proposition 3.4. *Suppose* $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$. *The following are equivalent:*

1. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* Π_1*-conservative over* $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.
2. $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is inconsistent.*

Proof. ($1 \Rightarrow 2$): Since $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Rfn}(\mathrm{Prov}_T) \vdash \mathrm{Con}^2_T$, we get $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}^2_T$ by the Π_1-conservativity. Then,

$$T + \mathrm{Con}_T \vdash \mathrm{Rfn}_{\Pi_1}(\mathrm{Prov}_T) \vdash \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}^2_T.$$

Hence, we have $T \vdash \mathrm{Prov}_T(\ulcorner \neg \mathrm{Con}_T \urcorner) \to \neg \mathrm{Con}_T$. By Löb's theorem, we have $T \vdash \neg \mathrm{Con}_T$. Therefore, $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$ is inconsistent because Con^2_T implies Con_T.

($2 \Rightarrow 1$): Suppose that $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$ is inconsistent. Then, $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ is also inconsistent and is trivially Π_1-conservative over $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$. ☐

The Σ_1-conservativity of $\mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T)$ is studied in Section 5. The following corollary follows from Propositions 3.2, 3.3, and 3.4 and Fact 2.5.

Corollary 3.5. *There exists a Rosser provability predicate* $\mathrm{Pr}^{\mathrm{R}}_T(x)$ *of* T *satisfying the following two conditions:*

1. *For every* $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$, $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* Γ*-conservative over* $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T)$. *Moreover,* $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* $\mathcal{B}(\Sigma_1)$*-conservative over* $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.
2. *If* $T + \mathrm{Con}_T$ *is consistent, then* $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is not* Π_1*-conservative over* $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.

The paper [12] also investigated a sufficient condition of $\mathrm{Pr}_T^{\mathrm{R}}(x)$ for $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ to prove Con_T. Proposition 5.5 of that paper states that if our logic is formulated so that φ and $\neg\neg\varphi$ are identical for all $\mathcal{L}_A$-formulas φ, then $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$. The proof of this proposition can be divided into two parts: first, if φ and $\neg\neg\varphi$ are identical for all φ, then the following conditions C1 and C2 hold for all Rosser provability predicates $\mathrm{Pr}_T^{\mathrm{R}}(x)$, and second, for any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ satisfying C1 and C2, we have $T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$.

C1 : $T \vdash \neg\mathrm{Con}_T \to \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \lor \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)$ for all $\mathcal{L}_A$-formulas φ.

C2 : $T \vdash \neg\big(\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \land \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)\big)$ for all $\mathcal{L}_A$-formulas φ.

We refine the second statement as follows.

Proposition 3.6. *If* $\mathrm{Pr}_T^{\mathrm{R}}(x)$ *satisfies C1, then* $T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$.

Proof. By the Fixed Point Lemma, we obtain a Π_1 sentence φ satisfying the following equivalence:

$$T \vdash \varphi \leftrightarrow \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \land \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \urcorner). \tag{3}$$

Since $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \varphi$, we have $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner)$ by (3), and hence $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner)$. Thus, we obtain

$$T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \varphi \leftrightarrow \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \urcorner). \tag{4}$$

Since $\neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)$ is also a Π_1 sentence,

$$T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \urcorner) \to \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner). \tag{5}$$

Since $\neg\varphi$ is a Σ_1 sentence, we have $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \to \neg\varphi$, and hence $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \varphi \to \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)$. By combining this with (4), we get

$$T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \urcorner) \to \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner).$$

By combining this with (5), we have

$$T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner).$$

Therefore,

$$T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \varphi \urcorner) \land \neg\mathrm{Pr}_T^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner).$$

By the condition C1, we conclude

$$T + \mathrm{Rfn}_{\Sigma_1 \cup \Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T.$$

□

Corollary 3.7. *If* $\mathrm{Pr}_T^{\mathrm{R}}(x)$ *satisfies C1, then* $\mathrm{Pr}_T^{\mathrm{R}}(x)$ *satisfies the following two properties:*

1. *For every* $\Gamma \in \{\Sigma_{n+1}, \Pi_{n+1} \mid n \geq 1\}$, $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ *is* Γ*-conservative over* $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.

2. *If $T + \mathrm{Con}_T$ is consistent, then $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ is not Π_1-conservative over $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.*

From Theorem 3.1 and Corollary 3.7, we obtained two different sufficient conditions, **D2** and C1, for Rosser provability predicates to satisfy Beklemishev's conservation theorem for $\Gamma \in \{\Sigma_{n+1}, \Pi_{n+1} \mid n \geq 1\}$. Here we consider Rosser provability predicates $\mathrm{Pr}^{\mathrm{R}}_{g'}(x)$ and $\mathrm{Pr}^{\mathrm{R}}_3(x)$, which are proved to exist in [13, Theorem 4.6] and [14, Theorem 11], respectively.

- Theorem 4.6 in [13] states that $\mathrm{Pr}^{\mathrm{R}}_{g'}(x)$ satisfies **D2** and $\mathsf{PA} \vdash \mathrm{Pr}^{\mathrm{R}}_{g'}(\ulcorner \neg \varphi \urcorner) \to \mathrm{Pr}^{\mathrm{R}}_{g'}(\ulcorner \neg \mathrm{Pr}^{\mathrm{R}}_{g'}(\ulcorner \varphi \urcorner) \urcorner)$. Also, Proposition 4.5 in [13] shows that such a Rosser provability predicate does not satisfy C1.
- Theorem 11 in [14] shows that $\mathrm{Pr}^{\mathrm{R}}_3(x)$ satisfies **D3** and **M**: "if $T \vdash \varphi \to \psi$, then $T \vdash \mathrm{Pr}^{\mathrm{R}}_3(\ulcorner \varphi \urcorner) \to \mathrm{Pr}^{\mathrm{R}}_3(\ulcorner \psi \urcorner)$". The proof of the theorem tells us that the predicate $\mathrm{Pr}^{\mathrm{R}}_3(x)$ also satisfies C1, but it follows from the second incompleteness theorem that $\mathrm{Pr}^{\mathrm{R}}_3(x)$ does not satisfy **D2**.

From these facts, we obtain that the conditions **D2** and C1 are generally incomparable with respect to Rosser provability predicates. In the next subsection, we also prove the existence of a Rosser provability predicate which satisfies **D2** but does not satisfy C1.

Recently, the condition C2 has also been studied. It is easy to see that every Rosser provability predicate satisfying **D2** also satisfies C2. It is proved in [14, Theorem 4] that if a provability predicate $\mathrm{Pr}_T(x)$ satisfies **D3** and **M**, then there exists a sentence φ such that $T \nvdash \neg\big(\mathrm{Pr}_T(\ulcorner \varphi \urcorner) \land \mathrm{Pr}_T(\ulcorner \neg \varphi \urcorner)\big)$. So, for example, the predicate $\mathrm{Pr}^{\mathrm{R}}_3(x)$ does not satisfy C2. In [10], the authors studied the modal logical aspect of Rosser provability predicates satisfying **M** and C2.

3.3 Rosser-type reflection principles not proving Con_T

Rosser provability predicates $\mathrm{Pr}^{\mathrm{R}}_T(x)$ in Corollaries 3.5 and 3.7 satisfy $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \vdash \mathrm{Con}_T$, and their conservation properties are based on Beklemishev's conservation theorem for $\mathrm{Prov}_T(x)$. On the other hand, it follows from our Theorem 3.1 that the conservation property also holds for Rosser provability predicates satisfying **D2**. The existence of Rosser provability predicates satisfying **D2** was in fact proved by Bernardi and Montagna [4] and Arai [1]. Here, we prove the existence of a Rosser provability predicate $\mathrm{Pr}^{\mathrm{R}}_T(x)$ satisfying more additional properties: the conservation theorem holds for $\mathrm{Pr}^{\mathrm{R}}_T(x)$ and it satisfies the condition C2, but $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \nvdash \mathrm{Con}_T$. This gives an alternative proof of Fact 2.6. Also, unlike C1, the condition C2 does not contribute to the provability of Con_T in $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$. Moreover, our predicate $\mathrm{Pr}^{\mathrm{R}}_T(x)$ satisfies that $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ is Π_1-conservative over $\mathsf{PA} + \mathrm{Con}_T$. We do not know whether $\mathsf{PA} + \mathrm{Con}_T$ can be replaced by $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_T)$.

Theorem 3.8. *There exists a Rosser provability predicate $\mathrm{Pr}^{\mathrm{R}}_T(x)$ satisfying the following conditions:*

1. *For any* $\Gamma \in \{\Sigma_n, \Pi_{n+1} \mid n \geq 1\}$*, we have that* $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* Γ*-conservative over* $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_T)$*.*

2. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is* Π_1*-conservative over* $\mathsf{PA} + \mathrm{Con}_T$*.*

3. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \nvdash \mathrm{Con}_T$*.*

4. $\mathrm{Pr}^{\mathrm{R}}_T(x)$ *satisfies* C2*.*

Before proving the theorem, we prepare some definitions. An $\mathcal{L}_A$-formula is said to be a *propositionally atomic* if it is either atomic or a quantified formula. Note that every $\mathcal{L}_A$-formula is a Boolean combination of propositionally atomic formulas. For each propositionally atomic formula φ, we prepare a propositional variable p_φ. We define the primitive recursive injection I from the set of all $\mathcal{L}_A$-formulas to propositional formulas as follows:

- $I(\varphi) :\equiv p_\varphi$ for any propositionally atomic φ,
- $I(\neg\varphi) :\equiv \neg I(\varphi)$,
- $I(\varphi \circ \psi) :\equiv= I(\varphi) \circ I(\psi)$ for $\circ \in \{\wedge, \vee, \to\}$.

Let X be a finite set of $\mathcal{L}_A$-formulas. An $\mathcal{L}_A$-formula φ is a *tautological consequence* (*t.c.*) of X if $I(\bigwedge X \to \varphi)$ is a tautology. Note that for any formula φ and finite set X of formulas, whether φ is a t.c. of X is primitive recursively determined. For each $m \in \omega$, let $P_{T,m} := \{\varphi \mid \mathbb{N} \models \exists y \leq \overline{m}\, \mathrm{Proof}_T(\ulcorner \varphi \urcorner, y)\}$. We may assume that PA can prove basic facts about these sets and notions. For example, PA proves "If φ is a t.c. of $P_{T,x}$, then φ is T-provable". The idea of constructing Rosser provability predicates satisfying **D2** using truth assignments of propositional logic is due to Arai [1], and the idea of using t.c. is due to Kurahashi [13].

We are ready to prove Theorem 3.8.

Proof. We define a $\Delta_1(\mathsf{PA})$-definable function $e(x)$ and a sequence $\{k_m\}_{m\in\omega}$ of numbers simultaneously in stages. Let $\mathrm{Pr}_e(x)$ and $\mathrm{Pr}^{\mathrm{R}}_e(x)$ be the formulas $\exists y(x = e(y))$ and $\mathrm{Pr}_e(x) \prec \mathrm{Pr}_e(\dot\neg x)$, respectively. By applying the formalized recursion theorem, we can use these formulas in the definition of e. We would like to prove that the formula $\mathrm{Pr}^{\mathrm{R}}_e(x)$ witnesses the statement of our theorem.

The definition of the function e consists of Procedures 1 and 2. The definition starts with Procedure 1. In Procedure 1, e outputs $\mathcal{L}_A$-formulas in stages referring to T-proofs based on the proof predicate $\mathrm{Proof}_T(x, y)$. A bell is prepared and may ring in this Procedure. After the bell rings, the definition goes to Procedure 2. In Procedure 2, e outputs all formulas. Such a definition of preparing a bell originated in Guaspari and Solovay [7].

We define the the function e as follows: Let $k_0 := 0$.

Procedure 1: The bell has not yet rung.

Stage 1.m: If $0 = 1$ is a t.c. of $P_{T,m}$, then ring the bell and go to Procedure 2.

If $0 = 1$ is not a t.c. of $P_{T,m}$, then we distinguish the following two cases:

- If $P_{T,m} = P_{T,m-1}$, define $k_{m+1} := k_m$ and go to Stage 1.$(m+1)$.
- If $\varphi \in P_{T,m} \setminus P_{T,m-1}$, we distinguish the following two cases:

  (i) : If there exist a number n, a Π_1 sentence π, and distinct formulas $\psi_0, \ldots, \psi_{n-1}$ such that φ is of the form

  $$\bigwedge_{i<n} \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi_i \urcorner) \to \psi_i\big) \to \pi$$

  and there exists a witness $r \leq m$ of the Σ_1 sentence $\mathrm{True}_{\Sigma_1}(\ulcorner \neg\pi \urcorner)$, then ring the bell and go to Procedure 2.

  (ii) : Otherwise, define $e(k_m) := \varphi$ and $k_{m+1} := k_m + 1$. Go to Stage 1.$(m+1)$.

Procedure 2: The bell has rung at Stage 1.m. We define a sequence $\{t_s\}$ of numbers and values $e(k_m), e(k_m+1), e(k_m+2), \ldots$ in stages. Let $\{\xi_s\}$ be the repetition-free sequence of all $\mathcal{L}_A$-formulas in ascending order of Gödel numbers. Define $t_0 := 0$.

Stage 2.s : We distinguish the following three cases:

(i'): If ξ_s is a t.c. of $P_{T,m-1}$, then define $e(k_m + t_s) := \xi_s$ and $t_{s+1} := t_s + 1$. Go to Stage 2.$(s+1)$.

(ii'): If ξ_s is not a t.c. of $P_{T,m-1}$ but $\neg\xi_s$ is a t.c. of $P_{T,m-1}$, then define $e(k_m + t_s) := \neg\xi_s$, $e(k_m + t_s + 1) := \xi_s$, and $t_{s+1} := t_s + 2$. Go to Stage 2.$(s+1)$.

(iii'): If neither ξ_s nor $\neg\xi_s$ is a t.c. of $P_{T,m-1}$, then for every $0 \leq a \leq m+1$, define $e(k_m + t_s + a) := \overbrace{\neg \ldots \neg}^{m+1-a}\xi_s$ and $t_{s+1} := t_s + m + 2$. Go to Stage 2.$(s+1)$.

We have finished the definition of e. Let $\mathrm{Bell}_e(x)$ be an $\mathcal{L}_A$-formula saying "the bell of e rings at Stage 1.x". We prove the properties of the function e in the following claims.

Claim 1. $\mathsf{PA} \vdash \exists x\, \mathrm{Bell}_e(x) \leftrightarrow \neg\mathrm{Con}_T$.

Proof. Argue in PA.

($\to$): Suppose that the bell rings at Stage 1.m. If $0 = 1$ is a t.c. of $P_{T,m}$, then $0 = 1$ is provable in T and T is inconsistent. So, it suffices to consider the case that the bell rings because of Case (i) in Procedure 1.

Suppose that we have numbers $n > 0$ and $r \leq m$, a Π_1 sentence π, and some distinct formulas $\varphi_0, \ldots, \varphi_{n-1}$ such that m is a T-proof of $\bigwedge_{i<n}\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big) \to \pi$ and r witnesses $\mathrm{True}_{\Sigma_1}(\ulcorner \neg\pi \urcorner)$. We prove the following subclaim.

Subclaim. *For each $i < n$, φ_i is a t.c. of $P_{T,m-1}$ or $\mathrm{Pr}_e(\ulcorner \neg\varphi_i \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \varphi_i \urcorner)$ holds.*

Proof. Suppose that φ_i is not a t.c. of $P_{T,m-1}$. We then have that $\varphi_i \notin P_{T,m-1}$ and thus φ_i is not output by e in Procedure 1. We distinguish the following two cases.

Case 1: $\neg\varphi_i$ is a t.c. of $P_{T,m-1}$.
We find s such that $\xi_s \equiv \varphi_i$. We have $e(k_m + t_s) = \neg\varphi_i$ by (ii'). We show that e does not output φ_i before Stage 2.s. It suffices to show that φ_i is not output in Stage 2.s_0 for any $s_0 < s$ by any of the cases (i'), (ii'), and (iii').

- Since φ_i is not a t.c. of $P_{T,m-1}$, we have that e does not output φ_i by (i').
- Let $s_0 < s$ be such that the condition of (ii') is met, then we have that $e(k_m + t_{s_0}) = \neg\xi_{s_0}$ and $e(k_m + t_{s_0} + 1) = \xi_{s_0}$. Since $\neg\xi_{s_0}$ is a t.c. of $P_{T,m-1}$ but φ_i is not, we have that $\varphi_i \not\equiv \neg\xi_{s_0}$. Also, we have $\varphi_i \not\equiv \xi_{s_0}$ because $s_0 < s$. Therefore, φ_i is not output in Stage 2.s_0 by (ii').
- Let $s_0 < s$ be such that $\varphi_i \equiv \overbrace{\neg \ldots \neg}^{c} \xi_{s_0}$ for some $c \leq m + 1$. Since $\neg\varphi_i$ is a t.c. of $P_{T,m-1}$, we have that either ξ_{s_0} or $\neg\xi_{s_0}$ is a t.c. of $P_{T,m-1}$. Thus, the condition of (iii') does not met for s_0. Hence, e does not output φ_i by (iii') at Stage 2.s_0.

We have shown that $\mathrm{Pr}_e(\ulcorner \neg\varphi_i \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \varphi_i \urcorner)$ holds.

Case 2: $\neg\varphi_i$ is not a t.c. of $P_{T,m-1}$.

We find s such that ξ_s is not a negated formula and $\varphi_i \equiv \overbrace{\neg \ldots \neg}^{c} \xi_s$ for some c. Then, for any $p < s$, ξ_p is not φ_i, and moreover φ_i is not obtained by adding negation symbols to ξ_p . Thus, e does not output φ_i before Stage 2.s. Since neither φ_i nor $\neg\varphi_i$ is a t.c. of $P_{T,m-1}$, for every $a \leq m + 1$, $e(k_m + t_s + a) = \overbrace{\neg \ldots \neg}^{m+1-a} \xi_s$ holds by (iii'). Since m is a T-proof of $\bigwedge_{i<n}\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big) \to \pi$, the Gödel number of φ_i is smaller than m, and thus we obtain that $c+1 \leq m+1$. We obtain

$$e(k_m + t_s + m - c) = \overbrace{\neg \ldots \neg}^{c+1} \xi_s = \neg\varphi_i$$

and

$$e(k_m + t_s + m - c + 1) = \overbrace{\neg \ldots \neg}^{c} \xi_s = \varphi_i.$$

It follows that $\mathrm{Pr}_e(\ulcorner \neg\varphi_i \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \varphi_i \urcorner)$ holds. □

If φ_i is a t.c. of $P_{T,m-1}$, then φ_i is provable in T. If φ_i is not a t.c. of $P_{T,m-1}$, then by the subclaim, we have that $\mathrm{Pr}_e(\ulcorner \neg\varphi_i \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \varphi_i \urcorner)$ holds. By formalized Σ_1-completeness, we have that $\mathrm{Pr}_e(\ulcorner \neg\varphi_i \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \varphi_i \urcorner)$ is provable in T. By witness comparison argument, $\neg\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner)$ is also provable in T. Therefore, $\bigwedge_{i<n}\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big)$ is provable in T. Since $\bigwedge_{i<n}\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big) \to \pi$ is T-provable, we obtain that π is also provable in T.

On the other hand, r is a witness of $\mathrm{True}_{\Sigma_1}(\ulcorner \neg\pi \urcorner)$, and thus $\neg\pi$ holds. Since $\neg\pi$ is a Σ_1 sentence, $\neg\pi$ is provable in T by formalized Σ_1-completeness. We conclude that T is inconsistent.

($\leftarrow$): Suppose that T is inconsistent. Let m be such that $0 = 1$ is a t.c. of $P_{T,m}$. Then, it follows that the bell must ring before Stage 1.$(m + 1)$. □

By the following claim, we have that the formulas $x = e(y)$, $\mathrm{Pr}_e(x)$, and $\mathrm{Pr}_e^{\mathrm{R}}(x)$ are a proof predicate of T, a provability predicate of T, and Rosser provability predicate of T, respectively.

Claim 2. $\mathsf{PA} \vdash \forall x\big(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_e(x)\big)$.

Proof. By the definition of e, it is easily shown that

$$\mathsf{PA} + \neg \exists x\, \mathrm{Bell}_e(x) \vdash \forall x(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_e(x)).$$

Since e outputs all $\mathcal{L}_A$-formulas in Procedure 2, we obtain

$$\mathsf{PA} + \exists x\, \mathrm{Bell}_e(x) \vdash \forall x\big(\mathrm{Fml}_{\mathcal{L}_A}(x) \leftrightarrow \mathrm{Pr}_e(x)\big).$$

Also, we have $\mathsf{PA} + \neg \mathrm{Con}_T \vdash \forall x\big(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Fml}_{\mathcal{L}_A}(x)\big)$. By combining these equivalences with Claim 1, we obtain

$$\mathsf{PA} + \exists x\, \mathrm{Bell}_e(x) \vdash \forall x(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_e(x)).$$

By the law of excluded middle, we conclude $\mathsf{PA} \vdash \forall x(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_e(x))$. □

Claim 3. *For any $n \in \omega$, $\mathsf{PA} \vdash \forall x(\mathrm{Bell}_e(x) \to x > \overline{n})$.*

Proof. By Claim 1 and the consistency of T, we have $\mathbb{N} \models \forall x\, \neg \mathrm{Bell}_e(x)$. Thus, $\mathsf{PA} \vdash \neg \mathrm{Bell}_e(\overline{n})$ for any $n \in \omega$. We then obtain $\mathsf{PA} \vdash \forall x(\mathrm{Bell}_e(x) \to x > \overline{n})$. □

The following claim is a key property of the function e. By Theorem 3.1, Clause 1 of the theorem holds for $\mathrm{Pr}_e^{\mathrm{R}}(x)$.

Claim 4. *Let ψ be any $\mathcal{L}_A$-formula. Then,* PA *proves the following statement: "If the bell rings at Stage $1.m$, then*

1. *$P_{T,m-1}$ is propositionally satisfiable,*
2. *if φ is a t.c. of $P_{T,m-1}$, then $\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner)$ holds,*
3. *if ψ is not a t.c. of $P_{T,m-1}$, then $\neg \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)$ holds."*

Proof. We proceed in PA. Assume that the bell rings at Stage $1.m$.

1. Suppose, toward a contradiction, that $P_{T,m-1}$ is not a propositionally satisfiable. Then, $0 = 1$ is a t.c. of $P_{T,m-1}$ and the bell rings at Stage $1.(m-1)$. This is a contradiction.

2. Suppose that φ is a t.c. of $P_{T,m-1}$. We find s such that $\xi_s \equiv \varphi$ in the sequence $\{\xi_s\}$. Then, we have $e(k_m + t_s) = \xi_s$ by (i'). We would like to show that e does not output $\neg\varphi$ before Stage $2.s$. Since $P_{T,m-1}$ is propositionally satisfiable by (1), we have $\neg\varphi \notin P_{T,m-1}$. Thus, e does not output $\neg\varphi$ before Stage $1.m$. Since $\neg\varphi$ is neither ξ_{s_0} nor $\neg\xi_{s_0}$ for all $s_0 < s$, we have that e does not output $\neg\varphi$ by (i') and (ii') before Stage $2.s$. Since φ is a t.c. of $P_{T,m-1}$, e also does not output $\neg\varphi$ by (iii'). Therefore, $\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner)$ holds.

3. Suppose that ψ is not a t.c. of $P_{T,m-1}$. As in the proof of subclaim in Claim 1, we can show that $\mathrm{Pr}_e(\ulcorner \neg\psi \urcorner) \preccurlyeq \mathrm{Pr}_e(\ulcorner \psi \urcorner)$ holds. The only difference

between the proofs is the part to show $c+1 \leq m+1$ in Case 2. In current case, it follows from the standardness of ψ. More precisely, in the case that for some s and c, ξ_s is not a negated formula and $\psi \equiv \overbrace{\neg \ldots \neg}^{c} \xi_s$, since c is a standard number, we obtain $c+1 \leq m+1$ by Claim 3. Therefore, we conclude that $\neg \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)$ holds. $\square$

We show that $\mathrm{Pr}_e^{\mathrm{R}}(x)$ satisfies the condition **D2**.

Claim 5. *For any $\mathcal{L}_A$-formulas φ and ψ,*

$$\mathsf{PA} \vdash \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \to \psi \urcorner) \to \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)\big).$$

Proof. Note that $\mathsf{PA} + \mathrm{Con}_T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \neg \mathrm{Prov}_T(\ulcorner \neg \varphi \urcorner)$. By combining this with Claim 2, $\mathsf{PA} + \mathrm{Con}_T \vdash \mathrm{Pr}_e(\ulcorner \varphi \urcorner) \to \neg \mathrm{Pr}_e(\ulcorner \neg \varphi \urcorner)$. It follows that $\mathsf{PA} + \mathrm{Con}_T \vdash \mathrm{Pr}_e(\ulcorner \varphi \urcorner) \leftrightarrow \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner)$ and hence $\mathsf{PA} + \mathrm{Con}_T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \leftrightarrow \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner)$. Since **D2** holds for $\mathrm{Prov}_T(x)$, we obtain

$$\mathsf{PA} + \mathrm{Con}_T \vdash \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \to \psi \urcorner) \to \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)\big).$$

Then, by Claim 1, it suffices to show

$$\mathsf{PA} + \exists x \, \mathrm{Bell}_e(x) \vdash \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \to \psi \urcorner) \to \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)\big).$$

We argue in $\mathsf{PA} + \exists x \, \mathrm{Bell}_e(x)$. Assume that the bell rings at Stage $1.m$. Suppose $\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \to \psi \urcorner)$ and $\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner)$ hold. Since φ and ψ are standard formulas, by Claim 4, both $\varphi \to \psi$ and φ are t.c.'s of $P_{T,m-1}$. Then, ψ is also a t.c. of $P_{T,m-1}$. By Claim 4 again, we conclude that $\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \psi \urcorner)$ holds. $\square$

Clause 2 of the theorem immediately follows from the following claim.

Claim 6. *For any Π_1 sentence π, if $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \vdash \pi$, then $\mathsf{PA} \vdash \neg \pi \to \neg \mathrm{Con}_T$.*

Proof. Suppose that $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}})$ proves π. Then, for some $n > 0$ and some distinct formulas $\varphi_0, \ldots, \varphi_{n-1}$, we have $T \vdash \bigwedge_{i<n} \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big) \to \pi$. We work in PA. Assume that $\neg \pi$ is true. Then, there exists the least witness r of $\mathrm{True}_{\Sigma_1}(\ulcorner \neg \pi \urcorner)$. Let m be the least T-proof of $\bigwedge_{i<n} \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi_i \urcorner) \to \varphi_i\big) \to \pi$ with $m \geq r$. Then, the bell must ring before Stage $1.(m+1)$. By Claim 1, T is inconsistent. $\square$

We show that Clause 3 of the theorem holds for $\mathrm{Pr}_e^{\mathrm{R}}(x)$.

Claim 7. $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \nvdash \mathrm{Con}_T$.

Proof. Suppose, toward a contradiction, that $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \vdash \mathrm{Con}_T$. Then, $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \vdash \mathrm{Rfn}(\mathrm{Pr}_T)$ holds by Goryachev's theorem and Claim 2. Since $T + \mathrm{Rfn}(\mathrm{Pr}_T) \vdash \mathrm{Con}_T^2$, we obtain $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \vdash \mathrm{Con}_T^2$. Since Con_T^2 is a Π_1 sentence, $\mathsf{PA} \vdash \neg \mathrm{Con}_T^2 \to \neg \mathrm{Con}_T$ by Claim 6. We then obtain $T \vdash \neg \mathrm{Con}_T$ by Löb's theorem. From the supposition, $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}})$ is inconsistent, and hence $T + \mathrm{Rfn}(\mathrm{Pr}_e^{\mathrm{R}}) \vdash 0 = 1$. Since $0 = 1$ is a Π_1 sentence, we obtain $\mathsf{PA} \vdash \neg 0 = 1 \to \neg \mathrm{Con}_T$ by Claim 6 again. Then, $\mathbb{N} \models \neg \mathrm{Con}_T$, a contradiction. $\square$

We finally show that the last clause of the theorem holds for $\mathrm{Pr}_e^{\mathrm{R}}(x)$.

Claim 8. *For any $\mathcal{L}_A$-formula φ,* $\mathsf{PA} \vdash \neg\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \land \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)\big)$.

Proof. Since $\mathsf{PA} \vdash \varphi \to (\neg\varphi \to 0 = 1)$, we have $\mathsf{PA} \vdash \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \to \big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner) \to \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner 0 = 1 \urcorner)\big)$ by Claim 5. Since $\mathsf{PA} \vdash \neg\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner 0 = 1 \urcorner)$, we obtain $\mathsf{PA} \vdash \neg\big(\mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \varphi \urcorner) \land \mathrm{Pr}_e^{\mathrm{R}}(\ulcorner \neg\varphi \urcorner)\big)$. □

This completes the proof of Theorem 3.8. □

4 Rosser provability predicates for which the conservation theorem does not hold

In this section, we study Rosser provability predicates for which the conservation theorem does not hold. That is, we provide a counterexample to Problem 2.7. In fact, we provide Rosser provability predicates having properties stronger than those required as counterexamples, namely, $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_T^{\mathrm{R}})$ is not Π_1-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}})$. We show this in two ways. In the first subsection, we prove that for each $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, there exists a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ such that Con_T witnesses the failure of the Π_1-conservativity of $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_T^{\mathrm{R}})$ over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}})$. In the second subsection, we prove the existence of a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ for which the Π_1-conservation theorem does not hold uniformly, that is, $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_T^{\mathrm{R}})$ is not Π_1-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}})$ for all $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$.

4.1 Rosser provability for which Con_T witnesses the lack of Π_1-conservativity

If $T + \mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ proves Con_T, then as shown in Proposition 3.2, for $\Gamma \in \{\Sigma_{n+1}, \Pi_{n+1} \mid n \geq 1\}$, the Π_1-conservativity of $\mathrm{Rfn}(\mathrm{Pr}_T^{\mathrm{R}})$ over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}})$ is equivalent to the provability of Con_T over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}})$. We show that we are free to control the smallest level of Rosser-type reflection principle that proves Con_T. We fix an effective sequence $\{\alpha_\Gamma\}_{\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}}$ such that each α_Γ is a Γ sentence provable in predicate logic which is not a Γ^d sentence. This sequence will also be used in Subsections 4.2 and 5.1.

Theorem 4.1. *For each $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, there exists a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T such that $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \mathrm{Con}_T$ and $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_T^{\mathrm{R}}) \nvdash \mathrm{Con}_T$.*

Proof. Throughout the proof, we fix $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$. For the fixed Γ, we define a $\Delta_1(\mathsf{PA})$-definable function f by using the formalized recursion theorem as in the proof of Theorem 3.8. The formulas $\mathrm{Pr}_f(x)$ and $\mathrm{Pr}_f^{\mathrm{R}}(x)$ based on f are also defined as in the proof of Theorem 3.8. We can effectively find a Π_1 sentence π and a Σ_1 sentence σ satisfying the following equivalences:

- $\mathsf{PA} \vdash \pi \leftrightarrow \neg \mathrm{Pr}^{\mathrm{R}}_f(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ and
- $\mathsf{PA} \vdash \sigma \leftrightarrow \mathrm{Pr}_f(\ulcorner \neg(\pi \wedge \alpha_{\Pi_n}) \urcorner) \preccurlyeq \mathrm{Pr}_f(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$.

Note that $\mathsf{PA} \vdash \sigma \to \pi$ holds. As in the proof of Theorem 3.8, the definition of f consists of Procedures 1 and 2. Let $k_0 := 0$.

Procedure 1: The bell has not yet rung.
Stage m: If $P_{T,m} = P_{T,m-1}$, then let $k_{m+1} := k_m$ and go to Stage $m+1$.
If $\varphi \in P_{T,m} \setminus P_{T,m-1}$, then depending on whether $\Gamma = \Sigma_n$ or $\Gamma = \Pi_n$, we provide each definition of f as follows.

Case 1: $\Gamma = \Sigma_n$.
We distinguish the following four cases:

(i): If φ is $\pi \wedge \alpha_{\Pi_n}$ or $\neg\neg(\pi \wedge \alpha_{\Pi_n})$, then define $f(k_m) := \pi \wedge \alpha_{\Pi_n}$ and $f(k_m + 1) := \sigma \wedge \alpha_{\Sigma_n}$. Ring the bell and go to Procedure 2.

(ii): If φ is $\neg(\pi \wedge \alpha_{\Pi_n})$ or $\neg(\sigma \wedge \alpha_{\Sigma_n})$, define $f(k_m) := \neg(\pi \wedge \alpha_{\Pi_n})$ and $f(k_m + 1) := \neg(\sigma \wedge \alpha_{\Sigma_n})$. Ring the bell and go to Procedure 2.

(iii): Else if φ is $\neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner \neg\varphi_i \urcorner) \prec \mathrm{Pr}_f(\ulcorner \varphi_i \urcorner)\big)$ for some j and some distinct Γ^d formulas $\varphi_0, \ldots, \varphi_{j-1}$ and f does not output $\varphi_0, \ldots, \varphi_{j-1}$ before stage m, then let $\varphi'_0, \ldots, \varphi'_{j-1}$ be the rearrangement of $\varphi_0, \ldots, \varphi_{j-1}$ in the descending order of length and define $f(k_m) := \neg(\pi \wedge \alpha_{\Pi_n})$ and $f(k_m + 1 + i) := \neg\varphi'_i$ for every $i < j$. Ring the bell and go to Procedure 2.

(iv): Otherwise, define $f(k_m) := \varphi$ and $k_{m+1} := k_m$. Go to Stage $m+1$.

Case 2: $\Gamma = \Pi_n$.
We replace (i) and (iii) of Case 1 by the following (i') and (iii'), respectively.

(i'): If φ is $\pi \wedge \alpha_{\Pi_n}$ or $\neg\neg(\sigma \wedge \alpha_{\Sigma_n})$, then define $f(k_m) := \pi \wedge \alpha_{\Pi_n}$, $f(k_m+1) := \sigma \wedge \alpha_{\Sigma_n}$ and $f(k_m+2) := \neg\neg(\sigma \wedge \alpha_{\Sigma_n})$. Ring the bell and go to Procedure 2.

(iii'): Else if φ is $\neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner \neg\varphi_i \urcorner) \prec \mathrm{Pr}_f(\ulcorner \varphi_i \urcorner)\big)$ for some j and some distinct Γ^d formulas $\varphi_0, \ldots, \varphi_{j-1}$ and f does not output $\varphi_0, \ldots, \varphi_{j-1}$ before stage m, then let $\varphi'_0, \ldots, \varphi'_{j-1}$ be the rearrangement of $\varphi_0, \ldots, \varphi_{j-1}$ in the descending order of length and define $f(k_m) := \pi \wedge \alpha_{\Pi_n}$ and $f(k_m + 1 + i) := \neg\varphi'_i$ for every $i < j$. Ring the bell and go to Procedure 2.

Procedure 2: The function f outputs all $\mathcal{L}_A$-formulas.
We finish the definition of the function f.
Let $\mathrm{Bell}_f(x)$ be an $\mathcal{L}_A$-formula saying "the bell of f rings at Stage x".

Claim 1. $\mathsf{PA} \vdash \exists x\, \mathrm{Bell}_f(x) \leftrightarrow \neg \mathrm{Con}_T$.

Proof. We discuss inside PA. The implication ($\leftarrow$) is easily followed from (i) or (i'), and so we prove the implication ($\rightarrow$). Suppose the bell rings at Stage m. We distinguish the following five cases.

Case 1: m is a T-proof of $\pi \wedge \alpha_{\Pi_n}$ or $\neg\neg(\pi \wedge \alpha_{\Pi_n})$.
By the definition of π, we have that $\neg \mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ is provable in T. Since $f(k_m) = \pi \wedge \alpha_{\Pi_n}$ and f does not output $\neg(\pi \wedge \alpha_{\Pi_n})$ before Stage m, we obtain that $\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ holds. By formalized Σ_1-completeness, this sentence is provable in T. Thus, T is inconsistent.

Case 2: m is a T-proof of $\neg(\pi \wedge \alpha_{\Pi_n})$ or $\neg(\sigma \wedge \alpha_{\Sigma_n})$.
Since α_{Π_n} is provable in predicate logic and σ implies π, we have that T proves $\neg(\sigma \wedge \alpha_{\Sigma_n})$ in both cases. Since $f(k_m) = \neg(\pi \wedge \alpha_{\Pi_n})$ and f does not output $\pi \wedge \alpha_{\Pi_n}$ before Stage m, we obtain that $\mathrm{Pr}_f(\ulcorner \neg(\pi \wedge \alpha_{\Pi_n}) \urcorner) \preccurlyeq \mathrm{Pr}_f(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ holds, that is, σ holds. Since σ is a Σ_1 sentence and α_{Σ_n} is provable in T, $\sigma \wedge \alpha_{\Sigma_n}$ is provable in T. It follows that T is inconsistent.

Case 3: m is a T-proof of $\neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner \neg\varphi_i \urcorner) \prec \mathrm{Pr}_f(\ulcorner \varphi_i \urcorner)\big)$ for some j and some distinct Γ^d formulas $\varphi_0, \ldots, \varphi_{j-1}$ and f does not output $\varphi_0, \ldots, \varphi_{j-1}$ before Stage m.
In this case, T proves $\neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner \neg\varphi_i \urcorner) \prec \mathrm{Pr}_f(\ulcorner \varphi_i \urcorner)\big)$. Let $\varphi'_0, \ldots, \varphi'_{j-1}$ be the rearrangement of $\varphi_0, \ldots, \varphi_{j-1}$ as above, then $f(k_m + 1 + i) = \neg\varphi'_i$ for every $i < j$. Note that if $\Gamma = \Sigma_n$, then $f(k_m) = \neg(\pi \wedge \alpha_{\Pi_n})$ and if $\Gamma = \Pi_n$, then $f(k_m) = \pi \wedge \alpha_{\Pi_n}$. Since α_{Π_n} is not a Σ_n sentence, we have that $f(k_m)$ is not a Γ^d sentence. Hence, $f(k_m)$ is distinct from each of Γ^d formulas $\varphi_0, \ldots, \varphi_{j-1}$. In addition, φ'_i is different from all of $\neg\varphi'_0, \ldots, \neg\varphi'_{i-1}$ for $i < j$ because of the order of the rearrangement. Therefore, we obtain that $\bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner \neg\varphi'_i \urcorner) \prec \mathrm{Pr}_f(\ulcorner \varphi'_i \urcorner)\big)$ holds, and this Σ_1 sentence is provable in T. We have that T is inconsistent.

Case 4: m is a T-proof of $\pi \wedge \alpha_{\Pi_n}$ or $\neg\neg(\sigma \wedge \alpha_{\Sigma_n})$.
We further distinguish the following two cases.

- m is a T-proof of $\pi \wedge \alpha_{\Pi_n}$: We obtain $f(k_m) = \pi \wedge \alpha_{\Pi_n}$. As in Case 1, $\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ and $\neg\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ are provable in T. It concludes that T is inconsistent.

- m is a T-proof of $\neg\neg(\sigma \wedge \alpha_{\Sigma_n})$: Since σ is provable in T and σ implies $\neg\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$, $\neg\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$ is provable in T. As in Case 1, it is shown that T proves $\mathrm{Pr}_f^{\mathrm{R}}(\ulcorner \pi \wedge \alpha_{\Pi_n} \urcorner)$. Thus, T is inconsistent.

$\square$

As in the proof of Theorem 3.8, we obtain the following claim. We then have that $x = f(y)$, $\mathrm{Pr}_f(x)$, and $\mathrm{Pr}_f^{\mathrm{R}}(x)$ are proof predicate, provability predicate, and Rosser provability predicate of T, respectively.

Claim 2. $\mathsf{PA} \vdash \forall x\big(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_f(x)\big)$.

We prove the first statement of the theorem.

Claim 3. $\mathsf{PA} + \mathrm{Rfn}_\Gamma(\mathrm{Pr}^{\mathrm{R}}_f) \vdash \mathrm{Con}_T$.

Proof. We distinguish the following two cases.

Case 1: $\Gamma = \Sigma_n$.
We show that $\mathsf{PA} + \neg\mathrm{Con}_T$ proves

$$\big(\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\pi \land \alpha_{\Pi_n})\urcorner) \land (\pi \land \alpha_{\Pi_n})\big) \lor \big(\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\sigma \land \alpha_{\Sigma_n}\urcorner) \land \neg(\sigma \land \alpha_{\Sigma_n})\big).$$

Then, the claim immediately follows since $\neg(\pi \land \alpha_{\Pi_n})$ and $\sigma \land \alpha_{\Sigma_n}$ are Σ_n.

We argue in $\mathsf{PA} + \neg\mathrm{Con}_T$. By Claim 1, the bell rings at Stage m. We distinguish the following two cases.

Case 1.1 : The bell rings because of (i) at Stage m.
In this case, $f(k_m) = \pi \land \alpha_{\Pi_n}$ and $f(k_m + 1) = \sigma \land \alpha_{\Sigma_n}$ and f does not output $\neg(\pi \land \alpha_{\Pi_n})$ and $\neg(\sigma \land \alpha_{\Sigma_n})$ before Stage m. Since $\pi \land \alpha_{\Pi_n}$ and $\neg(\sigma \land \alpha_{\Sigma_n})$ are distinct, $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi\land\alpha_{\Pi_n}\urcorner)$ and $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\sigma\land\alpha_{\Sigma_n}\urcorner)$ hold. Since $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi\land\alpha_{\Pi_n}\urcorner)$ implies $\neg\sigma$, we obtain $\neg(\sigma \land \alpha_{\Sigma_n})$. Thus, we have that $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\sigma \land \alpha_{\Sigma_n}\urcorner) \land \neg(\sigma \land \alpha_{\Sigma_n})$ holds.

Case 1.2 : The bell rings because of (ii) or (iii) at Stage m.
In this case, $f(k_m) = \neg(\pi \land \alpha_{\Pi_n})$ and f does not output $\neg\neg(\pi \land \alpha_{\Pi_n})$ and $\pi \land \alpha_{\Pi_n}$ before Stage m. We then have that $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\pi \land \alpha_{\Pi_n})\urcorner)$ and σ hold. Since σ implies π, we obtain $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\pi \land \alpha_{\Pi_n})\urcorner) \land (\pi \land \alpha_{\Pi_n})$ holds.

Case 2: $\Gamma = \Pi_n$.
We show that $\mathsf{PA} + \neg\mathrm{Con}_T$ proves

$$\big(\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi \land \alpha_{\Pi_n}\urcorner) \land \neg(\pi \land \alpha_{\Pi_n})\big) \lor \big(\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\sigma \land \alpha_{\Sigma_n})\urcorner) \land (\sigma \land \alpha_{\Sigma_n})\big).$$

We argue in PA. Suppose that the bell rings at Stage m. We distinguish the following two cases.

Case 2.1 : The bell rings because of (i') or (iii') at Stage m.
We have $f(k_m) = \pi \land \alpha_{\Pi_n}$ and f does not output $\neg(\pi \land \alpha_{\Pi_n})$ before Stage m, and so $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi \land \alpha_{\Pi_n}\urcorner)$ holds. Since $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi \land \alpha_{\Pi_n}\urcorner)$ implies $\neg\pi$, we obtain $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\pi \land \alpha_{\Pi_n}\urcorner) \land \neg(\pi \land \alpha_{\Pi_n})$.

Case 2.2 : The bell rings because of (ii) at Stage m.
We have $f(k_m) = \neg(\pi \wedge \alpha_{\Pi_n})$ and $f(k_m + 1) = \neg(\sigma \wedge \alpha_{\Sigma_n})$. Since f does not output $\neg\neg(\sigma \wedge \alpha_{\Sigma_n})$ and $\pi \wedge \alpha_{\Pi_n}$ before Stage m, $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\sigma \wedge \alpha_{\Sigma_n})\urcorner)$ and σ hold. We then obtain that $\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\neg(\sigma \wedge \alpha_{\Sigma_n})\urcorner) \wedge (\sigma \wedge \alpha_{\Sigma_n})$ holds. □

We prove the second statement of the theorem.

Claim 4. $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_f) \nvdash \mathrm{Con}_T$.

Proof. Suppose, towards a contradiction, that $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_f) \vdash \mathrm{Con}_T$. By the second incompleteness theorem, we have $T \nvdash \mathrm{Con}_T$, and thus we find some $j \geq 1$ and some distinct Γ^d formulas $\varphi_0, \ldots, \varphi_{j-1}$ such that

$$T \vdash \bigwedge_{i<j} (\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\varphi_i\urcorner) \to \varphi_i) \to \mathrm{Con}_T. \tag{6}$$

If $T \vdash \varphi_{i_0}$ for some $i_0 < j$, then $T \vdash \mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\varphi_{i_0}\urcorner) \to \varphi_{i_0}$ and hence this is removed from the assumption of (6). So, we may assume $T \nvdash \varphi_i$ for all $i < j$. Note that $\bigwedge_{i<j}\big(\mathrm{Pr}_f(\ulcorner\neg\varphi_i\urcorner) \prec \mathrm{Pr}_f(\ulcorner\varphi_i\urcorner)\big)$ implies $\bigwedge_{i<j} \neg\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\varphi_i\urcorner)$, and $\bigwedge_{i<j} \neg\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\varphi_i\urcorner)$ implies $\bigwedge_{i<j}(\mathrm{Pr}^{\mathrm{R}}_f(\ulcorner\varphi_i\urcorner) \to \varphi_i)$. It follows from (6) that

$$T + \neg\mathrm{Con}_T \vdash \neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner\neg\varphi_i\urcorner) \prec \mathrm{Pr}_f(\ulcorner\varphi_i\urcorner)\big).$$

It is known that $\neg\mathrm{Con}_T$ is Π_1-conservative over T, so we obtain

$$T \vdash \neg \bigwedge_{i<j} \big(\mathrm{Pr}_f(\ulcorner\neg\varphi_i\urcorner) \prec \mathrm{Pr}_f(\ulcorner\varphi_i\urcorner)\big).$$

Since $T \nvdash \varphi_i$ for all $i < j$, the bell must ring at some stage in the standard model $\mathbb{N}$. By Claim 1, T is inconsistent, a contradiction. Therefore, we conclude that $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_f) \nvdash \mathrm{Con}_T$. □

We have proved Theorem 4.1. □

4.2 Rosser predicate for which Π_1-conservation does not hold uniformly

This subsection is devoted to proving the following theorem. Note that the following theorem also gives an alternative proof of Fact 2.6.

Theorem 4.2. *There exists a Rosser provability predicate* $\mathrm{Pr}^{\mathrm{R}}_T(x)$ *of* T *satisfying the following conditions:*

1. $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_T)$ *is not* Π_1*-conservative over* $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}^{\mathrm{R}}_T)$ *for any* $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$,
2. $T + \mathrm{Rfn}(\mathrm{Pr}^{\mathrm{R}}_T) \nvdash \mathrm{Con}_T$.

Proof. We define a $\Delta_1(\mathsf{PA})$-definable function g outputting all theorems of T. Formulas $\mathrm{Pr}_g(x)$ and $\mathrm{Pr}_g^{\mathrm{R}}(x)$ based on g are defined. By using the fixed point lemma, we can effectively find an effective sequence $\{\psi_\Gamma\}_{\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}}$ of Π_1 sentences such that:

- For $\Gamma = \Sigma_1$: ψ_{Σ_1} is $\neg \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \beta \land \alpha_{\Sigma_1} \urcorner)$ where β is a Σ_1 sentence satisfying

$$\mathsf{PA} \vdash \beta \leftrightarrow \mathrm{Pr}_g(\ulcorner \neg(\beta \land \alpha_{\Sigma_1}) \urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner \beta \land \alpha_{\Sigma_1} \urcorner).$$

- For $\Gamma \neq \Sigma_1$: ψ_Γ satisfies

$$\mathsf{PA} \vdash \psi_\Gamma \leftrightarrow \neg \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \psi_\Gamma \land \alpha_\Gamma \urcorner).$$

Note that each ψ_Γ is defined by using g. By using the formalized recursion theorem, we can use ψ_Γ in the definition of g. Here, we define the function g. Let $k_0 := 0$

PROCEDURE 1: The bell has not yet rung.
Stage m: If $P_{T,m} = P_{T,m-1}$, then let $k_{m+1} := k_m$ and go to Stage $m+1$.
If $\varphi \in P_{T,m} \setminus P_{T,m-1}$, then we distinguish the following three cases:

(i): If there exist Γ, j, and distinct Γ^d formulas $\gamma_0, \ldots, \gamma_{j-1}$ such that g does not output these formulas before Stage m and φ is $\bigwedge_{i<j} \bigl(\mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \gamma_i \urcorner) \to \gamma_i\bigr) \to \psi_\Gamma \land \alpha_\Gamma$, then we define $g(k_m)$ depending on whether $\Gamma = \Sigma_1$ or not as follows:

- If $\Gamma = \Sigma_1$, define $g(k_m) := \beta \land \alpha_{\Sigma_1}$.
- If $\Gamma \neq \Sigma_1$, define $g(k_m) := \psi_\Gamma \land \alpha_\Gamma$.

Let $\gamma'_0, \ldots, \gamma'_{j-1}$ be the rearrangement of $\gamma_0, \ldots, \gamma_{j-1}$ in the descending order of length, and define $g(k_m + 1 + i) := \neg \gamma'_i$ for each $i < j$. Then, ring the bell and go to Procedure 2.

(ii): If φ is one of $\beta \land \alpha_{\Sigma_1}$, $\neg(\beta \land \alpha_{\Sigma_1})$, and $\neg(\psi_\Gamma \land \alpha_\Gamma)$ for some $\Gamma \neq \Sigma_1$, then define $g(k_m) := \varphi$. Ring the bell and go to Procedure 2.

(iii): Otherwise, define $g(k_m) := \varphi$ and $k_{m+1} := k_m + 1$. Go to Stage $m+1$.

PROCEDURE 2: The function g outputs all $\mathcal{L}_A$-formulas.
We finish the construction of the function g. Let $\mathrm{Bell}_g(x)$ be an $\mathcal{L}_A$-formula saying "the bell of g rings at Stage x".

Claim 1. $\mathsf{PA} \vdash \exists x\, \mathrm{Bell}_g(x) \leftrightarrow \neg \mathrm{Con}_T$.

Proof. We argue in PA. The implication $(\leftarrow)$ is obvious by considering (ii), and so we prove the implication $(\rightarrow)$. Suppose that the bell rings at Stage m. We distinguish the following five cases.

Case 1: m is a T-proof of $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\gamma_i\urcorner) \to \gamma_i\big) \to \psi_{\Sigma_1} \wedge \alpha_{\Sigma_1}$ for some j and distinct Π_1 formulas $\gamma_0, \ldots, \gamma_{j-1}$ such that g does not output $\gamma_0, \ldots, \gamma_{j-1}$ before Stage m.
Let $\gamma'_0, \ldots, \gamma'_{j-1}$ be the rearrangement of $\gamma_0, \ldots, \gamma_{j-1}$ in the descending order of length. Then, $g(k_m) = \beta \wedge \alpha_{\Sigma_1}$ and $g(k_m + 1 + i) = \neg\gamma'_i$ for every $i < j$. Since $\beta \wedge \alpha_{\Sigma_1}$ is not Π_1 but γ'_i is Π_1, γ'_i is different from $\beta \wedge \alpha_{\Sigma_1}$. Also γ'_i is different from any of $\neg\gamma'_0, \ldots, \neg\gamma'_{i-1}$ because of the order of the rearrangement. Thus, we obtain $\bigwedge_{i<j}\big(\mathrm{Pr}_g(\ulcorner\neg\gamma'_i\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\gamma'_i\urcorner)\big)$. This Σ_1 sentence is also provable in T because of formalized Σ_1-completeness. Then, T proves $\bigwedge_{i<j}\neg\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\gamma'_i\urcorner)$, and hence T also proves $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\gamma_i\urcorner) \to \gamma_i\big)$. Since $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\gamma_i\urcorner) \to \gamma_i\big) \to \psi_{\Sigma_1} \wedge \alpha_{\Sigma_1}$ is T-provable, we obtain that ψ_{Σ_1} is T-provable. That is, $\neg\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ is T-provable.

Since we have $g(k_m) = \beta \wedge \alpha_{\Sigma_1}$ and the bell does not ring before Stage m, we obtain that $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ holds. By formalized Σ_1-completeness theorem, T proves $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$. Thus, T is inconsistent.

Case 2: m is a T-proof of $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\gamma_i\urcorner) \to \gamma_i\big) \to \psi_\Gamma \wedge \alpha_\Gamma$ for some $\Gamma \neq \Sigma_1$, j, and distinct Γ^d formulas $\gamma_0, \ldots, \gamma_{j-1}$.
Since we have $g(k_m) = \psi_\Gamma \wedge \alpha_\Gamma$ and the bell does not ring before Stage m, $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$ holds. By formalized Σ_1 completeness, T proves $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$. As in Case 1, it can be shown that T proves $\neg\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$. Therefore, T is inconsistent.

Case 3: m is a T-proof of $\beta \wedge \alpha_{\Sigma_1}$.
By the choice of β, T proves $\mathrm{Pr}_g(\ulcorner\neg(\beta \wedge \alpha_{\Sigma_1})\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$. Then, $\neg\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ is provable in T. Since g does not output $\neg(\beta \wedge \alpha_{\Sigma_1})$ before Stage m and $g(k_m) = \beta \wedge \alpha_{\Sigma_1}$, we have that $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ holds. Since $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ is a Σ_1 sentence, it is provable in T. Thus, T is inconsistent.

Case 4: m is a T-proof of $\neg(\beta \wedge \alpha_{\Sigma_1})$.
Since α_{Σ_1} is provable, we have that $\neg\beta$ is T-provable. Thus, $\neg\big(\mathrm{Pr}_g(\ulcorner\neg(\beta \wedge \alpha_{\Sigma_1})\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)\big)$ is provable in T. Since $g(k_m) = \neg(\beta \wedge \alpha_{\Sigma_1})$ and g does not output $\beta \wedge \alpha_{\Sigma_1}$ before Stage m, we have that $\mathrm{Pr}_g(\ulcorner\neg(\beta \wedge \alpha_{\Sigma_1})\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$ holds. By Σ_1-completeness, T proves $\mathrm{Pr}_g(\ulcorner\neg(\beta \wedge \alpha_{\Sigma_1})\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\beta \wedge \alpha_{\Sigma_1}\urcorner)$. Thus, T is inconsistent.

Case 5: m is a T-proof of $\neg(\psi_\Gamma \wedge \alpha_\Gamma)$ for some $\Gamma \neq \Sigma_1$.
As in Case 4, we have that $\neg\psi_\Gamma$ and $\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$ are provable in T. Since g does not output $\psi_\Gamma \wedge \alpha_\Gamma$ before Stage m, we obtain that $\mathrm{Pr}_g(\ulcorner\neg(\psi_\Gamma \wedge \alpha_\Gamma)\urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$ holds and it is provable in T. We then obtain that $\neg\mathrm{Pr}^{\mathrm{R}}_g(\ulcorner\psi_\Gamma \wedge \alpha_\Gamma\urcorner)$ is T-provable. Hence, T is inconsistent. □

As in Claim 2 of the proof of Theorem 3.8, we obtain the following claim. Thus, $\mathrm{Pr}^{\mathrm{R}}_g(x)$ is a Rosser provability predicate of T.

Claim 2. $\mathsf{PA} \vdash \forall x\big(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_g(x)\big)$

Since ψ_{Γ^d} is a Π_1 sentence for each $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, the following claim immediately implies that $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_g^{\mathrm{R}})$ is not Π_1-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}})$.

Claim 3. *For any* $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, *we have* $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \psi_{\Gamma^d}$ *and* $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}}) \nvdash \psi_{\Gamma^d}$.

Proof. We distinguish the following two cases.

Case 1 : $\Gamma = \Pi_1$.
Since $\beta \wedge \alpha_{\Sigma_1}$ is a Σ_1 sentence, we have $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \beta \wedge \alpha_{\Sigma_1} \urcorner) \to \beta \wedge \alpha_{\Sigma_1}$. It follows $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \neg \psi_{\Sigma_1} \to \beta$. Since $\mathrm{Pr}_g(\ulcorner \neg(\beta \wedge \alpha_{\Sigma_1}) \urcorner) \preccurlyeq \mathrm{Pr}_g(\ulcorner \beta \wedge \alpha_{\Sigma_1} \urcorner)$ implies $\neg \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \beta \wedge \alpha_{\Sigma_1} \urcorner)$, we obtain $T \vdash \beta \to \psi_{\Sigma_1}$. Thus, we have $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \psi_{\Sigma_1}$.

We shall prove $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_g^{\mathrm{R}}) \nvdash \psi_{\Sigma_1}$. Suppose, towards a contradiction, that $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \psi_{\Sigma_1}$. Since α_{Σ_1} is provable in predicate logic, $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \psi_{\Sigma_1} \wedge \alpha_{\Sigma_1}$. Then, for some j and distinct Π_1 formulas $\gamma_0, \ldots, \gamma_{j-1}$, we have

$$T \vdash \bigwedge_{i<j} \big(\mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \gamma_i \urcorner) \to \gamma_i\big) \to \psi_{\Sigma_1} \wedge \alpha_{\Sigma_1}. \tag{7}$$

As in the proof of Claim 4 of the proof of Theorem 4.1, we may assume that $T \nvdash \gamma_i$ for every $i < j$. Then, the bell must ring in the standard model $\mathbb{N}$ of arithmetic. By Claim 1, we have $\mathbb{N} \models \neg \mathrm{Con}_T$, a contradiction.

Case 2 : $\Gamma \neq \Pi_1$.
Since $\Gamma^d \supseteq \Pi_1$ and $\psi_{\Gamma^d} \wedge \alpha_{\Gamma^d}$ is a Γ^d sentence, we obtain $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \psi_{\Gamma^d} \wedge \alpha_{\Gamma^d} \urcorner) \to \psi_{\Gamma^d} \wedge \alpha_{\Gamma^d}$. Since ψ_{Γ^d} is equivalent to $\neg \mathrm{Pr}_g^{\mathrm{R}}(\ulcorner \psi_{\Gamma^d} \wedge \alpha_{\Gamma^d} \urcorner)$, we obtain $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \psi_{\Gamma^d}$. Also as in Case 1, we can prove $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}}) \nvdash \psi_{\Gamma^d}$. □

Finally, we prove the second clause of the theorem.

Claim 4. $T + \mathrm{Rfn}(\mathrm{Pr}_g^{\mathrm{R}}) \nvdash \mathrm{Con}_T$.

Proof. Suppose, towards a contradiction, that $T + \mathrm{Rfn}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \mathrm{Con}_T$. We then find $\Gamma \supsetneq \Pi_1$ such that $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}}) \vdash \mathrm{Con}_T$. By Proposition 3.2, we have that $T + \mathrm{Rfn}(\mathrm{Pr}_g^{\mathrm{R}})$ is Γ-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}})$. In particular, $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_g^{\mathrm{R}})$ is Π_1-conservative over $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}_g^{\mathrm{R}})$. This contradicts Claim 4. □

This completes our proof of Theorem 4.2. □

5 Σ_1-conservation property and Σ_1-soundness

Kreisel and Lévy [11, Theorem 20] showed that if T is Σ_1-sound, then so is $T + \mathrm{Rfn}(T)$ (see also [12, Lemma 6.1]). Then, for every theory S with $T + \mathrm{Rfn}(T) \vdash S \vdash T$, by Σ_1-completeness, it follows that the Σ_1-soundness of T implies the Σ_1-conservativity of S over T. Smoryński proved the converse implication in the case of $S = T + \mathrm{Con}_T$.

Theorem 5.1 (Smoryński [18, p. 197][19, p. 366])**.** *Con_T is Σ_1-conservative over T if and only if T is Σ_1-sound.*

Recall that we assumed that T always denotes a consistent theory (cf. the beginning of Section 2). The consistency of T is obviously needed to show the right-to-left implication of Theorem 5.1. We can generalize Smoryński's theorem as follows.

Corollary 5.2. *For each $n \in \omega$, the following are equivalent:*

1. *T is Σ_1-sound.*
2. *$T + \mathrm{Con}_T^n$ is consistent and for any S with $T + \mathrm{Rfn}(T) \vdash S \vdash T + \mathrm{Con}_T^n$, S is Σ_1-conservative over $T + \mathrm{Con}_T^n$.*
3. *$T + \mathrm{Con}_T^n$ is consistent and $T + \mathrm{Con}_T^{n+1}$ is Σ_1-conservative over $T + \mathrm{Con}_T^n$.*

Proof. ($1 \Rightarrow 2$) and ($2 \Rightarrow 3$) are straightforward. We prove ($3 \Rightarrow 1$). Suppose that $T + \mathrm{Con}_T^n$ is consistent and $T + \mathrm{Con}_T^{n+1}$ is Σ_1-conservative over $T + \mathrm{Con}_T^n$. Since Con_T^{n+1} is equivalent to $\mathrm{Con}_{T+\mathrm{Con}_T^n}$, by Smoryński's theorem, we have that $T + \mathrm{Con}_T^n$ is Σ_1-sound. Hence, T is Σ_1-sound. ◻

Here, we focus on the second clause of Corollary 5.2. Beklemishev's conservation theorem tells us that $T + \mathrm{Rfn}(T)$ is Σ_1-conservative over $T + \mathrm{Rfn}_{\Sigma_1}(T)$. Hence, the Σ_1-conservativity of a theory S with $T + \mathrm{Rfn}(T) \vdash S \vdash T + \mathrm{Rfn}_{\Sigma_1}(T)$ over $T + \mathrm{Rfn}_{\Sigma_1}(T)$ does not imply the Σ_1-soundness of T. Let $T_\omega := T + \{\mathrm{Con}_T^n \mid n \in \omega\}$, and we propose the following problem.

Problem 5.3. *If T_ω is consistent, does the Σ_1-conservativity of $T + \mathrm{Rfn}(T)$ over T_ω imply the Σ_1-soundness of T?*

Goryachev [5, Theorem 3] proved that $T + \mathrm{Rfn}(T)$ is Π_1-conservative over T_ω. By formalizing this result in PA, in particular, we have $\mathsf{PA} \vdash \mathrm{Con}_{T+\mathrm{Rfn}(T)} \leftrightarrow \mathrm{Con}_{T_\omega}$. Hence, if T_ω is consistent, then $T + \mathrm{Rfn}(T) \nvdash \mathrm{Con}_{T_\omega}$. It follows that Smoryński's theorem cannot be used to solve the problem affirmatively.

5.1 Σ_1-conservation property for Rosser provability predicates

It is also known that there is a relationship between Σ_1-soundness and Σ_1-conservativity of theories having some axioms based on a Rosser provability predicate as well. Let ρ be a Π_1 Rosser sentence of T based on a Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$, then the following theorem due to Švejdar is an improvement of Smoryński's theorem because it is known that $\mathsf{PA} \vdash \mathrm{Con}_T \to \rho$.

Theorem 5.4 (Švejdar (cf. [15, Exercise 5.2(b)])). *ρ is Σ_1-conservative over T if and only if T is Σ_1-sound.*

Since $T + \mathrm{Rfn}(T) \vdash T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}}) \vdash \rho$, it follows from Švejdar's theorem that the Σ_1-soundness of T is equivalent to the Σ_1-conservativity of $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}})$ over T. In [12], it is shown that this equivalence also holds in the case that we replace $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}_T^{\mathrm{R}})$ by $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_T^{\mathrm{R}})$.

Theorem 5.5 (Kurahashi [12, Theorem 6.2]). *For any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T, $\mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}_T^{\mathrm{R}})$ is Σ_1-conservative over T if and only if T is Σ_1-sound.*

As shown in the last section, Beklemishev's theorem does not hold in general for Rosser provability predicates and $\Gamma \supseteq \Pi_1$. In this section, in the case of $\Gamma = \Sigma_1$, we show that the Σ_1-conservation property holds for Rosser provability predicates if and only if T is Σ_1-sound.

Theorem 5.6. *The following are equivalent:*

1. *T is Σ_1-sound.*
2. *For any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T, any $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, and any theory S with $T + \mathrm{Rfn}(T) \vdash S \vdash T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$, S is Σ_1-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.*
3. *For any Rosser provability predicate $\mathrm{Pr}_T^{\mathrm{R}}(x)$ of T, there exists $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$ such that $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}_T^{\mathrm{R}})$ is Σ_1-conservative over $T + \mathrm{Rfn}_\Gamma(\mathrm{Pr}_T^{\mathrm{R}})$.*

Proof. The implications $(1 \Rightarrow 2)$ and $(2 \Rightarrow 3)$ are easy. We prove the contrapositive of $(3 \Rightarrow 1)$. Suppose that T is not Σ_1-sound. Then, there exists a Σ_1 sentence θ such that $T \vdash \theta$ and $\mathbb{N} \not\models \theta$. We may assume that θ is of the form $\exists x \delta(x)$ for some Δ_0 formula $\delta(x)$. We define a $\Delta_1(\mathsf{PA})$-definable function h. Formulas $\mathrm{Pr}_h(x)$ and $\mathrm{Pr}_h^{\mathrm{R}}(x)$ based on h are defined as in the previous sections. We can effectively find an effective sequence $\{\chi_\Gamma\}_{\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}}$ of sentences such that:

- For $\Gamma = \Sigma_1$: χ_{Σ_1} is a Σ_1 sentence satisfying

$$\mathsf{PA} \vdash \chi_{\Sigma_1} \leftrightarrow \big(\mathrm{Pr}_h(\ulcorner \neg(\chi_{\Sigma_1} \wedge \alpha_{\Sigma_1})\urcorner) \vee \theta \prec \mathrm{Pr}_h(\ulcorner \chi_{\Sigma_1} \wedge \alpha_{\Sigma_1} \urcorner)\big).$$

- For $\Gamma \neq \Sigma_1$: χ_Γ is a Π_1 sentence satisfying

$$\mathsf{PA} \vdash \chi_\Gamma \leftrightarrow \neg\big(\mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma)\urcorner) \vee \theta\big).$$

We define the function h as follows: Let $k_0 := 0$.

PROCEDURE 1: The bell has not yet rung.
Stage m: If $P_{T,m} = P_{T,m-1}$, then let $k_{m+1} := k_m$ and go to Stage $m+1$.
If $\varphi \in P_{T,m} \setminus P_{T,m-1}$, then we distinguish the following three cases.

(i): If $\forall x \leq k_m \neg\delta(x)$ holds and φ is $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \gamma_i \urcorner) \to \gamma_i\big) \to \chi_\Gamma \wedge \alpha_\Gamma$ for some Γ, j, and some distinct Γ^d formulas $\gamma_0, \ldots, \gamma_{j-1}$ such that h does not output them before stage m, then define $h(k_m) := \chi_\Gamma \wedge \alpha_\Gamma$ and $h(k_m + 1 + i) := \neg\gamma'_i$ for every $i < j$. Here, $\gamma'_0, \ldots, \gamma'_{j-1}$ is the rearrangement of $\gamma_0, \ldots, \gamma_{j-1}$ in the descending order of length. Ring the bell and go to Procedure 2. When $j = 0$, the above description intends that φ is of the form $\chi_\Gamma \wedge \alpha_\Gamma$.

(ii): If φ is $\neg(\chi_\Gamma \wedge \alpha_\Gamma)$ for some Γ, then define $h(k_m) := \neg(\chi_\Gamma \wedge \alpha_\Gamma)$. Ring the bell and go to Procedure 2.

(iii): Otherwise, define $h(k_m) := \varphi$ and $k_{m+1} := k_m + 1$. Go to Stage $m+1$.

PROCEDURE 2: The function h outputs all formulas.
We finished the definition of h.
Let $\mathrm{Bell}_h(x)$ be an $\mathcal{L}_A$-formula saying "the bell of h rings at Stage x".

Claim 1. $\mathsf{PA} \vdash \exists x\, \mathrm{Bell}_h(x) \leftrightarrow \neg\mathrm{Con}_T$.

Proof. Argue in PA. We only give a proof of the implication $(\rightarrow)$. Suppose that the bell rings at Stage m. We distinguish the following two cases.

Case 1: $\forall x \leq k_m \neg\delta(x)$ holds and m is a T-proof of $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \gamma_i \urcorner) \to \gamma_i\big) \to \chi_\Gamma \wedge \alpha_\Gamma$ for some Γ, j, and distinct Γ^d formulas $\gamma_0, \ldots, \gamma_{j-1}$:
Since $\chi_\Gamma \wedge \alpha_\Gamma$ is not Γ^d but every γ_i for $i < j$ is Γ^d, we have that each of $\gamma_0, \ldots, \gamma_{j-1}$ is different from $\chi_\Gamma \wedge \alpha_\Gamma$. In the same argument as in the proof of Claim 1 of Theorem 4.2, we obtain $\bigwedge_{i<j}\big(\mathrm{Pr}_h(\ulcorner \neg\gamma_i \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \gamma_i \urcorner)\big)$ holds and it is provable in T by formalized Σ_1-completeness. Thus, T proves $\bigwedge_{i<j} \neg\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \gamma_i \urcorner)$, and also proves $\bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \gamma_i \urcorner) \to \gamma_i\big)$. Hence, we get that χ_Γ is T-provable. Then, regardless of whether $\Gamma = \Sigma_1$ or not, we have that

$$\neg\big(\mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma) \urcorner) \vee \theta\big) \tag{8}$$

is T-provable.

Note that $h(k_m) = \chi_\Gamma \wedge \alpha_\Gamma$, h does not output $\neg(\chi_\Gamma \wedge \alpha_\Gamma)$ before Stage m, and $\forall x \leq k_m \neg\delta(x)$ holds. So, we have that

$$\mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma) \urcorner) \vee \theta$$

holds and this Σ_1 sentence is provable in T. By combining this with (8), we obtain that T is inconsistent.

Case 2: m is a T-proof of $\neg(\chi_\Gamma \wedge \alpha_\Gamma)$.
Since α_Γ is T-provable, we have that $\neg\chi_\Gamma$ is T-provable. Since T proves θ, we have that $\neg\chi_{\Sigma_1}$ implies

$$\mathrm{Pr}_h(\ulcorner \chi_{\Sigma_1} \wedge \alpha_{\Sigma_1} \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_{\Sigma_1} \wedge \alpha_{\Sigma_1}) \urcorner) \vee \theta.$$

Therefore, regardless of whether $\Gamma = \Sigma_1$ or not, we obtain that T proves

$$\mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma) \urcorner) \vee \theta,$$

and hence T also proves

$$\mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner) \preccurlyeq \mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma) \urcorner). \tag{9}$$

Since $h(k_m) = \neg(\chi_\Gamma \wedge \alpha_\Gamma)$ and h does not output $\chi_\Gamma \wedge \alpha_\Gamma$ before Stage m, we obtain that

$$\mathrm{Pr}_h(\ulcorner \neg(\chi_\Gamma \wedge \alpha_\Gamma) \urcorner) \prec \mathrm{Pr}_h(\ulcorner \chi_\Gamma \wedge \alpha_\Gamma \urcorner)$$

holds and this is provable in T. By combining this with (9), the inconsistency of T follows. ☐

As in the previous proofs, the following claim holds, and hence $\mathrm{Pr}^{\mathrm{R}}_h(x)$ is a Rosser provability predicate of T.

Claim 2. $\mathsf{PA} \vdash \forall x\big(\mathrm{Prov}_T(x) \leftrightarrow \mathrm{Pr}_h(x)\big)$.

We finally prove that $\mathrm{Pr}^{\mathrm{R}}_h(x)$ uniformly lacks Σ_1-conservation property.

Claim 3. *For any $\Gamma \in \{\Sigma_n, \Pi_n \mid n \geq 1\}$, there exists a Σ_1 sentence σ such that $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \sigma$ and $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}^{\mathrm{R}}_h) \nvdash \sigma$.*

Proof. We distinguish the following two cases.

Case 1: $\Gamma = \Pi_1$.
Let σ be the Σ_1 sentence χ_{Σ_1}. Since $\Gamma^d = \Sigma_1$ and $\chi_{\Sigma_1} \wedge \alpha_{\Sigma_1}$ is a Σ_1 sentence, we obtain $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \chi_{\Sigma_1} \wedge \alpha_{\Sigma_1} \urcorner) \to \chi_{\Sigma_1}$. Since $T \vdash \theta$, it is easily shown that $T \vdash \neg\chi_{\Sigma_1} \to \mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \chi_{\Sigma_1} \wedge \alpha_{\Sigma_1} \urcorner)$. Hence, we obtain $T + \mathrm{Rfn}_{\Sigma_1}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \chi_{\Sigma_1}$.

Suppose, towards a contradiction, that $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \chi_{\Sigma_1}$. Then, $T + \mathrm{Rfn}_{\Pi_1}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \chi_{\Sigma_1} \wedge \alpha_{\Sigma_1}$, and hence for some j and distinct Π_1 sentences $\gamma_0, \ldots, \gamma_{j-1}$, we have $T \vdash \bigwedge_{i<j}\big(\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner \gamma_i \urcorner) \to \gamma_i\big) \to (\chi_{\Sigma_1} \wedge \alpha_{\Sigma_1})$. We may assume that for every $i < j$, γ_i is not provable in T. Since, $\mathbb{N} \not\models \theta$, we have that the bell must ring at some stage in $\mathbb{N}$. By Claim 1, T is inconsistent. A contradiction.

Case 2: $\Gamma \neq \Pi_1$.
Let σ be the Σ_1 sentence

$$\mathrm{Pr}_h(\ulcorner\neg(\chi_{\Gamma^d} \wedge \alpha_{\Gamma^d})\urcorner) \vee \theta \prec \mathrm{Pr}_h(\ulcorner\chi_{\Gamma^d} \wedge \alpha_{\Gamma^d}\urcorner).$$

Since $T \vdash \theta$, we obtain $T \vdash \sigma \leftrightarrow \chi_{\Gamma^d}$.

By the definition of χ_{Γ^d}, we obtain

$$\mathsf{PA} \vdash \neg\mathrm{Pr}^{\mathrm{R}}_h(\ulcorner\chi_{\Gamma^d} \wedge \alpha_{\Gamma^d}\urcorner) \to \chi_{\Gamma^d}.$$

Since $\Gamma^d \supseteq \Pi_1$, we have that $\chi_{\Gamma^d} \wedge \alpha_{\Gamma^d}$ is a Γ^d sentence. We then obtain

$$T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \mathrm{Pr}^{\mathrm{R}}_h(\ulcorner\chi_{\Gamma^d} \wedge \alpha_{\Gamma^d}\urcorner) \to \chi_{\Gamma^d}$$

Thus, we get $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \chi_{\Gamma^d}$, and hence $T + \mathrm{Rfn}_{\Gamma^d}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \sigma$.

We prove $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}^{\mathrm{R}}_h) \nvdash \sigma$. Suppose, towards a contradiction, that $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \sigma$. Then, $T + \mathrm{Rfn}_{\Gamma}(\mathrm{Pr}^{\mathrm{R}}_h) \vdash \chi_{\Gamma^d} \wedge \alpha_{\Gamma^d}$. As in Case 1, this implies the inconsistency of T, a contradiction. □

We have finished the proof of Theorem 5.6. □

Funding

The second author was supported by JSPS KAKENHI (grant numbers JP19K14586 and JP23K03200).

Acknowledgements

The authors would like to thank two anonymous referees for their valuable comments and suggestions.

References

[1] Toshiyasu Arai. Derivability conditions on Rosser's provability predicates. *Notre Dame Journal of Formal Logic*, 31(4):487–497, 1990.

[2] Lev D. Beklemishev. Notes on local reflection principles. *Theoria*, 63(3):139–146, 1997.

[3] Lev D. Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys*, 60(2):197–268, 2005.

[4] Claudio Bernardi and Franco Montagna. Equivalence relations induced by extensional formulae: classification by means of a new fixed point property. *Fundamenta Mathematicae*, 124:221–233, 1984.

[5] S. V. Goryachev. Interpretability of various extensions of arithmetic. *Mathematical Notes*, 40:821–827, 1986.

[6] S. V. Goryachev. Arithmetic with a local reflection principle for Rosser provability formulas. *Mathematical Notes*, 46(3):689–694, 1989.

[7] D. Guaspari and R.M. Solovay. Rosser sentences. *Annals of Mathematical Logic*, 16(1):81–99, 1979.

[8] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic.* Perspectives in Mathematical Logic. Berlin: Springer-Verlag, 1993.

[9] Richard Kaye. *Models of Peano arithmetic*, volume 15 of *Oxford Logic Guides*. Oxford: Clarendon Press, 1991.

[10] Haruka Kogure and Taishi Kurahashi. Arithmetical completeness theorems for monotonic modal logics. *Annals of Pure and Applied Logic*, 174(7):Paper No. 103271, 26, 2023.

[11] Georg Kreisel and Azriel Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 14:97–142, 1968.

[12] Taishi Kurahashi. Henkin sentences and local reflection principles for Rosser provability. *Annals of Pure and Applied Logic*, 167(2):73–94, 2016.

[13] Taishi Kurahashi. Rosser provability and normal modal logics. *Studia Logica*, 108(3):597–617, 2020.

[14] Taishi Kurahashi. Rosser provability and the second incompleteness theorem. In *Advances in mathematical logic*, volume 369 of *Springer Proc. Math. Stat.*, pages 77–97. Springer, Singapore, [2021] ©2021.

[15] Per Lindström. *Aspects of incompleteness.*, volume 10 of *Lect. Notes Log.* Natick, MA: Association for Symbolic Logic, 2nd ed. edition, 2003.

[16] John Barkley Rosser. Extensions of some theorems of Gödel and Church. *Bulletin of the American Mathematical Society*, 42:494, 1936.

[17] Craig Smoryński. The incompleteness theorems. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 821–865. North-Holland, 1977.

[18] Craig Smoryński. Calculating self-referential statements. *Fundamenta Mathematicae*, 109:189–210, 1980.

[19] Craig Smoryński. Fifty years of self-reference in arithmetic. *Notre Dame Journal of Formal Logic*, 22:357–374, 1981.