The Journal of Symbolic Logic
Volume 86, Number 3, September 2021

MÜNCHHAUSEN PROVABILITY

JOOST J. JOOSTEN

Abstract. By Solovay's celebrated completeness result [31] on formal provability we know that the provability logic **GL** describes exactly all provable structural properties for any sound and strong enough arithmetical theory with a decidable axiomatisation. Japaridze generalised this result in [22] by considering a polymodal version GLP of **GL** with modalities $[n]$ for each natural number n referring to ever increasing notions of provability. Modern treatments of GLP tend to interpret the $[n]$ provability notion as "provable in a base theory T together with all true Π^0_n formulas as oracles." In this paper we generalise this interpretation into the transfinite. In order to do so, a main difficulty to overcome is to generalise the syntactical characterisations of the oracle formulas of complexity Π^0_n to the hyper-arithmetical hierarchy. The paper exploits the fact that provability is Σ^0_1 complete and that similar results hold for stronger provability notions. As such, the oracle sentences to define provability at level α will recursively be taken to be consistency statements at lower levels: provability through provability whence the name of the paper. The paper proves soundness and completeness for the proposed interpretation for a wide class of theories, namely for any theory that can formalise the recursion described above and that has some further very natural properties. Some remarks are provided on how the recursion can be formalised into second order arithmetic and on lowering the proof-theoretical strength of these systems of second order arithmetic.

§1. Introduction.

As mentioned in the abstract, by Solovay's celebrated completeness result [31] on provability we know that the provability logic **GL** describes exactly all provable structural properties for any sound and strong enough arithmetical theory with a decidable axiomatisation. Japaridze generalised this result in [22] by considering a polymodal version GLP of **GL** with modalities $[n]$ for each natural number n referring to ever increasing notions of provability.

Japaridze considered an arithmetical interpretation of the logic GLP where the $[n]$ referred to a natural formalisation of "provable over the base theory T using at most n nested applications of the ω-rule." Beklemishev introduced in [6] the logics GLP_Λ that are like GLP only that they now include a sequence of provability predicates $[\alpha]$ of ever increasing strength for each ordinal α below some fixed ordinal Λ. In [18] the authors generalised Japaridze's result into the transfinite by providing an interpretation of GLP_Λ for recursive Λ into second order arithmetic by allowing for $[\alpha]$ at most α nestings of the omega rule, thereby providing a first arithmetical interpretation of GLP_Λ for $\Lambda > \omega$. In a recent paper [9] Beklemishev and Pakhomov provide an alternative interpretation in first order arithmetic enriched with a collection of ever more expressive truth predicates indexed by the ordinals.

Modern treatments of GLP_ω tend to interpret the $[n]$ provability notion as "provable in a base theory T together with all true Π^0_n formulas." Let us call

Received September 4, 2019.

2020 *Mathematics Subject Classification*. 03F03, 03F35, 03F45.

Key words and phrases. provability logics, ordinal analysis, second order arithmetic, transfinite polymodal provability.

0022-4812/21/8603-0007
DOI:10.1017/jsl.2021.44

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

this the *truth-interpretation* here. The main reason for the popularity of the truth-interpretation is that the resulting provability hierarchies run in phase with the arithmetical hierarchy and they imply good preservation properties between different consistency statements giving rise to the so-called *reduction property*. In particular, due to these good properties Beklemishev was able to set GLP_ω to work to perform proof-theoretical analyses of Peano Arithmetic and its kin [3–5]. Below we shall give more circumstantial evidence to why the truth interpretation is optimal.

As mentioned, the first arithmetical interpretation of transfinite polymodal provability logic [18] was, like Japaridze's original approach, based on iterating applications of the omega rule. Although it was observed in [23] that soundness of the interpretation is sufficient for the purpose of an ordinal analysis, the paper also contained a completeness proof in such general lines that it can be applied to a wide range of interpretations.

It seemed, however, that the omega-rule interpretation does not have all the desirable properties to make it directly a useful tool for ordinal analyses. Even though various known fragments of second order arithmetic like ATR_0, $\Pi^1_1 - \mathrm{CA}_0$, and $\Pi^1_1 - \mathrm{CA}_0 + \mathrm{Bar\ Induction}$ can be characterised [10, 12] in terms of reflection principles using versions of the omega rule interpretation of GLP_Λ, the fine-structure between various consistency statements could not be proven.

One possible reason may be that the omega provability predicates do not tie up with the arithmetical hierarchy and Turing jumps as observed in [24, Lemma 9]. A more concrete and serious objection is given in an unpublished simple observation from Fernández Duque: using only one application of the omega-rule one can prove any induction axiom so that the one-consistency of primitive recursive arithmetic in the omega-rule sense suffices to prove the consistency of Peano arithmetic.

In short, the truth interpretation of GLP_ω has better properties than the omega-rule interpretation. However, one advantage of the omega-rule interpretation is its amenability to transfinite generalisations. The formalisation of the truth interpretation relies on a syntactical characterisation of the arithmetical hierarchy in terms of the Σ^0_n formulas. It remained unclear how to generalise this in a canonical way to the hyperarithmetical setting or beyond without extending the language in a way that often seems rather ad-hoc.

The idea of this paper to overcome this is very simple yet turns out to be rather powerful. The *Friedman–Goldfarb–Harrington* theorem (FGH) tells us that for a wide range of theories, in a sense, the canonical consistency predicate is Π^0_1 complete. Thus, instead of using a true Π^0_1 sentence as oracle for the $[1]_T$ provability predicate in the truth interpretation, one can use a provably equivalent consistency statement.

Via a generalisation of the FGH theorem proven in [24, 26] one can see that the consistency notion corresponding to [1] provability is in a sense Π^0_2 complete and so on. Thus, it makes sense to consider the following recursion as in [24]: provability at level n means provable from an oracle which is a consistency statement of level m for some $m < n$. It feels like lifting oneself up from the swamp by pulling ones hairs as the *Baron von Münchhausen* did. Moreover, the recursion lends itself to an easy transfinite generalisation and that is exactly what this paper does. Before we close the introduction with an overview of how the current paper does so, we would like to point out how this paper fits in the landscape of related literature thereby trying to provide an ample justification for it.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

Ordinal analysis via polymodal provability logics seems to have various benefits over other methods of ordinal analysis. An important benefit is it allows to tell different incomplete theories apart at the lowest possible level of Π^0_1 sentences. It is good to recall that the classical Π^1_1 proof theoretical ordinal will not even discern theories at the level of Σ^1_1-level. Another benefit may seem the modularity of ordinal analysis: the ordinal analysis of different theories will all share the same template and re-use various tools and theorems.

We see another stronghold in the fact that the approach relates various different fields in a natural way. In particular, the closed formulas of GLP—called *worms*—are important in this. Worms can be used to denote various notions central to foundational issues. For one, they are simple and well-behaved elements from a well-behaved logic. Even though the logic GLP is known to be PSPACE-complete [29] it is Kripke incomplete. However, natural topological semantics do exist [1, 8, 11, 20, 21] even though it is known to depend on strong cardinal assumptions for various natural topological spaces [2].

Moreover, the closed fragment of GLP_Λ is very well behaved, well studied, and in particular does allow for natural relational semantics [15, 16, 21]. In addition, and this provides a second interpretation of worms, the worms are known to define a well-ordered relation as studied in [6, 7, 13, 17] and thus can provide for ordinal notation systems [6, 13, 14].

Some simple worms are just consistency statements which are known to be related to reflection principles so that by classical results they are related to fragments of arithmetic [27]. Thus, worms—apart from being privileged elements of a decidable logic—can denote both ordinals and fragments of arithmetic. A possibly more important use, however, lies in their relation to Turing progressions: each Turing progression below ε_0 can be approximated by the arithmetical interpretation of a GLP_ω worm. The relation goes even that far so that points in a universal modal model for the closed fragment of GLP_ω can be seen as arithmetical theories axiomatised by Turing progressions [25] so that the model displays all conservation results between the different theories. It is these four different possible denotations for worms that make them so versatile and make new interpretations of GLP_Λ as the current paper so promising.

1.1. Plan of the paper. Section 2 provides some useful lemmata and settles on notation which otherwise is quite standard so that it can be skipped by the initiate readers only to come back to it when needed. Then, in Section 3 the central provability notion of this paper is introduced: one-Münchhausen provability. The usage of the word “one” in there refers to the fact that provability at level α is allowed to use a single oracle sentence of a lower level consistency statement.

Section 4 mainly dwells on the fact that in general we cannot prove that different Münchhausen provability predicates are provably equivalent even if they are so on the low levels. It is observed that we do have uniqueness in case the object theory and the meta theory are provably the same.

Section 5 then proceeds to prove soundness for one-Münchhausen provability for a large class of theories and Section 6 proves arithmetical completeness. In Section 7 it is sketched how one-Münchhausen provability can be formalised in second order arithmetic. The formalisation requires a substantial amount of transfinite induction

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

both in the object and meta theory so that applications to ordinal analysis will become difficult. Finally, in Section 8 some first steps are taken on how to weaken the needed strength of the object and meta theory. By allowing for multiple oracles sentences instead of just one, soundness can be proven without any transfinite induction.

§2. Preliminaries. In this section we dwell succinctly on the necessary notions from both formal arithmetic and modal provability logics. Apart from proving a few new observations, we mainly settle on notation and refer to the literature for details.

2.1. Arithmetic. This paper deals with interpretations of transfinite provability logic. Even though the set-up of such interpretations starts schematically so that our analysis applies to a wide range of theories, we will in particular have second order arithmetic in mind. We refer the reader to standard references for details [5, 19, 30] and only include some minimal comments for expository purposes.

For first-order arithmetic, we shall work with theories with identity in the language $\{0, 1, \exp, +, \cdot, <\}$ of arithmetic where exp denotes the unary function $x \mapsto 2^x$. We define $\Delta^0_0 = \Sigma^0_0 = \Pi^0_0$ formulas (also referred to as *elementary formulas*) as those where all quantifiers occur bounded, that is we only allow quantifiers of the form $\forall\, x{<}t$ or $\exists\, x{<}t$ where t is some term not containing x. We inductively define Π^0_{n+1}/ Σ^0_{n+1} formulas as allowing a block of universal/existential quantifiers up-front a Σ^0_n/ Π^0_n formula. The union of these classes is called the *arithmetical formulas* and denoted by Π^0_ω.

If P is a predicate, the classes relativized to P are defined the same with the sole difference that we consider the predicate P as an atomic formula. We flag relativisation by including the predicate in brackets after the class like, for example, in $\Pi^0_1(P)$.

Peano Arithmetic (PA) contains the basic axioms describing the non-logical symbols together with induction formulas I_φ for any formula φ where as always $I_\varphi := \varphi(0) \wedge \forall x(\varphi(x) \to \varphi(x+1)) \to \forall x \varphi(x)$. When Γ is a complexity class, by $\mathrm{I}\Gamma$ we denote the theory which is like PA except that induction is restricted to formulas in Γ. The theory $\mathrm{I}\Delta^0_0$ is also referred to as *elementary arithmetic*[1] or *Kalmar elementary arithmetic* (EA).

In this paper we also mention collection axioms B_φ which basically state that the range of a function with finite domain is finite: $B_\varphi := \forall z{<}y\, \exists x \varphi(z,x) \to \exists u\, \forall z{<}y\, \exists\, x{<}u \varphi(z,x)$. Again, for a formula class Γ, by $\mathrm{B}\Gamma$ we denote the set of collection axioms for formulas from Γ.

Second order arithmetic is an extension of first order arithmetic where we now add second order set variables together with a binary symbol $\in$ for membership. Instead of extending identity to second order terms we stipulate that second order identity is governed by extensionality: $X = Y :\Leftrightarrow \forall x\ (x \in X \leftrightarrow x \in Y)$. The formula classes Σ^1_n and Π^1_n are defined as their first-order counterpart only that we now count second

[1] In the literature it is more common to work with a formulation of EA in the language without exponentiation. For the purpose of this paper, the differences are not essential.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

order quantification alternations. Likewise, by Π^1_ω we denote the class of all second order formulas.

The strength of various fragments of second order arithmetics is in large determined by their set existence axioms. The *comprehension axiom* for φ tells us that φ not containing x defines a set: $\exists X \forall x (x \in X \leftrightarrow \varphi)$. The second order system ACA_0 contains the defining axioms for the first-order non-logical symbols together with set-induction $0 \in X \wedge \forall x (x{\in}X \rightarrow x{+}1{\in}X) \rightarrow \forall x\ x{\in}X$ and comprehension for all arithmetical formulas.

The theory ACA_0 is conservative over PA for first-order formulas. In [18] the system ECA_0 is introduced as ACA_0 except that comprehension is restricted to Δ^0_0 formulas. In [10, Lemma 3.2] it is proven that ECA_0 is conservative over EA for first-order formulas.

We will tacitly assume that when we are given a theory T, we are actually given a decidable formula τ that binumerates the axioms of T. That is to say, χ is an axiom of T if and only if[2] $T \vdash \tau(\chi)$. For each theory T we denote by $\Box_T$ the unary Σ^0_1-predicate that defines provability in T. That is, $\mathbb{N} \models \varphi$ if and only if φ is provable in T. When we write $\Box_T \varphi(\dot{x})$ we denote the formula with free variable x that expresses that for each number x, the formula $\varphi(\overline{n})$ is provable in T. Here, $\overline{n}$ denotes the numeral of n which is a syntactical expression denoting n, for example defined as $\overline{0} = 0; \overline{x+1} = \overline{x} + 1$.

The Friedman–Goldfarb–Harrington Theorem (FGH for short) states that for any computably enumerable theory U, the corresponding formalised provability predicate is provably Σ^0_1-complete provided U is consistent. Since the theorem provides an important tool in this paper, let us give a precise formulation.

Theorem 2.1 (Friedman–Goldfarb–Harrington). *Let U be a computably enumerable theory with corresponding provability predicate $\Box_U$. We have that for any Σ^0_1 formula $\sigma(x)$, there is a Σ^0_1 formula $\rho(x)$ so that*

$$\mathrm{EA} \vdash \Diamond_U \top \rightarrow \forall x \, \big(\sigma(x) \leftrightarrow \Box_U \rho(\dot{x})\big).$$

The theorem was given its name in [32] in acknowledgment to the intellectual parents. Generalisations to other arithmetical provability predicates were studied in [24, 26]. In particular, the quantification over Σ^0_1 formulas (in the language without exponentiation however) can be made internal in EA and the ρ is obtained from σ by means of an elementary function. A very useful corollary of the FGH theorem is so-called *weak closure* of provability under disjunctions.

Corollary 2.2. *Let U be a computably enumerable theory with corresponding provability predicate $\Box_U$. We have*

$$\mathrm{EA} \vdash \forall \varphi \forall \psi \exists \chi \, (\Box_U \varphi \vee \Box_U \psi \leftrightarrow \Box_U \chi).$$

2.2. Transfinite provability logic. Even though via the FGH theorem the provability predicate $\Box_T$ is in a sense Σ_1 complete for a wide variety of theories, the

[2] We shall refrain from making a difference between syntactical objects and their Gödel numbers when the context allows us so.

provable structural behaviour of the predicate can be described with well-behaved PSPACE decidable propositional modal logics.

The simplest modal logics have one unary modal operator $\Box$ which syntactically behaves like negation. The dual modality $\Diamond$ can be seen as an abbreviation of $\neg\Box\neg$. The basic logic **K** is axiomatised by all propositional tautologies (in the signature with $\Box$) and all so-called distribution axioms $\Box(A \to B) \to (\Box A \to \Box B)$. The rules of **K** are modus ponens and Necessitation: from A conclude $\Box A$.

The logic **K4** arises by adding the transitivity axioms to **K**: $\Box A \to \Box\Box A$. Gödel–Löb's logic **GL** arises to adding Löb's axiom scheme to **K**: $\Box(\Box A \to A) \to \Box A$. It is known that **GL** is a proper extension of **K4** and that it exactly describes the provable structural properties of the provability predicate for a wide range of theories.

In this paper we are interested in provability logics of a collection of provability predicates $[\alpha]$ of increasing strength indexed by ordinals α. For the finite ordinals, this logic was discovered by Japaridze in [22]. We now present this logic, which would be GLP_ω in our notation as given in the following definition.

Definition 2.3. For Λ an ordinal or the class of all ordinals, the logic GLP_Λ is given by the following axioms:

1. all propositional tautologies,
2. Distributivity: $[\xi](\varphi \to \psi) \to ([\xi]\varphi \to [\xi]\psi)$ for all $\xi < \Lambda$,
3. Transitivity: $[\xi]\varphi \to [\xi][\xi]\varphi$ for all $\xi < \Lambda$,
4. Löb: $[\xi]([\xi]\varphi \to \varphi) \to [\xi]\varphi$ for all $\xi < \Lambda$,
5. Negative introspection: $\langle\zeta\rangle\,\varphi \to \langle\xi\rangle\,\varphi$ for $\xi < \zeta < \Lambda$,
6. Monotonicity: $\langle\xi\rangle\,\varphi \to [\zeta]\,\langle\xi\rangle\,\varphi$ for $\xi < \zeta < \Lambda$.

The rules are Modes Ponens and Necessitation for each modality: $\dfrac{\varphi}{[\xi]\varphi}$.

The following lemma is proven in [7].

Lemma 2.4. *The logic* GLP_Λ *is conservative over* $\mathsf{GLP}_{\Lambda'}$ *for* $\Lambda' < \Lambda$.

The lemma is particularly useful in proofs where you only have access to reasoning up to GLP'_Λ and tells you that any statement formulated in this fragment can actually be proven there. We shall use this result throughout the paper, mostly without explicit mention. Let us now prove some basic properties that shall be needed later in the paper.

Lemma 2.5. 1. $\mathsf{GLP} \vdash [\alpha]\langle\beta\rangle\top$ *whenever* $\alpha > \beta$.

2. *For* $\alpha > \beta$ *we have nn* $\vdash \langle\alpha\rangle\top \to \Big(\langle\beta\rangle\varphi \leftrightarrow \langle\alpha\rangle\langle\beta\rangle\varphi\Big)$.

3. *For* $\alpha \geq \beta > 0$ *we have* $\mathsf{GLP} \vdash \langle\alpha\rangle\top \ \to\ \Big(\langle\beta\rangle\phi \vee \Box\psi\Big) \ \leftrightarrow\ \Big(\langle\beta\rangle(\phi \vee \Box\psi)\Big)$.

Proof. We reason in GLP.

For *Item* 1: If $\langle\beta\rangle\top$, then $[\alpha]\langle\beta\rangle\top$ by the negative introspection axiom. In case $[\beta]\bot$ we get by an *ex falso* under the $[\beta]$ modality that $[\beta]\langle\beta\rangle\top$ whence $[\alpha]\langle\beta\rangle\top$ by monotonicity.

For *Item* 2 we work under the assumption that $\langle\alpha\rangle\top$. From $\langle\beta\rangle\varphi$ we get, since $\beta < \alpha$, that $[\alpha]\langle\beta\rangle\varphi$ so from $\langle\alpha\rangle\top$ we get $\langle\alpha\rangle\langle\beta\rangle\varphi$. For the other direction, from $\langle\alpha\rangle\langle\beta\rangle\varphi$ we get by monotonicity that $\langle\beta\rangle\langle\beta\rangle\varphi$ whence by transitivity we obtain the required $\langle\beta\rangle\varphi$.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

For *Item* 3: We now work under the assumption that $\langle\alpha\rangle\top$. The only case to consider in the $\to$ direction is when $\Box\psi$ holds. Then, $\Box\Box\psi$ whence $[\alpha]\Box\psi$ which together with $\langle\alpha\rangle\top$ yields $\langle\alpha\rangle\Box\psi$ whence $\langle\beta\rangle\big(\phi\vee\Box\psi\big)$.

For the $\leftarrow$ direction we need to prove $\langle\beta\rangle\big(\phi\vee\Box\psi\big)\to\langle\beta\rangle\phi\vee\Box\psi$. So, suppose $\langle\beta\rangle\big(\phi\vee\Box\psi\big)$ and $\neg\Box\psi$ whence $[\beta]\neg\Box\psi$. But since $\langle\beta\rangle\big(\phi\vee\Box\psi\big)$ we must have $\langle\beta\rangle\phi$ and by weakening $\langle\beta\rangle\phi\vee\Box\psi$. ⊣

2.3. Transfinite induction and its kin. In various arguments we will have to prove that a statement φ holds for all ordinals α. Often we will prove this by transfinite recursion on α. However, in certain cases, transfinite induction is not available. In such cases there is a technique called *reflexive induction*.

The principle of reflexive induction can syntactically be seen as twice weakening regular transfinite induction. Recall that transfinite induction for a formula φ is

$$\mathsf{TI}_\varphi \ := \ \forall\alpha\big(\forall\beta{<}\alpha\varphi(\beta)\to\varphi(\alpha)\big) \ \to \ \forall\alpha\varphi(\alpha),$$

and for a set of formulas Γ the principle $\mathsf{TI}(\Gamma)$ denotes the collection of all TI_φ for $\varphi\in\Gamma$. As a first weakening one could consider the rule-based version: from $T\vdash\forall\alpha\big(\forall\beta{<}\alpha\varphi(\beta)\to\varphi(\alpha)\big)$, conclude $T\vdash\forall\alpha\varphi(\alpha)$. Now, one can change the antecedent to $T\vdash\forall\alpha\big(\Box_T\forall\beta{<}\alpha\varphi(\dot\beta)\to\varphi(\alpha)\big)$ to arrive at reflexive induction. However, it turns out that by doing so, it has lost all its strength. That is, the resulting principle is provable in almost any theory:

Theorem 2.6 (Reflexive induction). *Let T be any theory capable of coding syntax. If $T\vdash\forall\alpha\Big(\Box_T\big(\forall\beta<\dot\alpha\ \varphi(\beta)\big)\to\varphi(\alpha)\Big)$, then $T\vdash\forall\alpha\varphi(\alpha)$.*

Although this principle is well known since Schmerl's work [28] we include a proof to emphasize that the principle actually does not rely at all on the fact that $<$ is a well-order. As a matter of fact, the proof goes through for any kind of relation and basically boils down to an application of Löb's Theorem.

Proof. We shall see that from the assumption

$$T\vdash\forall\alpha\Big(\Box_T\big(\forall\beta{<}\dot\alpha\ \varphi(\beta)\big)\to\varphi(\alpha)\Big),$$

we get $T\vdash\Box_T\forall\alpha\varphi(\alpha)\to\forall\alpha\varphi(\alpha)$ so that the conclusion $T\vdash\forall\alpha\varphi(\alpha)$ follows by Löb's Theorem.

Thus, we reason in T, pick α arbitrary, assume $\Box_T\forall\alpha\varphi(\alpha)$, or equivalently $\Box_T\forall\theta\varphi(\theta)$, and set out to prove $\varphi(\alpha)$. But using $\Box_T\big(\forall\beta{<}\dot\alpha\ \varphi(\beta)\big)\to\varphi(\alpha)$ in the last step of the following reasoning, we clearly have

$$\begin{array}{rcl}\Box_T\forall\theta\varphi(\theta) & \to & \Box_T\forall\theta\forall\beta{<}\theta\ \varphi(\beta)\\ & \to & \forall\theta\,\Box_T\forall\beta{<}\dot\theta\ \varphi(\beta)\\ & \to & \Box_T\forall\beta{<}\dot\alpha\ \varphi(\beta)\\ & \to & \varphi(\alpha).\end{array}$$

⊣

On occasion, in this paper we will have to combine regular transfinite induction and reflexive induction. We call this amalgamate *transfinite reflexive induction*.

Lemma 2.7 (Transfinite reflexive induction). *Let T be a theory with a sufficient amount of transfinite induction as specified below and let $\prec$ be a well-order in T.*

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

If

$$T \vdash \forall \alpha\Big(\forall \beta{\prec}\alpha\, \varphi(\beta) \;\wedge\; \Box_T\big(\forall \beta{\prec}\dot{\alpha}\, \varphi(\beta)\big) \;\to\; \varphi(\alpha)\Big),$$

then

$$T \vdash \forall \alpha\, \varphi(\alpha).$$

To prove transfinite reflexive induction for φ it suffices that T is capable of coding syntax and proves transfinite induction for formulas of the form $\Box_T \chi \to \varphi$.

Proof. To start our proof we assume

$$T \vdash \forall \alpha\Big(\forall \beta{\prec}\alpha\, \varphi(\beta) \;\wedge\; \Box_T\big(\forall \beta{\prec}\dot{\alpha}\, \varphi(\beta)\big) \;\to\; \varphi(\alpha)\Big). \tag{1}$$

We will prove by transfinite induction on α that

$$T \vdash \forall \alpha \,\Big(\Box_T \forall \beta{\prec}\dot{\alpha}\, \varphi(\beta) \to \varphi(\alpha)\Big), \tag{2}$$

so that the result $T \vdash \forall \alpha\, \varphi(\alpha)$ follows by reflexive induction (Lemma 2.6). Proving (2) for $\alpha = 0$ amounts to showing that $T \vdash \varphi(0)$ which follows directly from (1).

For the inductive step, we reason in T, fix some $\alpha > 0$, assume that

$$\forall \beta{\prec}\alpha \,\Big(\Box_T \forall \gamma{\prec}\dot{\beta}\, \varphi(\gamma) \to \varphi(\beta)\Big), \tag{3}$$

and set out to prove

$$\Box_T \forall \gamma{\prec}\dot{\alpha}\, \varphi(\gamma) \to \varphi(\alpha). \tag{4}$$

To this end, we further assume that $\Box_T \forall \gamma{\prec}\dot{\alpha}\, \varphi(\gamma)$, so that certainly we have $\forall \beta{\prec}\alpha \;\; \Box_T \forall \gamma{\prec}\dot{\beta}\, \varphi(\gamma)$. Combining the latter with (3) yields $\forall \beta{\prec}\alpha \;\; \varphi(\beta)$. This, together with our assumption $\Box_T \forall \gamma{\prec}\dot{\alpha}\, \varphi(\gamma)$ is the antecedent of (1) so that we may conclude $\varphi(\alpha)$ which finishes the proof. ⊣

§3. Theories for Single Oracle Münchhausen provability.

Throughout this section, we fix some ordinal Λ and understand that all other ordinals denoted in this section are majorized by Λ.

3.1. Single Oracle Münchhausen provability.

We are interested in theories T that can formalize a provability notion so that provably in T the following recursion holds:

$$[\zeta]^{\Lambda}_{T}\phi \;\;:\Leftrightarrow\;\; \Box_T \phi \;\vee\; \exists \psi\, \exists \xi{<}\zeta \,\big(\langle \xi \rangle^{\Lambda}_{T} \psi \;\wedge\; \Box_T(\langle \xi \rangle^{\Lambda}_{T} \psi \to \phi)\big). \tag{5}$$

Here, $\Box_T \varphi$ will denote a standard predicate on the natural numbers expressing "the formula (with Gödel number) φ is provable in the theory T." Further, it is understood that $\langle \xi \rangle^{\Lambda}_{T}$ stands for $\neg [\xi]^{\Lambda}_{T} \neg$.

Rather than exposing a concrete theory where this recursion is formalizable in a particular way and provable, we will define a class of theories that are able to define and prove this recursion and have some additional desirable properties.

Next we shall see which properties of the predicates $[\zeta]^{\Lambda}_{T}$ can be proven from the mere recursion defined in (5). It will turn out that under some fairly general conditions we can prove the collection of predicates $[\zeta]^{\Lambda}_{T}$ for $\zeta < \Lambda$ to provide a sound interpretation for GLP_{Λ}.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

In Section 6 we shall see that by requiring slightly more on our predicate and theory, this will give us arithmetical completeness.

In principle it would make sense to study (5) at a higher level of generality. For example, T could be some version of set-theory allowing for uncountable Λ. As long as (5) is provable together with some additional conditions, most of the results of this paper will carry over. It would be natural to require $\Box_T$ to be such that all **GL** theorems are schematically provable in T in such a setting.

Of course, this generalised setting would require one to overcome various substantial technical problems. For one, if each element in the uncountable ordering should be denotable by a syntactical term, the basic language itself should be uncountable so that coding and ordinal representations should be rethought. However, such a generalisation falls outside the scope of the current paper.

3.2. Theories amenable for Single Oracle Münchhausen provability. For the sake of readability we shall often not distinguish between an ordinal $\alpha < \Gamma$, a notation for such an α, or even an arithmetization of such a notation for α. We shall, however, be explicit about the difference between the ordering $<$ on the ordinals and the arithmetization $\prec$ of this ordering on ordinals.

Definition 3.1. Let T be a theory and let Λ denote an ordinal equipped with a representation in the language of T with corresponding represented ordering $\prec$. For this representation, it is required that

$$\begin{array}{l} T \vdash \text{“}\prec \text{ is transitive, right-discrete, and has a minimal element,”} \\ T \vdash (\xi \prec \zeta) \to [\zeta]^{\Lambda}_{T}(\xi \prec \zeta), \\ T \vdash \neg(\xi \prec \zeta) \to [\zeta]^{\Lambda}_{T}\neg(\xi \prec \zeta), \\ \xi < \zeta < \Lambda \text{ implies}^{3}\ T \vdash \xi \prec \zeta. \end{array}$$

We call T a *Single Oracle* Λ *-Münchhausen Theory*—or a Λ *-One-Münchhausen Theory* for short—whenever there is a binary predicate $[\xi]^{\Lambda}_{T}\varphi$ with free variables ξ and φ so that

$$T \vdash \forall \phi \, \forall \zeta {\prec} \Lambda \Big(\, [\zeta]^{\Lambda}_{T}\phi \; \leftrightarrow \; \Box_T \phi \; \vee \; \exists \psi \, \exists \, \xi {\prec} \zeta \, \big(\langle \xi \rangle^{\Lambda}_{T} \psi \; \wedge \; \Box_T \big(\, \langle \xi \rangle^{\Lambda}_{T} \psi \to \phi \big) \big) \Big).$$

In this case, we call the binary predicate $[\xi]^{\Lambda}_{T}\varphi$ a corresponding 1-Münchhausen provability predicate.

The “One” in “Λ-One-Münchhausen Theory” refers to the fact that provability $[\zeta]^{\Lambda}_{T}$ at level ζ makes use of one single oracle sentence $\langle \xi \rangle^{\Lambda}_{T}\psi$. In Section 8 we shall see variations where we allow various oracle sentences to occur.

Often shall we simply drop the, or some of the indices of $[\xi]^{\Lambda}_{T}$ like for example in $[\xi]_T\varphi$ in case the ordinal Λ is clear from the context. To shorten nomenclature further, we shall mostly simply speak of 1-Münchhausen theories and the corresponding 1-Münchhausen provability. Often, when we speak of 1-Münchhausen theories we implicitly assume that we have fixed some 1-Münchhausen provability predicate $[\alpha]\varphi$.

[3] This requirement can be dropped if we are happy with a soundness proof where all ordinals are internally quantified. In this case we assume that each $\alpha < \Gamma$ has a natural representation in T so that it makes sense to speak about the soundness of the necessitation rule.

Observation 3.2. *Since any* 1-*Münchhausen theory T proves that there is a* $\prec$-*minimal element, we shall use the notation* 0 *for this element even if the natural number* (*or object*) *representing this minimal element is not the natural number zero. Likewise, from right-discreteness we know that for any element* $\alpha \prec \Lambda$, *there is a next bigger element that we shall suggestively call* $\alpha + 1$. *In analogy, we shall denote* $0 + 1$ *by* 1, $1 + 1$ *by* 2, $2 + 1$ *by* 3, *et cetera.*

The following observation is immediate.

Lemma 3.3. *Let T be a Single Oracle* Λ-*Münchhausen Theory with corresponding* 1-*Münchhausen provability predicate* $[\alpha]^{\Lambda}_{T}$. *We have that*

$$T \vdash \forall\varphi \left([0]^{\Lambda}_{T}\varphi \leftrightarrow \Box_T\varphi\right).$$

When working with sound theories in the language of arithmetic, we know that all the corresponding 1-Münchhausen consistency statements are actually true:

Proposition 3.4. *Let T be a sound Single Oracle* Λ-*Münchhausen Theory in the language of arithmetic with corresponding* 1-*Münchhausen provability predicate* $[\alpha]^{\Lambda}_{T}$. *Then for each* $\xi \prec \Lambda$ *we have* $\mathbb{N} \models \langle\xi\rangle^{\Lambda}_{T}\top$.

Proof. By a simple case distinction. In case $\xi = 0$, we get from a hypothetical $\mathbb{N} \models [0]_T\bot$ together with the soundness and the above lemma that $\mathbb{N} \models \Box_T\bot$ so that $T \vdash \bot$ which cannot be.

In case $\xi \succ 0$, suppose for a contradiction that $\mathbb{N} \models [\xi]_T\bot$. Then, using soundness of T, we only need to consider the case that

$$\mathbb{N} \models \exists\psi\, \exists\zeta{\prec}\xi \left(\langle\zeta\rangle_T\psi \wedge \Box_T(\langle\zeta\rangle_T\psi \to \bot)\right),$$

so that for some ordinal $\zeta \prec \xi$ and some formula ψ we have $\mathbb{N} \models \langle\zeta\rangle_T\psi$. Also $\mathbb{N} \models \Box_T(\langle\zeta\rangle_T\psi \to \bot)$ so that $T \vdash \langle\zeta\rangle_T\psi \to \bot$ whence by soundness of T we see that $\mathbb{N} \models \neg\langle\zeta\rangle_T\psi$ which is a contradiction. ⊣

We note that the above argument does not use transfinite induction.

§4. On uniqueness of Münchhausen provability.

The definition of 1-Münchhausen provability allows for various different 1-Münchhausen predicates to exist. Of course, it would be highly desirable that the defining equivalence (5) for 1-Münchhausen provability defined a T provably unique predicate. We can prove uniqueness of the predicate via an external induction up to any level below ω.

Lemma 4.1. *Let T be a sound Single Oracle* Λ-*Münchhausen Theory for* $\Lambda \geq \omega$, *with corresponding* 1-*Münchhausen provability predicates* $[\alpha]^{\Lambda}_{T}$ *and* $\overline{[\alpha]}^{\Lambda}_{T}$. *We have for any natural number n that*

$$T \vdash \forall\varphi \left([n]^{\Lambda}_{T}\varphi \leftrightarrow \overline{[n]}^{\Lambda}_{T}\varphi\right).$$

Proof. We proceed by an external induction where the base case follows directly from Lemma 3.3. We shall omit super and sub indices.

For the inductive step, we reason in T, fix some formula φ, fix the $(n+1)$th element in the $\prec$ ordering, and assume $[n+1]\varphi$. In the non-trivial case, there is some formula ψ and an element $\tilde{m}{\prec}n+1$ so that $\langle\tilde{m}\rangle\psi$ and $\Box(\langle\tilde{m}\rangle\psi \to \varphi)$. Here we end our reasoning inside T. Since we can prove that any element $\prec$-below the externally given $n+1$ is either the zero-th, or the first, or ...or, the n-th element, we know that $\tilde{m}$ corresponds to some natural number $m < n+1$. Thus, we can appeal to the external induction hypothesis that tells us that

$$T \vdash \forall\psi\ ([m]\psi \leftrightarrow \overline{[m]}\psi), \tag{6}$$

and consequently

$$T \vdash \Box\forall\psi\ ([m]\psi \leftrightarrow \overline{[m]}\psi). \tag{7}$$

These two ingredients are sufficient to conclude $\overline{[m]}\psi$. Of course the other direction goes exactly the same. $\dashv$

Let us make some observations about this simple proof. First, we observe that we could only conclude (7) from (6) by necessitation since the meta-theory as in $T \vdash \ldots$ is the same as the object-theory as in $\Box_T$. Second, we observe that we only had access to the inductive hypothesis since we can express in the language of first order logic that being smaller than the $(n+1)$th element implies being equal to one of the zero-th, or ..., or the nth element. Of course, we cannot generalize this to the first limit ordinal and hence our external induction cannot be extended to the transfinite.

If we wish to generalize our argument to the transfinite, we should replace our external induction by an internal one. Of course, then in our meta-theory, we should have access to transfinite induction. However, we only see how to continue the proof in the case where the object theory equals the meta-theory and consequently also has the same amount of transfinite induction.

Lemma 4.2. *Let T be a theory that proves the recursion from (5) for two predicates $[\zeta]_U$ and $\overline{[\zeta]}_U$. We further suppose that T proves the basic facts about the ordering $\langle \Lambda, \prec\rangle$. Also, we assume that T proves transfinite $\Pi_1([\alpha], \overline{[\alpha]})$ induction.*

If T and U are T-provably equivalent, then we have that $[\zeta]_U$ and $\overline{[\zeta]}_U$ are T-provably equivalent predicates.

Proof. We have chosen a formulation where T and U are different from the outset so that we clearly see at what point we need to assume that T is T-provably equivalent to U.

Thus, we reason in T and will as a first attempt prove by transfinite $\Pi_1([\alpha], \overline{[\alpha]})$ induction that

$$\forall\zeta\,\forall\varphi\ ([\zeta]_U\varphi \ \leftrightarrow\ \overline{[\zeta]}_U\varphi).$$

For $\zeta = 0$ the equivalence is obvious. Thus, we fix some $\zeta \succ 0$ and focus on one implication the other being analogous. Thus, we assume that $[\zeta]_U\varphi$ and set out to prove $\overline{[\zeta]}_U\varphi$.

From the assumption $[\zeta]_U\varphi$ we find—in the non-trivial case—some formula ψ and ordinal $\xi \prec \zeta$ so that $\langle\xi\rangle_U\psi$ and $\Box_U(\langle\xi\rangle_U\psi \to \varphi)$. The inductive hypothesis now will tell us that $\langle\xi\rangle_U\psi \leftrightarrow \overline{\langle\xi\rangle}_U\psi$.

However, there is no way that we know that this equivalence is *provable*, that is, that we have $\Box_U\Big(\langle\xi\rangle_U\psi \leftrightarrow \overline{\langle\xi\rangle}_U\psi\Big)$. The latter would be needed to conclude $\Box_U(\overline{\langle\xi\rangle}_U\psi \to \varphi)$ so that $\overline{[\zeta]}_U\varphi$.

The problem cannot be solved by strengthening the induction to for example

$$\forall \phi\ \big[([\zeta]_U \phi \leftrightarrow \overline{[\zeta]}_U \phi)\ \wedge\ \Box_U([\dot{\zeta}]_U \phi \leftrightarrow \overline{[\dot{\zeta}]}_U \phi)\big],$$

since then the problem will simply come back but now under a box.

However, when $T = U$ we have access to transfinite reflexive induction as formulated in Lemma 2.7. That is, in order to show that $\forall \varphi\ ([\zeta]_U \varphi \ \leftrightarrow\ \overline{[\zeta]}_U \varphi)$ for a particular ζ we may assume both $\forall\, \xi \prec \zeta\, \forall \varphi\ ([\zeta]_U \varphi \ \leftrightarrow\ \overline{[\zeta]}_U \varphi)$ and also $\Box_U\big(\forall\, \xi \prec \dot{\zeta}\, \forall \varphi\ ([\zeta]_U \varphi \ \leftrightarrow\ \overline{[\zeta]}_U \varphi)\big)$ which makes that the proof now goes through easily. ⊣

This lemma tells us that solutions to the recursion equivalence (5) possibly need not be provably unique if the object theory U is different from the meta theory T or in case we do not have the sufficient amount of transfinite induction available. Actually, we conjecture that it is possible to come up with theories T and U and with two predicates $[\xi]_U$ and $\overline{[\xi]}_U$ both satisfying the recursive equivalence such that T does not prove the equivalence of $[\xi]_U$ and $\overline{[\xi]}_U$. We observe that not having provably unique fixpoints need not necessarily be a big problem and similar phenomena occur with for example Rosser fixpoints.

However, as we shall see in Section 5, we also need the object theory to be equal to the meta theory if we wish to prove the soundness of GLP_Λ with respect to the $[\zeta]^\Lambda_U$ predicates. In particular, the arithmetical soundness of the Necessitation rule requires the object and meta theory to be equal.

In case the object theory is not equal to the meta-theory, we can only prove a weak form of uniqueness as expressed in the following lemma.

Lemma 4.3. *Let T be a theory that proves the recursion expressed in Equation* (5) *for two predicates $[\zeta]_U$ and $\overline{[\zeta]}_V$ with V possibly different from U. We further suppose that T proves the basic facts about the ordering $\langle \Lambda, \prec \rangle$. In case T proves the arithmetical soundness of GLP_Λ for both predicates $[\zeta]_U$ and $\overline{[\zeta]}_V$, then* (*omitting subscripts*)

$$T \vdash \forall\, \alpha \prec \Lambda \Big(\ (\langle \alpha \rangle \top \leftrightarrow \overline{\langle \alpha \rangle} \top)\ \longrightarrow \forall \varphi\, \exists \psi\ ([\alpha]\varphi \leftrightarrow \overline{[\alpha]}\psi)\Big).$$

Proof. We reason[4] in the theory T, consider some $\alpha \prec \Lambda$ and sentence φ, assume that $\langle \alpha \rangle \top \leftrightarrow \overline{\langle \alpha \rangle} \top$, and claim that there exists ψ such that $[\alpha]\varphi \leftrightarrow \overline{[\alpha]}\psi$. If $[\alpha]\varphi$, then we put to be $\top$. Further, assume that $\neg[\alpha]\varphi$. In this case we have $\neg[\alpha]\bot$, i.e., $\langle \alpha \rangle \top$. Thus we have $\neg\overline{[\alpha]}\bot$ and hence we could put ψ to be $\bot$. ⊣

Of course, this proof is not constructive and as such little informative since the ψ formula is not uniformly obtained from φ. In [26] this question is taken up and addressed.

So far, in this section we have seen that with the techniques presented here we cannot prove that 1-Münchhausen provability predicates are uniquely defined by the recursion in (5) and we actually conjecture that non-equivalent solutions do exist. Only in the finite ordinals we can prove uniqueness. This allows us to relate the provability notions from this paper to similar ones from the literature. The most

[4] We thank the referee for pointing out this simple proof.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

prominent example is given by the predicate

$$[n]^{\mathsf{True}}_T\varphi \quad \text{which stands for} \quad \exists\,\pi{\in}\Pi^0_1\,\Big(\mathsf{True}_{\Pi^0_1}(\pi)\wedge\Box_T\big(\pi\to\varphi\big)\Big).$$

Furthermore, in [24] a reading is given where the modal operators $[n]\varphi$ are interpreted as follows.

$$\begin{aligned}[0]^{\Box}_T\phi &\;:=\Box_T\phi,\quad \text{and}\\ [n+1]^{\Box}_T\phi &\;:=\;\Box_T\phi\;\vee\;\exists\,\psi\bigvee_{0\leq m\leq n}\Big(\langle m\rangle^{\Box}_T\psi\;\wedge\;\Box(\langle m\rangle^{\Box}_T\psi\to\phi)\Big).\end{aligned} \tag{8}$$

Soundness for this interpretation in PA was proven and a strong relation was given to the truth provability predicates $[n]^{\mathsf{True}}_T$. The next lemma is a strengthening on the one hand since we weaken the base theory to EA and a weakening on the other hand since we only consider two modalities.

Lemma 4.4. *Let T be a theory that contains* EA. *We have that*

1. $\mathrm{EA}\vdash\forall\varphi\,([1]^{\Box}_T\varphi\leftrightarrow[1]^{\mathsf{True}}_T\varphi)$;
2. GLP_2 *is sound for T when interpreting* $[0]$ *as* $\Box_T$ *and* $[1]$ *as* $[1]^{\Box}_T$;
3. *In case that moreover T proves the Σ^0_1-collection principle we have* $T\vdash\forall\,\varphi\,\forall\psi\,\exists\chi\,\big([1]^{\Box}_T\varphi\vee[1]^{\Box}_T\psi\leftrightarrow[1]^{\Box}_T\chi\big)$.

Proof. It is easy to prove inside EA that $[0]^{\Box}\varphi\leftrightarrow\Box\varphi$ (we omit the subscripts). Likewise, $[0]^{\Box}\varphi\to[1]^{\Box}\varphi$ and $[1]^{\Box}\bot\to[1]^{\Box}\varphi$ are easy to prove. With these ingredients the first item easily follows: one direction is obvious since any oracle sentence of the form $\Diamond\psi$ is in Π^0_1. The other direction is immediate in case $[0]^{\Box}\bot$ and in the case $\langle 1\rangle^{\Box}\top$ it follows from the FGH theorem since under the consistency assumption, any Π^0_1 formula is equivalent and provably so to a formula of the form $\Diamond\psi$.

The second item follows from the first since the statement holds for the $[n]^{\mathsf{True}}_T$ provability predicates (see, e.g., [5]).

The third item is implicit in [24] and explicitly stated and proven in [26] for the $[1]^{\mathsf{True}}_T$ predicate which suffices by the first item of this lemma. ⊣

Via an easy external induction we can prove that (8) and (5) define provably equivalent predicates for all natural numbers. That is to say, if T is a 1-Münchhausen theory, then for each natural number n we have that

$$T\vdash\forall\varphi\,\big([n]^{\Box}_T\varphi\leftrightarrow[n]^{\Lambda}_T\varphi\big) \tag{9}$$

for any 1-Münchhausen provability predicate $[\alpha]^{\Lambda}_T$. Moreover, in Lemma 4.1 we know that any 1-Münchhausen provability predicate $[\alpha]^{\Lambda}_T$ will be uniquely defined up to ω. For later in the paper, we formulate the following corollary:

Corollary 4.5. *Let T be a* Λ-1-*Münchhausen theory with* $\Lambda>2$ *and corresponding* 1-*Münchhausen provability predicate* $[\alpha]^{\Lambda}_T$. *Moreover, let T contain* $\mathsf{B}\Sigma^0_1$.

1. GLP_2 *is sound for T when interpreting* $[0]$ *as* $[0]^{\Lambda}_T$ *and* $[1]$ *as* $[1]^{\Lambda}_T$;
2. $T\vdash\forall\,\varphi\,\forall\psi\,\exists\chi\,\big([1]^{\Lambda}_T\varphi\vee[1]^{\Lambda}_T\psi\leftrightarrow[1]^{\Lambda}_T\chi\big)$.

Proof. This follows directly from (9) and Lemma 4.4. ⊣

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

§5. Arithmetical soundness for one-Münchhausen provability. In this section we will consider Λ-One-Münchhausen theories T and their corresponding Λ-One-Münchhausen provability predicates for some fixed ordinal Λ represented in T. We shall see that from the mere defining recursion on the provability predicate we can obtain soundness of GLP_Λ.

Many arguments in this section require transfinite induction. As we have observed in Section 4 this means that the base theory should also prove a decent amount of transfinite induction. In Section 8 we shall see how the need of transfinite induction can be circumvented by slightly altering the defining recursion.

5.1. Basic properties. Let us start the soundness proof by some basic observations that need very little arithmetical strength to be proven. In particular, the following facts do not require transfinite induction.

Lemma 5.1. *Let T be a Λ-One-Münchhausen theory with corresponding provability predicate $[\xi]^{\Lambda}_{T}$. We have the following.*

1. $T \vdash \forall \xi \, \forall \chi \, \big([\xi]_T \bot \to [\xi]_T \chi\big)$ *and more in general,*
2. $T \vdash \forall \xi \, \forall \varphi, \chi \, \Big([\xi]_T \varphi \, \wedge \, \Box_T(\varphi \to \chi) \to [\xi]_T \chi\Big)$,
3. $T \vdash \forall \varphi \, \forall \psi \, \Big([\xi]_T \varphi \, \wedge \, \Box_T \psi \to [\xi]_T(\varphi \wedge \psi)\Big)$,
4. $T \vdash \exists x \, [\xi]_T \varphi(\dot{x}) \; \to \; [\xi]_T \exists x \varphi(x)$.

Proof. Clearly, the first item follows from the second, so we reason in T and assume $[\xi]_T \varphi$. Thus, in the non-trivial case, for some ψ and for some $\zeta \prec \xi$ we have $\langle \zeta \rangle_T \psi$ and $\Box_T(\langle \zeta \rangle_T \psi \to \varphi)$. Clearly, since $\Box_T(\varphi \to \chi)$, we have also $\Box_T(\langle \zeta \rangle_T \psi \to \chi)$ so that $[\xi]_T \chi$.

The third item follows from the second since in case of $\Box_T \psi$ we also have $\Box_T\big(\varphi \to (\varphi \wedge \psi)\big)$.

The fourth item follows by an easy case distinction on ξ being zero or not and both cases essentially follow from the fact that provably $\exists x \Box_T \varphi(\dot{x}) \to \Box_T \exists x \varphi(x)$. ⊣

From our defining recursion (5), we get the axiom of negative introspection and the axiom of monotonicity almost for free.

Lemma 5.2. *Let $\xi < \zeta < \Lambda$ be ordinals in a Λ-One-Münchhausen theory T. We have*

1. $T \vdash \forall \varphi \, \big(\langle \xi \rangle_T \varphi \; \to \; [\zeta]_T \langle \xi \rangle_T \varphi\big)$;
2. $T \vdash \forall \varphi \, \big([\xi]_T \varphi \; \to \; [\zeta]_T \varphi\big)$.

Proof. Item 1 is immediate since $\Box_T(\langle \xi \rangle^{\Box}_{T} \varphi \to \langle \xi \rangle^{\Box}_{T} \varphi)$ using the fact that $\xi < \zeta$ implies $T \vdash \xi \prec \zeta$. Likewise, Item 2 follows directly from the definition since provably $\eta \prec \xi \to \eta \prec \zeta$ (recall that we required that Münchhausen theories prove the transitivity of $\prec$ and moreover, $\xi < \zeta$ implies $T \vdash \xi \prec \zeta$). ⊣

It is easy yet important to observe that we actually have a formalized version of the previous lemma where we internally quantify over the ordinals. As such, the formalized lemma can be used for example in an induction where possibly non-standard ordinals are called upon.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

LEMMA 5.3. *Let T be a Λ-One-Münchhausen theory. We have*

1. $T \vdash \forall \xi {\prec} \zeta {\prec} \Lambda\, \forall \varphi\, \big(\langle \xi \rangle_T \varphi \ \to\ [\zeta]_T \langle \xi \rangle_T \varphi\big)$;
2. $T \vdash \forall \xi {\prec} \zeta {\prec} \Lambda\, \forall \varphi\, \big([\xi]_T \varphi \ \to\ [\zeta]_T \varphi\big)$.

These cross axioms are for many interpretations of GLP_Λ actually the harder axioms to prove sound. But in the Münchhausen interpretations they come almost for free.

The above lemma can also be interpreted that any 1-Münchhausen provability predicate is monotone in the ordinal parameter. We note that it is not trivial to see that the 1-Münchhausen provability predicate is monotone in the underlying base theory: Suppose that, for example, we have a formulation of elementary arithmetic and axiomatic set theory so that provably EA $\subset$ ZFC. This means that for any formula φ we have $\Box_{\mathrm{EA}}\varphi \to \Box_{\mathrm{ZFC}}\varphi$. Is it now easy to see that we also have the expected $[1]_{\mathrm{EA}}\varphi \to [1]_{\mathrm{ZFC}}\varphi$?

Let us suppose that $[1]_{\mathrm{EA}}\varphi$ because of some $\langle 0\rangle_{\mathrm{EA}}\psi$ with $\Box_{\mathrm{EA}}(\langle 0\rangle_{\mathrm{EA}}\psi \to \varphi)$. A priori it is not at all clear how this information will yield us a ψ' so that $\Box_{\mathrm{ZFC}}\Big(\langle 0\rangle_{\mathrm{ZFC}}\psi' \to \varphi\Big)$ and furthermore $\langle 0\rangle_{\mathrm{ZFC}}\psi'$: where would we get so much ZFC consistency strength from?[5]

At this point we can prove the soundness of the necessitation rule.

LEMMA 5.4. *Let T be a Λ-One-Münchhausen theory with corresponding* 1-*Münchhausen provability predicate $[\alpha]^\Lambda_T$. For any $\alpha \prec \Lambda$ we have that if $T \vdash \varphi$, then $T \vdash [\alpha]^\Lambda_T\varphi$.*

PROOF. We will only show $\frac{\varphi}{\Box_T\varphi}$. This is sufficient since necessitation for larger ordinals $\frac{\varphi}{[\alpha]_T\varphi}$ follows from the monotonicity of the predicate in α. But, as always $T \vdash \varphi$ can be expressed as a Σ^0_1 sentence which is true whence by Σ^0_1 completeness we get $T \vdash \Box_T\varphi$. ⊣

We shall now prove the remaining GLP axioms to be sound. The following lemma, which was proven in [18], tells us that we don't need to care about Löb's axiom $[\xi]^\Box([\xi]^\Box\varphi \to \varphi) \to [\xi]^\Box\varphi$.

LEMMA 5.5. *Let $\mathsf{GL}^{\blacksquare}$ denote the extension of GL with a new operator $\blacksquare$ and the following axioms for all formulas ϕ, and ψ:*

1. $\vdash \Box\phi \to \blacksquare\phi$,
2. $\vdash \blacksquare(\phi \to \psi) \to (\blacksquare\phi \to \blacksquare\psi)$ *and,*
3. $\vdash \blacksquare\phi \to \blacksquare\blacksquare\phi$.

Then, for all ϕ,

$$\mathsf{GL}^{\blacksquare} \vdash \blacksquare(\blacksquare\phi \to \phi) \to \blacksquare\phi.$$

Consequently, we only need to focus on the transitivity axioms $[\xi]\varphi \to [\xi][\xi]\varphi$ and distribution axioms $[\xi](\varphi \to \psi) \to ([\xi]\varphi \to [\xi]\psi)$ in our soundness proof. It is

[5] We have that ZFC is much stronger than EA, whence provably $\Diamond_{\mathrm{EA}}\chi \to \Box_{\mathrm{ZFC}}\Diamond_{\mathrm{EA}}\chi$. Consequently, in this particular example we could take $\psi' = \Diamond_{\mathrm{EA}}\psi$: in case $\Box_{\mathrm{ZFC}}\bot$ we trivially have $\Box_{\mathrm{ZFC}}\varphi$ and $\Diamond_{\mathrm{ZFC}}\top \to (\Diamond_{\mathrm{EA}}\psi \leftrightarrow \Diamond_{\mathrm{ZFC}}\Diamond_{\mathrm{EA}}\psi)$. However, for general $T \subset U$ we cannot use the same formula $\Diamond_T\psi$ to guarantee $[1]_T\varphi \to [1]_U\varphi$.

in this part where we need to assume that the object and meta theory are equal so that we have access to transfinite reflexive induction as formulated in Lemma 2.7.

5.2. On weak closure. A basic modal principle has that provability commutes with conjunctions: $\Box A \wedge \Box B \leftrightarrow \Box(A \wedge B)$. To have this reflected in our arithmetical soundness proof of 1-Münchhausen provability we shall need to combine two oracle calls $\langle\beta\rangle\varphi'$ and $\langle\beta\rangle\psi'$ into a single one. Clearly, we cannot expect consistency statements to be closed under conjunctions (or dually, provability statements to be closed under disjunctions). However, weak closure under conjunctions in the sense of $\forall\varphi'\forall\psi'\exists\chi\ (\langle\beta\rangle\varphi' \wedge \langle\beta\rangle\psi' \leftrightarrow \langle\beta\rangle\chi)$ as expressed in its dual form in Corollary 2.2 is sufficient for our purpose.

This subsection is dedicated to proving the weak closure property for 1-Münchhausen provability. We shall see how the well-known proof for the FGH theorem can be generalised to the current setting.

The standard proof for the FGH theorem goes via witness comparison statements of the form: one witness/proof occurs before another witness/proof. To generalise this, we will introduce the notion of an α-proof.

Definition 5.6. Let T be a single oracle Λ-Münchhausen Theory with corresponding 1-Münchhausen provability predicate $[\alpha]^{\Lambda}_{T}$. An α *-proof* x for φ is defined inside T as a triple $x = \langle\beta, \varphi', p\rangle$ so that either $\mathsf{Proof}_T(p,\varphi)$ or otherwise $(\beta \prec \alpha) \wedge (\langle\beta\rangle\varphi') \wedge \mathsf{Proof}_T(\langle\beta\rangle\varphi' \to \varphi)$. We will write $\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$.

Under the assumption that our predicate $[\beta]^{\Lambda}_{T}$ is sound for $\mathsf{GLP}_{\alpha+1}$ we can prove that the $[\alpha]^{\Lambda}_{T}$ predicate is complete for any Boolean combination of statements of the form $\mathsf{Proof}^{\alpha}_{T}(p,\psi)$. This is reflected in the following lemma.

Lemma 5.7. *Let T be a single oracle* Λ*-Münchhausen Theory with corresponding* 1*-Münchhausen provability predicate* $[\alpha]^{\Lambda}_{T}$ *so that T proves all the instantiations of axioms of* $\mathsf{GLP}_{\alpha+1}$. *Then we have that*

1. $T \vdash \mathsf{Proof}^{\alpha}_{T}(x,\varphi) \to [\alpha]^{\Lambda}_{T}\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$;
2. $T \vdash \neg\mathsf{Proof}^{\alpha}_{T}(x,\varphi) \to [\alpha]^{\Lambda}_{T}\neg\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$.

Proof. The proof is easy and basically follows from the soundness for $\mathsf{GLP}_{\alpha+1}$. Let us briefly comment on the second item. So, suppose that $\neg\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$. In case x is not a triple or not a triple of the required format, we can express this in a decidable way so that we also know this under $\Box_T$ whence also under $[\alpha]_T$. We now omit subscripts on the boxes in the remainder of this proof.

For $x = \langle\beta, \varphi', p\rangle$ of the right format we know from $\neg\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$ that (a.) $\neg\mathsf{Proof}_T(p,\varphi)$ and (b.) $\neg(\beta \prec \alpha)$ or $\neg\langle\beta\rangle\varphi'$ or $\neg\mathsf{Proof}^{\alpha}_{T}(p, \langle\beta\rangle\varphi' \to \varphi)$. We recall that $\neg\langle\beta\rangle\varphi'$ is just $[\beta]\neg\varphi'$ and by soundness we know $[\beta]\neg\varphi' \to [\alpha][\beta]\neg\varphi'$. Thus, for each of the disjuncts in (b.) we can obtain that disjunct and whence the whole disjunction under the $[\alpha]$ predicate. Likewise, we get $(a.)$ under the $[\alpha]$ predicate. By soundness, the $[\alpha]$ predicate commutes with conjunctions so that we obtain $[\alpha]^{\Lambda}_{T}\neg\mathsf{Proof}^{\alpha}_{T}(x,\varphi)$. ⊣

The notion of α-proofs allows us to compare them so that we can prove our main theorem.

Theorem 5.8. *Let T be a single oracle* Λ*-Münchhausen Theory with corresponding* 1*-Münchhausen provability predicate* $[\alpha]^{\Lambda}_{T}$ *so that T proves transfinite* $\Sigma^0_1([\alpha]^{\Lambda}_{T})$

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

induction and all the instantiations of axioms of $\mathsf{GLP}_{\alpha+1}$. *We then have*

$$T \vdash\!\vdash \forall\alpha\,\forall\varphi\,\forall\psi\,\exists\chi\,\Big([\alpha]^{\bigwedge}_{T}\varphi \vee [\alpha]^{\bigwedge}_{T}\psi \quad \leftrightarrow [\alpha]^{\bigwedge}_{T}\chi\Big).$$

Proof. We omit sub and superscripts, reason in T, fix φ and ψ arbitrary, and consider the following fixpoint:

$$\rho \leftrightarrow ([\alpha]\varphi \vee [\alpha]\psi) \leq [\alpha]\rho,$$

where $([\alpha]\varphi \vee [\alpha]\psi) \leq [\alpha]\rho$ is short for

$$\exists x\Big(\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big) \wedge \forall\, y{<}x\neg\mathsf{Proof}^{\alpha}(y,\rho)\Big).$$

We now claim that $([\alpha]\varphi \vee [\alpha]\psi) \leftrightarrow [\alpha]\rho$. In case that $[\alpha]\bot$ this is trivial so we may work under the assumption that $\langle\alpha\rangle\top$ and set out to prove both directions.

$\rightarrow$: For a contradiction we suppose $([\alpha]\varphi \vee [\alpha]\psi)$ but $\neg[\alpha]\rho$. Then certainly $([\alpha]\varphi \vee [\alpha]\psi) \leq [\alpha]\rho$. In particular for some x, we have $\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big) \wedge \forall\, y{<}x\neg\mathsf{Proof}^{\alpha}(y,\rho)$. Using Lemma 5.7 we may now conclude $[\alpha]\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big) \wedge \forall\, y{<}x[\alpha]\neg\mathsf{Proof}^{\alpha}(y,\rho)$. From the latter conjunct, $\Sigma^0_1([\alpha]^{\bigwedge}_{T})$ induction and closure of the $[\alpha]$ predicate under conjunctions we obtain $[\alpha]\Big(\neg\mathsf{Proof}^{\alpha}(0,\rho) \wedge \cdots \wedge \neg\mathsf{Proof}^{\alpha}(x-1,\rho)\Big)$ whence $[\alpha]\forall\, y{<}\dot{x}\neg\mathsf{Proof}^{\alpha}(y,\rho)$. Collecting our insights under the $[\alpha]$ we see that

$$[\alpha]\exists x\Big(\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big) \wedge \forall\, y{<}x\neg\mathsf{Proof}^{\alpha}(y,\rho)\Big),$$

which is by the definition of our fixpoint nothing but $[\alpha]\rho$ which is a contradiction.

$\leftarrow$: For a contradiction we suppose $[\alpha]\rho$ but $\neg([\alpha]\varphi \vee [\alpha]\psi)$ and recall that we may employ the additional assumption that $\langle\alpha\rangle\top$. From $[\alpha]\rho$ and $\neg([\alpha]\varphi \vee [\alpha]\psi)$ we obtain

$$\exists y\Big(\mathsf{Proof}^{\alpha}(y,\rho) \wedge \forall\, x{<}y{+}1\neg\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big)\Big).$$

By a reasoning analogous to the proof of the other implication we conclude

$$[\alpha]\exists y\Big(\mathsf{Proof}^{\alpha}(y,\rho) \wedge \forall\, x{<}y{+}1\neg\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big)\Big).$$

From the properties of $<$ being a total order, we obtain

$$[\alpha]\neg\exists x\Big(\big(\mathsf{Proof}^{\alpha}(x_0,\varphi) \vee \mathsf{Proof}^{\alpha}(x_1,\psi)\big) \wedge \forall\, y{<}x\neg\mathsf{Proof}^{\alpha}(y,\rho)\Big),$$

which is nothing but $[\alpha]\neg\rho$. Combining this with our assumption $[\alpha]\rho$ yields via the soundness of $\mathsf{GLP}_{\alpha+1}$ that $[\alpha]\bot$ but that contradicts our additional assumption $\langle\alpha\rangle\top$. ⊣

We emphasise that both Lemma 5.7 and Theorem 5.8 assume that soundness for $\mathsf{GLP}_{\alpha+1}$ holds. This assumption justifies that in particular we may use the fact that the $[\alpha]$ provability predicate commutes with conjunctions. Indeed, soundness for $\mathsf{GLP}_{\alpha+1}$ is a rather strong assumption, however, at the place where Theorem 5.8 is used in Item 5 of Theorem 5.9 below, we shall indeed have established the required soundness.

5.3. Soundness. Now that we have proven weak closure of one-Münchhausen provability under disjunctions we can finally prove arithmetical soundness.

Theorem 5.9. *Let T be a Λ-One-Münchhausen theory and let $[\alpha]^{\Lambda}_{T}$ be a corresponding provability predicate. If T proves transfinite $\Pi^0_1([\alpha]^{\Lambda}_T)$ and $\Sigma^0_1([\alpha]^{\Lambda}_T)$ induction we have that*

1. *T proves that all the rules and axioms of* GLP *are sound w.r.t. T by interpreting $[\alpha]$ as $[\alpha]^{\Lambda}_T$; in particular*
2. *Distributivity*: $T \vdash \forall\alpha\,\forall\varphi\,\forall\psi\,\Big([\alpha]^{\Lambda}_T(\varphi \to \psi) \to ([\alpha]^{\Lambda}_T\varphi \to [\alpha]^{\Lambda}_T\psi)\Big)$;
3. *Closure under conjunctions*:
$$T \vdash \forall\alpha\,\forall\varphi\,\forall\psi\,\Big([\alpha]^{\Lambda}_T\varphi \wedge [\alpha]^{\Lambda}_T\psi \;\;\leftrightarrow\;\; [\alpha]^{\Lambda}_T(\varphi \wedge \psi)\Big);$$
4. *Transitivity*: $T \vdash \forall\alpha\,\forall\varphi\,\Big([\alpha]^{\Lambda}_T\varphi \to [\alpha]^{\Lambda}_T[\alpha]^{\Lambda}_T\varphi\Big)$;
5. *Weak closure under disjunctions*:
$$T \vdash \forall\alpha\,\forall\varphi\,\forall\psi\,\exists\chi\,\Big([\alpha]^{\Lambda}_T\varphi \vee [\alpha]^{\Lambda}_T\psi \;\;\leftrightarrow [\alpha]^{\Lambda}_T\chi\Big).$$

Proof. If we wish to prove Item 1, we should prove the soundness of the rules and of the axioms.

As to the rules, the only rules of GLP are modus ponens and a necessitation rule for each modality: $\dfrac{\varphi}{[\xi]^{\Box}_T\varphi}$. As pointed out in Lemma 5.4 the soundness of the necessitation rules follows from necessitation for $\Box_T$ and by monotonicity, Lemma 5.3. As always, the soundness of modus ponens is immediate.

In the remainder of our proof we shall thus focus on the axioms. Since we proved the correctness of the negative introspection axioms—axioms of the form $\langle\beta\rangle\varphi \to [\alpha]\langle\beta\rangle\varphi$ for $\beta < \alpha$—and of the monotonicity axioms—axioms of the form $[\beta]\varphi \to [\alpha]\varphi$ for $\beta < \alpha$—without any induction in Lemma 5.3 and since by Lemma 5.5 we may disregard Löb's axiom, we set out to prove the remaining axioms which are just the distribution and the transitivity axioms to complete a proof of Item 1. In other words, to complete the proof of Item 1 we should prove Items 2 and 4.

To prove that both items hold up to a certain level $\alpha < \Lambda$ we proceed by an internal transfinite reflexive induction on α as expressed in Lemma 2.7. We need to prove both items simultaneously since they depend on each other. As a matter of fact, to get the proof going we will need to do some induction building and prove Items 2–4 of the proof simultaneously by a transfinite reflexive induction on α.

Thus, we will reason in T and shall mostly omit the subscript T and the superscript Λ in the remainder of this proof. The base case of the theorem is known to hold via the soundness of **GL** and the FGH theorem.

For the reflexive inductive step, we are to prove our four items (Items 2–4) at level α assuming that we have access to all four items at any level $\beta \prec \alpha$ and we also have these four items under a regular provability predicate $\Box_T$ at any level $\beta' \prec \alpha$. As we observed before, Item 1 at level α (soundness of GLP_α) follows directly from Items 2–4 for levels $\beta \prec \alpha$. Thus, we may in our inductive step assume that we have access—and T-provably so—to all GLP_α reasoning. Let us thus focus on the first item to prove:

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

Item 3: $\forall\varphi\,\forall\psi\ \Big([\alpha]^{\Lambda}_{T}\varphi \wedge [\alpha]^{\Lambda}_{T}\psi \ \ \leftrightarrow \ \ [\alpha]^{\Lambda}_{T}(\varphi\wedge\psi)\Big)$. We fix some φ and ψ and assume $[\alpha]\varphi$ and $[\alpha]\psi$. We consider two cases. In the easy case, we have that at least one of $\Box\varphi$ or $\Box\psi$ holds in which case the result directly follows from Lemma 5.1.3.

In the remaining case, by the recursion equation for $[\alpha]$, we find ordinals $\beta, \beta' < \alpha$ and some formulas φ', ψ' so that $\langle\beta\rangle\varphi', \langle\beta'\rangle\psi', \Box\big(\langle\beta\rangle\varphi' \to \varphi\big)$, and $\Box\big(\langle\beta'\rangle\psi' \to \psi\big)$.

We first remark that w.l.o.g. we may assume $\beta' = \beta$. For, if, e.g., $\beta' < \beta$, then by Lemma 2.5.2. we see that $\langle\beta\rangle\top \ \to \ \big(\langle\beta'\rangle\psi' \ \leftrightarrow \ \langle\beta\rangle\langle\beta'\rangle\psi'\big)$ with $\langle\beta\rangle\varphi' \to \langle\beta\rangle\top$. Since we perform a transfinite *reflexive* induction, we also have our inductive hypotheses under a $\Box$ and in particular $\Box\big(\langle\beta\rangle\langle\beta'\rangle\psi' \to \langle\beta'\rangle\psi'\big)$. Thus, we see that $\langle\beta\rangle\langle\beta'\rangle\psi' \wedge \Box\big(\langle\beta\rangle\langle\beta'\rangle\psi' \to \psi\big)$ whence

$$\exists\psi''\ \Big(\langle\beta\rangle\psi'' \wedge \Box\big(\langle\beta\rangle\psi'' \to \psi\big)\Big).$$

So, we assume $\beta' = \beta < \alpha$, and by the inductive hypothesis (on Item 5), we find χ with $\langle\beta\rangle\chi \ \leftrightarrow \ \langle\beta\rangle\varphi' \wedge \langle\beta\rangle\psi'$ whence by the reflexive induction hypothesis also $\Box\big(\langle\beta\rangle\chi \ \leftrightarrow \ \langle\beta\rangle\varphi' \wedge \langle\beta\rangle\psi'\big)$. Consequently, we have that $\Box(\langle\beta\rangle\chi \to \varphi\wedge\psi)$ and we are done with the direction $[\alpha]\varphi \wedge [\alpha]\psi \to [\alpha](\varphi\wedge\psi)$. The other direction follows directly from Lemma 5.1 since $\Box\big((\varphi\wedge\psi)\to\varphi\big)$ and $\Box\big((\varphi\wedge\psi)\to\psi\big)$.

Item 2: $\forall\varphi\,\forall\psi\ \Big([\alpha]^{\Lambda}_{T}(\varphi\to\psi) \to ([\alpha]^{\Lambda}_{T}\varphi \to [\alpha]^{\Lambda}_{T}\psi)\Big)$. From the previous item we know that

$$[\alpha](\varphi\to\psi)\wedge[\alpha]\varphi \ \leftrightarrow \ [\alpha]\Big((\varphi\to\psi)\wedge\varphi\Big),$$

so that the result follows from Lemma 5.1.

Item 4: $\forall\varphi\ \Big([\alpha]^{\Lambda}_{T}\varphi \to [\alpha]^{\Lambda}_{T}[\alpha]^{\Lambda}_{T}\varphi\Big)$. While reasoning in T we assume $[\alpha]\varphi$ and only consider the non-trivial case. Thus, for some φ' and some $\beta \prec \alpha$ we get $\langle\beta\rangle\varphi'$ and $\Box(\langle\beta\rangle\varphi' \to \varphi)$. By negative introspection we get $[\alpha]\langle\beta\rangle\varphi'$. Since T is a 1-Münchhausen theory it proves some properties of the order $\prec$. In particular, from $\beta\prec\alpha$, we also get $[\alpha](\beta\prec\alpha)$. From $\Box(\langle\beta\rangle\varphi' \to \varphi)$ we obtain by applying successively provable Σ^0_1 completeness and monotonicity that $[\alpha]\Box(\langle\beta\rangle\varphi' \to \varphi)$. Since we already proved closure of the $[\alpha]$ predicate under conjunctions, we can collect all the information under the $[\alpha]$ and applying Lemma 5.3.4. we see that we have obtained $[\alpha][\alpha]\varphi$.

Item 5: $\forall\varphi\,\forall\psi\,\exists\chi\ \Big([\alpha]^{\Lambda}_{T}\varphi \vee [\alpha]^{\Lambda}_{T}\psi \ \ \leftrightarrow [\alpha]^{\Lambda}_{T}\chi\Big)$. Here we can now simply invoke Theorem 5.8. We stress a subtle issue here that the order in which we proved the items of this theorem is indeed quite important. In particular, at the point we have proved Items 1–4 for $[\alpha]$, we have established the soundness of $\mathsf{GLP}_{\alpha+1}$ which is needed among the assumptions of Theorem 5.8. ⊣

§6. Completeness of Münchhausen provability. In this section we shall prove that under some modest set of extra assumptions, we can obtain completeness of one-Münchhausen provability. Basically, this section consist of invoking a result from [18] and recasting it in our context. Let us first recall some definitions and results.

6.1. Uniform proof and provability predicates. The definitions and results from this subsection all come from [18] where an arithmetical completeness proof is given that is schematic in an abstract kind of provability predicates. A first step in defining these provability predicates consists of defining so-called Λ-*uniform proof and provability predicates over* T.

Definition 6.1. Let T be representable and Λ a linear order. Given a formula $\pi(c, \lambda, \phi)$, we introduce the notation $[c : \lambda]_\pi\phi = \pi(c, \lambda, \phi)$, as well as $[\lambda]_\pi\phi = \exists c[c : \lambda]_\pi\phi$. The dual notions $\langle c : \lambda\rangle_\pi\phi$ and $\langle\xi\rangle_\pi\phi$ are defined as $\neg\pi(c, \lambda, \neg\phi)$ and $\neg\exists c[c : \lambda]_\pi\neg\phi$ respectively.

A Λ-*uniform proof predicate over* T is a formula $\pi(c, \lambda, \phi)$ (with all free variables shown) satisfying

1. $T \vdash \mathrm{I}\Sigma^0_1(\pi)$;
2. $T \vdash \forall\lambda\forall\phi\ (\Box_T\phi \to [\lambda]_\pi\phi)$;
3. $T \vdash \forall\lambda\forall\phi\forall\psi\ \Big([\lambda]_\pi(\psi \to \phi) \wedge [\lambda]_\pi\psi \to [\lambda]_\pi\phi\Big)$;
4. $T \vdash \forall c\,\forall\lambda\,\forall\xi{\leq_\Lambda}\lambda\,\forall\phi\ \Big([c : \xi]\pi\phi \to [c : \lambda]\pi\phi\Big)$;
5. $T \vdash \forall c\,\forall\lambda\,\forall\phi\ \Big([c : \lambda]_\pi\phi \to [\lambda]\pi[\dot{c} : \dot{\lambda}]_\pi\dot{\phi}\Big)$;
6. $T \vdash \forall c\forall\lambda\,\forall\phi\ \Big(\langle c : \lambda\rangle_\pi\phi \to [\lambda]\pi\langle\dot{c} : \dot{\lambda}\rangle_\pi\dot{\phi}\Big)$;
7. $T \vdash \forall\lambda\,\forall\,\xi{<_\Lambda}\lambda\,\forall\phi\ \Big(\langle\xi\rangle_\pi\phi \to [\lambda]_\pi\langle\dot{\xi}\rangle_\pi\dot{\phi}\Big)$.

We say that π is *sound*[6] if, moreover, $\mathbb{N} \models \forall\lambda\forall\phi\ ([\lambda]_\pi\phi \to \phi)$.

A formula $\hat{\pi}$ is a Λ-*uniform provability predicate* over T if $T \vdash \hat{\pi} \leftrightarrow \exists c\ \pi$, where π is a Λ-uniform proof predicate.

Moreover, the provability predicates are required to require a modicum of good behaviour as captured in the following definition.

Definition 6.2. Let π be a Λ-uniform proof predicate over a theory T. We say that π is *normalized* if it is provable in T that for every λ we have that every λ-derivable formula has infinitely many λ-derivations and, whenever $[c : \lambda]\pi\phi$ and $[c : \lambda]\pi\psi$, it follows that $\phi = \psi$; in other words, every derivation must be a derivation of a single formula.

Modal formulas are linked to arithmetical ones via an arithmetic interpretation.

Definition 6.3. An *arithmetic interpretation* is a function[7] $f : \mathbb{P} \to \mathcal{S}^1_\omega$.

If π is a Λ-uniform proof predicate over T, we denote by f_π the unique extension of f such that $f_\pi(p) = f(p)$ for every propositional variable p, $f_\pi(\bot) = \bot$, f_π commutes with Booleans, and $f_\pi([\lambda]\phi) = [\overline{\lambda}]_\pi f_\pi(\phi)$.

The following uniform completeness theorem is proven in [18, Theorem 10.2] and provides us with an easy way to prove completeness for our current interpretation.

Theorem 6.4. *If* Λ *is a computable linear order,* T *is any sound, representable theory extending* RCA_0, π *is a sound, normalized,* Λ-*uniform proof predicate over* T, *and* ϕ *is any* $\mathcal{L}_\Box$-*formula,* $\mathsf{GLP}_\Lambda \vdash \phi$ *if and only if, for every arithmetic interpretation* f, $T \vdash f_\pi(\phi)$.

[6] Observe that for π to be sound, we must have that T itself was already sound.

[7] By $\mathbb{P}$ we denote the set of propositional variables and by $\mathcal{S}^1_\omega$ we denote the set of Π^1_ω sentences.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

6.2. Arithmetical completeness for Münchhausen provability. We can now combine the results from this paper and the previous subsection to see that under some extra conditions we obtain arithmetical completeness for one-Münchhausen provability.

Theorem 6.5 Arithmetical completeness. *Let Λ be a computable linear order, and T is any sound, representable one-Münchhausen theory extending* RCA_0 *with corresponding provability predicate $[\alpha]_T{}^{\Lambda}\varphi$ so that $T \vdash \mathrm{I}\Sigma^0_1([\alpha]_T{}^{\Lambda}\varphi)$. We then have that $[\alpha]_T{}^{\Lambda}\varphi$ is a uniform provability predicate and in particular,*

$$\mathsf{GLP}_{\Lambda} \vdash \varphi \iff \forall * \ T \vdash \varphi^*.$$

Proof. As always, the $*$ in the statement of the theorem is understood to range over arithmetical interpretations that map propositional variables to arbitrary sentences, so that $*$ commutes with the Boolean connectives and each modal formula $[\alpha]\psi$ is mapped to $[\overline{\alpha}]_T{}^{\Lambda}\psi^*$.

From our provability predicate (omitting sub and superscripts) $[\alpha]\varphi$ we will define a proof predicate $\pi(c, \lambda, \phi)$ for which we will observe that over T it is a normalized uniform proof predicate so that provably $\exists c\, \pi(c, \lambda, \phi) \leftrightarrow [\lambda]\phi$. To this end we define a slight variation of Definition 5.6

$$\pi(c,\lambda,\phi) \ := \ c = \langle c_0, c_1\rangle \wedge \begin{cases} \Big(c_0 = 0 \quad \wedge \quad \mathsf{Proof}_T(c_1, \phi)\Big) & \vee \\ \Big(c_0 = 1 \quad \wedge \quad c_1 = \langle \xi, \psi, p\rangle \ \wedge \ \xi \prec \lambda \ \wedge \\ \qquad\qquad\qquad \langle\xi\rangle^{\Box}\psi \wedge \mathsf{Proof}_T(p, \langle\xi\rangle^{\Box}\psi \to \phi)\Big). \end{cases}$$

It is straightforward to see that, indeed, $T \vdash \exists c\, \pi(c, \lambda, \phi) \leftrightarrow [\lambda]\phi$. Since Proof_T is a normalized proof predicate, so is π. Thus, we should only check Properties 1–7 from Definition 6.1. Property 1 is one of the assumptions of the theorem and Properties 2, 3, and 7 follow directly from the arithmetical soundness of one-Münchhausen provability. Property 4 follows since T is a one-Münchhausen theory whence proves transitivity of $\prec$. Properties 5 and 6 are a direct consequence of the definition of π and the soundness of the one-Münchhausen provability predicate in very much the same spirit as the proof of Lemma 5.7. ⊣

§7. Some notes on the formalisation of one-Münchhausen provability. Throughout this paper we have been talking about Münchhausen provability predicates and proving all sorts of properties of them. The reserved reader may now question whether there exist one-Münchhausen theories with corresponding one-Münchhausen provability predicates at all. In this section we sketch how to formalize a Münchhausen provability predicate in second order arithmetic.

Just as in [18] we start our formalization by reserving a set parameter X where we will collect all the pairs $\langle\alpha, \varphi\rangle$ of ordinals α and formulas φ so that $[\alpha]\varphi$ holds. Next, we will write down a predicate that all and only the correct pairs $\langle\alpha, \varphi\rangle$ are in X. Thus, we write the recursion for one-Münchhausen provability replacing every occurrence of $[\alpha]\varphi$ by $\langle \alpha, \varphi\rangle \in X$ and consequently replacing $\langle\alpha\rangle\varphi$ by $\langle \alpha, \neg\varphi\rangle \notin X$. We define any set satisfying our predicate to be an 1–IMC for *Iterated one-Münchhausen Class*.

By naively doing so, a problem arises namely that we get occurrences of the set variable X under the regular provability predicate $\Box_T$. By using numerals we can speak under a box about numbers that 'live outside the box'. However, we do not have any syntactical artefact to denote arbitrary sets. A possibly way out here would be to resort to *oracle-provability* as introduced in [10]. Thus, for one-Münchhausen provability, the predicate would look something like:

$$\begin{aligned} &\mathsf{1\text{–}IMC}(X,\alpha) := \\ &\quad \forall\, \xi{\leq}\alpha\, \forall\varphi\Big[\langle \xi, \varphi\rangle \in X \quad\leftrightarrow\quad \Big(\Box_T\varphi \vee \exists\psi\,\exists\,\zeta{<}\xi\ \big(\langle\zeta,\neg\psi\rangle \notin X \wedge \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Box_{T|X}(\langle \zeta, \neg\psi\rangle \notin X \to \varphi)\big)\Big)\Big]. \end{aligned}$$

With such a predicate we can then define:

$$[\alpha]_{T,1}\varphi := \forall X\Big(\mathsf{1\text{–}IMC}(X,\alpha) \to \langle\alpha,\varphi\rangle \in X\Big).$$

However, it is not clear if such a predicate will satisfy the required recursive equation since the relation between oracle provability and regular provability is not yet entirely understood in all its details.

For these and other reasons we choose a different approach. We will anticipate that hopefully/probably the 1–IMC predicate will define a unique set. Then, under the box we can just use any set that satisfies $\mathsf{IMC}(X)$. Of course, the fixpoint theorem allows us to do so. In the formalisation of Münchhausen provability we will closely follow [18]. As such we allow ourselves to be rather sketchy and refer to [18] for the details.

Definition 7.1. We define the predicate $\mathsf{1\text{–}IMC}(X,\gamma)$ using the fixpoint theorem so that it satisfies (provably in ECA_0) the following recursion.

$$\begin{aligned} &\mathsf{1\text{–}IMC}\ \ (X,\gamma) \ \longleftrightarrow \\ &\quad \Big(\forall\,\alpha{\preceq}\gamma\ \forall\varphi\Big[\quad \langle\alpha,\varphi\rangle \in X \ \leftrightarrow \\ &\qquad\qquad \Box_U\varphi \vee \exists\,\beta{\prec}\alpha\exists\psi\Big(\langle\beta,\neg\psi\rangle \notin X \wedge \\ &\qquad\qquad\quad \Box_U\big[\exists X(\mathsf{1\text{–}IMC}(X,\dot\beta) \wedge \langle\dot\beta,\neg\dot\psi\rangle \notin X) \to \varphi\big]\Big)\Big]\Big). \end{aligned}$$

With this Iterated one-Münchhausen Class predicate we define our one-Münchhausen predicate as

$$[\alpha]_U\varphi := \forall X\Big(\mathsf{1\text{–}IMC}(X,\alpha) \to \langle\alpha,\varphi\rangle \in X\Big).$$

It is clear that our definition supposes that we fix an ordinal notation system for some ordinal Λ and that all our ordinal quantifications are restricted to this Λ. We observe that

$$\langle\alpha\rangle\varphi := \exists X\Big(\mathsf{1\text{–}IMC}(X,\alpha) \wedge \langle\alpha,\neg\varphi\rangle \notin X\Big).$$

Consequently we can rewrite the defining recursion for Iterated one-Münchhausen Classes as

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

$$\begin{aligned}\mathsf{1\text{–}IMC}\ (X,\gamma) \longleftrightarrow \\ \Big(\forall\,\alpha{\preceq}\gamma\ \forall\varphi\Big[\quad & \langle\alpha,\varphi\rangle\in X\ \leftrightarrow\\ & \Box_U\varphi\vee\exists\,\beta{\prec}\alpha\exists\psi\big(\langle\beta,\neg\psi\rangle\notin X\wedge\\ & \quad\Box_U\big[\langle\beta\rangle\psi\to\varphi\big]\big)\Big]\Big).\end{aligned}$$

It is clear that 1–IMC depends on the base theory U and on the ordinal representation Λ but for the sake of readability we suppress these dependencies in our notation. We remark that 1–IMC(X,γ) is of complexity Π^0_2 with free set variable X. Our predicate $[\alpha]\varphi$ has a universal quantifier ranging over all sets that are iterated Münchhausen classes. Of course, we would hope that indeed such classes are uniquely defined if they exists at all.

In order to express this, we will fix the following notation:

$$X\equiv_\alpha Y\ \ :=\ \ \forall\,\beta{\preceq}\alpha\,\forall\varphi\Big(\langle\beta,\varphi\rangle\in X\ \longleftrightarrow\ \langle\beta,\varphi\rangle\in Y\Big),$$

and

$$\exists^{\leq 1\restriction\alpha}X\ \mathsf{IMC}(X,\alpha)\ \ :=\ \ \forall X\,\forall Y\ \Big(\mathsf{IMC}(X,\alpha)\wedge\mathsf{IMC}(Y,\alpha)\ \longrightarrow\ X\equiv_\alpha Y\Big).$$

We can now state and prove a key ingredient in proving that our formalisation satisfies the defining recursion for Münchhausen provability.

Lemma 7.2. *Let U be a theory extending ECA_0. We have that*

$$ACA_0+\mathsf{wo}(\alpha)\vdash\forall\beta{\prec}\alpha\,\exists^{\leq 1\restriction\beta}X\ \mathsf{IMC}(X,\beta).$$

Proof. We prove by transfinite induction that $\mathsf{IMC}(X,\beta)\wedge\mathsf{IMC}(Y,\beta)\to X\equiv_\beta Y$ where X and Y are unbounded set variables. Note that this is an arithmetical formula so that ACA$_0$ can prove transfinite induction up to α for this formula since we assumed $\mathsf{wo}(\alpha)$. ⊣

Now that we have uniqueness we proceed as in [18, Theorem 4.3] to observe in Theorem 7.3 that we actually may perform transfinite induction for second order formulas as long as the second order formulas are restricted to the IMCs. As in [18] by $\mathbf{\Pi}^1_\omega\restriction\theta$ we denote the fragment $\mathbf{\Pi}^1_\omega$ of second order arithmetic where all second-order quantifiers are of the form $\forall X\ (\theta(X)\to\phi)$ or $\exists X\ (\theta(X)\wedge\phi)$.

Theorem 7.3. *Given a formula $\theta(X)\in\mathbf{\Pi}^1_\omega$,*

$$\mathrm{ACA}_0\vdash\forall\Lambda\ \Big(\exists^{\leq 1}X\ \theta(X)\wedge\mathtt{wo}(\Lambda)\to\mathtt{TI}(\Lambda,\mathbf{\Pi}^1_\omega\restriction\theta)\Big).$$

We are now ready to prove that our formalisation satisfies the required recursion.

Theorem 7.4. *Let T be any presentable theory extending ECA_0. We have*

$$\begin{aligned}ACA_0+\mathsf{wo}(\beta)+\exists X\,\mathsf{1\text{–}IMC}(X,\beta)\vdash\forall\,\alpha{\preceq}\beta\ \Big[[\alpha]_T\varphi\ \leftrightarrow\ \Box_T\varphi\vee\exists\psi\,\exists\gamma\\ \times\Big(\gamma\prec\alpha\wedge\ \langle\gamma\rangle_T\psi\ \wedge\Box_T\big(\langle\gamma\rangle_T\psi\ \to\ \varphi\big)\Big)\Big].\end{aligned}$$

Proof. By transfinite induction on α as in [18]. Note that we need the existence of a 1–IMC for the → direction. By Theorem 7.3 we have access to the transfinite induction in ACA$_0$ since we proved uniqueness for 1–IMC's. ⊣

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

We refer the reader to [18, Lemma 6.6] from where we can conclude that adding the assertion of the existence of a one-Münchhausen class to T yields a theory that is equi-consistent with T. Of course, in stronger theories like ATR_0 we can simply prove the existence of a one-Münchhausen class up to α for any well ordering α.

§8. Weakening the base theory: Münchhausen provability. In this paper we have introduced the notion of one-Münchhausen provability for which we have proven arithmetical sound and completeness. Furthermore, we have shown in Theorem 7.4 that the notion can be formalised in second order arithmetic. However, the theory where the formalisation takes place is quite strong. In particular, it requires a fair amount of transfinite induction. As pointed out, this proof theoretic strength is consequently also required in the object theory which is not desirable. Via various tricks, one can lower the required proof theoretic strength of the object and meta-theory. A first step in doing so is via the introduction of *Münchhausen provability*. Further tricks are presented and worked out in [26].

To define Münchhausen provability we will start out with a very similar but slightly different recursion equivalence:

$$\begin{aligned}[]
[\alpha]^{\boxtimes}_{T}\varphi \;\leftrightarrow\; \Box_T\varphi \vee \exists\sigma\,\exists\tau\,\Big(&|\sigma|=|\tau| \;\wedge\; \forall\, i{<}|\tau|\,\tau_i{\prec}\alpha \;\wedge\; \forall\, i{<}|\sigma|\,\langle\tau_i\rangle^{\boxtimes}_{T}\sigma(i)\\
&\wedge\; \Box_T\big(\forall\, i{<}|\sigma|\,\langle\tau_i\rangle^{\boxtimes}_{T}\sigma(i) \;\to\; \varphi\big)\Big).
\end{aligned} \tag{10}$$

In this recursive equivalence we understand that σ is a finite sequence of formulas with $|\sigma|$ denoting the length of the sequence and $\sigma(i)$ denoting the ith element of the sequence. Likewise, τ is understood as being a sequence of ordinals all bounded by α. We will write either $\tau(i)$ or τ_i for the ith element of τ. Moreover, $\langle\alpha\rangle^{\boxtimes}$ is as always to be read a shorthand for $\neg[\alpha]^{\boxtimes}\neg$.

One of the main complications in proving the arithmetical soundness of one-Münchhausen provability in the previous section was in the proof of the closure of provability under conjunctions, that is, $[\alpha]\varphi \wedge [\alpha]\psi \leftrightarrow [\alpha](\varphi\wedge\psi)$. The proof of this required a weak closure of consistency under conjunctions— $\forall\varphi,\psi\,\exists\chi\Big(\langle\alpha\rangle\varphi \wedge \langle\alpha\rangle\psi \leftrightarrow \langle\alpha\rangle\chi\Big)$—so that the conjunction of two oracle sentences could be conceived as a single oracle sentence. However, in the new recursive equivalence as we just defined in (10), the closure of oracles under conjunctions is built into the definition.

A further complication in proving the arithmetical soundness of one-Münchhausen provability in the previous sections was caused by the fact that weak closure under conjunctions of consistency needed to be verified under a box. This was obtained by requiring a fair amount of transfinite induction and by requiring that the object and meta-theory be equal. In this last section we shall see that these requirements can also be circumvented.

The defining equation (10) begs for a notational simplification. From now on, the Greek letter σ shall be reserved to denote sequences of formulas and the Greek letter τ shall be reserved to denote sequences of ordinals. As such, we settle upon the notational convention that $\tau \prec \alpha$ is short for $\forall\, i{<}|\tau|\,\tau_i{\prec}\alpha$ and $\langle\tau\rangle^{\boxtimes}_{T}\sigma$ is short for $|\sigma|=|\tau| \;\wedge\; \forall\, i{<}|\sigma|\,\langle\tau_i\rangle^{\boxtimes}_{T}\sigma(i)$. Since we shall require that provably $|\sigma|=|\tau| \to$

$\Box_T|\sigma| = |\tau|$, the defining recursion can be recasted as

$$[\alpha]^{\boxtimes}_{T}\varphi \ \leftrightarrow\ \Box_T\varphi \vee \exists\sigma\, \exists\, \tau{\prec}\alpha \ \Big(\langle\tau\rangle^{\boxtimes}_{T}\sigma \ \wedge\ \Box_T\big(\langle\tau\rangle^{\boxtimes}_{T}\sigma \ \to\ \varphi\big)\Big). \tag{11}$$

Although we still cannot prove that different predicates that provably satisfy (11) are provably equivalent, at least proving soundness of GLP_Λ for such predicates becomes an easy matter. Let us first define some important notions as before but now for Münchhausen provability instead of one-Münchhausen provability.

Definition 8.1. Let us call a theory T a Λ-Münchhausen theory whenever we can define a predicate $[\alpha]^{\boxtimes^{\Lambda}}_{T}$ so that T proves (11) together with

$T \vdash$ "$\prec$ is transitive, right-discrete, and has a minimal element, "
$T \vdash (\xi \prec \zeta) \to [\zeta]^{\boxtimes^{\Lambda}}_{T}(\xi \prec \zeta)$,
$T \vdash \neg(\xi \prec \zeta) \to [\zeta]^{\boxtimes^{\Lambda}}_{T}\neg(\xi \prec \zeta)$,
$\xi < \zeta < \Lambda$ implies $T \vdash \xi \prec \zeta$.

Moreover, it is understood that T has a simple coding machinery for finite sequence of objects so that the obvious facts about length and concatenation provably hold. For example, $T \vdash |\tau| = n \to \Box_T|\tau| = n$, etc.

In this case we call $[\alpha]^{\boxtimes^{\Lambda}}_{T}$ a $T(\Lambda)$-Münchhausen provability predicate.

When the theory T and the ordinal Λ are clear from the context, we shall simply speak of a Münchhausen theory and of a Münchhausen provability predicate. On occasion we might only mention the ordinal Λ or only the theory T and speak of, for example, a Λ-Münchhausen theory and a T-Münchhausen provability predicate respectively. As with one-Münchhausen provability we see that the interaction axioms become trivial to prove for any Münchhausen provability predicate. In what follows we will revisit and simplify the soundness proof.

Lemma 8.2. *Let T be a Λ-Münchhausen theory with corresponding predicate $[\alpha]^{\boxtimes^{\Lambda}}_{T}$. Omitting sub and superscripts, we have that*

1. $T \vdash \forall\alpha\, \forall\varphi\, \forall\, \beta{\prec}\alpha{\prec}\Lambda\ \big([\beta]^{\boxtimes}\varphi \to [\alpha]^{\boxtimes}\varphi\big)$,
2. $T \vdash \forall\alpha\, \forall\varphi\, \forall\, \beta{\prec}\alpha{\prec}\Lambda\ \big(\langle\beta\rangle^{\boxtimes}\varphi \to [\alpha]^{\boxtimes}\langle\beta\rangle^{\boxtimes}\varphi\big)$, *and more in general*
3. $T \vdash \forall\alpha\, \forall\sigma\, \forall\, \tau{\prec}\alpha{\prec}\Lambda\ \big(\langle\tau\rangle^{\boxtimes}\sigma \to [\alpha]^{\boxtimes}\langle\tau\rangle^{\boxtimes}\sigma\big)$.

Proof. The proof is straightforward and completely analogous to the proof of Lemma 5.3. Let us just shortly comment on the second item. So, we reason in T and pick a formula φ and ordinals α and β as indicated, assuming $\langle\beta\rangle^{\boxtimes}\varphi$. We now consider the sequence σ_φ of length 1 whose only element is the formula φ. Likewise, we consider the sequence τ_β of length 1 whose only element is the ordinal β. Clearly, $T \vdash \langle\tau_\beta\rangle^{\boxtimes}\sigma_\varphi \to \langle\beta\rangle^{\boxtimes}\varphi$ so that $[\alpha]^{\boxtimes}\langle\beta\rangle^{\boxtimes}\varphi$ follows. ⊣

Contrary to the case of 1-Münchhausen provability it becomes now an easy exercise to see that each (internally quantified) provability predicate satisfies the distribution axioms for the basic modal logic **K**. Moreover, necessitation is also a routine matter. Before we prove this, we first need a technical easy lemma similar to Lemma 5.1 whose proof is immediate.

Lemma 8.3. *Let T be a Λ-Münchhausen theory with corresponding predicate $[\alpha]^{\boxtimes^{\Lambda}}_{T}$. Again, omitting sub and superscripts, we have that*

$$U \vdash \forall\, \alpha \prec \Lambda\, \forall\, \varphi, \psi, \chi\, \Big([\alpha]^{\boxtimes}\psi \wedge \Box\varphi \wedge \Box(\varphi \wedge \psi \to \xi) \;\to\; [\alpha]^{\boxtimes}\xi \Big).$$

With this technical lemma at hand it becomes very easy to see that each Münchhausen provability predicate $[\alpha]^{\boxtimes^{\Lambda}}_{T}$ defines a normal[8] modal logic.

Lemma 8.4. *Let T be a Λ-Münchhausen theory with corresponding predicate $[\alpha]^{\boxtimes^{\Lambda}}_{T}$. Again, omitting sub and superscripts, we have that*

1. $T \vdash \forall\, \alpha \prec \Lambda\, \forall \varphi, \forall \psi\, \Big([\alpha]^{\boxtimes}(\varphi \to \psi) \;\to\; \big([\alpha]^{\boxtimes}\varphi \to [\alpha]^{\boxtimes}\psi\big)\Big)$, *and*
2. *for any ordinal α below Λ, if $T \vdash \varphi$, then $T \vdash [\alpha]^{\boxtimes}\varphi$.*

Proof. The proof of the second item is easy and identical to the proof Lemma 5.4. It is in the first item where we see that working with sequences of formulas instead of formulas in our oracles is essential. So, let us reason in T and fix α and φ as stated. We assume $[\alpha]^{\boxtimes}(\varphi \to \psi)$ and $[\alpha]^{\boxtimes}\varphi$ and need to prove $[\alpha]^{\boxtimes}\psi$.

The case that both $\Box(\varphi \to \psi)$ and $\Box\varphi$ hold is trivial and in case one of them holds, Lemma 8.3 provides a proof.

So, in the remaining and only non-trivial case, we find two pairs of sequences σ_{φ} with τ_{φ} and $\sigma_{\varphi\to\psi}$ with $\tau_{\varphi\to\psi}$ so that $\tau_{\varphi} \prec \alpha \wedge \langle\tau_{\varphi}\rangle\sigma_{\varphi} \;\wedge\; \Box(\langle\tau_{\varphi}\rangle\sigma_{\varphi} \to \varphi)$ and also $\tau_{\varphi\to\psi} \prec \alpha \wedge \langle\tau_{\varphi\to\psi}\rangle\sigma_{\varphi\to\psi} \;\wedge\; \Box\Big(\langle\tau_{\varphi\to\psi}\rangle\sigma_{\varphi\to\psi} \to (\varphi \to \psi)\Big)$. We now consider the concatenation $\tau_{\varphi} \star \tau_{\varphi\to\psi}$ of both τ-sequences and likewise $\sigma_{\varphi} \star \sigma_{\varphi\to\psi}$ denotes the concatenation of both σ-sequences. Clearly, we have $|\tau_{\varphi} \star \tau_{\varphi\to\psi}| = |\sigma_{\varphi} \star \sigma_{\varphi\to\psi}|$ and $\tau_{\varphi} \star \tau_{\varphi\to\psi} \prec \alpha$. Likewise, from our assumptions it is easy to observe that $\langle\tau_{\varphi} \star \tau_{\varphi\to\psi}\rangle\sigma_{\varphi} \star \sigma_{\varphi\to\psi}$ and $\Box\Big(\langle\tau_{\varphi} \star \tau_{\varphi\to\psi}\rangle\sigma_{\varphi} \star \sigma_{\varphi\to\psi} \to \psi\Big)$ so that indeed $[\alpha]^{\boxtimes}\psi$. ⊣

As a consequence of our previous lemmas, we know that all reasoning of the modal logic **K** can be applied to any Münchhausen provability predicate. We now turn to the transitivity axiom to conclude that each predicate $[\alpha]^{\boxtimes}$ actually is sound for **K4**. Before proving this, we need one easy technical observation.

Lemma 8.5. *Let T be a Λ-Münchhausen theory with corresponding predicate $[\alpha]^{\boxtimes}$. We have that*

$$T \vdash \exists x\, [\alpha]^{\boxtimes}\varphi(\dot{x}) \;\to\; [\alpha]^{\boxtimes}\exists x\, \varphi(x).$$

Proof. We reason in T and assume that for some x we gave $[\alpha]^{\boxtimes}\varphi(\dot{x})$. Thus, for some (possibly empty) σ and some ordinal β (less than α in case σ is non-empty) we have $\langle\alpha\rangle^{\boxtimes}\sigma$ and $\Box\big(\langle\beta\rangle^{\boxtimes}\sigma \to \varphi(\dot{x})\big)$ whence also $\Box\big(\langle\beta\rangle^{\boxtimes}\sigma \to \exists x\varphi(x)\big)$ as was to be shown. ⊣

We can now prove the soundness of the transitivity axiom.

[8] It is in this lemma that we see that working with a single β would not have worked directly. That is, if we had defined $[\alpha]^{\boxtimes}_{T}\varphi \;\leftrightarrow\; \Box_{T}\varphi \vee \exists\sigma\,\exists\,\beta \prec \alpha\, \Big(\forall\, i < |\sigma|\, \langle\beta\rangle^{\boxtimes}_{T}\sigma(i) \;\wedge\; \Box_{T}\big(\forall\, i < |\sigma|\, \langle\beta\rangle^{\boxtimes}_{T}\sigma(i) \;\to\; \varphi\big)\Big)$. The distributivity axiom can then only be proved if we can work with the largest consistency statement. Thus, something like Lemma 2.5.2. should be available. For that, the soundness of GLP_{β} would be needed and we are back at the transfinite induction template again.

Lemma 8.6. *Let T be a Λ-Münchhausen theory with corresponding predicate $[\alpha]^{\boxtimes}$. We have that*

$$T \vdash \forall\, \alpha{\prec}\Lambda\, \forall\, \varphi\, \Big([\alpha]^{\boxtimes}\varphi \;\to\; [\alpha]^{\boxtimes}[\alpha]^{\boxtimes}\varphi \Big).$$

Proof. The proof is very similar to Item 4 of Theorem 5.9 but now, there is no need for induction since we already know our predicate to be sound for **K** reasoning. Thus, we reason in T, fix some ordinal $\alpha \prec \Lambda$ and formula φ, and assume $[\alpha]^{\boxtimes}\varphi$. Now either $\Box\varphi$ or there is some sequence of ordinals $\tau \prec \alpha$ and sequence σ so that $\langle\beta\rangle^{\boxtimes}\sigma$ and $\Box\Big(\langle\beta\rangle^{\boxtimes}\sigma \to \varphi\Big)$. In the first case, we get from $\Box\varphi$ that $\Box\Box\varphi$ whence by applying monotonicity twice that $[\alpha]^{\boxtimes}[\alpha]^{\boxtimes}\varphi$. Thus we focus on the second case and fix a particular sequences τ and σ so that

1. $\tau \prec \alpha$;
2. $\langle\tau\rangle^{\boxtimes}\sigma$;
3. $\Box\Big(\langle\tau\rangle^{\boxtimes}\sigma \to \varphi\Big)$.

From the first item, we get by assumptions on Münchhausen theories that $[\alpha]^{\boxtimes}(\tau \prec \alpha)$. From the second item we get by negative introspection that $[\alpha]^{\boxtimes}\langle\tau\rangle^{\boxtimes}\sigma$. From the third item we get $\Box\Box\Big(\langle\tau\rangle^{\boxtimes}\sigma \to \varphi\Big)$ whence $[\alpha]^{\boxtimes}\Box\Big(\langle\tau\rangle^{\boxtimes}\sigma \to \varphi\Big)$. Collecting these three consequences and applying provable closure of provability under conjunctions we obtain

$$\exists\sigma\exists\tau\; [\alpha]^{\boxtimes}\Big(\tau{\prec}\alpha \;\wedge\; \langle\tau\rangle^{\boxtimes}\sigma \;\wedge\; \Box\big(\langle\tau\rangle^{\boxtimes}\sigma \to \varphi\big)\Big),$$

so that by Lemma 8.5 we conclude

$$[\alpha]^{\boxtimes}\exists\sigma\; \exists\tau{\prec}\alpha\; \Big(\langle\tau\rangle^{\boxtimes}\sigma \;\wedge\; \Box\big(\langle\tau\rangle^{\boxtimes}\sigma \to \varphi\big)\Big),$$

which implies $[\alpha]^{\boxtimes}[\alpha]^{\boxtimes}\varphi$ as was to be shown. ⊣

In the light of Lemma 5.5 we may now conclude arithmetical soundness for GLP_Λ for Münchhausen provability.

Theorem 8.7. *Let T be a Λ-Münchhausen theory and let $[\alpha]^{\boxtimes}_T$ be a corresponding Münchhausen provability predicate. Then, GLP_Λ is sound for T when the $[\alpha]$-modalities $(\alpha \prec \Lambda)$ are interpreted as $[\alpha]^{\boxtimes}_T$.*

Proof. As always we prove by induction on a GLP_Λ proof that if $\mathsf{GLP}_\Lambda \vdash \varphi$, then for any arithmetical realization $*$ we have that $T \vdash \varphi^*$. ⊣

It is clear how the completeness proof and formalisation can be adapted to the new provability notion. Actually, it seems that in a sense Münchhausen provability is more fundamental than one-Münchhausen provability. We have chosen to start this paper with one-Münchhausen provability instead for two reasons. Firstly, the defining recursion for one-Münchhausen provability is slightly easier and more perspicuous. But secondly, it is important to be aware of the tension between provable properties and provable provable properties in the notion of one-Münchhausen provability and how this tension can be mitigated via transfinite reflexive induction.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

Acknowledgments. We would like to thank the anonymous referee for his report which contained various excellent suggestions to improve the presentation of the paper. More seriously, the referee spotted a genuine error in an earlier draft of this paper. Not only did the referee spot the error but also did he provide a suggestion on how to repair it. Indeed, we were able to repair the proof following the suggestions. Independently, the error was also pointed out to me by David Fernández Duque to whom I am most grateful for his patience to wait for this paper to come out. Further, I would like to thank Lev Beklemishev, Volodya Shavrukov, Albert Visser, and the attendants of my *Seminari Cuc*. I am grateful to Ana Borges for asking the right questions and suggesting improvements on notation.

REFERENCES

[1] J. P. Aguilera and D. Fernández-Duque, *Strong completeness of provability logic for ordinal spaces*, this Journal, vol. 82 (2017), no. 2, pp. 608–628.

[2] J. Bagaria, M. Magidor, and H. Sakai, *Reflection and indescribability in the constructible universe*. ***Israel Journal of Mathematics***, vol. 208 (2015), no. 1, pp. 1–11.

[3] L. D. Beklemishev, *Proof-theoretic analysis by iterated reflection*. ***Archive for Mathematical Logic***, vol. 42 (2003), pp. 515–552.

[4] ———, *Provability algebras and proof-theoretic ordinals, I*. ***Annals of Pure and Applied Logic***, vol. 128 (2004), pp. 103–124.

[5] ———, *Reflection principles and provability algebras in formal arithmetic*. ***Uspekhi Matematicheskikh Nauk***, vol. 60 (2005), no. 2, pp. 3–78 (in Russian). English translation in: ***Russian Mathematical Surveys***, vol. 60 (2005), no. 2, pp. 197–268.

[6] ———, *Veblen hierarchy in the context of provability algebras*, ***Proceedings of the Twelfth International Congress of Logic, Methodology and Philosophy of Science*** (P. Hájek, L. Valdés-Villanueva, and D. Westerståhl, editors), Kings College Publications, London, 2005, pp. 65–78.

[7] L. D. Beklemishev, D. Fernández-Duque, and J.J. Joosten, *On provability logics with linearly ordered modalities*. ***Studia Logica***, vol. 102 (2014), pp. 541–566.

[8] L. D. Beklemishev and D. Gabelaia, *Topological completeness of the provability logic*GLP. ***Annals of Pure and Applied Logic***, vol. 164 (2013), no. 12, pp. 1201–1223.

[9] L. D. Beklemishev and F. N. Pakhomov, *Reflection algebras and conservation results for theories of iterated truth*, preprint, 2019, arXiv:1908.10302.

[10] A. Cordón Franco, D. Fernández-Duque, J.J. Joosten, and F. Lara Martín, *Predicativity through transfinite reflection*, this Journal, vol. 82 (2017), no. 3, pp. 787–808.

[11] D. Fernández-Duque, *The polytopologies of transfinite provability logic*. ***Archive for Mathematical Logic***, vol. 53 (2014), no. 3–4, pp. 385–431.

[12] ———, *Impredicative consistency and reflection*, preprint, 2017, arXiv:1509.04547.

[13] ———, *Worms and spiders*: *Reflection calculi and ordinal notation systems*. ***Journal of Applied Logics—IfCoLoG Journal of Logics and Their Applications***, vol. 4 (2017), no. 10, pp. 3277–3356.

[14] D. Fernández-Duque and D. Hermo Reyes, *A self-contained provability calculus for*, ***Proceedings of the 26th International Workshop on Logic, Language, Information, and Computation, Wollic 2019,*** (R. Iemhoff, M. Moortgat, and R. J. G. B. de Queiroz, editors), Lecture Notes in Computer Science, vol. 11541, Springer, Berlin, 2019, pp. 195–207.

[15] D. Fernández-Duque and J.J. Joosten, *Kripke models of transfinite provability logic*, ***Advances in Modal Logic*** (T. Bolander, T. Braüner, S. Ghilardi, and L. Moss, editors), vol. 9, College Publications, London, 2012, pp. 185–199.

[16] ———, *Models of transfinite provability logics*, this Journal, vol. 78 (2013), no. 2, pp. 543–561.

[17] ———, *Well-orders in the transfinite Japaridze algebra*. ***Logic Journal of the IGPL***, vol. 22 (2014), no. 6, pp. 933–963.

[18] ———, *The omega-rule interpretation of transfinite provability logic*. ***Annals of Pure and Applied Logic***, vol. 169 (2018), no. 4, pp. 333–371.

[19] P. Hájek and P. Pudlák, ***Metamathematics of First Order Arithmetic***, Springer, Berlin, 1993.

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press

[20] T.F. Icard, III, *A topological study of the closed fragment of* GLP. ***Journal of Logic and Computation***, vol. 21 (2011), pp. 683–696.

[21] K.N. Ignatiev, *On strong provability predicates and the associated modal logics*, this Journal, vol. 58 (1993), pp. 249–290.

[22] G. Japaridze, *The polymodal provability logic*, ***Intensional Logics and Logical Structure of Theories: Material from the Fourth Soviet-Finnish Symposium on Logic***, Metsniereba, Telaviv, 1988 (in Russian).

[23] J. J. Joosten, *Π^0_1-ordinal analysis beyond first-order arithmetic*. ***Mathematical Communications***, vol. 18 (2013), pp. 109–121.

[24] ———, *Turing jumps through provability*, ***Evolving Computability—11th Conference on Computability in Europe, CiE 2015,*** (A. Beckmann, V. Mitrana, and M. I. Soskova, editors), Lecture Notes in Computer Science, vol. 9136, Springer, New York, 2015, pp. 216–225.

[25] ———, *Turing–Taylor expansions of arithmetic theories*. ***Studia Logica***, vol. 104 (2016), pp. 1225–1243.

[26] ———, *Transfinite Turing jumps through provability*, preprint, 2021, arXiv (soon).

[27] G. Kreisel and A. Lévy, *Reflection principles and their use for establishing the complexity of axiomatic systems*. ***Zeitschrift für mathematische Logik und Grundlagen der Mathematik***, vol. 14 (1968), pp. 97–142.

[28] U. R. Schmerl, *A fine structure generated by reflection formulas over primitive recursive arithmetic*, ***Logic Colloquium '78 (Mons, 1978)***, Studies in Logic and the Foundations of Mathematics, vol. 97, North-Holland, Amsterdam, 1979, pp. 335–350.

[29] I. Shapirovsky, *PSPACE-decidability of Japaridze's polymodal logic*, ***Advances in Modal Logic*** (C. Areces and R. Goldblatt, editors), vol. 8, College Publications, London, 2008, pp. 289–304.

[30] S.G. Simpson, ***Subsystems of Second Order Arithmetic***, Cambridge University Press, New York, 2009.

[31] R.M. Solovay, *Provability interpretations of modal logic*. ***Israel Journal of Mathematics***, vol. 28 (1976), pp. 33–71.

[32] A. Visser, *Faith & falsity: A study of faithful interpretations and false Σ^0_1-sentences*. ***Annals of Pure and Applied Logic***, vol. 131 (2005), no. 1–3, pp. 103–131.

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF BARCELONA
BARCELONA, SPAIN
E-mail: jjoosten@ub.edu
URL: http://www.phil.uu.nl/˜jjoosten

https://doi.org/10.1017/jsl.2021.44 Published online by Cambridge University Press