Predicativity of the Mahlo Universe in Type Theory

Peter Dybjer[1] and Anton Setzer[2]

[1] Dept. of Computer Science and Engineering,
Chalmers University of Technology and University of Gothenburg, Sweden
http://www.cse.chalmers.se/~peterd/
[2]Dept. of Computer Science, Swansea University, UK
https://www.swansea.ac.uk/staff/a.g.setzer/ https://csetzer.github.io/

Abstract

We present a constructive, predicative justification of Setzer's Mahlo universe in type theory. Our approach is closely related to Kahle and Setzer's axiomatization of an extended predicative Mahlo universe in Feferman's Explicit Mathematics, a framework with direct access to the collection of partial functions. However, we here work directly in Martin-Löf type theory, a theory where all functions are total. We analyze Setzer's original version of the Mahlo universe, as opposed to the version derived in previous work through the modeling of Explicit Mathematics with an extended predicative Mahlo universe in type theory. We provide meaning explanations which extend and adapt those in Martin-Löf's article *Constructive Mathematics and Computer Programming* to cover the proof-theoretically much stronger Mahlo universe. In this way, we aim to resolve a long-standing discussion on whether the Mahlo universe is predicatively justifiable. We also construct four models in set-theoretic metalanguage that provide mathematical support for the meaning explanations. We prove that they are indeed models of the type theory in question and discuss their relationship to the meaning explanations. This research is a substantial step in the predicative justification of the consistency of proof-theoretically strong theories. Our work thus contributes to a revised Hilbert program, aiming to overcome the limitations implied by Gödel's incompleteness theorem, namely that there is no mathematical proof of the consistency of mathematical theories based on finitary methods, except for very weak theories.

1 Introduction

In this article we commemorate that it is now more than 100 years ago since the publication of Hermann Weyl's *Das Kontinuum* [48] - the first systematic development of predicative mathematics. Weyl introduced the notion of predicativity given the natural numbers. In this paper we discuss an extended notion of predicativity including Mahlo notions in type theory.

Martin-Löf's first published paper on type theory was entitled "An intuitionistic type theory: predicative part" [28]. This theory had an infinite hierarchy of universes. Its proof-theoretic strength was determined to be Γ_0 [20, 21], the limit of predicativity (given the natural numbers) in Feferman and Schütte's sense [47, 25, 24, 18, 43, 42]. In his article *Constructive Mathematics and Computer Programming* [29] Martin-Löf added W-types, and the theory became impredicative in the sense of Feferman and Schütte. Nevertheless, Martin-Löf still considered the theory predicative in an extended sense. A reason for this was that the theory was provided with *meaning explanations* that suggest how the types and terms of the theory are built up from below. They explain how the objects of the theory are trees that are built by a well-founded process of repeated lazy evaluation of expressions to canonical form.

Higher universes in type theory. Martin-Löf type theory was later extended with several higher universe constructions, such as Palmgren's universe operators, the super universe [33], Rathjen's superjump universes [40], and Setzer's Mahlo universe [44]. All these extensions were intended to be constructive and predictive in the sense of Martin-Löf's meaning explanations. However, the predicativity of the Mahlo universe was not so clear, especially after Palmgren [33] discovered that adding a natural elimination rule for it led to an inconsistency. Maybe Mahlo is a natural limit of Martin-Löf's extended predicativity as we conjectured in our paper on a finite axiomatisation of inductive-recursive definitions [14]?

A universe in type theory is a type closed under all standard type formers, such as $\Pi, \Sigma, 0, 1, 2, \mathrm{N}, \mathrm{W}$, and the identity type I. Universes can either be formulated à la Russell, where an element $A : \mathrm{U}$ is also a type A, or à la Tarski, where an element $a : \mathrm{U}$ is a "name" or "code" of a type A and there is a decoding map T such that $\mathrm{T}\, a = A$.

A super universe is a universe closed under an operator on families of sets that maps a universe $(\mathrm{U}_n, \mathrm{T}_n)$ to the next universe $(\mathrm{U}_{n+1}, \mathrm{T}_{n+1})$ in the hierarchy. One can then form an operator mapping a super universe to the next and form a super^2 universe closed under this operator. This process can be iterated, and thus one obtains super^n universes. More generally, one can define universes closed under arbitrary operators on families of sets. A Mahlo universe is a universe that contains all universes generated by family operators. Moreover, the latter are *subuniverses* of the Mahlo universe. One can show that these subuniverses arise as special cases of the inductive-recursive definitions in our theory **IR** [14]. This theory is formulated as an extension of Martin-Löf's logical framework [32], where there is a type Set of "sets" in Martin-Löf's sense: "to know a *set* is to know how the elements of the set are formed and how equal elements are formed", a phrase indicating that sets should be inductively (or inductive-recursively) generated. Therefore, we will refer to $\Pi, \Sigma, 0, 1, 2, \mathrm{N}, \mathrm{W}, \mathrm{I}$, etc., as *set formers* rather than type formers, when we work in this version of type theory. (We remark that Martin-Löf's notion of "set" is different from the notion of "h-set" in homotopy type theory.)

Let f be an operator on families of sets split into two components (f_0, f_1) where f_0 returns the index set and f_1 returns the family (see the paragraph on families of sets on p. 5 for full details of this notation). Then we can define a subuniverse $\mathrm{U}\, f_0\, f_1 : \mathrm{Set}$ with decoding $\mathrm{T}\, f_0\, f_1 : \mathrm{U}\, f_0\, f_1 \to \mathrm{Set}$ as an instance of an inductive-recursive definition in **IR**. In this way Set encodes Setzer's Mahlo universe [44]. We call it an *external* Mahlo universe to contrast it with the *internal* Mahlo universe that arises if we introduce a set $\mathsf{M} : \mathrm{Set}$ with the Mahlo property. This M goes beyond inductive-recursive definitions in **IR**.

When Martin-Löf extended his meaning explanations to the 1986 version based on a logical framework [32], he did not *not* stipulate that "to know a *type* is to know how the objects of the type are formed and how equal objects are formed". (We refer to Martin-Löf's Leiden lectures [26, 27] for a comprehensive account of the philosophical foundations of intuitionistic type theory with the distinction between types and sets.) The type Set is to be understood as "open" to extension with new inductive(-recursively) defined sets when we need them. Hence, it is not natural to add an elimination rule for it. In contrast to this, $\mathsf{M} : \mathrm{Set}$ is to be understood as "closed". Nevertheless, as Palmgren showed, adding a natural elimination rule for it leads to an inconsistency. This paradox may make us doubt that the internal Mahlo universe is a good predicative set according to Martin-Löf's conception.

Nevertheless, here we argue that the Mahlo universe is after all predicative and constructive by giving Martin-Löf style meaning explanations for it. Our argument can be applied both to the external and to the internal Mahlo universe, although we only discuss the somewhat simpler external version.

Plan of the paper. We begin by constructing a set-theoretic model of logical framework-based type theory with Set as a Mahlo universe and $\mathrm{U}\, f_0\, f_1$: Set with decoding $\mathrm{T}\, f_0\, f_1$: $\mathrm{U}\, f_0\, f_1 \to$ Set as subuniverses. This model is an adaptation of the model of **IR** [14], where we interpret the the type-theoretic function spaces as the sets of all set-theoretic functions. We work in classical set theory **ZFC** with a Mahlo cardinal M and an inaccessible cardinal I above it. (Note that we use the term "set" both for sets in Martin-Löf type theory and for sets in the set-theoretic metalanguage, so we hope this will not lead to confusion.) We interpret the collection of all types as V_{I} and Set as V_{M}. Let in set theory $\mathcal{F}am(V) = \{(X, Y) \mid X \in V, Y : X \to V\}$ be the families of sets in V. An operator on families of sets in the type theory is interpreted as a function $f : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})$. We then use the Mahlo property of M to show that there is an inaccessible cardinal $\kappa_f < \mathsf{M}$ such that $f : \mathcal{F}am(\mathrm{V}_{\kappa_f}) \to \mathcal{F}am(\mathrm{V}_{\kappa_f})$ and interpret the subuniverses $\mathrm{U}\, f_0\, f_1$ as $\mathcal{U}\, f_0\, f_1 = \mathrm{V}_{\kappa_f}$ à la Russell, that is, the decoding $\mathrm{T}\, f_0\, f_1$ is interpreted as the injection $\mathcal{T}\, f_0\, f_1 : \mathrm{V}_{\kappa_f} \hookrightarrow \mathrm{V}_{\mathsf{M}}$.

This first model is unnecessarily large. Therefore, we construct a second set-theoretic model where we interpret the inductive-recursively defined type-theoretic subuniverses $(\mathrm{U}\, f_0\, f_1, \mathrm{T}\, f_0\, f_1)$ in set theory in terms of inductively generated graphs $\mathcal{T}\, f_0\, f_1$ with domain $\mathcal{U}\, f_0\, f_1$ in the standard set-theoretic way following Allen [9]. However, in order to interpret Set as an inductively defined set $\mathcal{S}et$ we make use of the ideas behind Kahle and Setzer's extended predicative Mahlo universe [23, 16] in Feferman's theory of Explicit Mathematics [19]. The key point is that in order to add a subuniverse $\mathcal{U}\, f_0\, f_1$ to $\mathcal{S}et$ it suffices to require that the family operator f on families of sets is total on families over the subuniverse $\mathcal{U}\, f_0\, f_1$ itself. Although this may seem impredicative, we show that it results in an inductive definition of $\mathcal{S}et \subseteq \mathrm{V}_{\mathsf{M}}$. Moreover, we show that $\mathcal{T}\, f_0\, f_1 : \mathcal{U}\, f_0\, f_1 \to \mathcal{S}et$.

We also construct a third set-theoretic model. This is a variation of the second model that is closer to the model of the extended predicative Mahlo universe in Explicit Mathematics, where the function f ranges over arbitrary untyped terms denoting partial functions. To approximate this in set theory, we first replace $f : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ by arbitrary sets $f \in \mathrm{V}_{\mathsf{I}}$. Then we show that this can be further restricted to $f \in \mathrm{V}_{\mathsf{M}}$, and obtain as a small variant a fourth set theoretic model. Note that $\mathcal{S}et$ and therefore also the function space $\mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$ are not guaranteed to be elements of V_{M} and this model refers to local approximations of $\mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$. (We conjecture that $\mathcal{S}et \notin \mathrm{V}_{\mathsf{M}}$ provided M is the smallest Mahlo cardinal.)

The final step is to provide Martin-Löf style meaning explanations inspired by the second (and third and fourth) set-theoretic models.

The usual situation in type theory is that the meaning explanation for a type former is determined by the formation rule and the introduction rules, and the computation rules for the elimination constant are given by the equality rules. However, in the case of the Mahlo universe, this pattern is broken. If we take the formation rule for the subuniverses $\mathrm{U}\, f_0\, f_1$ as a type-checking condition (a *matching condition*), then we get a non-wellfounded type-checking process, because of Palmgren's paradox. (We remark that "type checking" here refers to the matching of canonical terms with canonical types in Martin-Löf's meaning explanations, and not to the type-checking of judgments in intensional type theory, as implemented in proof assistants.)

Instead, we let the second set-theoretic model suggest the type-checking conditions. As an example, we give one of the type-checking conditions for the judgment A : Set. If A has the canonical form $\mathrm{U}\, f_0\, f_1$, then we check whether

$$f_0\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x)) : \mathrm{Set}$$

in the context $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$, and

$$f_1\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x))\, y : \mathrm{Set}$$

in the context $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1, y : f_0\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x))$.

Note that we are only type-checking for arguments in the image of $\mathrm{T}\, f_0\, f_1 : \mathrm{U}\, f_0\, f_1 \to \mathrm{Set}$ and this avoids the circularity of taking $\mathrm{U}\, f_0\, f_1 : \mathrm{Set}$ as an argument for type-checking during the process of type checking $\mathrm{U}\, f_0\, f_1 : \mathrm{Set}$. Nevertheless, the formation rule for $\mathrm{U}\, f_0\, f_1$ can be justified on this basis. Here f_0 and f_1 are the two components of a function $f : \mathrm{Fam(Set)} \to \mathrm{Fam(Set)}$, see below for the precise definition.

In the conclusion and related work section we make some general remarks about the relationship between the meaning explanations and various mathematical models. We also discuss the relationship between our work and Rathjen's articles on the limits of Martin-Löf type theory [37, 39].

There is also an appendix where we show implementations in Agda of the internal Mahlo universe and Palmgren's paradox.

Mahlo universes in Explicit Mathematics. Let us explain the relationship between the present article and previous work on the extended predicative Mahlo universe in Explicit Mathematics. The aim of Kahle and Setzer [23] was to introduce the Mahlo universe "from below" so that the definition has an extended predicative character. While the subuniverses of the Mahlo universe in type theory are defined for arbitrary *total* functions on families of sets, Kahle and Setzer define them for arbitrary *partial* functions. However, while the latter are not directly available in Martin-Löf type theory, they are available in Explicit Mathematics, a framework developed by Solomon Feferman [19] and further explored by Gerhard Jäger and coworkers. Kahle and Setzer extended Explicit Mathematics with axioms for an *extended predicative Mahlo universe*.

This approach was further investigated by the authors [16] with the aim of providing a more explicit link to Martin-Löf's conception of predicativity. To this end, a model of Explicit Mathematics with an extended predicative Mahlo universe was implemented in an extension of Martin-Löf type theory. The set of untyped terms of Explicit Mathematics was implemented as a set in type theory, and on this basis the other basic notions were implemented. The extended predicative Mahlo universe was then implemented by a certain strong indexed inductive-recursive definition that goes beyond the authors' theories of indexed inductive-recursive definitions [12, 15]. Finally, we argued for the predicativity of our extension by providing Martin-Löf style meaning explanations for it.

When we model Explicit Mathematics with an extended predicative Mahlo universe in type theory, we get something rather different from Setzer's original Mahlo universe in type theory. In this article, we return to the original formulation and show how to provide direct meaning explanations for it.

Remarks on the notation. As already mentioned, a possible source of confusion is that we use the term "set" to denote both sets (in Martin-Löf's sense) in our version $\mathbf{TT}^{\mathsf{M}}$ of type theory and as sets in the set-theoretic metalanguage. We distinguish notationally between the type Set in type theory and its set-theoretic interpretation $\mathcal{S}et$ and between type-theoretic families in $\mathrm{Fam}(V)$ and set-theoretic ones in $\mathcal{F}am(V)$. Similarly, we distinguish between type-theoretic subuniverses $(\mathrm{U}\, f_0\, f_1, \mathrm{T}\, f_0\, f_1)$ and their set-theoretic interpretations $(\mathcal{U}\, f_0\, f_1, \mathcal{T}\, f_0\, f_1)$. Moreover, we distinguish notationally between the type-theoretic Cartesian products (dependent function spaces $(x : \sigma) \to \tau$) and disjoint unions (dependent products $\Sigma x : \sigma.\tau$) and their

interpretations in terms of $\prod$ and $\sum$ in set theory. Apart from these distinctions, we usually overload notation and use identical notation for type-theoretic concepts and their set-theoretic interpretation. For example, function application is always written $f\,a$ in $\mathbf{TT}^{\mathsf{M}}$ and often also in set theory, although sometimes we write $f(a)$ in set theory. Similarly, we write $f : A \to B$ or $g : \prod_{X:A} B(x)$ instead of $f \in (A \to B)$ and $g \in \prod_{X \in A} B(x)$ in set theory.

Notations for families of sets. In type theory, by $(X, Y) : \mathrm{Fam(Set)}$ we mean $X : \mathrm{Set}$, $Y : X \to \mathrm{Set}$. By $f : \mathrm{Fam(Set)} \to \mathrm{Fam(Set)}$ we mean the two components

$$\begin{array}{lcl} f_0 & : & (X : \mathrm{Set}) \to (Y : X \to \mathrm{Set}) \to \mathrm{Set} \\ f_1 & : & (X : \mathrm{Set}) \to (Y : X \to \mathrm{Set}) \to f_0\ X\ Y \to \mathrm{Set} \end{array}$$

In set theory we have already defined $\mathcal{F}am(V)$ above. After introducing $f : \mathcal{F}am(V) \to \mathcal{F}am(V)$, we use f_0, f_1 for the two set-theoretic components of that function, that is,

$$\begin{array}{lcllcl} f_0 & : & \prod_{X:V} \prod_{Y:X\to V} V & f_0\ X\ Y & = & \pi_0(f\ (X,Y)) \\ f_1 & : & \prod_{X:V} \prod_{Y:X\to V} (f_0\ X\ Y \to V) & f_1\ X\ Y\ Z & = & \pi_1(f\ (X,Y))\ Z \end{array}$$

Set-theoretic universes à la Tarski are given as $(\mathcal{U}, \mathcal{T}) \in \mathcal{F}am(V)$ for some V. When referring to a universe, we often just refer to $\mathcal{U}$ and leave $\mathcal{T}$ implicit. We also define $\mathcal{TF}am(\mathcal{U}) := \mathcal{TF}am(\mathcal{U}, \mathcal{T}) := \{(x, y) \mid x \in \mathcal{U} \wedge y : \mathcal{T}\,x \to \mathcal{U}\}$. If $f : \mathcal{TF}am(\mathcal{U}) \to \mathcal{TF}am(\mathcal{U})$ then we define its two components

$$\begin{array}{lcllcl} f_0 & : & \prod_{x:\mathcal{U}} \prod_{y:\mathcal{T}x\to\mathcal{U}} \mathcal{U} & f_0\ x\ y & = & \pi_0(f\ (x,y)) \\ f_1 & : & \prod_{x:\mathcal{U}} \prod_{y:\mathcal{T}x\to\mathcal{U}} (\mathcal{T}\ (f_0\ x\ y) \to \mathcal{U}) & f_1\ x\ y\ z & = & \pi_1(f\ (x,y))\ z \end{array}$$

In type theory, universes à la Tarski are given as $(U, T) : \mathrm{Fam(Set)}$, where we again often refer to the universe as U and leave T implicit. By $(x, y) : \mathrm{TFam}(U, T)$ or $(x, y) : \mathrm{TFam}(U)$ we mean $x : U$ and $y : T\,x \to U$, and by $f : \mathrm{TFam}(U) \to \mathrm{TFam}(U)$ we mean its two components

$$\begin{array}{lcl} f_0 & : & (x : U) \to (y : T\,x \to U) \to U \\ f_1 & : & (x : U) \to (y : T\,x \to U) \to T\ (f_0\ x\ y) \to U \end{array}$$

Agda code and Git repository. All display style Agda code has been type-checked in Agda and directly imported via the literal Agda framework into this paper. The Agda code is available in the Git repository [17]. Here the reader can also find full definitions of the standard set formers of Martin-Löf type theory and the closure of the external Mahlo universe under those. Moreover, the repository also includes an html version which doesn't require installation of Agda.

2 Type theory with an external Mahlo universe

We work in a version of Martin-Löf type theory based on a logical framework (see, for example, [32, 14]). This logical framework is a typed lambda calculus with dependent function types written $(x : \sigma) \to \tau$. Moreover, it has a type Set of sets in Martin-Löf's sense, and each object $A : \mathrm{Set}$ is also a type A.

On this basis, we can introduce formation, introduction, and elimination rules for the set formers by adding the respective constants with their types. The equality rules are represented by equations between expressions of the same type.

The proof assistant Agda [8] can be used for implementing this version of Martin-Löf type theory. For example, here is a definition of the set of natural numbers with the primitive recursion combinator R in Agda:

```
data ℕ : Set where
  O : ℕ
  s : ℕ → ℕ

R : {C : ℕ → Set} → C O → ((n : ℕ) → C n → C (s n)) → (c : ℕ) → C c
R d e O     = d
R d e (s n) = e n (R d e n)
```

(The curly braces in $\{C : \mathbb{N} \to \mathsf{Set}\}$ specify that C is an implicit argument.)

We now implement the inductive-recursive definition of the subuniverses $\mathrm{U}\, f_0\, f_1 : \mathsf{Set}$ with decodings $\mathrm{T}\, f_0\, f_1 : \mathrm{U}\, f_0\, f_1 \to \mathsf{Set}$ in Agda. Recall that such a subuniverse is closed under standard set formers such as $\Pi, \Sigma, 0, 1, 2, \mathrm{N}, \mathrm{W}$, and I, and also under arbitrary operators f on families of sets (with components f_0 and f_1). It has two constructors c_0 and c_1 that express closure under the two components f_0 and f_1, respectively. We omit the closure rules under the standard set formers and only display the Agda code for closure under f.

```
data U (f₀ : (X₀ : Set) → (X₀ → Set) → Set)
       (f₁ : (X₀ : Set) → (X₁ : X₀ → Set) → f₀ X₀ X₁ → Set) : Set where
  c₀ : (x₀ : U f₀ f₁)
     → (T f₀ f₁ x₀ → U f₀ f₁)
     → U f₀ f₁
  c₁ : (x₀ : U f₀ f₁)
     → (x₁ : (T f₀ f₁ x₀ → U f₀ f₁))
     → T f₀ f₁ (c₀ x₀ x₁)
     → U f₀ f₁

T : (f₀ : ((X₀ : Set) → (X₀ → Set) → Set))
    (f₁ : ((X₀ : Set) → (X₁ : X₀ → Set) → f₀ X₀ X₁ → Set))
  → U f₀ f₁ → Set
T f₀ f₁ (c₀ x₀ x₁) = f₀ (T f₀ f₁ x₀) (λ z → T f₀ f₁ (x₁ z))
T f₀ f₁ (c₁ x₀ x₁ t) = f₁ (T f₀ f₁ x₀) (λ z → T f₀ f₁ (x₁ z)) t
```

In this way, the type Set implements the external Mahlo universe: it contains $\mathrm{U}\, f_0\, f_1$ and this is a subuniverse à la Tarski of Set with a decoding map $\mathrm{T}\, f_0\, f_1$.

We will refer to the basic type theory as **TT**. It consists of the following parts:

- Martin-Löf's logical framework, that is, dependent type theory with dependent function types, a type Set, and for each $A : \mathsf{Set}$ a type A of its elements.
- Constants and equations for the standard set formers: $\Pi, \Sigma, 0, 1, 2, \mathrm{N}, \mathrm{W}$, and I. We have shown the Agda code for N and leave it to the reader to define the others.

We then extend our theory with the following:

- Constants and equations for the subuniverse à la Tarski as shown in the Agda code above. These consist of the typing rules for the set former U with decoding T and the constructors c_0 and c_1 that are codes for the family operator (f_0, f_1) with their decoding equations. Moreover, there are constructors for codes for the standard set formers with their decoding equations, but these are not displayed in the Agda code above.

We call the resulting theory $\mathbf{TT}^{\mathsf{M}}$.

3 First set-theoretic model

In our article on the finite axiomatization of induction-recursion [14] we showed the consistency of the theory **IR** by constructing a model in classical set theory where function types are interpreted as full set-theoretic function spaces. We remark again that the theory presented in the previous section is a subtheory of **IR** and hence we can refer to the same model. However, to prepare for the second *extended predicative* model, we give an alternative presentation of such a set-theoretic model.

ZFC with one Mahlo cardinal and one inaccessible above has much more proof theoretic strength than what is actually needed. Setzer [46] created a model of the internal Mahlo universe in Kripke-Platek set theory with one recursively Mahlo ordinal and finitely many admissibles above KPM^+. Together with [44] this shows that the proof-theoretic strength of the internal Mahlo universe (which does not make use of the logical framework) is that of KPM^+. Note that KPM^+ is a slight extension of the theory KPM analysed by Rathjen [35, 36]. The type theory of the external Mahlo universe using the logical framework can be interpreted in the type theory of the external Mahlo universe and therefore also in KPM^+.

Since our aim here is to motivate the meaning explanations given in Section 6, we will work in the more familiar setting of ZFC with a strongly Mahlo cardinal M and a strongly inaccessible cardinal I above it. It should be possible with some extra work to construct variants of our models in weaker set theories such as KPM^+.

We recall some definitions.

Definition 3.1. *A cardinal I is strongly inaccessible iff it is transfinite, a strong limit, and regular:*

$$\aleph_0 < \mathsf{I} \qquad \frac{\alpha < \mathsf{I}}{2^\alpha < \mathsf{I}} \qquad \frac{\alpha < \mathsf{I} \quad \beta : \alpha \to \mathsf{I}}{\bigvee_{i<\alpha} \beta_i < \mathsf{I}}$$

We will use the fact that if I is strongly inaccessible, then V_I is closed under the interpretation of the standard type-theoretic set formers.

Definition 3.2. *A cardinal M is strongly Mahlo iff each normal (strictly monotone and continuous) function $h : \mathsf{M} \to \mathsf{M}$ has a strongly inaccessible fixed point κ_h.*

It follows that M is strongly inaccessible and $h : \kappa_h \to \kappa_h$. In the presence of GHC, strongly and weakly inaccessible and strongly and weakly Mahlo coincide, and in the sequel, we will only say "inaccessible" and "Mahlo".

Mahlo cwfs. Categorically, the version of type theory based on the logical framework is modelled by a category with families (cwf) [11] with extra structure. We only give an overview here and refer to [10, 22] for details.

A cwf has four components $(\mathcal{C}tx, \mathcal{H}om, \mathcal{T}ype, \mathcal{T}m)$, where $\mathcal{C}tx$ and $\mathcal{H}om$ denote the set of objects and the family of morphisms of the category of contexts, and $\mathcal{T}ype$ and $\mathcal{T}m$ are the components of the family-valued functor from the category of contexts to the category of families of terms indexed by types. The extra structure for the logical framework is a Π-structure for modelling dependent function types and a structure for modelling the type Set, such that for each set A there is a type of elements $\mathrm{El}(A)$. Moreover, we need extra structure for all the standard set formers $\Pi, \Sigma, 0, 1, 2, \mathrm{N}, \mathrm{W}, \mathrm{I}$ and the subuniverse set former U with decoding map T. This amounts to requiring one constant for each formation rule, one for each introduction rule and one for each elimination rule. These are subject to certain equations expressed by the

equality rules for the respective set formers. We call a cwf with all this extra structure, a *Mahlo cwf*.

We refer to Hofmann [22] for the interpretation of Martin-Löf type theory in cwfs with extra structure corresponding to the type theory. Hofmann only spells out the details for type theory with dependent function types. This is the crucial part of the interpretation of the formal system $\mathbf{TT}^{\mathsf{M}}$ in an arbitrary Mahlo cwf. We have not carried out the details of this generalisation, but we do not expect any difficulties.

The first model as a Mahlo cwf. Let M be Mahlo and $\mathsf{I} > \mathsf{M}$ be inaccessible:

- Let $\mathcal{C}tx = \mathrm{V}_{\mathsf{I}}$ and $\mathcal{H}om(\Delta, \Gamma) = \Delta \to \Gamma$, the set of functions from Δ to Γ.
- Let $\mathcal{T}ype = \mathrm{V}_{\mathsf{I}}, \mathcal{T}ype(\Gamma) = \Gamma \to \mathcal{T}ype$, and $\mathcal{T}m(\Gamma, A) = \prod_{\gamma \in \Gamma} A\,\gamma$ for $A \in \mathcal{T}ype(\Gamma)$.
- V_{I} is closed under dependent function types because of inaccessibility of I.
- Let $\mathcal{S}et = \mathrm{V}_{\mathsf{M}} \in \mathrm{V}_{\mathsf{I}}$ and $\mathrm{V}_{\mathsf{M}} \subseteq \mathrm{V}_{\mathsf{I}}$.
- V_{M} is closed under the standard set formers because of inaccessibility of M.
- The structure for the subuniverses will be given below.

We refer to Aczel [6] for details of the set-theoretic interpretation of Martin-Löf type theory. Aczel also shows that this interpretation can be carried out in CZF (Aczel's constructive version of ZF) with the regular extension axiom and suitable universe axioms. Palmgren [34] implemented an interpretation of Martin-Löf type theory in Aczel's iterative set model of CZF in the proof assistant Agda. However, Mahlo universes are not covered in these works.

In order to interpret the subuniverses we prove the following.

Theorem 3.3. *Let $f : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})$. Then there exists an inaccessible cardinal κ_f such that $f : \mathcal{F}am(\mathrm{V}_{\kappa_f}) \to Fam(\mathrm{V}_{\kappa_f})$. Furthermore, if we define $\mathcal{U}\, f_0\, f_1 = \mathrm{V}_{\kappa_f} \in \mathrm{V}_{\mathsf{M}}$ and $\mathcal{T}\, f_0\, f_1 : \mathrm{V}_{\kappa_f} \hookrightarrow \mathrm{V}_{\mathsf{M}}$, $\mathcal{T}\, f_0\, f_1\, x = x$ then $(\mathcal{U}, \mathcal{T})$ is a Russell-style model of the Tarski-style subuniverse closed under f.*

Proof: We define a normal function $h_f(\alpha)$ for $\alpha < \mathsf{M}$ by transfinite recursion:

$$\begin{aligned} h_f(0) &= 0 \\ h_f(\alpha+1) &= \bigvee_{x \in \mathcal{F}am(\mathrm{V}_\alpha)} \mathrm{rank}(f(x)) \vee (h_f(\alpha)+1) \\ h_f(\lambda) &= \bigvee_{\alpha<\lambda} h_f(\alpha) \quad \text{for } \lambda \text{ limit ordinal} \end{aligned}$$

It follows by induction on α that $h_f(\alpha) < \mathsf{M}$: The cases $\alpha = 0$ and $\alpha = \lambda$ are immediate by the induction hypothesis and regularity of M. In the case of $\alpha = \alpha' + 1$ we have by inaccessibility of M that $\mathrm{card}(\mathcal{F}am(\mathrm{V}_\alpha)) < \mathsf{M}$, $\mathrm{rank}(f(x)) < \mathsf{M}$ for $x \in \mathcal{F}am(\mathrm{V}_\alpha)$, and therefore $h_f(\alpha+1) < \mathsf{M}$ by the induction hypothesis and regularity of M.

Since M is Mahlo and h_f is normal, there exists an inaccessible $\kappa_f < \mathsf{M}$ such that $h_f : \kappa_f \to \kappa_f$. We conclude that $f : \mathcal{F}am(\mathrm{V}_{\kappa_f}) \to \mathcal{F}am(\mathrm{V}_{\kappa_f})$. Since κ_f is inaccessible, V_{κ_f} is closed under the interpretation of the standard type-theoretic set formers.

4 Second set-theoretic model

We now provide an alternative interpretation where $\mathcal{T}ype \subseteq \mathrm{V}_{\mathsf{I}}, \mathcal{S}et \subseteq \mathrm{V}_{\mathsf{M}}$, and the graph of the decoding function $\mathcal{T}\, f_0\, f_1 \subseteq \mathrm{V}_{\kappa_f} \times \mathrm{V}_{\kappa_f}$ are all inductively generated. Moreover, we define $\mathcal{U}\, f_0\, f_1$ as the domain of $\mathcal{T}\, f_0\, f_1$ and prove $\mathcal{T}\, f_0\, f_1 : \mathcal{U}\, f_0\, f_1 \to \mathcal{S}et$.

Rule sets. Inductive definitions in set theory are often presented in terms of fixed points of monotone operators. However, here we instead use Aczel's *rule sets* [1]. This lets us write set-theoretic rules so that they look like syntactic inference rules. Since each rule set determines a monotone operator on subsets of the base set of the rule set, this is only a presentation issue. We recall some definitions:

Definition 4.1. *A rule*

$$\frac{u}{v}$$

on a base set U is a pair of sets $u \subseteq U$ (of premises) and $v \in U$ (the conclusion).

Let Φ be a set of rules on U. A set w is Φ-closed iff

$$\frac{u}{v} \in \Phi \text{ and } u \subseteq w \text{ implies } v \in w.$$

There is a least Φ-closed set

$$\mathcal{I}(\Phi) = \bigcap \{w \subseteq U \mid w\ \Phi\text{-closed}\},$$

the set inductively defined by Φ [1].

Inductive definition of the subuniverses. Let M be a Mahlo cardinal, $f : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})$, and $\kappa_f < \mathsf{M}$ be an inaccessible cardinal as in Theorem 3.3 in Section 3.

Let c_0 and c_1 be set-theoretic encodings of the constructors for $\mathcal{U}\, f_0\, f_1$ that express closure under f_0 and f_1. One version would be to define

$$\mathrm{c}_0\, x\, y := (0, x, y) \quad \mathrm{c}_1\, x\, y\, t := (1, x, y, t)$$

(to be precise, $(0, x, y) := (0, (x, y))$ and similarly for $(1, x, y, t)$).

We define the subuniverses by first inductively generating the graphs of the decoding functions $\mathcal{T}\, f_0\, f_1$ by a rule set on $\mathrm{V}_{\kappa_f} \times \mathrm{V}_{\kappa_f}$. This rule set has rules for closure under f_0 and f_1 with respective codes c_0 and c_1:

$$\{\frac{\{(x, X)\} \cup \{(y\, z, Y\, z) | z \in X\}}{(\mathrm{c}_0\, x\, y, f_0\, X\, Y)} \mid x, X \in \mathrm{V}_{\kappa_f}, y, Y : X \to \mathrm{V}_{\kappa_f}\}$$

$$\cup$$

$$\{\frac{\{(x, X)\} \cup \{(y\, z, Y\, z) | z \in X\}}{(\mathrm{c}_1\, x\, y\, t, f_1\, X\, Y\, t)} \mid x, X \in \mathrm{V}_{\kappa_f}, y, Y : X \to \mathrm{V}_{\kappa_f}, t \in f_0\, X\, Y\}$$

In addition to that, we have rules expressing that these subuniverses à la Tarski are closed under the standard set formers. For example, the rule set for closure under Σ and $\mathbb{N}$ is as follows:

$$\{\frac{\{(x, X)\} \cup \{(y\, z, Y\, z) | z \in X\}}{(\widehat{\Sigma}\ x\ y, \sum z \in X.Y\ z)} \mid x, X \in \mathrm{V}_{\kappa_f}, y, Y : X \to \mathrm{V}_{\kappa_f}\} \cup \{\frac{\emptyset}{(\widehat{\mathbb{N}}, \omega)}\}$$

The code constructors $\widehat{\Sigma}$ and $\widehat{\mathbb{N}}$ are defined in a similar way to c_0 and c_1.

We can now prove by induction on the rule set that $\mathcal{T}\, f_0\, f_1$ is a function. Let $\mathcal{U}\, f_0\, f_1 \subseteq \mathrm{V}_{\kappa_f}$ be the domain of $\mathcal{T}\, f_0\, f_1$. Hence, $\mathcal{T}\, f_0\, f_1 : \mathcal{U}\, f_0\, f_1 \to \mathrm{V}_{\kappa_f}$.

Note that we have not yet defined $\mathcal{S}et$ and that $\mathcal{T}\, f_0\, f_1 : \mathcal{U}\, f_0\, f_1 \to \mathrm{V}_{\kappa_f}$ is defined for an arbitrary operator f on $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ rather than on $\mathcal{F}am(\mathcal{S}et)$. This is analogous to Kahle and Setzer's *preuniverses* $\mathsf{PU}[a, f]$ in Explicit Mathematics [23, 16]. These are defined for arbitrary $a \in \mathrm{Tm}$ and $f : \mathrm{Tm} \to \mathrm{Tm}$ and not only for $a \in \Re$ and $f : \Re \to \Re$, where Tm is the set of untyped closed terms and $\Re \subseteq \mathrm{Tm}$ is the external Mahlo universe.

Inductive definition of the external Mahlo universe. The following is a rule set on V_{M} that (together with rules expressing that $\mathcal{S}et$ is a universe à la Russell closed under the standard set formers) inductively generates the external Mahlo universe $\mathcal{S}et \subseteq \mathrm{V}_{\mathsf{M}}$:

$$\{\frac{\begin{array}{c}\{f_0\,(\mathcal{T}\, f_0\, f_1\, x_0)\,((\mathcal{T}\, f_0\, f_1) \circ x_1) \mid (x_0, x_1) \in \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)\} \\ \cup\, \{f_1\,(\mathcal{T}\, f_0\, f_1\, x_0)\,((\mathcal{T}\, f_0\, f_1) \circ x_1)\, t \mid (x_0, x_1) \in \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1), t \in f_0\,(\mathcal{T}\, f_0\, f_1\, x_0)\,((\mathcal{T}\, f_0\, f_1) \circ x_1)\}\end{array}}{\mathcal{U}\, f_0\, f_1} \mid f : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})\}$$

The premises of this rule set express that the family operator f maps a family $(x_0, x_1) \in \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$ to a family $(X_0, X_1) \in \mathcal{F}am(\mathcal{S}et)$, where $X_0 = \mathcal{T}\, f_0\, f_1\, x_0$ and $X_1 = (\mathcal{T}\, f_0\, f_1) \circ x_1$. This corresponds to Kahle and Setzer's *independence condition* [23, 16].

As already mentioned, we need to add rules for $\mathcal{S}et$ that express closure under all standard set formers. For example, the rule set for closure under Σ and $\mathbb{N}$ is as follows:

$$\{\frac{\{X\} \cup \{Y\, z \mid z \in X\}}{\sum z \in X.Y\ z} \mid X \in \mathrm{V}_{\mathsf{M}}, Y : X \to \mathrm{V}_{\mathsf{M}}\} \cup \{\frac{\emptyset}{\omega}\}$$

Theorem 4.2. *Let $f : \mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$ and define $f' : \mathcal{F}am(\mathrm{V}_{\mathsf{M}}) \to \mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ by $f'(X, Y) = f(X, Y)$ if $(X, Y) \in \mathcal{F}am(\mathcal{S}et)$ and $f'(X, Y) = (\emptyset, \emptyset)$ otherwise. Then $\mathcal{T}\, f'_0\, f'_1 : \mathcal{U}\, f'_0\, f'_1 \to \mathcal{S}et$.*

Proof: We prove by induction on the rule set for $\mathcal{T}\, f'_0\, f'_1$ that if $(x, X) \in \mathcal{T}\, f'_0\, f'_1$, then $X \in \mathcal{S}et$. This implies the theorem. Assume $x, X \in \mathrm{V}_{\kappa_f}$, $y, Y : X \to V_{\kappa_f}$, $(x, X) \in \mathcal{T}\, f'_0\, f'_1$, $\forall z \in X.(y\, z, Y\, z) \in \mathcal{T}\, f'_0\, f'_1$. By induction hypothesis we have $X \in \mathcal{S}et$, $\forall z \in X.Y\, z \in \mathcal{S}et$. Then $(X, Y) \in \mathcal{F}am(\mathcal{S}et)$ and therefore $f'(X, Y) = f(X, Y) \in \mathcal{F}am(\mathcal{S}et)$. Hence $f'_0\, X\, Y = f_0\, X\, Y \in \mathcal{S}et$, and $f'_1\, X\, Y\, t = f_1\, X\, Y\, t \in \mathcal{S}et$ for $t \in f'_0\, X\, Y$ in the conclusion of the second rule. The case of the basic set formers follows similarly by IH and the closure of $\mathcal{S}et$ under the basic set formers.

Inductive definition of the collection of types. The following is a rule set on V_{I} that inductively generates the collection of types $\mathcal{T}ype \subseteq \mathrm{V}_{\mathsf{I}}$:

$$\{\frac{\emptyset}{\mathcal{S}et}\} \cup \{\frac{\emptyset}{X} \mid X \in \mathcal{S}et\} \cup \{\frac{\{X\} \cup \{Y\, x \mid x \in X\}}{\prod_{x \in X}\ Y\, x} \mid X \in \mathrm{V}_{\mathsf{I}}, Y : X \to \mathrm{V}_{\mathsf{I}}\}$$

The second model as a Mahlo cwf.

- Each context in $\mathcal{C}tx$ has a length n. We define the set of contexts of length n by induction on n:

  - The empty context $1 = \{()\} \in \mathcal{C}tx$ is the only context of length 0, where $() = \emptyset$ is the empty sequence.
  - Contexts of length $n+1$ have the form $\sum_{\gamma\in\Gamma} A\,\gamma \in \mathcal{C}tx$, where $\Gamma \in \mathcal{C}tx$ has length n and $A \in \mathcal{T}ype(\Gamma) = \Gamma \to \mathcal{T}ype$. The elements are sequences (γ, a), where $\gamma \in \Gamma$ and $a \in A\,\gamma$.
- $\mathcal{H}om(\Delta, \Gamma) = \Delta \to \Gamma$.
- $\mathcal{T}m(\Gamma, A) = \prod_{\gamma\in\Gamma} A\,\gamma$.
- $\mathcal{T}ype$ is defined as closed under the Cartesian product $\prod$ of families of sets in $\mathrm{V_I}$. It follows that the cwf has a structure for Π-types.
- $\mathcal{T}ype$ is defined to contain and include Set, that is, $\mathrm{Set} \in \mathcal{T}ype$ and $\mathrm{Set} \subseteq \mathcal{T}ype$.
- $\mathcal{S}et$ is defined to be closed under the standard set formers.

We interpret the subuniverse closed under $f : \mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$ as $\mathcal{U}\,f_0'\,f_1'$ with decoding $\mathcal{T}\,f_0'\,f_1' : \mathcal{U}\,f_0'\,f_1' \to \mathcal{S}et$ as in Theorem 4.2.

- To prove $\mathcal{U}\,f_0'\,f_1' \in \mathcal{S}et$, we assume $(x_0, x_1) \in \mathcal{T}\mathcal{F}am(\mathcal{U}\,f_0'\,f_1')$ and define $X_0 = \mathcal{T}\,f_0'\,f_1' x_0$ and $X_1 = (\mathcal{T}\,f_0'\,f_1') \circ x_1$. Hence $(X_0, X_1) \in \mathcal{F}am(\mathcal{S}et)$ and all the premises of the rule that adds $\mathcal{U}\,f_0'\,f_1'$ to $\mathcal{S}et$ in the inductive generation of $\mathcal{S}et$ are satisfied.
- $\mathcal{T}\,f_0'\,f_1' : \mathcal{U}\,f_0'\,f_1' \to \mathcal{S}et$ for $f : \mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$ follows by Theorem 4.2.
- $\mathcal{U}\,f_0'\,f_1' \in \mathcal{S}et$ and $\mathcal{T}\,f_0'\,f_1' : \mathcal{U}\,f_0'\,f_1' \to \mathcal{S}et$ is a universe à la Tarski closed under all standard set formers and under the family operator (f_0, f_1) with codes $(\mathrm{c}_0, \mathrm{c}_1)$. The typing rules for the constructors c_0 and c_1 of $\mathcal{U}\,f_0'\,f_1'$ and the equality rules for $\mathcal{T}\,f_0'\,f_1'$ are immediate from the rule set for $\mathcal{T}\,f_0'\,f_1'$.

5 Third and fourth set-theoretic models

In the previous Section 4 we referred to the set of functions $f : \mathcal{F}am(\mathrm{V_M}) \to \mathcal{F}am(\mathrm{V_M})$ when defining $\mathcal{U}\,f_0\,f_1 \in \mathrm{V_M}$. In the extended predicative Mahlo universe [23] this impredicativity was avoided by referring to the set of partial functions in Explicit Mathematics. Moreover, we transferred this construction to type theory by implementing a model of Explicit Mathematics [16].

In this section, we use a similar idea to construct a variation of the second set-theoretic model, where the partial functions in Explicit Mathematics are approximated by arbitrary sets in set theory. We use the fact that we can define $(f\,a)$ for arbitrary sets f and not only for functions in set theory:

$$f\,a := \bigcup\{x \mid (a, x) \in f\}$$

We thus replace the partial functions in Explicit Mathematics by arbitrary sets in $\mathrm{V_I}$ in set theory.

Let $f_0, f_1 : \mathrm{V_I}$ and define $\mathcal{T}\,f_0\,f_1$ as the set inductively generated by the following rule set on $\mathrm{V_M} \times \mathrm{V_M}$. As before, this rule set needs to be augmented by rules for closure under the standard set formers (in the previous version we didn't need the conditions that the result are in $\mathrm{V_M}$, since that was guaranteed by $f : \mathcal{F}am(\mathrm{V_M}) \to \mathcal{F}am(\mathrm{V_M})$):

$$\{\frac{\{(x, X)\} \cup \{(y\,z, Y\,z) | z \in X\}}{(\mathrm{c}_0\,x\,y, f_0\,X\,Y)} \mid x, X \in \mathrm{V_M},\ y, Y : X \to \mathrm{V_M},\ f_0\,X\,Y \in \mathrm{V_M}\}$$

$$\bigcup$$

$$\{\frac{\{(x,X)\}\cup\{(y\,z,Y\,z)|z\in X\}}{(\mathrm{c}_1\,x\,y\,t,f_1\,X\,Y\,t)}\mid x,X\in \mathrm{V}_{\mathsf{M}},\ y,Y:X\to \mathrm{V}_{\mathsf{M}},\ f_0\,X\,Y\in \mathrm{V}_{\mathsf{M}},t\in f_0\,X\,Y,\ f_1\,X\,Y\,t\in \mathrm{V}_{\mathsf{M}}\}$$

As before, we can prove that $\mathcal{T}\,f_0\,f_1$ is a function. Let $\mathcal{U}\,f_0\,f_1 := \mathrm{dom}(\mathcal{T}\,f_0\,f_1)\subseteq \mathrm{V}_{\mathsf{M}}$. We get $\mathcal{T}\,f_0\,f_1 : \mathcal{U}\,f_0\,f_1\ \to \mathrm{V}_{\mathsf{M}}$.

We define for $f_0,f_1\in \mathrm{V}_{\mathsf{I}}$ the two components of the lifting of a partial function (f_0,f_1) from $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ to $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ to a partial function $(f\mathcal{T}_0,f\mathcal{T}_1)$ from $\mathcal{TF}am(\mathcal{U}\,f_0,f_1)$ to $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$. Let $x=(x_0,x_1)$:

$$\begin{array}{lcl} f\mathcal{T}_0(x) & := & f_0\,(\mathcal{T}\,f_0\,f_1\,x_0)\,((\mathcal{T}\,f_0\,f_1)\circ x_1)\\ f\mathcal{T}_1(x,t) & := & f_1\,(\mathcal{T}\,f_0\,f_1\,x_0)\,((\mathcal{T}\,f_0\,f_1)\circ x_1)\,t\\ f\mathcal{T}(x) & := & (f\mathcal{T}_0(x),\lambda t\in f\mathcal{T}_0(x).f\mathcal{T}_1(x,t))\end{array}$$

As in [16, 23] we define $(\mathcal{T}\,f_0\,f_1)$ to be independent of V_{M}, if the conditions

$$f_0\,X\,Y\in \mathrm{V}_{\mathsf{M}}\qquad f_1\,X\,Y\,t\in \mathrm{V}_{\mathsf{M}}$$

are always fulfilled:

$$\mathit{Indep}(\mathcal{T}\,f_0\,f_1) :\Leftrightarrow \forall x\in\mathcal{TF}am(\mathcal{U}\,f_0\,f_1).f\mathcal{T}_0(x)\in \mathrm{V}_{\mathsf{M}}\wedge\forall t\in f\mathcal{T}_0(x).f\mathcal{T}_1(x,t)\in \mathrm{V}_{\mathsf{M}}$$

Lemma 5.1. *Let $f_0,f_1\in \mathrm{V}_{\mathsf{I}}$ and $\mathit{Indep}(\mathcal{T}\,f_0\,f_1)$. Then $\mathcal{U}\,f_0\,f_1\in \mathrm{V}_{\mathsf{M}}$ and $\mathcal{T}\,f_0\,f_1 : \mathcal{U}\,f_0\,f_1\to \mathrm{V}_{\mathsf{M}}$.*

Proof. Since the base set of the rule set for $\mathcal{T}\,f_0\,f_1$ is $\mathrm{V}_{\mathsf{M}}\times \mathrm{V}_{\mathsf{M}}$ it follows immediately that $\mathcal{T}\,f_0\,f_1 : \mathcal{U}\,f_0\,f_1\to \mathrm{V}_{\mathsf{M}}$.

To show $\mathcal{U}\,f_0\,f_1\in \mathrm{V}_{\mathsf{M}}$ we show that there is an inaccessible κ_f such that $\mathcal{T}\,f_0\,f_1\subseteq \mathrm{V}_{\kappa_f}\times \mathrm{V}_{\kappa_f}$.

Define a normal function $h_f:\mathsf{M}\to\mathsf{M}$ by transfinite recursion:

$$\begin{array}{lcl} h_f(0) & := & 0\\ h_f(\alpha+1) & := & (\bigvee_{x\in\mathcal{TF}am(\mathcal{U}\,f_0\,f_1)\cap \mathrm{V}_\alpha}\mathrm{rank}(f\mathcal{T}(x)))\vee(h_f(\alpha)+1)\\ h_f(\lambda) & := & \bigvee_{\alpha<\lambda}h_f(\alpha)\ \text{ for } \lambda \text{ limit ordinal}\end{array}$$

It follows by induction on α that $h_f(\alpha)<\mathsf{M}$.

Let κ_f be an inaccessible fixed point of h_f. Then V_{κ_f} is closed under the standard set formers. Furthermore, if the premises of the main rules for $\mathrm{c}_0\,x\,y$ or $\mathrm{c}_1\,x\,y\,t$ are in $\mathcal{T}\,f_0\,f_1\cap \mathrm{V}_{\kappa_f}$, then the conclusion is also in $\mathcal{T}\,f_0\,f_1\cap \mathrm{V}_{\kappa_f}$. Therefore, $\mathcal{T}\,f_0\,f_1\subseteq \mathrm{V}_{\kappa_f}$, $\mathcal{T}\,f_0\,f_1\in \mathrm{V}_{\kappa_f+1}\subseteq \mathrm{V}_{\mathsf{M}}$. □

We now define $\mathcal{S}et$ by the following rule set on V_{I} (which needs to be augmented by rules for closure under the standard set formers):

$$\{\ \frac{\{f\mathcal{T}_0(x)\mid x\in\mathcal{TF}am(\mathcal{U}\,f_0\,f_1)\}\cup\ \{f\mathcal{T}_1(x,t)\mid x\in\mathcal{TF}am(\mathcal{U}\,f_0\,f_1),t\in f\mathcal{T}_0(x)\}}{\mathcal{U}\,f_0\,f_1}\mid f_0,f_1\in \mathrm{V}_{\mathsf{I}}\}$$

Lemma 5.2. $\mathcal{S}et\subseteq \mathrm{V}_{\mathsf{M}}$

Proof. Proof by induction on the rule set for $\mathcal{S}et$. For the standard set formers this follows from the induction hypothesis. Assume that $\mathcal{U}\,f_0\,f_1\in\mathcal{S}et$ is introduced by its rule. Then by the induction hypothesis applied to the assumptions of the rule we have $\mathit{Indep}(\mathcal{T}\,f_0\,f_1)$ and therefore by Lem. 5.1 we have $\mathcal{U}\,f_0\,f_1\in \mathrm{V}_{\mathsf{M}}$. □

Lemma 5.3. *Let $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ be introduced by its rule, and $t \in \mathcal{U}\, f_0\, f_1$. Then $\mathcal{T}\, f_0\, f_1\, t \in \mathcal{S}et$.*

Proof. $u \in \mathcal{U}\, f_0\, f_1 \Leftrightarrow \exists t.(u,t) \in \mathcal{T}\, f_0\, f_1$, and then $t = \mathcal{T}\, f_0\, f_1\, u$. We show by induction on $(u,t) \in \mathcal{T}\, f_0\, f_1$ that $t \in \mathcal{S}et$.
For the standard set formers this follows by the induction hypothesis and closure of $\mathcal{S}et$ under the basic set constructions.
Assume $(u,t) = (\mathrm{c}_0\, x_0\, x_1, f_0\, X_0\, X_1)$ introduced by its rule. Then $(x_0, X_0) \in \mathcal{T}\, f_0\, f_1$ and $\forall t' \in X_0.(x_1\, t', X_1\, t') \in \mathcal{T}\, f_0\, f_1$. By $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ introduced by its rule it follows that $t = f_0\, X_0\, X_1 = f\mathcal{T}_0(x) \in \mathcal{S}et$.
Assume $(u,t) = (\mathrm{c}_1\, x_0\, x_1\, t, f_1\, X_0\, X_1\, t')$ introduced by its rule. Then $(x_0, X_0) \in \mathcal{T}\, f_0\, f_1$, and $\forall t'' \in X_0.(x_1\, t'', X_1\, t'') \in \mathcal{T}\, f_0\, f_1$. Furthermore, $t' \in f_0\, X_0\, X_1$. By $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ introduced by its rule it follows that $t = f_1\, X_0\, X_1\, t' = f\mathcal{T}_1(x, t') \in \mathcal{S}et$.

□

Lemma 5.4. *$\mathcal{S}et$ is a model of the external Mahlo universe.*

Proof. $\mathcal{S}et$ is by assumption closed under the standard set formers.
Assume $f = (f_0, f_1) : \mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$. Then $f \in \mathrm{V}_{\mathsf{I}}$. We need to show that $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ and $(\mathrm{c}_0\, f_0\, f_1, \mathrm{c}_1\, f_0\, f_1) : \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1) \to \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$.
Claim: $\forall (u,t) \in \mathcal{T}\, f_0\, f_1.t \in \mathcal{S}et \qquad (1)$
Proof of (1) by induction on $(u,t) \in \mathcal{T}\, f_0\, f_1$:
If (u,t) is introduced by basic set constructions this follows by induction hypothesis.
Let $(u,t) = (\mathrm{c}_0\, x_0\, x_1, f_0\, X_0,\, X_1)$ introduced by its rule. By induction hypothesis $X_0 \in \mathcal{S}et$ and for $z \in X_0$ we have $X_1\, z \in \mathcal{S}et$. Therefore, $t = f_0\, X_0\, X_1 \in \mathcal{S}et$.
Let $(u,t) = (\mathrm{c}_1\, x_0\, x_1\, t', f_0\, X_0,\, X_1\, t')$ be introduced by its rule. By the induction hypothesis $X_0 \in \mathcal{S}et$ and for $z \in X$ we have $X_1\ z \in \mathcal{S}et$. Furthermore, $t' \in f_0\, X_0\, X_1$. Therefore $f_1\, X_0\, X_1\, t' \in \mathcal{S}et$.
This concludes the proof of (1).

It follows that if $(x_0, x_1) \in \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$ then $(\mathcal{T}\, f_0\, f_1\, x_0, (\mathcal{T}\, f_0\, f_1) \circ x_1) \in \mathcal{F}am(\mathcal{S}et)$, and therefore the assumptions of the rule for $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ are fulfilled and therefore $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$. Therefore, $\mathcal{I}ndep(\mathcal{U}\, f_0\, f_1)$, and we get $(\mathrm{c}_0\, f_0\, f_1, \mathrm{c}_1\, f_0\, f_1) : \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1) \to \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$ □

We show that we can restrict $f_0, f_1 \in \mathrm{V}_{\mathsf{I}}$ to $f_0, f_1 \in \mathrm{V}_{\mathsf{M}}$:
We show first that $\mathcal{U}\, f_0\, f_1$ depends only on the restrictions of f_0, f_1 to $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$:
We define a more general version of $(f\mathcal{T}_0, f\mathcal{T}_1)$, namely the lifting of a partial function (f'_0, f'_1) from $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ to $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$ to a partial function $(f'\mathcal{T}_{f_0,f_1,0}, f'\mathcal{T}_{f_0,f_1,1})$ from $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0, f_1)$ to $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$:

$$f'\mathcal{T}_{f_0,f_1,0}(x) := f'_0\, (\mathcal{T}\, f_0\, f_1\, x_0)\, ((\mathcal{T}\, f_0\, f_1) \circ x_1)$$
$$f'\mathcal{T}_{f_0,f_1,1}(x,t) := f'_1\, (\mathcal{T}\, f_0\, f_1\, x_0)\, ((\mathcal{T}\, f_0\, f_1) \circ x_1)\, t$$

Lemma 5.5. *Assume*

$$\forall x \in \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1).f\mathcal{T}_0(x) = f'\mathcal{T}_{f_0,f_1,0}(x) \wedge \forall t \in f\mathcal{T}_0(x).f\mathcal{T}_1(x,t) = f'\mathcal{T}_{f_0,f_1,1}(x,t)$$

Then $\mathcal{T}\, f_0\, f_1 = \mathcal{T}\, f'_0\, f'_1$.

Proof. One shows by straightforward induction on $(u,t) \in \mathcal{T}\, f_0\, f_1$ that $(u,t) \in \mathcal{T}\, f'_0\, f'_1$. Then we show by induction that for $(u,t) \in \mathcal{T}\, f'_0\, f'_1$ we have $(u,t) \in \mathcal{T}\, f_0\, f_1$. □

We say (f_0, f_1) and (f'_0, f'_1) coincide on $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$ if the assumptions of the lemma 5.5 are fulfilled.

Corollary 5.6. *Assume $\mathcal{U}\, f_0\, f_1 \in \mathcal{S}et$ is introduced by its rule. Let*

$$\begin{array}{lcl} f_0\restriction & := & \{(X, Y, f_0\, X\, Y) \mid (x, X) \in \mathcal{T}\, f_0\, f_1, y, Y : X \to \mathrm{V}_{\mathsf{M}}.\forall z \in X.(y\, z, Y\, z) \in \mathcal{T}\, f_0\, f_1\} \\ f_1\restriction & := & \{(X, Y, t', f_1\, X\, Y\, t) \mid (x, X) \in \mathcal{T}\, f_0\, f_1, y, Y : X \to \mathrm{V}_{\mathsf{M}}.\forall z \in X.(y\, z, Y\, z) \in \mathcal{T}\, f_0\, f_1, t' \in f_0\, X\, Y\} \end{array}$$

Then (f_0, f_1) and $(f_0\restriction, f_1\restriction)$ coincide on $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1)$ and $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1) = \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\restriction\ f_1\restriction)$. Furthermore, $f_0\restriction, f_1\restriction \in \mathrm{V}_{\mathsf{M}}$.

Fourth set-theoretic model. This is obtained by replacing the condition $f_0, f_1 \in \mathrm{V}_{\mathsf{I}}$ by $f_0, f_1 \in \mathrm{V}_{\mathsf{M}}$.

Corollary 5.7. *Let $\mathcal{S}et'$ be defined by the same rules as for $\mathcal{S}et$, but replacing the condition $f_0, f_1 \in \mathrm{V}_{\mathsf{I}}$ by $f_0, f_1 \in \mathrm{V}_{\mathsf{M}}$. Then $\mathcal{S}et' = \mathcal{S}et$ and therefore $\mathcal{S}et'$ is a model of the external Mahlo universe.*

Note that if $(f_0, f_1) : \mathcal{F}am(\mathcal{S}et) \to \mathcal{F}am(\mathcal{S}et)$, it is not necessarily the case that $f_0, f_1 \in \mathrm{V}_{\mathsf{M}}$, since it is not necessarily the case that $\mathcal{S}et \in \mathrm{V}_{\mathsf{M}}$ (provided that M is the least Mahlo cardinal). However, we get $f_0\restriction, f_1\restriction \in \mathrm{V}_{\mathsf{M}}$.

Replacing $\mathcal{S}et$ by a Tarski Universe. If one wants to replace this model by a Tarski style model for $\mathcal{S}et$, one needs to replace the rule set for the external Mahlo rules by a rule set generating the graph of the decoding function $\mathcal{E}l$ for $\mathcal{S}et$ in a similar way as the inductive definition of the decoding function $\mathcal{T}\, f_0\, f_1$ for the subuniverses $\mathcal{U}\, f_0\, f_1$. These rules generate pairs (a, A) where a is a code and A is the result of applying the decoding function $\mathcal{E}l$ to it.

The reader might expect that the closure rule under $\mathcal{U}\, f_0\, f_1$ should have the conclusion $(\mathrm{u}\, f_0\, f_1, \mathcal{U}\, f_0\, f_1)$. However, that would not work: f_0, f_1 range over all elements in $\mathcal{F}am(\mathrm{V}_{\mathsf{M}})$, and therefore $\mathrm{u}\, f_0\, f_1$ is not an element of V_{M}. Actually, if one could define this, one could define a variant of Palmgren's paradox since one can extract from the code of $\mathrm{u}\, f_0\, f_1$ the functions f_0, f_1. One solution would be to replace it by $\mathrm{u}\, f_0\restriction\ f_1\restriction$, as defined in Corollary 5.6. Since $\mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\, f_1) = \mathcal{T}\mathcal{F}am(\mathcal{U}\, f_0\restriction\ f_1\restriction)$, that definition would create a function.

6 Meaning explanations

We shall now present informal meaning explanations for our theory $\mathbf{TT}^{\mathsf{M}}$. Meaning explanations for extensional type theory were introduced by Martin-Löf [29] and elaborated on in the book *Intuitionistic Type Theory* [30]. Further discussion of the meaning of the logical framework based version of type theory with its distinction between *types* and *sets* can be found in Martin-Löf's Leiden lectures [26, 27].

We will not elaborate on the meaning explanations for the basic type theory $\mathbf{TT}$, where we mostly follow [29]. A difference is that we do not explain type equality extensionally (two types are equal iff they have the same elements) but in the same way as equality of elements of a universe [13]. Regarding the logical framework, we explain Set in the same way as the universes in [29]. However, adding closure under subuniverses in the extended theory $\mathbf{TT}^{\mathsf{M}}$, so that Set becomes an external Mahlo universe, requires special provisions.

To explain the meaning of the judgment $a : A$ we first specify the *canonical forms* and the *computation rules* associating each term with its canonical form. Canonical forms are terms of the form $c\, a_1 \cdots a_n$, where c is a *constructor*. Note that we have *lazy* canonical terms – it is not required that $a_1, \ldots, a_n$ are canonical. If a has canonical form $c\, a_1 \cdots a_m$ and A has canonical form $C\, b_1 \cdots b_n$, then the meaning of $a : A$ is specified by *matching conditions* which state

whether the element constructor c matches the type constructor C, and if so, the conditions on the subterms. For example, the matching conditions for natural numbers are that 0 matches N, and $\mathrm{succ}\, a$ matches N, under the condition $a : \mathrm{N}$. The matching condition for W-types is that $\mathrm{sup}\, a\, b$ matches $\mathrm{W}\, A\, B$ under the condition that $a : A$ and $b : B\, a \to \mathrm{W}\, A\, B$.

We assume that the canonical forms, the computation rules, and the matching conditions are already specified for **TT** and only list those specific to $\mathbf{TT}^{\mathsf{M}}$.

New canonical terms. These are terms of the form $\mathrm{U}\, f_0\, f_1, \mathrm{c}_0\, u\, t$, and $\mathrm{c}_1\, u\, t\, b$, where f_0, f_1, u, t, b are terms (not necessarily canonical). U is a set constructor and c_0 and c_1 are element constructors.

New computation rules. These correspond to the two new equality rules for T. The canonical form of $\mathrm{T}\, f_0\, f_1\, a$ is v if either

- the canonical form of a is $\mathrm{c}_0\, u\, t$, and the canonical form of $f_0\, (\mathrm{T}\, f_0\, f_1\, u)\, ((\mathrm{T}\, f_0\, f_1) \circ t)$ is v;
- or the canonical form of a is $\mathrm{c}_1\, u\, t\, b$, and the canonical form of $f_1\, (\mathrm{T}\, f_0\, f_1\, u)\, ((\mathrm{T}\, f_0\, f_1) \circ t)\, b$ is v.

New matching conditions.

- The canonical form $\mathrm{U}\, f_0\, f_1$ matches the canonical form Set. The judgment is valid under the conditions that
$$f_0\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x)) : \mathrm{Set}$$
in the context $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$, and
$$f_1\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x))\, b : \mathrm{Set}$$
in the context $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1, b : f_0\, (\mathrm{T}\, f_0\, f_1\, u)\, (\lambda x.\mathrm{T}\, f_0\, f_1\, (t\, x))$.
- We then have matching conditions corresponding to the U-introduction rules.
  - The canonical form $\mathrm{c}_0\, u\, t$ matches the canonical form $\mathrm{U}\, f_0\, f_1$. The judgment is valid under the condition that $u : \mathrm{U}\, f_0\, f_1$ and $t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$.
  - The canonical form $\mathrm{c}_1\, u\, t\, b$ matches the canonical form $\mathrm{U}\, f_0\, f_1$. The judgment is valid under the conditions that $u : \mathrm{U}\, f_0\, f_1$ and $t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$, and $b : f_0\, u\, t$.
  - Since $\mathrm{U}\, f_0\, f_1$ is closed under all the standard set formers, we also have matching conditions for each of them.

  Moreover, in all cases we check the conditions for f_0 and f_1 in $\mathrm{U}\, f_0\, f_1$ as in the first item above.

Well-foundedness. The repeated process of lazily computing canonical forms and checking matching conditions must be well-founded. For example, the judgment $a : \mathrm{N}$ is only valid if the process of computing successive canonical forms of a produces finitely many successors and ends with a final matching $0 : \mathrm{N}$. If this process produces an infinite sequence of successors, then the judgment is not valid. Similarly, the judgment $c : \mathrm{W}\, A\, B$ must generate a well-founded tree of matchings of canonical forms. The root of the tree is the matching of $\mathrm{sup}\, a\, b : \mathrm{W}\, A\, B$

and the subtrees are the matchings of the canonical forms of $a : A$ and of $b\,x : \mathrm{W}\,A\,B$ for each $x : B\,a$.

Well-foundedness is a non-trivial issue for the Mahlo universe Set. In the second set-theoretic model we invoked the Mahlo cardinal M and the inaccessible $\mathsf{I} > \mathsf{M}$ in order to bound the size of the inductive definitions of $\mathrm{U}\,f_0\,f_1$ and Set, and of the collection of all types. The rules in these inductive definitions are set-theoretic renderings of the matching conditions in the meaning explanations. The set-theoretic inductive generation process mimics the repeated matching process in the meaning explanations and must be well-founded.

We emphasize that the matching condition for $\mathrm{U}\,f_0\,f_1$: Set deviates from the standard pattern where the matching conditions are immediate from the formation and introduction rules. If we followed that pattern here, then U-formation would require us to check the more general condition f : Fam(Set) $\to$ Fam(Set), and because of Palmgren's paradox the well-foundedness condition would be broken. The matching for $\mathrm{U}\,f_0\,f_1$: Set would have subtree matchings $f_0\,X_0\,X_1$: Set and $f_1\,X_0\,X_1\,t$: Set for (X_0, X_1) : Fam(Set) and $t : f_0\,X_0\,X_1$. Now, let $f = (\pi_0, \pi_1)$ be the identity operator on Fam(Set), that is, $\pi_0\,X_0\,X_1 = X_0$ and $\pi_1\,X_0\,X_1 = X_1$. Then the matching $\mathrm{U}\,\pi_0\,\pi_1$: Set has subtrees $\pi_0\,X_0\,X_1 = X_0$: Set for each X_0 : Set and $\pi_1\,X_0\,X_1\,t = X_1\,t$: Set for each $X_1 : X_0 \to$ Set and $t : X_0$. One of these subtrees is $X_0 = \mathrm{U}\,\pi_0\,\pi_1$: Set and we have a loop.

In our approach, we avoid the circularity by only matching for arguments in the image of the subuniverse and thus we avoid matching for $\mathrm{U}\,f_0\,f_1$: Set.

Matching conditions for equality judgments. The matching conditions for $a : A$ can be extended to typed equality judgments $a = a' : A$.

For example, for natural numbers we have $a = a' : A$ is valid if A has canonical form N and if a and a' both have the canonical form 0; or if a has canonical form $\mathrm{succ}\,b$ and a' has canonical form $\mathrm{succ}\,b'$ under the condition $b = b' :$ N.

The crucial matching condition for the external Mahlo universe Set is that $a = a' : A$ is valid if A has canonical form Set and a has canonical form $\mathrm{U}\,f_0\,f_1$ and a' has canonical form $\mathrm{U}\,f_0'\,f_1'$ under the condition that

$$f_0\,(\mathrm{T}\,f_0\,f_1\,u)\,(\lambda x.\mathrm{T}\,f_0\,f_1\,(t\,x)) = f_0'\,(\mathrm{T}\,f_0'\,f_1' u)\,(\lambda x.\mathrm{T}\,f_0'\,f_1'(t\,x)) : \mathrm{Set}$$

in the context $u : \mathrm{U}\,f_0\,f_1, t : \mathrm{T}\,f_0\,f_1\,u \to \mathrm{U}\,f_0\,f_1$, and

$$f_1\,(\mathrm{T}\,f_0\,f_1\,u)\,(\lambda x.\mathrm{T}\,f_0\,f_1\,(t\,x))\,b = f_0'\,(\mathrm{T}\,f_0'\,f_1' u)\,(\lambda x.\mathrm{T}\,f_0'\,f_1'(t\,x))\,b : \mathrm{Set}$$

in the context $u : \mathrm{U}\,f_0\,f_1, t : \mathrm{T}\,f_0\,f_1\,u \to \mathrm{U}\,f_0\,f_1, b : f_0\,(\mathrm{T}\,f_0\,f_1\,u)\,(\lambda x.\mathrm{T}\,f_0\,f_1\,(t\,x))$.

Matching conditions for other judgment forms. There are analogous matching conditions for the judgments A type and $A = A'$.

7 Justification of the rules

We now justify the correctness of the rules of $\mathbf{TT}^{\mathsf{M}}$ with respect to the meaning explanations. We only justify the rules that are new with respect to **TT**, that is, the rules for the subuniverses displayed as Agda code. This informal justification is similar to the justification of the rules in the second set-theoretic interpretation. We proceed with the details.

All rules assume that $f_0 : (X_0 : \mathrm{Set}) \to (X_0 \to \mathrm{Set}) \to \mathrm{Set}$ and $f_1 : (X_0 : \mathrm{Set}) \to (X_1 : X_0 \to \mathrm{Set}) \to f_0\,X_0\,X_1 \to \mathrm{Set}$.

Justification of the U-formation rule. We need to justify that $\mathrm{U}\, f_0\, f_1 : \mathrm{Set}$. We first make use of the matching condition for this case:

$$f_0\,(\mathrm{T}\, f_0\, f_1\, u)\,((\mathrm{T}\, f_0\, f_1) \circ t) : \mathrm{Set}$$

for $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$ and

$$f_1\,(\mathrm{T}\, f_0\, f_1\, u)\,((\mathrm{T}\, f_0\, f_1) \circ t)\, y : \mathrm{Set}$$

for $u : \mathrm{U}\, f_0\, f_1, t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1, b : f_0\,(\mathrm{T}\, f_0\, f1\, u)\,((\mathrm{T}\, f_0\, f_1) \circ t)$.

However, we know by the simultaneous justification of the typing rule for T, that $\mathrm{T}\, f_0\, f_1\, u : \mathrm{Set}$ and $(\mathrm{T}\, f_0\, f_1) \circ t : \mathrm{U}\, f_0\, f_1 \to \mathrm{Set}$. Hence, from the typings of f_0 and f_1 it follows that $\mathrm{U}\, f_0\, f_1 : \mathrm{Set}$.

As already discussed above, a judgment is only valid provided we get a well-founded process of lazily computing canonical forms and checking matching conditions. The insight that this process is well-founded is aided by the set-theoretic interpretation of Set as inductively defined by Aczel-style rules that correspond to the matching conditions above.

Justification of the U-introduction rules. These follow immediately from the matching conditions for c_0 and c_1 and for the matching conditions for the codes for the standard set formers.

Justification of the typing rule for T. To justify that $\mathrm{T}\, f_0\, f_1 : \mathrm{U}\, f_0\, f_1 \to \mathrm{Set}$ we use the computation rules for $\mathrm{T}\, f_0\, f_1\, a$. There is one case for each constructor for U.

- If the canonical form of a is $\mathrm{c}_0\, u\, t$, then the canonical form of $\mathrm{T}\, f_0\, f_1\, a$ is the same as the canonical form of $f_0\,(\mathrm{T}\, f_0\, f_1\, u)\,((\mathrm{T}\, f_0\, f_1) \circ t)$. However, since $u : \mathrm{U}\, f_0\, f_1$ and $t : \mathrm{T}\, f_0\, f_1\, u \to \mathrm{U}\, f_0\, f_1$ by assumption, we need to check that $\mathrm{T}\, f_0\, f_1\, u : \mathrm{Set}$ and $(\mathrm{T}\, f_0\, f_1) \circ t : \mathrm{U}\, f_0\, f_1 \to \mathrm{Set}$ in order to justify that $\mathrm{T}\, f_0\, f_1\, a : \mathrm{Set}$. Here we rely on the fact that the process of lazy computation of canonical forms of the elements of $\mathrm{U}\, f_0\, f_1$ and the associated checking of matching conditions is well-founded.

- The case where the canonical form of a is $\mathrm{c}_1\, a_0\, t\, b$ is justified in a similar way.

- There is one case for each constructor of codes for standard set formers.

This justification corresponds to the proof by rule induction that the second set-theoretic model validates the typing rule for T in Theorem 4.2. It is tempting to say that the informal justification for this typing rule is "by induction" on the generation of the elements of $\mathrm{U}\, f_0\, f_1$. However, this is misleading, since such meta-mathematical induction cannot be invoked in these *pre-mathematical* justifications.

Justification of the equality rules for T. This is a typed equality (although the type is not displayed in the Agda code) which is an easy consequence of the typing rule for T. Moreover, since the two sides of the equations have the same canonical form, it is immediate that they are equal.

8 Conclusion and related research

We have provided meaning explanations for Setzer's Mahlo universe in type theory. Thus, we claim that it is a predicative notion in Martin-Löf's extended sense.

The external Mahlo universe Set differs from other universe constructions in type theory in that we cannot directly use the formation rule for the subuniverses as a matching condition when we explain their meaning. Instead, we propose a modified condition that yields a well-founded matching process. With the modified condition, we avoid the circularity that breaks well-foundedness and can no longer justify the elimination rule that leads to the inconsistency discovered by Palmgren.

We remark that the situation here is different from the extended predicative Mahlo universe in Explicit Mathematics, which does have an elimination rule [16].

The reader may wonder why our theory $\mathbf{TT}^{\mathsf{M}}$ has subuniverses à la Tarski, but an external Mahlo universe Set à la Russell. One reason is that we base our theory on Martin-Löf's logical framework (and its implementation in Agda), where Set is à la Russell, but where the inductive-recursive definition of the subuniverses has to be implemented à la Tarski. A more fundamental reason is that universes à la Russell reflect set *constructors* such as $\Pi, \Sigma, \mathrm{N}, \ldots$. However, the subuniverse construction is an example of an inductive-recursive definition that reflects the family operators f_0, f_1 and these are not set constructors.

Our analysis of predicativity can be adapted to the internal Mahlo universe $\mathrm{M} : \mathrm{Set}$. The analysis is analogous and only slightly more complicated, since M is a universe à la Tarski with decoding $\mathrm{S} : \mathsf{M} \to \mathrm{Set}$. We refer to the Appendix for an Agda implementation.

We would like to make some remarks on the role of our second set-theoretic model as a meta-mathematical counterpart to the pre-mathematical meaning explanations. The reader may be surprised that we refer to several non-constructive and impredicative features: set-theoretic function spaces that include uncomputable functions, powersets, the axiom of choice, and classical large cardinals. The reason is that although we do not want to rely on either of these principles, the mathematical work needed for building a set-theoretic model adds precision and insight to the informal meaning explanations.

Our set-theoretic models could be constructivized by working in Aczel's predicative CZF with the regular extension axiom and suitable axioms for universes, including axioms for Mahlo universes. However, less is gained than it seems since it would mean that we take similar principles for granted as those which we wish to analyse. The constructivity and predicativity of CZF (and its extensions) relies on its interpretation in Martin-Löf type theory [3, 5, 7]. In order to interpret axioms for Mahlo universes in CZF we need Mahlo universes in type theory.

Moreover, in Section 3 we discussed the possibility of working in Kripke-Platek set theory, but deemed it sufficient for our purpose to work in the more familiar classical set theory ZFC with inaccessible and Mahlo cardinals.

Yet another possibility is to support the informal meaning explanations by constructing a realizability model [2, 4, 12]. Such a model may seem more satisfactory constructively, since it starts with a set of terms (the realizers) and explicitly formalizes the computation process. Although this would provide a more fine-grained formal analysis of the meaning explanations, its advantage is to some extent illusory. Such an approach also depends on the metalanguage, which needs strong features to ensure the existence of certain inductive and inductive-recursive definitions employed by the model.

The limits of Martin-Löf type theory according to Rathjen. In [37] and its slightly extended version [39] (see also [38]) Michael Rathjen investigates the constructive principles of

Martin-Löf type theory. There are several points of contact between his articles and ours.

First and foremost, Rathjen aims at establishing an upper bound on predicativity in Martin-Löf's sense, while we try to improve the lower bound by arguing that Setzer's Mahlo universe is indeed predicative in the sense of Martin-Löf's meaning explanations. Since the two frameworks are different, we don't claim to establish a lower bound in a mathematical sense for what strength can be reached using Rathjen's setting. Whether Rathjen's framework establishes an upper bound on any future extension of Martin-Löf type theory is of course an open-ended question subject to debate (see the reviews [31] and [45]).

Another similarity is that both Rathjen's articles and the present one make use of set-theoretic modelling of type-theoretic universes. Like us, Rathjen points out that the Mahlo universe constitutes a substantial step beyond previously defined higher universes such as superuniverses and superjump universes. On page 421 of [39] he writes:

> Setzer's theory is stronger than those based on higher type universes. It provides an important step for expanding the realm of Martin-Löf type theory. The difference between **TTM** and the systems above is that **TTM** introduces a new construction principle which is not foreshadowed in Martin-Löf's original papers. This is witnessed by the fact that models for Setzer's Mahlo universe are generated by a non-monotonic inductive definition (see Section 6) and, furthermore, by an observation due to Palmgren which shows it to be incompatible with elimination rules for the universe. In a sense, **TTM** means a paradigm shift to a new Martin-Löf type theory in that the rules for forming the elements of a type are no longer required to be monotonic.

While Rathjen views the Mahlo universe as generated by a non-monotone inductive definition, we show here how to generate the Mahlo universe and its subuniverses with their decodings by rule sets in Aczel's sense.

Rathjen's upper bound is given by the theory $\mathbf{T} := \mathbf{KP^r} + \forall x.\exists M.(x \in M \wedge M \prec_1 V)$, where $\mathbf{KP^r}$ is Kripke Platek set theory, but with the foundation scheme restricted to sets, and the additional axiom states that every set is contained in a transitive set which is a Σ_1 elementary substructure of the set-theoretic universe V. Rathjen motivates his upper bound by an analysis of the (possibly non-monotone) inductive definitions conjectured to be sufficient for modelling any future extension of Martin-Löf type theory. We refer to Rathjen's articles for details and discussion.

Furthermore, Rathjen [37] proves in Theorem 5 that $\mathbf{T}$ has the same proof theoretic strength as $(\mathbf{\Pi^1_2 - CA})\restriction$, that is, Π^1_2-comprehension with induction on natural numbers restricted to sets. Our lower bound is the strength of the Mahlo universe $|\mathrm{KPM}^+|$ ([46, 44]), which is well below $|(\mathbf{\Pi^1_2 - CA})\restriction|$.

It would be interesting to further investigate the relationship between Rathjen's ideas and ours. For example, we could try to make use of Richter and Aczel's double inductive definitions [41] (Def. 7.1). However, this is beyond the scope of the present paper.

A Agda implementation of the internal Mahlo universe

In Section 2 we defined the external Mahlo universe. The Git repository [17] also contains a version with closure under the basic set constructions.

We now also define an internal Mahlo universe in Agda. We will use it in Appendix B to implement Palmgren's proof that adding a natural elimination rule for the internal Mahlo universe leads to an inconsistency. This proof uses that the Mahlo universe is closed under the

set formers $\bot$ and $\to$. We therefore include constructors for their codes in the definition of the Mahlo universe. (The full definition should include closure under all standard set formers, but they are omitted here.)

We first define the empty type $\bot$:

```
data ⊥ : Set where

¬ : Set → Set
¬ X = X → ⊥
```

The internal Mahlo universe closed under the subuniverse set former U and under $\bot$ and $\to$ is defined as follows. Note that we have not included the closure of the subuniverse under the standard set formers in the code since this plays no role in the proof of Palmgren's paradox.

```
data M : Set where
  U'     : (f₀ : (x₀ : M) → (S x₀ → M) → M)
           (f₁ : (x₀ : M) → (x₁ : S x₀ → M) → S (f₀ x₀ x₁) → M)
           → M
  ⊥'     : M
  _→'_ : M → M → M

S : M → Set
S (U' f₀ f₁) = U f₀ f₁
S ⊥'         = ⊥
S (a →' b) = S a → S b

data U (f₀ : (x₀ : M) → (S x₀ → M) → M)
       (f₁ : (x₀ : M) → (x₁ : S x₀ → M) → S (f₀ x₀ x₁) → M) : Set where
  c0 : (x₀ : U f₀ f₁) → (S (T f₀ f₁ x₀) → U f₀ f₁)
       → U f₀ f₁
  c1 : (x₀ : U f₀ f₁) → (x₁ : (S (T f₀ f₁ x₀) → U f₀ f₁))
       → S (T f₀ f₁ (c0 x₀ x₁))
       → U f₀ f₁

T : (f₀ : (x₀ : M) → (S x₀ → M) → M)
    (f₁ : (x₀ : M) → (x₁ : S x₀ → M) → S (f₀ x₀ x₁) → M)
    → U f₀ f₁ → M
T f₀ f₁ (c0 x₀ x₁) = f₀ (T f₀ f₁ x₀) (λ x₀ → T f₀ f₁ (x₁ x₀))
T f₀ f₁ (c1 x₀ x₁ t) = f₁ (T f₀ f₁ x₀) (λ x₀ → T f₀ f₁ (x₁ x₀)) t
```

B Agda implementation of Palmgren's paradox

We implement a proof in Agda of Palmgren's paradox [33], that is, the inconsistency of the Mahlo universe with a natural elimination rule. Our presentation is an adaptation of the proof of the inconsistency of an elimination rule for the axiomatic Mahlo universe in Explicit Mathematics in [16]. There we defined a general recursion operator and then a fixed point of the function that maps a set to its negation. We refer to that article for more explanation.

We use the Agda code for the internal Mahlo universe M in appendix A and add Palmgren's elimination rule to it:

```
M-elim : {C : M → Set}
       → (d_u : (f₀ : ((x₀ : M) → (S x₀ → M) → M))
              → (f₁ : ((x₀ : M) → (x₁ : S x₀ → M) → S (f₀ x₀ x₁) → M))
              → C (U' f₀ f₁))
       → (d⊥ : C ⊥')
       → (d→ : (x y : M) → C x → C y → C (x →' y))
       → (x₀ : M) → C x₀
M-elim d_u d⊥ d→ (U' f₀ f₁) = d_u f₀ f₁
M-elim d_u d⊥ d→ ⊥'         = d⊥
M-elim d_u d⊥ d→ (a →' b) = d→ a b (M-elim d_u d⊥ d→ a)
                                   (M-elim d_u d⊥ d→ b)
```

In the version in Explicit Mathematics, where the argument of U simply was a function $f : \mathsf{M} \to \mathsf{M}$, we used the elimination rule to define:

$$\begin{array}{ll} \mathrm{ap} : \mathsf{M} \to \mathsf{M} \to \mathsf{M} & \mathrm{emb} : (\mathsf{M} \to \mathsf{M}) \to \mathsf{M} \\ \mathrm{ap}\ (\mathrm{U}\ f)\ x = f\ x & \mathrm{emb}\ f = \mathrm{U}\ f \end{array}$$

and obtained ap (emb f) $x = f\ x$. In this way, we could simulate the untyped lambda calculus and define the Y-combinator. Thus, we could define the fixed point of $\lambda x.x \to' \bot'$ and thus derive an inconsistency.

In type theory this is slightly more complicated since the argument of U' is $f : \mathrm{TFam}(\mathsf{M}) \to \mathrm{TFam}(\mathsf{M})$ or more precisely its components (f_0, f_1).

So we need to lift functions $f : \mathsf{M} \to \mathsf{M}$ to $f' : \mathrm{TFam}(\mathsf{M}) \to \mathrm{TFam}(\mathsf{M})$.
In order to do this, we define two dummy elements:

- dum will be used to lift $x : \mathsf{M}$ to $(x, \mathsf{dum}) : \mathrm{TFam}(\mathsf{M})$
- dum' will be used to lift a function $f : \mathsf{M} \to \mathsf{M}$ to $(f_0, \mathsf{dum}') : \mathrm{TFam}(\mathsf{M}) \to \mathrm{TFam}(\mathsf{M})$, where $f_0 = \lambda u\ \_ \to f\ u$.

```
dum : {x : M} → S x → M
dum a = ⊥'

dum' : {x : M → M}(x₀ : M) (x₁ : S x₀ → M) → S (x x₀) → M
dum' _ _ _ = ⊥'
```

We define ap : M such that ap (U' f g) $x = f\ x$ dum.

```
ap : M → M → M
ap = M-elim (λ f₀ _ x → f₀ x dum)
            (λ _ → ⊥')
            (λ _ _ _ _ _ → ⊥')
```

We define emb : (M → M) → M, s.t. ap (emb f) $y = f\ y$:

```
emb : (M → M) → M
emb f = U' (λ u _ → f u) dum'
```

Now we can define the Y-combinator:

```
y : (M → M) → M
y k = emb (λ x → k (ap x x))

Y : (M → M) → M
Y k = ap (y k) (y k)
```

We define negation

```
I : M → M
I x = x →' ⊥'
```

and its fixed point:

```
a : M
a = Y I

A : Set
A = S a
```

We have definitionally $a = a \to' \bot'$ and therefore $A = \neg A$. However, a doesn't normalise (because it reduces to a term definitionally equal to $a \to' \bot'$ and we therefore get an infinite reduction sequence). Because of this, Agda doesn't infer these equalities. Instead, we define

```
p₁ : A → ¬ A
p₁ x = x

p₂ : ¬ A → A
p₂ x = x
```

and the inconsistency follows:

```
p₃ : ¬ A
p₃ x = p₁ x x

p₄ : A
p₄ = p₂ p₃

inconsistent : ⊥
inconsistent = p₃ p₄
```

References

[1] Peter Aczel. An introduction to inductive definitions. In John Barwise, editor, *Handbook of Mathematical Logic*, chapter C.7, pages 739–782. North Holland, 1977. doi:10.1016/S0049-237X(08)71120-0.

[2] Peter Aczel. The strength of Martin-Löf's type theory with one universe. In S. Miettinen and J. Väänanen, editors, *Proceedings of the Symposium on Mathematical Logic (Oulu 1974)*, pages 1–32, 1977. Report No 2 of Dept. Philosophy, University of Helsinki.

[3] Peter Aczel. The type theoretic interpretation of constructive set theory. In A. MacIntyre, L. Pacholski, and J. Paris, editors, *Logic Colloquium '77*, pages 55–66. North-Holland, 1978. doi:10.1016/S0049-237X(08)71989-X.

[4] Peter Aczel. Frege structures and the notions of proposition, truth, and set. In H. Jerome Keisler Jon Barwise and Kenneth Kunen, editors, *The Kleene Symposium*, volume 101, pages 31–59. North-Holland, 1980. doi:10.1016/S0049-237X(08)71252-7.

[5] Peter Aczel. The type theoretic interpretation of constructive set theory: inductive definitions. In Barcan Marcus et al., editors, *Logic, Methodology, and Philosophy of Science VII*, pages 17–49. Elsevier Science Publishers, 1986. doi:10.1016/S0049-237X(08)71989-X.

[6] Peter Aczel. On relating type theories and set theories. In Thorsten Altenkirch, Wolfgang Naraschewski, and Bernhard Reus, editors, *Types for Proofs and Programs, International Workshop TYPES '98, Kloster Irsee, Germany, March 27-31, 1998, Selected Papers*, volume 1657 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 1998. doi:10.1007/3-540-48167-2_1.

[7] Peter Aczel and Michael Rathjen. Notes on constructive set theory. Technical Report 40, Institut Mittag-Leffler, The Royal Swedish Academy of Sciences, 2000/2001. ISSN: 1103-467X. ISRN IML-R–40-00/01–SE. Available from https://ncatlab.org/nlab/files/AczelRathjenCST.pdf.

[8] Agda Community. Welcome to Agda's documentation!, Retrieved 15 January 2023. URL: https://agda.readthedocs.io/.

[9] Stuart Allen. A non-type-theoretic definition of Martin-Löf's types. In *Proceedings of the Symposium on Logic in Computer Science (LICS '87), Ithaca, New York, USA, June 22-25, 1987*, pages 215–221, 1987. Available from https://ecommons.cornell.edu/server/api/core/bitstreams/c714d33b-6c35-4958-9e54-20845ba11ca4/content.

[10] Simon Castellan, Pierre Clairambault, and Peter Dybjer. Categories with families: Unityped, simply typed, and dependently typed. In Claudia Casadio and Philip J. Scott, editors, *Joachim Lambek: The Interplay of Mathematics, Logic, and Linguistics*, Outstanding Contributions to Logic. Springer, 2021. doi:10.1007/978-3-030-66545-6_5.

[11] Peter Dybjer. Internal type theory. In *TYPES '95, Types for Proofs and Programs*, number 1158 in Lecture Notes in Computer Science, pages 120–134. Springer, 1996. doi:10.1007/3-540-61780-9_66.

[12] Peter Dybjer. A general formulation of simultaneous inductive-recursive definitions in type theory. *Journal of Symbolic Logic*, 65(2):525 – 549, June 2000. doi:10.2307/2586554.

[13] Peter Dybjer. Program testing and the meaning explanations of intuitionistic type theory. In Peter Dybjer, Sten Lindström, Erik Palmgren, and Göran Sundholm, editors, *Epistemology versus Ontology - Essays on the Philosophy and Foundations of Mathematics in Honour of Per Martin-Löf*, volume 27 of *Logic, Epistemology, and the Unity of Science*, pages 215–241. Springer, 2012. doi:10.1007/978-94-007-4435-6_11.

[14] Peter Dybjer and Anton Setzer. A finite axiomatization of inductive-recursive definitions. In Jean-Yves Girard, editor, *Typed Lambda Calculi and Applications*, volume 1581 of *Lecture Notes in Computer Science*, pages 129–146. Springer, April 1999. doi:10.2307/2586554.

[15] Peter Dybjer and Anton Setzer. Indexed induction-recursion. *Journal of Logic and Algebraic Programming*, 66:1 – 49, 2006. doi:10.1016/j.jlap.2005.07.001.

[16] Peter Dybjer and Anton Setzer. The extended predicative Mahlo universe in Martin-Löf type theory. *Journal of Logic and Computation*, May 2023. doi:10.1093/logcom/exad022.

[17] Peter Dybjer and Anton Setzer. Agda Code Weyl Paper 2024 Dybjer Setzer, April 2024. URL: https://github.com/csetzer/agdaCodeWeylVolume2024SetzerDybjer. HTML version of Code available at https://csetzer.github.io/articles/weylVolume2024/agdaCodeHtml/loadAll.html.

[18] Solomon Feferman. Systems of predicative analysis. *The Journal of Symbolic Logic*, 29(1):pp. 1–30, 1964. URL: http://www.jstor.org/stable/2269764.

[19] Solomon Feferman. A language and axioms for explicit mathematics. In J. Crossley, editor, *Algebra and Logic*, volume 450 of *Lecture Notes in Mathematics*, page 87–139. Springer, 1975. doi:10.1007/BFb0062852.

[20] Solomon Feferman. Iterated inductive fixed-point theories: Application to Hancock's conjecture.

In George Metakides, editor, *Patras Logic Symposion*, volume 109 of *Studies in Logic and the Foundations of Mathematics*, pages 171 – 196. Elsevier, 1982. doi:10.1016/S0049-237X(08)71364-8.

[21] Peter Hancock. *Ordinals and interactive programs*. PhD thesis, LFCS, University of Edinburgh, 2000. URL: https://era.ed.ac.uk/bitstream/handle/1842/376/ECS-LFCS-00-421.pdf?sequence=2&isAllowed=y.

[22] Martin Hofmann. Syntax and semantics of dependent types. In *Semantics and logics of computation*, volume 14 of *Publications of the Newton Institute*, pages 79–130. Cambridge University Press, 1997. doi:10.1007/978-1-4471-0963-1_2.

[23] Reinhard Kahle and Anton Setzer. An extended predicative definition of the Mahlo universe. In Ralf Schindler, editor, *Ways of Proof Theory*, Ontos Series in Mathematical Logic, pages 309 – 334, Berlin, Boston, 2010. De Gruyter. doi:10.1515/9783110324907.315.

[24] Georg Kreisel. La prédicativité. *Bulletin de la Société Mathématique de France*, 88:371–391, 1960. Available from http://archive.numdam.org/article/BSMF_1960__88__371_0.pdf. URL: http://eudml.org/doc/86990.

[25] Georg Kreisel et al. Ordinal logics and the characterization of informal concepts of proof. In *Proceedings of the International Congress of Mathematicians*, volume 14, pages 289–299. Cambridge University Press, 1958. URL: https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM1958/ICM1958.ocr.pdf.

[26] Per Martin-Löf. Philosophical aspects of intuitionistic type theory. Lectures given at the Faculteit der Wijsbegeerte, Rijksuniversiteit Leiden, 23 September – 16 December, 1993, edited and transferred to LaTeXby Ansten Klev. URL: https://pml.flu.cas.cz/uploads/PML-LeidenLectures93.pdf.

[27] Per Martin-Löf. Sets, types, and categories. Lecture given at the Symposium of Constructive Type Theory, Rijksuniversiteit Leiden, 6 February 2004, edited and transferred to LaTeXby Ansten Klev. URL: https://pml.flu.cas.cz/uploads/PML-Leiden06Feb04.pdf.

[28] Per Martin-Löf. An intuitionistic theory of types: predicative part. In H.E. Rose and J.C. Shepherdson, editors, *Logic Colloquium '73, Proceedings of the Logic Colloquium*, volume 80 of *Studies in Logic and the Foundations of Mathematics*, pages 73–118. North-Holland, 1975. doi:10.1016/S0049-237X(08)71945-1.

[29] Per Martin-Löf. Constructive mathematics and computer programming. In L. Jonathan Cohen, Jerzy Łoś, Helmut Pfeiffer, and Klaus-Peter Podewski, editors, *Logic, Methodology and Philosophy of Science VI, Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science, Hannover 1979*, volume 104 of *Studies in Logic and the Foundations of Mathematics*, pages 153–175. North-Holland, 1982. doi:10.1016/S0049-237X(09)70189-2.

[30] Per Martin-Löf. *Intuitionistic type theory*, volume 1 of *Studies in Proof Theory*. Bibliopolis, 1984. URL: http://www.cs.cmu.edu/afs/cs/user/crary/www/819-f09/Martin-Lof80.pdf.

[31] Gregori E. Mints. Review on: Rathjen, Michael. The constructive Hilbert program and the limits of Martin-Löf type theory. Synthese 147, No 1, 81-120, 2005, 2005. Zentralblatt Math, Zbl 1108.03056. URL: https://zbmath.org/1108.03056.

[32] Bengt Nordström, Kent Petersson, and Jan Smith. *Programming in Martin-Löf's Type Theory: an Introduction*. Oxford University Press, 1990. Book out of print. Online version available via http://www.cs.chalmers.se/Cs/Research/Logic/book/.

[33] Erik Palmgren. On universes in type theory. In G. Sambin and J.M. Smith, editors, *Twenty-five years of constructive type theory: Proceedings of a Congress held in Venice, October 1995*, volume 36, pages 191 – 204. Oxford University Press, 1998. https://global.oup.com/academic/product/twenty-five-years-of-constructive-type-theory-9780198501275?cc=gb&lang=en&.

[34] Erik Palmgren. From type theory to setoids and back. *Math. Struct. Comput. Sci.*, 32(10):1283–1312, 2022. doi:10.1017/S0960129521000189.

[35] Michael Rathjen. Ordinal notations based on a weakly Mahlo cardinal. *Archive for Mathematical Logic*, 29:249 – 263, 1990. doi:10.1007/BF01651328.

[36] Michael Rathjen. Proof-theoretic analysis of KPM. *Archive of Mathematical Logic*, 30(5 – 6):377 – 403, September 1991. doi:10.1007/BF01621475.

[37] Michael Rathjen. The constructive Hilbert program and the limits of Martin-Löf type theory. *Synthese*, 147:81 – 120, 2005. doi:10.1007/978-1-4020-8926-8_17.

[38] Michael Rathjen. Universes and the limits of Martin-Löf type theory, 15 March 2008. Slides of talk presented at Russell'08 Proof Theory meets Type Theory, Swansea. https://csetzer.github.io/russell08/index.html. URL: https://csetzer.github.io/russell08/abstracts/index.html.

[39] Michael Rathjen. The constructive Hilbert program and the limits of Martin-Löf type theory. In Sten Lindström, Erik Palmgren, Krister Segerberg, and Viggo Stoltenberg-Hansen, editors, *Logicism, Intuitionism, and Formalism: What has Become of Them?*, volume 341 of *Synthese Library*, pages 397–433, Dordrecht, 2009. Springer Netherlands. doi:10.1007/978-1-4020-8926-8_17.

[40] Michael Rathjen, Edward Griffor, and Erik Palmgren. Inaccessibility in constructive set theory and type theory. *Annals of Pure and Applied Logic*, 94(1-3):181–200, 1998. doi:10.1016/S0168-0072(97)00072-9.

[41] Wayne Richter and Peter Aczel. Inductive definitions and reflecting properties of admissible ordinals. In J. E. Fenstad and P. G. Hinman, editors, *Generalized recursion theory*, volume 79 of *Studies in Logic and the Foundations of Mathematics*, pages 301 – 381, Amsterdam, 1973. North-Holland. doi:10.1016/S0049-237X(08)70592-5.

[42] Kurt Schütte. Eine Grenze Für die Beweisbarkeit der Transfiniten Induktion in der Verzweigten Typenlogik. *Archive for Mathematical Logic*, 7:45–60, 1964. doi:10.1007/BF01972460.

[43] Kurt Schütte. Predicative well-orderings. In J.N. Crossley and M.A.E. Dummett, editors, *Formal Systems and Recursive Functions*, volume 40 of *Studies in Logic and the Foundations of Mathematics*, pages 280–303. Elsevier, 1965. URL: https://www.sciencedirect.com/science/article/pii/S0049237X0871694X, doi:https://doi.org/10.1016/S0049-237X(08)71694-X.

[44] Anton Setzer. Extending Martin-Löf type theory by one Mahlo-universe. *Arch. Math. Log.*, 39:155 – 181, 2000. doi:10.1007/s001530050140.

[45] Anton Setzer. Review on: Rathjen, Michael. The constructive Hilbert program and the limits of Martin-Löf type theory. Synthese 147, No 1, 81-120, 2005, 2006. Mathrev, MR2182643 (2006m:03023). URL: https://mathscinet.ams.org/mathscinet/article?mr=2182643.

[46] Anton Setzer. Universes in type theory part I – Inaccessibles and Mahlo. In A. Andretta, K. Kearnes, and D. Zambella, editors, *Logic Colloquium '04*, Lecture Notes in Logic, pages 123 – 156. Association of Symbolic Logic, Lecture Notes in Logic 29, Cambridge University Press, 2008. doi:10.1017/CBO9780511721151.009.

[47] Nik Weaver. Predicativity beyond Gamma_0, 2009. arXiv:math/0509244.

[48] Hermann Weyl. *Das Kontinuum. Kritische Untersuchungen über die Grundlagen der Analysis*. Veit, Leipzig, 1918. doi:10.1515/9783112451144.

.