On Universes in Type Theory

Erik Palmgren[1]
Uppsala University

1 Introduction

The notion of a *universe of types* was introduced into constructive type theory by Martin-Löf (1975). According to the propositions-as-types principle inherent in type theory, the notion plays two rôles. The first is as a collection of sets or types closed under certain type constructions. The second is as a set of constructively given infinitary formulas. In this paper we discuss the notion of universe in type theory and suggest and study some useful extensions. We assume familiarity with type theory as presented in e.g. (Martin-Löf 1984).

Universes have been effective in expanding the realm of constructivism. One example is constructive category theory where type universes take the rôles of Grothendieck universes of sets, in handling large categories. A more profound example is Aczel's (1986) type-theoretic interpretation of constructive set theory (CZF). It is done by coding $\in$-diagrams into well-order types, with branching over an arbitrary type of the universe. The latter generality is crucial to interpret the separation axiom. The introduction of universes and well-orders (W-types) in conjunction gives a great proof-theoretic strength. This has provided constructive justification of strong subsystems of second order arithmetic studied by proof-theorists (see Griffor and Rathjen (1994) and Setzer (1993), and for some early results, see Palmgren (1992)). At present, it appears that the most easily justifiable way to increase the proof-theoretic strength of type theory is to introduce ever more powerful universe constructions. We will give two such extensions in this paper. Besides contributing to the understanding of subsystems of second order arithmetic and pushing the limits of inductive definability, such constructions provide intuitionistic analogues of large cardinals (Rathjen *et al.* to appear). A third new use of universes is to facilitate the incorporation of classical reasoning into constructive type theory. We introduce a universe of classical propositions and prove a conservation result for 'Π_2-formulas'. Extracting programs from classical proofs is then tractable within type theory.

The next section gives an introduction to the notion of universe. The central part of the paper is Section 3 where we introduce a universe forming operator and a super universe closed under this operator. Section 4 summarises what is known

[1]The research reported herein was supported partly by the Swedish Research Councils for Natural Sciences (NFR) and Engineering Sciences (TFR), and partly by the EU Project Twinning: Proof Theory and Computation (contract SC1*-CT91-0724 (TSTS)). *Author's current address:* Department of Mathematics, Chalmers University of Technology and the University of Göteborg, S-412 96 Göteborg. E-mail: epalmgr@math.chalmers.se.

about the proof-theoretic strength of this extension, mainly results due to M. Rathjen. In Section 5 we introduce the notion of higher order universe operators. While all of the preceding development is predicative, it is also possible to define impredicative theories using universes. In Section 6 we point out some dangers in combining such ideas with elimination rules. In particular, we discuss Setzer's Mahlo universe. Finally, in Section 7 we construct the classical universe.

2 Universes

From an abstract point of view a type universe is simply a type of types closed under certain type constructions. Being a type of types can be formulated in essentially two ways (Martin-Löf 1984): à la Tarski, by introducing a type of codes U for types and a decoding function $T(\cdot)$:

$$U\ type \qquad \frac{a \in U}{T(a)\ type}$$

or, alternatively, à la Russell by simply introducing U and identifying codes and types

$$U\ type \qquad \frac{A \in U}{A\ type}.$$

The Russell formulation should be regarded as an informal version of the Tarski formulation, but is too unclear when nesting universe constructions, e.g. as in the super universe. Thus we use the Tarski formulation for complete precision.

Modern presentations of type theory employ a so called *logical framework* (Nordström *et al.* 1990). This a typed lambda calculus with a dependent function space construction (Π-types) and a universe of types $(Set, El(\cdot))$. The types of this universe are called *sets*. In this framework different type theories can be specified by giving closure conditions to the sets, and by introducing constants and computation rules to types constructed from sets. Later extensions have also Σ-types or *records*. We shall here present the rules for the extensions of type theory in the older, more readable style of Martin-Löf (1984), as far as possible. (In section 5 we need however a logical framework with Σ-types.)

In Martin-Löf's type theory two different conceptions of universes occur. The first captures the idea of reflection of the judgement forms $A\ set$ and $A = B$ into a hierarchy of universes (U_n, T_n) *externally* indexed by $n = 1, 2, 3, \ldots$. That is, whenever $A\ set$ then in some universe U_n, there is a code a so that $T_n(a) = A$, and if $A = B$ with $T_n(a) = A$ and $T_n(b) = B$, then $a = b \in U_n$. This is as in Martin-Löf (1975; 1982) albeit there it is formulated à la Russell. The second idea, which is preferable, is to uniformly construct universes above earlier universes (hinted at in Martin-Löf (1984), p. 89).

Universes as full reflections. We view the formation of the hierarchy (U_1, T_1), $(U_2, T_2), \ldots$ as a process. At first there is no universe. Then we introduce a universe U_1 of codes for all basic sets,

$$U_1\ set \qquad \frac{x \in U_1}{T_1(x)\ set}$$

$$n_0^1, n_1^1, n^1 \in U_1 \qquad T_1(n_0^1) = N_0 \quad T_1(n_1^1) = N_1 \quad T_1(n^1) = N,$$

where N_0 is the empty set, N_1 is the set with single element 0_1, and N is the set of natural numbers. Furthermore we assume that it is *closed under* Π-formation:

$$\frac{\begin{array}{c}(x \in T_1(a)) \\ \vdots \\ a \in U_1 \qquad b \in U_1\end{array}}{\pi(a,(x)b) \in U_1} \qquad \frac{\begin{array}{c}(x \in T_1(a)) \\ \vdots \\ a \in U_1 \qquad b \in U_1\end{array}}{T_1(\pi(a,(x)b)) = (\Pi x \in T_1(a))T_1(b),}$$

and we also assume that (U_1, T_1) is similarly closed under $\Sigma, I, +$ and other set formers, if desired. Hence for every set A formed without universes there is $a \in U_1$ so that $T_1(a) = A$. At this stage there is no difference between the two versions of universes. We have that if $A = B$ is formed without the use of universes, then $a = b \in U_1$ for some a and b such that $T_1(a) = A$, $T_1(b) = B$, by the usual equalities that come with every canonical constant. Then at the next stage we introduce a new universe (U_2, T_2) closed under the set formers $\Pi, \Sigma, \ldots$ with new codes for all sets

$$n_0^2, n_1^2, n^2 \in U_2 \qquad T_2(n_0^2) = N_0 \quad T_2(n_1^2) = N_1 \quad T_2(n^2) = N.$$

But U_1 and $T_1(x)$ are sets of the previous stage, so we must also introduce

$$u_1^2 \in U_2 \qquad T_2(u_1^2) = U_1 \qquad \frac{x \in U_1}{t_1^2(x) \in U_2}.$$

We have a host of new set equalities to reflect in U_2: $T_1(n_k^1) = N_k$ and $T_1(n^1) = N$ give

$$t_1^2(n_k^1) = n_k^2 \in U_2 \qquad t_1^2(n^1) = n^2 \in U_2$$

and since $T_1(\pi(a,(x)b)) = (\Pi x \in T_1(a))T_1(b)$ we should assume

$$t_1^2(\pi(a,(x)b)) = \pi(t_1^2(a),(x)t_1^2(b)) \in U_2$$

and so on for all codes for set formers. We may also express this as: t_1^2 is a homomorphism with respect to set constructors, extending $t_1^2(n_k^1) = n_k^2$ and $t_1^2(n^1) = n^2 \in U_2$. At each step in the construction of the hierarchy of universes we introduce new codes and equalities between codes for sets and set equalities which can be formed. Having completed the hierarchy $(U_1, T_1), (U_2, T_2), \ldots, (U_n, T_n), \ldots$, $n < \omega$ we notice that any proof in the resulting system can only use finitely many universes (proofs are finite and the universes are *externally* indexed) and hence the reflection principle holds for both sets and set equalities.

If we try to iterate this process into the transfinite we run into something like

$$u_1^\omega, u_2^\omega, u_3^\omega, \ldots \in U_\omega$$

a universe which has infinitely many introduction rules. Thus it is impossible to formulate an elimination rule without having some kind of internal indexing of the universes.

Universes as uniform constructions. Here we do not reflect set equalities. This allows us to simply inject the codes for sets from an earlier universe into the next. We construct a hierarchy of universes $(U_1, T_1), (U_2, T_2), \ldots$ stepwise. Assume that (U_n, T_n) has already been constructed, then

$$U_{n+1}\ set \qquad \frac{x \in U_{n+1}}{T_{n+1}(x)\ set}$$

and

$$u_n \in U_{n+1} \qquad T_{n+1}(u_n) = U_n$$

$$\frac{x \in U_n}{t_n(x) \in U_{n+1}} \qquad \frac{x \in U_n}{T_{n+1}(t_n(x)) = T_n(x)}.$$

Thus $t_n(a)$ is now considered to be a canonical element in U_{n+1}, and is regarded as a copy of a in U_{n+1}. As an example, note that the code for U_j, $j < n+1$ in U_{n+1} is $t_n(t_{n-1}(\ldots t_{j+1}(u_j)\ldots))$. We furthermore assume that (U_{n+1}, T_{n+1}) is closed under the same set formers as (U_n, T_n). The construction of (U_{n+1}, T_{n+1}) depends thus only on *the family* (U_n, T_n). Observe that we still reflect the judgement form $A\ set$.

It seems that the idea of universes as full reflections is difficult to formulate for transfinite hierarchies. The usefulness of reflecting equalities of sets is not clear. Thus we shall only consider hierarchies of universes built using the uniform construction.

Remark The formation of the next universe was formulated as an operator in the domain-theoretic model (Palmgren 1993) of partial type theory. This leads to the formalisation of universe operators in the next section.

3 Universe Operators and Super Universes

By having universe formation as an operator and a *super universe* closed under this operator we may form transfinite level universes much the same way as we may form transfinite sets using an ordinary universe. The universe forming operation acts on families of sets. We can form a universe $(U(A,(x)B), T(A,(x)B))$ above *any family of sets* $(A,(x)B)$

$$\frac{\begin{array}{c}(x \in A)\\ \vdots \\ A\ set \qquad B\ set\end{array}}{U(A,(x)B)\ set} \qquad \frac{a \in U(A,(x)B)}{T(A,(x)B,a)\ set}.$$

We assume that $(U(A,(x)B), T(A,(x)B))$ is closed under the usual set formers $\Pi, \Sigma, +, Id$. That the universe is above the family $(A,(x)B)$ is expressed by

$$*(A,(x)B) \in U(A,(x)B) \qquad T(A,(x)B, *(A,(x)B)) = A$$

$$\frac{a \in A}{\ell(A,(x)B,a) \in U(A,(x)B)} \qquad \frac{a \in A}{T(A,(x)B,\ell(A,(x)B,a)) = B(a/x)}$$

Thus $*(A,(x)B)$ is a code for A in $U(A,(x)B)$, and $\ell(A,(x)B,a)$ is a copy of the code a in A for $B(a/x)$. If we assume that (U_0,T_0) is some basic family of sets, then we can define the hierarchy of universes as follows

$$U_{n+1} := U(U_n,(x)T_n(x)) \qquad T_{n+1}(a) := T(U_n,(x)T_n(x),a),$$

and let $u_n := *(U_n,(x)T_n(x))$ and $t_n(a) := \ell(U_n,(x)T_n(x),a)$.

The super universe. We now consider a universe (V,S) — *the super universe* — which in addition to being closed under the set formers $\Pi, \Sigma, +, I$ is also closed under universe formation. Moreover we assume that it contains basic sets. The closure under the universe operator is given by

$$\frac{\begin{array}{c}(x \in S(a)))\\ \vdots \\ a \in V \qquad b \in V\end{array}}{u(a,(x)b) \in V} \qquad \frac{\begin{array}{c}(x \in S(a)))\\ \vdots \\ a \in V \qquad b \in V\end{array}}{S(u(a,(x)b)) = U(S(a),(x)S(b))}$$

$$\frac{\begin{array}{c}(x \in S(a))\\ \vdots \\ a \in V \qquad b \in V \qquad c \in S(u(a,(x)b))\end{array}}{t(a,(x)b,c) \in V} \qquad \frac{\begin{array}{c}(x \in S(a))\\ \vdots \\ a \in V \qquad b \in V \qquad c \in S(u(a,(x)b))\end{array}}{S(t(a,(x)b,c)) = T(S(a),(x)S(b),c)}$$

Note that $u(a,(x)b)$ and $t(a,(x)b,c)$ are canonical elements. The term $t(a,(x)b,\cdot)$ injects codes from the universe $U(S(a),(x)S(b))$ into V. The super universe has an inductive structure and it is not difficult to formulate an elimination rule for it.

Transfinite hierarchies. Examples of a transfinite sets can easily be constructed using recursion and a universe (cf. Martin-Löf (1975), p. 83). Transfinite level universes are, however, more complicated to construct since they are to be given as families of sets. Consider the set of all codes for families in the super universe V

$$\mathcal{F}_V := (\Sigma x \in V)[S(x) \to V].$$

In the following, let $\langle \cdot,\cdot \rangle$ denote pairing and p,q the first and second projection respectively. Define $B_V(c) := S(p(c))$ for $c \in \mathcal{F}_V$, the base of the family coded by c, and $F_V(c,x) := S(Ap(q(c),x))$ for $x \in B_V(c)$, the family of sets over $B_V(c)$ coded by $c \in \mathcal{F}_V$. We shall define $\hat{u} \in \mathcal{F}_V \to \mathcal{F}_V$, such that

$$B_V(\hat{u}(c)) = U(B_V(c),(x)F_V(c,x)) \quad F_V(\hat{u}(c),w) = T(B_V(c),(x)F_V(c,x),w).$$

and this is achieved by

$$\hat{u} := (\lambda c)\langle u(p(c), (x)Ap(q(c), x)), (\lambda w)t(p(c), (x)Ap(q(c), x), w)\rangle.$$

Hence if c is a code for a universe, then $\hat{u}(c)$ is a code for the universe above c.

Let c_0 be a code for a suitable basic family of sets. By recursion we can define $\hat{u}^n(c_0)$ $(n \in N)$ the codes for finite iterates of universes above c_0. Then

$$U_\omega := U((\Sigma n \in N)B_V(\hat{u}^n(c_0)), (z)F_V(\hat{u}^{p(z)}(c_0), q(z)))$$

is a universe of transfinite level, and it is straightforward to find its code in the super universe.

4 Proof-theoretic Strength

Type theory with one universe closed under the W-set, named $\mathbf{ML}_1\mathbf{W}$, is proof-theoretically very strong, among the theories that have so far been given a complete constructive justification. Recall that the W-set is a general inductive set former, by which one may construct the Brouwer ordinals as well-founded trees which branch over a given family of sets. A slight weakening of $\mathbf{ML}_1\mathbf{W}$ has the strength of Kripke–Platek set theory extended with a principle corresponding to the existence of an inaccessible cardinal (Griffor and Rathjen 1994). Independently, Setzer (1993) determined the strength of the full theory.

It is interesting to note that universes give strength already without W-sets. Let $\gamma_0 = \varepsilon_0$ and let $\gamma_{n+1} = \varphi_{\gamma_n}0$, where $(\varphi_\alpha)_\alpha$ are the Veblen functions, i.e. $\varphi_0(\xi) = \omega^\xi$ and, for $\alpha > 0$, φ_α is the enumeration function for the common fixed points of the functions φ_β $(\beta < \alpha)$. Aczel (1977) showed that the strength of type theory with one universe is γ_1. Hancock's conjecture (cf. Martin-Löf (1975)) stated that the strength of type theory with n universes is γ_n, and was proved by Feferman (1982). From this it follows that the strength of type theory with arbitrarily many finite level universes is the limit of $(\gamma_n)_n$, i.e. Γ_0. The latter result was achieved independently by Aczel.

In a previous version of the present paper we interpreted an intuitionistic version of a theory ATR using an internally indexed hierarchy (U_n, T_n) $(n \in N)$. The classical version of this theory has strength Γ_{ε_0} (cf. Simpson 1982). Subsequently, Rathjen has obtained sharp results for theories involving the super universe. One ingredient in the proof of the lower bound of the super universe is (a relativised version of) the interpretation of ATR. We summarise his results. Let **MLU** denote the type theory with the universe operator U of Section 3 and no elimination rules for U. Let **MLS** be type theory with the universe operator U and the super universe closed under this operator, as in Section 3, and no elimination rules. The variant of **MLS** where the operator U may only act on families from the super universe is called **MLS** $\uparrow$. Let $(\Phi_\alpha)_\alpha$ be defined just as the hierarchy of Veblen functions, except that $\Phi_0(\xi) = \Gamma_\xi$.

Theorem 4.1. (Rathjen 1997)

(i) $|\mathbf{MLU}| = \Gamma_0$
(ii) $|\mathbf{MLS} \upharpoonright | = \Phi_{\varepsilon_0}(0)$
(iii) $|\mathbf{MLS}| = \Phi_{\Gamma_0}(0)$

The strength of **MLS** with W-types has also been determined by Rathjen (1997). We refer to Griffor and Rathjen (1994), Palmgren (1992), Rathjen *et al.* (to appear), Setzer (1993; 1995) and the next section for further proof-theoretic results.

Remark The ordinal Γ_0 is usually called the Feferman–Schütte bound for predicativity. The *proof-theorist's notion of predicativity* is based on the idea that an ordinal is predicative if it can be reached by a certain autonomous progression of theories starting from Peano arithmetic. This is to be contrasted with what we could call the *constructivist's notion of predicativity,* which recognises a construction as predicative if it has a clear inductive structure, e.g. W-sets and super universes. Note for example that the theory **MLS** goes well beyond Γ_0. Not too many theories of strength between Γ_0 and the Howard ordinal have been found. According to the results above, universes seem to provide natural examples of such theories.

5 Higher Order Universe Operators

The notion of universe operator can be extended to all finite orders. To formulate them we use a logical framework $(Set, El(\cdot))$ with Σ-types. The Σ-types are written in boldface $(\mathbf{\Sigma}x \in B)C$ and their associated pairing function, left and right projections are denoted by $\lfloor \cdot, \cdot \rfloor$, $\mathbf{p}$ and $\mathbf{q}$ respectively. Where no confusion can arise we write A instead of $El(A)$ to simplify the presentation.

Definition 5.1 *Construct an externally indexed hierarchy of types*

$$O_0 = Set, \qquad F_n = (\mathbf{\Sigma}A \in V)(A)O_n,$$
$$O_{n+1} = (F_n)F_n.$$

Then O_n is the type of operators of order n, *and F_n is the* type of families of operators of order n.

The Theories $\mathbf{ML}^n$, $\mathbf{n = 0, 1, 2, \ldots}$. We define this sequence of theories inductively. Basic type theory with Π, Σ, $+$ and I-sets, and the basic sets N_0, N_1 and N is $\mathbf{ML}^0$. We define a type theory $\mathbf{ML}^{n+1}$ by adding to the theory $\mathbf{ML}^n$, the new functions $U_0^n, \ldots, U_n^n$, $T_0^n, \ldots, T_n^n$, $\ell_0^n, \ldots, \ell_n^n$, $*_0^n, \ldots, *_n^n$, $u_0^n, \ldots, u_{n-1}^n$ and $t_0^n, \ldots, t_{n-1}^n$. The pair U_k^n, T_k^n is used to construct a family of operators of order k from given families operators of orders $k, k+1, \ldots, n$. Thus U_0^n, T_0^n will construct the actual universe. The constants $\ell_k^n, *_k^n$ are lifting functions analogous to ℓ and $*$ for the universe operator of Section 3. The constants u_k^n, t_k^n signify the application of an operator of level k to a family of operators of level $k-1$. All these functions are canonical (constructors), except the T_k^n:s. Their axiomatisation is as follows.

Let $A_n \in Set, B_n \in (A_n)O_n, \ldots, A_0 \in Set, B_0 \in (A_0)O_0$. Write $\mathcal{P}$ for the sequence of parameters $A_n, B_n, \ldots, A_0, B_0$. For $k = 0, \ldots, n$, we assume the following rules:

$$U_k^n(\mathcal{P}) \in Set \qquad \frac{a \in U_k^n(\mathcal{P})}{T_k^n(\mathcal{P}, a) \in O_k}$$

$$\frac{a \in A_k}{\ell_k^n(\mathcal{P}, a) \in U_k^n(\mathcal{P})} \qquad \frac{a \in A_k}{T_k^n(\mathcal{P}, \ell_k^n(\mathcal{P}, a)) = B_k(a)}$$

$$*_k^n(\mathcal{P}) \in U_0^n(\mathcal{P}) \qquad T_0^n(\mathcal{P}, *_k^n(\mathcal{P})) = A_k.$$

We assume rules that states that $U_0^n(\mathcal{P}), T_0^n(\mathcal{P}, \cdot)$ is a universe closed under Π, Σ, $+$ and I and that it contains the basic sets N_0, N_1 and N. Below we abbreviate $U_k^n(\mathcal{P})$ by U_k^n and $T_k^n(\mathcal{P}, \cdot)$ by $T_k^n(\cdot)$. For $k = 1, \ldots, n$ we have the following rules for application of operators. Introduce codes for the code sets resulting when applying o to $(a, (x)b)$:

$$\frac{\begin{array}{c} \\ o \in U_k^n \qquad a \in U_0^n \qquad \begin{array}{c}(x \in T_0^n(a)) \\ \vdots \\ b \in U_{k-1}^n\end{array}\end{array}}{u_{k-1}^n(\mathcal{P}, o, a, (x)b) \in U_0^n,}$$

and under the same assumptions we have the equality

$$T_0^n(u_{k-1}^n(\mathcal{P}, o, a, (x)b)) = \mathbf{p}\Big(T_k^n(o)(\lfloor T_0^n(a), (x)T_{k-1}^n(b)\rfloor)\Big) \in Set.$$

Furthermore

$$\frac{\begin{array}{c}(x \in T_0^n(a)) \\ \vdots \\ o \in U_k^n \qquad a \in U_0^n \qquad b \in U_{k-1}^n \qquad z \in T_0^n(u_{k-1}^n(\mathcal{P}, o, a, (x)b))\end{array}}{t_{k-1}^n(\mathcal{P}, o, a, (x)b, z) \in U_{k-1}^n,}$$

and under same assumptions we have the equality

$$T_{k-1}^n(t_{k-1}^n(\mathcal{P}, o, a, (x)b, z)) = \mathbf{q}\Big(T_k^n(o)(\lfloor T_0^n(a), (x)T_{k-1}^n(b)\rfloor)\Big)(z) \in O_{k-1}.$$

This concludes the axiomatisation of $\mathbf{ML}^{n+1}$.

Remark Note that the axiomatisation of $\mathbf{ML}^{n+1}$ is not minimal, since for $k < n$, U_k^n, T_k^n can do the job of U_k^{n-1}, T_k^{n-1} by letting $U_k^{n-1} = U_k^n(N_0, (x)B)$, where $B(x)$ $(x \in N_0)$ is an empty family, and similarly for T_k^{n-1} etc.

Example 5.2 (Universe operator.) The theory $\mathbf{ML}^1$ is simply **MLU**. Clearly $U_0^0(A, B)$ and $T_0^0(A, B, \cdot)$ is a universe above the family of sets A, B.

Example 5.3 (Super universe operator.) Define an operator $Q_1 \in O_1 = (F_0)F_0$ of order one by letting

$$Q_1(\lfloor A, B\rfloor) = \lfloor U_0^0(A,B), (x)T_0^0(A,B,x)\rfloor.$$

Then $U_0^1(N_1, (x)Q_1, A, B)$ and $T_0^1(N_1, (x)Q_1, A, B, \cdot)$, with Q_1 as in Example 5.2, defines a super universe above the family of sets A, B. Indeed, letting $\mathcal{P} = N_1, (x)Q_1, A, B$ then $u_0^1(\mathcal{P}, \ell_1^1(\mathcal{P}, 0_1), a, (x)b)$ corresponds to the canonical expression $u(a, (x)b)$ of Section 3, and $t_0^1(\mathcal{P}, \ell_1^1(\mathcal{P}, 0_1), a, (x)b, z)$ corresponds to $t(a, (x)b, z)$. The theory $\mathbf{ML}^2$ also allows the formation of universes closed under arbitrary prescribed family of operators.

Example 5.4 Here is an example of the use of $\mathbf{ML}^3$. Let $Q_2 \in (F_1)O_1$ be defined by

$$Q_2(\lfloor I, J\rfloor)(\lfloor A, B\rfloor) = \lfloor U_0^1(I,J,A,B), (x)T_0^1(I,J,A,B,x)\rfloor.$$

This gives a super universe operator from a given family I, J of (universe) operators. Define from this an operator $\hat{Q}_2 \in (F_1)F_1$ of order two, by letting $\hat{Q}_2(\lfloor I, J\rfloor) = \lfloor N_1, (x)Q_2(\lfloor I, J\rfloor)\rfloor$. Let $\mathcal{P}'$ be the sequence $N_1, (x)\hat{Q}_2, N_1, (x)Q_1$, A, B. Then $\mathbf{M} = U_0^2(\mathcal{P}')$, $\mathbf{S}(\cdot) = T_0^2(\mathcal{P}', \cdot)$ is a universe closed under this operator as well. Let $\mathbf{Q} = U_1^2(\mathcal{P}')$ and $\mathbf{F}(a)(X,Y) = \mathbf{p}(T_1^2(\mathcal{P}', a)(\lfloor X, Y\rfloor))$, $\mathbf{G}(a)(X,Y)(c) = \mathbf{q}(T_1^2(\mathcal{P}', a)(\lfloor X, Y\rfloor))(c)$. Then $(\mathbf{Q}; \mathbf{F}, \mathbf{G})$ represents the universe of universe operators (cf. Rathjen *et al.* (to appear)).

There are interesting proof-theoretic applications of this kind of theories. In the presence of W-sets the universes become type-theoretic counterparts of large cardinals. Setzer (1995) gives a type theory whose strength exceeds Kripke-Platek (KP) set theory together with a Mahlo cardinal. His universe construction is however impredicative, see Section 6. Rathjen (1997) considers a theory $\mathbf{MLF_W}$ which is essentially $\mathbf{ML}^2$ extended with W-sets and where all universes are closed under W-sets. The strength is that of a KP set theory with a Mahlo cardinal, but with restricted set induction. The corresponding theory without W-sets, **MLF**, seems considerably harder to analyse, nevertheless its strength has been conjectured (Rathjen 1997). Rathjen *et al.* (to appear) presents an extension of Aczel's constructive set theory which encompasses constructive versions of Mahlo's π-numbers. A constructive justification of this set theory is obtained by an interpretation in the type theory **MLQ**. This theory may in turn be interpreted into $\mathbf{ML}^3$, if we expand it and its universes with W-sets.

Example 5.5 (The theory $\mathbf{ML}^{m+1}$.) To give a further example of the use of the higher operators, we show how some operator

$$Q \in (F_n)(F_{n-1})\cdots(F_k)F_k,$$

$n > k$, may be internalised. Notice its mixed order. Write $\overline{Q}(G_n, \ldots, G_k)$ for $\mathbf{p}(Q(G_n, \ldots, G_k))$ — the code set — and $\tilde{Q}(G_n, \ldots, G_k)$ for $\mathbf{q}(Q(G_n, \ldots, G_k))$ — the decoding function. We need to lift Q to $\hat{Q} \in (F_n)F_n$ by putting

$$\hat{Q} = (G_n)\lfloor N_1, (x)(G_{n-1})\lfloor N_1, \ldots \lfloor N_1, (x)(G_k)Q(G_n, \ldots, G_k)\rfloor \cdots \rfloor\rfloor.$$

Let $\mathcal{P} = A_m, B_m, \ldots, A_{n+1}, B_{n+1}, \ldots, A_0, B_0$, $m > n$, be a set of parameters, where $B_{n+1}(\hat{q}) = \hat{Q}$ for some $\hat{q} \in A_{n+1}$ and $\lfloor A_i, B_i \rfloor \in F_i$. We show that Q is indeed internal to the universe given by $U_i \equiv U_i^m(\mathcal{P})$ and $T_i(\cdot) \equiv T_i^m(\mathcal{P}, \cdot)$. We write $t_i(r, a, b, c), \ell_i(d)$ for $t_i^m(\mathcal{P}, r, a, b, c), \ell_i^m(\mathcal{P}, d)$, respectively. Suppose that

$$a_n \in U_0, b_n \in (T_0(a_n))U_n, \ldots, a_k \in U_0, b_k \in (T_0(a_k))U_k.$$

Define $\overline{q}(a_n, b_n, \ldots, a_k, b_k) \in U_k$ to be

$$u_k(t_{k+1}(\cdots t_n(\ell_{n+1}(\hat{q}), a_n, b_n, 0_1), \ldots, a_{k+1}, b_{k+1}, 0_1), a_k, b_k).$$

Then a straightforward calculation shows that $T_0(\overline{q}(a_n, b_n, \ldots, a_k, b_k))$ is

$$\overline{Q}(\lfloor T_0(a_n), (x)T_n(b_n(x)) \rfloor, \ldots, \lfloor T_0(a_k), (x)T_k(b_k(x)) \rfloor).$$

Moreover define $\tilde{q}(a_n, b_n, \ldots, a_k, b_k, z) \in U_k$ to be

$$t_k(t_{k+1}(\cdots t_n(\ell_{n+1}(\hat{q}), a_n, b_n, 0_1), \ldots, a_{k+1}, b_{k+1}, 0_1), a_k, b_k, z).$$

Then $T_k(\tilde{q}(a_n, b_n, \ldots, a_k, b_k, z))$ is

$$\tilde{Q}(\lfloor T_0(a_n), (x)T_n(b_n(x)) \rfloor, \ldots, \lfloor T_0(a_k), (x)T_k(b_k(x)) \rfloor, z).$$

Remark The theories $\mathbf{ML}^n$ were suggested by the author in October 1989. A more recent development is Dybjer's general scheme for inductive-recursive definitions (Dybjer, to appear). It captures the super universe construction, and a further generalisation captures also $\mathbf{ML}^n$. The scheme was partly inspired by Mendler's (1991) categorical interpretation of universes, which in turn took as a motivation the super universe of Section 3.

6 Stepping into the Impredicative

Impredicative theories can be formulated very clearly using universes. However, such universes have no inductive structure as we shall see in two examples.

Consider type theory with one universe $(\ddot{U}, \ddot{T})$ extended by codes for second order universal quantification $\ddot{\forall}$ in the following manner. Letting $\mathcal{P}(a) := \ddot{T}(a) \to \ddot{U}$, we adopt the introduction rules

$$\frac{\begin{array}{c}(X \in \mathcal{P}(a)) \\ \vdots \\ a \in \ddot{U} \qquad b \in \ddot{U}\end{array}}{\ddot{\forall}(a, (X)b) \in \ddot{U}} \qquad \frac{\begin{array}{c}(X \in \mathcal{P}(a)) \\ \vdots \\ a \in \ddot{U} \qquad b \in \ddot{U}\end{array}}{\ddot{T}(\ddot{\forall}(a, (X)b)) = (\Pi X \in \mathcal{P}(a))\ddot{T}(b)}.$$

We thus add second order quantification over each set a in $\ddot{U}$. As is wellknown, the second order existential quantifier is definable from the universal quantifier. It is straightforward to see that the full comprehension principle is valid in this universe.

We note that the universe $(\ddot{U}, \ddot{T})$ is in some sense *non-wellfounded.* Indeed, assume one imposes the natural elimination rule for the universe by assuming U-elimination (cf. Palmgren (1992), p. 95) extended with a clause for the $\ddot{\forall}$-case

$$h(\ddot{\forall}(a,(X)b)) = d_{\ddot{\forall}}(a,(X)b,h(a),(\lambda X)h(b)). \tag{6.1}$$

Then we obtain an inconsistent theory with non normalising terms. Let n_0 and n_1 be codes for the sets N_0 and N_1, respectively. Using (6.1) we define a term $h(z) \in \ddot{U}$ $(z \in \ddot{U})$ such that $h(\ddot{\forall}(a,(x)b)) = Ap(g, Ap((\lambda x)h(b), (\lambda y)\varphi))$ where $g \in \ddot{U} \to \ddot{U}$ is an arbitrary function and $\varphi \in \ddot{U}$. Now letting $\varphi := \ddot{\forall}(n_1,(x)Ap(x,0_1))$, we have

$$h(\varphi) = Ap(g, Ap((\lambda x)h(Ap(x,0_1)),(\lambda y)\varphi)) = Ap(g,h(\varphi)).$$

Hence $h(\varphi)$ is a fixed point of g. Letting $g := (\lambda x)(x \to n_0)$, this leads to outright inconsistency, since we then obtain a set $A = \ddot{T}(h(\varphi))$, such that $A = \neg A$. If we instead take $g := (\lambda x)(x \to x)$, the equation $A = A \to A$ emerges. From this we obtain a nonterminating term, by considering it as a model of untyped λ-calculus. The problematic point with the above universe is that it occurs negatively in one of its own introduction rules.

Another, proof-theoretically more interesting example, is Setzer's Mahlo universe (M, S) (Setzer 1995). Here one crucial introduction rule is

$$\frac{f \in \mathcal{F}_M \to \mathcal{F}_M}{u_f \in M},$$

where $\mathcal{F}_M = (\Sigma x \in M)[S(x) \to M]$. Similarly to the above we can prove that it is inconsistent with the natural elimination rule. This rule is analogous to the one for $(\ddot{U}, \ddot{T})$ but we have instead

$$h(u_f) = d_u(f,(\lambda x)h(f(x))).$$

For any $g \in M \to M$ we define $g^+ \in \mathcal{F}_M \to \mathcal{F}_M$ by $g^+(w) = \langle g(p(w)),(\lambda x)n_1\rangle$, and for any $f \in \mathcal{F}_M \to \mathcal{F}_M$ we define $f^- \in M \to M$ by $f^-(a) = p(f\langle a,(\lambda x)n_1\rangle)$. (The particular choice of n_1 is not important, any other code would do.) Thus $(g^+)^-(a) = g(a)$. By the natural elimination rule, there exists $h \in M \to M$ such that

$$h(u_f) = h(f^-(u_f)) \to n_0 \qquad (f \in \mathcal{F}_M \to \mathcal{F}_M).$$

Now put $f = ((\lambda x \in M)x)^+$. Then by the above $f^-(u_f) = u_f$, so $h(u_f) = h(u_f) \to n_0$. Hence $A = \neg A$, for some A and analogously to the above $B = B \to B$ for some B. We summarise the results as a theorem.

Theorem 6.1 *Let T be a type theory with either the second order universe or with Setzer's Mahlo universe. Then T becomes inconsistent and non-normalising, when adding the natural elimination rules.*

We remark that Setzer did not himself consider an elimination rule for his universe (Setzer 1995). However, it seems reasonable from a predicative point of view to require that any set introduced in type theory should be consistent with the natural elimination rules generated by the introduction rules.

7 Classical Universes Within Type Theory

The A-translation is a combination of Gödel's negative translation with Dragalin and Friedman's wellknown syntactic translation. This translation gives an easy method for proving conservativity of Π^0_2-sentences of many classical theories over their intuitionistic counterpart. We shall here use a universe of classical propositions to obtain a semantic version of this method. The idea is to extend type theory with a universe of propositions for which classical logic holds. It is in a precise sense a (small-) complete boolean algebra with prescribed falsity.

The A-translation for the $\forall$, $\wedge$, $\rightarrow$ fragment of minimal logic (Berger and Schwichtenberg 1995) has a particularly simple form. We shall make a semantic version of this translation. Our starting point is a Martin-Löf type theory with a universe of sets (U, T). Here it will be useful to think of a code $a \in U$ as a (constructively) given infinitary formula, and the decoding $T(a)$ as its canonical Tarski semantics. We extend this type theory with a universe of propositions $(U_{\neg\neg}, T_A)$ for each set A. We define it as follows:

$$U_{\neg\neg}\ set \qquad \frac{A\ set \quad b \in U_{\neg\neg}}{T_A(b)\ set}.$$

The absurdity of this universe will be A, and for each set p of U we introduce a new proposition ${}^g p$ into the new universe.

$$\bot \in U_{\neg\neg} \qquad T_A(\bot) = A$$

$$\frac{p \in U}{{}^g p \in U_{\neg\neg}} \qquad \frac{p \in U}{T_A({}^g p) = (T(p) \rightarrow A) \rightarrow A}.$$

We assume closure only under implication, conjunction and universal quantification over small sets:

$$\frac{b \in U_{\neg\neg} \quad c \in U_{\neg\neg}}{b \supset c \in U_{\neg\neg}} \qquad \frac{b \in U_{\neg\neg} \quad c \in U_{\neg\neg}}{T_A(b \supset c) = T_A(b) \rightarrow T_A(c)}$$

$$\frac{b \in U_{\neg\neg} \quad c \in U_{\neg\neg}}{b \wedge c \in U_{\neg\neg}} \qquad \frac{b \in U_{\neg\neg} \quad c \in U_{\neg\neg}}{T_A(b \wedge c) = T_A(b) \times T_A(c)}$$

$$\frac{\begin{array}{c}(x \in T(s)) \\ \vdots \\ s \in U \qquad b \in U_{\neg\neg}\end{array}}{\forall(s, (x)b) \in U_{\neg\neg}} \qquad \frac{\begin{array}{c}(x \in T(s)) \\ \vdots \\ s \in U \qquad b \in U_{\neg\neg}\end{array}}{T_A(\forall(s, (x)b)) = (\Pi x \in T(s)) T_A(b)}.$$

We admit also proof by induction on this universe, a principle which is no stronger than recursion on an ordinary universe. This is then the semantic version of the A-translation for minimal logic. The basic results are proved similarly as in the syntactic case.

Theorem 7.1 *The universe $(U_{\neg\neg}, T_A)$ satisfies stability and ex falso quod libet, i.e. there are constructions for $T_A(\neg\neg b \supset b)$ and $T_A(\bot \supset b)$ for any $b \in U_{\neg\neg}$.*

Proof By induction on the universe. □

Theorem 7.2. (Π_2-conservation) *Let $p(x, y) \in U$ $(x \in R, y \in S)$ be a family of small sets over small sets $R = T(r)$ and $S = T(s)$. If for all small A, $T_A(\forall(r, (x)\neg\forall(s, (y)\neg^g p(x, y))))$ is true, then $(\Pi x \in R)(\Sigma y \in S)T(p(x, y))$ is true.*

Proof For any given $x \in R$, substitute for A the set $(\Sigma y \in S)T(p(x, y))$ and then proceed as in the familiar syntactic proof. □

Note that $T_A(b)$ does not in general follow from $T(b)$. It is possible to make intricate analyses for what b this in fact is the case, by generalising results from the syntactic situation. Here we only observe that $T_A({}^g i(s, x, y))$ holds whenever $I(T(s), x, y)$ holds and that the translation of Peano's fourth axiom $(n + 1 \neq 0)$ is valid. Moreover the induction schemata for natural numbers and W-sets, with branching over any small family of sets, are valid in translated form. This means that in the classical universe we may use higher type arithmetic and the mentioned induction schemes. It seems to be an interesting task to investigate what further principles are valid.

The semantic version of the A-translation was completely formalised using the proof support system ALF, and tested on a small program extraction problem. This was done in cooperation with U. Berger. The advantage of the semantic version is that it is possible to work entirely within one theory, and that classical and constructive methods may be mixed.

8 Acknowledgements

I am grateful to Per Martin-Löf for encouragement to pursue the construction of universes, and to Ed Griffor and Michael Rathjen for discussions. Section 2 and 3 formed part of my PhD Thesis. Section 7 was written (and implemented) while visiting the Ludwig-Maximilians Universität in Munich. Thanks go to Helmut Schwichtenberg, Ulrich Berger and Anton Setzer for their hospitality. I thank Peter Hancock and an anonymous referee for valuable comments on the content and presentation of this paper.

Bibliography

Aczel, P. (1977). The strength of Martin-Löf's intuitionistic type theory with one universe. In: S. Miettinen and J. Väänänen (eds.) *Proc. Symp. on Mathematical Logic (Oulo 1974),* pp. 1 – 32, Report no. 2, Dept. of Philosophy, University of Helsinki.

Aczel, P. (1986). The type theoretic interpretation of constructive set theory: inductive definitions. In: R.B. Marcus *et al.* (eds.) *Logic, Methodology and Philosophy of Science VII,* pp. 17 – 49. North-Holland, Amsterdam.

Berger, U. and Schwichtenberg, H. (1995). Program extraction from classical proofs. In: D. Leivant (ed.) *Logic and Computational Complexity, Indianapolis*

1994, Lectures Notes in Computer Science, vol. 960, pp. 77 – 97. Springer, Berlin.

Dybjer, P. (to appear). A general formulation of simultaneous inductive-recursive definitions in type theory. *J. Symbolic Logic.*

Feferman, S. (1982). Iterated inductive fixed-point theories: application to Hancock's conjecture. In: G. Metakides (ed.) *Patras Logic Symposium*, pp. 171 – 196. North-Holland, Amsterdam.

Griffor, E. and Rathjen, M. (1994). The strength of some Martin-Löf type theories. *Arch. Math. Logic* **33**, pp. 347 – 385.

Martin-Löf, P. (1975). An intuitionistic theory of types: predicative part. In: H.E. Rose and J. Shepherdson (eds.) *Logic Colloquium '73,* pp. 73 – 118. North-Holland, Amsterdam.

Martin-Löf, P. (1982). Constructive mathematics and computer programming. In: L.J. Cohen *et al.* (eds.) *Logic, Methodology and Philosophy of Science VI,* pp. 153 – 175. North-Holland, Amsterdam.

Martin-Löf, P. (1984). *Intuitionistic Type Theory.* Bibliopolis, Naples.

Mendler, N.P. (1991). Predicative type universes and primitive recursion. In: *Proceedings of the Sixth Annual Symposium on Logic in Computer Science,* pp. 173 – 184. IEEE Computer Society Press.

Nordström, B., Peterson, K. and Smith, J.M. (1990). *Programming in Martin-Löf's Type Theory.* Oxford University Press, Oxford.

Palmgren, E. (1991). *On Fixed Points, Inductive Definitions and Universes in Martin-Löf's Type Theory.* Dissertation, Uppsala.

Palmgren, E. (1992). Type-theoretic interpretation of strictly positive, iterated inductive definitions. *Arch. Math. Logic* **32**, pp. 75 – 99.

Palmgren, E. (1993). An information system interpretation of Martin-Löf's partial type theory with universes. *Inform. and Comput.* **106**, pp. 26 – 60.

Rathjen, M. (1997). *The strength of some universe constructions in Martin-Löf type theory.* Abstract, March 25 1997.

Rathjen, M., Griffor, E. and Palmgren, E. (to appear). Inaccessibility in constructive set theory and type theory. *Ann. Pure Appl. Logic.*

Setzer, A. (1993). *Proof Theoretical Strength of Martin-Löf Type Theory with W-type and one Universe.* Dissertation, Munich. (revised version to appear in *Ann. Pure Appl. Logic*).

Setzer, A. (1995). *A type theory for one Mahlo universe.* Manuscript, September 1995. Abstract in *Bull. Symbolic Logic* 3(1997), pp. 128 – 129.

Simpson, S.G. (1982). Σ^1_1 and Π^1_1 transfinite induction. In: D. van Dalen *et al.* (eds.) *Logic Colloquium '80,* pp. 239 – 253. North-Holland, Amsterdam.