Proof Systems for Retracts in Simply Typed Lambda Calculus

Colin Stirling
School of Informatics
University of Edinburgh
cps@inf.ed.ac.uk

Abstract. This paper concerns retracts in simply typed lambda calculus assuming $\beta\eta$-equality. We provide a simple tableau proof system which characterises when a type is a retract of another type and which leads to an exponential decision procedure.

1 Introduction

Type $\rho$ is a retract of type $\tau$ if there are functions $C : \rho \rightarrow \tau$ and $D : \tau \rightarrow \rho$ with $D \circ C = \lambda x. x$. This paper concerns retracts in the case of simply typed lambda calculus [1]. Various questions can be asked. The decision problem is: given $\rho$ and $\tau$, is $\rho$ a retract of $\tau$? Is there an independent characterisation of when $\rho$ is a retract of $\tau$? Is there an inductive method, such as a proof system, for deriving assertions of the form “$\rho$ is a retract of $\tau$”? If so, can one also construct (inductively) the witness functions $C$ and $D$?

Bruce and Longo [2] provide a simple proof system that solves when there are retracts in the case that $D \circ C = \beta \lambda x. x$. The problem is considerably more difficult if $\beta$-equality is replaced with $\beta\eta$-equality. De Liguoro, Piperno and Statman [3] show that the retract relation with respect to $\beta\eta$-equality coincides with the surjection relation: $\rho$ is a retract of $\tau$ iff for any model there is a surjection from $\tau$ to $\rho$. They also provide a proof system for the affine case (when each variable in $C$ and $D$ occurs at most once) assuming a single ground type. Regnier and Urzyczyn [9] extend this proof system to cover multiple ground types. The proof systems yield simple inductive nondeterministic algorithms belonging to NP for deciding whether $\rho$ is an affine retract of $\tau$. Schubert [10] shows that the problem of affine retraction is NP-complete and how to derive witnesses $C$ and $D$ from the proof system in [9]. Under the assumption of a single ground type, decidability of when $\rho$ is a retract of $\tau$ is shown by Padovani [8] by explicit witness construction (rather than by a proof system) of a special form.

More generally, decidability of the retract problem follows from decidability of higher-order matching in simply typed lambda calculus [13]: $\rho$ is a retract of $\tau$ iff the equation $\lambda z^\rho. x_1^{\rho \rightarrow \tau} (x_2^{\rho \rightarrow \tau} z) =_{\beta\eta} \lambda z^\rho. z$ has a solution (the witnesses $D$ and $C$ for $x_1$, $x_2$). Since the complexity of matching is non-elementary [15] this decidability result leaves open whether there is a better algorithm, or even a proof.

1 For a full version see http://www.homepages.inf.ed.ac.uk/cps/ret.pdf
system, for the problem. In the case of $\beta$-equality matching is no guide to solvability: the retract problem is simply solvable whereas $\beta$-matching is undecidable [4].

In this paper we provide an independent solution to the retract problem. We show it is decidable by exhibiting sound and complete tableau proof systems. We develop two proof systems for retracts, one for the (slightly easier) case when there is a single ground type and the other for when there are multiple ground types. Both proof systems appeal to paths in terms. Their correctness depend on properties of such paths. We appeal to a dialogue game between witnesses of a retract to prove such properties: a similar game-theoretic characterisation of $\beta$-reduction underlies decidability of matching.

In Section 2 we introduce retracts in simply typed lambda calculus and fix some notation for terms as trees and for their paths. The two tableau proof systems for retracts are presented in Section 3 where we also briefly examine how they generate a decision procedure for the retract problem. In Section 4 we sketch the proof of soundness of the tableau proof systems (and completeness and further details are provided in the full version).

2 Preliminaries

Simple types are generated from ground types using the binary function operator $\rightarrow$. We let $a, b, o, \ldots$ range over ground types and $\rho, \sigma, \tau, \ldots$ range over simple types. Assuming $\rightarrow$ associates to the right, so $\rho \rightarrow \sigma \rightarrow \tau$ is $\rho \rightarrow (\sigma \rightarrow \tau)$, if a type $\rho$ is not a ground type then it has the form $\rho_1 \rightarrow \ldots \rightarrow \rho_n \rightarrow a$. We say that $a$ is the target type of $a$ and of any type $\rho_1 \rightarrow \ldots \rightarrow \rho_n \rightarrow a$.

Simply typed terms in Church style are generated from a countable set of typed variables $x^\sigma$ using lambda abstraction and function application [1]. We write $S^\tau$, or sometimes $S : \sigma$, to mean term $S$ has type $\sigma$. The usual typing rules hold: if $S^\tau$ then $\lambda x^\sigma.S^\tau : \sigma \rightarrow \tau$; if $S^\sigma \rightarrow \tau$ and $U^\sigma$ then $(S^\sigma \rightarrow U^\tau) : \tau$. In a sequence of unparenthesised applications we assume that application associates to the left, so $SU_1 \ldots U_k$ is $((\ldots (SU_1) \ldots )U_k)$. Another abbreviation is $\lambda z_1 \ldots z_m$ for $\lambda z_1 \ldots \lambda z_m$.

Usual definitions of when a variable occurrence is free or bound and when a term is closed are assumed.

We also assume the usual dynamics of $\beta$ and $\eta$-reductions and the consequent $\beta\eta$-equivalence between terms (as well as $\alpha$-equivalence). Confluence and strong normalisation ensure that terms reduce to (unique) normal forms. Moreover, we assume the standard notion of $\eta$-long $\beta$-normal form (a term in normal form which is not an $\eta$-reduct of some other term) which we abbreviate to lnf. The syntax of such terms reflects their type: a lnf of type $\alpha$ is a variable $x^\alpha$, or $x U_1 \ldots U_k$ where $x^{\rho_1 \ldots \rho_n - \alpha}$ and each $U_i^{\rho_i}$ is a lnf; a lnf of type $\rho_1 \rightarrow \ldots \rightarrow \rho_n \rightarrow a$ has the form $\lambda x_1^{\rho_1} \ldots x_n^{\rho_n} S$, where $S^\alpha$ is a lnf.

The following definition introduces retracts between types [2, 3].

**Definition 1.** Type $\rho$ is a retract of type $\tau$, written $\models \rho \preceq \tau$, if there are terms $C : \rho \rightarrow \tau$ and $D : \tau \rightarrow \rho$ such that $D \circ C =_{\beta\eta} \lambda x^\alpha.x$. 
The witnesses $C$ and $D$ to a retract can always be presented as lns. We can think of $C$ as a “coder” and $D$ as a “decoder” [9]. Assume $\rho = \rho_1 \to \ldots \to \rho_1 \to a$ and $\tau = \tau_1 \to \ldots \to \tau_n \to a$: in a retract the types must share target type [9]. We instantiate the bound $\rho_i$ variables in a decoder $D$ to $D(z_1^{\rho_1}, \ldots, z_n^{\rho_n})$, often abbreviated to $D(\overline{z})$, and the bound variable of type $\rho$ in $C$ to $C(x^\tau)$; so, $\models \rho \leq \tau$ if $D(z_1^{\rho_1}, \ldots, z_n^{\rho_n})(C(x^\tau)) = \beta_0 x z_1 \ldots z_i$. From [9], we can restrict a decoder to be of the form $\lambda f^\tau. f S_1^{\tau_1} \ldots S_n^{\tau_n}$ with $f$ as head variable and a coder $C(x)$ has the form $\lambda y_1^{\tau_1} \ldots y_n^{\tau_n}. H(x T_1^{\rho_1} \ldots T_n^{\rho_n})$.

**Definition 2.** We say that the decoder $D(z_1, \ldots, z_i) = \lambda f^\tau. f S_1^{\tau_1} \ldots S_n^{\tau_n}$ and the coder $C(x) = \lambda y_1^{\tau_1} \ldots y_n^{\tau_n}. H(x T_1^{\rho_1} \ldots T_n^{\rho_n})$ are canonical witnesses for $\rho \leq \tau$ if $D(\overline{z})(C(x)) = \beta_0 x z_1 \ldots z_i$ and they obey the following properties:

1. variables $f, z_1, \ldots, z_i$ occur only once in $D(\overline{z})$,
2. $x$ occurs only once in $C(x)$,
3. $H$ is $\varepsilon$ if $\rho$ and $\tau$ are constructed from a single ground type,
4. if $T_1^{\rho_1}$ contains an occurrence of $y_1$ then it is the head variable of $T_1^{\rho_1}$, $z_i$ occurs in $S_i^{\tau_1}$ and $T_i^{\rho_i}$ contains no other occurrences of any $y_k$, $1 \leq k \leq n$.

The next result follows from observations in [3, 9].

**Proposition 1.** $\models \rho \leq \tau$ iff there exist canonical witnesses for $\rho \leq \tau$.

So, if there is only a single ground type then $C(x)$ can be restricted to have the form $\lambda y_1^{\tau_1} \ldots y_n^{\tau_n}. x T_1^{\rho_1} \ldots T_i^{\rho_i}$ with $x$ as a head variable [3].

**Example 1.** From [3]. Let $\rho = \rho_1 \to \rho_2 \to o$ where $\rho_1 = \rho_2 = \sigma \to o$ and let $\tau = \tau_1 \to o$ where $\tau_1 = \sigma \to (o \to o \to o) \to o$ and $\sigma$ is arbitrary. It follows that $\models \rho \leq \tau$. A decoder $D(z_1^{\rho_1}, z_2^{\rho_2})$ is $\lambda f^\tau. f (\lambda u^\sigma. v^o. o. u(z_1 u)(z_2 u))$ and a coder $C(x^\tau)$ is $\lambda y_1^{\tau_1}. x (\lambda w^\sigma. y^o. w^o. u^o. w(z_1 y)(z_2 y))$; so, $(D(z_1, z_2))(C(x)) \to_\beta x (\lambda w^\sigma. z_1 y)(\lambda w^\sigma. z_2 y) = \beta_0 x z_1 z_2$. $\square$

**Example 2.** From [9] with multiple ground types. Let $\rho = \rho_1 \to \rho_2 \to a$ where $\rho_1 = b \to a$, $\rho_2 = a$ and let $\tau = \tau_1 \to a$ where $\tau_1 = b \to (a \to o \to a) \to a$. A decoder is $D(z_1^{\rho_1}, z_2^{\rho_2}) = \lambda f^\tau. f (\lambda u^b. a^o. u(z_1 u)(z_2 u))$ and a coder $C(x^\tau)$ is $\lambda y_1^{\tau_1}. y^b. \gamma^o. w^a. x (\lambda v^\sigma. y^o. (\lambda w^\sigma. w(z_1 y)(z_2 y))(v z))$; so, $(D(z_1, z_2))(C(x)) \to_\beta x (\lambda v^\sigma. z_1 y)(\lambda v^\sigma. z_2 y) = \beta_0 x z_1 z_2$. $\square$

Terms are represented as special kinds of tree (that we call binding trees in [12, 14]) with dummy lambdas and an explicit binding relation. A term of the form $y^a$ is represented as a tree with a single node labelled $y^a$. In the case of $y U_1 \ldots U_k$, when $y^{\rho_1 \ldots \rho_k} \to a$, we assume that a dummy lambda with the empty sequence of variables is placed directly above any subterm $U_i$ in its tree representation if $\rho_i$ is a ground type. With this understanding, the tree for $y U_1 \ldots U_k$ consists of a root node labelled $y^{\rho_1 \ldots \rho_k} \cdot u$ and $k$-successor trees representing $U_1, \ldots, U_k$. We also use the abbreviation $\lambda \overline{y}$ for $\lambda y_1 \ldots y_m$ for $m \geq 0$, so $\overline{y}$ is possibly the empty sequence of variables in the case of a dummy lambda. The
Fig. 1. $D(z_1, z_2)$ and $C(x)$ of Example 1

tree representation of $\lambda \overline{y}.S : \rho_1 \rightarrow \ldots \rightarrow \rho_k \rightarrow a$ consists of a root node labelled $\lambda \overline{y}$ and a single successor tree for $S^a$. The trees for $C(x)$ and $D(z_1, z_2)$ of Example 1, where we have omitted the types, are in Figure 1.

We say that a node is a lambda (variable) node if it is labelled with a lambda abstraction (variable). The type (target type) of a variable node is the type (target type) of the variable at that node and the type (target type) of a lambda node is the type (target type) of the subterm rooted at that node.

The other elaboration is that we assume an extra binary relation $\downarrow$ between nodes in a tree that represents binding; that is, between a node labelled $\lambda y_1 \ldots y_n$ and a node below it labelled $y_j$ (that it binds). A binder $\lambda y$ is such that either $y$ is empty and therefore is a dummy lambda and cannot bind a variable occurrence or $\overline{y} = y_1 \ldots y_k$ and $\lambda \overline{y}$ can only then bind variable occurrences of the form $y_i$, $1 \leq i \leq k$. Consequently, we also employ the following abbreviation $n \downarrow m$ if $n \downarrow m$ and $n$ is labelled $\lambda y_1 \ldots y_k$ and $m$ is labelled $y_i$. In Figure 1 we have not included the binding relation; however, for instance, (2) $\downarrow$ (7).

**Definition 3.** Lambda node $n$ is a descendant ($k$-descendant) of $m$ if either $m \downarrow m'$ ($m \downarrow_k m'$), $n$ is a successor of $m'$ for some $m'$ and the target types of $m$, $m'$ and $n$ are the same, or $n'$ is a descendant ($k$-descendant) of $m$ and $n$ is a descendant of $n'$ for some $n'$. 
We assume a standard presentation of nodes of a tree as sequences of integers: an initial sequence, typically ε, is the root node; if n is a node and m is the ith successor of n then m = ni. For the sake of brevity we have not followed this approach in Figure 1 where we have presented each node as a unique integer (i).

**Definition 4.** A path of the tree of a term of type σ is a sequence of nodes \( \pi = n_1, \ldots, n_k \) where \( n_1 \) is the root of the tree, each \( n_{i+1} \) is a successor of \( n_i \) and if \( n_j \) is a variable node then for some \( i < j, n_i \downarrow n_j \) (hence is a closed path).

For paths \( m = m_1, \ldots, m_i \) and \( \pi = n_1, \ldots, n_k \) of type σ we write \( m \sqsubseteq \pi \) if for some \( i > 0 \), for all \( h \leq 2i, m_h = n_h \), \( m_{2i+1} = n_{2i}p, n_{2i+1} = n_{2i}q \) and \( p < q \).

A (closed) subtree of a tree of a term of type σ is a set of paths \( P \) of type σ such that if \( m, \pi \) are distinct paths in \( P \) then \( m \sqsubseteq \pi \) or \( \pi \sqsubseteq m \).

A path \( \pi = n_1, \ldots, n_k \) is a contiguous sequence of nodes in a tree of a term starting at the root; for \( i \geq 1 \), each \( n_{2i-1} \) is a lambda node and each \( n_{2i} \) is a variable node (whose binder occurs earlier in the path). Path \( \overline{\pi} \) is before \( \pi \), \( \overline{\pi} \sqsubseteq \pi \), if they have a common even length prefix and then differ as to their successors (the one in \( \overline{\pi} \) before that in \( \pi \)). These paths could, therefore, be in the same term: therefore, a closed subtree is a set of such paths.

**Definition 5.** A path \( \pi = n_1, \ldots, n_i \) is k-minimal provided that for each binding node \( n_i \) there are at most \( k \) distinct nodes \( n_j, i < j \leq l \), such that \( n_i \downarrow n_j \). A subtree \( P \) is k-minimal if each path in \( P \) is k-minimal.

Not every path or subtree is useful in a term. So, we define when a path or subtree is realisable meaning that their nodes are “accessible” [7] or “reachable” [6] in an applicative context.

**Definition 6.** Assume \( \pi = n_1, \ldots, n_i \) is a path of odd length of a closed term \( T \) of type \( \sigma \), \( m \) is the node below \( n_i \) in \( T \) and \( T' \) is the term \( T \) when the variable \( u^\tau \) at node \( m \) is replaced with a fresh free variable \( z^\gamma \). We say that \( \pi \) is realisable if there is a closed term \( U = \lambda y^\sigma.yS_1 \ldots S_k \) such that \( UT' = \beta_\gamma \lambda x.zW_1 \ldots W_q \) for some \( q \geq 0 \).

**Definition 7.** Assume \( P \) is a subtree of closed term \( T \) of type \( \sigma \) where each path has even length, \( n_1, \ldots, n_q \) are the leaves of \( P \) and \( T_i, 1 \leq i \leq q \), is the term \( T \) when the variable \( u^\tau_i \) at \( n_i \) is replaced with a fresh free variable \( z^\gamma_i \). We say that \( P \) is realisable if there is a closed term \( U = \lambda y^\sigma.yS_1 \ldots S_k \) such that for each \( i, UT_i = \beta_\gamma \lambda x.z_i W_1 \ldots W_{q_i} \) for \( q_i \geq 0 \).

Next we define two useful operations on paths, restriction relative to a suffix and the subtype after a prefix.

**Definition 8.** Assume that \( \pi = n_1, \ldots, n_p \) is a path, \( \sigma = \sigma_1 \rightarrow \ldots \rightarrow \sigma_k \rightarrow a \), \( n_i \) is a lambda node of type \( \sigma \) and \( w = n_1, \ldots, n_p \) is a suffix of \( \pi \).

1. The suffix \( w \) admits \( \sigma_j, 1 \leq j \leq k \), if either there is no \( n_q, i \leq q \leq p \), such that \( n_i \downarrow_j n_q \) or there is a \( j \)-descendant \( n_q \) of \( n_i \) whose type is \( \tau_1 \rightarrow \ldots \rightarrow \tau_i \rightarrow a \) and for some \( r \) there is not a \( t : q < t \leq p \) such that \( n_q \downarrow_r n_t \) and \( a \) is the target type of \( \tau_r \).
2. The restriction of \( \sigma \) to \( w, \sigma \mid w \), is defined as \( \sigma_w \) where
   
   \[ a_w = a, \]
   
   \[ (\sigma_j \rightarrow \ldots \rightarrow \sigma_k \rightarrow a)_w = \text{if } w \text{ admits } \sigma_j \rightarrow \sigma_j \rightarrow \ldots \rightarrow \sigma_k \rightarrow a \text{ then } \sigma_j \rightarrow \ldots \rightarrow \sigma_k \rightarrow a, \]
   
   otherwise, \( \text{else } \sigma_j \rightarrow \ldots \rightarrow \sigma_k \rightarrow a \)_w.

**Definition 9.** Assume that \( \pi = n_1, \ldots, n_p \) is a path of type \( \sigma \). For a prefix \( w \) of \( \pi \) we define the subtype of \( \sigma \) after \( w, w(\sigma) \):

- if \( w = \varepsilon \) (the empty prefix) then \( \sigma \),
- if \( w = n_1, \ldots, n_q, q \leq p \), then the type of node \( n_q \).

We also define a canonical presentation of a (prefix or suffix of a) path \( \pi = n_1, \ldots, n_k \) of type \( \sigma \) as a word \( w \). If \( w \) is the empty prefix we write \( w = \varepsilon \). Otherwise, \( w = (w_1, \ldots, w_i) \), \( j \leq k \), where for each \( i \geq 0 \), \( w_{k+1} = n_{2i+1} \) and if \( n_{k} \downarrow_{m} n_{2i+1} \), then \( w_{k+1} = n_{k+1} \). Thus, we distinguish between \( w = \varepsilon \) (the empty word) and \( w = (\varepsilon) \) the prefix of length 1 consisting of the root node. Also, we can present a subtree as a set of words. Words will occur in our proof systems as presentations of paths. For example, \( w = (\varepsilon, 1, 11, 112, 1112) \) of type \( \tau \) as in Example 1 represents the path labelled \( \lambda f, f, \lambda uv, v, \lambda \) of \( D(z_1, z_2) \) in Figure 1 when its root is \( \varepsilon \). To illustrate Definitions 8 and 9, for the prefix \( w = (\varepsilon, 11, 112) \) of type \( \tau \) as in Example 1 and \( \pi = (\varepsilon, 11, 112) \) of type \( \tau \) we have \( w' = (\varepsilon, 1, 11) \) and the suffix \( w'' = (11, 112, 1112) \) of \( w \) we have \( w''(\sigma) = \tau_1 \) where \( \tau_1 = \sigma \rightarrow (o \rightarrow o \rightarrow o) \rightarrow o \) as in Example 1 and \( \tau_1 \mid w'' = \sigma \rightarrow o \); word \( w'' \) of type \( \tau_1 \) has labelling \( \lambda u \rightarrow o \rightarrow o \rightarrow o, v, \lambda \); so, \( w'' \) admits the first component \( \sigma \) of \( \tau_1 \) but not the second \( (o \rightarrow o \rightarrow o) \). The final element of \( w' \) is the same as the first element of \( w'' \); in such a case we define their concatenation to be \( w \).

**Definition 10.** The concatenation of (a prefix) \( v \) and (a suffix) \( w, v \cdot w \), is:

\[ \varepsilon \cdot w = w; \text{ if } v_k = w_1 \text{ then } v_1, \ldots, v_k \cdot w_1, \ldots, w_n = v_1, \ldots, v_k, w_2, \ldots, w_n. \]

## 3 Proof Systems for Retracts

We now develop goal directed tableau proof systems for showing retracts. By inverting the rules one has more classical axiomatic systems: we do it this way because it thereby provides an immediate nondeterministic decision procedure for deciding retracts. We present two such proof systems: a slightly simpler system for the restricted case when there is a single ground type and one for the general case.

### 3.1 Single Ground Type

Assertions in our proof system are of two kinds. First is \( \rho \leq \tau \) with meaning \( \rho \) is a retract of \( \tau \). The second has the form \([\rho_1, \ldots, \rho_k] \leq \tau \) which is based on the “product” as defined in [3]. We follow [9] in allowing reordering of components of types since \( \rho \rightarrow \sigma \rightarrow \tau \) is isomorphic to \( \sigma \rightarrow \rho \rightarrow \tau \). Instead we could include explicit rules for reordering (as with the axiom in [3]). Moreover, we assume that \([\rho_1, \ldots, \rho_k] \) is a multi-set and so elements can be in any order.
The proof rules are given in Figure 2. There is a single axiom $I$, identity, a weakening rule $W$, a covariance rule $C$, and two product rules $P_1$ and $P_2$. The rules are goal directed: for instance, $C$ allows one to decompose the goal $\delta \rightarrow \rho \leq \sigma \rightarrow \tau$ into the two subgoals $\delta \leq \sigma$ and $\rho \leq \tau$. $I$, $W$ and $C$ (or their variants) occur in the proof systems for affine retracts (when variables in witnesses can only occur at most once) [3, 9]. The new rules are the product rules: $P_2$ appeals to $k$-minimal realisable paths (presented as words), and the restriction operator of Definition 8. The proof system does not require the axiom A4 of [3], $\sigma (\sigma \rightarrow a) \rightarrow a$; all instances are provable using $W$ and $C$.

**Definition 11.** A successful proof tree for $\rho \leq \tau$ is a finite tree whose root is labelled with the goal $\rho \leq \tau$, the successor nodes of a node are the result of an application of one of the rules to it, and each leaf is labelled with an axiom. We write $\vdash \rho \leq \tau$ if there is a successful proof tree for $\rho \leq \tau$.

For some intuition about the product rules assume $\rho = \rho_1 \rightarrow \ldots \rightarrow \rho_l \rightarrow a$ and $\tau = \tau_1 \rightarrow \ldots \rightarrow \tau_n \rightarrow a$. Now, $\vdash \rho \leq \tau$ if there are canonical, Definition 2, witnesses $D(z_1^{\rho_1}, \ldots, z_l^{\rho_l}) = \lambda f^\tau . f S_{\tau_1}^{\rho_1} \ldots S_{\tau_n}^{\rho_l}$. Since we can reorder components of $\rho$ and $\tau$ we can assume that $z_1$ is in $S_{\tau_1}^{\rho_1}$. Suppose $z_1, \ldots, z_k$, where $k > 1$, are in $S_{\tau_1}^\rho$ and so $y_1$ must occur in $T_{\tau_1}^{\rho_1}, \ldots, T_{\tau_1}^{\rho_l}$. Therefore, there is a common coder $S_{\tau_1}^{\rho_1}(x_1/z_1, \ldots, x_k/z_k)$ and $k$ decoders $T_i(z_i)$ where $z_i = z_i^{\rho_1} \ldots z_i^{\rho_l}$ and $y_1, \ldots, y_k$ are the components of $\rho_i$ such that $T_i(z_i)(S_{\tau_1}^{\rho_1}(x_1, \ldots, x_k)) = y_i x_i z_i$ (which is similar to the product in [3]). In $S_{\tau_1}^{\rho_1}(x_1/z_1, \ldots, x_k/z_k)$ there are distinct odd length paths $w_1, \ldots, w_k$ of type $\tau_1$ to the lambda nodes above $x_1, \ldots, x_k$. These paths are realisable, Definition 6, because each $x_i$ belongs to the normal form of $T_i(z_i)(S_{\tau_1}^{\rho_1}(x_1, \ldots, x_k))$. Using a combinatorial argument, see the full version, $S_{\tau_1}^{\rho_1}$ can be chosen so that these words are $k$-minimal and
by reordering $\rho$’s components $w_1 \sqsubset \ldots \sqsubset w_k$. We may not be able to reduce to the subgoals $\rho_1 \sqsubseteq \tau_1, \ldots, \rho_k \sqsubseteq \tau_1$ as $w_i$ may prescribe the form of $T_i(\tau_i)$: if $T_i(\tau_i) = \lambda f^\tau_1 . f S_1^i \ldots S_m^i$ then path $w_i$ may prevent $S_i^j$ containing elements of $\tau_i$; so, this may restrict the possible distribution of $\tau_i$ within the subterms $S_1^i, \ldots, S_m^i$ which is captured using $\tau_i \downharpoonright w_i$.

An example proof tree is in Figure 3 for the retract of Example 1 (which is not affine). Rule $P_1$ is applied to the root and then $P_2$ to the first subgoal where $w_1 = (\varepsilon, 2, 21)$ and $w_2 = (\varepsilon, 2, 22)$. Let $\sigma’ = \sigma \rightarrow (o \rightarrow o \rightarrow o) \rightarrow o$. Now, $\sigma’ \downharpoonright w_1 = \sigma \rightarrow a = \sigma’ \downharpoonright w_2$; in both cases only the first component of $\sigma’$ is admitted.

3.2 Multiple Ground Types

We extend the proof system to include multiple ground types. Again, assertions are of the two kinds $\rho \sqsubseteq \tau$ and $[\rho_1, \ldots, \rho_k] \sqsubseteq \tau$. However, we now assume that to be a well-formed assertion $\rho \sqsubseteq \tau$ both $\rho$ and $\tau$ must share the same target type (which is guaranteed when there is a single ground type). The rules for this assertion are as before the axiom $I$, weakening $W$, covariance $C$ and the product rule $P_1$ in Figure 2: however, $C$ carries the requirement that the target types of $\delta$ and $\sigma$ coincide. The other product rule $P_2’$, presented in Figure 4, is different: the arity of $\rho_1 \rightarrow \ldots \rightarrow \rho_n \rightarrow a$ is the maximum of $n$ and the arities of each $\rho_i$ where a ground type $a$ has arity 0.

$$P_2’ \quad \frac{[\rho_1, \ldots, \rho_k] \sqsubseteq \sigma}{\rho_1 \sqsubseteq v_1(\sigma) \downharpoonright w_1 \ldots \rho_k \sqsubseteq v_k(\sigma) \downharpoonright w_k} \quad \text{where}$$

- $k’$ is the maximum of $k$ and $h^2$ where $h$ is the arity of $\sigma$
- there is a $k’$-minimal realisable subtree $U$ of type $\sigma$ where each path has even length (which can be $\emptyset$),
- each $v_i$ is $\varepsilon$, a prefix of a path in $U$ of odd length or the extension of a path in $U$ with a single node,
- $v_i^\uparrow w_1 \sqsubseteq \ldots \sqsubseteq v_i^\uparrow w_k$ and each $v_i^\uparrow w_i$ is a $k’$-minimal realisable path of type $\sigma$ of odd length and if $U \neq \emptyset$, $v_i^\uparrow w_i$ extends some path in $U$.

**Fig. 4.** Product proof rule for multiple ground types
(b → a) → a → a ≤ (b → (a → a → a) → a) → a

\[ [b → a, o] ≤ b → (a → o → a) → a \quad a ≤ a \]

\[ b → a ≤ b → a \quad o ≤ o \]

Fig. 5. A proof tree for Example 2

In \([n_1, \ldots, n_k] \leq \sigma\) it is not required that \(n_j\) and \(\sigma\) share the same target type. Instead rule \(P_2^n\) requires that \(n_i\) and \(\tau_i\), see Definition 9, do share target types: for the concatenation \(v_i^n w_i\), see Definition 10. The specialisation to the case of the single ground type is when \(U = \emptyset\) and \(v = \epsilon\).

Let \(\rho = n_1 → \ldots → n_i → a\) and \(\tau = \tau_1 → \ldots → \tau_i → a\). So, \(\models \rho ≤ \tau\) if and only if there are canonical witnesses \(D(z_1^{n_1}, \ldots, z_k^{n_k}) = \lambda y^{n_1} f^T z_1^{n_1} \ldots S_{i-n}^{n_i}\) and \(C(x) = \lambda y^{n_1} \ldots y^{n_k} T(x T_1^{n_1} \ldots T_2^n)\). Assume \(z_1, \ldots, z_k\), where \(k ≥ 1\), occur in \(S_{i-n}^{n_i}\). There is a path \(v\) in \(C(x)\) to the node above \(x\) which determines a subtree \(U\) of \(S_{i-n}^{n_i}\). The head variable in \(T_2^n\) bound in \(v\) has the same target type as \(n_i\). There are distinct paths \(v_1^n w_1, \ldots, v_k^n w_k\) of odd length to the lambda nodes above \(z_1, \ldots, z_k\) in \(S_{i-n}^{n_i}\); \(v_i\) is decided by the meaning of the head variable in \(T_2^n\); so, \(v_i(\tau_i)\) has the same target type as \(n_i\). The rest of the path is the tail of \(w_i\); so we need to consider whether \(\models n_i ≤ v_i(\tau_i) \downarrow w_i\).

Figure 5 is the proof tree for the retract in Example 2. There is an application of \(P_1^n\) followed by \(P_2^n\). In the application of \(P_2^n\) the subtree \(U = \{\epsilon, 2\}\), \(v_1 = \epsilon\), \(w_1 = (\epsilon, 2, 21) = v_1^n w_1\), \(v_2 = (\epsilon, 2, 22) = v_2^n w_2\) when \(w_2 = (22)\). So, \(v_1(b → (a → o → a) → a) \downarrow w_1 = b → a\) as the first component is admitted (unlike the second); and \(v_2(b → (a → o → a) → a) = o = o \downarrow w_2\).

### 3.3 Complexity

The proof systems provide nondeterministic decision procedures for checking retracts. Each subgoal of a proof rule has smaller size than the goal. Hence, by focussing on one subgoal at a time a proof witness can be presented in PSPACE. However, this does not take into account checking that a subgoal obeys the side conditions in the case of the product rules. Given any type \(\sigma\), there are boundedly many realisable \(k\)-minimal paths (with an upper bound of \(k^n\) where \(n\) is size of \(\sigma\)). So, this means that overall the decision procedure requires at most EXPSPACE.

### 4 Soundness and Completeness

To show soundness and completeness of our proof systems, we define a dialogue game \(G(D(\overline{\pi}), C(x))\) played by a single player \(\forall\) on the trees of potential witnesses for a retract that characterises when \((D(\overline{\pi}))C(x) = \beta_{\gamma} x\overline{\pi}\), similar to game semantics [5]. The game is defined in the full version of the paper.
To provide intuition for the reader we briefly describe $G(D(z_1, z_2), C(x))$ where these terms are from Figure 1. Play starts at node (0), the binder $\lambda f$ at that node is associated with $C(x)$ rooted at (12); so, the next position is at node (1) and therefore jumps to (12); the binder at (12) $\lambda y$ is associated with node (2) (the successor of (1)). Play proceeds to (13) and $v$ chooses to go left or right; suppose it is left, so play is then at (14); nodes (13) and (14) are part of the normal form, see Definition ???. Play descends to (15) and, therefore, jumps to (2); so, with the binder at (2), $u$ is associated with the the subtree at (16) and $v$ with the subtree at (18). Play proceeds to (3) and so jumps to (18); now, $s$ is associated with (4) and $t$ with (8). Play proceeds to (19) and so jumps to (4), descends to (5) and then to (6) and then to (7) and jumps to (16) before finishing at (17). This play captures the path $x\lambda w.z_1 w$ of the normal form.

Some of the key properties, defined in the full version, we appeal to in the correctness proofs below associate subtrees with realisable paths and vice versa. For instance, as illustrated in the play above the path rooted at (0) downto (7) is associated with the subtree rooted at (12) and with leaves (17) and (19). Let $\rho = \rho_1 \to \ldots \to \rho_1 \to a$ and $\tau = \tau_1 \to \ldots \to \tau_n \to a$ and let $\tau_1 = \sigma = \sigma_1 \to \ldots \to \sigma_m \to b$.

**Theorem 1.** (Soundness) If $\vdash \rho \leq \tau$ then $\models \rho \leq \tau$.

**Proof.** By induction on the depth of a proof. For the base case, the result is clear for a proof that uses the axiom $I$. So, assume the result for all proofs of depth $< d$. Consider now a proof of depth $d$. We proceed by examining the first rule that is applied to show $\vdash \rho \leq \tau$. If it is $W$ or $C$ the result follows using the same arguments as in [3]. Assume the rule is $W$ and suppose $\models \rho \leq \tau$. Therefore there are terms $D_1$ and $C_1$ such that $D_1^{-\rho}(C_1^{-\rho}) = \rho_1 \ x$. Now $D_1^{(\delta^{-\rho})-\rho} = \lambda f^{\sigma^{-\rho}}g^{\sigma^{-\rho}}a^{\sigma^{-\rho}}D_1(fy)$ and $C_1^{\rho^{-\rho}}x = \lambda s^{\sigma^{-\rho}}C_1(x)$ are witnesses for $\models \rho \leq \sigma \to \tau$. Assume that the rule is $C$, so $\models \delta \leq \sigma$ and $\models \rho \leq \sigma$. So there are terms $D_1, C_1, D_2, C_2$ such that $D_1^{\rho^{-\rho}}(C_1^{-\rho}) = \rho_1 x$ and $D_2^{\rho^{-\rho}}(C_2^{-\rho}) = \rho_1 x$. Now $D_1^{\rho^{-\rho}}(\delta^{-\rho}) = \lambda x^{\rho_1}D_2(x(D_1y))$ and $C_1^{\delta^{-\rho}}\delta^{-\rho} = \lambda u^{\rho_1}D_2(u(C_1z))$ are witnesses for $\models \delta \to \rho \leq \sigma \to \tau$.

Consider next that the first rule is $P_1$. So after $P_1$ there is either an application of $P_2$ or $P_3$; in the former case, there are $k$-minimal realisable paths $w_1 \subseteq \ldots \subseteq w_k$ of odd length of type $\sigma$ such that $\vdash \rho \leq \sigma \mid w_i$; in the latter case, there is a $k'$-minimal realisable subtree $U$ of type $\sigma$ where each path has even length; and there are paths $v_1^i w_1 \subseteq \ldots \subseteq v_{k'}^i w_k$ where each element is a $k'$-minimal realisable path of type $\sigma$ of odd length and if $U \neq \emptyset$, it extends some path in $U$ and where each $v_i$ is $e$, a prefix of a path in $U$ of odd length path or an extension of a path in $U$ with a single node and $\vdash \rho \leq \rho_i \mid w_i$; where $k'$ is the maximum of $k$ and the square of the arity of $\sigma$. So, by the induction hypothesis there are terms $D_i(z_i)$ and $C_i(x_i)$ such that $D_i(z_i)(C_i(x_i)) = \beta_0 x_i z_i$, witnesses for $\rho_1 \leq \sigma \mid w_i$ or $\rho_1 \leq \rho_i \sigma \mid w_i$, and terms $D'(z_{k+1} \ldots, z_i)$ and $C'(x')$ such that $D'(z_{k+1} \ldots, z_i)(C'(x')) = \beta_0 x' z_{k+1} \ldots z_i$, witnesses for $\rho_{k+1} \to \ldots \to \rho_1 \to a \leq \tau'$ where $\tau' = \tau_2 \to \ldots \to \tau_n \to a$. We assume that all these terms are canonical witnesses. The term $D'(z_{k+1} \ldots, z_i)$
is \( \lambda f^{\tau}.fS_2^{\tau_2} \ldots S_n^{\tau_n} \) and \( C(x') \) is \( \lambda y_2^{\tau_2} \ldots y_n^{\tau_n}.H'(x'T_{k+1}^{\rho_k} \ldots T_1^{\rho_1}) \) where \( H' = \varepsilon \) if the rule applied was \( P_2 \).

We need to show that there are terms \( D(z_1, \ldots, z_l) \) and \( C(x) \) that are witnesses for \( \models P \leq \tau. D(\pi) \) will have the form \( \lambda f^{\tau}.fS_1^{\tau_1} \ldots S_n^{\tau_n} \) and \( C(x) \) the form \( \lambda y_1^{\tau_1} \ldots y_n^{\tau_n}.H(xT_1^{\rho_1} \ldots T_1^{\rho_1}) \) where \( H = \varepsilon \) in the case of a single ground type. All that remains is to define \( S_1^{\tau_1} \) so it contains \( z_1, \ldots, z_k, T_1^{\rho_1}, \ldots, T_1^{\rho_1} \) and \( H \) (as an extension of \( H' \)). If \( U = \emptyset \) then \( H = H' \). Otherwise, let \( u \) be an odd length path such that \( U \) is associated with (so, its head variable is \( y_1^{\tau_1} \)). \( H \) consists of the suffix of \( u \) followed by the subtree \( H' \). The head variable of each \( T_i^{\rho_i} \) is \( y_i \) in the case of the single ground type and \( \lambda x \) in the general case (which is either \( y_1 \) or bound in \( u \)). We assume that \( S_i' \) is the subterm of \( S_i^{\tau_i} \) that is rooted at the initial vertex of the path \( w_i \); which is \( S_i^{\tau_i} \) itself in the single ground type. To complete these terms we require that \( T_i^{\rho_i} (S_i'(z_1, \ldots, z_k)) = \beta_n z_i \). Therefore, removing lambda abstraction over variables \( z_i \) and changing \( z_i \) to \( x_i \), we require that \( T_i(\pi_i)(S_i'(x_1, \ldots, x_k)) = \beta_n x_i \pi_i \). We construct a term \( C''(x_i) \) that occurs after the path \( w_i \) in \( S_i' \) (and which has root \( x_i \) when there is a single ground type). We also complete \( T_i(\pi_i) \) whose initial part is the tree \( U_i \) associated with the path \( w_i \).

First, we examine the single ground type case. So, \( S_i' \) will have the form \( \lambda u_1 \ldots u_m.S_i', C''(x_i) \) the form \( x_iC_1'' \ldots C_{m_i}'' \), and \( T_i(\pi_i) \) the form \( \lambda f^{\tau_i}.fV_i^1 \ldots V_i^m \). Assume \( D_i(\pi_i) = \lambda g_i^{w_i}.g_iW_i^1 \ldots W_i^m \) and \( C_i(x_i) = \lambda u_1 \ldots u_{i_j}.x_iC_1 \ldots C_{i_j} \). Assume \( w_i \) admits \( \sigma_i \); therefore, for some \( r : 1 \leq r \leq m, i_j = r \) (so, \( V_i^r \)) may contain occurrences of variables in \( \pi_i \). If \( w_i \) does not occur in the path \( w_i \) then we set \( V_i^r = W_i^r \). Otherwise, there is a non-empty subpath \( w_{ir} \) of \( w_i \) generated by \( u_{ir} \), and a subtree \( U_{ir} \) of \( V_i^r \) associated with \( w_{ir} \). Each \( C_j^i \) contains a single \( u_{is} \) (as head variable). Assume \( C_1^i \) contains \( u_r \). Assume that the path in \( W_i^r \) to the lambda node above \( z_{is} \) is \( w_{is}' \). If we can build the same path in \( V_i^r \) (by copying nodes of \( C_j^i \) to \( C_{is}^i \)) then we are done (letting \( V_i^r \) include this path followed by the subterm of \( W_i^r \) rooted at \( z_{is} \)). Otherwise, we initially include \( w_{ir} \) in \( C_{is}^i \) and then try to build \( w_{is}' \) in \( V_i^r \) by copying nodes of \( C_j^i \) to \( C_{is}^i \) in \( V_i^r \) and, therefore, in \( U_{ir} \), there is a path whose prefix except for its final variable vertex is the same as a prefix of \( w_{is}' \) and then differ. In the game \( G(C_{is}^i, V_i^r) \), play jumps from that variable in \( V_i^r \) to a lambda node in \( w_{ir} \). By definition of admits, there is a binder \( \nu \) labelled \( \lambda \pi \) in \( w_{ir} \) such that for some \( q \) not \((\nu' \downarrow q) \) for all nodes \( u_{is}' \) after \( \nu' \) in \( w_i \) (and in \( w_{is} \)). Therefore, we add a variable node labelled \( v_q \) to the end of \( w_{ir} \) in \( C_{is}^i \); so play jumps to a lambda node in \( V_i^r \) which is a successor of a leaf of \( U_{ir} \); below this node, we build the path \( w_{is}' \) except for its root node (by adding further nodes to \( C_{is}^i \) and add the subtree rooted at \( z_{is} \) in \( W_i^r \) to \( V_i^r \)).

For the general case, assume \( v_i(\sigma) = \sigma_1' \rightarrow \ldots \rightarrow \sigma_m' \rightarrow b \). So, \( S_i^{\sigma(\pi)} \) will have the form \( \lambda u_1 \ldots u_m.S_i', C''(x_i) \) the form \( H_i(x_iC_1'' \ldots C_{m_i}'' \ldots C_{s_i}'' \ldots C_1 \ldots C^i) \) and \( T_i(\pi_i) \) the form \( \lambda f^{\tau_i}.fV_i^1 \ldots V_i^m \). Assume \( D_i(\pi_i) = \lambda g_i^{w_i}.g_iW_i^1 \ldots W_i^m \) and \( C_i(x_i) = \lambda u_1 \ldots u_{i_j}.H_i(x_iC_1 \ldots C_{i_j}) \). We set \( H_i = H_i' \). Then we proceed in a similar fashion to the single base type case. If some \( u_{is}' \) does not occur in the path \( w_i \) then
$V^*_i = W^*_i$; otherwise we need to build similar paths to $z_{is}$ in $W^*_i$ in $V^*_i$ (by copying vertices from $C^*_i$ to $C^*_i$ and using that $w_i$ admits $(v_i(\sigma))_r$. □

The proof of completeness (by induction on the size of $\rho$) is easier.

**Theorem 2.** (Completeness) If $\models \rho \sqsubseteq \tau$ then $\vdash \rho \sqsubseteq \tau$.

5 Conclusion

We have provided tableau proof systems that characterise when a type is a retract of another type in simply typed lambda calculus (with respect to $\beta\eta$-equality). They offer a a nondeterministic decision procedure for the retract problem in EXPSACE: it may be possible to improve on the rather crude $k$-minimality bounds used on paths within the proof systems. Given the constructive proof of correctness, we also expect to be able to extract witnesses for a retract from a successful tableau proof tree (similar in spirit to [10]).

References