Foundational Nonuniform (Co)datatypes for Higher-Order Logic

Jasmin Christian Blanchette*, Fabian Meier†, Andrei Popescu‡, and Dmitriy Traytel†

*Inria & LORIA, Nancy, France, and Max-Planck-Institut für Informatik, Saarbrücken, Germany
†Institute of Information Security, Department of Computer Science, ETH Zürich, Switzerland
‡School of Science and Technology, Middlesex University London, UK

Abstract—Nonuniform (or “nested” or “heterogeneous”) datatypes are recursively defined types in which the type arguments vary recursively. They arise in the implementation of finger trees and other efficient functional data structures. We show how to reduce a large class of nonuniform datatypes and codatatypes to uniform types in higher-order logic. We programmed this reduction in the Isabelle/HOL proof assistant, thereby enriching its specification language. Moreover, we derive (co)recursion and (co)induction principles based on a weak variant of parametricity.

I. INTRODUCTION

Inductive (or algebraic) datatypes—often simply called datatypes—are a central feature of typed functional programming languages and of most proof assistants. A simple example is the type of finite lists over a type parameter α, specified as follows (in a notation inspired by Standard ML):

\[
\alpha \text{ list} = \text{Nil} \mid \text{Cons} \ (\alpha \ text{ list})
\]

A datatype is uniform if the recursive occurrences of the datatype have the same arguments as the definition itself, as is the case for list; otherwise, the datatype is nonuniform. Nonuniform types are also called “nested” or “heterogeneous” in the literature. Powerlists are nonuniform:

\[
\alpha \text{ plist} = \text{Nil} \mid \text{Cons} \ ((\alpha \times \alpha) \ plist)
\]

The type \(\alpha \text{ plist}\) is freely generated by the constructors \(\text{Nil}: \alpha \text{ plist}\) and \(\text{Cons}: \alpha \rightarrow (\alpha \times \alpha) \text{ plist} \rightarrow \alpha \text{ plist}\). When \(\text{Cons}\) is applied several times, the product type constructors \((\times)\) accumulate to create pairs, pairs of pairs, and so on. Thus, any powerlist of length 3 will have the form

\[
\text{Cons} \ (\text{Cons} \ (\text{Cons} \ (\text{Nil} \ (\text{Nil} \ (\alpha \text{ list})))) \ (\alpha \text{ list} \ (\alpha \text{ list})))
\]

Nonuniform datatypes arise in the implementation of efficient functional data structures, such as finger trees [24], and they underlie Okasaki’s bootstrapping and implicit recursive slowdown optimization techniques [37]. Yet many programming languages and proof assistants lack proper support for such types. For example, even though Standard ML allows nonuniform definitions, a typing restriction disables interesting recursive definitions. As for proof assistants, Agda, Coq, Lean, and Matita allow nonuniform definitions, but they are built into the logic (dependent type theory), with all the risks and limitations that this entails [12, Section 1].

For systems based on higher-order logic such as HOL4, HOL Light, and Isabelle/HOL, no dedicated support exists for nonuniform types, probably because they are widely believed to lie beyond the logic’s ML-style polymorphism. Building on the well-understood metatheory of uniform datatypes (Section II), we disprove this folklore belief by showing how to define a large class of nonuniform datatypes by reduction to their uniform counterparts within higher-order logic (Section III).

Our constructions allow variations along several axes. They cater for mutual definitions:

\[
\alpha \text{ ptree} = \text{Node} \ (\alpha \text{ pforest})
\]

\[
\alpha \text{ pforest} = \text{Nil} \mid \text{Cons} \ (\alpha \text{ ptree} \ ((\alpha \times \alpha) \text{ pforest})
\]

They allow multiple recursive occurrences, with different type arguments:

\[
\alpha \text{ plist}' = \text{Nil} \mid \text{Cons}_1 \ (\alpha \text{ plist}) \mid \text{Cons}_2 \ (\alpha \times \alpha) \text{ plist'}
\]

They allow multiple type arguments, which may all vary:

\[
(\alpha, \beta) \text{ tplist} = \text{Nil} \beta \mid \text{Cons} \ ((\alpha \times \alpha, \text{unit} + \beta) \text{ tplist})
\]

Moreover, they allow the presence of datatypes, codatatypes, and other well-behaved type constructors both around the type arguments and around the recursive type occurrences:

\[
\alpha \text{ streel} = \text{Node} \ ((\alpha \text{ fset} \ \text{stree}) \text{ fset})
\]

Here, \(\text{fset}\) is the type constructor associated with finite sets.

Furthermore, the constructions can be extended to coinductive (or coalgebraic) datatypes—codatatypes. Codatatypes are generally non-well-founded, allowing infinite values. They are often used to model the datatypes of languages with a nonstrict (lazy) evaluation strategy, such as Haskell, and they can be very convenient for some proving tasks. The codatatype definition

\[
\alpha \text{ pstream} = \text{Cons} \ ((\alpha \times \alpha) \text{ pstream})
\]

introduces “powerstreams,” with infinite values of the form

\[
\text{Cons} \ (\text{Cons} \ (\text{Cons} \ (\text{Nil} \ (\text{Nil} \ (\alpha \text{ list})))) \ (\alpha \text{ list} \ (\alpha \text{ list})))
\]

Unlike dependent type theory, higher-order logic requires all types to be nonempty (inhabited). To introduce a new type, we must both provide a construction in terms of existing types and prove its nonemptiness. For example, a datatype specification analogous to the \(\text{pstream}\) codatatype above should be rejected. In previous work [13], we showed how to decide the nonemptiness problem for uniform types—including mutually recursive specifications and arbitrary mixtures of datatypes and codatatypes—by viewing the definitions as a grammar,
with the defined types as nonterminals. Here, we extend this result to nonuniform types (Section IV). This is achieved by encoding the nonuniformities in a generalized grammar, which can decide nonemptiness of the sets that arise in the construction of the nonuniform types.

Once a datatype has been introduced, users want to define functions that recurse on it and carry out proofs by induction involving these functions—and similarly for codatatypes. A uniform datatype definition generates an induction theorem involving these functions—and similarly for codatatypes. A functions that recurse on it and carry out proofs by induction by encoding the nonuniformities in a generalized grammar, with the defined types as nonterminals. Here, we extend this parameterized by a polymorphic property \( \text{bool} \) because the second and third occurrences of the variable \( Q \) should look like this:

\[
\forall Q. \ Q \text{Nil} \land (\forall x. x. Q \text{xs} \Rightarrow Q \ (\text{Cons} \ x \text{xs})) \Rightarrow \forall ys. Q \ ys
\]

However, this formula is not typable in higher-order logic, because the second and third occurrences of the variable \( Q \) in need different types: \((\alpha \times \alpha) \text{pl list} \to \text{bool} \) versus \( \alpha \text{pl list} \to \text{bool} \). Our solution is to replace the theorem by a procedure parameterized by a polymorphic property \( \varphi_\alpha : \alpha \text{pl list} \to \text{bool} \) (Section V). For \text{pl list}, the procedure transforms a proof goal of the form \( \varphi_\alpha \ys \) into two subgoals \( \varphi_\alpha \text{Nil} \) and \( \forall x. x. \varphi_\alpha \times \alpha \text{xs} \Rightarrow \varphi_\alpha \ (\text{Cons} \ x \text{xs}) \). A weak form of parametricity is needed to recursively transfer properties about \( \varphi_\alpha \) to properties about \( \varphi_\alpha \times \alpha \). Our approach to (co)recursion is similar (Section VI).

All the constructions and derivations are formalized in the Isabelle/HOL proof assistant and form the basis of high-level commands that let users define nonuniform types and (co)recursive functions on them and reason (co)inductively about them (Section VII). The commands are foundational: Unlike all previous implementations of nonuniform types in proof assistants, they require no new axioms or extensions of the logic. An example involving \( \lambda \)-terms in De Bruijn notation demonstrates the practical potential of our approach.

Our main contributions are the following:

- We designed a reduction of nonuniform datatypes to uniform datatypes within the relatively weak higher-order logic, including recursion and induction.
- We adapted the constructions to support codatatypes as well, exploiting dualities.
- We coded the reduction in a proof assistant based on higher-order logic, yielding a first implementation of nonuniform datatypes without dependent types.

The formal proofs, the source code, and the examples are publicly available [11].

II. PRELIMINARIES

A. Higher-Order Logic

We consider classical higher-order logic (HOL) with Hilbert choice, the axiom of infinity, and rank-1 polymorphism. HOL is based on Church’s simple type theory [14]. It is the logic of Gordon’s original HOL system [18] and of its many successors and emulators, notably HOL4, HOL Light, and Isabelle/HOL.

Primitive types are built from type variables \( \alpha, \beta, \ldots \), a type \( \text{bool} \) of Booleans, and an infinite type \( \text{ind} \) using the function type constructor; thus, \((\text{bool} \to \text{ind}) \to \text{ind} \) is a type. Primitive constants are equality \( = : \alpha \to \alpha \to \text{bool} \), the Hilbert choice operator, and 0 and \text{Suc} for \text{ind}.

Terms are built from constants and variables by means of typed \( \lambda \)-abstraction and application.

A polymorphic type is a type \( T \) that contains type variables. If \( T \) is polymorphic with variables \( \forall \alpha \in \text{F} \), we write \( \bar{T} \) instead of \( T \). Formulas are closed terms of type \( \text{bool} \). The logical connectives and quantifiers on formulas are defined using the primitive constants—e.g., True as \( (\lambda x : \text{bool} \cdot x) = (\lambda x : \text{bool} \cdot x) \). Polymorphic formulas are thought of as universally quantified over their type variables. For example, \( \forall x : \alpha. x = x \) really means \( \forall Q. \forall x. x = x \). Nested type quantifications such as \( (\forall Q. \ldots) \Rightarrow (\forall Q. \ldots) \) are not expressible. Since HOL was designed to support mathematical reasoning, we will express the concepts in standard mathematical language. Occasionally, when we hit the limitations of HOL, we will indicate so.

The only primitive mechanism for defining new types in HOL is type definition: For any existing type \( \bar{T} \) and predicate \( P : \bar{T} \to \text{bool} \) such that \( \{ x : \bar{T} \mid P x \} \) is nonempty, we can introduce a type \( \bar{S} \) isomorphic to \( \{ x : \bar{T} \mid P x \} \). Upon meeting the definition \( \bar{S} = \{ x : \bar{T} \mid P x \} \), the system first requires the user to prove \( \exists x : \bar{T}, P x \) and then introduces the type \( \bar{S} \), the projection \( \text{Rep}_\alpha : \bar{S} \to \bar{T} \), and the embedding \( \text{Abs}_\alpha : \bar{T} \to \bar{S} \) such that \( \forall x. P (\text{Rep}_\alpha x), \forall x. \text{Abs}_\alpha (\text{Rep}_\alpha x) = x, \) and \( \forall x. P x \Rightarrow \text{Rep}_\alpha (\text{Abs}_\alpha x) = x \). The nonemptiness check is necessary because all types in HOL must be nonempty [18], [38].

Thus, unlike dependent type theory, HOL does not have (co)datatypes as primitives. However, datatypes [5], [19], [20], [33] and, more recently, codatatypes [43] are supported via derived specification mechanisms. Users can write fixpoint definitions such in ML-style notation, and the system automatically defines the type using nonrecursive type definitions (ultimately appealing to the infinite type \( \text{ind} \) and the function space); defines the constructors and related operators; and proves characteristic properties, such as injectivity of constructors, induction theorems, and recursion theorems.

B. Bounded Natural Functors

In this paper, we take uniform (co)datatypes for granted, thus assuming the availability of types such as \( \alpha \text{list} \). Often it is useful to think not in terms of polymorphic types, but in terms of type constructors. For example, \( \text{list} \) is a type constructor in one variable, sum \((+) \) and product types \((\times) \) are binary type constructors. Most type constructors are not only operators on types but have a richer structure, that of bounded natural functors (BNFs) [43].

We write \([n]\) for \{1, ..., n\} and \(\alpha \text{set} \) for the powertype of \( \alpha \), consisting of sets of \( \alpha \) elements; it is isomorphic to \( \alpha \to \text{bool} \). An \( n \)-ary BNF is a tuple \( F = (F, \text{map}_F, (\text{set}_F)^{\leq [n]} , \text{bd}_F) \), where

- \( F \) is an \( n \)-ary type constructor;
- \( \text{map}_F : (\alpha_1 \to \beta_1) \to \cdots \to (\alpha_n \to \beta_n) \to \tau F \to \tau \beta F; \)
- \( \text{set}_F : \tau F \to \alpha_i \text{ set} \) for \( i \in [n]; \)
- \( \text{bd}_F \) is an infinite cardinal number

satisfying the following properties:

- \((F, \text{map}_F) \) is an \( n \)-ary weak-pullback-preserving functor.
Each $\text{set}_F$ is a natural transformation between the functor $(F, \text{map}_F)$ and the powerset functor $(\text{set}, \text{image})$.

$\forall i \in [n]. \forall a \in \text{set}_F x. \ f_i a = g_i a \Rightarrow \text{map}_F \ F x = \text{map}_F \ G x.$

$\forall i \in [n]. \forall x : (\alpha_1, \alpha_2) \ F. \ |\text{set}_F x| \leq b d_F.$

For example, list is a unary BNF, where $\text{map}_{\text{list}}$ is the standard map function, $\text{set}_{\text{list}}$ collects all elements occurring in its argument, and $b d_{\text{list}}$ is the cardinality of the natural numbers $\mathit{nat}$.

BNFs are closed under uniform (least and greatest) fixpoint definitions [43] and the nonemptiness problem for BNFs constructed by basic functors, fixpoints and composition is decidable [13]. These crucial properties enable a modular approach to mixing and nesting uniform (co)datatypes and deciding if a high-level specification yields valid, i.e., nonempty, HOL types. In addition, BNFs display predicator and relator structure [41]. The predicator $\text{pred}_F : (\alpha_1 \to \text{bool}) \to \cdots \to (\alpha_n \to \text{bool}) \to F \to \text{bool}$ and the relator $\text{rel}_F : (\alpha_1 \to \beta_1 \to \text{bool}) \to \cdots \to (\alpha_n \to \beta_n \to \text{bool}) \to F \to \overline{F} \ F \to \text{bool}$, are defined from $\text{set}_F$ and $\text{map}_F$ as follows:

$\text{pred}_F \ F x \iff \forall i \in [n]. \forall a \in \text{set}_F x. \ P a$;

$\text{rel}_F \ F x y \iff (\exists \ z \in [n]. \text{set}_F x \subseteq \{(a, b) : R_i a b\}) \land $ 

$\text{map}_F \ \text{fst} x = x \land \text{map}_F \ \text{snd} y = y$, where $\text{fst}$ and $\text{snd}$ are standard projection functions on the product type $\times$.

For list, $\text{pred}_{\text{list}} \ P x s$ states that $P$ holds for all elements of the list $x s$ and $\text{rel}_{\text{list}} \ R x s y s$ states that $xs$ and $ys$ have the same length and are element-wise related by $R$.

Relators and predicatures are useful to track parametricity [40], [44]. A polymorphic constant $c : \pi F$ is parametric if, for all relations $R_i : \alpha_i \to \beta_i \to \text{bool}$ for each $i \in [n]$ we have $\text{rel}_F \ F c c$—i.e., every two instances of $c$ are related by the lifting of the relations associated with the component types. Parametricity applies not only to BNFs but also to any combination of BNFs using the function space. For polymorphic functions $f : \pi F \to \pi G$ between two BNFs, $f$ is parametric if and only if $f$ is a natural transformation (Appendix A).

III. (CO)DATATYPE DEFINITIONS

Before describing the reduction of nonuniform (co)datatypes to uniform (co)datatypes in full generality, we start with a simple example that conveys the main idea. The reduction proceeds by defining a larger uniform datatype and carving out a subset that is isomorphic to the desired nonuniform type. To prove that the constructed type is the intended one, we establish the isomorphism between the defined nonuniform type and the right-hand side of its specification.

A. An Example: Powerlists

Okasaki [37, Section 10.1.1] sketches how to mimic nonuniform datatypes using uniform datatypes. He approximates powerlists by the following definitions:

$\text{datatype} \ a \ \text{sh} = \text{Leaf} a \ | \ \text{Node} (a \ \text{sh} \times a \ \text{sh})$

$\text{datatype} \ a \ \text{raw} = \text{Nil}_0 \ | \ \text{Cons}_0 (a \ \text{sh}) (a \ \text{raw})$

The type $a \ \text{raw}$ corresponds to lists of binary trees $a \ \text{sh}$. It is larger than powerlists in two ways: (1) $a \ \text{sh}$ allows non-full binary trees, which cannot arise in a powerlist; (2) $a \ \text{raw}$ imposes no restriction on the depth of the binary trees, whereas a powerlist has elements successively of depth $0, 1, 2, \ldots$.

Okasaki considers these mismatches as one of two disadvantages of the above encoding. The other disadvantage is that the encoding requires users to insert Leaf and Node coercions to convert an element such as $(a, b), (c, d) : (\times a \times) \times (\times a \times)$ to Node (Leaf $a$, Leaf $b$), Node (Leaf $c$, Leaf $d$) : $a \ sh$.

We overcome the first disadvantage by using a type definition. From the raw type, we select exactly those inhabitants that correspond to powerlists. To achieve this, we define two predicatures, $\mathit{ok} : a \ sh \to \text{bool}$ and $\mathit{ok} : a \ sh \to \text{bool}$, as the least predicatures satisfying the following rules:

$\mathit{ok} n \text{Nil}_0 \mathit{ok} n \text{Cons}_0 (x s x s) \Rightarrow \mathit{ok} n (\mathit{Cons}_0 x s x s)$

The second disadvantage is addressed by hiding the internal construction of $a \ \text{plist}$. We define the powerlist constructors $\text{Nil} : a \ \text{plist}$ and $\text{Cons} : a \to (a \times a) \ \text{plist} \to a \ \text{plist}$ in terms of $\text{Nil}_0$ and $\text{Cons}_0$. These definitions will require some additional machinery on the raw type.

B. General Type Construction

We assume that the desired nonuniform datatype has a single constructor. Separate constructors are easy to introduce as syntactic sugar [10, Section 4]). For powerlists, the single constructor definition is $a \ \text{plist} = \mathcal{C} \mathcal{t} \mathcal{r} \ \text{plist} (\text{unit} + a \times (a \times a) \ \text{plist})$. It corresponds to finding a least solution (up to isomorphism) to the type fixpoint equation $\alpha \ \text{plist} \simeq (\alpha, (a) F) G$ with $a F = a \times a$ and $\alpha G = \text{unit} + a \times a$.

We generalize this setting in multiple dimensions. First, we support a simultaneous definition of an arbitrary number $i$ of mutual nonuniform datatypes. For example, $\text{ptree}$ and $\text{pforest}$ from Section I are given by the system of fixpoint equations

$\alpha \ \text{ptree} \simeq (\alpha, (a, F_1) \ \text{ptree}, (a, F_2) \ \text{pforest}) \ G_1,$

$\alpha \ \text{pforest} \simeq (\alpha, (a, F_3) \ \text{ptree}, (a, F_4) \ \text{pforest}) \ G_2$

where $(\alpha, \beta, \gamma) G_1 = a \times \gamma, (\alpha, \beta, \gamma) G_2 = \text{unit} + \beta \times \gamma, a F_1 = a F_2 = a F_3 = a, \alpha F_4 = a \times a$. We assume that all $G$'s depend on the same type variables, even though the dependence may be spurious, as in the case of $G_1$ and $\beta$.

Second, a type may occur several times on the right-hand side of a definition. We support an arbitrary number $j$ of such occurrences. This feature is used in the $\text{pplist}$ type: $a \ \text{pplist} \simeq (a, (a, F_1) \ \text{pplist}, (a, F_2) \ \text{pplist}) \ G$, where $(\alpha, \beta, \gamma) G = \text{unit} + a \times a \gamma, a F_1 = a, \alpha F_2 = a \times a$.

Finally, the construction supports an arbitrary number $k$ of type parameters. The parameter changes may differ for different type parameters, such as in the $\text{tplist}$ example:

$(\alpha, \beta) \ \text{tplist} \simeq (\alpha, (\alpha, (a, F_1) \ \text{tplist}, (a, F_2) \ \text{tplist}) \ G$, where $(\alpha, \beta, \gamma) G = \beta + a \times a, (\alpha, \beta) F_1 = a \times a, \alpha F_2 = \text{unit} + B$. As before for the $G$'s, all $F$'s may depend on all type parameters of the specified nonuniform type.

In the sequel, the indices $i$, $j$, and $k$ range over $[i], [j], \text{and } [k]$, respectively. Moreover, we abbreviate indexed sequences
using a horizontal bar; for example, $\alpha$ stands for $\alpha_1, \ldots, \alpha_k$, and $\overline{\alpha} F_1$ stands for $\overline{\alpha} F_{11}, \ldots, \overline{\alpha} F_{1k}$. It should be clear from the context which index is omitted.

A definition of $i$ mutual nonuniform datatypes $T_i$ is a system of $i$ type fixpoint equations

$$\overline{\alpha} \; T_i \simeq \left( \overline{\alpha}, \left( \overline{\alpha} F_1 \right) T_{\sigma(1)}, \ldots, \left( \overline{\alpha} F_k \right) T_{\sigma(j)} \right) G_i$$

where the $G_i$’s are $(k+j)$-ary BNFs, the $F_{jk}$’s are $k$-ary BNFs, and $\sigma : [j] \to [i]$ is a monotone surjective function that expresses which of the $i$ mutual types belong to which recursive occurrence. The construction generalizes Okasaki’s idea and yields $k$-ary BNFs $T_i$ that are least solutions (up to isomorphism) to equation (1). A uniform datatype definition [10] is a special case with $j = i$, $\sigma(i) = i$, and $\overline{\alpha} F_{jk} = \alpha_k$.

We start by defining the shape types $\overline{\alpha} \; sh_k$ that overapproximate the recursive changes to the type arguments. There are $k$ shape types, corresponding to the $k$ type arguments, and they are mutually recursive uniform datatypes:

$$\overline{\alpha} \; sh_k = \text{Leaf}_k \alpha_k \mid \text{Node}_{ik} \left( \overline{\alpha} \; sh \; F_{ik} \right) \mid \cdots \mid \text{Node}_{jk} \left( \overline{\alpha} \; sh \; F_{jk} \right)$$

For $\text{plist}$, the $sh$ type is $sh$. In general, each recursive occurrence may change the type arguments in a different way; this is reflected in the different Node constructors.

Next, we define $i$ uniform mutually recursive datatypes $\overline{\alpha} \; raw_i$ that recurse through the $G_i$’s in the same way as the $T_i$’s do, except that they keep the type arguments unchanged. The immediate $\overline{\alpha} sh$ arguments to $G_i$ are replaced by $\overline{\alpha} sh$:

$$\overline{\alpha} \; raw_i = \text{Raw}_i \left( \left[ \overline{\alpha} \; sh \right], \overline{\alpha} \; raw_{\sigma(1)} \right) \cdots \overline{\alpha} \; raw_{\sigma(j)} \right) G_i$$

For every $i$, we specify subsets of the types $\overline{\alpha} raw_i$ that are isomorphic to the nonuniform types $T_i$, by defining predicates $\delta_k$ that characterize the allowed shapes and their changes in the recursion. As in the powerlist example, the definition of $\delta_k$ relies on auxiliary predicates $\overline{\delta}_{jk} : [j] \; \text{list} \to \overline{\alpha} \; sh_k \to \text{bool}$ on the shape types. The type of $\overline{\delta}_{jk}$ shows an important difference to the $\text{plist}$ example: The first argument is not just a natural number denoting the depth of a full tree but has more structure. We call it the shadow of the shape and let $\Delta$ stand for $[j] \; \text{list}$. The additional structure is necessary because different Node constructors may occur in a single shape element. These occurrences in the full shape trees are layered: All Node constructors right above the Leaf constructors belong to a fixed occurrence $j$. The next layer of Nodes may belong to a different fixed occurrence $f$. The shadow summarizes the occurrence indices. Consider $\text{Con}_s \; u \left( \text{Con}_x v \left( \text{Con}_y \; w \left( \text{Con}_z \; \text{Nil} \right) \right) \right) : \text{plist}$. This order of constructors forces $x$’s type to be $\alpha F = \alpha F_1 F_2 F_3$, with $\alpha F_1 = \alpha$ and $\alpha F_2 = \alpha \times \alpha$. Consequently, $x$ is embedded into $\alpha sh$ as $\text{Node}_2 \left( \text{map}_F F_2 \text{Node}_2 \left( \text{map}_F \text{Node}_1 \left( \text{map}_F \text{Leaf} x \right) \right) \right)$, whose shadow is $[2, 2, 1]$.

Formally, the predicates $\overline{\delta}_{jk}$ are defined together as the least predicates satisfying the rules

$$\overline{\delta}_{jk} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) x \cdots \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) u \Rightarrow \overline{\delta}_{jk} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) F_j \left( \alpha \right) F_{jk} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) f \Rightarrow \overline{\delta}_{jk} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) \left( j \, u \right) \left( \text{Node}_{jk} \alpha \right) f$$

where $\alpha$ and $\alpha$ are notations for Nil and Cons. To access the recursive components of $sh$, we rely on the predicates associated with the $F$’s. Predicators are monotone. The $i$ mutual predicates $\delta_k : \Delta \to \overline{\alpha} \; raw_i \to \text{bool}$ are defined similarly:

$$\text{pred}_{\overline{\delta}_{jk}} \left( \overline{\delta}_{jk} \right) \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) u \cdots \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) u \left( \overline{\alpha} \; raw_{\sigma(1)} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) \left( j \, u \right) \right) g \Rightarrow \overline{\delta}_{jk} \left( \alpha \right) \left( \text{Leaf}_{j} \alpha \right) \left( j \, u \right) g$$

We access the $k$ immediate components of the shape type and the $j$ recursive components of the raw type through the predicator. We write that an element $r$ of type $\overline{\alpha} \; raw_i$ has shadow $u$ if $\delta_k u \; r$ holds.

Finally, the nonuniform type $T_i$ can be defined as the subset of $\overline{\alpha} \; raw_i$ that satisfies the $\delta_k$ predicate for the empty shadow: $\overline{\alpha} \; T_i = \{ r : \overline{\alpha} \; raw_i \mid \delta_k r \}$.

This property is sufficient to prove that $T_i$ is a BNF. By virtue of being a BNF, $T_i$ can appear around type arguments and recursive type occurrences in future uniform or nonuniform (co)datatype definitions.

**C. Nonuniform Constructors**

If the type $T_i$ is the nonuniform type that we intended to construct, it should satisfy equation (1). We prove this isomorphism by defining a constructor $\text{Ctor}_i : \left( \overline{\alpha}, \left( \overline{\alpha} F_1 \right) T_{\sigma(1)}, \ldots, \left( \overline{\alpha} F_k \right) T_{\sigma(j)} \right) G_i \to \overline{\alpha} \; T_i$ and a destructor $\text{Dt}_{\sigma(j)} : \overline{\alpha} \; T_i \to \left( \overline{\alpha}, \left( \overline{\alpha} F_1 \right) T_{\sigma(1)}, \ldots, \left( \overline{\alpha} F_k \right) T_{\sigma(j)} \right) G_i$ by showing that they are inverses of each other.

Figure 1 gives diagrammatic definitions of $\text{Ctor}_i$ (by composing the functions on the outer arrows) and $\text{Dt}_{\sigma(j)}$ (by composing the functions on the inner arrows). All shape and raw types occurring in the diagram are annotated with their shadows. Abs$_k$ can be applied only to raw elements with shadow $[]$. The unLeaf$_k$ and unRaw$_i$ functions are inverses of the corresponding constructors satisfying unLeaf$_k$ (Leaf$_k$ $\alpha$) = $\alpha$ and unRaw$_i$ (Raw$_i$ $\alpha$) = $\alpha$. Note that unLeaf$_k$ is underspecified and (just as Abs$_k$) may be applied only to Leaf$_k$ shape elements with shadow $[]$. Moreover, the definition must bridge the gap between the types $\overline{\alpha} F_j \overline{\alpha} raw_{\sigma(j)}$ of shadow $[]$ and $\overline{\alpha} \; raw_{\sigma(j)}$ of shadow $[j]$ (the rightmost arrows in Figure 1). This must happen recursively (even though the constructor $\text{Ctor}_i$ itself is not recursive), by inlining the additional $F$s into a new layer of the shape type (right above the Leaf constructors) and therefore requires a generalization that transforms an arbitrary shadow $u$ into $\alpha \cdot j$ (i.e., the list $u$ with the element $j$ appended to it). For each fixed $j$, inlining is implemented by means of $i$ mutual primitively recursive functions $\overline{\delta}_{j} : \overline{\alpha} \; F_j \overline{\alpha} \; raw_{\sigma(j)} \to \overline{\alpha} \; sh_k \circ \overline{\alpha} \; sh_k$ on the shape type.
Inlining is injective. We define the (partial) inverse operations \( \downarrow_{ji} : \Delta \rightarrow \overline{\alpha} \text{raw}_{ji} \rightarrow (\overline{\alpha} F^j) \text{raw}_{ji} \) and \( \uparrow_{ji} : \Delta \rightarrow \overline{\alpha} \text{sh}_{ji} \rightarrow (\overline{\alpha} F^j) \text{sh}_{ji} \), which are useful when defining the destructors \( \text{dtor} \). The additional shadow parameter in \( \Box_{ji} \) denotes how many more layers to destruct until we arrive at the last layer of Nodes (with only Leaf constructors below).

\[
\begin{align*}
\Box_{ji} \emptyset (\text{Node}_{ji} \, f) &= \text{Leaf}_{k} ((\text{map}_{F^j} \, \text{unLeaf}_{j}) \, f) \\
\Box_{ji} (\ell \cdot u) (\text{Node}_{ji} \, f) &= \text{Node}_{ji} ((\text{map}_{F^j} \, \Box_{ji} \ell) \cdot u) \, f \\
\downarrow_{ji} u (\text{Raw}_{ji} \, g) &= \\
& \text{Raw}_{ji} (\text{map}_{G_{ji}} \, \Box_{ji} \ell) \cdot u (\downarrow_{j, r, i} \, (1 \cdot u)) \ldots (\downarrow_{j, r, i} \, \emptyset) \cdot g)
\end{align*}
\]

We establish the expected behavior of \( \uparrow_{ji} \) and \( \downarrow_{ji} \) with respect to shadows and prove that they are mutually inverse. The proofs proceed by induction on the \textit{raw} type using very similar omitted auxiliary lemmas for \( \Box_{ji} \) and \( \Box_{ji} \).

**Lemma 2:**
1) \( \text{ok}_{ji} \cdot u \cdot r \Rightarrow \text{ok}_{ji} (u \cdot r) \cup (\uparrow_{ji} \cdot r) \); 3) \( \text{ok}_{ji} \cdot u \Rightarrow \downarrow_{ji} u (\uparrow_{ji} \cdot r) = r \);
2) \( \text{ok}_{ji} \cdot (u \cdot r) \Rightarrow \text{ok}_{ji} (u \cdot (\downarrow_{ji} \cdot r)) \); 4) \( \text{ok}_{ji} \cdot (u \cdot r) \Rightarrow \text{ok}_{ji} (\downarrow_{ji} \cdot (u \cdot r)) = r \).

Since every pair of arrows in Figure 1 is mutually inverse (when applied to elements of the right shadow), we obtain our desired isomorphism property for \( \Box_{ji} \) and \( \Box_{ji} \).

**Theorem 3:** \( \text{dtor} (\text{Ctor} (g)) = g \) and \( \text{Ctor} (\text{dtor} (g)) = g \).

Finally, we prove characteristic theorems for \( T_{i} \)’s BNF constants. We focus on the property that \( \text{map}_{R_{i}} \) commutes with the constructor \( \text{Ctor} \). The theorems for the relator, the predicate, and the set functions are proved analogously.

**Theorem 4:** \( \text{map}_{F^i} \, \overline{F} (\text{Ctor} \, g) = \text{Ctor} \, (\text{map}_{R_{i}} \, \overline{F} \, g) \) where \( \overline{\alpha} R_{i} = (\overline{\alpha}, (\overline{\alpha} F^i) \, T_{i}(\emptyset), \ldots, (\overline{\alpha} F^i) \, T_{i}([j \mapsto i])) \) \( G_{i} \) and \( \text{map}_{R_{i}} \) is the map function associated to this composite BNF.

The proof of Theorem 4 relies on commutation properties of \( \text{map}_{\text{raw}_{i}} \) and \( \uparrow_{ji} \) and of \( \text{map}_{\text{sh}_{i}} \) and \( \Box_{ji} \) that can be proved by induction. This is a pervasive pattern when defining recursive functions on nonuniform datatypes.

**Lemma 5:** \( \text{map}_{\text{sh}_{i}} \, \Box_{ji} \, s) = \Box_{ji} \, (\text{map}_{\text{sh}_{i}} \, \text{map}_{F_{i}} \, f) \, s \) and \( \text{map}_{F_{i}} \, \overline{\alpha} \cup \emptyset \, r) = \uparrow_{ji} \, (\text{map}_{\text{raw}_{i}} \, \text{map}_{F_{i}} \, f) \, r) \).

### D. Nonuniform Codatatypes

The construction can be gracefully extended to support types whose elements may be infinitely deep: codatatypes. Given a definition as in equation (1), codatatypes are exactly the types \( T_{i} \) that are the greatest solution to this equation.

This change in semantics needs to be reflected only at the \textit{raw} level. Accordingly, the \textit{raw} types are defined as an uniform mutual codatatype definition. The shape types remain unchanged, even since in an infinitely deep object all type arguments are finite (but unbounded) type expressions.

The subsequent changes are similarly mild: the predicates \( \text{ok}_{i} \) are now defined as a mutual greatest (or coinductive) fixpoint of the same introduction rule; the functions \( \uparrow_{ji} \) and \( \downarrow_{ji} \) are defined by primitive corecursion, using the same equations as in the datatype case though.

All theorems from Subsections III-B and III-C hold exactly as stated also for codatatypes. The proofs however are different: for example while propositions 1) and 2) of Lemma 2 were proved by induction on \( r \) for datatypes, for codatatypes the proof proceeds by coinduction on the now coinductive definitions of \( \text{ok}_{i} \). Similarly, the equational statements (e.g., 3) and 4) of Lemma 2 or the raw part of Lemma 5) are proved by coinduction on the \( = \) predicate.

**IV. THE NONEMPENESS PROBLEM**

All types in HOL are required to be nonempty. This is a fundamental design decision connected to the presence of Hilbert choice in HOL [18], [38]. As we are developing more sophisticated high-level datatype specification mechanisms, the problem of establishing nonemptiness of the introduced types becomes more difficult.

For nonuniform (co)datatypes \( T_{i} \) specified mutually recursively, the question is whether \( T_{i} \) are indeed valid HOL types, i.e., are nonempty. We are interested in an answer that is both automatic, i.e., is given without asking the user to perform any proof, and complete, i.e., does not reject any valid types.

In previous work, we designed a solution for mutual uniform (co)datatypes [13]. It is based on storing, for each BNF \( \overline{\alpha} K \) with \( \overline{\alpha} = (\alpha_{1}, \ldots, \alpha_{n}) \), complete information on its conditional nonemtness, i.e., on which combinations of nonemptiness assumptions for the argument types \( \alpha_{i} \), would be sufficient to guarantee nonemptiness of \( \overline{\alpha} K \). For example, if \( n = 3 \) and \( \overline{\alpha} K \) is \( \alpha_{1} \text{stream} + \alpha_{2} \times \alpha_{3} \), then for \( \overline{\alpha} K \) to be nonempty it suffices that either \( \alpha_{1} \), or both \( \alpha_{2} \) and \( \alpha_{3} \) be nonempty. We say that \{\{1\}\} and \{\{2,3\}\}, or, simply, \{1\} and \{2,3\} are witnesses for the nonemptiness of \( \alpha_{1} \text{stream} + \alpha_{2} \times \alpha_{3} \).

The above discussion assumes that \( K \) operates on possibly nonempty collections of elements (which, technically, as a type constructor, it does not, since the HOL type variables are assumed to range over nonempty types). To model this, we employ the \( \text{set}_{\alpha} \) operators to capture the action of \( K \) on \( \text{sets} \), as the homonymous constant \( K : \alpha_{1} \text{set} \rightarrow \ldots \rightarrow \alpha_{n} \text{set} \rightarrow (\overline{\alpha} K) \text{set} \), defined by \( K \alpha_{1} \ldots \alpha_{n} \{x : \overline{\alpha} K \mid \forall i \in [n], \text{set}_{\alpha_{i}} x \subseteq A_{i} \} \). This way, the constant \( K \) operates on sets like the type
constructor \( K \) operates on types. And since sets can be empty, we are able to express witnesshood:

Given \( I \subseteq [n] \), we call \( I \) a witness for \( K \) if, for all sets \( \mathcal{A}, \forall i \in I, A_i \neq \emptyset \) implies \( K \mathcal{A} \neq \emptyset \). A set \( \mathcal{I} \subseteq [n] \) set of witnesses for \( K \) is called perfect if for all witnesses \( I \subseteq [n] \) there exists \( I \in \mathcal{I} \) such that \( I \subseteq J \). Thus, a perfect set of witnesses for \( K \) is one where no witness is missed, in that any witness \( J \) is equal to, or improved by, an \( I \in \mathcal{I} \).

We fix a definition of \( i \) mutual nonuniform datatypes \( T_i \), as depicted in equation (1) from Section III-B. We assume the involved BNF’s, namely, each \( G \) and each \( F_i \), are endowed with perfect sets of witnesses \( \mathcal{I}(G) \) and \( \mathcal{I}(F_i) \). We will show how effectively to construct perfect sets of witnesses for the \( T_i \)’s. On the one hand, this allows us to decide when the \( T_i \)’s are nonempty, hence valid HOL types—if and only if their perfect sets are nonempty. On the other hand, this equips the \( T_i \)’s with infrastructure needed to establish nonemptiness in future (co)datatypes that may use them as parameters.

To identify the witnesses for the \( T_i \)’s, we try a similar approach to what we did for uniform datatypes. There, we define a context-free grammar except that its productions act on define a context-free set-grammar (which is like a standard approach to what we did for uniform datatypes. There, we derivations following the definitions

\[
\begin{align*}
\sigma T_i & \overset{\text{dctor}}{\rightarrow} (\sigma, \sigma T_1, \ldots, \sigma T_i) G_i \\
n & \in \mathcal{I}
\end{align*}
\]

with each \( T_i \) deriving sets containing the nonterminals \( T_i \) and the terminals \( \alpha_k \) allowed by witnesses of \( G_i \).

For the nonuniform case, when applying recursively productions following the definitions

\[
\begin{align*}
\sigma T_i & \overset{\text{dctor}}{\rightarrow} (\sigma, (\sigma F_1) T_{\sigma(1)}, \ldots, (\sigma F_i) T_{\sigma(i)}) G_i \\
n & \in \mathcal{I}
\end{align*}
\]

we see that the \( T_i \)’s become applied to larger and larger polynomial expressions involving the \( F_i \)’s. To keep the set-grammar finite, we take a more abstract view, retaining from the \( F_i \)’s only their witness-relevant information, obtained by suitably combining their perfect sets \( \mathcal{I}(F_i) \). We define the set \( \text{PolyWit} \) of polynomial witness sets (polywits for short), inductively as follows:

- If \( k \in [k] \), then \( \{k\} \in \text{PolyWit} \).
- If \( (j,k) \in [j] \times [k] \) and \( p_1, \ldots, p_k \in \text{PolyWit} \), then \( (p_1, \ldots, p_k) : (\sigma F_i) \in \text{PolyWit} \).

Note that polywits are sets of subsets of \( [k] \). In the second clause above, we used the composition \( (p_1, \ldots, p_k) \cdot (\sigma F_i) \), which is defined as \( \bigcup_{j \in \mathcal{I}(F_i)} \bigcup_{l \in [j]} \{l \} \bigcup_{l \in [p_k]} \{l \} \). This composition captures the computation of witnesses for composed BNFs, namely, for the composition of \( F_i \) with the BNFs corresponding to the polywits \( p_1, \ldots, p_k \).

We fix a set of tokens, \( \text{Tok} = \{t_i \mid i \in [i] \} \), to represent symbolically the \( T_i \)’s. We define the set-grammar \( \text{Gr} = (\text{Term}, \text{NTerm}, \text{Prod}) \) as follows. Its terminals \( \text{Term} \) are \( [k] \), i.e., one number \( k \in [k] \) for each type variable \( \alpha_k \). The nonterminals \( \text{NTerm} \) are either polywits or have the form \((p_1, \ldots, p_k)I_i \), where each \( p_k \) is a polywit and \( i \in [i] \). There are two types of productions:

1. \( p \mapsto I \), where \( p \in \text{PolyWit} \) and \( I \in p \)
2. \( \mathcal{I}t_i \mapsto \Gamma_j \) where \( i \in [i] \), \( J \in \mathcal{I}(G_i) \) and \( \Gamma_j = \{p_k \mid k \in J \cap [k] \} \cup \{\{\mathcal{I} \cdot \mathcal{I}(F_{j_1}) \cdot \mathcal{I}(F_{j_2}) \cdot \mathcal{I}(F_{j_3}) \} \mid k + j \in J \} \)

The first type of production selects witnesses from polywits. The second type mirrors the recursion in the definition of the \( T_i \)’s by following the destructors and selecting the terminals and nonterminals according to the witnesses of \( G_i \).

Let \( \text{Lang}_i(\text{Gr}) \) be the language, i.e., set of subsets of \( [k] \), generated by \( \text{Gr} \) starting from the nonterminal \( \{\{\alpha_1\}, \ldots, \{\alpha_k\}\} \) (the token for \( T_i \) applied to the trivial polywits for its argument types). Similarly, let \( \text{Lang}_{\text{PolyWit}}(\text{Gr}) \) be the language cogenerated by \( \text{Gr} \)—allowing infinite chains of productions, i.e., allowing infinite derivation trees—again, starting from \( \{\{\alpha_1\}, \ldots, \{\alpha_k\}\} \).

**Theorem 6:**

1.1) the definition is valid in HOL (i.e., the specified types are nonempty) if and only if \( \text{Lang}_i(\text{Gr}) \neq \emptyset \);
1.2) \( \text{Lang}_i(\text{Gr}) \) is a perfect set of witnesses for \( T_i \).
2) If we interpret the definition as specifying mutual codatatypes, then:

1.1) the definition is always valid in HOL (and in fact \( \text{Lang}_{\text{PolyWit}}(\text{Gr}) \neq \emptyset \) always holds);
1.2) \( \text{Lang}_{\text{PolyWit}}(\text{Gr}) \) is a perfect set of witnesses for \( T_i \).

Note how the formulation of the theorem distinguishes the nonemptiness subproblem from the witness problem. This is because the \( T_i \)’s cannot be registered as types without knowing their nonemptiness, more precisely, without knowing the nonemptiness of their representing predicates \( \text{ok}_i \) from the raw types \( \text{raw}_i \). In the codatatype case, nonemptiness always holds thanks to the greatest fixpoint nature of the construction.

Since the \( \text{raw}_i \)’s are BNFs, we already have a perfect set of witnesses for them, but those usually will not satisfy \( \text{ok}_i \) (and even if they were, they may not give a perfect set for \( T_i \)). So to prove the theorem we adapt the notion of witness from types to predicates and show that the languages (co)generated by \( \text{Gr} \) offer perfect sets for \( \text{ok}_i, u \) for any shadow \( u \). We generalize from \( [\ldots] \) to arbitrary \( u \) because the shadow increases along applications of the \( \text{raw} \) destructors. Appendix B gives details.

For any finite set-grammar \( \text{Gr} \), the languages \( \text{Lang}_i(\text{Gr}) \) and \( \text{Lang}_{\text{PolyWit}}(\text{Gr}) \) are effectively computable by fixpoint iteration [13]. Moreover, in [13, Section 4.3] we show that iteration only
needs a number of steps equal to the number of nonterminals. However, for uniform datatypes this number is precisely that of mutual types, i, in our nonuniform case, it is the larger number $j \times k \times |\text{PolyWit}|$, where $|\text{PolyWit}| = O(2^k)$. Fortunately, in practice the declared types do not have many type variables, so the doubly exponential blowup in k is not problematic.

As an example, consider the following contrived definition of the nonuniform codatatype of $(\alpha_1,\alpha_2)$-alternating streams:

$$(\alpha_1, \alpha_2) \alter \equiv C \alpha_1 ((\alpha_2, \alpha_1) \alter) \mid D \alpha_2 ((\alpha_2, \alpha_1) \alter)$$

Thus, we have $i = 1$, $j = 1$, $k = 2$, $\sigma$ is the unique function from $[1]$ to $[1]$, $(\alpha_1, \alpha_2) F_{11} = \alpha_2$, $(\alpha_1, \alpha_2) F_{12} = \alpha_1$ and $(\alpha_1, \alpha_2, \alpha_3) G_1 = \alpha_1 \times \alpha_3 + \alpha_2 \times \alpha_3$. Thus, $\{\{2\}\}$, $\{\{1\}\}$ and $\{\{1,3\}, \{2,3\}\}$ are perfect sets of witnesses for $F_{11}$, $F_{21}$ and $G_1$, respectively. Figure 2 shows two infinite derivation trees from the initial nonterminal $(\{1\}, \{2\}) t_1$ in the grammar $G_r$ associated to this definition, where we write $\{1\}$ and $\{2\}$ for the polywits $\{\{1\}\}$ and $\{\{2\}\}$. The trees repeat the same pattern after reaching $(\{1\}, \{2\}) t_1$. In the left tree, the top production is $(\{1\}, \{2\}) t_1 \rightarrow (\{1\}, \{2\}) t_1$; this is a valid production of type 2, based on $G_1$’s witness $\{1,3\}$. By contrast, the tree’s other production of type 2, $(\{2\}, \{1\}) t_1 \rightarrow (\{1\}, \{2\}) t_1$, uses the other witness, $J = \{2,3\}$. The frontiers of the two trees are $\{1\}$ and $\{2\}$, respectively. In fact, $\{\{1\}\}$ forms a perfect set of witnesses for $\alter$.

Even though $\alter$ is always nonempty, since it is a codatatype, determining a perfect set of witnesses is important for maintaining a complete solution to the overall nonemptiness problem. If we used an imperfect set of witnesses such as $\{\{1,2\}\}$, we would reject valid datatypes such as $\alpha \text{fractal} = (\alpha, ((\alpha, \alpha) \alter) \text{fractal}) \alter$, where we must know that $\{\{1\}\}$ is a witness for $\alter$ to infer nonemptiness.

V. (Co)induction Principles

In a proof assistant, high-level abstractions must be matched by a reasoning apparatus. Next we discuss how reasoning principles for nonuniform (co)datatypes can be inferred in HOL. To avoid cluttering the ideas with too many technicalities, in this and the next section we discuss the restricted situation of a single (co)datatype $\alpha T$ defined as fixpoint isomorphism $\alpha T \simeq (\alpha, \alpha F T) G$ as in Section III-B but with $i = j = k = 1$. We will reuse all the infrastructure defined in Section III-B while omitting all indices except for when they are needed, e.g., for distinguishing between the two set operators for $G$.

(Co)induction involves reasoning about the elements of a (co)datatype and those of its (co)recursive components. BNFs allow us to capture components abstractly, in terms of the “set” operators. For example, for any element $r$ of the uniform (co)datatype $\alpha \text{raw}$, its components are the elements $r' \in \text{set}_G(\text{unRaw} r)$—because, in its fixpoint definition, $\text{raw}$ appears recursively as the second argument of $G$.

A. Induction

Induction for uniform datatypes can be smoothly expressed in HOL. For example, the induction principle for $\alpha \text{raw}$ is the following HOL theorem, which we will refer to as $\text{Ind}_{\text{raw}}$:

$$\forall \alpha. (\forall r : \alpha \text{raw}. (\forall r' \in \text{set}_G(\text{unRaw} r). Q r') \Rightarrow Q r) \Rightarrow (\forall r : \alpha \text{raw}. Q r)$$

It states that, to show that a predicate $Q$ holds for all $\alpha \text{raw}$, it suffices to show that $Q$ holds for any element $r$ given that $Q$ holds for the recursive components $\text{set}_G(\text{unRaw} r)$ of $r$.

As we remarked in Section I, a verbatim translation of $\text{Ind}_{\text{raw}}$ for $T$ would not be typable, since $Q$ would be a variable used with two different types. But even if we change $Q$ from a quantified variable to a polymorphic predicate $Q : \alpha \text{raw} \rightarrow \text{bool}$ and remove the outer $\forall$), the formula would be unsound, due to the cross-type nature of the $T$-components: Whereas $t$ has type $\alpha T$, its components $t'$ have type $\alpha F T$. For example, if $Q$ is vacuously false on the type $\alpha T$, we could use such an induction theorem (with $\alpha$ instantiated to $\text{nat}$) to wrongly infer that $Q$ is true on $\alpha T$.

On the other hand, for each polymorphic predicate $P : \alpha T \rightarrow \text{bool}$, we can hope to prove the following inference rule in HOL, where for clarity we make explicit the universal quantification over the type variable $\alpha$, occurring both in the assumption and the conclusion

$$\forall \alpha. \forall t : \alpha T. (\forall r' \in \text{set}_G(\text{unRaw} t). P t') \Rightarrow P t$$

Let us try to prove this rule sound. All we have at our disposal is the representation type $\alpha \text{raw}$ and its induction principle. So we should try to reduce $\text{Ind}_{\text{raw}}$ to $\text{Ind}_P$ along the embedding–projection pair $\text{Rep} : \alpha T \rightarrow \alpha \text{raw}$ and $\text{Abs} : \alpha \text{raw} \rightarrow \alpha T$, where the predicate $\text{ok} []$ describes the image of $\text{Rep}$.

We start by defining $Q$ to be $\text{Abs} \circ P$ and try to prove $\forall r. \text{ok} [] r \Rightarrow Q r$ using $\text{Ind}_{\text{raw}}$, hoping to be able to connect the hypothesis of $\text{Ind}_{\text{raw}}$ with that of $\text{Ind}_P$. We quickly encounter the following problem, depicted in Figure 3.\footnote{There and in forthcoming figures we replace all map$_G f g$ arrow annotations with arrows carrying two labels The types should make clear which function represents which argument of map$_G$. For uniformity, we put the first argument $f$ to the right of the arrow’s direction and the second argument $g$ to the left.}
corresponds to $t : a T$ (via the embedding–projection pair); then $T$-induction speaks about the $T$-components $r' : a F T$ of $t$, which do not correspond to the raw-components $r : a$ of $r$, but rather to elements $r'' : a F$ of the form $\downarrow r$. This mismatch is a consequence of our representation technique: To represent $T$‘s destructor using raw’s destructor we needed to apply the “correction” $\downarrow r$ for the nonuniformity. In order to cope with it, we appeal to the shape type $a sh T$, which is in the simplified setting essentially $a + a F + a F^2 + \ldots$, and thus includes all the types inhabited by $t$, its components, the components’ components and so on.

So we weaken our goal, trying to prove that $P$ holds not on all types $a T$, but only on types of the form $a sh T$—for $Q$, this means switching from $a sh T$ to $a sh raw$. As shown in Figure 4, now we have a way to travel from the type $a sh F$ back to the type $a sh$—namely, by applying Node to level the nonuniformity $F$ into the larger type $sh$. For this to work, we need $Q$ to reflect $\text{map}_\text{raw} \text{Node}$, i.e., have $Q(\text{map}_\text{raw} \text{Node} r'')$ imply $Q r''$.

Another issue is that $r'' = \text{map}_\text{raw} \text{Node} r''$ itself is not in the image of $\text{Rep}$: $r''$ has shadow $[1]$ instead of the required $[]$. We must generalize our goal to arbitrary shadows, i.e., to $\forall u. \forall u r \Rightarrow Q' u r$, for a suitable predicate $Q': \Delta \rightarrow a \text{sh raw} \rightarrow \text{bool}$ that extends $Q$ in that $Q'[\downarrow []] = Q$. To this end, we define $\downarrow: \Delta \rightarrow a \text{sh raw} \rightarrow a \text{sh raw}$, an operator that generalizes the trip from $r'$ to $r''$ to $\text{map}_\text{raw} \text{Node}$ described above to an arbitrary shadow $u$, and $\downarrow$, the cumulative iteration of $\downarrow$: $\downarrow u r = \text{map}_\text{raw} \text{Node} (\downarrow u r)$ $\downarrow u r = s$ $\downarrow (u + 1) r = \downarrow u (\downarrow u r)$

To see the intuition of these operators, we can regard the elements of both $\beta$ and $\beta$ as trees whose leaves are elements of $\beta$ and whose nodes branch according to $F$. Then $\downarrow$ traverses elements of $a \text{sh}$ until it reaches their innermost nodes (with only leaves, i.e., elements of $a$, as subtrees), and then immerses them as top nodes in the inner shape layer. The additional shadow argument $u$ is needed in order to identify when an innermost tree has been reached (since we only count on $\downarrow u r$ being well-behaved if $\downarrow u r$ holds).

The $sh$ counterparts of the above, $\downarrow: \Delta \rightarrow a \text{sh} \rightarrow a \text{sh}$ and $\downarrow$ are defined similarly (using $\text{map}_\text{sh}$ instead of $\text{map}_\text{raw}$). The key property of the “immerse” family of operators is that they commute with raw’s destructor in the following sense.

**Lemma 7:** The left subdiagram in Figure 5 is commutative.

Now, taking $Q' u r$ to be $Q(\downarrow u r)$ does the job. Namely, $Q'$ can be proved by raw-induction on $r$, since it achieves the desired correspondence between the raw-components and the $T$-components, namely, between the leftmost and rightmost edges of Figure 5’s diagram. That the correspondence works is ensured by the diagram’s commutativity, as a composition of two commutative subdiagrams: the left by the above lemma and the right by the definition of $\text{dtor}$.

Thus, assuming the hypothesis of $\text{Ind}_P$, we have proved $\forall a. \forall t_1, t_2 : a T. P t_1 t_2 \Rightarrow \text{rel}_G(\downarrow) P (\text{map}_T t_1) (\text{map}_T t_2)$ $\forall a. \forall t : a T. P t_1 t_2 \Rightarrow t_1 = t_2$

For this rule to be sound, $P : a T \rightarrow a T \rightarrow \text{bool}$ should again interact well with injective functions, however this time in the opposite direction. We say that $P$ is injective-parametric (IMP), if $P t_1 t_2$ implies $P (\text{map}_T f t_1) (\text{map}_T g t_2)$ for all $t_1, t_2 : a T$ and injective functions $f, g : a \rightarrow \beta$. (Both $\text{Leaf}$ and $\text{Node}$ are injective.) In conclusion, we have obtained:

**Theorem 8:** If $P$ is IMP, then $\text{Ind}_P$ is derivable in HOL.

IMP is related but significantly weaker than (arbitrary) parametricity, which for $P$ would mean $P t \Leftrightarrow P (\text{map}_T f t)$ for all $t$ and arbitrary functions $f : a \rightarrow \beta$. (Both $\text{Leaf}$ and $\text{Node}$ are injective.) In conclusion, we have obtained:

**Theorem 8:** If $P$ is IMP, then $\text{Ind}_P$ is derivable in HOL.

B. Coinduction

We have designed the above infrastructure, consisting of the “immerse” operators, to work equally well for the codatatype as it does for the datatype. Namely, when $a T$ is a codatatype, these operators are defined in the same way and can be used to derive the soundness of a nonuniform coinduction rule under similar assumptions to the induction case (from the corresponding uniform coinduction on the raw type):

$$\forall a. \forall t_1, t_2 : a T. P t_1 t_2 \Rightarrow \text{rel}_G(\downarrow) P (\text{map}_T t_1) (\text{map}_T t_2)$$

**Coind**

For this rule to be sound, $P : a T \rightarrow a T \rightarrow \text{bool}$ should again interact well with injective functions, however this time in the opposite direction. We say that $P$ is injective-parametric (IMP), if $P t_1 t_2$ implies $P (\text{map}_T f t_1) (\text{map}_T g t_2)$ for all $t_1, t_2 : a T$ and injective functions $f, g : a \rightarrow \beta$. (Both $\text{Leaf}$ and $\text{Node}$ are injective.) In conclusion, we have obtained:

**Theorem 9:** If $P$ is IMP, then $\text{Coind}_P$ is derivable in HOL.

Unlike IMP, IAP disallows the usage of the universal quantifier in $P$, while it allows the existential quantifier. This is a quite desirable symmetry: Induction requires the universal quantifier to perform generalization over non-inductive parameters. For coinduction, the existential quantifier takes this role.

VI. (Co)Recursion Principles

For nonuniform (co)datatypes to be practically useful, there must exist some infrastructure supporting (co)recursive function definitions. We start with datatypes and consider the following simple recursive function on powerlists:

- **split** : $(\alpha \times \beta) \rightarrow \alpha \text{plist} \times \beta \text{plist}$
- **split Nil** = $(\text{Nil, Nil})$
- **split (Cons ab xs)** = let $(\text{as, bs}) = \text{split} (\text{map}_\text{plist} \text{swap} \text{xs})$ in $(\text{Cons (fst ab) as}, \text{Cons (snd ab) bs})$
Here, the pattern matched variable \( x_s \) has type \(((\alpha \times \beta) \times (\alpha \times \beta))\) `plist` and the auxiliary `swap` function is defined as `swap ((a_1, b_1), (a_2, b_2)) = ((a_1, a_2), (b_1, b_2))`. The function split uses polymorphic recursion: its type on the right hand side of the specification is different from the one on the left hand side. More precisely, the recursive call is applied to an argument of type \(((\alpha \times \alpha) \times (\beta \times \beta))\) `plist`. None of the existing tools for defining recursive functions in higher-order logic can handle polymorphic recursion—the gap we are about to fill.

Also note that split is not primitively recursive in the standard sense: the recursive call is applied to a modified pattern matched argument `map` `swap` `xs`. However, the modification happens through the `map` function, which does not change the length of `xs`. Hence, such generalized primitively recursive specifications, which modify the arguments of the recursive call merely through a map function, are safely terminating.

### A. Generalized Primitive Recursion

Following the foundational approach, primitively recursive specifications in HOL are reduced to nonrecursive definitions using a recursion combinator [10]. The equally expressive but slightly less convenient primitively iterative specifications can be reduced too, using a simpler fold combinator. For a uniform datatype \( \alpha T = \text{Ctor} ((\alpha, \alpha T) G) \) (e.g., \( \alpha T = \alpha \text{list} \) with \( (\alpha, \beta) G = \text{unit} + \alpha \times \beta \)) the fold combinator has type

\[
((\alpha, \beta) G \rightarrow \beta) \rightarrow \alpha T \rightarrow \beta
\]

A function \( f = \text{fold} \, b \) for some fixed \( b : (\alpha, \beta) G \rightarrow \beta \), satisfies the characteristic recursive equation \( f \circ (\text{Ctor} \, g) = b \circ (\text{map} \, G \, f \, g) \). We call \( b \) the blueprint of \( f \). Note that \( b \) describes how to combine the results of the recursive calls into a new result of type \( \beta \). The recursion combinator’s blueprint, of type \( (\alpha, \alpha T \times \beta) G \rightarrow \beta \), generalizes fold’s blueprint by providing access to the original \( \alpha T \) values, in addition to the results of the recursive calls. Although our ideas support recursion, we focus on iteration to simplify the presentation.

For a nonuniform datatype \( \alpha T = (\alpha, \alpha F T) G \) (as before for simplicity we consider the setting \( i = j = k = 1 \)), the natural generalization of fold would be a combinator of type

\[
\forall Y. (\forall \alpha. (\alpha, \alpha F Y) G \rightarrow \alpha Y) \rightarrow \beta T \rightarrow \beta Y
\]

where the universally quantified type constructor \( Y \) captures the positions where \( \alpha \) have to be replaced by \( \alpha F \), since the recursive calls will be applied to a term of type \( \alpha F T \). The explicit universal quantification over \( \alpha \) indicates that the blueprint needs to be truly polymorphic in \( \alpha \).

Bird and Paterson [9] observe that the above combinator is not practical, since the primitive iteration scheme it provides is very restrictive: it forces the type argument \( \beta \) of \( T \) to be fully polymorphic. In fact, neither the split function, nor a simple summation of a powerlist storing natural numbers can be expressed using that fold. To overcome the limitation, they propose a generalized fold of type

\[
\forall X Y. (\forall \alpha. (\alpha X, \alpha F Y) G \rightarrow \alpha Y) \rightarrow (\forall \alpha. \alpha X F \rightarrow \alpha F X) \rightarrow \beta X T \rightarrow \beta Y
\]

where the second argument enables recursive functions of a more refined type \( \beta X T \rightarrow \beta Y \) by providing a distributive law \( a : \forall \alpha. \alpha X F \rightarrow \alpha F X \) which we call the (argument) `swapper`. Bird and Paterson require \( X \) and \( Y \) to be functors and the two function arguments \( b \) and \( a \) to fold to be natural transformations. The function \( f = \text{fold} \, b \, a \) is then a natural transformation too and adheres to the characteristic equation

\[
f \circ (\text{Ctor} \, g) = b \circ (\text{map} \, G \, \text{id} \circ (f \circ \text{map}_F \, a) \circ g) \tag{2}
\]

It is straightforward to allow functors \( X \) and \( Y \) to be of arbitrary arity \( n \) instead of 1. The function split can then be defined by setting, \( n = 2 \), \((\alpha, \beta) G = \text{unit} + \alpha \times \beta, \alpha F = \alpha \times \alpha, (\alpha, \beta) X = \alpha \times \beta, (\alpha, \beta) Y = \alpha \text{plist} \times \beta \text{plist}, a = \text{swap}, \) and

\[
\begin{align*}
  b \, (\text{Inl} \, ()) &= (\text{Nil}, \text{Nil}) \\
  b \, (\text{Inr} \, (ab, abs)) &= \text{let} \, (as, bs) = abs \rightarrow in \, (\text{Cons} \, (\text{fst} \, ab) \, as, \text{Cons} \, (\text{snd} \, ab) \, bs)
\end{align*}
\]

where `Inl` and `Inr` are the standard embeddings of `+`. For simplicity, the rest of the section assumes \( n = 1 \).

We propose an even more flexible fold combinator which replaces the functor \( F \), which is fixed in the nonuniform datatype specification and shows up in the recursive calls, with another arbitrary functor \( V \) of the same arity as \( F \) (here, 1):

\[
\forall X Y. (\forall \alpha. (\alpha X, \alpha V Y) G \rightarrow \alpha Y) \rightarrow (\forall \alpha. \alpha X F \rightarrow \alpha V X) \rightarrow \beta X T \rightarrow \beta Y
\]

This allows the recursive calls to return a type \( \alpha V Y \) instead of the fixed \( \alpha F Y \). The combinator satisfies the same characteristic equation (2) (with the more general types).

All those expressive combinators for nonuniform types share one problem: in higher-order logic neither type constructor quantification nor type variable quantification that happens not at the outermost level is possible. Thus, it is impossible to define the fold constants for nonuniform datatypes in HOL.

Instead, we follow a similar route as for induction. We devise a recursion procedure that takes (here: unary) BNFs
V, X, Y,\textsuperscript{2} a blueprint \( b : (\alpha X, \alpha V Y) \rightarrow \alpha Y \) and a swapper \( a : \alpha X F \rightarrow \alpha V X \) as input and produces a function \( f : \alpha X T \rightarrow \alpha Y \) satisfying equation (2) as output.

Internally, the procedure defines a recursive function using \( b \) and \( a \) on the \texttt{raw} type and lifts it to the nonuniform type. To perform such a lifting for induction, the inductive property \( P \) had to be a polymorphic IAP term. For recursion, we require both \( b \) and \( a \) to be polymorphic injective-parametric terms, i.e., parametric only for relations that are graphs of injective function. This is a weaker assumption than Bird and Paterson’s naturality assumption (e.g., \( \text{map}_F (\text{map}_X f) = \text{map}_X (\text{map}_Y f) \circ a \) for \( a \)). On (bounded) natural functors injective-parametricity implies the weak naturality assumption that demands for the above equation to hold only for injective functions \( f \). Consequently, \( f \) will also only be a natural function for injective functions. However, our construction is closed: If both \( b \) and \( a \) are fully parametric in some type parameters, \( f \) is fully parametric in those as well.

The definition of \( f \) proceeds in four steps. First, we define a shape type \( \text{sh}_V \) for \( V \) analogously to \( \text{sh} \) for \( F \), including the constructors \( \text{Leaf}_V \), \( \text{Node}_V \), their inverses \( \text{unLeaf}_V \), \( \text{unNode}_V \), and the functions \( \text{map}_X, \text{map}_Y \), and \( \text{map}_2 \). Second, we lift \( a \) to shapes \( \alpha \) : \( \Delta \rightarrow \alpha X s \rightarrow \alpha \text{sh}_V X \) by recursion on the shadow:

\[
\pi \Gamma \downarrow = \text{map}_X \text{Leaf}_V \circ \text{unLeaf} \\
\pi (1 \uparrow u) = \text{map}_X \text{Node}_V \circ a \circ \text{map}_F (\pi u) \circ \text{unNode}.
\]

Third, we define a \texttt{raw} version of our function \( f_{\text{raw}} : \Delta \rightarrow \alpha X T \rightarrow \alpha \text{sh}_V Y \) by primitive recursion:

\[
f_{\text{raw}} u (\text{Raw} g) = b (\text{map}_G (\pi u) (\text{map}_2 \text{unNode}_V \circ f_{\text{raw}} (1 \uparrow u)) g)
\]

The generalization to \( \text{sh}_V \) in the return type of \( f_{\text{raw}} \) is similar to the generalization performed for induction. Finally, we define the function \( f \) as \( f = \text{unLeaf}_V \circ f_{\text{raw}} \circ \text{Rep} \).

Figure 6 justifies why the above definitions make sense by proving equation (2). When reading the diagram, some of the arrows labeled by injective functions, such a \( \text{Leaf}_V \) and \( \text{Node}_V \) (possibly under further maps), need to be inverted for the diagram to make sense. Elements of the two highlighted types have shadow \([1]\). All other elements of types \( \text{sh}, \text{sh}_V \), and \( \text{raw} \) occurring in the diagram have shadow \([1]\).

Equation (2) is the outermost pentagon, which is filled by commutative diagrams starting by unfolding the definitions of \( f \) (twice), \( \text{Ctor} \), and \( \text{map}_F \) as well as the recursive specification of \( f_{\text{raw}} \). The quadrilateral \( \text{5} \) follows from the naturality for injective functions (\( \text{Leaf}_V \)) of \( b \) and \( \text{6} \) from the recursive specification of \([1]\). The remaining commutative pentagon \( \text{7} \) crucially relates \( f_{\text{raw}} \) and \( \uparrow \) (similar to Lemma 5 for \( \text{map}_2 \)). The proof of that fact follows by induction. Therefore, the property used in \( \text{7} \) for shadow \([1]\) needs to be generalized to an arbitrary \( u \) and requires an auxiliary fact about \( a \) and \( \text{map}_2 \) alongside with the facts that \( \pi \) and \( f_{\text{raw}} \) preserve \( \text{sh} \) \text{sh} and \( ok \). The proofs rely on the injective-parametricity of \( a \) and \( b \).

Lemma 10:

\[\text{3} \text{Strictly speaking, the boundedness assumption are not needed for } X \text{ and } Y, \text{ which implies that } \alpha \text{ set is permitted to occur in those type expressions.}\]
equation $f(Ctor\ g) = b(\text{map}_G\ \text{id}(f \circ \text{map}_T\ a)\ g)$ holds for a fixed injective-parametric blueprint $b$ and swapper $a$.

The code for step 2 constructs the low-level types, terms, and lemma statements presented in Sections III to VI and proves the lemmas using specialized tactics—ML programs that generalize the proofs from the formalization. In principle, the tactics should always succeed, but it is necessary to execute them to obtain the highest level of trustworthiness. Assuming Isabelle’s inference kernel is correct, bugs in the execute them to obtain the highest level of trustworthiness. Assuming Isabelle’s inference kernel is correct, bugs in the new commands might lead to run-time failures but never to logical inconsistencies. For step 3, we were able to generalize and reuse the infrastructure for uniform types that performs the same lifting from low to high level [10, Sections 3–6].

Step 4 takes the form of six main commands available to the users and making definitions and reasoning about nonuninorm types nearly as convenient as for uniform types.

The nonuniform \_co\_datatypenonuniform \_co\_induct commands can be used to define nonuniform types. For example, the following definition introduces a type of $\lambda$-terms over variables drawn from $\alpha$, with De Bruijn notation for bound variables [8]:

\[
\text{nonuniform\_datatype\ } \alpha\ tm = \\
\quad \text{Var}\ \alpha \mid \text{App}\ (\alpha\ tm)\ (\alpha\ tm) \mid \text{Lam}\ ((\text{unit} + \alpha)\ tm)
\]

Entering a $\lambda$-abstraction (Lam) creates a new variable, which is accommodated by the extended type $\text{unit} + \alpha$ consisting of the values $\text{Inl}(\cdot)$ (the new variable) and $\text{Inr} x$ for all $x : \alpha$.

The command performs the type construction and computes a nonempty witness. Then it defines the constructors Var, App, Lam and corresponding destructors and derives characteristic theorems about the constructors, the destructors, and the BNF constants $\text{map}_{\text{tm}}, \text{pred}_{\text{tm}}, \text{rel}_{\text{tm}}, \text{set}_{\text{tm}}$.

The nonuniform\_prim\_\_co\_recursive \_induct commands allow users to define primitively (co)recursive functions, by specifying their (co)recursive equations.\(^3\) For example, the following definition introduces a function join that “flattens” a term whose variables are themselves terms:

\[
\text{nonuniform\_prim\_recursive join} : \alpha\ tm\ \alpha\ tm \rightarrow \alpha\ tm
\]

\[
\text{join (Var } x \text{) } = x
\]

\(^3\)The implementation of these two commands is incomplete at the time of this writing. We do not foresee any difficulties beyond those which we met for the other commands and expect to finish the implementation in the weeks following the submission deadline. Our archive [11] will be updated.

The command extracts blueprints and swappers from the user-specified equations and emits parametricity proof obligations that must be discharged by the user. In the example, the swapper is the $\lambda$-expression of type $\text{unit} + \alpha\ tm \rightarrow (\text{unit} + \alpha)\ tm$ that is passed to the outer $\text{map}_{\text{tm}}$. Once the proofs are complete, the command derives a low-level characteristic theorem about the defined function. Then it derives the equations specified by the user from this theorem.

One of the most basic operations on $\lambda$-terms is substitution: $\text{subst} : (\alpha \rightarrow \beta\ tm) \rightarrow \alpha\ tm \rightarrow \beta\ tm$. Due to the limitation that arguments to recursive functions must be BNFs, we cannot define higher-order functions like $\text{subst}$ that depend on a type variable that changes in the recursive calls. But we can define $\text{subst}$ as a composition: $\text{subst } \sigma = \text{join} \circ \text{map}_{\text{tm}}\ \sigma$.

The nonuniform\_co\_induct commands can be used to prove a lemma (or a set of lemmas for mutual definitions) by (co)induction. For example, the command

\[
\text{nonuniform\_induction s in subst\_subst}:
\]

\[
\text{subst } \tau (\text{subst } \sigma\ s) = \text{subst } (\text{subst } \tau \circ \sigma)\ s
\]

emits proof obligations for parametricity and the three cases of the induction on $s$ in $\alpha$. Often, the parametricity proofs can be delegated to Isabelle’s Transfer tool [26]. Once the obligations are discharged, the stated property is derived and stored under the specified name (subst\_subst). For the technical reason explained in Section V, the derivation can be performed only by an Isabelle command, not by a proof method as is done for uniform (co)datatypes [10]. The main advantage of proof methods is that they can be invoked on an arbitrary proof goal in the middle of a proof.

We conclude with a codatatype example: We prove two alternative definitions of the constant powerstream equivalent. All required proofs are fully automatic after specifying the trivial bisimulation relation $R l r \equiv \exists x x s. l = \text{const } x \land r = \text{map}_{\text{pstream}} (\lambda_\leftarrow\ x)\ xs$ in the coinduction proof.

\[
\text{nonuniform\_codatatypenonuniform\_prim\_\_co\_\_recursive const : } \alpha \rightarrow \alpha\ pstream
\]

\[
\text{const } x = \text{Cons } x (\text{const } (x, x))
\]
nonuniform_coinduct R in const_alt:
  const x = mapistream (λₚ : α x) xs

VIII. DISCUSSION AND RELATED WORK

a) Inspiration: For representation, we generalized Okasaki’s construction [37] to arbitrary datatypes. Nordhoff et al. [36] have used this construction partially (defining the sh and raw types but without introducing a new nonuniform type) in their Isabelle/HOL formalization of 2-3 finger trees. The corresponding reduction of nonuniform to uniform codatatypes appears not to have been studied in the literature.

For recursion, we refined Bird and Paterson’s generalized fold combinators [9] in several ways, including weakening the parametricity/naturality condition and enabling non-functor target domains. In turn, Bird and Paterson had improved on the standard sheaf-functor approach from category theory [29].

Our (co)induction principles take advantage of the BNF structure including set operators and relators, and form a lighter alternative to fibration-based approaches [17], [22] for the category of sets and functions.

b) Comparison with Other Proof Assistants: Our work shows that nonuniform (co)datatypes and the associated polymorphic (co)recursion [21], [34] can be supported in the minimalistic rank-one polymorphic framework of HOL, and therefore made available in HOL-based provers, which cover about half of the theorem proving community.

The dependent type theory (DTT) camp, represented by theorem provers such as Agda, Coq, Matita, and Lean, has sophisticated type systems built into their mechanized logic, including native nonuniform datatypes. Several case studies in these provers exploit nonuniformity [4], [15], [25], [35], [42].

Compared with the DTT systems, our support for nonuniform types in HOL has some limitations. First, dependent families of (nonuniform) types cannot be expressed in HOL. Second, Agda supports self-nested (co)datatypes. For example, a definition such as a bush = BNil | BCons a (a bush bush) is beyond scope of our results, since the recursive occurrence of bush is nested in itself. Third, our (co)induction principles have some restrictions concerning (a weak form of) parametricity. The reason is that we cannot perform well-founded induction across types in HOL. While practical predicates about functional programs obey them, the restrictions are not necessary semantically. Appendix D presents an axiomatic extension that allows HOL to “see” cross-type. However, adding axioms, no matter how provably correct they may be, goes against the main tenets of HOL.

Our approach has some advantages as well, stemming from its category-theory-awareness. First, arbitrary parameter types, not just (co)datatypes, can be plugged in the specifications for nonuniform (co)datatypes, either inside or outside of the recursive occurrences in the specification. For example, the type stree from Section I is possible because the type a fset, of finite sets is a BNF. This is not possible with DTT, where datatypes are restricted to a predefined grammar.

Second, since the foundational approach compels us to maintain the functorial structure to justify fixpoint definitions, users can enjoy default map functions and relators, as well as some polytypic properties either out of the box or within the immediate reach. Thus, our nonuniform recursion principle delivers parametric functions, i.e., natural transformations. Moreover, the fusion laws [9] (Appendix C), known to be important in reasoning about functional programs, rely heavily on functoriality and naturality, and they are immediate in our framework. In contrast, in DTT, very little structure is available for nonuniform datatypes after definition. In particular, map functions and relators are missing and can be difficult to add.

c) Other Work: The pioneering work of Bird and his collaborators on nonuniform datatypes [7], [8], [9] has been extended into several directions, including structures for efficient functional programming [23], [24], [30], datatypes with references [16], as well as work directly relevant for DTT proof assistants: reduction to W-types and container types [1], typed term rewriting frameworks for total programming [2], [3], [31], intensional-DTT induction [32]. Our current contribution was concerned with bootstrapping nonuniform datatypes in HOL on a sound and compositional basis. Only time will tell if Isabelle/HOL users, or more generally the HOL community of users and researchers, will embrace nonuniform datatypes and their applications to a similar scale as in advanced programming languages and type theories.

Acknowledgment: We thank Peter Lammich for pointing us to his work on finger trees and for formalizing Okasaki’s construction for powerlists and Johannes Hölzl for explaining us Lean’s support for nonuniform datatypes.

REFERENCES

A. (Weak) Parametricity and Naturality

\[ c : \mathcal{P} \rightarrow \mathcal{R} \text{ is parametric rel}_F \mathcal{R} c c \text{ for all } \mathcal{R} \]
\[ f : \mathcal{P} \rightarrow \mathcal{G} \text{ is parametric (P) rel}_F \mathcal{R} f f \text{ for all } \mathcal{R} \]

Function space relation \( R \Rightarrow S \) defined as \( \lambda f. (\forall a b. R a b = S (f a) (g b)) \). In other words two functions are related by \( R \Rightarrow S \) if for all R-related inputs, the outputs are S-related.

\[ f : \mathcal{P} \rightarrow \mathcal{G} \text{ injective-parametric (IP) rel}_F \mathcal{R} f f \text{ for all } \mathcal{R} \] that are graphs of injective functions, i.e., left-total, singlevalued relations.

\[ f \text{ natural transformation (NAT) } f \circ \text{map}_F \mathcal{R} = \text{map}_G \mathcal{R} \circ f \text{ for all } \mathcal{G} \]

\[ \text{predicate } P : \mathcal{P} \rightarrow \text{bool} \text{ (injective)-parametric (rel}_F \mathcal{R} \Rightarrow \Gamma) P \text{ P for all } \mathcal{R} \] (that are graphs of injective functions).

\[ \text{predicate } P : \mathcal{P} \rightarrow \text{bool} \text{ injective-monotone parametric (IMP) rel}_F \mathcal{R} \Rightarrow (\Rightarrow) P \text{ P for all } \mathcal{R} \] that are graphs of injective functions.

\[ \text{predicate } P : \mathcal{P} \rightarrow \text{bool} \text{ injective-antitone parametric (IAP) } \text{ when } (\Rightarrow) \text{ is replaced by } (\Leftarrow) \]

\[ P = \text{ NAT} \]
\[ \Rightarrow IP \]
\[ \Rightarrow IMP \]
\[ \Rightarrow IAP \]
\[ \Rightarrow \text{IMP and IAP } \Rightarrow \text{IP} \]

B. Proof Sketch of Theorem 6

Since the types \( T_i \) do not exist yet, we need to work on the representation types \( \text{raw}_i \), and prove a property about the predicates \( \text{ok}_i \) that represent the to-be-defined types \( T_i \). For this, we introduce a refinement of the notion of witness: Given a shadow \( u : \Delta \), a set \( I \subseteq \{i\} \) is an \textit{u-witness} for \( \text{raw}_i \) if, for all sets \( \mathcal{A}, \forall k \in I. \mathcal{A} \neq \emptyset \implies \exists r \in \text{raw}_i \mathcal{A}. \text{ok}_k u r \). We prove:

(3) \( \text{Lang}_i(\mathcal{G}) \) (or \( \text{Lang}_\infty(\mathcal{G}) \)) is a perfect sets of \textit{u-witnesses} for the (co)datatype \( \text{raw}_i \).

To prove (3), we are looking for a connection between the grammar \( \mathcal{G} \) and \( \text{raw}_i \)'s (co)recursivity specification, in the direction of the destructor

\[ \pi \text{raw}_i \mapsto_{\text{unraw}} \pi \text{sh}, \pi \text{raw}_{\sigma(1)}, \cdots, \pi \text{raw}_{\sigma(j)} \]

By the definition of \( \text{ok}_i \), if \( r : \pi \text{raw}_i \) is such that \( \text{ok}_k u r \), then \( \text{dtor}_r \) has, for each \( j \in [|i|] \), its \( j \)-components \( \sigma_j : \pi \text{raw}_{\sigma(j)} \) satisfying \( \text{ok}_k (j u) r' \). As discussed in Section III-B, the shadow increment between \( u \) and \( j u \) encodes the application of the components \( F_j = (F_{j1}, \ldots, F_{jk}) \), reflected in the grammar’s productions of type 2. This suggests a recursive translation of shadows into polymorphs:

\[ \gamma_k [\{i\}] \]
\[ \gamma_k (j u) = (\gamma_1 u, \ldots, \gamma_n u) \cdot \sigma (F_j) \]

Now we can formulate a generalization of (3), taking into account arbitrary shadows, not just \( \{\} \). For each nonterminal \( x \), we write \( \text{Lang}_i(\mathcal{G}) \) (or \( \text{Lang}_\infty(\mathcal{G}) \)) for the language (co)generated by \( x \).

(4) For all \( u : \Delta \), \( \text{Lang}_{[\gamma_1 u, \ldots, \gamma_n u]}(\mathcal{G}) \) (or \( \text{Lang}_{\infty(\gamma_1 u, \ldots, \gamma_n u)}(\mathcal{G}) \)) is a perfect set of \textit{u-witnesses} for the (co)datatype \( \text{raw}_i \).
Finally, (4) can be proved using standard (uniform) (co)induction and (co)recursion, moving back and forth between $Gr$-derivation trees and the raw $w_i$’s.

For datatypes: That every $I \in \text{Lang}_{(\gamma_k u_0, \ldots, \gamma_u v)}$ (Gr) is an $u$-witnesses follows by structural induction on its derivation tree in $Gr$. Conversely, that for every $u$-witness $J$, we have $I \subseteq J$ for some $I \in \text{Lang}_{(\gamma_k u_0, \ldots, \gamma_u v)}$ (Gr) follows by induction on the definition of $ok_u$.

For codatatypes: To prove that every $I_0 \in \text{Lang}_{\omega_{\text{nu}}}$ is an $u_0$-witness, we let $\overline{J}$ be such that $\forall i \in I_0, A_i \neq \emptyset$. With $u_0, I_0$ and $\overline{J}$ fixed, let $Tr$ be a (possibly infinite) derivation tree of $(\gamma_k u_0, \ldots, \gamma_u v)$ in Gr—thus having $I_0$ as the set of terminals on its frontier. Let $\Delta_{n_i}$ consists of all shadows having $u_0$ as a prefix and such that the nonterminal $(\gamma u) t_j$ occurs in the tree $Tr$, where we write $\gamma u$ for $(\gamma_1 u, \ldots, \gamma_k u)$.

Mutually corecursively (by primitive raw-corecursion), we define the functions $w_i : \Delta_{n_i} \to \overline{\sigma} \text{raw}_{w_i}$, by

$$w_i u = \text{map}_{G_i} \overline{\sigma} \text{raw}_{w_i}$$

where the element $g_u \in (\overline{\sigma} \text{sh}, \Delta_{n_0, r_0(1)}, \ldots, \Delta_{n_0, r_0(j)}) G_i$ is defined as follows: Let $(\gamma u) t_j \to \Gamma_j$ be the (type 2) production in $Tr$ corresponding to the terminal $(\gamma u) t_i$.

- From the definition of $\Gamma_j$ it follows that, for each $j$ such that $k+j \in J$, the nonterminal $(\gamma_j u, \ldots, \gamma_u v) t_j$, i.e., $(\gamma_j (\gamma u) \ldots, \gamma_u (\gamma u) t_j)$, is also in $Tr$; hence $\gamma u \to \Gamma_j$.

- Also from the definition of $\Gamma_j$ it follows that, for each $k \in [k] \cap J$, the nonterminal $\gamma_k u$ is also in $Tr$. Since only type 1 productions are applicable to polymorphic nonterminals, let $\gamma_k u \to \Gamma_k$ be the production from $Tr$ applied to $\gamma_k u$. Then $I$ is included in $Tr$’s frontier, i.e., $I \subseteq I_0$. Then, picking some elements $a_i \in A_i$ for $i \in I$, we can define shapes $s_k \in sh_k \overline{J}$ for each $k \in [k]$ that are full trees, i.e., that $\overline{\sigma} k$, $s_k$ holds.

Thus, we have constructed the elements $\gamma u \in \Delta_{n_0, r_0(j)}$ for each $j$ such that $k+j \in J$ and $s_j \in sh_j (\overline{J})$ (in particular, $s_j : \overline{\sigma} sh_j$) for each $k \in [k]$. Since $J$ is a witness for $G_i$, we obtain our desired element $g_u \in (\overline{\sigma} sh, \Delta_{n_0, r_0(1)}, \ldots, \Delta_{n_0, r_0(j)}) G_i$, which concludes the definition of the $w_i$’s. Because of our choices in the definition, it is now routine to prove:

- by rule coinduction on the definition of the $ok_i$’s, that $\overline{\sigma} k u (w_i u)$ holds;
- by rule induction on the definition of the set operators for raw, that $\overline{\sigma} \text{raw}_{w,k} (w_i u) \subseteq A_i$ holds, which means that $w_i u \in \overline{\sigma} \text{raw}_{w_i} \overline{A}$ holds.

This concludes the proof that $I_0$ is a witness.

Conversely, to prove that for every $u$-witness $J$, we have $I \subseteq J$ for some $I \in \text{Lang}_{\omega_{\text{nu}}}$ (Gr), we construct a (possibly infinite) derivation tree whose frontier includes $J$. The construction proceeds corecursively, by extracting each time the next production to be applied from $J$ and the $G_i$’s witnesses.

C. Fusion Laws

We fix a nonuniform datatype $\alpha \ T = (\alpha \, \alpha \, F \ T) \ G$. Let us write $\text{NURec}(X,Y,a,b)$ for the polymorphic function $f : \alpha X T \to \alpha Y$ defined by nonuniform recursion from a blueprint $b : (\alpha X, \alpha Y) G \to \alpha Y$ and a swapper $a : \alpha X F \to \alpha X V$, as in Section VI-A. (Of course, $\text{NURec}$ cannot be a HOL combinator—it is just a meta-level notation.)

Theorem 11: The following hold:

Fold Fusion: If $\kappa : \alpha Y \to \alpha Y'$ is such that $\kappa \circ b = b' \circ \text{map}_G \ k$, then $\kappa \circ \text{NURec}(X,Y,a,b) = \text{NURec}(X,Y',a,b')$.

Map Fusion: If $\kappa : \alpha X' \to \alpha X$ is such that $\kappa \circ a' = a \circ \kappa$, then $\text{NURec}(X,Y,a,b) \circ \text{map}_F \ k = \text{NURec}(X',Y,a',b)$.

Proof. TODO: By nonuniform induction; things are suitably parametric (equality of parametric / IAP functions); maybe draw diagrams, writing $f$ and $f'$ for the recursively defined functions.

The duals of the fusion laws hold when $\alpha T \equiv (\alpha \, \alpha \, F \ T) \ G$ is a uniform codatatype as in Section VI-B.

D. Cross-Type Induction Schema

As discussed in the paper, a main restriction of our work is induction for nonuniform types, where we require IA-parametricity of the predicate. Here, we show how a gentle, provably consistent axiomatic extension of HOL removes this restriction. The axiom does not refer to nonuniform datatypes, or the intricate construction leading to them, or even to BNFs. However, it is easy to see that the rule is safe: the change in type is impossible to justify. The problem is the change in type parametricity of the predicate. Here, we show how a gentle, uniform codatatype as in Section VI-B.

We fix the types $\alpha T$, $\alpha F$ and $M$ (with the notations $T$ and $F$ not connected to nonuniform datatypes).

Let $P : \alpha T \to \text{bool}$ be a polymorphic predicate, for which we want to prove $\forall \alpha. \forall t : \alpha T. P t \text{ a natural approach would be induction using a measure } \mu : \alpha T \to M \text{ which decreases w.r.t. a well-founded relation } r : M \text{ set } \to M \text{ set } \to \text{bool}. \text{ But what if the measure decreases by changing the type, as in } r (m (t') (m t)), \text{ where } t : \alpha T \text{ and } t' : \alpha F T ? \text{ This is still acceptable, since well-foundedness should still operate across the types } \alpha F^n T. \text{ Formally, we would like to have the following rule, where } w f r \text{ states that } r \text{ is well-founded.}

$$\forall \alpha. \forall t : \alpha T. \ w f r \land (\forall t' : \alpha F T. r (m (t') (m t) ) \Rightarrow P t') \Rightarrow P t$$

WFInd_{T,F,r,m,P}

In HOL, this type of induction in HOL across varying types is impossible to justify. The problem is the change in type during the descent: the elements $t'$ smaller than $t$ (via $m$, w.r.t. $r$) do not dwell the same type as $t$, $\alpha T$, but a different type $(\alpha F) \ T$. And induction in HOL across varying types is impossible (unless, as we have seen, we require parametricity). However, it is easy to see that the rule is safe:

Theorem 12: The rule schema $\text{WFInd}$ is sound in the standard models of HOL [39] and in the ground models.
of Isabelle/HOL [27], hence is consistent with HOL and Isabelle/HOL.  

Proof. A standard model of HOL fixes a universe \( \mathcal{U} \) of sets with good closure properties (e.g., closed under function spaces) and interprets a type constructor such as \( T \) as a function on this universe, \( [T] : \mathcal{U} \to \mathcal{U} \), a type such as \( M \) as an element of the universe, \( [M] \in \mathcal{U} \), etc. Moreover, a polymorphic constant such as \( m : \alpha T \to M \) is interpreted as a \( \mathcal{U} \)-indexed family \( ([m]_A)_{A \in \mathcal{U}} \). Crucially, it interprets the function-space type constructor as the set of all functions between the interpretation of its arguments and the type \( \text{nat} \) as a countable set \([\text{nat}]\), which with \([0] \) and \([\text{Suc}] \) is isomorphic to the natural numbers. This means that the scheme WFInd can be justified inside \( \mathcal{U} \) as follows: Assuming its conclusion is false and repeatedly using its hypothesis, we obtain the infinite sequence \( (A_i) \) and \( (b_i) \) such that \( A_{i+1} = [F](A_i) \), \( b_i \in [T](A_i) \), \( [m](A_i)(b_i) = \text{True} \) and \( \lambda \alpha. [\text{Suc}] \). The derivation takes place by instantiating the parameters of WFInd\(_T,F,r,m,p\) using those of \( \text{Ind}_T^\alpha \). We take:

- \( T \) and \( F \) to be the nonuniform datatypes and its nesting BNF
- \( M \) to be the datatype \( M = \text{MCons} \). \( \text{unit}, M \) \( G \)
- \( r \) to be the immediate subterm relation associated to \( M \), namely \( r = \{ (m', m) | m' \in \text{set}_{G_2}(\text{MCons} m) \} \)
- \( m \) to be the composition \( \text{rawmeas} \circ \text{Rep}_T \), where \( \text{Rep}_T : \alpha T \to \alpha \text{raw} \) is the representation function for \( T \) and \( \text{rawmeas} : \alpha \text{raw} \to M \) sends any \( r \) to its recursive depth:

\[
\text{rawmeas} (\text{Raw} g) = \text{MCons} (\text{map}_G (\lambda \alpha. (\cdot)) \text{rawmeas} g)
\]

It is not hard to verify the assumptions of WFInd\(_T,F,r,m,p\). \( \Box \)

In summary, the unrestricted versions of nonuniform induction is made available via a consistent axiomatic extension of HOL. The users can choose between enabling this axiom or using the more restricted principle we were able to prove entirely in HOL.

\footnote{The reason why we treat Isabelle/HOL specially is that it allows ad hoc overloading of constants intertwined with type definitions, which is problematic in the standard HOL semantics [27, 28].}