

Chapter 3

Recursion in the Pointfree Style

How useful from a programmer's point of view are the abstract concepts presented in the previous chapter? Recall that a table was presented — table 2.1 — which records an analogy between abstract type notation and the corresponding data-structures available in common, imperative languages.

This analogy is precisely our point of departure for extending the abstract notation towards a most important field of programming: *recursion*.

3.1 Motivation

Let us consider a very common data-structure in programming: “linked-lists”. In PASCAL one will write

```
L =  $\hat{N}$ ;  
N = record  
    first: A;  
    next:  $\hat{N}$   
end;
```

to specify such a data-structure L . This consists of a pointer to a *node* (N), where a node is a record structure which puts some predefined type A together with a pointer to another node, and so on. In the C programming language, every $x \in L$ will be declared as

```
L x;
```

in the context of datatype definition

```
typedef struct N {  
    A first;
```

```

struct N *next;
} *L;

```

and so on.

What interests us in such “first year programming course” datatype declarations? Records and pointers have already been dealt with in table 2.1. So we can use this table to find the abstract version of datatype L , by replacing pointers by the “ $1 + \dots$ ” notation and records (*structs*) by the “ $\dots \times \dots$ ” notation:

$$\begin{cases} L = 1 + N \\ N = A \times (1 + N) \end{cases} \quad (3.1)$$

We obtain a system of two equations on unknowns L and N , in which L 's dependence on N can be removed by substitution:

$$\begin{aligned} & \begin{cases} L = 1 + N \\ N = A \times (1 + N) \end{cases} \\ \equiv & \quad \{ \text{substituting } L \text{ for } 1 + N \text{ in the second equation} \} \\ & \begin{cases} L = 1 + N \\ N = A \times L \end{cases} \\ \equiv & \quad \{ \text{substituting } A \times L \text{ for } N \text{ in the first equation} \} \\ & \begin{cases} L = 1 + A \times L \\ N = A \times L \end{cases} \end{aligned}$$

System (3.1) is thus equivalent to:

$$\begin{cases} L = 1 + A \times L \\ N = A \times (1 + N) \end{cases} \quad (3.2)$$

Intuitively, L abstracts the “possibly empty” linked-list of elements of type A , while N abstracts the “non-empty” linked-list of elements of type A . Note that L and N are independent of each other, but also that each depends on itself. Can we solve these equations in a way such that we obtain “solutions” for L and N , in the same way we do with school equations such as, for instance,

$$x = 1 + \frac{x}{2} \quad ? \quad (3.3)$$

Concerning this equation, let us recall how we would go about it in school mathematics:

$$x = 1 + \frac{x}{2}$$

$$\begin{aligned}
&\equiv \quad \left\{ \text{adding } -\frac{x}{2} \text{ to both sides of the equation} \right\} \\
&\quad x - \frac{x}{2} = 1 + \frac{x}{2} - \frac{x}{2} \\
&\equiv \quad \left\{ -\frac{x}{2} \text{ cancels } \frac{x}{2} \right\} \\
&\quad x - \frac{x}{2} = 1 \\
&\equiv \quad \left\{ \text{multiplying both sides of the equation by } 2 \text{ etc.} \right\} \\
&\quad 2 \times x - x = 2 \\
&\equiv \quad \left\{ \text{subtraction} \right\} \\
&\quad x = 2
\end{aligned}$$

We very quickly get solution $x = 2$. However, many steps were omitted from the actual calculation. This unfolds into the longer sequence of more elementary steps which follows, in which notation $a - b$ abbreviates $a + (-b)$ and $\frac{a}{b}$ abbreviates $a \times \frac{1}{b}$, for $b \neq 0$:

$$\begin{aligned}
&\quad x = 1 + \frac{x}{2} \\
&\equiv \quad \left\{ \text{adding } -\frac{x}{2} \text{ to both sides of the equation} \right\} \\
&\quad x - \frac{x}{2} = \left(1 + \frac{x}{2}\right) - \frac{x}{2} \\
&\equiv \quad \left\{ + \text{ is associative} \right\} \\
&\quad x - \frac{x}{2} = 1 + \left(\frac{x}{2} - \frac{x}{2}\right) \\
&\equiv \quad \left\{ -\frac{x}{2} \text{ is the additive inverse of } \frac{x}{2} \right\} \\
&\quad x - \frac{x}{2} = 1 + 0 \\
&\equiv \quad \left\{ 0 \text{ is the unit of addition} \right\} \\
&\quad x - \frac{x}{2} = 1 \\
&\equiv \quad \left\{ \text{multiplying both sides of the equation by } 2 \right\} \\
&\quad 2 \times \left(x - \frac{x}{2}\right) = 2 \times 1 \\
&\equiv \quad \left\{ 1 \text{ is the unit of multiplication} \right\} \\
&\quad 2 \times \left(x - \frac{x}{2}\right) = 2 \\
&\equiv \quad \left\{ \text{multiplication distributes over addition} \right\}
\end{aligned}$$

$$\begin{aligned}
& 2 \times x - 2 \times \frac{x}{2} = 2 \\
\equiv & \quad \{ 2 \text{ cancels its inverse } \frac{1}{2} \} \\
& 2 \times x - 1 \times x = 2 \\
\equiv & \quad \{ \text{multiplication distributes over addition} \} \\
& (2 - 1) \times x = 2 \\
\equiv & \quad \{ 2 - 1 = 1 \text{ and } 1 \text{ is the unit of multiplication} \} \\
& x = 2
\end{aligned}$$

Back to (3.2), we would like to submit each of the equations, *e.g.*

$$L = 1 + A \times L \tag{3.4}$$

to a similar reasoning. Can we do it? The analogy which can be found between this equation and (3.3) goes beyond pattern similarity. From chapter 2 we know that many properties required in the reasoning above hold in the context of (3.4), provided the “=” sign is replaced by the “ \cong ” sign, that of set-theoretical isomorphism. Recall that, for instance, + is associative (2.46), 0 is the unit of addition (2.79), 1 is the unit of multiplication (2.81), multiplication distributes over addition (2.50) *etc.* Moreover, the first step above assumed that addition is compatible (monotonic) with respect to equality,

$$\begin{array}{rcl}
a & = & b \\
c & = & d \\
\hline
a + c & = & b + d
\end{array}$$

a fact which still holds when numeric equality gives place to isomorphism and numeric addition gives place to coproduct:

$$\begin{array}{rcl}
A & \cong & B \\
C & \cong & D \\
\hline
A + C & \cong & B + D
\end{array}$$

— recall (2.44) for isos f and g .

Unfortunately, the main steps in the reasoning above are concerned with two basic *cancellation properties*

$$\begin{aligned}
x + b = c & \equiv x = c - b \\
x \times b = c & \equiv x = \frac{c}{b} \quad (b \neq 0)
\end{aligned}$$

which hold about numbers but do not hold about datatypes. In fact, neither products nor

coproducts have arbitrary inverses ¹, and so we cannot “calculate by cancellation”. How do we circumvent this limitation?

Just think of how we would have gone about (3.3) in case we didn’t know about the *cancellation properties*: we would be bound to the x by $1 + \frac{x}{2}$ substitution plus the other properties. By performing such a substitution over and over again we would obtain. . .

$$\begin{aligned}
 x &= 1 + \frac{x}{2} \\
 \equiv & \quad \{ x \text{ by } 1 + \frac{x}{2} \text{ substitution followed by simplification } \} \\
 x &= 1 + \frac{1 + \frac{x}{2}}{2} = 1 + \frac{1}{2} + \frac{x}{4} \\
 \equiv & \quad \{ \text{the same as above} \} \\
 x &= 1 + \frac{1}{2} + \frac{1 + \frac{x}{2}}{4} = 1 + \frac{1}{2} + \frac{1}{4} + \frac{x}{8} \\
 \equiv & \quad \{ \text{over and over again, } n\text{-times} \} \\
 & \dots \\
 \equiv & \quad \{ \text{simplification} \} \\
 x &= \sum_{i=0}^n \frac{1}{2^i} + \frac{x}{2^{n+1}} \\
 \equiv & \quad \{ \text{sum of } n \text{ first terms of a geometric progression} \} \\
 x &= (2 - \frac{1}{2^n}) + \frac{x}{2^{n+1}} \\
 \equiv & \quad \{ \text{let } n \rightarrow \infty \} \\
 x &= (2 - 0) + 0 \\
 \equiv & \quad \{ \text{simplification} \} \\
 x &= 2
 \end{aligned}$$

Clearly, this is a much more complicated way of finding solution $x = 2$ for equation (3.3). But we would have loved it in case it were the only known way, and this is precisely what happens with respect to (3.4). In this case we have:

$$\begin{aligned}
 L &= 1 + A \times L \\
 \equiv & \quad \{ \text{substitution of } 1 + A \times L \text{ for } L \}
 \end{aligned}$$

¹The initial and terminal datatypes do have inverses — 0 is its own “additive inverse” and 1 is its own “multiplicative inverse” — but not all the others.

$$\begin{aligned}
& L = 1 + A \times (1 + A \times L) \\
\equiv & \quad \{ \text{distributive property (2.50)} \} \\
& L \cong 1 + A \times 1 + A \times (A \times L) \\
\equiv & \quad \{ \text{unit of product (2.81) and associativity of product (2.32)} \} \\
& L \cong 1 + A + (A \times A) \times L \\
\equiv & \quad \{ \text{by (2.82), (2.84) and (2.87)} \} \\
& L \cong A^0 + A^1 + A^2 \times L \\
\equiv & \quad \{ \text{another substitution as above and similar simplifications} \} \\
& L \cong A^0 + A^1 + A^2 + A^3 \times L \\
\equiv & \quad \{ \text{after } (n+1)\text{-many similar steps} \} \\
& L \cong \sum_{i=0}^n A^i + A^{n+1} \times L
\end{aligned}$$

Bearing a large n in mind, let us deliberately (but temporarily) ignore term $A^{n+1} \times L$. Then L will be isomorphic to the sum of n -many contributions A^i ,

$$L \cong \sum_{i=0}^n A^i$$

each of them consisting of i -long tuples, or *sequences*, of values of A . (Number i is said to be the *length* of any sequence in A^i .) Such sequences will be denoted by enumerating their elements between square brackets, for instance the *empty sequence* $[\]$ which is the only inhabitant in A^0 , the two element sequence $[a_1, a_2]$ which belongs to A^2 provided $a_1, a_2 \in A$, and so on. Note that all such contributions are mutually disjoint, that is, $A^i \cap A^j = \emptyset$ wherever $i \neq j$. (In other words, a sequence of length i is never a sequence of length j , for $i \neq j$.) If we join all contributions A^i into a single set, we obtain the set of all *finite sequences* on A , denoted by A^* and defined as follows:

$$A^* \stackrel{\text{def}}{=} \bigcup_{i \geq 0} A^i \tag{3.5}$$

The intuition behind taking the limit in the numeric calculation above was that term $\frac{x}{2^{n+1}}$ was getting smaller and smaller as n went larger and larger and, “in the limit”, it could be ignored. By analogy, taking a similar limit in the calculation just sketched above will mean that, for a “sufficiently large” n , the sequences in A^n are so long that it is very unlikely that we will ever use them! So, for $n \rightarrow \infty$ we obtain

$$L \cong \sum_{i=0}^{\infty} A^i$$

Because $\sum_{i=0}^{\infty} A^i$ is isomorphic to $\bigcup_{i=0}^{\infty} A^i$ (see exercise 20), we finally have:

$$L \cong A^*$$

All in all, we have obtained A^* as a solution to equation (3.4). In other words, datatype L is isomorphic to the datatype which contains all finite sequences of some predefined datatype A . This corresponds to the HASKELL `[a]` datatype, in general. Recall that we started from the “linked-list datatype” expressed in PASCAL or C. In fact, wherever the C programmer thinks of linked-lists, the HASKELL programmer will think of finite sequences.

But, what does equation (3.4) mean in fact? Is A^* the only solution to this equation? Back to the numeric field, we know of equations which have more than one solution — for instance $x = \frac{x^2+3}{4}$, which admits two solutions 1 and 3 —, which have no solution at all — for instance $x = x + 1$ —, or which admit an infinite number of — for instance $x = x$.

We will address these topics in the next section about *inductive* datatypes and in chapter 7, where the formal semantics of recursion will be made explicit. This is where the “limit” constructions used informally in this section will be shown to make sense.

3.2 Introducing inductive datatypes

Datatype L as defined by (3.4) is said to be *recursive* because L “recurs” in the definition of L itself². From the discussion above, it is clear that set-theoretical equality “=” in this equation should give place to set-theoretical isomorphism (“ \cong ”):

$$L \cong 1 + A \times L \tag{3.6}$$

Which isomorphism $L \xleftarrow{in} 1 + A \times L$ do we expect to witness (3.4)? This will depend on which particular solution to (3.4) we are thinking of. So far we have seen only one, A^* . By recalling the notion of *algebra* of a datatype (section 2.18), so we may rephrase the question as: which algebra

$$A^* \xleftarrow{in} 1 + A \times A^*$$

do we expect to witness the tautology which arises from (3.4) by replacing unknown L with solution A^* , that is

$$A^* \cong 1 + A \times A^* \quad ?$$

²By analogy, we may regard (3.3) as a “recursive definition” of number 2.

It will have to be of the form $in = [in_1, in_2]$ as depicted by the following diagram:

$$\begin{array}{ccc}
 1 & \xrightarrow{i_1} 1 + A \times A^* & \xleftarrow{i_2} A \times A^* \\
 & \searrow in_1 & \swarrow in_2 \\
 & & A^*
 \end{array}
 \quad (3.7)$$

Arrows in_1 and in_2 can be guessed rather intuitively: $in_1 = [\]$, which will express the “NIL pointer” by the empty sequence, at A^* level, and $in_2 = cons$, where $cons$ is the standard “left append” sequence constructor, which we for the moment introduce rather informally as follows:

$$\begin{aligned}
 cons &: A \times A^* \longrightarrow A^* \\
 cons(a, [a_1, \dots, a_n]) &= [a, a_1, \dots, a_n]
 \end{aligned}
 \quad (3.8)$$

In a diagram:

$$\begin{array}{ccc}
 1 & \xrightarrow{i_1} 1 + A \times A^* & \xleftarrow{i_2} A \times A^* \\
 & \searrow [\] & \swarrow cons \\
 & & A^*
 \end{array}
 \quad (3.9)$$

Of course, for in to be iso it needs to have an inverse, which is not hard to guess,

$$out \stackrel{\text{def}}{=} (! + \langle hd, tl \rangle) \cdot (=_{[\]}?) \quad (3.10)$$

where sequence operators hd (*head of a nonempty sequence*) and tl (*tail of a nonempty sequence*) are (again informally) described as follows:

$$\begin{aligned}
 hd &: A^* \longrightarrow A \\
 hd[a_1, a_2, \dots, a_n] &= a_1
 \end{aligned}
 \quad (3.11)$$

$$\begin{aligned}
 tl &: A^* \longrightarrow A^* \\
 tl[a_1, a_2, \dots, a_n] &= [a_2, \dots, a_n]
 \end{aligned}
 \quad (3.12)$$

Showing that in and out are each other inverses is not a hard task either:

$$\begin{aligned}
 & in \cdot out = id \\
 \equiv & \quad \{ \text{definitions of } in \text{ and } out \} \\
 & [\] \cdot (! + \langle hd, tl \rangle) \cdot (=_{[\]}?) = id \\
 \equiv & \quad \{ +\text{-absorption (2.41) and (2.15)} \}
 \end{aligned}$$

$$\begin{aligned}
& [[], cons \cdot \langle hd, tl \rangle] \cdot (= []?) = id \\
\equiv & \quad \{ \text{property of sequences: } cons(hd\ s, tl\ s) = s \} \\
& [[], id] \cdot (= []?) = id \\
\equiv & \quad \{ \text{going pointwise (2.60)} \} \\
& \left\{ \begin{array}{l} = []\ a \Rightarrow [[], id] (i_1\ a) \\ \neg(= []\ a) \Rightarrow [[], id] (i_2\ a) \end{array} \right. = a \\
\equiv & \quad \{ \text{+-cancellation (2.38)} \} \\
& \left\{ \begin{array}{l} = []\ a \Rightarrow \underline{[]} a \\ \neg(= []\ a) \Rightarrow \underline{id}\ a \end{array} \right. = a \\
\equiv & \quad \{ a = [] \text{ in one case and identity function (2.9) in the other} \} \\
& \left\{ \begin{array}{l} a = [] \Rightarrow a \\ \neg(a = []) \Rightarrow a \end{array} \right. = a \\
\equiv & \quad \{ \text{property } (p \rightarrow f, f) = f \text{ holds} \} \\
& a = a
\end{aligned}$$

A comment on the particular choice of terminology above: symbol *in* suggests that we are going inside, or constructing (synthesizing) values of A^* ; symbol *out* suggests that we are going out, or destructing (analyzing) values of A^* . We shall often resort to this duality in the sequel.

Are there more solutions to equation (3.6)? In trying to implement this equation, a HASKELL programmer could have written, after the declaration of type A , the following datatype declaration:

```
data L = Nil () | Cons (A,L)
```

which, as we have seen in section 2.18, can be written simply as

```
data L = Nil | Cons (A,L) (3.13)
```

and generates diagram

$$\begin{array}{ccc}
1 & \xrightarrow{i_1} & 1 + A \times L \xleftarrow{i_2} A \times L \\
& \searrow \underline{Nil} & \downarrow in' \\
& & L
\end{array}
\quad (3.14)$$

leading to algebra $in' = [\underline{Nil}, Cons]$.

HASKELL seems to have generated another solution for the equation, which it calls L . To avoid the inevitable confusion between this symbol denoting the newly created datatype and symbol L in equation (3.6), which denotes a mathematical variable, let us use symbol T to denote the former (T stands for “type”). This can be coped with very simply by writing T instead of L in (3.13):

$$\text{data } T = \text{Nil} \mid \text{Cons } (A, T) \quad (3.15)$$

In order to make T more explicit, we will write in_T instead of in' .

Some questions are on demand at this point. First of all, what is datatype T ? What are its inhabitants? Next, is $T \xleftarrow{in_T} 1 + A \times T$ an iso or not?

HASKELL will help us to answer these questions. Suppose that A is a primitive numeric datatype, and that we add `deriving Show` to (3.15) so that we can “see” the inhabitants of the T datatype. The information associated to T is thus:

```
Main> :i T
-- type constructor
data T

-- constructors:
Nil :: T
Cons :: (A,T) -> T

-- instances:
instance Show T
instance Eval T
```

By typing `Nil`

```
Main> Nil
Nil :: T
```

we confirm that `Nil` is itself an inhabitant of T , and by typing `Cons`

```
Main> Cons
<<function>> :: (A,T) -> T
```

we realize that `Cons` is not so (as expected), but it can be used to build such inhabitants, for instance:

```
Main> Cons(1,Nil)
Cons (1,Nil) :: T
```

or

```
Main> Cons(2,Cons(1,Nil))
Cons (2,Cons (1,Nil)) :: T
```

etc. We conclude that *expressions* involving *Nil* and *Cons* are inhabitants of type T . Are these the *only* ones? The answer is *yes* because, by design of the HASKELL language, the constructors of type T will remain fixed once its declaration is interpreted, that is, no further constructor can be added to T . Does in_T have an inverse? Yes, its inverse is coalgebra

$$\begin{aligned} out_T : T &\longrightarrow 1 + A \times T \\ out_T Nil &= i_1 NIL \\ out_T(Cons(a,l)) &= i_2(a,l) \end{aligned} \quad (3.16)$$

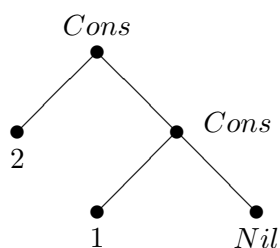
which can be straightforwardly encoded in HASKELL using the `Either` realization of $+$ (recall sections 2.9 and 2.18):

```
outT :: T -> Either () (A,T)
outT Nil = Left ()
outT (Cons(a,l)) = Right(a,l)
```

In summary, isomorphism

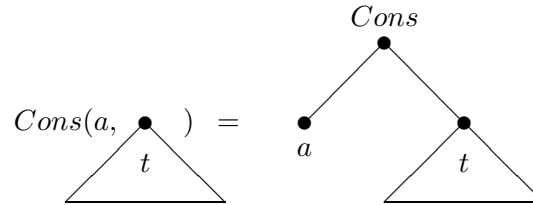
$$\begin{array}{ccc} & \xrightarrow{out_T} & \\ T & \cong & 1 + A \times T \\ & \xleftarrow{in_T} & \end{array} \quad (3.17)$$

holds, where datatype T is inhabited by symbolic expressions which we may visualize very conveniently as trees, for instance



picturing expression $Cons(2, Cons(1, Nil))$. *Nil* is the empty tree and *Cons* may be

regarded as the operation which adds a new root and a new branch, say a , to a tree t :



The choice of symbols \mathbb{T} , Nil and Cons was rather arbitrary in (3.15). Therefore, an alternative declaration such as, for instance,

$$\text{data } \mathbb{U} = \text{Stop} \mid \text{Join } (A, \mathbb{U}) \quad (3.18)$$

would have been perfectly acceptable, generating another solution for the equation under algebra $[\underline{\text{Stop}}, \text{Join}]$. It is easy to check that (3.18) is but a renaming of Nil to Stop and of Cons to Join . Therefore, both datatypes are isomorphic, or “abstractly the same”.

Indeed, any other datatype X *inductively* defined by a constant and a binary constructor accepting A and X as parameters will be a solution to the equation. Because we are just renaming symbols in a consistent way, all such solutions are abstractly the same. All of them capture the abstract notion of a *list* of symbols.

We wrote “inductively” above because the set of all expressions (trees) which inhabit the type is defined by induction. Such types are called *inductive* and we shall have a lot more to say about them in chapter 7.

Exercise 26. Obviously,

```
either (const []) (:)
```

does not work as a HASKELL realization of the mediating arrow in diagram (3.9). What do you need to write instead?

□

3.3 Observing an inductive datatype

Suppose that one is asked to express a particular *observation* of an inductive such as \mathbb{T} (3.15), that is, a function of signature $B \longleftarrow^f \mathbb{T}$ for some target type B . Suppose, for

instance, that A is \mathbb{N}_0 (the set of all non-negative integers) and that we want to add all elements which occur in a T -list. Of course, we have to ensure that addition is available in \mathbb{N}_0 ,

$$\begin{aligned} add : \mathbb{N}_0 \times \mathbb{N}_0 &\longrightarrow \mathbb{N}_0 \\ add(x, y) &\stackrel{\text{def}}{=} x + y \end{aligned}$$

and that $0 \in \mathbb{N}_0$ is a value denoting “the addition of nothing”. So constant arrow $\mathbb{N}_0 \xleftarrow{0} 1$ is available. Of course, $add(0, x) = add(x, 0) = x$ holds, for all $x \in \mathbb{N}_0$. This property means that \mathbb{N}_0 , together with operator add and constant 0 , forms a *monoid*, a very important algebraic structure in computing which will be exploited intensively later in this book. The following arrow “packaging” \mathbb{N}_0 , add and 0 ,

$$\mathbb{N}_0 \xleftarrow{[0, add]} 1 + \mathbb{N}_0 \times \mathbb{N}_0 \quad (3.19)$$

is a convenient way to express such a structure. Combining this arrow with the algebra

$$T \xleftarrow{in_T} 1 + \mathbb{N}_0 \times T \quad (3.20)$$

which defines T , and the function f we want to define, the target of which is $B = \mathbb{N}_0$, we get the almost closed diagram which follows, in which only the dashed arrow is yet to be filled in:

$$\begin{array}{ccc} T & \xleftarrow{in_T} & 1 + \mathbb{N}_0 \times T \\ f \downarrow & & \vdots \downarrow \\ \mathbb{N}_0 & \xleftarrow{[0, add]} & 1 + \mathbb{N}_0 \times \mathbb{N}_0 \end{array} \quad (3.21)$$

We know that $in_T = [Nil, Cons]$. A pattern for the missing arrow is not difficult to guess: in the same way f bridges T and \mathbb{N}_0 on the lefthand side, it will do the same job on the righthand side. So pattern $\dots + \dots \times f$ comes to mind (recall section 2.10), where the “ \dots ” are very naturally filled in by identity functions. All in all, we obtain diagram

$$\begin{array}{ccc} T & \xleftarrow{[Nil, Cons]} & 1 + \mathbb{N}_0 \times T \\ f \downarrow & & \downarrow id + id \times f \\ \mathbb{N}_0 & \xleftarrow{[0, add]} & 1 + \mathbb{N}_0 \times \mathbb{N}_0 \end{array} \quad (3.22)$$

which pictures the following property of f

$$f \cdot [Nil, Cons] = [0, add] \cdot (id + id \times f) \quad (3.23)$$

and is easy to convert to pointwise notation:

$$\begin{aligned}
& f \cdot [\underline{Nil}, Cons] = [\underline{0}, add] \cdot (id + id \times f) \\
\equiv & \quad \{ (2.40) \text{ on the lefthand side, (2.41) and identity } id \text{ on the righthand side} \} \\
& [f \cdot \underline{Nil}, f \cdot Cons] = [\underline{0}, add \cdot (id \times f)] \\
\equiv & \quad \{ \text{either structural equality (2.58)} \} \\
& \begin{cases} f \cdot \underline{Nil} = \underline{0} \\ f \cdot Cons = add \cdot (id \times f) \end{cases} \\
\equiv & \quad \{ \text{going pointwise} \} \\
& \begin{cases} (f \cdot \underline{Nil})x = \underline{0}x \\ (f \cdot Cons)(a, x) = (add \cdot (id \times f))(a, x) \end{cases} \\
\equiv & \quad \{ \text{composition (2.6), constant (2.12), product (2.22) and definition of } add \} \\
& \begin{cases} f Nil = 0 \\ f(Cons(a, x)) = a + f x \end{cases}
\end{aligned}$$

Note that we could have used $out_{\mathbb{T}}$ in diagram (3.21),

$$\begin{array}{ccc}
\mathbb{T} & \xrightarrow{out_{\mathbb{T}}} & 1 + \mathbb{N}_0 \times \mathbb{T} \\
f \downarrow & & \downarrow id + id \times f \\
\mathbb{N}_0 & \xleftarrow{[\underline{0}, add]} & 1 + \mathbb{N}_0 \times \mathbb{N}_0
\end{array} \tag{3.24}$$

obtaining another version of the *definition* of f ,

$$f = [\underline{0}, add] \cdot (id + id \times f) \cdot out_{\mathbb{T}} \tag{3.25}$$

which would lead to exactly the same pointwise recursive definition:

$$\begin{aligned}
& f = [\underline{0}, add] \cdot (id + id \times f) \cdot out_{\mathbb{T}} \\
\equiv & \quad \{ (2.41) \text{ and identity } id \text{ on the righthand side} \} \\
& f = [\underline{0}, add \cdot (id \times f)] \cdot out_{\mathbb{T}} \\
\equiv & \quad \{ \text{going pointwise on } out_{\mathbb{T}} \text{ (3.16)} \} \\
& \begin{cases} f Nil = ([\underline{0}, add \cdot (id \times f)] \cdot out_{\mathbb{T}}) Nil \\ f(Cons(a, x)) = ([\underline{0}, add \cdot (id \times f)] \cdot out_{\mathbb{T}})(a, x) \end{cases} \\
\equiv & \quad \{ \text{definition of } out_{\mathbb{T}} \text{ (3.16)} \}
\end{aligned}$$

$$\begin{aligned}
& \begin{cases} f \text{ Nil} = ([\underline{0}, \text{add} \cdot (\text{id} \times f)] \cdot i_1) \text{ Nil} \\ f(\text{Cons}(a, x)) = ([\underline{0}, \text{add} \cdot (\text{id} \times f)] \cdot i_2)(a, x) \end{cases} \\
\equiv & \quad \{ \text{+-cancellation (2.38)} \} \\
& \begin{cases} f \text{ Nil} = \underline{0} \text{ Nil} \\ f(\text{Cons}(a, x)) = (\text{add} \cdot (\text{id} \times f))(a, x) \end{cases} \\
\equiv & \quad \{ \text{simplification} \} \\
& \begin{cases} f \text{ Nil} = 0 \\ f(\text{Cons}(a, x)) = a + f x \end{cases}
\end{aligned}$$

Pointwise f mirrors the structure of type \mathbb{T} in having has many definition clauses as constructors in \mathbb{T} . Such functions are said to be defined *by induction on* the structure of their input type. If we repeat this calculation for \mathbb{N}_0^* instead of \mathbb{T} , that is, for

$$\text{out} = (! + \langle \text{hd}, \text{tl} \rangle) \cdot (=_{[]}?)$$

— recall (3.10) — taking place of $\text{out}_{\mathbb{T}}$, we get a “more algorithmic” version of f :

$$\begin{aligned}
f &= [\underline{0}, \text{add}] \cdot (\text{id} + \text{id} \times f) \cdot (! + \langle \text{hd}, \text{tl} \rangle) \cdot (=_{[]}?) \\
\equiv & \quad \{ \text{+-functor (2.42), identity and } \times \text{-absorption (2.25)} \} \\
f &= [\underline{0}, \text{add}] \cdot (! + \langle \text{hd}, f \cdot \text{tl} \rangle) \cdot (=_{[]}?) \\
\equiv & \quad \{ \text{+-absorption (2.41) and constant } \underline{0} \} \\
f &= [\underline{0}, \text{add} \cdot \langle \text{hd}, f \cdot \text{tl} \rangle] \cdot (=_{[]}?) \\
\equiv & \quad \{ \text{going pointwise on guard } (=_{[]}?) \text{ (2.60) and simplifying} \} \\
f l &= \begin{cases} l = [] \Rightarrow \underline{0} l \\ \neg(l = []) \Rightarrow (\text{add} \cdot \langle \text{hd}, f \cdot \text{tl} \rangle) l \end{cases} \\
\equiv & \quad \{ \text{simplification} \} \\
f l &= \begin{cases} l = [] \Rightarrow 0 \\ \neg(l = []) \Rightarrow \text{hdl} + f(\text{ttl}) \end{cases}
\end{aligned}$$

The outcome of this calculation can be encoded in HASKELL syntax as

```

f l | l == [] = 0
    | otherwise = head l + f (tail l)

```

or

```
f l = if l == []
      then 0
      else head l + f (tail l)
```

both requiring the equality predicate “==” and destructors “head” and “tail”.

3.4 Synthesizing an inductive datatype

The issue which concerns us in this section dualizes what we have just dealt with: instead of analyzing or *observing* an inductive type such as \mathbb{T} (3.15), we want to be able to synthesize (generate) particular inhabitants of \mathbb{T} . In other words, we want to be able to specify functions with signature $B \xrightarrow{f} \mathbb{T}$ for some given source type B . Let $B = \mathbb{N}_0$ and suppose we want f to generate, for a given natural number $n > 0$, the list containing all numbers less or equal to n in decreasing order

$$\text{Cons}(n, \text{Cons}(n-1, \text{Cons}(\dots, \text{Nil})))$$

or the empty list Nil , in case $n = 0$.

Let us try and draw a diagram similar to (3.24) applicable to the new situation. In trying to “re-use” this diagram, it is immediate that arrow f should be reversed. Bearing duality in mind, we may feel tempted to reverse all arrows just to see what happens. Identity functions are their own inverses, and $\text{in}_{\mathbb{T}}$ takes the place of $\text{out}_{\mathbb{T}}$:

$$\begin{array}{ccc} \mathbb{T} & \xleftarrow{\text{in}_{\mathbb{T}}} & 1 + \mathbb{N}_0 \times \mathbb{T} \\ f \uparrow & & \uparrow \text{id} + \text{id} \times f \\ \mathbb{N}_0 & \xrightarrow{\dots\dots\dots} & 1 + \mathbb{N}_0 \times \mathbb{N}_0 \end{array}$$

Interestingly enough, the bottom arrow is the one which is not obvious to reverse, meaning that we have to “invent” a particular destructor of \mathbb{N}_0 , say

$$\mathbb{N}_0 \xrightarrow{g} 1 + \mathbb{N}_0 \times \mathbb{N}_0$$

fitting in the diagram and *generating* the particular computational effect we have in mind. Once we do this, a recursive definition for f will pop out immediately,

$$f = \text{in}_{\mathbb{T}} \cdot (\text{id} + \text{id} \times f) \cdot g \quad (3.26)$$

which is equivalent to:

$$f = [\underline{\text{Nil}}, \text{Cons} \cdot (\text{id} \times f)] \cdot g \quad (3.27)$$

Because we want $f\ 0 = Nil$ to hold, g (the actual generator of the computation) should distinguish input 0 from all the others. One thus decomposes g as follows,

$$\mathbb{N}_0 \xrightarrow{=0?} \mathbb{N}_0 + \mathbb{N}_0 \xrightarrow{!+h} 1 + \mathbb{N}_0 \times \mathbb{N}_0$$

\xrightarrow{g}

leaving h to fill in. This will be a *split* providing, on the lefthand side, for the value to be *Cons*'ed to the output and, on the righthand side, for the “seed” to the next recursive call. Since we want the output values to be produced contiguously and in decreasing order, we may define $h = \langle id, pred \rangle$ where, for $n > 0$,

$$pred\ n \stackrel{\text{def}}{=} n - 1 \quad (3.28)$$

computes the *predecessor* of n . Altogether, we have synthesized

$$g = (! + \langle id, pred \rangle) \cdot (=0?) \quad (3.29)$$

Filling this in (3.27) we get

$$\begin{aligned} f &= [\underline{Nil}, Cons \cdot (id \times f)] \cdot (! + \langle id, pred \rangle) \cdot (=0?) \\ \equiv & \quad \{ \text{+ -absorption (2.41) followed by } \times \text{-absorption (2.25) etc.} \} \\ f &= [\underline{Nil}, Cons \cdot \langle id, f \cdot pred \rangle] \cdot (=0?) \\ \equiv & \quad \{ \text{going pointwise on guard } =_0? \text{ (2.60) and simplifying} \} \\ f\ n &= \begin{cases} n = 0 & \Rightarrow Nil \\ \neg(n = 0) & \Rightarrow Cons(n, f(n - 1)) \end{cases} \end{aligned}$$

which matches the function we had in mind:

```
f n | n == 0    = Nil
    | otherwise = Cons(n, f(n-1))
```

We shall see briefly that the constructions of the f function adding up a list of numbers in the previous section and, in this section, of the f function generating a list of numbers are very standard in algorithm design and can be broadly generalized. Let us first introduce some standard terminology.

3.5 Introducing (list) catas, anas and hylos

Suppose that, back to section 3.3, we want to *multiply*, rather than add, the elements occurring in lists of type T (3.15). How much of the program synthesis effort presented there can be reused in the design of the new function?

It is intuitive that only the bottom arrow $\mathbb{N}_0 \xleftarrow{[0,add]} 1 + \mathbb{N}_0 \times \mathbb{N}_0$ of diagram (3.24) needs to be replaced, because this is the only place where we can specify that target datatype \mathbb{N}_0 is now regarded as the carrier of another (multiplicative rather than additive) monoidal structure,

$$\mathbb{N}_0 \xleftarrow{[\perp,mul]} 1 + \mathbb{N}_0 \times \mathbb{N}_0 \quad (3.30)$$

for $mul(x, y) \stackrel{\text{def}}{=} x \cdot y$. We are saying that the argument list is now to be reduced by the multiplication operator and that output value 1 is expected as the result of “nothing left to multiply”.

Moreover, in the previous section we might have wanted our number-list generator to produce the list of even numbers smaller than a given number, in decreasing order (see exercise 29). Intuition will once again help us in deciding that only arrow g in (3.26) needs to be updated.

The following diagrams generalize both constructions by leaving such bottom arrows unspecified,

$$\begin{array}{ccc} \mathbb{T} & \xrightarrow{out_{\mathbb{T}}} & 1 + \mathbb{N}_0 \times \mathbb{T} \\ f \downarrow & & \downarrow id+id \times f \\ B & \xleftarrow{g} & 1 + \mathbb{N}_0 \times B \end{array} \quad \begin{array}{ccc} \mathbb{T} & \xleftarrow{in_{\mathbb{T}}} & 1 + \mathbb{N}_0 \times \mathbb{T} \\ f \uparrow & & \uparrow id+id \times f \\ B & \xrightarrow{g} & 1 + \mathbb{N}_0 \times B \end{array} \quad (3.31)$$

and express their duality (*cf.* the directions of the arrows). It so happens that, for each of these diagrams, f is uniquely dependent on the g arrow, that is to say, each particular instantiation of g will determine the corresponding f . So both g s can be regarded as “seeds” or “genetic material” of the f functions they uniquely define³.

Following the standard terminology, we express these facts by writing $f = \langle g \rangle$ with respect to the lefthand side diagram and by writing $f = \llbracket g \rrbracket$ with respect to the righthand side diagram. Read $\langle g \rangle$ as “the \mathbb{T} -*catamorphism* induced by g ” and $\llbracket g \rrbracket$ as “the \mathbb{T} -*anamorphism* induced by g ”. This terminology is derived from the Greek words $\kappa\alpha\tau\alpha$ (*cata*) and $\alpha\nu\alpha$ (*ana*) meaning, respectively, “downwards” and “upwards” (compare with the direction of the f arrow in each diagram). The exchange of parentheses “ $\langle \rangle$ ” and “ $\llbracket \rrbracket$ ” in double parentheses “ $\langle \rangle$ ” and “ $\llbracket \rrbracket$ ” is aimed at expressing the duality of both concepts.

We shall have a lot to say about catamorphisms and anamorphisms of a given type such as \mathbb{T} . For the moment, it suffices to say that

- the \mathbb{T} -catamorphism induced by $B \xleftarrow{g} 1 + \mathbb{N}_0 \times B$ is the unique function $B \xleftarrow{\langle g \rangle} \mathbb{T}$

³The theory which supports the statements of this paragraph will not be dealt with until chapter 7.

which obeys to property (or is defined by)

$$\langle g \rangle = g \cdot (id + id \times \langle g \rangle) \cdot out_{\mathbb{T}} \quad (3.32)$$

which is the same as

$$\langle g \rangle \cdot in_{\mathbb{T}} = g \cdot (id + id \times \langle g \rangle) \quad (3.33)$$

- given $B \xrightarrow{g} 1 + \mathbb{N}_0 \times B$ the \mathbb{T} -anamorphism induced by g is the unique function $B \xrightarrow{\langle g \rangle} \mathbb{T}$ which obeys to property (or is defined by)

$$\langle g \rangle = in_{\mathbb{T}} \cdot (id + id \times \langle g \rangle) \cdot g \quad (3.34)$$

From (3.31) it can be observed that \mathbb{T} can act as a mediator between any \mathbb{T} -anamorphism and any \mathbb{T} -catamorphism, that is to say, $B \xleftarrow{\langle g \rangle} \mathbb{T}$ composes with $\mathbb{T} \xleftarrow{\langle h \rangle} C$, for some $C \xrightarrow{h} 1 + \mathbb{N}_0 \times C$. In other words, a \mathbb{T} -catamorphism call always observe (consume) the output of a \mathbb{T} -anamorphism. The latter produces a list of \mathbb{N}_0 s which is consumed by the former. This is depicted in the diagram which follows:

$$\begin{array}{ccc} B & \xleftarrow{g} & 1 + \mathbb{N}_0 \times B \\ \langle g \rangle \uparrow & & \uparrow id + id \times \langle g \rangle \\ \mathbb{T} & \xleftarrow{in_{\mathbb{T}}} & 1 + \mathbb{N}_0 \times \mathbb{T} \\ \langle h \rangle \uparrow & & \uparrow id + id \times \langle h \rangle \\ C & \xrightarrow{h} & 1 + \mathbb{N}_0 \times C \end{array} \quad (3.35)$$

What can we say about the $\langle g \rangle \cdot \langle h \rangle$ composition? It is a function from B to C which resorts to \mathbb{T} as an *intermediate* data-structure and can be subject to the following calculation (cf. outermost rectangle in (3.35)):

$$\begin{aligned} \langle g \rangle \cdot \langle h \rangle &= g \cdot (id + id \times \langle g \rangle) \cdot (id + id \times \langle h \rangle) \cdot h \\ \equiv & \quad \{ \text{+-functor (2.42)} \} \\ \langle g \rangle \cdot \langle h \rangle &= g \cdot ((id \cdot id) + (id \times \langle g \rangle) \cdot (id \times \langle h \rangle)) \cdot h \\ \equiv & \quad \{ \text{identity and } \times\text{-functor (2.28)} \} \\ \langle g \rangle \cdot \langle h \rangle &= g \cdot (id + id \times \langle g \rangle) \cdot \langle h \rangle \cdot h \end{aligned}$$

This calculation shows how to define $C \xleftarrow{\langle g \rangle \cdot \langle h \rangle} B$ in one go, that is to say, doing without any intermediate data-structure:

$$\begin{array}{ccc}
 B & \xleftarrow{g} & 1 + \mathbb{N}_0 \times B \\
 \langle g \rangle \cdot \langle h \rangle \uparrow & & \uparrow id + id \times \langle g \rangle \cdot \langle h \rangle \\
 C & \xrightarrow{h} & 1 + \mathbb{N}_0 \times C
 \end{array} \tag{3.36}$$

As an example, let us see what comes out of $\langle g \rangle \cdot \langle h \rangle$ for h and g respectively given by (3.29) and (3.30):

$$\begin{aligned}
 \langle g \rangle \cdot \langle h \rangle &= g \cdot (id + id \times \langle g \rangle \cdot \langle h \rangle) \cdot h \\
 \equiv & \quad \{ \langle g \rangle \cdot \langle h \rangle \text{ abbreviated to } f \text{ and instantiating } h \text{ and } g \} \\
 f &= [\underline{1}, mul] \cdot (id + id \times f) \cdot (! + \langle id, pred \rangle) \cdot (=_{0?}) \\
 \equiv & \quad \{ +-functor (2.42) \text{ and identity } \} \\
 f &= [\underline{1}, mul] \cdot (! + (id \times f) \cdot \langle id, pred \rangle) \cdot (=_{0?}) \\
 \equiv & \quad \{ \times\text{-absorption (2.25) and identity } \} \\
 f &= [\underline{1}, mul] \cdot (! + \langle id, f \cdot pred \rangle) \cdot (=_{0?}) \\
 \equiv & \quad \{ +-absorption (2.41) \text{ and constant } \underline{1} (2.15) \} \\
 f &= [\underline{1}, mul \cdot \langle id, f \cdot pred \rangle] \cdot (=_{0?}) \\
 \equiv & \quad \{ \text{McCarthy conditional (2.59)} \} \\
 f &= (=_{0?}) \rightarrow \underline{1}, mul \cdot \langle id, f \cdot pred \rangle
 \end{aligned}$$

Going pointwise, we get — via (2.59) —

$$\begin{aligned}
 f 0 &= [\underline{1}, mul \cdot \langle id, f \cdot pred \rangle](i_1 0) \\
 &= \quad \{ +-cancellation (2.38) \} \\
 & \quad \underline{1} 0 \\
 &= \quad \{ \text{constant function (2.12)} \} \\
 & \quad 1
 \end{aligned}$$

and

$$\begin{aligned}
 f(n+1) &= [\underline{1}, mul \cdot \langle id, f \cdot pred \rangle](i_2(n+1)) \\
 &= \quad \{ +-cancellation (2.38) \}
 \end{aligned}$$

$$\begin{aligned}
& mul \cdot \langle id, f \cdot pred \rangle (n + 1) \\
= & \{ \text{pointwise definitions of } split, \text{ identity, predecessor and } mul \} \\
& (n + 1) \times f n
\end{aligned}$$

In summary, f is but the well-known factorial function:

$$\begin{cases} f 0 = 1 \\ f(n + 1) = (n + 1) \times f n \end{cases}$$

This result comes to no surprise if we look at diagram (3.35) for the particular g and h we have considered above and recall a popular “definition” of factorial:

$$n! = \underbrace{n \times (n - 1) \times \dots \times 1}_{n \text{ times}} \quad (3.37)$$

In fact, $[[h]] n$ produces T-list

$$Cons(n, Cons(n - 1, \dots Cons(1, Nil)))$$

as an intermediate data-structure which is consumed by $[[g]]$, the effect of which is but the “replacement” of $Cons$ by \times and Nil by 1, therefore accomplishing (3.37) and realizing the computation of factorial.

The moral of this example is that a function as simple as factorial can be *decomposed* into two components (producer/consumer functions) which share a common intermediate inductive datatype. The producer function is an anamorphism which “represents” or produces a “view” of its input argument as a value of the intermediate datatype. The consumer function is a catamorphism which reduces this intermediate data-structure and produces the final result. Like factorial, many functions can be handsomely expressed by a $[[g]] \cdot [[h]]$ composition for a suitable choice of the intermediate type, and of g and h . The intermediate data-structure is said to be *virtual* in the sense that it only exists as a means to induce the associated pattern of recursion and disappears by calculation.

The composition $[[g]] \cdot [[h]]$ of a T-catamorphism with a T-anamorphism is called a T-*hylomorphism*⁴ and is denoted by $[[g, h]]$. Because g and h fully determine the behaviour of the $[[g, h]]$ function, they can be regarded as the “genes” of the function they define. As we shall see, this analogy with biology will prove specially useful for algorithm analysis and classification.

Exercise 27. A way of computing n^2 , the square of a given natural number n , is to sum up the n first odd numbers. In fact, $1^2 = 1$, $2^2 = 1 + 3$, $3^2 = 1 + 3 + 5$, etc., $n^2 = (2n - 1) + (n - 1)^2$. Following this hint, express function

$$sq n \stackrel{\text{def}}{=} n^2 \quad (3.38)$$

⁴This terminology is derived from the Greek word $\nu\lambda\omicron\sigma$ (hylos) meaning “matter”.

as a T-hylomorphism and encode it in HASKELL.

□

Exercise 28. Write function x^n as a T-hylomorphism and encode it in HASKELL.

□

Exercise 29. The following function in HASKELL computes the T-sequence of all even numbers less or equal to n :

```
f n = if n <= 1
      then Nil
      else Cons(m, f(m-2))
      where m = if even n then n else n-1
```

Find its “genetic material”, that is, function g such that $f = \llbracket g \rrbracket$ in

$$\begin{array}{ccc}
 \mathbf{T} & \xleftarrow{\text{in}_{\mathbf{T}}} & 1 + \mathbf{N}_0 \times \mathbf{T} \\
 \uparrow \llbracket g \rrbracket & & \uparrow \text{id} + \text{id} \times \llbracket g \rrbracket \\
 \mathbf{N}_0 & \xrightarrow{g} & 1 + \mathbf{N}_0 \times \mathbf{N}_0
 \end{array}$$

□

3.6 Inductive types more generally

So far we have focussed our attention exclusively to a particular inductive type \mathbf{T} (3.20) — that of finite sequences of non-negative integers. This is, of course, of a very limited scope. First, because one could think of finite sequences of other datatypes, *e.g.* Booleans or many others. Second, because other datatypes such as trees, hash-tables *etc.* exist which our notation and method should be able to take into account.

Although a generic theory of arbitrary datatypes requires a theoretical elaboration which cannot be explained at once, we can move a step further by taking the two observations above as starting points. We shall start from the latter in order to talk generically about inductive types. Then we introduce parameterization and functorial behaviour.

Suppose that, as a mere notational convention, we abbreviate every expression of the form “ $1 + \mathbb{N}_0 \times \dots$ ” occurring in the previous section by “ $F \dots$ ”, e.g. $1 + \mathbb{N}_0 \times B$ by $F B$, e.g. $1 + \mathbb{N}_0 \times T$ by $F T$

$$\begin{array}{ccc} & \xrightarrow{\text{out}_T} & \\ T & \cong & F T \\ & \xleftarrow{\text{in}_T} & \end{array} \quad (3.39)$$

etc. This is the same as introducing a datatype-level operator

$$F X \stackrel{\text{def}}{=} 1 + \mathbb{N}_0 \times X \quad (3.40)$$

which maps every datatype A into datatype $1 + \mathbb{N}_0 \times A$. Operator F captures the pattern of recursion which is associated to so-called “right” lists (of non-negative integers), that is, lists which grow to the right. The slightly different pattern $G X \stackrel{\text{def}}{=} 1 + X \times \mathbb{N}_0$ will generate a different, although related, inductive type

$$X \cong 1 + X \times \mathbb{N}_0 \quad (3.41)$$

— that of so-called “left” lists (of non-negative integers). And it is not difficult to think of the pattern which merges both right and left lists and gives rise to bi-linear lists, better known as *binary trees*:

$$X \cong 1 + X \times \mathbb{N}_0 \times X \quad (3.42)$$

One may think of many other expressions $F X$ and guess the inductive datatype they generate, for instance $H X \stackrel{\text{def}}{=} \mathbb{N}_0 + \mathbb{N}_0 \times X$ generating non-empty lists of non-negative integers (\mathbb{N}_0^+). The general rule is that, given an inductive datatype definition of the form

$$X \cong F X \quad (3.43)$$

(also called a domain equation), its pattern of recursion is captured by a so-called *functor* F .

3.7 Functors

The concept of a functor F , borrowed from category theory, is a most generic and useful device in programming⁵. As we have seen, F can be regarded as a datatype constructor

⁵The category theory practitioner must be warned of the fact that the word *functor* is used here in a too restrictive way. A proper (generic) definition of a functor will be provided later in this

which, given datatype A , builds a more elaborate datatype $F A$; given another datatype B , builds a similarly elaborate datatype $F B$; and so on. But what is more important and has the most beneficial consequences is that, if F is regarded as a functor, then its data-structuring effect extends smoothly to functions in the following way: suppose that $B \xleftarrow{f} A$ is a function which observes A into B , which are parameters of $F A$ and $F B$, respectively. By definition, if F is a functor then $F B \xleftarrow{F f} F A$ exists for every such f :

$$\begin{array}{ccc} A & \cdots & F A \\ f \downarrow & & \downarrow F f \\ B & \cdots & F B \end{array}$$

$F f$ extends f to F -structures and will, by definition, obey to two very basic properties: it commutes with identity

$$F id_A = id_{(F A)} \quad (3.44)$$

and with composition

$$F(g \cdot h) = (F g) \cdot (F h) \quad (3.45)$$

Two simple examples of a functor follow:

- Identity functor: define $F X = X$, for every datatype X , and $F f = f$. Properties (3.44) and (3.45) hold trivially just by removing symbol F wherever it occurs.
- Constant functors: for a given C , define $F X = C$ (for all datatypes X) and $F f = id_C$, as expressed in the following diagram:

$$\begin{array}{ccc} A & \cdots & C \\ f \downarrow & & \downarrow id_C \\ B & \cdots & C \end{array}$$

Properties (3.44) and (3.45) hold trivially again.

In the same way functions can be unary, binary, *etc.*, we can have functors with more than one argument. So we get binary functors (also called *bifunctors*), ternary functors *etc.* Of course, properties (3.44) and (3.45) have to hold for every parameter of an n -ary functor. For a binary functor B , for instance, equation (3.44) becomes

$$B(id_A, id_B) = id_{B(A,B)} \quad (3.46)$$

Data construction	Universal construct	Functor	Description
$A \times B$	$\langle f, g \rangle$	$f \times g$	Product
$A + B$	$[f, g]$	$f + g$	Coproduct
B^A	\overline{f}	f^A	Exponential

Table 3.1: Datatype constructions and associated operators.

and equation (3.45) becomes

$$B(g \cdot h, i \cdot j) = B(g, i) \cdot B(h, j) \quad (3.47)$$

Product and coproduct are typical examples of bifunctors. In the former case one has $B(A, B) = A \times B$ and $B(f, g) = f \times g$ — recall (2.22). Properties (2.29) and (2.28) instantiate (3.46) and (3.47), respectively, and this explains why we called them the functorial properties of product. In the latter case, one has $B(A, B) = A + B$ and $B(f, g) = f + g$ — recall (2.37) — and functorial properties (2.43) and (2.42). Finally, exponentiation is a functorial construction too: assuming A , one has $F X \stackrel{\text{def}}{=} X^A$ and $F f \stackrel{\text{def}}{=} \overline{f \cdot ap}$ and functorial properties (2.73) and (2.74). All this is summarized in table 3.1.

Such as functions, functors may compose with each other in the obvious way: the composition of F and G , denoted $F \cdot G$, is defined by

$$(F \cdot G)X \stackrel{\text{def}}{=} F(GX) \quad (3.48)$$

$$(F \cdot G)f \stackrel{\text{def}}{=} F(Gf) \quad (3.49)$$

3.8 Polynomial functors

We may put constant, product, coproduct and identity functors together to obtain so-called *polynomial functors*, which are described by polynomial expressions, for instance

$$F X = 1 + A \times X$$

— recall (3.6). A polynomial functor is either

- a constant functor or the identity functor, or
- the (finitary) product or coproduct (sum) of other polynomial functors, or
- the composition of other polynomial functors.

So the effect on arrows of a polynomial functor is computed in an easy and structured way, for instance:

$$\begin{aligned}
 Ff &= (1 + A \times X)f \\
 &= \{ \text{sum of two functors where } A \text{ is a constant and } X \text{ is a variable} \} \\
 &\quad (1)f + (A \times X)f \\
 &= \{ \text{constant functor and product of two functors} \} \\
 &\quad id_1 + (A)f \times (X)f \\
 &= \{ \text{constant functor and identity functor} \} \\
 &\quad id_1 + id_A \times f \\
 &= \{ \text{subscripts dropped for simplicity} \} \\
 &\quad id + id \times f
 \end{aligned}$$

So, $1 + A \times f$ denotes the same as $id_1 + id_A \times f$, or even the same as $id + id \times f$ if one drops the subscripts.

It should be clear at this point that what was referred to in section 2.10 as a “symbolic pattern” applicable to both datatypes and arrows is after all a functor in the mathematical sense. The fact that the same polynomial expression is used to denote both the data and the operators which structurally transform such data is of great conceptual economy and practical application. For instance, once polynomial functor (3.40) is assumed, the diagrams in (3.31) can be written as simply as

$$\begin{array}{ccc}
 T & \xrightarrow{out_T} & FT \\
 f \downarrow & & \downarrow Ff \\
 B & \xleftarrow{g} & FB
 \end{array}
 \quad
 \begin{array}{ccc}
 T & \xleftarrow{in_T} & FT \\
 f \uparrow & & \uparrow Ff \\
 B & \xrightarrow{g} & FB
 \end{array}
 \quad (3.50)$$

It is useful to know that, thanks to the isomorphism laws studied in chapter 2, every polynomial functor F may be put into the canonical form,

$$\begin{aligned}
 FX &\cong C_0 + (C_1 \times X) + (C_2 \times X^2) + \cdots + (C_n \times X^n) \\
 &= \sum_{i=0}^n C_i \times X^i
 \end{aligned}
 \quad (3.51)$$

and that *Newton’s binomial formula*

$$(A + B)^n \cong \sum_{p=0}^n {}^n C_p \times A^{n-p} \times B^p
 \quad (3.52)$$

can be used in such conversions. These are performed up to isomorphism, that is to say, after the conversion one gets a different but isomorphic datatype. Consider, for instance, functor

$$F X \stackrel{\text{def}}{=} A \times (1 + X)^2$$

(where A is a constant datatype) and check the following reasoning:

$$\begin{aligned}
 F X &= A \times (1 + X)^2 \\
 &\cong \{ \text{law (2.87)} \} \\
 &\quad A \times ((1 + X) \times (1 + X)) \\
 &\cong \{ \text{law (2.50)} \} \\
 &\quad A \times ((1 + X) \times 1 + (1 + X) \times X) \\
 &\cong \{ \text{laws (2.81), (2.31) and (2.50)} \} \\
 &\quad A \times ((1 + X) + (1 \times X + X \times X)) \\
 &\cong \{ \text{laws (2.81) and (2.87)} \} \\
 &\quad A \times ((1 + X) + (X + X^2)) \\
 &\cong \{ \text{law (2.46)} \} \\
 &\quad A \times (1 + (X + X) + X^2) \\
 &\cong \{ \text{canonical form obtained via laws (2.50) and (2.88)} \} \\
 &\quad \underbrace{A}_{C_0} + \underbrace{A \times 2 \times X}_{C_1} + \underbrace{A}_{C_2} \times X^2
 \end{aligned}$$

Exercise 30. Synthesize the isomorphism $A + A \times 2 \times X + A \times X^2 \xleftarrow{\nu} A \times (1 + X^2)$ implicit in the above reasoning.

□

3.9 Polynomial inductive types

An inductive datatype is said to be *polynomial* wherever its pattern of recursion is described by a polynomial functor, that is to say, wherever F in equation (3.43) is polynomial. For instance, datatype T (3.20) is polynomial ($n = 1$) and its associated polynomial

functor is canonically defined with coefficients $C_0 = 1$ and $C_1 = \mathbb{N}_0$. For reasons that will become apparent later on, we shall always impose $C_0 \neq 0$ to hold in a *polynomial* datatype expressed in canonical form.

Polynomial types are easy to encode in HASKELL wherever the associated functor is in canonical polynomial form, that is, wherever one has

$$\begin{array}{c} \mathbb{T} \\ \swarrow \cong \\ \text{in}_{\mathbb{T}} \quad \sum_{i=0}^n C_i \times \mathbb{T}^i \end{array} \quad (3.53)$$

Then we have

$$\text{in}_{\mathbb{T}} \stackrel{\text{def}}{=} [f_1, \dots, f_n]$$

where, for $i = 1, n$, f_i is an arrow of type $\mathbb{T} \longleftarrow C_i \times \mathbb{T}^i$. Since n is finite, one may expand exponentials according to (2.87) and encode this in HASKELL as follows:

```
data T = C0 |
        C1 (C1, T) |
        C2 (C2, (T, T)) |
        ... |
        Cn (Cn, (T, ..., T))
```

Of course the choice of symbol C_i to realize each f_i is arbitrary⁶. Several instances of polynomial inductive types (in canonical form) will be mentioned in section 3.13. Section 3.16 will address the conversion between inductive datatypes induced by so-called *natural transformations*.

The concepts of catamorphism, anamorphism and hylomorphism introduced in section 3.5 can be extended to arbitrary polynomial types. We devote the following sections to explaining catamorphisms in the polynomial setting. Polynomial anamorphisms and hylomorphisms will not be dealt with until chapter 7.

3.10 F-algebras and F-homomorphisms

Our interest in polynomial types is basically due to the fact that, for polynomial F , equation (3.43) always has a particularly interesting solution which corresponds to our notion of a recursive datatype.

⁶A more traditional (but less close to (3.53)) encoding will be

$$\text{data T} = \text{C0} \mid \text{C1 C1 T} \mid \text{C2 C2 T T} \mid \dots \mid \text{Cn Cn T} \dots \text{T} \quad (3.54)$$

delivering every constructor in curried form.

In order to explain this, we need two notions which are easy to understand: first, that of an *F-algebra*, which simply is any function α of signature $A \xleftarrow{\alpha} F A$. A is called the *carrier* of F-algebra α and contains the values which α manipulates by computing new A -values out of existing ones, according to the F-pattern (the “type” of the algebra). As examples, consider $[\underline{0}, add]$ (3.19) and in_{\top} (3.20), which are both algebras of type $F X = 1 + \mathbb{N}_0 \times X$. The type of an algebra clearly determines its form. For instance, any algebra α of type $F X = 1 + X \times X$ will be of form $[\alpha_1, \alpha_2]$, where α_1 is a constant and α_2 is a binary operator. So monoids are algebras of this type⁷.

Secondly, we introduce the notion of an *F-homomorphism* which is but a function observing a particular F-algebra α into another F-algebra β :

$$\begin{array}{ccc}
 A & \xleftarrow{\alpha} & F A \\
 f \downarrow & & \downarrow F f \\
 B & \xleftarrow{\beta} & F B
 \end{array}
 \quad f \cdot \alpha = \beta \cdot (F f)
 \quad (3.55)$$

Clearly, f can be regarded as a structural translation between A and B , that is, A and B have a similar structure⁸. Note that — thanks to (3.44) — identity functions are always (trivial) F-homomorphisms and that — thanks to (3.45) — these homomorphisms compose, that is, the composition of two F-homomorphisms is an F-homomorphism.

3.11 F-catamorphisms

An F-algebra can be epic, monic or both, that is, iso. Iso F-algebras are particularly relevant to our discussion because they describe solutions to the $X \cong F X$ equation (3.43). Moreover, for polynomial F a particular iso F-algebra always exists, which is denoted by $\mu F \xleftarrow{in} F \mu F$ and has special properties. First, its carrier is the smallest among the carriers of other iso F-algebras, and this is why it is denoted by μF — μ for “minimal”⁹. Second, it is the so-called *initial* F-algebra. What does this mean?

It means that, for every F-algebra α there exists one and only one F-homomorphism between in and α . This unique arrow mediating in and α is therefore determined by α itself, and is called the *F-catamorphism* generated by α . This construct, which was introduced in 3.5, is in general denoted by $(\alpha)_F$:

⁷But not every algebra of this type is a monoid, since the type of an algebra only fixes its syntax and does not impose any properties such as associativity, *etc.*

⁸Cf. *homomorphism* = *homo* (the same) + *morphos* (structure, shape).

⁹ μF means the least fixpoint solution of equation $X \cong F X$, as will be described in chapter 7.

$$\begin{array}{ccc}
 \mu F & \xleftarrow{in} & F \mu F \\
 f = \langle \alpha \rangle_F \downarrow & & \downarrow F \langle \alpha \rangle_F \\
 A & \xleftarrow{\alpha} & F A
 \end{array} \tag{3.56}$$

We will drop the F subscript in $\langle \alpha \rangle_F$ wherever deducible from the context, and often call α the “gene” of $\langle \alpha \rangle_F$.

As happens with *splits*, *eithers* and transposes, the uniqueness of the catamorphism construct is captured by a universal property established in the class of all F -homomorphisms:

$$k = \langle \alpha \rangle \Leftrightarrow k \cdot in = \alpha \cdot F k \tag{3.57}$$

According to the experience gathered from section 2.12 onwards, a few properties can be expected as consequences of (3.57). For instance, one may wonder about the “gene” of the identity catamorphism. Just let $k = id$ in (3.57) and see what happens:

$$\begin{aligned}
 id &= \langle \alpha \rangle \Leftrightarrow id \cdot in = \alpha \cdot F id \\
 &= \{ \text{identity (2.10) and } F \text{ is a functor (3.44)} \} \\
 id &= \langle \alpha \rangle \Leftrightarrow in = \alpha \cdot id \\
 &= \{ \text{identity (2.10) once again} \} \\
 id &= \langle \alpha \rangle \Leftrightarrow in = \alpha \\
 &= \{ \alpha \text{ replaced by } in \text{ and simplifying} \} \\
 id &= \langle in \rangle
 \end{aligned}$$

Thus one finds out that the genetic material of the identity catamorphism is the initial algebra in . Which is the same as establishing the *reflection property* of catamorphisms:

Cata-reflection :

$$\begin{array}{ccc}
 \mu F & \xleftarrow{in} & F \mu F \\
 \langle in \rangle \downarrow & & \downarrow F \langle in \rangle \\
 \mu F & \xleftarrow{in} & F \mu F
 \end{array} \quad \langle in \rangle = id_{\mu F} \tag{3.58}$$

In a more intuitive way, one might have observed that $\langle in \rangle$ is, by definition of in , the unique arrow mediating μF and itself. But another arrow of the same type is already known: the identity $id_{\mu F}$. So these two arrows must be the same.

Another property following immediately from (3.57), for $k = \langle \alpha \rangle$, is

Cata-cancellation :

$$(\downarrow\alpha) \cdot in = \alpha \cdot F(\downarrow\alpha) \quad (3.59)$$

Because *in* is iso, this law can be rephrased as follows

$$(\downarrow\alpha) = \alpha \cdot F(\downarrow\alpha) \cdot out \quad (3.60)$$

where *out* denotes the inverse of *in*:

$$\begin{array}{ccc} & \xrightarrow{out} & \\ \mu F & \cong & F \mu F \\ & \xleftarrow{in} & \end{array}$$

Now, let *f* be F-homomorphism (3.55) between F-algebras α and β . How does it relate to $(\downarrow\alpha)$ and $(\downarrow\beta)$? Note that $f \cdot (\downarrow\alpha)$ is an arrow mediating μF and B . But B is the carrier of β and $(\downarrow\beta)$ is the unique arrow mediating μF and B . So the two arrows are the same:

Cata-fusion :

$$\begin{array}{ccc} \mu F & \xleftarrow{in} & F \mu F & f \cdot (\downarrow\alpha) = (\downarrow\beta) \quad \text{if} \quad f \cdot \alpha = \beta \cdot F f & (3.61) \\ (\downarrow\alpha) \downarrow & & \downarrow F(\downarrow\alpha) & & \\ A & \xleftarrow{\alpha} & F A & & \\ f \downarrow & & \downarrow F f & & \\ B & \xleftarrow{\beta} & F B & & \end{array}$$

Of course, this law is also a consequence of the universal property, for $k = f \cdot (\downarrow\alpha)$:

$$\begin{aligned} f \cdot (\downarrow\alpha) = (\downarrow\beta) & \Leftrightarrow (f \cdot (\downarrow\alpha)) \cdot in = \beta \cdot F(f \cdot (\downarrow\alpha)) \\ & \Leftrightarrow \{ \text{composition is associative and F is a functor (3.45)} \} \\ & \quad f \cdot ((\downarrow\alpha) \cdot in) = \beta \cdot (F f) \cdot (F(\downarrow\alpha)) \\ & \Leftrightarrow \{ \text{cata-cancellation (3.59)} \} \\ & \quad f \cdot \alpha \cdot F(\downarrow\alpha) = \beta \cdot F f \cdot F(\downarrow\alpha) \\ & \Leftrightarrow \{ \text{require } f \text{ to be a F-homomorphism (3.55)} \} \\ & \quad f \cdot \alpha \cdot F(\downarrow\alpha) = f \cdot \alpha \cdot F(\downarrow\alpha) \wedge f \cdot \alpha = \beta \cdot F f \\ & \Leftarrow \{ \text{simplify} \} \\ & \quad f \cdot \alpha = \beta \cdot F f \end{aligned}$$

The presentation of the *absorption* property of catamorphisms entails the very important issue of parameterization and deserves to be treated in a separate section, as follows.

3.12 Parameterization, type functors and cata-absorption

By analogy with what we have done about *splits* (product), *eithers* (coproduct) and *transposes* (exponential), we now look forward to identifying F-catamorphisms which exhibit functorial behaviour.

Suppose that one wishes to square all numbers which are members of lists of type \mathbb{T} (3.20). It can be checked that

$$\langle \langle \underline{Nil}, Cons \cdot (sq \times id) \rangle \rangle \quad (3.62)$$

will do this for us, where $\mathbb{N}_0 \xleftarrow{sq} \mathbb{N}_0$ is given by (3.38). This catamorphism, which converted to pointwise notation is nothing but function h which follows

$$\begin{cases} h Nil = Nil \\ h(Cons(a, l)) = Cons(sq a, h l) \end{cases}$$

maps type \mathbb{T} to itself. This is because sq maps \mathbb{N}_0 to \mathbb{N}_0 . Now suppose that, instead of sq , one would like to apply a given function $B \xleftarrow{f} \mathbb{N}_0$ (for some B other than \mathbb{N}_0) to all elements of the argument list. It is easy to see that it suffices to replace f for sq in (3.62). However, the output type no longer is \mathbb{T} , but rather type $\mathbb{T}' \cong 1 + B \times \mathbb{T}'$.

Types \mathbb{T} and \mathbb{T}' are very close to each other. They share the same “shape” (recursive pattern) and only differ with respect to the type of elements — \mathbb{N}_0 in \mathbb{T} and B in \mathbb{T}' . This suggests that these two types can be regarded as instances of a more generic list datatype `List`

$$\text{List } X \xleftarrow[\text{in}=[\underline{Nil}, Cons]]{\cong} 1 + X \times \text{List } X \quad (3.63)$$

in which the type of elements X is allowed to vary. Thus one has $\mathbb{T} = \text{List } \mathbb{N}_0$ and $\mathbb{T}' = \text{List } B$.

By inspection, it can be checked that, for every $B \xleftarrow{f} A$,

$$\langle \langle \underline{Nil}, Cons \cdot (f \times id) \rangle \rangle \quad (3.64)$$

maps `List A` to `List B`. Moreover, for $f = id$ one has:

$$\begin{aligned} & \langle \langle \underline{Nil}, Cons \cdot (id \times id) \rangle \rangle \\ = & \quad \{ \text{by the } \times\text{-functor-id property (2.29) and identity } \} \\ & \langle \langle \underline{Nil}, Cons \rangle \rangle \\ = & \quad \{ \text{cata-reflection (3.58)} \} \\ & id \end{aligned}$$

3.12. PARAMETERIZATION, TYPE FUNCTORS AND CATA-ABSORPTION 85

Therefore, by defining

$$\text{List } f \stackrel{\text{def}}{=} ([\text{Nil}, \text{Cons} \cdot (f \times id)])$$

what we have just seen can be written thus:

$$\text{List } id_A = id_{\text{List } A}$$

This is nothing but law (3.44) for F replaced by List . Moreover, it will not be too difficult to check that

$$\text{List } (g \cdot f) = \text{List } g \cdot \text{List } f$$

also holds — *cf.* (3.45). Altogether, this means that List can be regarded as a functor.

In programming terminology one says that $\text{List } X$ (the “lists of X ’s datatype”) is *parametric* and that, by instantiating parameter X , one gets ground lists such as lists of integers, booleans, *etc.* The illustration above deepens one’s understanding of parameterization by identifying the functorial behaviour of the parametric datatype along with its parameter instantiations.

All this can be broadly generalized and leads to what is commonly known by a *type functor*. First of all, it should be clear that the generic format

$$T \cong FT$$

adopted so far for the definition of an inductive type is not sufficiently detailed because it does not provide a parametric view of T . For simplicity, let us suppose (for the moment) that only one parameter is identified in T . Then we may factor this out via *type variable* X and write (overloading symbol T)

$$TX \cong B(X, TX)$$

where B is called the type’s *base functor*. Binary functor $B(X, Y)$ is given this name because it is the basis of the whole inductive type definition. By instantiation of X one obtains F . In the example above, $B(X, Y) = 1 + X \times Y$ and in fact $FY = B(\mathbb{N}_0, Y) = 1 + \mathbb{N}_0 \times Y$, recall (3.40). Moreover, one has

$$Ff = B(id, f) \tag{3.65}$$

and so every F -homomorphism can be written in terms of the base-functor of F , *e.g.*

$$f \cdot \alpha = \beta \cdot B(id, f)$$

instead of (3.55).

$\mathbb{T} X$ will be referred to as the *type functor* generated by B :

$$\underbrace{\mathbb{T} X}_{\text{type functor}} \cong \underbrace{B(X, \mathbb{T} X)}_{\text{base functor}}$$

We proceed to the description of its functorial behaviour — $\mathbb{T} f$ — for a given $B \xleftarrow{f} A$. As far as typing rules are concerned, we shall have

$$\frac{B \xleftarrow{f} A}{\mathbb{T} B \xleftarrow{\mathbb{T} f} \mathbb{T} A}$$

So we should be able to express $\mathbb{T} f$ as a $B(A, -)$ -catamorphism ($\llbracket g \rrbracket$):

$$\begin{array}{ccc} A & & \mathbb{T} A \xleftarrow{in_{\mathbb{T} A}} B(A, \mathbb{T} A) \\ f \downarrow & & \downarrow \mathbb{T} f = \llbracket g \rrbracket \quad \downarrow B(id, \mathbb{T} f) \\ B & & \mathbb{T} B \xleftarrow{g} B(A, \mathbb{T} B) \end{array}$$

As we know that $in_{\mathbb{T} B}$ is the standard constructor of values of type $\mathbb{T} B$, we may put it into the diagram too:

$$\begin{array}{ccc} A & & \mathbb{T} A \xleftarrow{in_{\mathbb{T} A}} B(A, \mathbb{T} A) \\ f \downarrow & & \downarrow \mathbb{T} f = \llbracket g \rrbracket \quad \downarrow B(id, \mathbb{T} f) \\ B & & \mathbb{T} B \xleftarrow{g} B(A, \mathbb{T} B) \\ & & \swarrow in_{\mathbb{T} B} \quad \dashrightarrow \\ & & B(B, \mathbb{T} B) \end{array}$$

The catamorphism's gene g will be synthesized by filling the dashed arrow in the diagram with the “obvious” $B(f, id)$, whereby one gets

$$\mathbb{T} f \stackrel{\text{def}}{=} \llbracket in_{\mathbb{T} B} \cdot B(f, id) \rrbracket \tag{3.66}$$

and a final diagram, where $in_{\mathbb{T} A}$ is abbreviated by in_A (ibid. $in_{\mathbb{T} B}$ by in_B):

$$\begin{array}{ccc} A & & \mathbb{T} A \xleftarrow{in_A} B(A, \mathbb{T} A) \\ f \downarrow & & \downarrow \mathbb{T} f = \llbracket in_B \cdot B(f, id) \rrbracket \quad \downarrow B(id, \mathbb{T} f) \\ B & & \mathbb{T} B \xleftarrow{in_B} B(B, \mathbb{T} B) \xleftarrow{B(f, id)} B(A, \mathbb{T} B) \end{array}$$

3.12. PARAMETERIZATION, TYPE FUNCTORS AND CATA-ABSORPTION 87

Next, we proceed to derive the useful law of *cata-absorption*

$$\llbracket g \rrbracket \cdot \mathbb{T} f = \llbracket g \cdot \mathbb{B}(f, id) \rrbracket \quad (3.67)$$

as a consequence of the laws studied in section 3.11. Our target is to show that, for $k = \llbracket g \rrbracket \cdot \mathbb{T} f$ in (3.57), one gets $\alpha = g \cdot \mathbb{B}(f, id)$:

$$\begin{aligned} & \llbracket g \rrbracket \cdot \mathbb{T} f = \llbracket \alpha \rrbracket \\ \Leftrightarrow & \quad \{ \text{type-functor definition (3.66)} \} \\ & \llbracket g \rrbracket \cdot \llbracket in_B \cdot \mathbb{B}(f, id) \rrbracket = \llbracket \alpha \rrbracket \\ \Leftarrow & \quad \{ \text{cata-fusion (3.61)} \} \\ & \llbracket g \rrbracket \cdot in_B \cdot \mathbb{B}(f, id) = \alpha \cdot \mathbb{B}(id, \llbracket g \rrbracket) \\ \Leftrightarrow & \quad \{ \text{cata-cancellation (3.59)} \} \\ & g \cdot \mathbb{B}(id, \llbracket g \rrbracket) \cdot \mathbb{B}(f, id) = \alpha \cdot \mathbb{B}(id, \llbracket g \rrbracket) \\ \Leftrightarrow & \quad \{ \mathbb{B} \text{ is a bi-functor (3.47)} \} \\ & g \cdot \mathbb{B}(id \cdot f, \llbracket g \rrbracket \cdot id) = \alpha \cdot \mathbb{B}(id, \llbracket g \rrbracket) \\ \Leftrightarrow & \quad \{ id \text{ is natural (2.11)} \} \\ & g \cdot \mathbb{B}(f \cdot id, id \cdot \llbracket g \rrbracket) = \alpha \cdot \mathbb{B}(id, \llbracket g \rrbracket) \\ \Leftrightarrow & \quad \{ (3.47) \text{ again, this time from left to right} \} \\ & g \cdot \mathbb{B}(f, id) \cdot \mathbb{B}(id, \llbracket g \rrbracket) = \alpha \cdot \mathbb{B}(id, \llbracket g \rrbracket) \\ \Leftarrow & \quad \{ \text{obvious} \} \\ & g \cdot \mathbb{B}(f, id) = \alpha \end{aligned}$$

The following diagram pictures this property of catamorphisms:

$$\begin{array}{ccccc} & & & & \xrightarrow{in_A} & & \mathbb{B}(A, \mathbb{T} A) \\ & & & & & & \downarrow \mathbb{B}(id, \mathbb{T} f) \\ & & & & & & \mathbb{T} C \\ & & & & & & \downarrow \mathbb{B}(id, \llbracket g \rrbracket) \\ & & & & & & \mathbb{B}(A, D) \\ & & & & & & \downarrow \mathbb{B}(f, id) \\ & & & & & & \mathbb{B}(C, D) \\ & & & & & & \downarrow g \\ & & & & & & D \end{array}$$

$\mathbb{T} A \xleftarrow{in_A} \mathbb{B}(A, \mathbb{T} A) \xrightarrow{\mathbb{B}(f, id)} \mathbb{B}(A, \mathbb{T} C) \xleftarrow{in_C} \mathbb{B}(C, \mathbb{T} C) \xrightarrow{\mathbb{B}(f, id)} \mathbb{B}(C, D) \xleftarrow{g} D$
 $\mathbb{T} A \xrightarrow{\mathbb{T} f} \mathbb{T} C \xrightarrow{\llbracket g \rrbracket} D$
 $\mathbb{B}(A, \mathbb{T} A) \xrightarrow{\mathbb{B}(id, \mathbb{T} f)} \mathbb{B}(A, \mathbb{T} C) \xrightarrow{\mathbb{B}(id, \llbracket g \rrbracket)} \mathbb{B}(A, D) \xrightarrow{\mathbb{B}(f, id)} \mathbb{B}(C, D) \xrightarrow{g} D$

It remains to show that (3.66) indeed defines a functor. This can be verified by checking properties (3.44) and (3.45) for $F = \mathbb{T}$:

- Property **type-functor-id**, cf. (3.44):

$$\begin{aligned}
& \top id \\
= & \quad \{ \text{by definition (3.66)} \} \\
& (\mathit{in}_B \cdot B(id, id)) \\
= & \quad \{ B \text{ is a bi-functor (3.46)} \} \\
& (\mathit{in}_B \cdot id) \\
= & \quad \{ \text{identity and cata-reflection (3.58)} \} \\
& id
\end{aligned}$$

- Property **type-functor**, cf. (3.45) :

$$\begin{aligned}
& \top (f \cdot g) \\
= & \quad \{ \text{by definition (3.66)} \} \\
& (\mathit{in}_B \cdot B(f \cdot g, id)) \\
= & \quad \{ id \cdot id = id \text{ and } B \text{ is a bi-functor (3.47)} \} \\
& (\mathit{in}_B \cdot B(f, id) \cdot B(g, id)) \\
= & \quad \{ \text{cata-absorption (3.67)} \} \\
& (\mathit{in}_B \cdot B(f, id)) \cdot \top g \\
= & \quad \{ \text{again cata-absorption (3.67)} \} \\
& (\mathit{in}_B) \cdot \top f \cdot \top g \\
= & \quad \{ \text{cata-reflection (3.58) followed by identity} \} \\
& \top f \cdot \top g
\end{aligned}$$

3.13 A catalogue of standard polynomial inductive types

The following table contains a collection of standard polynomial inductive types and associated base type bi-functors, which are in canonical form (3.53). The table contains two extra columns which may be used as bookmarks for equations (3.65) and (3.66), respec-

3.13. A CATALOGUE OF STANDARD POLYNOMIAL INDUCTIVE TYPES⁸⁹

tively ¹⁰:

Description	$\mathbb{T} X$	$\mathbb{B}(X, Y)$	$\mathbb{B}(id, f)$	$\mathbb{B}(f, id)$
“Right” Lists	List X	$1 + X \times Y$	$id + id \times f$	$id + f \times id$
“Left” Lists	LList X	$1 + Y \times X$	$id + f \times id$	$id + id \times f$
Non-empty Lists	NList X	$X + X \times Y$	$id + id \times f$	$f + f \times id$
Binary Trees	BTree X	$1 + X \times Y^2$	$id + id \times f^2$	$id + f \times id$
“Leaf” Trees	LTree X	$X + Y^2$	$id + f^2$	$f + id$

(3.68)

All type functors \mathbb{T} in this table are unary. In general, one may think of inductive datatypes which exhibit more than one type parameter. Should n parameters be identified in \mathbb{T} , then this will be based on an $n + 1$ -ary base functor \mathbb{B} , *cf.*

$$\mathbb{T}(X_1, \dots, X_n) \cong \mathbb{B}(X_1, \dots, X_n, \mathbb{T}(X_1, \dots, X_n))$$

So, every $n + 1$ -ary polynomial functor $\mathbb{B}(X_1, \dots, X_n, X_{n+1})$ can be identified as the basis of an inductive n -ary type functor (the convention is to stick to the canonical form and reserve the last variable X_{n+1} for the “recursive call”). While type bi-functors ($n = 2$) are often found in programming, the situation in which $n > 2$ is relatively rare. For instance, the combination of leaf-trees with binary-trees in (3.68) leads to the so-called “full tree” type bi-functor

Description	$\mathbb{T}(X_1, X_2)$	$\mathbb{B}(X_1, X_2, Y)$	$\mathbb{B}(id, id, f)$	$\mathbb{B}(f, g, id)$
“Full” Trees	FTree(X_1, X_2)	$X_1 + X_2 \times Y^2$	$id + id \times f^2$	$f + g \times id$

(3.69)

As we shall see later on, these types are widely used in programming. In the actual encoding of these types in HASKELL, exponentials are normally expanded to products according to (2.87), see for instance

```
data BTree a = Empty | Node(a, (BTree a, BTree a))
```

Moreover, one may chose to curry the type constructors as in, *e.g.*

```
data BTree a = Empty | Node a (BTree a) (BTree a)
```

Exercise 31. Write as a catamorphisms

- the function which counts the number of elements of a non-empty list (type NList in (3.68)).

¹⁰Since $(id_A)^2 = id_{(A^2)}$ one writes id^2 for id in this table.

- the function which computes the maximum element of a binary-tree of natural numbers.

□

Exercise 32. Characterize the function which is defined by $(\llbracket \square, h \rrbracket)$ for each of the following definitions of h :

$$h(x, (y_1, y_2)) = y_1 ++ [x] ++ y_2 \quad (3.70)$$

$$h = ++ \cdot (singl \times ++)$$
 (3.71)

$$h = ++ \cdot (++ \times singl) \cdot swap \quad (3.72)$$

assuming $singl\ a = [a]$. Identify in (3.68) which datatypes are involved as base functors.

□

Exercise 33. Write as a catamorphism the function which computes the *frontier* of a tree of type LTree (3.68), listed from left to right.

□

3.14 Functors and type functors in HASKELL

The concept of a (unary) functor is provided in HASKELL in the form of a particular class exporting the `fmap` operator:

```
class Functor f where
  fmap :: (a -> b) -> (f a -> f b)
```

So `fmap g` encodes `F g` once we declare `F` as an instance of class `Functor`. The most popular use of `fmap` has to do with HASKELL lists, as allowed by declaration

```
instance Functor [] where
  fmap f []      = []
  fmap f (x:xs) = f x : fmap f xs
```

in language's *Standard Prelude*.

In order to encode the type functors we have seen so far we have to do the same concerning their declaration. For instance, should we write

```
instance Functor BTree
  where fmap f =
        cataBTree ( inBTree . (id |- (f >< id)) )
```

concerning the binary-tree datatype of (3.68) and assuming appropriate declarations of `cataBTree` and `inBTree`, then `fmap` is overloaded and used across such binary-trees.

Bi-functors can be added easily by writing

```
class BiFunctor f where
  bmap :: (a -> b) -> (c -> d) -> (f a c -> f b d)
```

Exercise 34. Declare all datatypes in (3.68) in HASKELL notation and turn them into HASKELL type functors, that is, define `fmap` in each case.

□

Exercise 35. Declare datatype (3.69) in HASKELL notation and turn it into an instance of class `BiFunctor`.

□

3.15 The mutual-recursion law

The theory developed so far for building (and reasoning about) recursive functions doesn't cope with mutual recursion. As a matter of fact, the pattern of recursion of a given `cata(ana,hylo)morphism` involves only the recursive function being defined (even though more than once, in general, as dictated by the relevant base functor). As a counter-example, consider the following pair of mutually dependent functions over \mathbb{N}_0 , written in HASKELL notation:

$$\begin{aligned} f\ 0 &= 1 \\ f(n+1) &= g\ n \\ \\ g\ 0 &= 1 \\ g(n+1) &= f\ n + g\ n \end{aligned}$$

Note that the recursive branch of f delegates the evaluation to g and that that of g does the same with respect to f . In general, a system of n mutually recursive functions

$$\begin{cases} f_1 = \phi_1(f_1, \dots, f_n) \\ \vdots \\ f_n = \phi_n(f_1, \dots, f_n) \end{cases}$$

involves n functions and n functional combinators ϕ_1, \dots, ϕ_n which can be of different “shape”. Below we address the special case in which all functions are defined over the same inductive datatype, as is the case of f and g above. With no loss of generality, we stick to pairs of functions, the extension to $n > 2$ functions being straightforward.

Since we want to reason at pointfree level, our first step is to write the pointfree-transforms of f and g , respectively

$$\begin{aligned} f \cdot [\underline{0}, \text{suc}] &= [\underline{1}, g] \\ g \cdot [\underline{0}, \text{suc}] &= [\underline{1}, + \cdot \langle f, g \rangle] \end{aligned}$$

whose underlying inductive type

$$\mathbb{N}_0 \begin{array}{c} \xrightarrow{\quad} \\ \cong \\ \xleftarrow{\quad} \\ \text{in} = [\underline{0}, \text{suc}] \end{array} \underbrace{1 + \mathbb{N}_0}_{F \mathbb{N}_0} \quad (3.73)$$

is such that $F f = id + f$. The f, g mutual dependence can be made more explicit by forcing f to depend also on itself, via \times -cancellation:

$$\begin{aligned} f \cdot [\underline{0}, \text{suc}] &= [\underline{1}, \pi_2 \cdot \langle f, g \rangle] \\ g \cdot [\underline{0}, \text{suc}] &= [\underline{1}, + \cdot \langle f, g \rangle] \end{aligned}$$

Reverse $+$ -absorption will further enable us to rewrite the above into

$$\begin{aligned} f \cdot \text{in} &= [\underline{1}, \pi_2] \cdot F \langle f, g \rangle \\ g \cdot \text{in} &= [\underline{1}, +] \cdot F \langle f, g \rangle \end{aligned}$$

leading to diagrams

$$\begin{array}{ccc} \mathbb{N}_0 & \xleftarrow{\text{in}} & F \mathbb{N}_0 \\ f \downarrow & & \downarrow F \langle f, g \rangle \\ \mathbb{N} & \xleftarrow{[\underline{1}, \pi_2]} & F(\mathbb{N} \times \mathbb{N}) \end{array} \quad \begin{array}{ccc} \mathbb{N}_0 & \xleftarrow{\text{in}} & F \mathbb{N}_0 \\ g \downarrow & & \downarrow F \langle f, g \rangle \\ \mathbb{N} & \xleftarrow{[\underline{1}, +]} & F(\mathbb{N} \times \mathbb{N}) \end{array}$$

Individually, none of these two functions is a catamorphism, since the recursive call involves $\langle f, g \rangle$, instead of f or g alone. This suggests that, perhaps, the combined function

$\langle f, g \rangle$ is expressible as a catamorphism. This is indeed true and is captured by the so-called *mutual-recursion law*, also referred to as “Fokkinga’s law”. This law states that, in general, given inductive type T and two functions

$$\begin{array}{ccc} T & \xleftarrow{in} & F T \\ f \downarrow & & \downarrow F \langle f, g \rangle \\ A & \xleftarrow{h} & F(A \times B) \end{array} \quad \begin{array}{ccc} T & \xleftarrow{in} & F T \\ g \downarrow & & \downarrow F \langle f, g \rangle \\ B & \xleftarrow{k} & F(A \times B) \end{array}$$

then $\langle f, g \rangle$ is catamorphism

$$\begin{array}{ccc} T & \xleftarrow{in} & F T \\ \langle f, g \rangle \downarrow & & \downarrow F \langle f, g \rangle \\ A \times B & \xleftarrow{\langle h, k \rangle} & F(A \times B) \end{array}$$

In symbols, the mutual recursion law is stated as follows

$$\begin{cases} f \cdot in = h \cdot F \langle f, g \rangle \\ g \cdot in = k \cdot F \langle f, g \rangle \end{cases} \equiv \langle f, g \rangle = \langle \langle h, k \rangle \rangle \quad (3.74)$$

and is easy to check:

$$\begin{aligned} & \langle f, g \rangle = \langle \langle h, k \rangle \rangle \\ \equiv & \quad \{ \text{cata-universal (3.57)} \} \\ & \langle f, g \rangle \cdot in = \langle h, k \rangle \cdot F \langle f, g \rangle \\ \equiv & \quad \{ \times\text{-fusion (2.24) twice} \} \\ & \langle f \cdot in, g \cdot in \rangle = \langle h \cdot F \langle f, g \rangle, k \cdot F \langle f, g \rangle \rangle \\ \equiv & \quad \{ (2.56) \} \\ & f \cdot in = h \cdot F \langle f, g \rangle \quad \wedge \quad g \cdot in = k \cdot F \langle f, g \rangle \end{aligned}$$

Altogether, since $f = \pi_1 \cdot \langle f, g \rangle$ and $g = \pi_2 \cdot \langle f, g \rangle$, one is able to apply catamorphism theory to f and g *indirectly* through catamorphism $\langle f, g \rangle$.

Let us see the application of this law to our starting example, that is, for $h = [\underline{1}, \pi_2]$ and $k = [\underline{1}, +]$. We get:

$$\begin{aligned} \langle f, g \rangle &= \langle \langle [\underline{1}, \pi_2], [\underline{1}, +] \rangle \rangle \\ \equiv & \quad \{ \text{exchange law (2.47)} \} \end{aligned}$$

$$\begin{aligned}
\langle f, g \rangle &= ([\langle \underline{1}, \underline{1} \rangle, \langle \pi_2, + \rangle]) \\
\equiv & \{ \text{going pointwise and denoting } \langle f, g \rangle \text{ by } fg \} \\
& \begin{cases} fg\ 0 = (1, 1) \\ fg\ (n + 1) = (y, x + y) \text{ where } (x, y) = fg\ n \end{cases}
\end{aligned}$$

(It can be observed that $f = \pi_2 \cdot fg$ is a version of the Fibonacci function.)

As another example of application of (3.74), consider the following recursive predicate which checks whether a (non-empty) list is ordered,

$$\begin{aligned}
ord : A^+ &\longrightarrow 2 \\
ord\ [a] &= \text{TRUE} \\
ord\ (\text{cons}(a, l)) &= a \geq (\text{listMax}\ l) \wedge (ord\ l)
\end{aligned}$$

where \geq is assumed to be a total order on datatype A and

$$\text{listMax} = ([\text{id}, \text{max}]) \tag{3.75}$$

computes the greatest element of a given list of A s:

$$\begin{array}{ccc}
A^+ & \xleftarrow{[\text{singl}, \text{cons}]} & A + A \times A^+ \\
\text{listMax} \downarrow & & \downarrow \text{id} + \text{id} \times \text{listMax} \\
A & \xleftarrow{[\text{id}, \text{max}]} & A + A \times A
\end{array}$$

(In the diagram, $\text{singl}\ a = [a]$.)

Predicate ord is not a catamorphism because of the presence of $\text{listMax}\ l$ in the recursive branch. However, the following diagram depicting ord

$$\begin{array}{ccc}
A^+ & \xleftarrow{[\text{singl}, \text{cons}]} & A + A \times A^+ \\
ord \downarrow & & \downarrow \text{id} + \text{id} \times (\text{listMax}, ord) \\
2 & \xleftarrow{[\text{TRUE}, \alpha]} & A + A \times (A \times 2)
\end{array}$$

(where $\alpha(a, (m, b)) \stackrel{\text{def}}{=} a \geq m \wedge b$) suggests the possibility of using the mutual recursion law. One only has to find a way of letting listMax depend also on ord , which isn't difficult: for any $A^+ \xrightarrow{g} B$, one has

$$\begin{array}{ccc}
A^+ & \xleftarrow{[\text{singl}, \text{cons}]} & A + A \times A^+ \\
\text{listMax} \downarrow & & \downarrow \text{id} + \text{id} \times (\text{listMax}, g) \\
A & \xleftarrow{[\text{id}, \text{max} \cdot (\text{id} \times \pi_1)]} & A + A \times (A \times B)
\end{array}$$

where the extra presence of g is cancelled by projection π_1 .

For $B = 2$ and $g = ord$ we are in position to apply Fokkinga's law and obtain:

$$\begin{aligned} \langle listMax, ord \rangle &= (\langle [id, max \cdot (id \times \pi_1)], [TRUE, \alpha] \rangle) \\ &= \{ \text{exchange law (2.47)} \} \\ &(\langle [id, TRUE], \langle max \cdot (id \times \pi_1), \alpha \rangle \rangle) \end{aligned}$$

Of course, $ord = \pi_2 \cdot \langle listMax, ord \rangle$. By denoting the above synthesized catamorphism by aux , we end up with the following version of ord :

$$ord\ l = \begin{array}{l} \text{let } (a, b) = aux\ l \\ \text{in } b \end{array}$$

where

$$\begin{aligned} aux &: A^+ \longrightarrow A \times 2 \\ aux\ [a] &= (a, TRUE) \\ aux\ (cons(a, l)) &= \begin{array}{l} \text{let } (m, b) = aux\ l \\ \text{in } (max(a, m), (a > m \wedge b)) \end{array} \end{aligned}$$

3.15.1 “Banana-split”: a corollary of the mutual-recursion law

Let $h = i \cdot F \pi_1$ and $k = j \cdot F \pi_2$ in (3.74). Then

$$\begin{aligned} f \cdot in &= (i \cdot F \pi_1) \cdot F \langle f, g \rangle \\ \equiv & \{ \text{composition is associative and } F \text{ is a functor} \} \\ f \cdot in &= i \cdot F (\pi_1 \cdot \langle f, g \rangle) \\ \equiv & \{ \text{by } \times\text{-cancellation (2.20)} \} \\ f \cdot in &= i \cdot F f \\ \equiv & \{ \text{by cata-cancellation} \} \\ f &= \langle i \rangle \end{aligned}$$

Similarly, from $k = j \cdot F \pi_2$ we get

$$g = \langle j \rangle$$

Then, from (3.74), we get

$$\langle \langle i \rangle, \langle j \rangle \rangle = \langle \langle i \cdot F \pi_1, j \cdot F \pi_2 \rangle \rangle$$

that is

$$\langle \langle i \rangle, \langle j \rangle \rangle = \langle (i \times j) \cdot \langle F \pi_1, F \pi_2 \rangle \rangle \quad (3.76)$$

by (reverse) \times -absorption (2.25).

This law provides us with a very useful tool for “parallel loop” inter-combination: “loops” $\langle i \rangle$ and $\langle j \rangle$ are fused together into a single “loop” $\langle (i \times j) \cdot \langle F \pi_1, F \pi_2 \rangle \rangle$. The need for this kind of calculation arises very often. Consider, for instance, the function which computes the average of a non-empty list of natural numbers,

$$average \stackrel{\text{def}}{=} (/) \cdot \langle sum, length \rangle \quad (3.77)$$

where sum and $length$ are the expected \mathbb{N}^+ catamorphisms:

$$\begin{aligned} sum &= \langle [id, +] \rangle \\ length &= \langle [\underline{1}, succ \cdot \pi_2] \rangle \end{aligned}$$

As defined by (3.77), function $average$ performs two independent traversals of the argument list before division $(/)$ takes place. Banana-split will fuse such two traversals into a single one (see function aux below), thus leading to a function which will run “twice as fast”:

$$\begin{aligned} average\ l &= x/y \\ \text{where } (x, y) &= aux\ l \\ aux[a] &= (a, 1) \\ aux(cons(a, l)) &= (a + x, y + 1) \\ \text{where } (x, y) &= aux\ l \end{aligned} \quad (3.78)$$

Exercise 36. Calculate (3.78) from (3.77). Which of these two versions of the same function is easier to understand?

□

3.16 Inductive datatype isomorphism

not yet available

3.17 Bibliography notes

It is often the case that the expressive power of a particular programming language or paradigm is counter-productive in the sense that too much freedom is given to programmers. Sooner or later, these will end up writing unintelligible (authorship dependent) code which will become a burden to whom has to maintain it. Such has been the case of imperative programming in the past (inc. assembly code), where the unrestricted use of `goto` instructions eventually gave place to `if-then-else`, `while` and `repeat structured` programming constructs.

A similar trend has been observed over the last decades at a higher programming level: arbitrary recursion and/or (side) effects have been considered harmful in functional programming. Instead, programmers have been invited to structure their code around generic program devices such as eg. *fold/unfold* combinators, which bring discipline to recursion. One witnesses progress in the sense that the loss of freedom is balanced by the increase of formal semantics and the availability of program calculi.

Such disciplined programming combinators have been extended from list-processing to other inductive structures thanks to one of the most significant advances in programming theory over the last decade: the so-called *functorial* approach to datatypes which originated mainly from [MA86], was popularized by [Mal90] and reached textbook format in [BdM97]. A comfortable basis for exploiting *polymorphism* [Wad89], the “datatypes as functors” moto has proved beneficial at a higher level of abstraction, giving birth to *polytypism* [JJ96].

The literature on *anas*, *catas* and *hylos* is vast (see eg. [MH95], [JJ98], [GHA01]) and it is part of a broader discipline which has become known as the *mathematics of program construction* [Bac04]. This chapter provides an introduction to such a discipline. Only the calculus of catamorphisms has been presented. The corresponding theory of anamorphisms and hylomorphisms demands further mathematical machinery and will won't be dealt with before chapters 5 and 7.