When is a function a fold or an unfold?

Jeremy Gibbons
University of Oxford

Graham Hutton
University of Nottingham

Thorsten Altenkirch
University of Nottingham

April 29, 2002

Abstract

We give a necessary and sufficient condition for when a set-theoretic function can be written using the recursion operator fold, and a dual condition for the recursion operator unfold. The conditions are simple, practically useful, and generic in the underlying datatype.

1 Introduction

The recursion operator fold encapsulates a common pattern for defining programs that *consume* values of a *least* fixpoint type such as finite lists. Dually, the recursion operator unfold encapsulates a common pattern for defining programs that *produce* values of a *greatest* fixpoint type such as infinite lists or streams. Theory and applications of fold abound — see [11, 4] for recent surveys — while in recent years it has become increasingly clear that the less well-known concept of unfold is just as useful [5, 6, 10, 12, 14].

Given the interest in fold and unfold, it is natural to ask when a program can be written using one of these operators. Surprisingly little is known about this question. This article gives a complete answer for the special case in which programs are total functions between sets. In particular, we give a necessary and sufficient condition for when a set-theoretic function can be written using fold, and a dual condition for unfold. The conditions are simple, practically useful, and generic in the underlying datatype. However, our proofs are set-theoretic, and make essential use of the Axiom of Choice; hence the result does not generalize to categories of constructive functions[1].

2 Fold and unfold

In this section we review the categorical treatment of fold and unfold in terms of initial algebras and final coalgebras; for further details see [17, 19, 13, 1].

Suppose that we fix a category $\mathcal{C}$ and a functor $F : \mathcal{C} \to \mathcal{C}$. An *algebra* is pair (A, f) comprising an object A and an arrow $f : F\,A \to A$, and a *homomorphism* $h : (A, f) \to (B, g)$ from one such algebra to another is an arrow $h : A \to B$ such that the following square commutes:

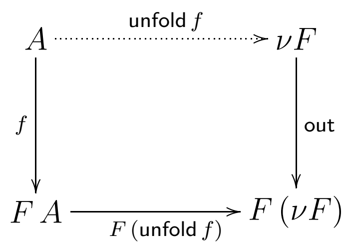
$$\begin{array}{ccc} F\,A & \xrightarrow{F\,h} & F\,B \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle g} \\ A & \xrightarrow[h]{} & B \end{array}$$

An *initial algebra* is an initial object in the category with algebras as objects and homomorphisms as arrows. We write $(\mu F, \mathsf{in})$ for an initial algebra, and $\mathsf{fold}\,f$ for the unique homomorphism $h : (\mu F, \mathsf{in}) \to (A, f)$ from the initial algebra to any other algebra (A, f). That is, $\mathsf{fold}\,f$ is defined as the unique arrow that makes the following square commute:

$$\begin{array}{ccc} F\,(\mu F) & \xrightarrow{F\,(\mathsf{fold}\,f)} & F\,A \\ {\scriptstyle \mathsf{in}}\downarrow & & \downarrow{\scriptstyle f} \\ \mu F & \xrightarrow[\mathsf{fold}\,f]{} & A \end{array}$$

The dual notions of *coalgebra*, *cohomomorphism*, and *terminal coalgebra* are defined similarly. We write $(\nu F, \mathsf{out})$ for a terminal coalgebra, and $\mathsf{unfold}\,f$ for the unique cohomomorphism $h : (A, f) \to (\nu F, \mathsf{out})$ from any coalgebra (A, f) to the terminal coalgebra. That is, $\mathsf{unfold}\,f$ is defined as the unique arrow that makes the following square commute:

[1] Such as the effective topos or the category of ω-sets.



In the literature, $\mathsf{fold}\, f$ and $\mathsf{unfold}\, f$ are sometimes written as $(\!| f |\!)$ and $[\!(f)\!]$, and called *catamorphisms* and *anamorphisms* respectively.

2.1 Example: finite lists

Suppose that we define a functor $L : \mathcal{SET} \to \mathcal{SET}$ by $L\,A = \mathbf{1} + (\mathbb{N} \times A)$ and $L\,f = \mathsf{id}_{\mathbf{1}} + (\mathsf{id}_{\mathbb{N}} \times f)$, where $\mathbb{N}$ is the set of natural numbers. Then an algebra is a pair (A, f) comprising a set A and a function $f : \mathbf{1} + (\mathbb{N} \times A) \to A$. Functions of this type can always be uniquely decomposed into the form $f = [g, h]$ for some other functions $g : \mathbf{1} \to A$ and $h : \mathbb{N} \times A \to A$. A homomorphism $f : (A, [g, h]) \to (B, [i, j])$ is a function $f : A \to B$ such that $f \cdot g = i$ and $f \cdot h = j \cdot (\mathsf{id}_{\mathbb{N}} \times f)$.

The functor L has an initial algebra defined by $(\mu L, \mathsf{in}) = (\mathit{List}(\mathbb{N}), [\mathit{nil}, \mathit{cons}])$, where $\mathit{List}(A)$ is the set of all finite lists with elements drawn from A, and $\mathit{nil} : \mathbf{1} \to \mathit{List}(\mathbb{N})$ and $\mathit{cons} : \mathbb{N} \times \mathit{List}(\mathbb{N}) \to \mathit{List}(\mathbb{N})$ are constructor functions for this set. Given any other set A and two functions $i : \mathbf{1} \to A$ and $j : \mathbb{N} \times A \to A$, the function $\mathsf{fold}\,[i, j] : \mathit{List}(\mathbb{N}) \to A$ is uniquely defined by the following two equations:

$$
\begin{array}{lcl}
\mathsf{fold}\,[i, j] \cdot \mathit{nil} & = & i \\
\mathsf{fold}\,[i, j] \cdot \mathit{cons} & = & j \cdot (\mathsf{id}_{\mathbb{N}} \times \mathsf{fold}\,[i, j])
\end{array}
$$

That is, $\mathsf{fold}\,[i, j]$ processes a list by replacing the *nil* constructor at the end of the list by the function i, and each *cons* constructor within the list by the function j. For example, the function $\mathit{sum} : \mathit{List}(\mathbb{N}) \to \mathbb{N}$ that sums a list of naturals can be defined by $\mathit{sum} = \mathsf{fold}\,[\mathit{zero}, \mathit{plus}]$, where $\mathit{zero} : \mathbf{1} \to \mathbb{N}$ and $\mathit{plus} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ are given by $\mathit{zero}\,() = 0$ and $\mathit{plus}\,(x, y) = x + y$.

We will use this datatype in examples later. For notational simplicity, we will write '$[\,]$' for $\mathit{nil}\,()$, and '$x : xs$' for $\mathit{cons}\,(x, xs)$. Thus, we might have written the above definition of fold more perspicuously as:

$$
\begin{array}{lcl}
(\mathsf{fold}\,[i, j])\,[\,] & = & i \\
(\mathsf{fold}\,[i, j])\,(x : xs) & = & j\,(x, (\mathsf{fold}\,[i, j])\,xs)
\end{array}
$$

2.2 Example: infinite lists

Suppose that we define a functor $S : \mathcal{SET} \rightarrow \mathcal{SET}$ by $SA = \mathbb{N} \times A$ and $Sf = \mathsf{id}_{\mathbb{N}} \times f$. Then a coalgebra is a pair (A, f) comprising a set A and a function $f : A \rightarrow \mathbb{N} \times A$. Functions of this type can always be uniquely decomposed into the form $f = \langle g, h \rangle$ for some other functions $g : A \rightarrow \mathbb{N}$ and $h : A \rightarrow A$. A cohomomorphism $f : (A, \langle g, h \rangle) \rightarrow (B, \langle i, j \rangle)$ is a function $f : A \rightarrow B$ such that $i \cdot f = g$ and $j \cdot f = f \cdot h$.

The functor S has a terminal coalgebra $(\nu S, \mathsf{out}) = (\mathit{Stream}(\mathbb{N}), \langle \mathit{head}, \mathit{tail} \rangle)$, where $\mathit{Stream}(A)$ is the set of all streams (infinite lists) with elements drawn from A, and $\mathit{head} : \mathit{Stream}(\mathbb{N}) \rightarrow \mathbb{N}$ and $\mathit{tail} : \mathit{Stream}(\mathbb{N}) \rightarrow \mathit{Stream}(\mathbb{N})$ are destructor functions for this set. Given any other set A and two functions $g : A \rightarrow \mathbb{N}$ and $h : A \rightarrow A$, the function $\mathsf{unfold}\, \langle g, h \rangle : A \rightarrow \mathit{Stream}(\mathbb{N})$ is uniquely defined by the following two equations:

$$
\begin{array}{lll}
\mathit{head} \cdot \mathsf{unfold}\, \langle g, h \rangle & = & g \\
\mathit{tail} \cdot \mathsf{unfold}\, \langle g, h \rangle & = & \mathsf{unfold}\, \langle g, h \rangle \cdot h
\end{array}
$$

That is, $\mathsf{unfold}\, \langle g, h \rangle$ produces a stream by using the function g to produce the *head* of the stream, and the function h to generate another value that is then itself unfolded in the same way to produce the *tail* of the stream. For example, the function $\mathit{from} : \mathbb{N} \rightarrow \mathit{Stream}(\mathbb{N})$, which produces a stream of naturals ascending in steps of one, can be defined by $\mathit{from} = \mathsf{unfold}\, \langle \mathsf{id}_{\mathbb{N}}, \mathit{succ} \rangle$ where $\mathit{succ} : \mathbb{N} \rightarrow \mathbb{N}$ is given by $\mathit{succ}\, x = x + 1$.

3 When is an arrow a fold or an unfold?

The fold operator encapsulates a common pattern for defining an arrow of type $\mu F \rightarrow A$. It is natural then to ask when an arrow of this type can be written using fold. More precisely, when can an arbitrary arrow $h : \mu F \rightarrow A$ be written in the form $h = \mathsf{fold}\, f$ for some other arrow $f : FA \rightarrow A$?

A technically complete, but nonetheless unsatisfactory, answer to this question is provided by the universal property of the fold operator [17], which can be stated as the following equivalence:

$$h = \mathsf{fold}\, f \quad \Leftrightarrow \quad h \cdot \mathsf{in} = f \cdot F\, h$$

The $\Rightarrow$ direction of this equivalence states that $\mathsf{fold}\, f$ is a homomorphism from the initial algebra $(\mu F, \mathsf{in})$ to another algebra (A, f), while the $\Leftarrow$ direction states that any other homomorphism h between these two algebras must be equal to $\mathsf{fold}\, f$. Taken as a whole, the universal property expresses the fact that $\mathsf{fold}\, f$ is the unique homomorphism from $(\mu F, \mathsf{in})$ to (A, f).

The universal property provides a complete answer to our question — h can be written in the form $\mathsf{fold}\, f$ precisely when $h \cdot \mathsf{in} = f \cdot F\, h$ — but is less helpful than it might be because it requires that we already know f. Given a specific h, however, the universal property can often be used to guide the construction of an appropriate f [11], but we do not consider this a completely satisfactory answer either, because this approach is only a heuristic, and it is sometimes difficult to apply in practice.

The problem with the universal property is that it concerns an *intensional* aspect of h, namely the function f that forms part of its implementation. Often a condition based on purely *extensional* aspects is more useful. A partial answer to our question with purely extensional concerns is that every left invertible arrow $h : \mu F \to A$ can be written using fold [19]. Formally, if we assume that there exists an arrow $g : A \to \mu F$ such that $g \cdot h = \mathsf{id}_{\mu F}$, then the equation $h = \mathsf{fold}\, f$ can be solved for f as follows:

$$
\begin{array}{ll}
 & h \;=\; \mathsf{fold}\, f \\
\Leftrightarrow & \quad \{ \text{ universal property } \} \\
 & h \cdot \mathsf{in} \;=\; f \cdot F\, h \\
\Leftrightarrow & \quad \{ \text{ identities } \} \\
 & h \cdot \mathsf{in} \cdot \mathsf{id}_{F\,(\mu F)} \;=\; f \cdot F\, h \\
\Leftrightarrow & \quad \{ \text{ functors } \} \\
 & h \cdot \mathsf{in} \cdot F\,(\mathsf{id}_{\mu F}) \;=\; f \cdot F\, h \\
\Leftrightarrow & \quad \{ \text{ assumption } \} \\
 & h \cdot \mathsf{in} \cdot F\,(g \cdot h) \;=\; f \cdot F\, h \\
\Leftrightarrow & \quad \{ \text{ functors } \} \\
 & h \cdot \mathsf{in} \cdot F\, g \cdot F\, h \;=\; f \cdot F\, h \\
\Leftarrow & \quad \{ \text{ composition } \} \\
 & f \;=\; h \cdot \mathsf{in} \cdot F\, g
\end{array}
$$

In summary, we have derived the following implication:

$$
g \cdot h \;=\; \mathsf{id}_{\mu F} \quad \Rightarrow \quad h \;=\; \mathsf{fold}\,(h \cdot \mathsf{in} \cdot F\, g)
$$

As an example, the function $rev : List(\mathbb{N}) \to List(\mathbb{N})$ that reverses a list is its own inverse ($rev \cdot rev = \mathsf{id}_{List(\mathbb{N})}$), and hence it is immediate that rev can be written using fold by the above implication. Note, however, that this implication only provides a partial answer to our question, because the converse is not true in general. That is, not every arrow $h : \mu F \to A$ that can be written using fold is left invertible. For example, the function $\mathit{sum} : \mathit{List}(\mathbb{N}) \to \mathbb{N}$ was written using fold in the previous section, but is not left invertible.

Dually, the unfold operator also satisfies a universal property, which can be used to show that every right invertible arrow of type $A \to \nu F$ can be written using unfold [19]. For example, the function $\mathit{evenpos} : \mathit{Stream}(\mathbb{N}) \to \mathit{Stream}(\mathbb{N})$ that removes every other element from a stream has a right inverse (any function that inserts an element between each adjacent pair in a stream), and hence it is immediate that *evenpos* can be written using unfold. However, not every arrow $h : A \to \nu F$ that can be written using unfold is right invertible. For example, the function $\mathit{from} : \mathbb{N} \to \mathit{Stream}(\mathbb{N})$ was written using unfold in the previous section, but is not right invertible.

As far as we are aware, the invertibility results above are the only known results that state when arbitrary arrows of the correct type can be written using fold or unfold. We conclude this section by noting that much more progress has been made concerning specific kinds of arrows. For example, the fusion law states that the composition of a homomorphism and a fold can always be written as a fold, while the banana split law states that two folds applied to the same argument can always be written as a single fold [19].

4 When is a function a fold?

In this section we give a necessary and sufficient condition for when an arrow can be written using fold, for the special case of the category $\mathcal{SET}$ in which the arrows are total functions between sets. We dualize the result for unfolds in the following section.

The result depends on the following definition:

Definition 4.1 The *kernel* [16] of a function $f : A \to B$ is the set of pairs of elements that are identified by f:

$$\mathsf{ker}\, f \quad = \quad \{\ (a, a') \in A \times A \mid f\, a = f\, a'\ \}$$

The main result of this section is a necessary and sufficient condition for when an arbitrary arrow $h : \mu F \to A$ in $\mathcal{SET}$ can be written in the form $h = \mathsf{fold}\, f$ for some other arrow $f : F\, A \to A$.

Theorem 4.2 Suppose that $h : \mu F \to A$. Then

$$(\exists g : F\,A \to A.\quad h = \mathsf{fold}\,g)$$
$$\Leftrightarrow\quad (\mathsf{ker}\,(F\,h) \subseteq \mathsf{ker}\,(h \cdot \mathsf{in}))$$

The crux of the proof is the well-known observation that inclusion of kernels is equivalent to the existence of 'postfactors':

Lemma 4.3 Suppose that $f : A \to B$ and $h : A \to C$. Then

$$(\exists g : B \to C.\quad h = g \cdot f) \quad\Leftrightarrow\quad \mathsf{ker}\,f \subseteq \mathsf{ker}\,h \;\wedge\; B \to C \neq \emptyset$$

Proof The proof is straightforward. For the left-to-right direction, assume that $g : B \to C$ and $h = g \cdot f$; then clearly $B \to C \neq \emptyset$, and moreover,

$$
\begin{array}{ll}
 & (a, a') \in \mathsf{ker}\,f \\
\Leftrightarrow & \quad \{\ \text{kernels}\ \} \\
 & f\,a = f\,a' \\
\Rightarrow & \quad \{\ \text{Leibniz}\ \} \\
 & g\,(f\,a) = g\,(f\,a') \\
\Leftrightarrow & \quad \{\ \text{assumption}\ \} \\
 & h\,a = h\,a' \\
\Leftrightarrow & \quad \{\ \text{kernels}\ \} \\
 & (a, a') \in \mathsf{ker}\,h
\end{array}
$$

Conversely, assume that $\mathsf{ker}\,f \subseteq \mathsf{ker}\,h$ and $B \to C \neq \emptyset$, so that either $B = \emptyset$ or $C \neq \emptyset$. When $B = \emptyset$, let g be the unique function in $B \to C$; note that g is the 'empty function', and so $g \cdot f$ is empty too. Moreover, $A = \emptyset$ because of the type of f, so h is also the empty function and hence equal to $g \cdot f$. When $C \neq \emptyset$, we define $g\,b$ for b in the range of f by $g\,b = h\,a$ for some a with $f\,a = b$; this is a proper definition, because if there are two choices a, a' with $f\,a = f\,a' = b$, then $h\,a = h\,a'$ also by assumption. For b outside the range of f, we define $g\,b$ arbitrarily. By construction, this gives $h\,a = g\,(f\,a)$ for every a. □

Lemma 4.4

$$\mu F \to A \neq \emptyset \Rightarrow F\,A \to A \neq \emptyset$$

Proof We note that

$$F\,A \to A \neq \emptyset \Leftrightarrow (A = \emptyset \Rightarrow F\,A = \emptyset)$$

and show that the implication $A = \emptyset \Rightarrow F\,A = \emptyset$ holds:

$$
\begin{array}{ll}
& A = \emptyset \\
\Rightarrow & \quad \{ \ \mu F \to A \neq \emptyset \ \} \\
& \mu F = \emptyset \\
\Rightarrow & \quad \{ \ \mathsf{in} : F\,\mu F \to \mu F \ \} \\
& F\,\mu F = \emptyset \\
\Rightarrow & \quad \{ \ \mu F = \emptyset = A \ \} \\
& F\,A = \emptyset
\end{array}
$$

□

Proof of Theorem 4.2 Given the crucial lemma above, the proof of the theorem is almost embarrassingly simple:

$$
\begin{array}{ll}
& \exists g : F\,A \to A. \quad h = \mathsf{fold}\,g \\
\Leftrightarrow & \quad \{ \ \text{universal property} \ \} \\
& \exists g : F\,A \to A. \quad h \cdot \mathsf{in} = g \cdot F\,h \\
\Leftrightarrow & \quad \{ \ \text{Lemma 4.3} \ \} \\
& \mathsf{ker}\,(F\,h) \subseteq \mathsf{ker}\,(h \cdot \mathsf{in}) \ \wedge \ F\,A \to A \neq \emptyset \\
\Leftrightarrow & \quad \{ \ \text{Lemma 4.4 with } h : \mu F \to A \neq \emptyset \ \} \\
& \mathsf{ker}\,(F\,h) \subseteq \mathsf{ker}\,(h \cdot \mathsf{in})
\end{array}
$$

□

Remark 4.5 Note that use of Lemma 4.3 in the proof of Theorem 4.2 actually involves constructing the body of the fold. We therefore have a similar result, but explicitly mentioning the fold body: suppose that $h : \mu F \to A$; then

$$h = \mathsf{fold}\,(h \cdot \mathsf{in} \cdot g) \quad \Leftarrow \quad \mathsf{ker}\,(F\,h) \subseteq \mathsf{ker}\,(h \cdot \mathsf{in}) \ \wedge \ F\,A \to A \neq \emptyset$$

where $g : F\,A \to F\,\mu F$ is such that $h \cdot \mathsf{in} \cdot g \cdot F\,h$. However, this is an implication, not an equivalence; in particular, it cannot be used to show that a certain h cannot be expressed as a fold.

Remark 4.6 For the type $\mathit{List}(A)$ of finite lists with elements drawn from A, with constructors $\mathit{nil} : \mathbf{1} \to \mathit{List}(A)$ and $\mathit{cons} : A \times \mathit{List}(A) \to \mathit{List}(A)$, Theorem 4.2 reduces to stating that an arbitrary function $h : \mathit{List}(A) \to B$ can be written directly as a fold precisely when the lists that are identified by h are closed under cons, in the sense that for all x,

$$h\,\mathit{xs} \ = \ h\,\mathit{ys} \quad \Rightarrow \quad h\,(x : \mathit{xs}) \ = \ h\,(x : \mathit{ys})$$

Example 4.7 The function $sum : List(\mathbb{N}) \to \mathbb{N}$ of Section 2.1 satisfies the equations

$$\begin{array}{lcl} sum\,[\,] & = & 0 \\ sum\,(x : xs) & = & x + sum\,xs \end{array}$$

A simple calculation verifies that the lists identified by sum are closed under $cons$:

$$\begin{array}{ll} & sum\,(x : xs) \;=\; sum\,(x : ys) \\ \Leftrightarrow & \quad \{ \text{ definition of } sum \text{ } \} \\ & x + sum\,xs \;=\; x + sum\,ys \\ \Leftarrow & \quad \{ \text{ extensionality } \} \\ & sum\,xs \;=\; sum\,ys \end{array}$$

and hence sum can be written directly using fold.

Example 4.8 In contrast, if we define a function $stail : List(\mathbb{N}) \to List(\mathbb{N})$ (for 'safe tail') by the equations

$$\begin{array}{lcl} stail\,[\,] & = & [\,] \\ stail\,(x : xs) & = & xs \end{array}$$

then a simple counterexample verifies that the lists identified by $stail$ are not closed under $cons$: for example, with $xs = [\,]$ and $ys = 0 : [\,]$, we have $stail\,xs = [\,] = stail\,ys$, but $stail\,(1 : xs) = [\,] \neq 0 : [\,] = stail\,(1 : ys)$. Therefore $stail$ cannot be written directly as a fold.

Example 4.9 On the datatype $List(\mathbb{R})$ of finite lists of reals, consider the problem of computing $floorsum = floor \cdot rsum$, where $rsum : List(\mathbb{R}) \to \mathbb{R}$ sums a list of reals and $floor : \mathbb{R} \to \mathbb{Z}$ rounds a real r down to the largest integer at most r. Since the result is an integer, one might wonder whether the computation can be carried out as a fold to integers, thereby avoiding the computationally more expensive real arithmetic. One cannot: we have $floorsum\,(0.3 : [\,]) = floorsum\,(0.6 : [\,])$, yet $floorsum\,(0.5 : 0.3 : [\,]) \neq floorsum\,(0.5 : 0.6 : [\,])$.

On the other hand, the reverse composition $rsum \cdot map\,floor$, which floors every element of the list before summing, can be written as a fold: an argument similar to that in Example 4.7 applies. This is an instance of *deforestation* [23], an efficiency-improving transformation whereby two computations are combined into one and the intermediate data structure (here of type $List(\mathbb{Z})$) eliminated.

Remark 4.10 For the type $\mathit{Tree}(A)$ of binary trees with constructors $\mathit{leaf} : A \to \mathit{Tree}(A)$ and $\mathit{node} : \mathit{Tree}(A) \times \mathit{Tree}(A) \to \mathit{Tree}(A)$, Theorem 4.2 reduces to stating that an arbitrary function $h : \mathit{Tree}(A) \to B$ can be written directly as a fold precisely when the trees that are identified by h are closed under *node*, in the sense that for all t, u,

$$h\,t = h\,t' \wedge h\,u = h\,u' \quad \Rightarrow \quad h\,(\mathit{node}\,(t,u)) = h\,(\mathit{node}\,(t',u'))$$

Example 4.11 For another example of deforestation, consider $\mathit{flatsum} = \mathit{sum} \cdot \mathit{flatten}$, where $\mathit{flatten} : \mathit{Tree}(A) \to \mathit{List}(A)$ generates a list of the elements of a tree. The intermediate list in this computation can be eliminated, because

$$\begin{array}{cl} & \mathit{flatsum}\,(\mathit{node}\,(t,u)) \\ = & \quad \{ \text{ definition of } \mathit{flatsum} \ \} \\ & \mathit{sum}\,(\mathit{flatten}\,(\mathit{node}\,(t,u))) \\ = & \quad \{ \text{ definition of } \mathit{flatten} \ \} \\ & \mathit{sum}\,(\mathit{flatten}\,t \mathbin{+\!\!+} \mathit{flatten}\,u) \\ = & \quad \{ \ \mathit{sum} \text{ distributes over } \mathbin{+\!\!+} \ \} \\ & \mathit{flatsum}\,t + \mathit{flatsum}\,u \end{array}$$

from which we conclude that trees identified under *flatsum* are closed under *node*. (Here, '$\mathbin{+\!\!+}$' concatenates two lists.)

Example 4.12 The predicate $\mathit{bal} : \mathit{Tree}(A) \to \mathbb{B}$ that holds of tree iff it is *balanced* (with all the leaves at the same depth) is not a fold: with tree t being balanced and of depth 1, and tree u being balanced and of depth 2, both t and u are identified by *bal* (both yielding true), yet $\mathit{bal}\,(\mathit{node}\,(t,t)) \neq \mathit{bal}\,(\mathit{node}\,(t,u))$.

Example 4.13 However, the function $\mathit{dbal} : \mathit{Tree}(A) \to \mathbb{N} \times \mathbb{B}$ that computes a pair, the depth of the tree and whether it is balanced, is a fold. Because

$$\begin{array}{lcl} \mathit{depth}\,(\mathit{node}\,(t,u)) & = & 1 + \mathit{max}\,(\mathit{depth}\,t, \mathit{depth}\,u) \\ \mathit{bal}\,(\mathit{node}\,(t,u)) & = & \mathit{bal}\,t \ \wedge \ \mathit{bal}\,u \ \wedge \ \mathit{depth}\,t = \mathit{depth}\,u \end{array}$$

trees identified by *dbal* are closed under *node*. This is an example of a *mutumorphism* [7] or *almost homomorphism* [3, 8]; transforming a function into such a form is an important step towards constructing an efficient data-parallel algorithm for computing it.

5 When is a function an unfold?

Dualising Theorem 4.2 to unfold is straightforward. For our purposes, the appropriate dual of the notion of the kernel of a function is simply its *image*:

Definition 5.1 The *image* of a function $f : A \to B$ is the set of elements that are produced by f:

$$\mathsf{img}\, f \quad = \quad \{\, b \in B \mid \exists a : A. \quad f\, a = b \,\}$$

The duality between kernel pairs and images is perhaps not immediately evident, but is revealed by thinking relationally. In particular, if functions are viewed as relations in the obvious way, then the relational composition $f^{\circ} \cdot f$ of a function f with its converse f° is precisely the kernel of f, while the dual composition $f \cdot f^{\circ}$ is (the identity relation on) the image of f. (A more elaborate characterization of kernels is needed for relations in general, but it agrees with this one when restricted to total functions.)

We can now present our result for unfold, which gives a necessary and sufficient condition for when an arbitrary arrow $h : A \to \nu F$ in $\mathcal{SET}$ can be written in the form $h = \mathsf{unfold}\, g$ for some other arrow $g : A \to F\, A$.

Theorem 5.2 Suppose that $h : A \to \nu F$. Then

$$\begin{array}{l} (\exists g : A \to F\, A. \quad h = \mathsf{unfold}\, g) \\ \quad \Leftrightarrow \quad \mathsf{img}\,(F\, h) \supseteq \mathsf{img}\,(\mathsf{out} \cdot h) \end{array}$$

Again, the crux of the proof is a well-known observation, this time about factoring functions in the other direction.

Lemma 5.3 Inclusion of images is equivalent to the existence of 'prefactors': suppose that $f : B \to C$ and $h : A \to C$; then

$$(\exists g : A \to B. \quad h = f \cdot g) \quad \Leftrightarrow \quad \mathsf{img}\, f \supseteq \mathsf{img}\, h \ \wedge\ A \to B \neq \emptyset$$

Proof From left to right, assume that $g : A \to B$ and $h = f \cdot g$; then clearly $A \to B \neq \emptyset$, and moreover,

$$\begin{array}{ll} & c \in \mathsf{img}\, h \\ \Leftrightarrow & \quad \{ \text{ images } \} \\ & \exists a. \quad h\, a = c \\ \Leftrightarrow & \quad \{ \text{ assumption } \} \\ & \exists a. \quad f\,(g\, a) = c \\ \Rightarrow & \quad \{ \text{ images } \} \\ & c \in \mathsf{img}\, f \end{array}$$

Conversely, assume that $\mathsf{img}\, f \supseteq \mathsf{img}\, h$ and $A \to B \neq \emptyset$, so that either $A = \emptyset$ or $B \neq \emptyset$. When $A = \emptyset$, then h is the empty function; choose g to be the empty function too, so $f \cdot g$ is also empty and hence equal to h. When $B \neq \emptyset$, we define $g\, a$ for $a : A$ as follows. Let $c = h\, a$; by assumption, $c \in \mathsf{img}\, f$ too, so there exists a $b : B$ with $f\, b = c$, and we define $g\, a$ to be such a b. If there is more than one such b, it doesn't matter which we choose; by construction, we will have $h = f \cdot g$. □

We also have the dual to lemma 4.4:

Lemma 5.4

$$A \to \nu F \neq \emptyset \Rightarrow A \to F\, A \neq \emptyset$$

Proof We note that

$$A \to F\, A \neq \emptyset \Leftrightarrow (A \neq \emptyset \Rightarrow F\, A \neq \emptyset)$$

and show that the implication $A \neq \emptyset \Rightarrow F\, A \neq \emptyset$ holds:

$$\begin{array}{ll} & A \neq \emptyset \\ \Rightarrow & \quad \{\ A \to F\, A \neq \emptyset\ \} \\ & \nu F \neq \emptyset \\ \Rightarrow & \quad \{\ \mathsf{out} : \nu F \to F\, \nu F\ \} \\ & F\, \nu F \neq \emptyset \end{array}$$

We also have that

$$\begin{array}{ll} & A \neq \emptyset \\ \Rightarrow & \quad \{\ \text{constant function}\ \} \\ & \nu F \to A \neq \emptyset \\ \Rightarrow & \quad \{\ \text{functoriality of } F\ \} \\ & F\, \nu F \to F\, A \neq \emptyset \end{array}$$

Putting both results together we obtain:

$$F\, A \neq \emptyset$$

□

Proof of Theorem 5.2 Again, the proof is simple:

$$
\begin{array}{ll}
& \exists g : A \to F\,A. \quad h = \mathsf{unfold}\,g \\
\Leftrightarrow & \quad \{ \text{ universal property } \} \\
& \exists g : A \to F\,A. \quad \mathsf{out} \cdot h = F\,h \cdot g \\
\Leftrightarrow & \quad \{ \text{ Lemma 5.3 } \} \\
& \mathsf{img}\,(F\,h) \supseteq \mathsf{img}\,(\mathsf{out} \cdot h) \;\wedge\; A \to F\,A \neq \emptyset \\
\Leftrightarrow & \quad \{ \text{ Lemma 5.4 using } h : A \to \nu F \neq \emptyset \;\; \} \\
& \mathsf{img}\,(F\,h) \supseteq \mathsf{img}\,(\mathsf{out} \cdot h)
\end{array}
$$

□

Remark 5.5 For the type $\mathit{Stream}(\mathbb{N})$ of streams of naturals, with destructors $\mathit{head} : \mathit{Stream}(\mathbb{N}) \to \mathbb{N}$ and $\mathit{tail} : \mathit{Stream}(\mathbb{N}) \to \mathit{Stream}(\mathbb{N})$, Theorem 5.2 reduces to stating that an arbitrary function $h : A \to \mathit{Stream}(\mathbb{N})$ can be written as an unfold precisely when the *tail* of every stream producable by h is itself producable by h, in the sense that:

$$\mathsf{img}\,(\mathit{tail} \;\cdot\; h) \;\subseteq\; \mathsf{img}\,h$$

Example 5.6 Consider the function $\mathit{from} : \mathbb{N} \to \mathit{Stream}(\mathbb{N})$ defined in Section 2.2. Note that $(\mathit{tail} \cdot \mathit{from})\,n$ is the stream $[n+1, n+2, \ldots]$, and in general, $\mathsf{img}\,(\mathit{tail} \cdot \mathit{from})$ is the set of streams $\{\ [n+1, n+2, \ldots] \mid n \in \mathbb{N}\ \}$, which is in included in $\mathsf{img}\,\mathit{from}$, the set of streams $\{\ [n, n+1, \ldots] \mid n \in \mathbb{N}\ \}$. Therefore, *from* is expressible as an unfold.

Example 5.7 On the other hand, consider the function $\mathit{mults} : \mathbb{N} \to \mathit{Stream}(\mathbb{N})$ such that $\mathit{mults}\,n$ yields the stream of multiples $[0, n, n \times 2, n \times 3, \ldots]$ of the natural n. Here, $(\mathit{tail} \cdot \mathit{mults})\,n$ is the stream $[n, n \times 2, \ldots]$, and so $\mathsf{img}\,(\mathit{tail} \cdot \mathit{mults})$ is not included in $\mathsf{img}\,\mathit{mults}$, which includes only streams whose *head* is 0. Therefore, *mults* cannot be expressed directly as an unfold.

Remark 5.8 For the codatatype $\mathit{CoList}(A)$ of finite and infinite lists with elements drawn from A, with destructors $\mathit{hd} : \mathit{CoList}(A) \to \mathbf{1} + A$ and $\mathit{tl} : \mathit{CoList}(A) \to \mathbf{1} + \mathit{CoList}(A)$, Theorem 5.2 reduces to stating that an arbitrary function $h : B \to \mathit{CoList}(A)$ can be written as an unfold precisely when the *tl* of every colist producable by h is itself producable by h, in the sense that:

$$\mathsf{img}\,(\mathit{tl} \;\cdot\; h) \;\subseteq\; \mathsf{img}\,(\mathsf{id}_{\mathbf{1}} + h)$$

Example 5.9 The function *rle* : $\mathit{CoList}(A) \to \mathit{CoList}(A \times \mathbb{N})$ performs *run-length encoding*: equal adjacent elements of the input are grouped together, and output as a single element together with an occurrence count. It can be written as an unfold: a non-empty encoding of a colist (that is, the encoding of a non-empty colist) has a tail which is itself the encoding of another colist, namely the one resulting from removing from the original colist all initial elements equal to the head. (The corresponding function of type $\mathit{List}(A) \to \mathit{List}(A \times \mathbb{N})$ is a fold, too; we leave the proof as an exercise.)

Remark 5.10 For the codatatype $\mathit{CoTree}(A)$ of infinite binary trees with elements drawn from A, with destructors $\mathit{root} : \mathit{CoTree}(A) \to A$ and $\mathit{left}, \mathit{right} : \mathit{CoTree}(A) \to \mathit{CoTree}(A)$, Theorem 5.2 reduces to stating that an arbitrary function $h : B \to \mathit{CoTree}(A)$ can be written as an unfold precisely when the children of every cotree producable by h are themselves producable by h, in the sense that:

$$\begin{aligned} \mathsf{img}\,(\mathit{left} \cdot h) &\subseteq \mathsf{img}\,h \\ \mathsf{img}\,(\mathit{right} \cdot h) &\subseteq \mathsf{img}\,h \end{aligned}$$

Example 5.11 Consider the *position tree*, an infinite binary tree in which every node is labelled its *position* in the tree, a finite sequence of booleans recording the left and right turns from the root in order to reach that node. The function $\mathit{positions} : \mathbf{1} \to \mathit{CoTree}(\mathit{List}(\mathbb{B}))$ that yields this tree is not an unfold: $\mathsf{img}\,\mathit{positions}$ contains exactly one tree, with the empty sequence at its root, whereas the left and right children should have singleton lists at their roots.

On the other hand, the function $\mathit{posfrom} : \mathit{List}(\mathbb{B}) \to \mathit{CoTree}(\mathit{List}(\mathbb{B}))$, which generates the position tree starting from a given position, is an unfold: the left child of *posfrom bs* is *posfrom* (*true* : *bs*) and so is producable in turn by *posfrom*, as is the right child.

6 Conclusion

We have given the first complete results for when an arbitrary arrow can be written directly as a fold or unfold, for the special case of the category $\mathcal{SET}$. This is interesting from a theoretical point of view, as it increases our understanding of these important operators from functional programming. More importantly, though, we expect it to have practical applications in program optimisation. Various research groups [9, 15, 20] are working on *deforestation*. A well-structured program is typically factored into several phases, each phase generating a data structure consumed by the subsequent phase; deforestation fuses adjacent phases and eliminates the intermediate data structure. When performed as a compiler optimisation, it yields efficient object code without sacrificing the structure and clarity of the source code. Our results can be used to determine that two phases cannot be fused to a fold or an unfold. It might be possible to harness a testing framework like QuickCheck [2] to the task of finding counterexamples to the appropriate inclusions.

In future work we will investigate whether the results can be generalised to other categories, but at present it is not clear how this can be done. In particular, our proofs seem to make essential use of the axiom of choice, and so do not carry through to computational categories such as $\mathcal{C}PO$. This does not rule out all interesting categories, though; although our proof does not carry through to $\mathcal{P}fun$ or $\mathcal{R}EL$, it may be that the theorem itself or some variation on it does. Another fruitful direction for future work is to extend these results to other patterns of recursion, such as primitive (co-)recursion [18, 21] and course-of-value (co-)iteration [22].

Acknowledgements

We are grateful to Lambert Meertens, who pointed out to us that the crucial lemmas (Lemma 4.3 and Lemma 5.3) were equivalences, not just implications, thereby simplifying our proofs considerably. Graham Hutton and Thorsten Altenkirch are supported by ESPRIT Working Group *Applied Semantics*. During part of this work, Graham Hutton was also supported by EPSRC grant *Structured Recursive Programming*.

References

[1] R. Bird and O. de Moor. *Algebra of Programming*. Prentice Hall, 1997.

[2] K. Claessen and J. Hughes. Quickcheck: A lightweight tool for random testing of Haskell programs. In *Proc. 5th ACM SIGPLAN International Conference on Functional Programming*, September 2000.

[3] M. Cole. Parallel programming with list homomorphisms. *Parallel Processing Letters*, 5(2):191–203, 1995.

[4] J. Gibbons. Calculating functional programs. In *Summer School and Workshop on Algebraic and Coalgebraic Methods in the Mathematics of Program Construction*, Oxford, April 2000.

[5] J. Gibbons and G. Hutton. Proof methods for structured corecursive programs. In *Proc. 1st Scottish Functional Programming Workshop*, Stirling, Scotland, August 1999.

[6] J. Gibbons and G. Jones. The under-appreciated unfold. In *Proc. 3rd ACM SIGPLAN International Conference on Functional Programming*, Baltimore, Maryland, September 1998.

[7] M. M. Fokkinga. *Law and Order in Algorithmics*. PhD thesis, Universiteit Twente, 1992.

[8] S. Gorlatch. Extracting and implementing list homomorphisms in parallel program development. *Science of Computer Programming*, 33:1–27, 1999.

[9] Z. Hu, H. Iwasaki and M. Takeichi. Deriving structural hylomorphisms from recursive definitions. In *Proc. 1st ACM SIGPLAN International Conference on Functional Programming*, 1996.

[10] G. Hutton. Fold and unfold for program semantics. In *Proc. 3rd ACM SIGPLAN International Conference on Functional Programming*, Baltimore, Maryland, September 1998.

[11] G. Hutton. A tutorial on the universality and expressiveness of fold. *Journal of Functional Programming*, 9(4):355–372, July 1999.

[12] B. Jacobs, L. Moss, H. Reichel, and J. Rutten, editors. *Proc. of the First Workshop on Coalgebraic Methods in Computer Science*. Elsevier Science B.V., 1998. Electronic Notes in Theoretical Computer Science Volume 11.

[13] B. Jacobs and J. Rutten. A tutorial on (co)algebras and (co)induction. *Bulletin of the European Association for Theoretical Computer Science*, 62:222–259, 1997.

[14] B. Jacobs and J. Rutten, editors. *Proc. of the Second Workshop on Coalgebraic Methods in Computer Science*. Elsevier Science B.V., 1999. Electronic Notes in Theoretical Computer Science Volume 19.

[15] J. Launchbury and T. Sheard. Warm fusion: Deriving build-catas from recursive definitions. In *Proc. Conference on Functional Programming Languages and Computer Architecture*, ACM Press, 1995.

[16] S. Mac Lane. *Categories for the Working Mathematician*. Number 5 in Graduate Texts in Mathematics. Springer-Verlag, 1971.

[17] G. Malcolm. Algebraic data types and program transformation. *Science of Computer Programming*, 14(2-3):255–280, September 1990.

[18] L. Meertens. Paramorphisms. *Formal Aspects of Computing*, 4(5):413–424, 1992.

[19] E. Meijer, M. Fokkinga, and R. Paterson. Functional programming with bananas, lenses, envelopes and barbed wire. In J. Hughes, editor, *Proc. Conference on Functional Programming and Computer Architecture*, number 523 in LNCS. Springer-Verlag, 1991.

[20] A. Takano and E. Meijer. Shortcut deforestation in calculational form. In *Proc. Conference on Functional Programming Languages and Computer Architecture*, ACM Press, 1995.

[21] V. Vene and T. Uustalu. Functional programming with apomorphisms (corecursion). *Proceedings of the Estonian Academy of Sciences: Physics, Mathematics*, 47(3):147–161, 1998.

[22] V. Vene. *Categorical Programming with Inductive and Coinductive Types*. PhD thesis, Universitity of Tartu, 2000.

[23] P. Wadler. Deforestation: Transforming programs to eliminate trees. *Theoretical Computer Science*, 73:231–248, 1990.