
Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/28183/1/underperforming.pdf

Copyright and reuse:

The Nottingham ePrints service makes this work by researchers of the University of Nottingham available open access under the following conditions.

- Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners.
- To the extent reasonable and practicable the material made available in Nottingham ePrints has been checked for eligibility before being made available.
- Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
- Quotations or similar reproductions must be sufficiently acknowledged.

Please see our full end user licence at:
http://eprints.nottingham.ac.uk/end_user_agreement.pdf

A note on versions:

The version presented here may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher’s version. Please see the repository url above for details on accessing the published version and note that access may require a subscription.

For more information, please contact eprints@nottingham.ac.uk
The Under-Performing Unfold

A new approach to optimising corecursive programs

Jennifer Hackett  Graham Hutton
University of Nottingham
{jpjh, gmh}@cs.nott.ac.uk

Mauro Jaskelioff
Universidad Nacional de Rosario, Argentina
CIFASIS–CONICET, Argentina
jaskelioff@cifasis-conicet.gov.ar

Abstract
This paper presents a new approach to optimising corecursive programs by factorisation. In particular, we focus on programs written using the corecursion operator unfold. We use and expand upon the proof techniques of guarded coinduction and unfold fusion, capturing a pattern of generalising coinductive hypotheses by means of abstraction and representation functions. The pattern we observe is simple, has not been observed before, and is widely applicable. We develop a general program factorisation theorem from this pattern, demonstrating its utility with a range of practical examples.

Categories and Subject Descriptors D.1.1 [Programming Techniques]: Applicative (Functional) Programming

General Terms Theory, Proof Methods, Optimisation

Keywords fusion, factorisation, coinduction, unfolds

1. Introduction

When writing programs that produce data structures it is often natural to use the technique of corecursion [8, 20], in which subterms of the result are produced by recursive calls. This is particularly useful in lazy languages such as Haskell, as it allows us to process parts of the result without producing the rest of it. In this way, we can write programs that deal with large or infinite data structures and trust that the memory requirements remain reasonable.

However, while this technique may allow us to save on the number of recursive calls by producing data lazily, the time cost of each call or the space cost of the arguments can still be a problem. For this reason, it is necessary to examine various approaches to reducing these costs. A commonly-used approach is program fusion [4, 11, 31], where separate stages of a computation are combined into one to avoid producing intermediate data. For this paper, our focus will be on the opposite technique of program factorisation [7], where a computation is split into separate parts. In particular, we consider the problem of splitting a computation into the combination of a more efficient worker program that uses a different representation of data and a wrapper program that effects the necessary change of representation [10].

The primary contribution of this paper is a general factorisation theorem for programs in the form of an unfold [9], a common pattern of corecursive programming. By exploiting the categorical principle of duality, we adapt the theory of worker-wrapper factorisation for folds, which was developed initially by Hutton, Jaskelioff and Gill [15] and subsequently extended by Sculthorpe and Hutton [24]. We discuss the practical considerations of the dual theory, which differ from those of the original theory and avoid the need for strictness side conditions. In addition, we revise the theory to further improve its utility. As we shall see, the resulting theory for unfolds captures a common pattern of coinductive reasoning that is simple, has not been observed before, and is widely applicable. We demonstrate the application of the new theory with a range of practical examples of varying complexity.

The development of our new theory is first motivated using a small programming example, which we then generalise using category theory. We use only simple concepts for this generalisation, and only a minimal amount of categorical knowledge is required. Readers without a background in category theory should not be deterred. The primary application area for our theory is functional languages such as Haskell, however the use of categorical concepts means that our theory is more widely applicable.

2. Coinductive Types and Proofs

Haskell programmers will be familiar with recursive type definitions, where a type is defined in terms of itself. For example, the type of natural numbers can be defined as follows:

```
data N = Zero | Succ N
```

This definition states that an element of the type \( N \) is either \( \text{Zero} \) or \( \text{Succ} \ n \) for some \( n \) that is also an element of the type \( N \). More formally, recursive type definitions can be given meaning as fixed points of type-level equations. For example, we can read the above as defining \( N \) to be a fixed point of the equation \( X = 1 + X \), where 1 is the unit type and + is the disjoint sum of types.

Assuming that such a fixed point equation has at least one solution, there are two we will typically be interested in. First of all, there is the least fixed point, which is given by the smallest type that is closed under the constructors. This is known as an inductive type. In a set-theoretic context, the inductive interpretation of our definition of \( N \) is simply the set of natural numbers, with constructors \( \text{Zero} \) and \( \text{Succ} \) having the usual behaviour.

Alternatively, there is the greatest fixed point, which is given by the largest type that supports deconstruction of values by pattern matching. This is known as a coinductive type [12]. In a set-theoretic context, the coinductive interpretation of our definition of \( N \) is the set of naturals augmented with an infinite value \( \infty \) that is the solution to the equation \( \infty = \text{Succ} \ \infty \).

In general, coinductively-defined sets have infinite elements, while inductively-defined sets do not. In the setting of Haskell, however, types correspond to (pointed) complete partial orders.
(CPOs) rather than sets, where there is no distinction between inductive and coinductive type definitions as the two notions coincide [6]. For the purposes of this article we will use Haskell syntax as a metalanguage for programming in both set-theoretic and CPO contexts. It will therefore be necessary to distinguish between inductive and coinductive definitions, which we do so by using the keywords \texttt{data} and \texttt{codata} respectively.

For example, we could define an inductive type of lists and a coinductive type of streams as follows, using the constructor \((:)\) in both definitions for consistency with Haskell usage:

\[
\begin{align*}
\text{data} & \quad [a] & = & a : [a] \mid [] \\
\text{codata} & \quad \text{Stream} a & = & a : \text{Stream} a
\end{align*}
\]

If types are sets, the first definition gives finite lists while the second gives infinite streams. If types are CPOs, the first definition gives both finite and infinite lists, while the second gives infinite streams. Also note that in the context of sets, if streams were defined using \texttt{data} rather than \texttt{codata} the resulting type would be empty.

2.1 Coinduction

To reason about elements of inductive types one can use the technique of induction. Likewise, to reason about elements of coinductive types one can use the dual technique of coinduction. To formalise this precisely involves the notion of \textit{bisimulation} [8, 12]. In this section we shall give an informal presentation of \textit{guarded} coinduction [3, 28], a special case that avoids the need for such machinery. This form of coinduction is closely related to the unique fixed point principle developed by Hinze [14].

To prove an equality \(\text{lhs} = \text{rhs}\) between expressions of the same coinductive type using guarded coinduction, we simply attempt the proof in the usual way using equational reasoning. However, we may also make use of the \textit{coinductive hypothesis}, which allows us to substitute \text{lhs} for \text{rhs} (or vice-versa) provided that we only do so immediately underneath a constructor of the coinductive type. We say that such a use of the coinductive hypothesis is \textit{guarded}. For example, if we define the following functions that produce values of type \(\text{Stream} \mathbb{N}\)

\[
\begin{align*}
\text{from} & \quad n & = & n : \text{from} (n + 1) \\
\text{skips} & \quad n & = & n : \text{skips} (n + 2) \\
\text{double} & \quad (n : ns) & = & n * 2 : \text{double} ns
\end{align*}
\]

then we can show that \text{skips} \((n * 2)\) = \text{double} \((\text{from} n)\) for any natural number \(n\) using guarded coinduction:

\[
\begin{align*}
\text{skips} (n * 2) & = \{ \text{definition of \text{skips}} \} \\
& = \{ \text{arithmetic} \} \\
n * 2 \cdot \text{skips} ((n + 1) * 2) & = \{ \text{coinductive hypothesis} \} \\
n * 2 \cdot \text{double} (\text{from} (n + 1)) & = \{ \text{definition of \text{double}} \} \\
\text{double} (n : \text{from} (n + 1)) & = \{ \text{definition of \text{from}} \} \\
\text{double} (\text{from} n) & = \{ \text{coinductive hypothesis} \}
\end{align*}
\]

Despite the apparent circularity in using an instance of our desired result in the third step of the proof, the proof is guarded because the use of the coinductive hypothesis only occurs directly below the \((:)\) constructor. Therefore the reasoning is valid.

3. Example: Tabulating a Function

Consider the problem of \textit{tabulating} a function \(f :: \mathbb{N} \rightarrow a\) by applying it to every natural number in turn and forming a stream from the results. We would like to define a function that performs this task, specified informally as follows:

\[
\begin{align*}
\text{tabulate} &: (\mathbb{N} \rightarrow a) \rightarrow \text{Stream} a \\
\text{tabulate} f &= [f 0, f 1, f 2, f 3, \ldots]
\end{align*}
\]

The following definition satisfies this specification:

\[
\text{tabulate} f = f 0 : \text{tabulate} (f o (+1))
\]

However, this definition is inefficient, as with each recursive call the function argument becomes more costly to apply, as shown in the following expansion of the definition:

\[
\begin{align*}
\text{tabulate} f &= [f 0, (f o (+1)) 0, (f o (+1)) o (+1)) 0, \ldots]
\end{align*}
\]

The problem is that the natural number is recomputed from scratch each time by repeated application of \((+1)\) to 0. If we were to save the result and re-use it in future steps, we could avoid repeating work. The idea can be implemented by defining a new function that takes the current value as an additional argument:

\[
\begin{align*}
\text{tabulate}’ &: (\mathbb{N} \rightarrow (\mathbb{N} \rightarrow a)) \rightarrow \text{Stream} a \\
\text{tabulate}’ (f, n) &= f n : \text{tabulate}’ (f, n + 1)
\end{align*}
\]

The correctness of the more efficient implementation for tabulation can be captured by the following equation:

\[
\text{tabulate} f = \text{tabulate}’ (f, 0)
\]

which can written in point-free form as

\[
\text{tabulate} = \text{tabulate}’ o (\lambda f \rightarrow (f, 0))
\]

This equation can be viewed as a \textit{program factorisation}, in which \text{tabulate} is factored into the composition of \text{tabulate}’ and the function \((\lambda f \rightarrow (f, 0))\). This latter function effects a \textit{change of data representation} from the old argument type \(\mathbb{N} \rightarrow a\) to the new argument type \(\mathbb{N} \rightarrow (\mathbb{N} \rightarrow a)\). In order to try to prove the above equation, we proceed by guarded coinduction:

\[
\begin{align*}
\text{tabulate} f & = \{ \text{definition of \text{tabulate}} \} \\
& = \{ \text{coinductive hypothesis} \} \\
& = f 0 : \text{tabulate}’ (f o (+1), 0) \\
& = \{ \text{assumption} \} \\
& = f 0 : \text{tabulate}’ (f, 1) \\
& = \{ \text{definition of \text{tabulate}} \} \\
& = \text{tabulate}’ (f, 0)
\end{align*}
\]

To complete the proof, the assumption used in the third step \(\text{tabulate}’ (f o (+1), 0) = \text{tabulate}’ (f, 1)\) must be verified. We could attempt to prove this as follows:

\[
\begin{align*}
\text{tabulate}’ (f o (+1), 0) & = \{ \text{definition of \text{tabulate}}’ \} \\
& = (f o (+1)) 0 : \text{tabulate}’ (f o (+1), 1) \\
& = \{ \text{composition, arithmetic} \} \\
& = f 1 : \text{tabulate}’ (f o (+1), 1) \\
& = \{ \text{assumption} \} \\
& = f 1 : \text{tabulate}’ (f, 2) \\
& = \{ \text{definition of \text{tabulate}}’ \} \\
& = \text{tabulate}’ (f, 1)
\end{align*}
\]

Once again, however, the proof relies on an assumption that needs to be verified. We could continue like this \textit{ad infinitum} without ever actually completing the proof! We can avoid this problem by generalising our correctness property to

\[
\text{tabulate} (f o (+n)) = \text{tabulate}’ (f, n)
\]
which in the case of \( n = 0 \) simplifies to the original equation. The proof of the generalised property is now a straightforward application of guarded coinduction, with no assumptions required:

\[
\text{tabulate } (f \circ (+n)) = \begin{cases} \text{definition of tabulate} \\ (f \circ (+n)) 0 : \text{tabulate} (f \circ (+n) \circ (+1)) = \{ \text{simplication} \} \\ f n : \text{tabulate} (f \circ (+n+1)) = \{ \text{coinduction hypothesis} \} \\ f n : \text{tabulate}' (f, n+1) = \{ \text{definition of tabulate}' \} \\
\end{cases}
\]

It is often necessary to generalise coinductive hypotheses in this way for proofs by guarded coinduction, just as it is often necessary to generalise inductive hypotheses for proofs by induction.

4. Abstracting

The above example is an instance of a general pattern of optimisation that is simple yet powerful. We abstract from this example to the general case in two steps, firstly by generalising on the underlying datatypes involved and secondly by generalising on the pattern of coinductive definition that is used.

4.1 Abstracting on Datatypes

In the tabulation example, we replaced the original function of type \((N \to a) \to \text{Stream a}\) with a more efficient function of type \([N \to a, N] \to \text{Stream a}\), changing the type of the argument. Essentially, we used a “larger” type as a representation of a “smaller” type. We can generalise this idea to any two types where one serves as a representation of the other.

Suppose we have two types, \(a\) and \(b\), with conversion functions \(\text{abs} : b \to a\) and \(\text{rep} : a \to b\) such that \(\text{abs} \circ \text{rep} = \text{id}_a\). We can think of \(b\) as a larger type that faithfully represents the elements of the smaller type \(a\). Now suppose that we are given a function \(\text{old} : a \to c\), together with a more efficient version \(\text{new} : b \to c\) that acts on the larger type \(b\). Then the correctness of the more efficient version can be captured by the equation

\[
\text{old} = \text{new} \circ \text{rep}
\]

However, using the assumption that \(\text{abs} \circ \text{rep} = \text{id}_a\) we can strengthen this property by the following calculation:

\[
\text{old} = \text{new} \circ \text{rep} \\
\Leftrightarrow \{ \text{abs} \circ \text{rep} = \text{id}_a \} \\
\text{old} \circ \text{abs} \circ \text{rep} = \text{new} \circ \text{rep} \\
\Leftrightarrow \{ \text{cancelling \text{rep} on both sides} \} \\
\text{old} \circ \text{abs} = \text{new}
\]

In summary, if we wish to show that function \(\text{new}\) is correct, it suffices to show that \(\text{old} \circ \text{abs} = \text{new}\). This stronger correctness property may be easier to prove than the original version.

We now apply the above idea to our earlier example. In this case, the appropriate \(\text{abs}\) and \(\text{rep}\) functions are:

\[
\begin{align*}
\text{abs} &: (N \to a, N) \to (N \to a) \\
\text{abs} (f, n) &= f \circ (+n) \\
\text{rep} &: (N \to a) \to (N \to a, N) \\
\text{rep} f &= (f, 0)
\end{align*}
\]

The required relationship \(\text{abs} \circ \text{rep} = \text{id}\) follows immediately from the fact that 0 is the identity for addition. The above calculation can therefore be specialised to our example as follows:

\[
\forall f . \text{tabulate } f = \text{tabulate}' (f, 0) \\
\Leftrightarrow \{ \text{definition of \text{rep}} \}
\]

The final equation is precisely the generalised correctness property from the previous section, but has now been obtained from an abstract framework that is generic in the underlying datatypes.

4.2 Abstracting on Corecursion Pattern

For the next step in the generalisation, we make some assumptions about the corecursive structure of the functions that we are dealing with. In particular, we assume that they are instances of a specific pattern of corecursive definition called \textit{unfold} [9, 18].

4.2.1 Unfold for Streams

An unfold is a function that produces an element of a coinductive type as its result, producing all subterms of the result using recursive calls. This pattern can be abstracted into an operator, which we define in the case of streams as follows:

\[
\text{unfold} :: (a \to b) \to (a \to a) \to a \to \text{Stream b}
\]

\[
\text{unfold } h t x = h x \circ \text{unfold } h t (t x)
\]

The function \(\text{unfold} \ h \ t\) produces a stream from a seed value \(x\) by using the function \(h\) to produce the head of the stream from the seed, and applying the function \(t\) to produce a new seed that is used to generate the tail of the stream in the same manner. For efficiency we could choose to combine the \(h\) and \(t\) functions into a single function of type \(a \to (b, a)\), allowing for increase sharing between the two computations. However, we present a ‘tuple-free’ version because it leads to simpler equational reasoning.

By providing suitable definitions for \(h\) and \(t\), it is straightforward to redefine the functions \text{tabulate} and \text{tabulate}'

\[
\begin{align*}
\text{tabulate} &= \text{unfold } h t \\
\text{tabulate}' &= \text{unfold } h' t'
\end{align*}
\]

In this way, unfold allows us to factor out the basic steps in the computations. A similar unfold operator can be defined for any coinductive type. For example, for infinite binary trees

\[
\text{codata } \text{Tree } a = \text{Node } (\text{Tree } a) a (\text{Tree } a)
\]

the following definition for \text{unfold} \(l \ n \ r\) produces a tree from a seed value by using \(l\) and \(r\) to produce new seeds for the left and right subtrees, and \(n\) to produce the node value:

\[
\begin{align*}
\text{unfold} :: (a \to a) \to (a \to b) \to (a \to a) \to a \to \text{Tree } b \\
\text{unfold } l n r x = \text{Node } (\text{unfold } l n r (l x)) (n x) (\text{unfold } l n r (r x))
\end{align*}
\]

Once again, the \(l\), \(n\), and \(r\) functions could be combined.

4.2.2 Unfold Fusion

The unfold operator for any type has an associated \textit{fusion} law [18], which provides sufficient conditions for when the composition of
an unfold with another function can be expressed as a single unfold.

In the case of streams, the law is as follows:

**Theorem 1 (Unfold Fusion for Streams).** Given

\[
\begin{align*}
  h :: a \rightarrow c & \quad h' :: b \rightarrow c \\
  t :: a \rightarrow a & \quad t' :: b \rightarrow b
\end{align*}
\]

we have the following implication:

\[
\text{unfold } h \circ t = \text{unfold } h' \circ t'
\]

The proof is a simple application of guarded coinduction:

\[
\begin{align*}
  \text{unfold } h' \circ t' x &= \{ \text{definition of unfold} \} \\
  h' x : \text{unfold } h' \circ t' (t' x) &= \{ \text{coinduction hypothesis} \} \\
  h' x : \text{unfold } h \circ t (g (t' x)) &= \{ \text{first assumption: } h' = h \circ g \} \\
  h (g x) : \text{unfold } h \circ t (g (t' x)) &= \{ \text{second assumption: } g \circ t' = t \circ g \} \\
  h (g x) : \text{unfold } h \circ t (g x) &= \{ \text{definition of unfold} \} \\
  \text{unfold } h \circ t (g x) &= \{ \text{fusion} \} \\
  h' = h \circ g & \quad \wedge \quad \text{abs} \circ t' = t \circ \text{abs}
\end{align*}
\]

The fusion law provides sufficient conditions for when our strengthened correctness property \( \text{old} \circ \text{abs} = \text{new} \) holds. Assuming that \( \text{old} \) and \( \text{new} \) can both be expressed as unfolds, then:

\[
\text{old} \circ \text{abs} = \text{new}
\]

\[
\Leftrightarrow \{ \text{old} = \text{unfold } h \circ t, \text{new} = \text{unfold } h' \circ t' \}
\]

\[
\text{unfold } h \circ t = \text{unfold } h' \circ t'
\]

\[
= \{ \text{fusion} \}
\]

\[
\text{abs} \circ t' = t \circ \text{abs}
\]

4.3 **Unfold Factorisation for Streams**

Combining the two ideas of abstracting on the types and abstracting on the coinduction pattern, we obtain a general theorem for factorising functions defined using unfold for streams.

**Theorem 2 (Unfold Factorisation for Streams).** Given

\[
\begin{align*}
  \text{abs} :: b \rightarrow a & \quad \text{h} :: a \rightarrow c \\
  \text{rep} :: a \rightarrow b & \quad t :: a \rightarrow a \\
  t' :: b \rightarrow b
\end{align*}
\]

satisfying the assumptions

\[
\begin{align*}
  \text{abs} \circ \text{rep} &= \text{id}_a \\
  h' &= h \circ \text{abs} \\
  \text{abs} \circ t' &= t \circ \text{abs}
\end{align*}
\]

we have the factorisation

\[
\text{unfold } h \circ t = \text{unfold } h' \circ \text{rep}
\]

Using this result, we can split a function unfold \( h \circ t \) into the composition of a worker function unfold \( h' \circ t' \) that uses a different representation of data and a wrapper function \( \text{rep} \) that effects the necessary change of data representation.

We now apply this to the tabulate example. As we have already shown that \( \text{abs} \circ \text{rep} = \text{id} \), it is only necessary to verify the remaining two assumptions. We start with the first assumption:

\[
\begin{align*}
  h (\text{abs} (f, n)) &= \{ \text{definition of abs} \} \\
  h (f \circ (+n)) &= \{ \text{definition of } h \}
\end{align*}
\]

\[
(f \circ (+n)) 0 = \{ \text{simplification} \}
\]

\[
f n = \{ \text{definition of } h' \}
\]

\[
h' (f, n)
\]

Now, the second assumption:

\[
\begin{align*}
  t (\text{abs} (f, n)) &= \{ \text{definition of abs} \} \\
  t (f \circ (+n)) &= \{ \text{definition of } t \} \\
  f \circ (+n) \circ (+1) &= \{ \text{simplification} \} \\
  f \circ (+n + 1) &= \{ \text{definition of abs} \} \\
  \text{abs} (f, n + 1) &= \{ \text{definition of } t' \} \\
  \text{abs} (t' (f, n)) &= \{ \text{definition of } t \}
\end{align*}
\]

In conclusion, by generalising from our tabulation example we have derived a framework for factorising corecursive functions that are defined using the unfold operator for streams.

5. **Categorifying**

To recap, we have combined the idea of a change of data representation with an application of fusion to produce a factorisation theorem for stream unfolds. This theorem covers cases not covered by fusion alone. For example, attempting to prove the tabulate example correct simply using fusion fails in precisely the same way that our attempted proof using coinduction failed.

However, so far we have only concerned ourselves with the coinductive type of streams. If we wish to apply this technique to other coinductive types, it would seem that we must define unfold and prove its fusion law for every such type we intend to use. Thankfully this is not the case, as category theory provides a convenient generic approach to modelling coinductive types and their unfold operators using the notion of *final coalgebras* [18].

5.1 **Final Coalgebras**

Suppose that we fix a category \( C \) and a functor \( F : C \rightarrow C \) on this category. Then an \( F \)-coalgebra is a pair \( (A, f) \) consisting of an object \( A \) along with an arrow \( f : A \rightarrow F A \). We often omit the object \( A \) as it is implicit in the type of \( f \). A homomorphism between coalgebras \( f : A \rightarrow F A \) and \( g : B \rightarrow F B \) is an arrow \( h : A \rightarrow B \) such that \( F h \circ f = g \circ h \). This property is captured by the following commutative diagram:

\[
\begin{align*}
  &A &\xrightarrow{h} &B \\
  &\downarrow{f} &\downarrow{g} &\downarrow \\
  &FA &\xrightarrow{Fh} &FB
\end{align*}
\]

Intuitively, a coalgebra \( f : A \rightarrow F A \) can be thought of as giving a *behaviour* to elements of \( A \), where the possible behaviours are specified by the functor \( F \). For example, if we define \( F X = 1 + X \) on the category Set of sets and total functions, then a coalgebra \( f : A \rightarrow 1 + A \) is the transition function of a state machine in which each element of \( A \) is either a terminating state or has a single successor. In turn, a homomorphism corresponds to a behaviour-preserving mapping, in the sense that if we first apply the homomorphism \( h \) and then the target behaviour captured by \( g \), we
obtain the same result as if we apply the source behaviour captured by \( f \) and then apply \( h \) to the components of the result.

A final coalgebra, denoted \((\nu F, \text{out})\), is an \( F \)-coalgebra to which any other coalgebra has a unique homomorphism. If a final coalgebra exists, it is unique up to isomorphism. Given a coalgebra \( f : A \to F A \), the unique homomorphism from \( f \) to the final coalgebra \( \text{out} \) is denoted \( \text{unfold} \ f : A \to \nu F \). This \textit{uniqueness property} can be captured by the following equivalence:

\[
\text{unfold} \ f \iff F h \circ f = \text{out} \circ h
\]

We also have a fusion rule for unfold:

**Theorem 3 (Unfold Fusion for Final Coalgebras).** Given

\[
f : A \to F A \quad g : B \to F B \quad h : A \to B
\]

we have the following implication:

\[
\text{unfold} \ g \circ h = \text{unfold} \ f \iff F h \circ f = g \circ h
\]

The proof of this theorem can be conveniently captured by the following commutative diagram:

\[
\begin{array}{ccc}
A & \xrightarrow{f} & B \\
\downarrow{h} & & \downarrow{\text{unfold} \ g} \\
F A & \xrightarrow{\text{unfold} \ h} & F B
\end{array}
\]

\[
\begin{array}{ccc}
& & F (\text{unfold} \ g) \\
\nu F & \xleftarrow{\text{out}} & &
\end{array}
\]

The left square commutes by assumption while the right square commutes because \( \text{unfold} \ g \) is a coalgebra homomorphism. Therefore, the outer rectangle commutes, meaning that \( \text{unfold} \ g \circ h \) is a homomorphism from \( f \) to \( \text{out} \). Finally, because homomorphisms to the final coalgebra \( \text{out} \) are unique and \( \text{unfold} \ f \) is also such a homomorphism, the result \( \text{unfold} \ g \circ h = \text{unfold} \ f \) holds.

We illustrate the above concepts with a concrete example. Consider the functor \( F X = N \times X \) on the category \( \text{Set} \). This functor has a final coalgebra \((\text{Stream} \ N, (\text{head}, \text{tail}))\), consisting of the set \( \text{Stream} \ N \) of streams of natural numbers together with the function \((\text{head}, \text{tail}) : \text{Stream} \ N \to N \times \text{Stream} \) that combines the stream destructors \( \text{head} : \text{Stream} \ N \to N \) and \( \text{tail} : \text{Stream} \ N \to \text{Stream} \) \( N \). Given any set \( A \) and functions \( h : A \to N \) and \( t : A \to A \), the function \( \text{unfold}(h, t) : A \to \text{Stream} \ N \) is uniquely defined by the two equations

\[
\begin{align*}
\text{head} \circ \text{unfold}(h, t) & = h \\
\text{tail} \circ \text{unfold}(h, t) & = \text{unfold}(h, t) \circ t
\end{align*}
\]

which are equivalent to the more familiar definition of unfold using the stream constructor \((:)\) presented earlier:

\[
\text{unfold} \ h \ t \ x = h \ x : \text{unfold} \ h \ t \ (t \ x)
\]

We also note that the earlier fusion law for streams is simply a special case of the more general fusion law where \( F X = N \times X \).

The fusion precondition simplifies as follows

\[
\begin{align*}
F g \circ (h', t') & = (h, t) \circ g \\
\Leftrightarrow \{ \text{definition of } F, \text{products} \} & \{ h' \circ g \circ t' = (h \circ g, t \circ g) \} \\
\Leftrightarrow \{ \text{separating components} \} & h' = h \circ g \land g' \circ t = t \circ g
\end{align*}
\]

and the postcondition is clearly equivalent. All of the above also holds for \( \text{Stream} \ A \) for an arbitrary set \( A \).

It is now straightforward to generalise our earlier unfold factorisation theorem from streams to final coalgebras. Combining the general unfold fusion law with the same type abstraction idea from before, we obtain the following theorem.

**Theorem 4 (General Unfold Factorisation).** Given

\[
\begin{align*}
\text{abs} : B \to A & \quad f : A \to F A \\
\text{rep} : A \to B & \quad g : B \to F B
\end{align*}
\]

satisfying the assumptions

\[
\text{abs} \circ \text{rep} = \text{id}_A
\]

\[
F \text{abs} \circ \text{rep} = \text{abs} \circ g
\]

we have the factorisation

\[
\text{unfold} f = \text{unfold} g \circ \text{rep}
\]

that splits the original corecursive program \( \text{unfold} \ f \) into the composition of a worker \( \text{unfold} \ g \) and a wrapper \( \text{rep} \).

### 5.2 Exploiting Duality

Our results up to this point have been generic with respect to the choice of a category \( C \). This is helpful, because not only are the results general, they are also subject to \textit{duality}.

In category theory, the principle of duality states that if a property holds of all categories, then the \textit{dual} of that property must also hold of all categories. By applying this duality principle to our general unfold factorisation theorem, we obtain the following factorisation theorem for \( \text{fold} \)s, the categorical dual of \( \text{unfold} \)s:

**Theorem 5.** Given

\[
\begin{align*}
\text{abs} : A \to B & \quad f : F A \to A \\
\text{rep} : B \to A & \quad g : F B \to B
\end{align*}
\]

satisfying the assumptions

\[
\text{rep} \circ \text{abs} = \text{id}_A
\]

\[
g \circ F \text{abs} = \text{abs} \circ f
\]

we have the factorisation

\[
\text{fold} f = \text{rep} \circ \text{fold} g
\]

that splits the original recursive program \( \text{fold} \ f \) into the composition of a wrapper \( \text{rep} \) and a worker \( \text{fold} \ g \).

Note that now \( \text{rep} \) is required to be a left-inverse of \( \text{abs} \), rather than the other way around. If we swap their names to reflect this new situation, we see that this is a special case of the following general result, due to Scluthorpe and Hutton [24]:

**Theorem 6 (Worker-Wrapper Factorisation for Initial Algebras).** Given

\[
\begin{align*}
\text{abs} : B \to A & \quad f : F A \to A \\
\text{rep} : A \to B & \quad g : F B \to B
\end{align*}
\]

satisfying one of the assumptions

\[
(A) \hspace{1em} \text{abs} \circ \text{rep} = \text{id}_A
\]

\[
(B) \hspace{1em} \text{abs} \circ \text{rep} \circ f = \text{id}_A
\]

\[
(C) \hspace{1em} \text{fold} (\text{abs} \circ \text{rep} \circ f) = \text{fold} f
\]

and one of the conditions

\[
(1) \hspace{1em} g = \text{rep} \circ f \circ F \text{abs}
\]

\[
(2) \hspace{1em} g \circ F \text{rep} = \text{rep} \circ f \circ F \text{abs}
\]

\[
(3) \hspace{1em} f \circ F \text{abs} = \text{abs} \circ g
\]
we have the factorisation

\[ \text{fold } f = \text{abs } \circ \text{fold } g \]

If we now apply duality in turn to this theorem, we obtain an even more general version of our unfold factorisation theorem:

**Theorem 7** (Worker-Wrapper Factorisation for Final Coalgebras).

Given

\[ \text{abs }: B \rightarrow A \quad f : A \rightarrow F A \\
\text{rep }: A \rightarrow B \quad g : B \rightarrow F B \]

satisfying one of the assumptions

\( (A) \, \text{abs } \circ \text{rep } = \text{id}_A \)
\( (B) \, f \circ \text{abs } \circ \text{rep } = f \)
\( (C) \, \text{unfold } (f \circ \text{abs } \circ \text{rep }) = \text{unfold } f \)

and one of the conditions

\[ (1) \, g = F \circ \text{rep } \circ f \circ \text{abs } \]
\[ (2) \, F \circ \text{abs } \circ g = f \circ \text{abs } \]
\[ (3) \, F \circ \text{rep } \circ f = g \circ \text{rep } \]

\( (1') \, \text{unfold } B \circ g = \text{unfold } F (\text{rep } \circ f \circ \text{abs }) \)
\( (2') \, \text{unfold } g = \text{unfold } f \circ \text{abs } \)

we have the factorisation

\[ \text{unfold } f = \text{unfold } g \circ \text{rep } \]

At this point it would be reasonable to ask why we did not simply present the dualised theorem straight away. This is indeed possible, but we do not feel it would be a good approach. In particular, our systematic development that starts from a concrete example and then applies steps of abstraction, generalisation and dualisation provides both motivation and explanation for the theorem.

We now turn our attention to interpreting Theorem 7. First of all, assumptions (B) and (C) are simply generalised versions of our original assumption (A), in the sense that (A) \( \Rightarrow \) (B) \( \Rightarrow \) (C). Second, conditions (1) and (3) are alternatives to the original condition (2), providing a degree of flexibility for the user of the theorem to select the most convenient. In general these three conditions are unrelated, but any of them is sufficient to ensure that the theorem holds. Finally, the \( \beta \) conditions in the second group arise as weaker versions of the corresponding conditions in the first, i.e. (1) \( \Rightarrow \) (1\( ' \)) and (2) \( \Rightarrow \) (2\( ' \)), and given assumption (C) the two \( \beta \) conditions are equivalent. We omit proofs as they are dual to those in [24].

While the new assumptions and conditions are dual to those of the original theorem, they differ significantly in terms of their practical considerations, which we shall discuss now. Firstly, assumption (B) in the theorem for fold, i.e. \( \text{abs } \circ \text{rep } \circ f = f \), can be interpreted as “for any \( x \) in the range of \( f \), \( \text{abs } (\text{rep } x) = x \)”. This can therefore be proven by reasoning only about such \( x \). When applying the theorem for unfold, this kind of simple reasoning for assumption (B) is not possible, as \( f \) is now applied last rather than first and hence cannot be factored out of the proof.

Secondly, condition (2) in the fold case, i.e. \( \text{rep } \circ f = g \circ F \text{rep } \), allowed \( g \) to depend on a precondition set up by \( f \). If such a precondition is desired for the unfold case, condition (3) must be used. This has important implications for use of this theorem as a basis for optimisation, as we will often derive \( g \) based on a specification given by one of the conditions.

Finally, we note that proving (C\( ' \)), (1\( ' \)) or (2\( ' \)) for the fold case usually requires induction. To prove the corresponding properties for the unfold case will usually require the less widely-understood technique of coinduction. These properties may therefore turn out to be less useful in the unfold case for practical purposes, despite only requiring a technique of comparable complexity. If we want to use this theory to avoid coinduction altogether, assumption (C) and the \( \beta \) conditions are not applicable. Section 5.3 offers a way around this problem in the case of (C).

### 5.3 Refining Assumption (C)

As it stands, assumption (C) is expressed as an equality between two corecursive programs defined using unfold, and hence may be non-trivial to prove. However, we can derive an equivalent assumption that may be easier to prove in practice:

\[ \text{unfold } f = \text{unfold } (f \circ \text{abs } \circ \text{rep }) \]
\[ \Rightarrow \{ \text{uniqueness property of unfold } (f \circ \text{abs } \circ \text{rep }) \} \]
\[ out \circ \text{unfold } f = F (\text{unfold } f) \circ f \circ \text{abs } \circ \text{rep } \]
\[ \Rightarrow \{ \text{unfold } f \text{ is a homomorphism } \} \]
\[ out \circ \text{unfold } f = \text{out } \circ \text{unfold } f \circ \text{abs } \circ \text{rep } \]
\[ \Rightarrow \{ \text{out } \text{ is an isomorphism } \} \]
\[ \text{unfold } f = \text{unfold } f \circ \text{abs } \circ \text{rep } \]

We denote this equivalent version of assumption (C) as (C\( ' \)). As this new assumption concerns only the conversions \( \text{abs } \) and \( \text{rep } \) along with the original program \( \text{unfold } f \), it may be provable simply from the original program’s correctness properties.

Assumption (C\( ' \)) also offers a simpler proof of Theorem 7 than one obtains by dualising the proof in [24]. We start from this assumption and use the fact that in this context, conditions (1), (2) and (1\( ' \)) all imply (2\( ' \)):

\[ \text{unfold } f = \text{unfold } f \circ \text{abs } \circ \text{rep } \]
\[ \Rightarrow \{ 2\( ' \); unfold } g = \text{fold } f \circ \text{abs } \}
\[ \text{unfold } f = \text{fold } g \circ \text{rep } \]

The proof in the case of condition (3) remains the same as previously. We conclude by noting that the implication (B) \( \Rightarrow \) (C\( ' \)) is not as obvious as the original implication (B) \( \Rightarrow \) (C). Altering the theory in this manner thus “moves work” from proving the main result to proving the relationships between the conditions.

### 5.4 Applying the Theory in Haskell

The category that is usually used to model Haskell types and functions is CPO, the category of (pointed) complete partial orders and continuous functions. While a fold operator can be defined in this category, its uniqueness property carries a strictness condition [18]. As a result, the worker-wrapper theory for folds in CPO requires a strictness condition of its own [24]. However, in the case of unfold in CPO the uniqueness property holds with no side conditions [18] so our theorem can be freely used to reason about Haskell programs without such concerns.

### 6. Examples

We now present a collection of worked examples, demonstrating how our new factorisation theorem may be applied. Firstly, we revisit the tabulation example, and show that it is a simple application of the theory. Secondly, we consider the problem of flattening a list, where we reduce the time cost by delaying expensive operations and performing them in a batch. We believe that this will be a common use of our theory. Thirdly, we consider the problem of taking the initial segment of a list, which allows us to demonstrate how sometimes different choices of condition can lead to the same result. Finally, we consider the problem of flattening a tree. In all cases the proofs are largely mechanical and the main inspiration necessary is in the choice of a new data representation.

With the exception of the initial segment example, all of the following examples occur in the context of the category Set of sets.
and total functions. Working in Set results in simpler reasoning as we do not need to consider issues of partiality.

6.1 Example: Tabulating a Function

We can instantiate our theory as shown above to give a proof of the correctness of our tabulate example. The proof uses assumption (A) and condition (2). Therefore we see that this example is a simple application of the worker-wrapper machinery.

6.2 Example: Cycling a List

The function cycle takes a non-empty finite list and produces the stream consisting of repetitions of that list. For example:

cycle [1, 2, 3] = [1, 2, 3, 1, 2, 3, 1, 2, 3, ...]

One possible definition for cycle is as follows, in which we write \([a]^+\) for the type of non-empty lists of type \(a\):

cycle :: \([a]^+\) -> Stream a
cycle (x : xs) = x : cycle (xs ++ [x])

However, this definition is inefficient, as the append operator ++ takes linear time in the length of the input list. Recalling that Stream \(a\) is the final coalgebra of the functor \(F\), we can rewrite cycle as an unfold:

cycle = unfold h t
  where h xs = head xs
t xs = tail xs ++ [head xs]

The idea we shall apply to improve the performance of cycle is to combine several ++ operations into one, thus reducing the average cost. To achieve this, we create a new representation where the original list of type \([a]^+\) is augmented with a (possibly empty) list of elements that have been added to the end. We keep this second list in reverse order so that appending a single element is a constant-time operation. The rep and abs functions are as follows:

\[
\begin{align*}
  \text{rep} :: [a]^+ & \rightarrow ([a]^+, [a]) \\
  \text{rep} \; \text{xs} & = (\text{xs}, []) \\
  \text{abs} :: ([a]^+, [a]) & \rightarrow [a]^+ \\
  \text{abs} \; (\text{xs}, \text{ys}) & = \text{xs} \mathbin{\oplus} \text{reverse} \; \text{ys}
\end{align*}
\]

Given these definitions it is easy to verify assumption (A):

\[
\begin{align*}
  \text{abs} \; (\text{rep} \; \text{xs}) & = \{ \text{definition of \text{rep}} \} \\
  \& \text{abs} \; (\text{xs}, []) & = \{ \text{definition of \text{abs}} \} \\
  \& \text{xs} \mathbin{\oplus} \text{reverse} \; [] & = \{ \text{definition of \text{reverse}} \} \\
  \& \text{xs} \mathbin{\oplus} [] & = \{ [] \text{is unit of \oplus} \} \\
  \& \text{xs} &
\end{align*}
\]

For this example we take condition (2), i.e. \(F\) abs \circ g = f \circ abs\), as our specification of \(g\), once again specialising to the two conditions \(h \circ \text{abs} = h'\) and \(t \circ \text{abs} = t'\). From this we can calculate \(h'\) and \(t'\) separately. First we calculate \(h'\):

\[
\begin{align*}
  h' \; (\text{xs}, \text{ys}) & = \{ \text{specification} \} \\
  h \; (\text{abs} \; (\text{xs}, \text{ys})) & = \{ \text{definition of \text{abs}} \} \\
  h \; (\text{xs} \mathbin{\oplus} \text{reverse} \; \text{ys}) & = \{ \text{definition of \text{h}} \} \\
  \text{head} \; (\text{xs} \mathbin{\oplus} \text{reverse} \; \text{ys}) & = \{ \text{xs is nonempty} \} \\
  \text{head} \; \text{xs} &
\end{align*}
\]

Now we calculate a definition for \(t'\). Starting from the specification \(abs \circ t' = t \circ \text{abs}\), we calculate as follows:

\[
\begin{align*}
  t \; (\text{abs} \; (\text{xs}, \text{ys})) & = \{ \text{definition of \text{abs}} \} \\
  t \; (\text{xs} \mathbin{\oplus} \text{reverse} \; \text{ys}) & = \{ \text{case analysis} \} \\
  \text{case \text{xs}} \; \text{of} & \\
  & \{ \text{definition of \text{abs}} \} \\
  & \{ \text{definition of \text{reverse}, \oplus} \} \\
  & \{ \text{pulling \text{abs} out of cases} \} \\
  \text{abs} \; (\text{case \text{xs}} \; \text{of} & \\
  & \{ \text{definition of \text{abs}} \} \\
  & \{ \text{definition of \text{reverse}, \oplus} \} \\
  & \text{xs} \mathbin{\oplus} \text{reverse} \; (\text{xs} \mathbin{\oplus} \text{reverse} \; \text{ys}) & = \{ \text{xs, x : ys} \}
\end{align*}
\]

Hence, \(t'\) can be defined as follows:

\[
\begin{align*}
  t' \; ([\text{x}], \text{ys}) & = \text{reverse} \; (\text{x} : \text{ys}), [] & [\text{xs} = (\text{x}, \text{ys})] \\
  t' \; (\text{x} : \text{xs} \mathbin{\oplus} \text{ys}) & = (\text{x} : \text{xs}, \text{ys}) \\
\end{align*}
\]

In conclusion, by applying our worker-wrapper theorem, we have calculated a factorised version of cycle:

cycle = unfold h' \; t' \circ \text{rep}
  where h' \; (\text{xs}, \text{ys}) = \text{head} \; \text{xs}
  t' \; ([\text{x}], \text{ys}) = \text{reverse} \; (\text{x} : \text{ys}), []
  t' \; (\text{x} : \text{xs} \mathbin{\oplus} \text{ys}) = (\text{x}, \text{xs} : \text{ys})

which can be written directly as

\[
\begin{align*}
  \text{cycle} & = \text{cycle}' \circ \text{rep}
  \quad \text{where}
  \quad \text{cycle}' \; ([\text{x}], \text{ys}) & = \text{x} : \text{cycle}' \; (\text{reverse} \; (\text{x} : \text{ys}), [])
  \quad \text{cycle}' \; (\text{x} : \text{xs} \mathbin{\oplus} \text{ys}) & = \text{cycle}' \; (\text{xs}, \text{x} : \text{ys})
\end{align*}
\]

This version only performs a reverse operation once for every cycle of the input list, so the average cost to produce a single element is now constant. We believe that this kind of optimisation — in which costly operations are delayed and combined into a single operation — will be a common use of our theory.

6.3 Example: Initial Segment of a List

The function init takes a list and returns the list consisting of all the elements of the original list except the last one:

\[
\begin{align*}
  \text{init} :: [a] & \rightarrow [a] \\
  \text{init} \; [] & = \bot \\
  \text{init} \; [\text{x}] & = [\text{x}] \\
  \text{init} \; (\text{x} : \text{xs}) & = \text{x} : \text{init} \; \text{xs}
\end{align*}
\]

Here, \(\bot\) represents the failure of the function to produce a result.

(In Haskell we would not need to give this first case, but we make it explicit here for the purposes of reasoning.)
This example is particularly interesting as it does not require the function being optimised to be explicitly written as an unfold at any point. As the function in question is partial, the relevant category in this case is CPO rather than Set.

Each call of $init$ checks to see if the argument is empty. However, the argument of the recursive call can never be empty, as if it were then the second case would have been matched rather than the third. We would therefore like to perform this check only once. We can use unfold worker-wrapper to achieve this, by essentially using a “de-consed” list as our representation:

$$
\begin{align*}
\text{rep} & : [a] \rightarrow (a, [a]) \\
\text{rep} & : {} \rightarrow \bot \\
\text{rep} (x : xs) &= (x, xs) \\
abs & : (a, [a]) \rightarrow [a] \\
abs (x, xs) &= x : xs
\end{align*}
$$

In this case, assumption (A) fails:

$$
\begin{align*}
abs (\text{rep} []) &= \{ \text{definition of abs} \} \\
\bot &= \{ \text{abs is strict} \} \\
\neq [\ ]
\end{align*}
$$

If we were to rewrite $init$ as an unfold, we could then prove (B). However, we instead avoid this by using the alternative assumption (C'). This expands to $\text{init} \circ \text{abs} \circ \text{rep} = \text{init}$, which we prove by case analysis on the argument. For the empty list:

$$
\begin{align*}
\text{init} (\text{abs} (\text{rep} [])) &= \{ \text{definition of rep} \} \\
\text{init} (\bot) &= \{ \text{init is strict} \} \\
\bot &= \{ \text{definition of init} \} \\
\text{init} []
\end{align*}
$$

For the undefined value $\bot$:

$$
\begin{align*}
\text{init} (\text{abs} (\text{rep} \bot)) &= \{ \text{init, abs and rep are all strict} \} \\
\bot
\end{align*}
$$

Otherwise:

$$
\begin{align*}
\text{init} (\text{abs} (\text{rep} (x : xs))) &= \{ \text{definition of abs} \} \\
\text{init} (x, xs) &= \{ \text{definition of abs} \} \\
\text{init} (x : xs)
\end{align*}
$$

Using Condition (2β)

Firstly, we demonstrate a derivation of a new worker function that avoids writing $init$ explicitly in terms of unfold. The only condition that permits this is (2β), which expands to $\text{init}' = \text{init} \circ \text{abs}$. We can calculate the definition of $\text{init}'$ simply by applying this specification to $(x, xs)$:

$$
\begin{align*}
\text{init}' (x, xs) &= \{ \text{specification of init'} \} \\
\text{init} (x, xs) &= \{ \text{definition of abs} \} \\
\text{init} (x : xs)
\end{align*}
$$

As we used the specification of $\text{init}'$ in its own derivation, the above derivation only guarantees partial correctness. We should take a moment to convince ourselves that the resulting definition does actually satisfy the specification. In this case, both sides are clearly total, so there is no problem. In conclusion, we have derived an alternative definition for the function $init$

$$\begin{align*}
\text{init} &= \text{init}' \circ \text{rep} \\
\text{where}
\end{align*}$$

that only performs the check for the empty list once.

Note that we derived the more efficient version of $init$ using worker-wrapper factorisation for unfolds without ever writing the program in question as an unfold. This is a compelling argument for the flexibility of the theory, but we should also note that because of our choice of assumption and condition none of the extra structure of unfolds is needed, as the equality chain

$$
\begin{align*}
\text{init}' \circ \text{rep} &= \{ (2\beta) \} \\
\text{init} \circ \text{abs} \circ \text{rep} &= \{ (C') \} \\
\text{init}
\end{align*}
$$

holds regardless of whether $init$ and $\text{init}'$ are unfolds. In a sense, because we have chosen the weakest properties, the theory gives us less. This suggests that we should generally prefer to use stronger properties if possible. The use of partially-correct reasoning is also unsatisfactory, as it requires us to appeal to totality.

Using Condition (1)

If we write $init$ explicitly as an unfold, this example can also use condition (1). The unfold for (possibly finite) lists is:

$$
\begin{align*}
\text{unfold} : (a \rightarrow \text{Maybe} (b, a)) \rightarrow a \rightarrow [b] \\
\text{unfold} f x &= \text{case } f x \text{ of }
\end{align*}
$$

8 2014/1/13
We attempt to simplify this, noting that the type of \( x \) is \([a]\). Condition (1) gives us the explicit definition \( g(x) = \text{fold} \circ \text{rep} \).

Using this, we can define \( \text{init} \) as follows:

\[
\text{init} :: [a] \rightarrow [a] \\
\text{init} = \text{fold} \circ g
\]

where

\[
\begin{align*}
\text{f} & :: [a] \rightarrow \text{Maybe} (a, [a]) \\
\text{f} [\text{[]}] & = \text{Nothing} \\
\text{f} [\text{x}] & = \text{Just} (x, \text{xs})
\end{align*}
\]

In order to construct a function \( g \) such that \( \text{init} = \text{fold} \circ \text{rep} \), where \( \text{rep} \) is defined as before, we use worker-wrapper factorisation. Condition (1) gives us the explicit definition \( g = \text{F} \circ \text{rep} \circ f \circ \text{abs} \). Instantiating this for our particular \( \text{F} \), we have:

\[
g(x) = \text{case } f(a, [a]) \text{ of}
\begin{align*}
\text{Nothing} & \rightarrow \text{Nothing} \\
\text{Just} (y, x') & \rightarrow \text{Just} (y, \text{rep} x')
\end{align*}
\]

We attempt to simplify this, noting that the type of \( x \) is \((a, [a])\).

We calculate \( g \) separately for the input \((a, \text{[]})\):

\[
g(a, \text{[]}) = \{ \text{definition of } g \} \\
g(a, a) = \{ \text{definition of } g \}
\]

and for \((a, as)\), where \( as \) is non-empty:

\[
g(a, as) = \{ \text{definition of } g \} \\
g(a, a') = \{ \text{definition of } g \}
\]

We thus obtain the following definition of \( g \):

\[
g(a, [a]) = \text{Nothing} \\
g(a, a', as) = \text{Just} (a, (a', as))
\]

Because we are working in CPO, we must also consider the behaviour of the function \( g \) on the input \((a, \bot)\):

\[
g(a, \bot) = \{ \text{specification of } g \} \\
g(a, a', \text{as}) = \{ \text{definition of } g \}
\]

\[
\text{case } f(a : \bot) \text{ of}
\begin{align*}
\text{Nothing} & \rightarrow \text{Nothing} \\
\text{Just} (y, x') & \rightarrow \text{Just} (y, \text{rep} x') \\
\end{align*}
\]

However, because \( g \) is defined by pattern matching on the second component of the tuple, the equation \( g(a, \bot) = \bot \) clearly holds. Therefore, the above definition of \( g \) satisfies worker-wrapper condition (1) and so the factorisation

\[
\text{init} = \text{fold} \circ \text{rep}
\]

is correct. Note that \( \text{fold} \circ \text{rep} \) is precisely the \( \text{init} \) function that we derived above, now defined as an unfold.

While we had to write \( \text{init} \) explicitly as an unfold to perform this derivation, the calculation of the improved program was more straightforward than before, and largely mechanical.

### Remarks

This above example shows that the same optimised function can sometimes be obtained using different approaches. It is worth noting that this particular optimisation is also an instance of call-pattern specialisation [23], as implemented in the Glasgow Haskell Compiler. However, neither one of these approaches subsumes the other, it simply happens in this case that they coincide.

#### 6.4 Example: Flattening a Tree

Our final example concerns the left-to-right traversal of a binary tree. The naïve way to implement such a traversal is as follows, which corresponds to expressing the function as a fold:

\[
data \text{Tree } a = \text{Null} | \text{Fork} (\text{Tree } a) a (\text{Tree } a)
\]

\[
\text{flatten} :: \text{Tree } a \rightarrow [a]
\]

\[
\text{flatten } \text{Null} = \text{[]}
\]

\[
\text{flatten } (\text{Fork } t1 x t2) = \text{flatten } t1 + x + \text{flatten } t2
\]

However, this approach takes time quadratic in the number of nodes. By fusing this definition with \( id = \text{foldout } \text{out} \), we can transform the function into an unfold. Here is the fusion condition:

\[
(\text{out } \circ \text{flatten}) t = \text{case } g t \text{ of}
\begin{align*}
\text{Nothing} & \rightarrow \text{Nothing} \\
\text{Just} (x, x') & \rightarrow \text{Just} (x, \text{flatten } t')
\end{align*}
\]

By writing a function \( \text{removemin} \) satisfying the specification of \( g \), we obtain the following new definition of \( \text{flatten} \):

\[
\text{flatten} :: \text{Tree } a \rightarrow [a]
\]

\[
\text{flatten } = \text{fold } \text{unfold } \text{removemin}
\]

where

\[
\text{removemin } \text{Null} = \text{Nothing}
\]

\[
\text{removemin } (\text{Fork } t1 x t2) =
\begin{align*}
\text{case } \text{removemin } t1 \text{ of}
\text{Nothing} & \rightarrow \text{Just} (x, t2) \\
\text{Just} (y, t1') & \rightarrow \text{Just} (y, \text{Fork } t1' x t2)
\end{align*}
\]

This approach takes time proportional to \( n \times l \), where \( n \) is the number of nodes and \( l \) is the leftwards depth of the tree, i.e. the depth of the deepest node counting only left branches.

However, we can use our worker-wrapper approach to improve this further, exploiting the isomorphism between lists of rose trees (trees with an arbitrary number of children for each node) and binary trees. First we define the type of rose trees:
data RoseTree a = RoseTree a [RoseTree a]

Now we define a version of flatten for lists of rose trees, in which the list acts as a priority queue of the elements in the original tree, so that at each stage we remove the root of the first tree in the queue, and push all its children onto the front of the queue:

\[
\text{flatten'} :: [\text{RoseTree } a] \rightarrow [a] \\
\text{flatten'} = \text{unfold}
\]

where
\[
g(\text{RoseTree } x \ ts1 : ts2) = \text{Just } (x, ts1 \uplus ts2)
\]

We define \(\text{rep}\) and \(\text{abs}\) functions to convert between this priority queue representation and the original binary tree representation.

\[
\text{rep} :: \text{Tree } a \rightarrow [\text{RoseTree } a] \\
\text{rep} = \text{reverse} \circ \text{listify}
\]

where
\[
\text{listify } \text{Null} = [] \\
\text{listify } (\text{Fork } t1 \ ts \ t2) = \text{listify } t1
\]

\[
\text{abs} :: [\text{RoseTree } a] \rightarrow \text{Tree } a \\
\text{abs} = \text{delistify} \circ \text{reverse}
\]

where
\[
\text{delistify } [] = \text{Null} \\
\text{delistify } (\text{RoseTree } x \ ts1 : ts2) = \\
\text{Fork} \ (\text{delistify} \ ts2) \ (\text{abs } ts1)
\]

Essentially, \(\text{rep}\) pulls apart a tree along its left branch into a list, while \(\text{abs}\) puts the tree back together. The use of \(\text{reverse}\) is necessary to ensure that the leftmost node is at the head of the list.

The result is an alternate definition \(\text{flatten} = \text{flatten'} \circ \text{rep}\) that has comparable performance on balanced trees, but much better performance on trees with long left branches.

We now verify the correctness of this new definition using our worker-wrapper theorem. Firstly, we show that assumption (A) holds, i.e. \(\text{abs} \circ \text{rep} = \text{id}\). To do this we note that because \(\text{reverse}\) is self-inverse, \(\text{abs} \circ \text{rep} = \text{delistify} \circ \text{listify}\). We prove \(\text{delistify} \circ \text{listify} = \text{id}\) by induction on trees. First, the base case:

\[
\text{delistify } \text{listify } \text{Null} = \{ \text{definition of listify} \} \\
\text{delistify } [] = \{ \text{definition of delistify} \} \\
\text{Null}
\]

Then the inductive case:

\[
\text{delistify } (\text{listify } (\text{Fork } t1 \ ts t2)) = \\
\{ \text{definition of listify} \} \\
\text{delistify } (\text{RoseTree } x (\text{rep } ts) : \text{listify } t1) = \\
\{ \text{definition of delistify} \} \\
\text{Fork} \ (\text{delistify } \text{listify } t1) \ x \ (\text{abs } (\text{rep } ts)) = \\
\{ \text{abs} \circ \text{rep} = \text{delistify} \circ \text{listify} \} \\
\text{Fork} \ (\text{delistify } \text{listify } t1) \ x \ (\text{delistify } \text{listify } ts) = \\
\{ \text{inductive hypothesis} \} \\
\text{Fork } t1 x ts t2
\]

Now we must prove that \(g\) satisfies one of the worker-wrapper specifications. We choose condition (2) to verify, in this case:

\[
\text{removemin } (\text{abs } ts) = \\
\text{case } g ts \ of \\
\text{Nothing} \rightarrow \text{Nothing} \\
\text{Just } (x, ts') \rightarrow \text{Just } (x, \text{abs } ts')
\]

We verify this equation by induction on the length of the priority queue. For the base case when the queue is empty, we have:

\[
\text{removemin } (\text{abs } []) = \\
\{ \text{definition of abs} \}
\]

For the inductive case, rather than \(\text{RoseTree } x \ ts1 : ts2\) we use \(ts1 \uplus [\text{RoseTree } x \ ts2]\), as \(\text{abs}\) reverses its argument:

\[
\text{removemin } (\text{abs } (ts1 \uplus [\text{RoseTree } x \ ts2])) = \\
\{ \text{definition of abs} \} \\
\text{removemin } \ (\text{delistify} \ (\text{RoseTree } x \ ts2 : \text{reverse } ts1)) = \\
\{ \text{definition of delistify} \} \\
\text{removemin } \ (\text{delistify} \ (\text{reverse } ts1)) = \\
\{ \text{abs } \circ \text{delistify } \circ \text{reverse} \} \\
\text{removemin } (\text{Fork } (\text{abs } ts1) x (\text{abs } ts2)) = \\
\{ \text{definition of removemin} \} \\
\text{case } \text{removemin } (\text{abs } ts1) \ of \\
\text{Nothing} \rightarrow \text{Just } (x, \text{abs } ts2) \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{Fork } ts'1 x (\text{abs } ts2)) = \\
\{ \text{inductive hypothesis} \} \\
\text{case } \text{case } (g ts1) \ of \\
\text{Nothing} \rightarrow \text{Nothing} \\
\text{Just } (y, ts'1) \rightarrow \text{Just } (y, \text{abs } ts'1 x (\text{abs } ts2)) = \\
\{ \text{case of case, pattern matching} \} \\
\text{case } g ts1 \ of \\
\text{Nothing} \rightarrow \text{Just } (x, \text{abs } ts2) \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{Fork } ts'1 x (\text{abs } ts2)) = \\
\{ \text{definition of abs} \} \\
\text{case } g ts1 \ of \\
\text{Nothing} \rightarrow \text{Just } (x, \text{abs } ts2) \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{Fork } (\text{delistify } (\text{reverse } ts1)) x (\text{abs } ts2)) = \\
\{ \text{definition of abs} \} \\
\text{case } g ts1 \ of \\
\text{Nothing} \rightarrow \text{Just } (x, \text{abs } ts2) \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{abs } (ts1' + [\text{RoseTree } x \ ts2])) = \\
\{ \text{definition of } g, \text{ case} \}
\]

We must prove that this last expression is equal to

\[
\text{case } g \ (ts1 \uplus [\text{RoseTree } x \ ts2]) \ of \\
\text{Nothing} \rightarrow \text{Nothing} \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{abs } (ts1' + [\text{RoseTree } x \ ts2]))
\]

which we do by case analysis on ts1. When ts1 = [], we have:

\[
\text{case } g \ [] \ of \\
\text{Nothing} \rightarrow \text{Just } (x, \text{abs } ts2) \\
\text{Just } (y, ts'1) \rightarrow \\
\text{Just } (y, \text{abs } (ts1' + [\text{RoseTree } x \ ts2])) = \\
\{ \text{definition of } g, \text{ case} \}
\]
In 1998, Gibbons and Jones published the paper The Under-Appeared unfold [9], which gave the inspiration for the title of this paper. The paper argued that unfold was an underutilised programming tool, and justified this by presenting algorithms for breadth-first traversal using both fold and unfold, arguing that the unfold-based algorithms were clearer.

7.2 Unfold in Category Theory

It seems that categorical unfolds (also known as anamorphisms) were largely developed in parallel to unfold in programming languages. The earliest mention of categorial unfold appears to be in Hagino’s 1987 PhD thesis A Categorical Programming Language [18] and subsequently in Malcolm’s 1990 thesis Algebraic Data Types and Program Transformation [17]. Interestingly, neither of these make any linguistic distinction between folds and unfolds, using the same terminology to refer to both. It seems that this distinction was not made until Meijer et al.’s 1991 paper Functional Programming with Bananas, Lenses, Envelopes and Barbed Wire [18], which concerned folds and unfolds in CPO and essentially unified the programming language work with the categorial work.

7.3 Program Fusion

In functional programming, many successful program optimisations are based upon fusion, in which separate parts of a program are fused together to eliminate intermediate data structures. Fusion was first introduced by Wadler in 1990 by the name of deforestation [31]. Since then, it has been widely explored, especially in regards to specific recursion patterns.

A particularly successful example is foldr/foldl fusion, in which list-producers are defined in terms of a function build while list-consumers are defined in terms of foldr. Introduced by Gill, Launchbury and Peyton Jones [11], this pattern has been the subject of much research [2, 26, 30] and is implemented in GHC.

A more recent innovation by Coupts, Leshchinsky and Stewart is stream fusion, a way to optimise list-processing functions by changing the representation of lists [4]. This approach makes essential use of an unfold-like operator, and is related to worker-wraper factorisation as it involves a change of intermediate data type. However, stream fusion utilises the recursive structure of the underlying data, whereas our approach does not.

7.4 Program Factorisation

Compared to fusion-based techniques, program factorisation or “fission” seems far less well-explored. Gibbons’ 2006 paper, Fusion for Program Comprehension [7], presents an application of this idea which differs from our work in two ways. Firstly, Gibbons does not concern himself with program optimisation; rather, his intended use is understanding an already-written program by breaking it down into the combination of separate parts that are easier to comprehend. Secondly, while Gibbons’ approach is based upon applying fusion in reverse, the proof of our approach to program factorisation actually involves a forward application of fusion.

8. Conclusion

In this paper, we have presented a novel approach to optimising programs written in the form of an unfold, showing how a useful approach comes from the commonly-used idea of generalising a coinductive hypothesis. We have provided a general factorisation theorem that can be used either to guide the derivation of an optimised program or to prove such a program correct, resulting in a technique that we believe has wide applicability. We demonstrated the utility of our technique with a collection of examples.
8.1 Further Work

We have only considered programs in the form of an unfold, but there are other corecursive patterns that can be considered. One example is apomorphisms [29], which capture the idea of primitive corecursion, allowing construction of the result to short-cut by producing the remainder of the result in a single step. The apomorphism operator for streams can be defined as follows:

\[
apo h t x = h x : \text{case } t x \text{ of}
\]
\[
\text{Left } x' \rightarrow apo h t x'
\]
\[
\text{Right } x \rightarrow x
\]

Apomorphisms have their own fusion law, so it seems likely that they would have a useful worker-wrapper factorisation theorem.

Another possible direction is to adapt our method to deal with circular definitions. For example, we can write a circular definition of the infinite stream of Fibonacci numbers:

\[\text{fibs} = 0 : 1 : \text{zipWith } (+) \text{ (tail fibs)}\]

Coinductive techniques can be used to reason about circular definitions like this one, but whether a factorisation theorem analogous to ours exists for such definitions remains to be seen.

We could also consider extending this work to monadic and comonadic unfoldings. Monadic unfolds [21] are of particular interest; consider the monadic unfold operator for streams:

\[
\text{unfoldM :: Monad } m \Rightarrow (a \rightarrow m \ (b, a)) \rightarrow a \rightarrow m \ (\text{Stream } b)
\]

\[
\text{unfoldM } f x = \text{do } (b, x') \leftarrow f x
\]
\[
bs \leftarrow \text{unfoldM } f x'
\]
\[
\text{return } (b : bs)
\]

Any monad with a strict bind operator will fail to produce a result, instead simply bottoming out. Intuitively, the problem is that an infinite number of effects must be applied before the final result can be produced. Thus we see that there is a fundamental difference between ordinary and monadic unfolds that raises interesting questions concerning the worker-wrapper theory.

The theory we have presented is concerned with correctness. In order to reason about efficiency gains, we also need an operational theory, for which purposes we are currently exploring the use of improvement theory [19]. Finally, while we have noted that the proofs are often mechanical, we have yet to consider how our new theory may be mechanised. A team at the University of Kansas is currently working on the implementation of various worker-wrapper theories as an extension to the Glasgow Haskell Compiler [5, 25], with promising initial results.

Acknowledgments

The authors would like to thank Richard Bird for assistance with tracing the history of the unfold operator. We would also like to thank Philippa Cowdery for the observation in section 6.4 that the unfold-based \textit{flatten} could be derived from the fold-based \textit{flatten}.

References