Metamorphisms: Streaming Representation-Changers

Jeremy Gibbons

Computing Laboratory, University of Oxford
www.comlab.ox.ac.uk/oucl/work/jeremy.gibbons
jeremy.gibbons@comlab.ox.ac.uk

Abstract

Unfolds generate data structures, and *folds* consume them. A *hylomorphism* is a fold after an unfold, generating then consuming a *virtual data structure*. A *metamorphism* is the opposite composition, an unfold after a fold; typically, it will convert from one data representation to another. In general, metamorphisms are less interesting than hylomorphisms: there is no automatic *fusion* to *deforest* the intermediate virtual data structure. However, under certain conditions fusion is possible: some of the work of the unfold can be done before all of the work of the fold is complete. This permits *streaming metamorphisms*, and among other things allows conversion of *infinite data representations*. We present a theory of metamorphisms and outline some examples.

1 Introduction

Folds and *unfolds* in functional programming [1–3] are well-known tools in the programmer's toolbox. Many programs that consume a data structure follow the pattern of a fold; and dually, many that produce a data structure do so as an unfold. In both cases, the structure of the program is determined by the structure of the data it processes — to invert the architect Louis Sullivan's dictum [4], one might say that *function follows form*.

It is natural to consider also compositions of these operations. Meijer [5] coined the term *hylomorphism* for the composition of a fold after an unfold. The *virtual data structure* [6] produced by the unfold is subsequently consumed by the fold; the structure of that data determines the structure of both its producer and its consumer. Under certain rather weak conditions, the intermediate data structure may be eliminated or *deforested* [7], and the two phases fused into one slightly more efficient one.

In this paper, we consider the opposite composition, of an unfold after a fold. Programs of this form consume an input data structure using a fold, constructing some intermediate (possibly unstructured) data, and from this intermediary produce an output data structure using an unfold. Note that the two data structures may now be of different shapes, since they do not meet. Indeed, such programs may often be thought of as *representation changers*, converting from one structured representation of some abstract data to a different structured representation of the same data. Despite the risk of putting the reader off with yet another neologism of Greek origin, we cannot resist coining the term *metamorphism* for such compositions, because they typically metamorphose representations.

In general, metamorphisms are perhaps less interesting than hylomorphisms, because there is no nearly-automatic deforestation. Nevertheless, sometimes fusion is possible; under certain conditions, some of the unfolding may be performed before all of the folding is complete. This kind of fusion can be helpful for controlling the size of the intermediate data. Perhaps more importantly, it can allow conversions between infinite data representations. For this reason, we call such fused metamorphisms *streaming algorithms*; they are the main subject of this paper. We encountered them fortuitously while trying to describe some data compression algorithms [8], but have since realized that they are an interesting construction in their own right.

The remainder of this article is organized as follows. Section 2 summarizes the theory of folds and unfolds. Section 3 introduces metamorphisms, which are unfolds after folds. Section 4 presents a theory of streaming, which is the main topic of the paper. Section 5 provides an extended application of streaming, and Section 6 outlines two other applications and a unifying scheme described in more detail elsewhere. Finally, Section 7 discusses some ideas for generalizing the currently rather list-oriented theory, and describes related work. This article is an expansion of an earlier version [9] presented at *Mathematics of Program Construction 2004*.

2 Origami programming

We are interested in capturing and studying *recurring patterns of computation*, such as folds and unfolds. As has been strongly argued by the recently popular *design patterns* movement [10], identifying and exploring such patterns has many benefits: reuse of abstractions, rendering 'folk knowledge' in a more accessible format, providing a common vocabulary of discourse, and so on. What distinguishes patterns in functional programming from patterns in object-oriented and other programming paradigms is that the better 'glue' available in the former [11] allows the patterns to be expressed as *abstractions*

within the language, rather than having to resort to extra-linguistic prose and diagrams.

We use the notation of Haskell [12], the de facto standard lazy functional programming language, except that we take the liberty to use some typographic effects in formatting, and to elide some awkwardnesses such as type coercions and qualifications that are necessary for programming but that obscure the points we are trying to make. (The true story, warts and all, is included in an appendix.)

Most of this paper involves the datatype of lists:

$$\mathbf{data}\ [\alpha] \mathrel{\widehat{=}} [\,] \mid \alpha : [\alpha]$$

That is, the datatype $[\alpha]$ of lists with elements of type α consists of the empty list $[\,]$, and non-empty lists of the form $a : x$ with head a of type α (we write '$a :: \alpha$') and tail $x :: [\alpha]$. Note that the datatype consists of both finite and infinite lists; for simplicity, we avoid consideration of partial lists.

The primary patterns of computation over such lists are the *fold*, which consumes a list and produces some value:

$$\begin{aligned}
&\mathsf{foldr} && :: (\alpha \to \beta \to \beta) \to \beta \to [\alpha] \to \beta \\
&\mathsf{foldr}\ f\ b\ [\,] && \mathrel{\widehat{=}} b \\
&\mathsf{foldr}\ f\ b\ (a : x) && \mathrel{\widehat{=}} f\ a\ (\mathsf{foldr}\ f\ b\ x)
\end{aligned}$$

and the *unfold* [13], which produces a list from some seed:

$$\begin{aligned}
&\mathsf{unfoldr} && :: (\beta \to \mathsf{Maybe}\ (\alpha, \beta)) \to \beta \to [\alpha] \\
&\mathsf{unfoldr}\ f\ b && \mathrel{\widehat{=}} \mathbf{case}\ f\ b\ \mathbf{of} \\
& && \quad \mathsf{Just}\ (a, b') \to a : \mathsf{unfoldr}\ f\ b' \\
& && \quad \mathsf{Nothing} \to [\,]
\end{aligned}$$

Here, the datatype Maybe augments a type α with an additional value Nothing:

$$\mathbf{data}\ \mathsf{Maybe}\ \alpha \mathrel{\widehat{=}} \mathsf{Nothing} \mid \mathsf{Just}\ \alpha$$

We will later make use of the related sum datatype constructor Either. An element of type $\mathsf{Either}\ \alpha\ \beta$ consists either of an element of type α or of one of type β:

$$\mathbf{data}\ \mathsf{Either}\ \alpha\ \beta \mathrel{\widehat{=}} \mathsf{Left}\ \alpha \mid \mathsf{Right}\ \beta$$

The foldr pattern consumes list elements from right to left, following the way lists are constructed; as a variation on this, there is another fold which consumes elements from left to right:

$$\begin{aligned}
&\mathsf{foldl} && :: (\beta \to \alpha \to \beta) \to \beta \to [\alpha] \to \beta \\
&\mathsf{foldl}\ f\ b\ [\,] && \mathrel{\widehat{=}} b \\
&\mathsf{foldl}\ f\ b\ (a : x) && \mathrel{\widehat{=}} \mathsf{foldl}\ f\ (f\ b\ a)\ x
\end{aligned}$$

We also use the operator scanl, which is like foldl but which returns all partial results instead of just the final one:

$$\begin{aligned}
&\mathsf{scanl} && :: (\beta \to \alpha \to \beta) \to \beta \to [\alpha] \to [\beta] \\
&\mathsf{scanl}\ f\ b\ [\,] && \mathrel{\widehat{=}} [b] \\
&\mathsf{scanl}\ f\ b\ (a : x) && \mathrel{\widehat{=}} b : \mathsf{scanl}\ f\ (f\ b\ a)\ x
\end{aligned}$$

We introduce also a datatype of internally-labelled binary trees:

$$\mathbf{data}\ \mathsf{Tree}\ \alpha \mathrel{\widehat{=}} \mathsf{Node}\ (\mathsf{Maybe}\ (\alpha, \mathsf{Tree}\ \alpha, \mathsf{Tree}\ \alpha))$$

with fold operator

$$\begin{aligned}
&\mathsf{foldt} && :: (\mathsf{Maybe}\ (\alpha, \beta, \beta) \to \beta) \to \mathsf{Tree}\ \alpha \to \beta \\
&\mathsf{foldt}\ f\ (\mathsf{Node\ Nothing}) && \mathrel{\widehat{=}} f\ \mathsf{Nothing} \\
&\mathsf{foldt}\ f\ (\mathsf{Node}\ (\mathsf{Just}\ (a, t, u))) && \mathrel{\widehat{=}} f\ (\mathsf{Just}\ (a, \mathsf{foldt}\ f\ t, \mathsf{foldt}\ f\ u))
\end{aligned}$$

and unfold operator

$$\begin{aligned}
&\mathsf{unfoldt} && :: (\beta \to \mathsf{Maybe}\ (\alpha, \beta, \beta)) \to \beta \to \mathsf{Tree}\ \alpha \\
&\mathsf{unfoldt}\ f\ b && \mathrel{\widehat{=}} \mathbf{case}\ f\ b\ \mathbf{of} \\
&\qquad && \quad \mathsf{Nothing} && \to \mathsf{Node\ Nothing} \\
&\qquad && \quad \mathsf{Just}\ (a, b_1, b_2) && \to \mathsf{Node}\ (\mathsf{Just}\ (a, \mathsf{unfoldt}\ f\ b_1, \mathsf{unfoldt}\ f\ b_2))
\end{aligned}$$

It would perhaps be more elegant to define lists and their recursion patterns in the same style as done here for trees, but for consistency with the Haskell standard libraries we adopt their definitions instead. We could also condense the above code by using various higher-order combinators, but for accessibility we refrain from doing so.

The remaining notation will be introduced as it is encountered. For more examples of the use of these and related patterns in functional programming, see [14], and for the theory behind this approach to programming, see [15]; for a slightly different view of both in a relational setting, see [3].

3 Metamorphisms

In this section we present three simple examples of metamorphisms, or representation changers in the form of unfolds after folds. These three represent the entire spectrum of possibilities: it turns out that the first permits streaming automatically (assuming lazy evaluation), the second does so with some work, and the third does not permit it at all.

3.1 Reformatting lines

The classic application of metamorphisms is for dealing with *structure clashes* [16]: data is presented in a format that is inconvenient for a particular kind of processing, so it needs to be rearranged into a more convenient format. For example, a piece of text might be presented in 70-character lines, but required for processing in 60-character lines. Rather than complicate the processing by having to keep track of where in a given 70-character line a virtual 60-character line starts, good practice would be to separate the concerns of rearranging the data and of processing it. A *control-oriented* or *imperative* view of this task can be expressed in terms of coroutines: one coroutine, the processor, repeatedly requests the other, the rearranger, for the next 60-character line. A *data-oriented* or *declarative* view of the same task consists of describing the intermediate data structure, a list of 60-character lines. With lazy evaluation, the two often turn out to be equivalent; but the data-oriented view may be simpler, and is certainly the more natural view in functional programming.

We define the following Haskell functions.

$$
\begin{array}{lll}
\mathit{reformat} & :: & \mathit{Integer} \to [[\alpha]] \to [[\alpha]] \\
\mathit{reformat}\ n & \mathrel{\widehat{=}} & \mathit{writeLines}\ n \cdot \mathit{readLines} \\
 & & \\
\mathit{readLines} & :: & [[\alpha]] \to [\alpha] \\
\mathit{readLines} & \mathrel{\widehat{=}} & \mathsf{foldr}\ (+\!\!+)\ [\,] \\
\mathit{writeLines} & :: & \mathit{Integer} \to [\alpha] \to [[\alpha]] \\
\mathit{writeLines}\ n & \mathrel{\widehat{=}} & \mathsf{unfoldr}\ (\mathit{split}\ n) \quad \textbf{where}\ \mathit{split}\ n\ [\,] \mathrel{\widehat{=}} \mathsf{Nothing} \\
 & & \qquad\qquad\qquad\qquad\qquad\qquad\quad \mathit{split}\ n\ x \mathrel{\widehat{=}} \mathsf{Just}\ (\mathit{splitAt}\ n\ x)
\end{array}
$$

The function *readLines* is just what is called *concat* in the Haskell standard library; we have written it explicitly as a fold here to emphasize the program structure. The function *writeLines n* partitions a list into segments of length n, the last segment possibly being short. (The operator '$\cdot$' denotes function composition, and '$+\!\!+$' is list concatenation. The function *splitAt* from the Haskell libraries breaks a list in two at a given position.)

The function *reformat* fits our definition of a metamorphism, since it consists of an unfold after a fold. Because $+\!\!+$ is non-strict in its right-hand argument, *reformat* is automatically streaming when lazily evaluated: the first lines of output can be produced before all the input has been consumed. Thus, we need not maintain the whole concatenated list (the result of *readLines*) in memory at once, and we can even reformat infinite lists of lines.

3.2 Radix conversion

Converting fractions from positional notation in one radix to another is a change of representation. We define functions *radixConvert*, *fromBase* and *toBase* as follows:

$$\begin{array}{ll} \mathit{radixConvert} & :: (\mathit{Integer}, \mathit{Integer}) \to [\mathit{Integer}] \to [\mathit{Integer}] \\ \mathit{radixConvert}\ (b, b') & \mathrel{\widehat{=}} \mathit{toBase}\ b' \cdot \mathit{fromBase}\ b \\ \\ \mathit{fromBase} & :: \mathit{Integer} \to [\mathit{Integer}] \to \mathit{Rational} \\ \mathit{fromBase}\ b & \mathrel{\widehat{=}} \mathsf{foldr}\ \mathit{step}_b\ 0 \\ \\ \mathit{toBase} & :: \mathit{Integer} \to \mathit{Rational} \to [\mathit{Integer}] \\ \mathit{toBase}\ b & \mathrel{\widehat{=}} \mathsf{unfoldr}\ \mathit{next}_b \end{array}$$

where

$$\begin{array}{lll} \mathit{step}_b\ n\ x & \mathrel{\widehat{=}} (x + n) \div b & \\ \mathit{next}_b\ 0 & \mathrel{\widehat{=}} \mathsf{Nothing} & \\ \mathit{next}_b\ x & \mathrel{\widehat{=}} \mathsf{Just}\ (\lfloor y \rfloor, y - \lfloor y \rfloor) & \textbf{where}\ y \mathrel{\widehat{=}} b \times x \end{array}$$

Thus, *fromBase* b takes a (finite) list of digits and converts it into a fraction; provided the digits are all at least zero and less than b, the resulting fraction will be at least zero and less than one. For example,

$$\mathit{fromBase}\ 10\ [2, 5] = \mathit{step}_{10}\ 2\ (\mathit{step}_{10}\ 5\ 0) = {}^{1}\!/_{4}$$

Then *toBase* b takes a fraction between zero and one, and converts it into a (possibly infinite) list of digits in base b. For example,

$$\mathit{toBase}\ 2\ ({}^{1}\!/_{4}) = 0 : \mathsf{unfoldr}\ \mathit{next}_2\ ({}^{1}\!/_{2}) = 0 : 1 : \mathsf{unfoldr}\ \mathit{next}_2\ 0 = [0, 1]$$

Composing *fromBase* for one radix with *toBase* for another effects a change of radix.

At first blush, this looks very similar in structure to the reformatting example of Section 3.1. However, now the fold operator $step_b$ is strict in its right-hand argument. Therefore, *fromBase b* must consume its whole input before it generates any output — so these conversions will not work for infinite fractions, and even for finite fractions the entire input must be read before any output is generated.

Intuitively, one might expect to be able to do better than this. For example, consider converting the decimal fraction $[2, 5]$ to the binary fraction $[0, 1]$. The initial 2 alone is sufficient to justify the production of the first bit 0 of the output: whatever follows (provided that the input really does consist of decimal digits), the fraction lies between $^2/_{10}$ and $^3/_{10}$, and so its binary representation must start with a zero. We make this intuition precise in Section 4; it involves, among other steps, inverting the structure of the traversal of the input by replacing the foldr with a foldl.

Of course, digit sequences like this are not a good representation for fractions: many useful operations turn out to be uncomputable. In Section 5, we look at a better representation. It still turns out to leave some operations uncomputable (as any non-redundant representation must), but there are fewer of them.

3.3 Sorting

As a third introductory example, we consider tree-based sorting algorithms. One such sorting algorithm is a variation on Hoare's Quicksort [17]. What makes Quicksort particularly quick is that it is in-place, needing only logarithmic extra space for the control stack; but it is difficult to treat in-place algorithms functionally, so we ignore that aspect. Structurally, Quicksort turns out to be a hylomorphism: it unfolds the input list by repeated partitioning to produce a binary search tree, then folds this tree to yield the output list.

We use the datatype of binary trees from Section 2. We also use the following functions:

$$
\begin{array}{ll}
\mathit{partition} & :: [\alpha] \to \mathsf{Maybe}\ (\alpha, [\alpha], [\alpha]) \\
\mathit{partition}\ [\,] & \mathrel{\widehat{=}} \mathsf{Nothing} \\
\mathit{partition}\ (a : x) & \mathrel{\widehat{=}} \mathsf{Just}\ (a, \mathit{filter}\ (\leq a)\ x, \mathit{filter}\ (> a)\ x) \\
 & \\
\mathit{join} & :: \mathsf{Maybe}\ (\alpha, [\alpha], [\alpha]) \to [\alpha] \\
\mathit{join}\ \mathsf{Nothing} & \mathrel{\widehat{=}} [\,] \\
\mathit{join}\ (\mathsf{Just}\ (a, x, y)) & \mathrel{\widehat{=}} x +\!\!+ [a] +\!\!+ y
\end{array}
$$

(The function *filter* from the Haskell libraries returns those elements of a list that satisfy a given predicate.) The first partitions a non-empty list into a pivot

and the smaller and larger elements, and returns Nothing given an empty list; the second concatenates a pair of lists with a given element in between, and returns the empty list given Nothing. Given these auxiliary functions, we have

$$quicksort \mathrel{\widehat{=}} \mathsf{foldt}\ join \cdot \mathsf{unfoldt}\ partition$$

as a hylomorphism.

One can sort also as a tree metamorphism: the same type of tree is an intermediate data structure, but this time it is a *minheap* rather than a binary search tree: the element stored at each node is no greater than any element below that node. Moreover, this time the tree producer is a *list fold* (rather than a tree unfold) and the tree consumer is a *list unfold* (rather than a tree fold).

We use functions

$$\begin{array}{lll}
insert & :: & \alpha \to \mathsf{Tree}\ \alpha \to \mathsf{Tree}\ \alpha \\
insert\ a\ t & \mathrel{\widehat{=}} & merge\ (\mathsf{Node}\ (\mathsf{Just}\ (a, e, e)), t) \\
 & & \quad \mathbf{where}\ e \mathrel{\widehat{=}} \mathsf{Node\ Nothing} \\
 & & \\
splitMin & :: & \mathsf{Tree}\ \alpha \to \mathsf{Maybe}\ (\alpha, \mathsf{Tree}\ \alpha) \\
splitMin\ (\mathsf{Node}\ t) & \mathrel{\widehat{=}} & \mathbf{case}\ t\ \mathbf{of} \\
 & & \quad \mathsf{Nothing} \to \mathsf{Nothing} \\
 & & \quad \mathsf{Just}\ (a, u, v) \to \mathsf{Just}\ (a, merge\ (t, u)) \\
 & & \\
merge\ (t, \mathsf{Node\ Nothing}) & \mathrel{\widehat{=}} & t \\
merge\ (\mathsf{Node\ Nothing}, u) & \mathrel{\widehat{=}} & u \\
merge\ (\mathsf{Node}\ x, \mathsf{Node}\ y) & \mathrel{\widehat{=}} & \mathbf{if}\ \ a < b \\
 & & \quad \mathbf{then}\ \mathsf{Node}\ (a, t_2, merge\ (t_1, \mathsf{Node}\ y)) \\
 & & \quad \mathbf{else}\ \ \mathsf{Node}\ (b, u_2, merge\ (u_1, \mathsf{Node}\ x)) \\
 & & \quad\quad \mathbf{where}\ \mathsf{Just}\ (a, t_1, t_2) \mathrel{\widehat{=}} x \\
 & & \quad\quad\quad\quad \mathsf{Just}\ (b, u_1, u_2) \mathrel{\widehat{=}} y
\end{array}$$

The first inserts an element into a heap; the second splits a non-empty heap into its least element and the remainder, and returns Nothing given the empty heap. Both are defined in terms of the auxilliary function *merge*, which combines two heaps into one. Given these component functions, we have

$$heapsort \mathrel{\widehat{=}} \mathsf{unfoldr}\ splitMin \cdot \mathsf{foldr}\ insert\ (\mathsf{Node\ Nothing})$$

as a metamorphism. (Contrast this description of heapsort with the one given by Augusteijn [18] in terms of hylomorphisms, driving the computation by the shape of the intermediate tree rather than the two lists.)

Here, unlike in the reformatting and radix conversion examples, there is no hope for streaming: the second phase cannot possibly make any progress until the entire input is read, because the first element of the sorted output (which is the least element of the list) might be the last element of the input. Sorting is inherently a memory-intensive process, and cannot be performed on infinite lists.

4 Streaming

Of the three examples in Section 3, one automatically permits streaming and one can never do; only one, namely radix conversion, warrants further investigation in this regard. As suggested in Section 3.2, it ought to be possible to produce some of the output before all of the input is consumed. In this section, we see how this can be done, developing some general results along the way.

4.1 The streaming theorem

The second phase of the metamorphism involves producing the output, maintaining some state in the process. That state is initialized to the result of the first phase, obtained by folding the entire input, and evolves as the output is unfolded. Streaming must involve starting to unfold from an earlier state, the result of folding only some initial part of the input. Therefore, it is natural to consider metamorphisms in which the folding phase is an instance of foldl:

$$\mathsf{unfoldr}\ f \cdot \mathsf{foldl}\ g\ c$$

Essentially the problem is a matter of finding some kind of *invariant* of this state that determines the initial behaviour of the unfold. This idea is captured by the following definition.

Definition 1 *The 'streaming condition' for f and g is: for all a, b, c and c',*

$$f\ c = \mathsf{Just}\ (b, c') \ \Rightarrow\ f\ (g\ c\ a) = \mathsf{Just}\ (b, g\ c'\ a)$$

Informally, the streaming condition states the following: if c is a state from which the unfold would produce some output element (rather than merely the empty list), then so is the modified state $g\,c\,a$ for any a; moreover, the element b output from c is the same as that output from $g\ c\ a$, and the residual states c' and $g\ c'\ a$ stand in the same relation as the starting states c and $g\ c\ a$. In

other words, 'the next output produced' is invariant under consuming another input.

This invariant property is sufficient for the unfold and the fold to be fused into a single process, which alternates (not necessarily strictly) between consuming inputs and producing outputs. We define:

$$
\begin{array}{lll}
\mathit{stream} & :: & (\gamma \to \mathsf{Maybe}\,(\beta,\gamma)) \to (\gamma \to \alpha \to \gamma) \to \gamma \to [\alpha] \to [\beta] \\
\mathit{stream}\ f\ g\ c\ x & \mathrel{\widehat{=}} & \mathbf{case}\ f\ c\ \mathbf{of} \\
 & & \quad \mathsf{Just}\,(b,c') \to b : \mathit{stream}\ f\ g\ c'\ x \\
 & & \quad \mathsf{Nothing} \quad \to \mathbf{case}\ x\ \mathbf{of} \\
 & & \qquad\qquad a : x' \to \mathit{stream}\ f\ g\ (g\ c\ a)\ x' \\
 & & \qquad\qquad [\,] \quad\ \to [\,]
\end{array}
$$

Informally, $\mathit{stream}\ f\ g :: \gamma \to [\alpha] \to [\beta]$ involves a producer f and a consumer g; maintaining a state c, it consumes an input list x and produces an output list y. If f can produce an output element b from the state c, this output is delivered and the state revised accordingly. If f cannot, but there is an input a left, this input is consumed and the state revised accordingly. When the state is 'wrung dry' and the input is exhausted, the process terminates.

Formally, the relationship between the metamorphism and the streaming algorithm is given by the following theorem.

Theorem 2 (Streaming Theorem [8]) *If the streaming condition holds for f and g, then*

$$\mathit{stream}\ f\ g\ c\ x = \mathsf{unfoldr}\ f\ (\mathsf{foldl}\ g\ c\ x)$$

on finite lists x.

Proof. The proof is given in [8]. We prove a stronger theorem (Theorem 6) later. □

Note that the result relates behaviours on finite lists only: on infinite lists, the foldl never yields a result, so the metamorphism may not either, whereas the streaming process can be productive — indeed, that is the main point of introducing streaming in the first place.

As a simple example, consider the functions *unCons* and *snoc*, defined as follows:

$$\begin{aligned} &\mathit{unCons}\ [\,] &&\mathrel{\widehat{=}} \mathsf{Nothing} \\ &\mathit{unCons}\ (a : x) &&\mathrel{\widehat{=}} \mathsf{Just}\ (a, x) \\ &\mathit{snoc}\ x\ a &&\mathrel{\widehat{=}} x +\!\!+ [a] \end{aligned}$$

The streaming condition holds for *unCons* and *snoc*: $\mathit{unCons}\ x = \mathsf{Just}\ (b, x')$ implies $\mathit{unCons}\ (\mathit{snoc}\ x\ a) = \mathsf{Just}\ (b, \mathit{snoc}\ x'\ a)$. Therefore, Theorem 2 applies, and

$$\mathsf{unfoldr}\ \mathit{unCons} \cdot \mathsf{foldl}\ \mathit{snoc}\ [\,] = \mathit{stream}\ \mathit{unCons}\ \mathit{snoc}\ [\,]$$

on finite lists (but not infinite ones!). The left-hand side is a two-stage copying process with an unbounded intermediate buffer, and the right-hand side a one-stage copying queue with a one-place buffer.

4.2 Reversing the order of evaluation

In order to make a streaming version of radix conversion, we need to rewrite *fromBase b* as an instance of foldl rather than of foldr. Fortunately, there is a standard technique for doing this:

$$\mathsf{foldr}\ f\ b = \mathit{applyto}\ b \cdot \mathsf{foldr}\ (\cdot)\ \mathsf{id} \cdot \mathsf{map}\ f$$

where $\mathit{applyto}\ b\ f \mathrel{\widehat{=}} f\ b$. Because composition is associative with unit id, the foldr on the right-hand side can — by the First Duality Theorem [19] — be replaced by a foldl.

Although valid, this is not always very helpful. In particular, it can be quite inefficient — the fold now constructs a long composition of little functions of the form $f\ a$, and this composition typically cannot be simplified until it is eventually applied to b. However, it is often possible that we can find some *representation* of those little functions that admits composition and application in constant time. Reynolds [20] calls this transformation *defunctionalization*.

Theorem 3 *Given fold arguments $f :: \alpha \to \beta \to \beta$ and $b :: \beta$, suppose there is a type ρ of representations of functions of the form $f\ a$ and their compositions, with the following operations:*

- *a representation function $\mathit{rep} :: \alpha \to \rho$ (so that $\mathit{rep}\ a$ is the representation of $f\ a$);*

- *an abstraction function* $abs :: \rho \to \beta \to \beta$*, such that* $abs\ (rep\ a) = f\ a$*;*
- *an analogue* $\odot :: \rho \to \rho \to \rho$ *of function composition, such that* $abs\,(r \odot s) = abs\ r \cdot abs\ s$*;*
- *an analogue* $ident :: \rho$ *of the identity function, such that* $abs\ ident = \mathsf{id}$*;*
- *an analogue* $app_b :: \rho \to \beta$ *of application to* b*, such that* $app_b\ r = abs\ r\ b$*.*

Then

$$\mathsf{foldr}\ f\ b = app_b \cdot \mathsf{foldl}\ (\odot)\ ident \cdot \mathsf{map}\ rep$$

The foldl *and the map can be fused:*

$$\mathsf{foldr}\ f\ b = app_b \cdot \mathsf{foldl}\ (\circledast)\ ident$$

where $r \circledast a \mathrel{\widehat{=}} r \odot rep\ a$*.*

(Note that the abstraction function *abs* is used above only for stating the correctness conditions; it is not applied anywhere.)

For example, let us return to radix conversion, as introduced in Section 3.2. Recall that we had

$$\begin{array}{ll} radixConvert\ (b, b') & \mathrel{\widehat{=}}\ toBase\ b' \cdot fromBase\ b \\ fromBase\ b & \mathrel{\widehat{=}}\ \mathsf{foldr}\ step_b\ 0 \\ toBase\ b & \mathrel{\widehat{=}}\ \mathsf{unfoldr}\ next_b \end{array}$$

where

$$\begin{array}{lll} step_b\ n\ x & \mathrel{\widehat{=}} & (x + n) \div b \\ next_b\ 0 & \mathrel{\widehat{=}} & \mathsf{Nothing} \\ next_b\ x & \mathrel{\widehat{=}} & \mathsf{Just}\ (\lfloor y \rfloor, y - \lfloor y \rfloor) \quad \textbf{where}\ y \mathrel{\widehat{=}} b \times x \end{array}$$

The 'little functions' here are of the form $step_b\ n$, or equivalently, $(\div b) \cdot (+n)$. This class of functions is closed under composition:

$$\begin{array}{ll} & (step_c\ n \cdot step_b\ m)\ x \\ = & \quad \{\text{composition}\} \\ & step_c\ n\ (step_b\ m\ x) \\ = & \quad \{step\} \\ & ((x + m) \div b + n) \div c \\ = & \quad \{\text{arithmetic}\} \\ & (x + m + b \times n) \div (b \times c) \\ = & \quad \{\text{composition}\} \\ & ((\div (b \times c)) \cdot (+m + b \times n))\ x \end{array}$$

We therefore defunctionalize $step_b\ n$ to the pair (n, b), and define:

$$\begin{array}{ll} rep_b\ n & \widehat{=}\ (n, b) \\ abs\ (n, b)\ x & \widehat{=}\ (x + n) \div b \\ (n, c) \odot (m, b) & \widehat{=}\ (m + b \times n, b \times c) \\ (n, c) \circledast_b m & \widehat{=}\ (n, c) \odot rep_b\ m\ \widehat{=}\ (m + b \times n, b \times c) \\ ident & \widehat{=}\ (0, 1) \\ app\ (n, b) & \widehat{=}\ abs\ (n, b)\ 0\ \widehat{=}\ n \div b \end{array}$$

Theorem 3 then tells us that

$$fromBase\ b = app \cdot \mathsf{foldl}\ (\circledast_b)\ ident$$

4.3 Checking the streaming condition

We cannot quite apply Theorem 2 yet, because the composition of $toBase\ b'$ and the revised $fromBase\ b$ has the abstraction function app between the unfold and the fold. Fortunately, that app fuses with the unfold; more generally, any projection, and indeed any surjection, fuses with an unfold.

Lemma 4 (Unfold fusion)

$$\mathsf{unfoldr}\ f \cdot g = \mathsf{unfoldr}\ f' \Leftarrow f \cdot g = mapl\ g \cdot f'$$

where mapl is the map operation for the base functor of the list datatype:

$$\begin{array}{ll} mapl\ f\ \mathsf{Nothing} & \widehat{=}\ \mathsf{Nothing} \\ mapl\ f\ (\mathsf{Just}\ (a, b)) & \widehat{=}\ \mathsf{Just}\ (a, f\ b) \end{array}$$

Corollary 5 *If* $g \cdot g' = \mathsf{id}$ *then* $\mathsf{unfoldr} f \cdot g = \mathsf{unfoldr} f'$ *where* $f' \widehat{=} mapl\, g' \cdot f \cdot g$.

In our case,

$$\begin{array}{ll} & \mathsf{unfoldr}\ next_c \cdot app = \mathsf{unfoldr}\ nextapp_c \\ \Leftarrow & \quad \{\text{unfold fusion}\} \\ & next_c \cdot app = mapl\ app \cdot nextapp_c \end{array}$$

and

$$\begin{array}{ll} & \mathit{next}_c\ (\mathit{app}\ (n, r)) \\ = & \quad \{\mathit{app},\ \mathit{next}_c;\ \text{let}\ u \mathrel{\widehat{=}} \lfloor n \times c \div r \rfloor\} \\ & \textbf{if}\ n{=}{=}0\ \textbf{then}\ \mathsf{Nothing}\ \textbf{else}\ \mathsf{Just}\ (u, n{\times}c{\div}r - u) \\ = & \quad \{\mathit{app};\ \text{there is some leeway here (see below)}\} \\ & \textbf{if}\ n{=}{=}0\ \textbf{then}\ \mathsf{Nothing}\ \textbf{else}\ \mathsf{Just}\ (u, \mathit{app}\ (n - u{\times}r{\div}c, r{\div}c)) \\ = & \quad \{\mathit{mapl}\} \\ & \mathit{mapl}\ \mathit{app}\ (\textbf{if}\ n{=}{=}0\ \textbf{then}\ \mathsf{Nothing}\ \textbf{else}\ \mathsf{Just}\ (u, (n - u{\times}r{\div}c, r{\div}c))) \end{array}$$

Therefore we try defining

$$\begin{array}{l} \mathit{nextapp}_c\ (n, r) \mathrel{\widehat{=}} \textbf{if}\ n{=}{=}0\ \textbf{then}\ \mathsf{Nothing}\ \textbf{else}\ \mathsf{Just}\ (u, (n - u{\times}r{\div}c, r{\div}c)) \\ \qquad\qquad\qquad\qquad \textbf{where}\ u \mathrel{\widehat{=}} \lfloor n \times c \div r \rfloor \end{array}$$

Note that there was some leeway here: we had to partition $n{\times}c{\div}r - u$ into a numerator and denominator, and we chose $(n - u{\times}r{\div}c, r{\div}c)$ out of the many ways of doing this. One might perhaps have expected $(n{\times}c - u{\times}r, r)$ instead; however, this leads to a dead-end, as we show later. Note also that our choice involves generalizing from integer to rational components.

Having now massaged our radix conversion program into the correct format:

$$\mathit{radixConvert}\ (b, c) = \mathsf{unfoldr}\ \mathit{nextapp}_c \cdot \mathsf{foldl}\ (\circledast_b)\ \mathit{ident}$$

we may consider whether the streaming condition holds for $\mathit{nextapp}_c$ and $\circledast_b$; that is, whether

$$\begin{array}{ll} & \mathit{nextapp}_c\ (n, r) = \mathsf{Just}\ (u, (n', r')) \\ \Rightarrow & \\ & \mathit{nextapp}_c\ ((n, r) \circledast_b m) = \mathsf{Just}\ (u, (n', r') \circledast_b m) \end{array}$$

An element u is produced from a state (n, r) if and only if $n \neq 0$, in which case $u = \lfloor n \times c \div r \rfloor$. The modified state $(n, r) \circledast_b m$ evaluates to $(m + b \times n, b \times r)$. Since $n, b > 0$ and $m \geq 0$, this necessarily yields an element; this element v equals $\lfloor (m + b \times n) \times c \div (b \times r) \rfloor$. We have to check that u and v are equal. Sadly, in general they are not: since $0 \leq m < b$, it follows that v lies between u and $\lfloor (n + 1) \times c \div r \rfloor$, but these two bounds need not meet.

Intuitively, this can happen when the state has not completely determined the next output, and further inputs are needed in order to make a commitment to that output. For example, consider having consumed the first digit 6_{10} while converting the sequence $[6, 7]$ in decimal (representing 0.67_{10}) to ternary. The fraction 0.6_{10} is about 0.1210_3; nevertheless, it is evidently not safe to commit

to producing the digit 1, because the true result is greater than 0.2_3, and there is not enough information to decide whether to output a 1_3 or a 2_3 until the 7_{10} has been consumed as well.

This is a common situation with streaming algorithms: the producer function (*nextapp* above) needs to be more cautious when interleaved with consumption steps than it does when all the input has been consumed. In the latter situation, there are no further inputs to invalidate a commitment made to an output; but in the former, a subsequent input might invalidate whatever output has been produced. The solution to this problem is to introduce a more sophisticated version of streaming, which proceeds more cautiously while input remains, but switches to the normal more aggressive mode if and when the input is exhausted. That is the subject of the next section.

4.4 Flushing streams

The typical approach is to introduce a 'restriction'

$$\mathit{snextapp} \mathrel{\widehat{=}} \mathsf{guard}\ \mathit{safe}\ \mathit{nextapp}$$

of *nextapp* for some predicate *safe*, where

$$\mathsf{guard}\ p\ f\ x \mathrel{\widehat{=}} \mathbf{if}\ p\ x\ \mathbf{then}\ f\ x\ \mathbf{else}\ \mathsf{Nothing}$$

and to use *snextapp* as the producer for the streaming process. In the case of radix conversion, the predicate safe_c (dependent on the output base c) could be defined

$$\mathit{safe}_c\ (n, r) \mathrel{\widehat{=}} (\lfloor n \times c \div r \rfloor = \lfloor (n+1) \times c \div r \rfloor)$$

That is, the state (n, r) is safe for the output base c if these lower and upper bounds on the next digit meet; with this proviso, the streaming condition holds, as we checked above. (In fact, we need to check not only that the same elements are produced from the unmodified and the modified state, but also that the two residual states are related in the same way as the two original states. With the definition of $\mathit{nextapp}_c$ that we chose above, this second condition does hold; with the more obvious definition involving $(n \times c - u \times r, r)$ that we rejected, it does not.)

However, with this restricted producer the streaming process no longer has the same behaviour on finite lists as does the plain metamorphism: when the input is exhausted, the more cautious *snextapp* may have left some outputs

still to be produced that the more aggressive *nextapp* would have emitted. The streaming process should therefore switch into a final 'flushing' phase when all the input has been consumed.

This insight is formalized in the following generalization of *stream*:

$$
\begin{array}{l}
\mathit{fstream} :: (\gamma{\to}\mathsf{Maybe}\ (\beta,\gamma)) \to (\gamma{\to}\alpha{\to}\gamma) \to (\gamma{\to}[\beta]) \to \gamma \to [\alpha] \to [\beta] \\
\mathit{fstream}\ f\ g\ h\ c\ x \mathrel{\widehat{=}} \mathbf{case}\ f\ c\ \mathbf{of} \\
\qquad\qquad\qquad \mathsf{Just}\ (b, c') \to b : \mathit{fstream}\ f\ g\ h\ c'\ x \\
\qquad\qquad\qquad \mathsf{Nothing} \quad \to \mathbf{case}\ x\ \mathbf{of} \\
\qquad\qquad\qquad\qquad\qquad\qquad a : x' \to \mathit{fstream}\ f\ g\ h\ (g\ c\ a)\ x' \\
\qquad\qquad\qquad\qquad\qquad\qquad [\,] \quad\ \to h\ c
\end{array}
$$

The difference between *fstream* and *stream* is that the former has an extra argument, h, a 'flusher'; when the state is wrung as dry as it can be and the input is exhausted, the flusher is applied to the state to squeeze the last few elements out. This is a generalization, because supplying the trivial flusher that always returns the empty list reduces *fstream* to *stream*.

In fact, not only is *stream* expressible in terms of *fstream*, but the converse holds too: *fstream* can be expressed in terms of *stream*, using a state of a sum type and an extra 'end of input' marker to indicate when to switch into the flushing phase. So *stream* and *fstream* are actually equally expressive. To be precise, in order to simulate an instance of *fstream*, the state for the *stream* consists of either an element of the original state type γ, or a residual output list of type $[\beta]$; the input is the original input with a sentinel appended. Then

$$\mathit{fstream}\ f\ g\ h\ c\ x = \mathit{stream}\ f'\ g'\ (\mathsf{Left}\ c)\ (\mathit{map}\ \mathsf{Just}\ x +\!\!+ [\mathsf{Nothing}])$$

where

$$
\begin{array}{lll}
f'\ (\mathsf{Left}\ c) & \mathrel{\widehat{=}} \mathbf{case}\ f\ c\ \mathbf{of} \\
 & \qquad \mathsf{Just}\ (b, c') \to \mathsf{Just}\ (b, \mathsf{Left}\ c') \\
 & \qquad \mathsf{Nothing} \quad \to \mathsf{Nothing} \\
f'\ (\mathsf{Right}\ y) & \mathrel{\widehat{=}} \mathbf{case}\ y\ \mathbf{of} \\
 & \qquad b : y' \to \mathsf{Just}\ (b, \mathsf{Right}\ y') \\
 & \qquad [\,] \quad\ \to \mathsf{Nothing} \\
g'\ (\mathsf{Left}\ c)\ (\mathsf{Just}\ a) & \mathrel{\widehat{=}} \mathsf{Left}\ (g\ c\ a) \\
g'\ (\mathsf{Left}\ c)\ \mathsf{Nothing} & \mathrel{\widehat{=}} \mathsf{Right}\ (h\ c)
\end{array}
$$

Note that the consumer initiates the switch into the flushing state, and is never subsequently called. Of course, if the original input is infinite, the sentinel is never reached.

The relationship of metamorphisms to flushing streams is a little more complicated than that to ordinary streams. One way of expressing the relationship is via a generalization of $\mathsf{unfoldr}$, whose final action is to generate a whole tail of the resulting list rather than the empty list. This is an instance of *primitive corecursion* (called an *apomorphism* by Vene and Uustalu [21]), which is the categorical dual of *primitive recursion* (called a *paramorphism* by Meertens [22]).

$$\begin{array}{l}\mathsf{apol} \quad :: (\beta \to \mathsf{Maybe}\,(\alpha, \beta)) \to (\beta \to [\alpha]) \to \beta \to [\alpha] \\ \mathsf{apol}\; f\; h\; b \mathrel{\widehat{=}} \mathbf{case}\; f\; b\; \mathbf{of} \\ \qquad\qquad \mathsf{Just}\,(a, b') \to a : \mathsf{apol}\; f\; h\; b' \\ \qquad\qquad \mathsf{Nothing} \quad\; \to h\; b\end{array}$$

Informally, $\mathsf{apol}\; f\; h\; b = \mathsf{unfoldr}\; f\; b +\!\!+ h\; b'$, where b' is the final state of the unfold (if there is one — and if there is not, the value of $h\; b'$ is irrelevant), and $\mathsf{unfoldr}\, f = \mathsf{apol}\, f\, (\mathsf{const}\, [\,])$. On finite inputs, provided that the streaming condition holds, a flushing stream process yields the same result as the ordinary streaming process, but with the results of flushing the final state (if any) appended.

Theorem 6 (Flushing Stream Theorem) *If the streaming condition holds for f and g, then*

$$\mathit{fstream}\; f\; g\; h\; c\; x = \mathsf{apol}\; f\; h\; (\mathsf{foldl}\; g\; c\; x)$$

on finite lists x.

The proof uses the following lemma [8], which lifts the streaming condition from single inputs to finite lists of inputs.

Lemma 7 *If the streaming condition holds for f and g, then*

$$f\; c = \mathsf{Just}\,(b, c') \Rightarrow f\; (\mathsf{foldl}\; g\; c\; x) = \mathsf{Just}\,(b, \mathsf{foldl}\; g\; c'\; x)$$

for all b, c, c' and finite lists x.

Proof. A simple proof by induction on x suffices. The lemma can also be proved using the following fusion law for foldl, which itself may be proved by induction: if $h\; (g\; u\; v) = g'\; (h\; u)\; v$ for all u and v, then

$$h\; (\mathsf{foldl}\; g\; c\; x) = \mathsf{foldl}\; g'\; (h\; c)\; x$$

for all c and finite lists x. (If in addition h is strict, then the equation holds for partial and infinite lists x too; but we do not need that stronger result.) However, the proof using fusion is more complicated. $\Box$

The proof of Theorem 6 also uses the *approximation lemma* [23,24].

Lemma 8 (Approximation Lemma) *For finite, infinite or partial lists x and y,*

$$x = y \;\equiv\; \forall n.\ approx\ n\ x = approx\ n\ y$$

where

$$\begin{array}{lll} approx & :: & Int \to [\alpha] \to [\alpha] \\ approx\ (n+1)\ [\,] & \widehat{=} & [\,] \\ approx\ (n+1)\ (a : x) & \widehat{=} & a : approx\ n\ x \end{array}$$

Note that $approx\ 0\ x = \bot$ for any x, by case exhaustion.

Proof of Theorem 6 By Lemma 8, it suffices to show, for fixed f, g, h and for all n and finite x, that

$$\forall c.\ approx\ n\ (fstream\ f\ g\ h\ c\ x) = approx\ n\ (\mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ c\ x))$$

under the assumption that the streaming condition holds for f and g. We use a 'double induction' simultaneously over n and the length $\#x$ of x (or equivalently, a standard induction using the pointwise ordering on pairs $(n, \#x)$). The inductive hypothesis is that

$$\forall c.\ approx\ m\ (fstream\ f\ g\ h\ c\ y) = approx\ m\ (\mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ c\ y))$$

for any m, y such that $m < n \wedge \#y \leq \#x$ or $m \leq n \wedge \#y < \#x$. We then proceed by case analysis to complete the inductive step.

Case $f\ c = \mathsf{Just}\ (b, d)$**.** In this case, we make a subsidiary case analysis on n.

Subcase $n = 0$**.** Then the result holds trivially.

Subcase $n = n' + 1$**.** Then we have:

$$
\begin{array}{ll}
 & \mathit{approx}\ (n'+1)\ (\mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ c\ x)) \\
= & \quad \{\text{Lemma 7: } f\ (\mathsf{foldl}\ g\ c\ x) = \mathsf{Just}\ (b, \mathsf{foldl}\ g\ d\ x)\} \\
 & \mathit{approx}\ (n'+1)\ (b : \mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ d\ x)) \\
= & \quad \{\mathit{approx}\} \\
 & b : \mathit{approx}\ n'\ (\mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ d\ x)) \\
= & \quad \{\text{induction: } n' < n\} \\
 & b : \mathit{approx}\ n'\ (\mathit{fstream}\ f\ g\ h\ d\ x) \\
= & \quad \{\mathit{approx}\} \\
 & \mathit{approx}\ (n'+1)\ (b : \mathit{fstream}\ f\ g\ h\ d\ x) \\
= & \quad \{\mathit{fstream}\text{; case assumption}\} \\
 & \mathit{approx}\ (n'+1)\ (\mathit{fstream}\ f\ g\ h\ c\ x)
\end{array}
$$

Case $f\ c = \mathsf{Nothing}$**.** In this case, we make a subsidiary case analysis on x.

Subcase $x = a : x'$**.** Then

$$
\begin{array}{ll}
 & \mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ c\ (a : x')) \\
= & \quad \{\mathsf{foldl}\} \\
 & \mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ (g\ c\ a)\ x')) \\
= & \quad \{\text{induction: } \#x' < \#x\} \\
 & \mathit{fstream}\ f\ g\ h\ (g\ c\ a)\ x' \\
= & \quad \{\mathit{fstream}\text{; case assumption}\} \\
 & \mathit{fstream}\ f\ g\ h\ c\ (a : x')
\end{array}
$$

Subcase $x = [\,]$**.** Then

$$
\begin{array}{ll}
 & \mathsf{apol}\ f\ h\ (\mathsf{foldl}\ g\ c\ [\,]) \\
= & \quad \{\mathsf{foldl}\} \\
 & \mathsf{apol}\ f\ h\ c \\
= & \quad \{\text{case assumption}\} \\
 & h\ c \\
= & \quad \{\mathit{fstream}\text{; case assumption}\} \\
 & \mathit{fstream}\ f\ g\ h\ c\ [\,]
\end{array}
$$

4.5 Invoking the Flushing Stream Theorem

Theorem 6 gives conditions under which an apomorphism applied to the result of a foldl may be streamed. This seems of limited use, since such scenarios are not commonly found. However, they can be constructed from more common scenarios in which the apomorphism is replaced with a simpler unfold. One way is to introduce the trivial apomorphism, whose flusher always returns the empty list. A more interesting, and the most typical, way is via the following observation.

Lemma 9 *For any predicate p,*

$$\mathsf{apol}\ (\mathsf{guard}\ p\ f)\ (\mathsf{unfoldr}\ f) = \mathsf{unfoldr}\ f$$

Informally, the work of an unfold can be partitioned into 'cautious' production, using the more restricted producer $\mathsf{guard}\ p\ f$, followed by more 'aggressive' production using simply f when the more cautious producer blocks.

4.6 Radix conversion as a flushing stream

Returning for a final time to radix conversion, we see that

$$\begin{array}{ll} & \mathit{radixConvert}\ (b, c) \\ = & \quad \{\text{as derived in Section 4.3}\} \\ & \mathsf{unfoldr}\ \mathit{nextapp}_c \cdot \mathsf{foldl}\ (\circledast_b)\ \mathit{ident} \\ = & \quad \{\text{Lemma 9}\} \\ & \mathsf{apol}\ (\mathsf{guard}\ \mathit{safe}_c\ \mathit{nextapp}_c)\ (\mathsf{unfoldr}\ \mathit{nextapp}_c) \cdot \mathsf{foldl}\ (\circledast_b)\ \mathit{ident} \end{array}$$

For brevity, we define

$$\mathit{snextapp}_c\ (n, r) \mathrel{\widehat{=}} \mathsf{guard}\ \mathit{safe}_c\ \mathit{nextapp}_c$$

We verified in Sections 4.3 and 4.4 that the streaming condition holds for $\mathit{snextapp}_c$ and $\circledast_b$. Theorem 6 then tells us that we can convert from base b to base c using

$$\mathit{radixConvert}\ (b, c) = \mathit{fstream}\ \mathit{snextapp}_c\ (\circledast_b)\ (\mathsf{unfoldr}\ \mathit{nextapp}_c)\ (0, 1)$$

This program works for finite or infinite inputs, and is always productive. It does, however, always produce an infinite result, even when a finite result would be correct. For example, it will correctly convert $\frac{1}{3}$ from base 10 to base 2, but in converting from base 10 to base 3 it will produce an infinite tail of zeroes.

5 Continued fractions

Continued fractions are finite or infinite constructions of the form

$$b_0 + \cfrac{a_0}{b_1 + \cfrac{a_1}{b_2 + \cfrac{a_2}{b_3 + \cdots}}}$$

In our treatment, all the coefficients will be integers. Continued fractions provide an elegant representation of numbers, both rational and irrational. They have therefore been proposed by various authors [25–30] as a good format in which to carry out *exact real arithmetic*. Some of the algorithms for simple arithmetic operations on continued fractions can be seen as metamorphisms, and as we shall show here, they can typically be streamed.

We consider algorithms on *regular* continued fractions: ones in which all the numerators a_i are 1, and all the b_i coefficients (except perhaps b_0) are at least 1. We denote regular continued fractions more concisely in the form $\langle b_0, b_1, b_2, \ldots\rangle$. For example, the regular continued fraction for π starts with the coefficients $\langle 3, 7, 15, 1, 292, \ldots\rangle$. Peyton Jones [27] paraphrased this:

> about 3; but not really 3, more like $3 + 1/7$; but not really 7, more like $7 + 1/15$; but not really 15, more like $15 + 1/1 (= 16)$; but not really 1, more like $1 + 1/292$; ...

Finite continued fractions correspond to rationals; for example, $\langle 1, 2, 3, 4\rangle$ represents

$$1 + \cfrac{1}{2 + \cfrac{1}{3 + 1/4}} = \frac{43}{30}$$

The number of terms in the continued fraction representation of a rational a/b is related to the number of steps taken by Euclid's algorithm to compute the greatest common divisor of a and b.

Infinite continued fractions represent irrational numbers; but many interesting irrationals have periodic continued fraction representations, or at least ones exhibiting clear patterns:

$$\begin{aligned}
\sqrt{2} &= \langle 1, 2, 2, 2, 2 \ldots\rangle \\
(1+\sqrt{5})/2 &= \langle 1, 1, 1, 1, 1 \ldots\rangle \\
e &= \langle 2, 1, 2, 1, 1, 4, 1, 1, 6, 1, \ldots\rangle
\end{aligned}$$

The finite or infinite regular continued fractions that do not end with a 1 are in one-to-one correspondence with the real numbers — in contrast to digit sequences (which are redundant: $0.99\dot{9} = 1.0$) and rationals (which are both redundant: $1/2 = 2/4$ and incomplete: $\sqrt{2} =?$).

Given an infinite regular continued fraction, its finite prefixes are called its *convergents*; in a precise technical sense, these form the closest rational approximations to the real number represented by the entire continued fraction. Huygens used these convergents to find accurate gear ratios for a mechanical planetarium, and they explain the design decisions behind some calendars and musical scales. The first few convergents to π are

$$3, 22/7, 333/106, 355/113, \ldots$$

The relatively large term 292 coming next means that the previous convergent $\langle 3, 7, 15, 1\rangle = 355/113$ is a very close rational approximation to π; the simplest closer approximation is $\langle 3, 7, 15, 1, 292\rangle = 103993/33102$.

5.1 Converting continued fractions

We consider first conversions between rationals and finite regular continued fractions. To complete the isomorphism between these two sets, we need to augment the rationals with $1/0 = \infty$, corresponding to the empty continued fraction. We therefore introduce a type *ExtRat* of rationals extended with ∞.

Conversion from rationals to continued fractions is straightforward. Infinity, by definition, is represented by the empty fraction. A finite rational a/b has a first term given by $a \mathbin{\underline{div}} b$, the integer obtained by rounding the fraction down; this leaves a remainder of $(a \mathbin{\underline{mod}} b)/b$, whose reciprocal is the rational from which to generate the rest of the continued fraction. If the reciprocal is infinite, the remainder is zero, and the conversion is complete. Note that as a consequence of rounding the fraction down to get the first term, the remainder is between zero and one, and its reciprocal is at least one; therefore the next term (and by induction, all subsequent terms) will be at least one, yielding a regular continued fraction as claimed.

$$\begin{array}{l}
\textbf{type}\ \mathit{CF} \mathrel{\widehat{=}} [\mathit{Integer}] \\
\mathit{toCF} \quad :: \mathit{ExtRat} \to \mathit{CF} \\
\mathit{toCF} \quad \mathrel{\widehat{=}} \mathsf{unfoldr}\ \mathit{get} \\
\qquad\qquad \textbf{where}\ \mathit{get}\ x \mathrel{\widehat{=}} \textbf{if} \quad x{=}{=}\infty \\
\qquad\qquad\qquad\qquad\qquad\quad \textbf{then}\ \mathsf{Nothing} \\
\qquad\qquad\qquad\qquad\qquad\quad \textbf{else}\ \ \mathsf{Just}\ (\lfloor x \rfloor, 1/(x - \lfloor x \rfloor))
\end{array}$$

Converting in the opposite direction is more difficult: of course, we have seen that not all continued fractions correspond to rationals. However, finite ones do, and for these we can compute the rational using a fold — it suffices to fold with the inverse of *get* (or at least, what would have been the inverse of *get*, if foldr had been defined to take an argument of type $\mathsf{Maybe}\,(\alpha, \beta) \to \beta$, dualizing unfoldr).

$$\begin{aligned}&\mathit{fromCF} :: \mathit{CF} \to \mathit{ExtRat}\\ &\mathit{fromCF} \mathrel{\widehat{=}} \mathsf{foldr}\ \mathit{put}\ \infty \quad \textbf{where}\ \mathit{put}\ n\ y \mathrel{\widehat{=}} n + {}^{1}\!/_{y}\end{aligned}$$

Thus, $\mathit{fromCF} \cdot \mathit{toCF}$ is the identity on extended rationals, and $\mathit{toCF} \cdot \mathit{fromCF}$ is the identity on finite continued fractions. On infinite continued fractions, *fromCF* yields no result: *put* is strict, so the whole list of coefficients is required. One could compute an infinite sequence of rational approximations to the irrational value represented by an infinite continued fraction, by converting to a rational each of the convergents. But this is awkward, because the fold starts at the right, and successive approximations will have no common subexpressions — the process does not constitute a scan. It would be preferable if we could write *fromCF* as an instance of foldl; then the sequence of approximations would be computed with the corresponding scanl.

Fortunately, Theorem 3 comes to the rescue again. This requires defunctionalizations of functions of the form $\mathit{put}\ n$ and their compositions. For proper rationals, we reason:

$$\begin{aligned}&\mathit{put}\ n\ (\mathit{put}\ m\ {}^{a}\!/_{b})\\ =\quad &\{\mathit{put}\}\\ &\mathit{put}\ n\ (m + {}^{b}\!/_{a})\\ =\quad &\{\text{arithmetic}\}\\ &\mathit{put}\ n\ ({}^{m\times a+b}\!/_{a})\\ =\quad &\{\mathit{put}\}\\ &n + {}^{a}\!/_{m\times a+b}\\ =\quad &\{\text{arithmetic}\}\\ &(n \times (m \times a + b) + a)/(m \times a + b)\\ =\quad &\{\text{collecting terms; dividing through by } b\}\\ &((n \times m + 1) \times {}^{a}\!/_{b} + n)/(m \times {}^{a}\!/_{b} + 1)\end{aligned}$$

This is a ratio of integer-coefficient linear functions of ${}^{a}\!/_{b}$, sometimes known as a *rational function* or *linear fractional transformation* of ${}^{a}\!/_{b}$. The general form of such a function takes x to $(q\,x + r)/(s\,x + t)$ (denoting multiplication by juxtaposition for brevity), and can be represented by the four integers q, r, s, t.

For the improper rational ∞, we reason:

$$
\begin{array}{ll}
 & put\ n\ (put\ m\ \infty) \\
= & \quad \{put\} \\
 & put\ n\ (m + {}^{1}\!/_{\infty}) \\
= & \quad \{{}^{1}\!/_{\infty} = 0\} \\
 & put\ n\ m \\
= & \quad \{put\} \\
 & n + {}^{1}\!/_{m} \\
= & \quad \{\text{arithmetic}\} \\
 & (n \times m + 1)/m
\end{array}
$$

which agrees with the result for proper rationals, provided we take the reasonable interpretation that $(q \times \infty + r)/(s \times \infty + t) = {}^{q}\!/_{s}$.

Following Theorem 3, then, we choose four-tuples of integers as our representation; for reasons that will become clear, we write these four-tuples in the form $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$. The abstraction function *abs* applies the rational function:

$$abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ x \mathrel{\widehat{=}} {}^{q\,x+r}\!/_{s\,x+t}$$

which simplifies to ${}^{q}\!/_{s}$ when $x = \infty$, and to ∞ when $s\,x + t = 0$ (or when $x = \infty$ and $s = 0$). The representation function *rep* injects the integer n into the representation of *put* n:

$$rep\ n \mathrel{\widehat{=}} \left(\begin{smallmatrix} n & 1 \\ 1 & 0 \end{smallmatrix}\right)$$

The identity function is represented by *ident*:

$$ident \mathrel{\widehat{=}} \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$$

We verify that rational functions are indeed closed under composition, by

constructing the representation of function composition:

$$
\begin{array}{ll}
 & abs\ (\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \odot \left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right))\ x \\
= & \quad \{\text{requirement}\} \\
 & abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ (abs\ \left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right)\ x) \\
= & \quad \{abs\} \\
 & abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ ((q'\,x + r')/(s'\,x + t')) \\
= & \quad \{abs \text{ again}\} \\
 & (q\,(q'\,x{+}r') + r\,(s'\,x{+}t'))/(s\,(q'\,x{+}r') + t\,(s'\,x{+}t')) \\
= & \quad \{\text{collecting terms}\} \\
 & ((q\,q'{+}r\,s')\,x + (q\,r'{+}r\,t'))/((s\,q'{+}t\,s')\,x + (s\,r'{+}t\,t')) \\
= & \quad \{abs\} \\
 & abs\ \left(\begin{smallmatrix} q\,q'{+}r\,s' & q\,r'{+}r\,t' \\ s\,q'{+}t\,s' & s\,r'{+}t\,t' \end{smallmatrix}\right)\ x
\end{array}
$$

We therefore define

$$\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \odot \left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right) \mathrel{\widehat{=}} \left(\begin{smallmatrix} q\,q'{+}r\,s' & q\,r'{+}r\,t' \\ s\,q'{+}t\,s' & s\,r'{+}t\,t' \end{smallmatrix}\right)$$

Finally, we define an extraction function

$$
\begin{array}{l}
app\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \mathrel{\widehat{=}} abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ \infty \\
\qquad\qquad = q/_s
\end{array}
$$

(Notice that $\odot$ turns out to be matrix multiplication, and *ident* the unit matrix, which explains the choice of notation. These matrices are sometimes called *homographies*, and the rational functions they represent *homographic functions* or *Möbius transformations*. We discuss this connection further in Section 6.3.)

By Theorem 3 we then have

$$fromCF = app \cdot \mathsf{foldl}\ (\circledast)\ ident \quad \textbf{where}\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \circledast n \mathrel{\widehat{=}} \left(\begin{smallmatrix} n\,q{+}r & q \\ n\,s{+}t & s \end{smallmatrix}\right)$$

Of course, this still will not work for infinite continued fractions; however, we can now define

$$
\begin{array}{l}
fromCFi\ ::\ CF \to [ExtRat] \\
fromCFi \mathrel{\widehat{=}} \mathsf{map}\ app \cdot \mathsf{scanl}\ (\circledast)\ ident
\end{array}
$$

yielding the (infinite) sequence of finite convergents of an (infinite) continued fraction.

In Section 5.1, we derived the program

$$fromCF = app \cdot \mathsf{foldl}\ (\circledast)\ \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$$

for converting a finite continued fraction to an extended rational. In fact, we can compute an arbitrary rational function of a continued fraction, by starting this process with an arbitrary homography in place of the identity $\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$. This is because composition $\odot$ fuses with the fold:

$$\begin{array}{ll} & abs\ h\ (fromCF\ ns)\\ = & \quad \{fromCF\}\\ & abs\ h\ (app\ (\mathsf{foldl}\ (\circledast)\ ident\ ns))\\ = & \quad \{\text{specification of } app\}\\ & abs\ h\ (abs\ (\mathsf{foldl}\ (\circledast)\ ident\ ns)\ \infty)\\ = & \quad \{\text{requirement on } abs \text{ and } \odot\}\\ & abs\ (h \odot \mathsf{foldl}\ (\circledast)\ ident\ ns)\ \infty\\ = & \quad \{\text{fold fusion: } \odot \text{ is associative, and } ident \text{ its unit}\}\\ & abs\ (\mathsf{foldl}\ (\circledast)\ h\ ns)\ \infty\\ = & \quad \{\text{specification of } app \text{ again}\}\\ & app\ (\mathsf{foldl}\ (\circledast)\ h\ ns)\end{array}$$

For example, suppose we want to compute the rational $2/x-3$, where x is the rational represented by a particular (finite) continued fraction ns. We could convert ns to the rational x, then perform the appropriate rational arithmetic. Alternatively, we could convert ns to a rational as above, starting with the homography $\left(\begin{smallmatrix}0 & 2\\ 1 & -3\end{smallmatrix}\right)$ instead of $\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$, and get the answer directly. If we want the result as a continued fraction again rather than a rational, we simply post-apply $toCF$.

Of course, this will not work to compute rational functions of *infinite* continued fractions, as the folding will never yield a result. Fortunately, it is possible to applying streaming, so that terms of the output are produced before the whole input is consumed. This is the focus of the remainder of this section. The derivation follows essentially the same steps as were involved in radix conversion.

The streaming process maintains a state in the form of a homography, which represents the mapping from what is yet to be consumed to what is yet to be produced. The production steps of the streaming process choose a term to output, and compute a reduced homography for the remainder of the computation. Given a current homography $\left(\begin{smallmatrix}q & r\\ s & t\end{smallmatrix}\right)$, and a chosen term n, the reduced

homography $\left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right)$ is determined as follows:

$$
\begin{array}{ll}
 & (q\,x + r)/(s\,x + t) = n + 1/((q'\,x + r')/(s'\,x + t')) \\
\equiv & \quad \{\text{reciprocal}\} \\
 & (q\,x + r)/(s\,x + t) = n + (s'\,x + t')/(q'\,x + r') \\
\equiv & \quad \{\text{rearrange}\} \\
 & (s'\,x + t')/(q'\,x + r') = (q\,x + r)/(s\,x + t) - n \\
\equiv & \quad \{\text{incorporate } n \text{ into fraction}\} \\
 & (s'\,x + t')/(q'\,x + r') = (q\,x + r - n\,(s\,x + t))/(s\,x + t) \\
\equiv & \quad \{\text{collect } x \text{ and non-}x \text{ terms}\} \\
 & (s'\,x + t')/(q'\,x + r') = ((q - n\,s)\,x + r - n\,t)/(s\,x + t) \\
\Leftarrow & \quad \{\text{equating terms}\} \\
 & q' = s, r' = t, s' = q - n\,s, t' = r - n\,t
\end{array}
$$

That is,

$$\left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & -n \end{smallmatrix}\right) \odot \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$$

We therefore define

$$
\begin{aligned}
emit\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ n &\mathrel{\widehat{=}} \left(\begin{smallmatrix} 0 & 1 \\ 1 & -n \end{smallmatrix}\right) \odot \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \\
&= \left(\begin{smallmatrix} s & t \\ q - n\,s & r - n\,t \end{smallmatrix}\right)
\end{aligned}
$$

5.2.1 Making it a metamorphism

In most of what follows, we assume that we have a *completely regular* continued fraction, namely one in which every coefficient including the first is at least one. This implies that the value represented by the continued fraction is between one and infinity. Section 5.2.5 shows how to handle the first coefficient, in case it is less than one.

Given the representation of a rational function in the form of a homography h, we introduce the function *rfc* ('*r*ational *f*unction of a *c*ompletely regular continued fraction') to apply it as follows:

$$rfc\ h \mathrel{\widehat{=}} toCF \cdot app \cdot \mathsf{foldl}\ (\circledast)\ h$$

This is almost a metamorphism: *toCF* is indeed an unfold, but we must get rid of the projection function *app* in the middle. Unfold fusion (Corollary 5) applies, and we have

$$\mathsf{unfoldr}\ get \cdot app = \mathsf{unfoldr}\ geth$$

where *geth* (for '*get* on homographies') is defined by

$$geth\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \mathrel{\widehat{=}} \textbf{if}\ s{=}{=}0\ \textbf{then}\ \mathsf{Nothing}\ \textbf{else}\ \mathsf{Just}\ (n, emit\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ n) \\ \textbf{where}\ n \mathrel{\widehat{=}} q\ \underline{div}\ s$$

as can easily be verified.

This yields a metamorphism:

$$rfc\ h = \mathsf{unfoldr}\ geth \cdot \mathsf{foldl}\ (\circledast)\ h$$

5.2.2 Checking the streaming condition

Now we must check that the streaming condition holds for *geth* and $\circledast$. We require that when

$$geth\ h = \mathsf{Just}\ (n, h')$$

then, for any subsequent term m (which we can assume to be at least 1, this being a completely regular continued fraction),

$$geth\ (h \circledast m) = \mathsf{Just}\ (n, h' \circledast m)$$

Unpacking this, when $h = \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$ and $h' = \left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right)$, we have $s \neq 0$, $n = q\ \underline{div}\ s$, $q' = s$, and $s' = q\ \underline{mod}\ s$; moreover, $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \circledast m = \left(\begin{smallmatrix} m\,q+r & q \\ m\,s+t & s \end{smallmatrix}\right)$. We require among other things that $m\,s + t \neq 0$ and $(m\,q + r)\ \underline{div}\ (m\,s + t) = q\ \underline{div}\ s$. Sadly, this does not hold; for example, if $m = 1$ and s, t are positive,

$$\begin{array}{ll} & {}^{q+r}\!/_{s+t} < 1 + {}^{q}\!/_{s} \\ \equiv & \quad \{\text{multiply through}\} \\ & s\,(q + r) < (q + s)\,(q + t) \\ \equiv & \quad \{\text{algebra}\} \\ & r\,s < q\,t + s\,t + s^2 \end{array}$$

which fails if r is sufficiently large.

5.2.3 Cautious progress

As with the radix conversion algorithm in Section 4.3, the function that produces the next term of the output must be more cautious when it is interleaved with consumption steps that it may be after all the input has been consumed. The above discussion suggests that we should commit to an output only when it is safe from being invalidated by a later input; in symbols,

only when $(m\,q + r)\ \underline{div}\ (m\,s + t) = q\ \underline{div}\ s$ for any $m \geq 1$. This follows if s and t are non-zero and have the same sign, and if $(q + r)\ \underline{div}\ (s + t) = q\ \underline{div}\ s$, as a little calculation will verify.

(Another way of looking at this is to observe that the value represented by a completely regular continued fraction ranges between 1 and ∞, so the result of transforming it under a homography $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$ ranges between

$$abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ 1 = {}^{q+r}\!/_{s+t}$$

and

$$abs\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)\ \infty = {}^{q}\!/_{s}$$

if $s \neq 0$. If the two denominators have the same sign, the result ranges between these two; if they have different signs, it ranges outside them. Therefore, the first coefficient of the output is determined if the denominators have the same sign (which follows if s and t are non-zero and of the same sign) and the two fractions have the same integer parts.)

We therefore define *gets* (for 'safe *get*') to be $\mathsf{guard}\ same\ geth$, where guard is as defined in Section 4.4 and

$$same\ \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \mathrel{\widehat{=}} s\,t > 0 \wedge (q{+}r)\ \underline{div}\ (s{+}t) \mathrel{=\!=} q\ \underline{div}\ s$$

Note that whenever *gets* produces a value, *geth* produces the same value; but sometimes *gets* produces nothing when *geth* produces something. The streaming condition *does* hold for *gets* and $\circledast$, as the reader may now verify.

5.2.4 Flushing streams

It is not the case that $\mathsf{unfoldr}\ get \cdot app = \mathsf{unfoldr}\ gets$, of course, because the latter is too cautious. However, by Lemma 9 it does follow that

$$\mathsf{unfoldr}\ get \cdot app = apol\ gets\ (\mathsf{unfoldr}\ geth)$$

This cautiously produces elements while it it safe to do so, then throws caution to the winds and produces elements anyway when it ceases to be safe. Moreover, Theorem 6 applies to the cautious part, and so

$$\begin{aligned} rfc\ h &= \mathsf{unfoldr}\ get \cdot app \cdot \mathsf{foldl}\ (\circledast)\ h \\ &= fstream\ gets\ (\circledast)\ (\mathsf{unfoldr}\ geth)\ h \end{aligned}$$

This streaming algorithm can compute a rational function of a finite or infinite completely regular continued fraction, yielding a finite or infinite regular continued fraction as a result.

5.2.5 Handling the first term

A regular but not completely regular continued fraction may have a first term of 1 or less, invalidating the reasoning above. However, this is easy to handle, simply by consuming the first term immediately. We introduce a wrapper function rf:

$$\begin{aligned} rf\ h\ [\,] &\mathrel{\widehat{=}} rfc\ h\ [\,] \\ rf\ h\ (n : ns) &\mathrel{\widehat{=}} rfc\ (h \circledast n)\ ns \end{aligned}$$

This streaming algorithm can compute any rational function of a finite or infinite regular continued fraction, completely regular or otherwise.

5.3 Rational binary functions of continued fractions

The streaming process described in Section 5.2 allows us to compute a unary rational function $(a\,x + b)/(c\,x + d)$ of a single continued fraction x. The technique can be adapted to allow a binary rational function $(a\,x\,y + b\,x + c\,y + d)/(e\,x\,y + f\,x + g\,y + h)$ of continued fractions x and y. This does not fit into our framework of metamorphisms and streaming algorithms, because it combines two arguments into one result; nevertheless, much of the same reasoning can be applied. We simply outline the development here; we intend to elaborate on this in a companion paper.

5.3.1 Familiar operations as rational binary functions

Let us write the binary rational function $(a\,x\,y + b\,x + c\,y + d)/(e\,x\,y + f\,x + g\,y + h)$ of x and y as a *bihomography* or *tensor* applied to (x, y):

$$\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(x, y)$$

Suitable choices of the coefficients yield addition, subtraction, multiplication and division:

$$x + y = \left(\begin{smallmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)(x, y)$$
$$x - y = \left(\begin{smallmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)(x, y)$$
$$x \times y = \left(\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)(x, y)$$
$$x \div y = \left(\begin{smallmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{smallmatrix}\right)(x, y)$$

5.3.2 Consuming a term

Recall that for rational unary functions, the transformation on the coefficients needed when consuming a single term n of the argument is given by $\circledast$:

$$\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \circledast n = \left(\begin{smallmatrix} n\,q+r & q \\ n\,s+t & t \end{smallmatrix}\right)$$

For rational binary functions, the transformation on the coefficients needed when consuming a single term n of the first argument x is as follows:

$$\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right) \oslash n \mathrel{\widehat{=}} \left(\begin{smallmatrix} a\,n+b & a & c\,n+d & c \\ e\,n+f & e & g\,n+h & g \end{smallmatrix}\right)$$

Similarly, for consuming the first term m of the second argument y:

$$\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right) \oslash n \mathrel{\widehat{=}} \left(\begin{smallmatrix} a\,n+c & b\,n+d & a & b \\ e\,n+g & f\,n+h & e & f \end{smallmatrix}\right)$$

That is,

$$\left(\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right) \oslash n\right)(x, y) = \left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(n : x, y)$$

and

$$\left(\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right) \oslash n\right)(x, y) = \left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(x, n : y)$$

5.3.3 Producing a term

For rational unary functions, the transformation on the coefficients needed when producing a single term n of the result is given by the function *emit*:

$$\mathit{emit}\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) n = \left(\begin{smallmatrix} s & t \\ q-n\,s & r-n\,t \end{smallmatrix}\right)$$

The transformation on the coefficients needed when producing a single term n of the result is as follows:

$$\mathit{emit}\ \left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)\ n \mathrel{\widehat{=}} \left(\begin{smallmatrix} e & f & g & h \\ a-n\,e & b-n\,f & c-n\,g & d-n\,h \end{smallmatrix}\right)$$

That is,

$$\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(x,y) = n + \frac{1}{\left(\begin{smallmatrix} e & f & g & h \\ a-n\,e & b-n\,f & c-n\,g & d-n\,h \end{smallmatrix}\right)(x,y)}$$

5.3.4 When to commit?

For rational unary functions, the homography $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$ determines the next coefficient of the result when

$$s\,t > 0 \wedge (q+r)\ \underline{\mathit{div}}\ (s+t) = q\ \underline{\mathit{div}}\ s$$

Similarly, the extreme values of

$$\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(x,y)$$

can be determined from its values at the four corners consisting of x, y chosen from $1, \infty$:

$$\begin{aligned}
\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(1,1) &= {}^{a+b+c+d}\!/_{e+f+g+h} \\
\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(1,\infty) &= {}^{a+c}\!/_{e+g} \\
\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(\infty,1) &= {}^{a+b}\!/_{e+f} \\
\left(\begin{smallmatrix} a & b & c & d \\ e & f & g & h \end{smallmatrix}\right)(\infty,\infty) &= {}^{a}\!/_{e}
\end{aligned}$$

provided that the denominators are all non-zero. If the denominators all have the same sign, the result ranges between these four corner values; if in addition all four corner values have the same integer part, then this is the next coefficient of the result.

5.3.5 Which term to consume?

When it is not safe to commit to an output, an input term should be consumed; but for binary functions, a choice has to be made from which of the two inputs to consume a term. Various strategies have been proposed for making this choice. No choice is wrong, but some will be more effective than others. Vuillemin [28] consumes an element from both inputs together

(it will be seen that the two consumption operators commute, in the sense that $(h \oslash n) \oslash m = (h \oslash m) \oslash n$); equivalently, one could alternate between consumption from the left and consumption from the right. But these are likely to be suboptimal strategies, in that they consume more input terms than are necessary for producing a given number of output terms. Potts [31] describes some more elaborate strategies, aiming for a good compromise between effectiveness (consuming the terms that lead most quickly to outputs) and efficiency (making the choice quickly).

5.3.6 Productivity

Unfortunately, there is a problem of *productivity* with binary operations on continued fractions, that does not arise with unary operations: some operations are simply uncomputable. For example, computing $x - x$ for an irrational x ought to yield zero, but this cannot be determined without examining all of the infinite input. This problem has led Vuillemin [28] and others to consider *redundant* representations of continued fractions, so that such a computation can still be productive. It is necessarily the case then that ordinary exact two-way comparisons become uncomputable; however, 'loose' three-way comparisons (such as a test for zero, with a third case for when the argument is 'close to zero') can be implemented.

6 Other applications of streaming

In Sections 6.1 and 6.2, we briefly outline two other applications of streaming; we have described both in more detail elsewhere, and we refer the reader to those sources for the full stories. Section 6.3 outlines a common unification of a number of applications of streaming described in this article and elsewhere.

6.1 Digits of π

In [32] we present an *unbounded spigot algorithm* for computing the digits of π. This work was inspired by Rabinowitz and Wagon [33], who coined the term *spigot algorithm* for an algorithm that yields output elements incrementally and does not reuse them after they have been output — so the digits drip out one by one, as if from a leaky tap. (In contrast, most algorithms for computing approximations to π, including the best currently known, work inscrutably until they deliver a complete response at the end of the computation.) Although incremental, Rabinowitz and Wagon's algorithm is *bounded*, since one needs to decide at the outset how many digits are to be computed,

whereas our algorithm yields digits indefinitely. This observation is nothing to do with computational evaluation order: Rabinowitz and Wagon's algorithm is just as bounded in a lazy language.

The algorithm is based on the following infinite sum:

$$\pi = \sum_{i=0}^{\infty} \frac{(i!)^2 2^{i+1}}{(2i+1)!}$$

which expands out to

$$\pi = 2 + \frac{1}{3}\Big(2 + \frac{2}{5}\Big(2 + \frac{3}{7}\Big(\cdots\Big(2 + \frac{i}{2i+1}\Big(\cdots\Big)\Big)\Big)\Big)\Big)$$

One can view this expansion as an infinite composition of linear fractional transformations:

$$\pi = \Big(2 + \frac{1}{3}\times\Big)\Big(2 + \frac{2}{5}\times\Big)\Big(2 + \frac{3}{7}\times\Big)\cdots\Big(2 + \frac{i}{2i+1}\times\Big)\cdots$$

A streaming algorithm can convert this infinite sequence of linear fractional transformations (represented as homographies) into an infinite sequence of decimal digits. The consumption operator is matrix multiplication, written $\odot$ in Section 5.1. When a digit n is produced, the state $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$ should be transformed into a homography (q', r', s', t') such that

$$abs\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) x = n + abs\left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right) x \div 10$$

which simplifies to

$$\left(\begin{smallmatrix} q' & r' \\ s' & t' \end{smallmatrix}\right) = \left(\begin{smallmatrix} 10 & -10n \\ 0 & 1 \end{smallmatrix}\right) \odot \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$$

Any tail of the input sequence represents a value between 3 and 4, because

$$
\begin{array}{ll}
 & 3 \\
= & \quad \{\text{solution of recurrence}\} \\
 & x \quad \text{where} \quad x \mathrel{\widehat{=}} 2 + \frac{1}{3}x \\
= & \quad \{\text{expand recurrence}\} \\
 & 2 + \frac{1}{3}\Big(2 + \frac{1}{3}\Big(2 + \cdots\Big)\Big) \\
< & \quad \{\text{bound terms}\} \\
 & 2 + \frac{i}{2i+1}\Big(2 + \frac{i+1}{2i+3}\Big(2 + \cdots\Big)\Big) \\
< & \quad \{\text{bound terms}\} \\
 & 2 + \frac{1}{2}\Big(2 + \frac{1}{2}\Big(2 + \cdots\Big)\Big) \\
= & \quad \{\text{contract recurrence}\} \\
 & x \quad \text{where} \quad x \mathrel{\widehat{=}} 2 + \frac{1}{2}x \\
= & \quad \{\text{solution of recurrence}\} \\
 & 4
\end{array}
$$

Therefore, homography h determines the next digit when

$$\lfloor abs\ h\ 3 \rfloor \mathrel{\widehat{=}} \lfloor abs\ h\ 4 \rfloor$$

in which case, the digit is the common value of these two expressions. This reasoning gives us the following program:

$$pi \mathrel{\widehat{=}} stream\ prod\ (\odot)\ ident\ lfts$$

where

$$
\begin{array}{l}
lfts \quad \mathrel{\widehat{=}} [(\begin{smallmatrix} k & 4k+2 \\ 0 & 2k+1 \end{smallmatrix}) \mid k \leftarrow [1..]] \\
prod\ h \mathrel{\widehat{=}} \textbf{if } \lfloor abs\ h\ 4 \rfloor \mathrel{=\!=} n \textbf{ then } \mathsf{Just}\ (n, (\begin{smallmatrix} 10 & -10n \\ 0 & 1 \end{smallmatrix}) \odot h) \textbf{ else } \mathsf{Nothing} \\
\qquad\qquad \textbf{where } n \mathrel{\widehat{=}} \lfloor abs\ h\ 3 \rfloor
\end{array}
$$

(The definition of *lfts* is via a 'list comprehension', yielding the term before the bar for each value of k in $[1..]$.)

6.2 Arithmetic coding

Arithmetic coding [34] is a method for data compression. It can be more *effective* (compressing better) than rival schemes such as Huffman coding, while

still being as *efficient* (using little time and space). Moreover, it is well suited to *adaptive* encoding, in which the coding scheme evolves to match the text being encoded.

The basic idea of arithmetic encoding is simple. The message to be encoded is broken into *symbols*, such as characters, and each symbol of the message is associated with a semi-open *subinterval* of the unit interval $[0\,..\,1)$. Encoding starts with the unit interval, and *narrows* it according to the intervals associated with each symbol of the message in turn. The encoded message is the binary representation of some *fraction* chosen from the final 'target' interval.

In [8] we present a detailed derivation of arithmetic encoding and decoding. We merely outline the development of encoding here, to show where streaming fits in. Decoding follows a similar process to encoding, starting with the unit interval and homing in on the binary fraction, reconstructing the plaintext in the process; but we will not discuss it here.

The encoding process can be captured as follows. The type *Interval* represents intervals of the real line, usually *subunits* (subintervals of the unit interval):

$$\begin{aligned} &\mathit{unit} :: \mathit{Interval} \\ &\mathit{unit} \mathrel{\widehat{=}} [0\,..\,1) \end{aligned}$$

Narrowing an interval lr by a subunit pq yields a subinterval of lr, which stands in the same relation to lr as pq does to *unit*.

$$\begin{aligned} &\mathit{narrow} :: \mathit{Interval} \to \mathit{Interval} \to \mathit{Interval} \\ &\mathit{narrow}\ [l\,..\,r)\ [p\,..\,q) \mathrel{\widehat{=}} [l + (r{-}l) \times p\,..\,l + (r{-}l) \times q) \end{aligned}$$

We consider only non-adaptive encoding here for simplicity: adaptivity turns out to be orthogonal to streaming. We therefore assume a single fixed model, represented as a function from symbols to intervals:

$$\mathit{model} :: \mathit{Symbol} \to \mathit{Interval}$$

Encoding is a three-stage process: mapping symbols into intervals, narrowing intervals to a target interval, and generating the binary representation of a fraction within that interval (missing its final 1). (It is tempting to choose a fraction then convert it to binary; but unless this is done carefully, the fraction will not be one that has a short binary representation.)

$$\begin{aligned} &\mathit{encode} :: [\mathit{Symbol}] \to [\mathit{Bool}] \\ &\mathit{encode} \mathrel{\widehat{=}} \mathsf{unfoldr}\ \mathit{nextBit} \cdot \mathsf{foldl}\ \mathit{narrow}\ \mathit{unit} \cdot \mathsf{map}\ \mathit{model} \end{aligned}$$

where

$$
\begin{array}{l}
\mathit{nextBit}\ (l, r) \\
\quad \mid r \leq \tfrac{1}{2} \quad \widehat{=}\ \mathsf{Just}\ (\mathit{False}, \mathit{narrow}\ (0, 2)\ (l, r)) \\
\quad \mid \tfrac{1}{2} \leq l \quad \widehat{=}\ \mathsf{Just}\ (\mathit{True}, \mathit{narrow}\ (-1, 1)\ (l, r)) \\
\quad \mid l < \tfrac{1}{2} < r \ \widehat{=}\ \mathsf{Nothing}
\end{array}
$$

This can be turned into a metamorphism, as the map fuses with the fold.

As described, this is not a very efficient encoding method: the entire message has to be digested into a target interval before any of the fraction can be generated. However, the streaming condition holds, and bits of the fraction can be produced before all of the message is consumed:

$$
\begin{array}{l}
\mathit{encode} \quad :: [\mathit{Symbol}] \to [\mathit{Bool}] \\
\mathit{encode}\ m = \mathit{stream}\ \mathit{nextBit}\ \mathit{narrow}\ \mathit{unit} \cdot \mathsf{map}\ \mathit{model}
\end{array}
$$

(Again, the map fuses with the stream.)

6.3 A unified view of number representations

The observant reader may have noticed some similarities between three of the examples discussed in this article: radix conversion (Section 4), arithmetic on continued fractions (Section 5), and computing the digits of π (Section 6.1). Indeed, all three may be unified into a general scheme described by Potts and Edalat [35,31], and based on infinite sequences of linear fractional transformations. We summarize the ideas here.

6.3.1 The general scheme

The general idea is that a number x in some base interval I is represented by an infinite sequence $\alpha = \langle \alpha_1, \alpha_2, \ldots \rangle$ of linear fractional transformations. The elements of the matrix representing the transformation and the endpoints of the interval are drawn from some field. Each matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is non-singular, that is, $ad - bc \neq 0$.

The image $\alpha(J)$ of an interval J under a transformation α is defined in the obvious way. The ith approximant $\alpha|i$ of a number representation $\alpha = \langle \alpha_1, \alpha_2, \ldots \rangle$ is defined to be $\alpha_1(\alpha_2(\cdots \alpha_i(I)))$. Representation α is valid if these approximants are properly nested: $I = \alpha|0 \supseteq \alpha|1 \supseteq \alpha|2 \supseteq \cdots$, which follows if each term α_i satisfies $\alpha_i(I) \subseteq I$. A valid representation may still not represent a unique number (for example, if each α_i is the identity transformation);

representation α represents number x if the approximants $\alpha|i$ converge to a singleton set $\bigcap_i \alpha|i = \{x\}$.

6.3.2 Decimal representation

For example, decimal representation uses the field of rationals, the base interval $I = [0, 1]$, and terms $\left(\begin{smallmatrix} 1 & d \\ 0 & 10 \end{smallmatrix}\right)$ for $d \in 0, 1 \ldots 9$. Indeed, the terms are non-singular, and $\left(\begin{smallmatrix} 1 & d \\ 0 & 10 \end{smallmatrix}\right)[0, 1] = [{}^{d}\!/_{10}, {}^{d+1}\!/_{10}] \subseteq [0, 1]$, so all sequences of terms are valid. Moreover, each term shrinks an interval by a factor of ten, so any sequence of terms converges to a single representative.

6.3.3 Continued fractions

Completely regular continued fractions, in which the first coefficient is at least 1 like all the other coefficients, uses the field of rationals extended with ∞ (the type *ExtRat* of Section 5.1), the base interval $I = [1, \infty]$, and terms $\left(\begin{smallmatrix} b & 1 \\ 1 & 0 \end{smallmatrix}\right)$ with $b \in 1, 2 \ldots$. Again, the terms are non-singular, and $\left(\begin{smallmatrix} b & 1 \\ 1 & 0 \end{smallmatrix}\right)[1, \infty] = [b, b + 1] \subseteq [1, \infty]$. It can also be shown that any sequence of such terms converges to a single representative.

6.3.4 Digits of π

Rabinowitz and Wagon's spigot algorithm uses the field of rationals, the base interval $[3, 4]$, and terms $\left(\begin{smallmatrix} k & 4k+2 \\ 0 & 2k+1 \end{smallmatrix}\right)$ for $k = 1, 2 \ldots$. Again, the terms are non-singular, and

$$\begin{pmatrix} k & 4k+2 \\ 0 & 2k+1 \end{pmatrix} [3, 4] = [3{}^{1}\!/_{2} - {}^{3}\!/_{4k+2}, 4 - {}^{2}\!/_{2k+1}] \subseteq [3, 4]$$

so all sequences of terms are valid. Moreover, each term shrinks an interval by a factor of at least two, so any sequence converges to a single representative.

6.3.5 Other representations

Potts and Edalat [35,31] use the field of rationals, the base interval $I = [0, \infty]$, and terms the *positive* non-singular linear fractional transformations (that is, those with all four elements at least zero), which can be shown to maintain proper nesting of intervals. They use these to develop various kinds of representation, including redundant ones. The scheme also covers positional representations with non-integer [36, Exercise 1.2.8–35] and imaginary [37, Section 4.1] bases.

6.3.6 Streaming conversions

The single unified setting allows one to convert between these various representations using streaming algorithms. For example, the spigot algorithm for π is based on converting from the third representation above to the first. The state consists of a non-singular linear fractional transformation h. For a simple conversion, h can be initialized to the identity transform; but an arbitrary (non-singular) rational unary function can be applied at the same time, as discussed in Section 5.2. However, the invariant that $h(I) \subseteq I$ should be maintained throughout the conversion process. Consumption is simply a matter of multiplying matrix h by the next input term α_i. Production, if possible, involves factorizing h into an output term β_j and a revised state h' such that $h = \beta_j \cdot h'$, while maintaining the above invariant. Note that non-singularity of the state is another invariant, maintained by consumption and production of non-singular terms.

To illustrate, reconsider conversion of 0.67_{10} to ternary, as described in Section 4.3. After consuming the 6_{10}, the state is $\left(\begin{smallmatrix}1 & 6\\ 0 & 10\end{smallmatrix}\right)$. The available output digits are $\left(\begin{smallmatrix}1 & d\\ 0 & 3\end{smallmatrix}\right)$ for $d = 0, 1, 2$, but none of these is in fact possible. For example, $\left(\begin{smallmatrix}1 & 6\\ 0 & 10\end{smallmatrix}\right)$ factorizes into $\left(\begin{smallmatrix}1 & 0\\ 0 & 3\end{smallmatrix}\right)\left(\begin{smallmatrix}1 & 6\\ 0 & 10/3\end{smallmatrix}\right)$, but the putative revised state $\left(\begin{smallmatrix}1 & 6\\ 0 & 10/3\end{smallmatrix}\right)$ takes the base interval $[0, 1]$ to $[18/10, 21/10] \not\subseteq [0, 1]$; similar objections apply to the other two digits. Therefore no output can be produced at this point, and the next input term 7_{10} must be consumed.

7 Future and related work

The notion of metamorphisms in general and of streaming algorithms in particular arose out of our work on arithmetic coding [8]. Since then, we have seen the same principles cropping up in other areas, most notably in the context of various kinds of numeric representations: the radix conversion problem from Section 3.2, continued fractions as described in Section 5, and computations with infinite compositions of homographies as used in Section 6.1. Indeed, one might even see arithmetic coding as a kind of numeric representation problem.

7.1 Generic streaming

Our theory of metamorphisms could easily be generalized to other datatypes: there is nothing to prevent consideration of folds consuming and unfolds producing datatypes other than lists. However, we do not currently have any convincing examples.

Perhaps related to the lack of convincing examples for other datatypes, it is not clear what a generic theory of streaming algorithms would look like. List-oriented streaming relies essentially on foldl, which does not generalize in any straightforward way to other datatypes. (We have in the past attempted to show how to generalize scanl to arbitrary datatypes [38–40], and Pardo [41] has improved on these attempts; but we do not see yet how to apply those constructions here.)

However, the unfold side of streaming algorithms does generalize easily, to certain kinds of datatype if not obviously all of them. Consider producing a data structure of the type

$$\textbf{data}\ \mathit{Generic}\ \tau\ \alpha \mathrel{\widehat{=}} \mathit{Gen}\ (\mathsf{Maybe}\ (\alpha, \tau\ (\mathit{Generic}\ \tau\ \alpha)))$$

for some instance τ of the type class *Functor*. (Lists essentially match this pattern, with τ the identity functor. The type Tree of internally-labelled binary trees introduced in Section 2 matches too, with τ being the pairing functor. In general, datatypes of this form have an empty structure, and all non-empty structures consist of a root element and a τ-shaped collection of children.) It is straightforward to generalize the streaming condition to such types:

$$f\ c = \mathsf{Just}\ (b, c') \Rightarrow f\ (g\ c\ a) = \mathsf{Just}\ (b, \mathit{fmap}\ (\lambda u \to g\ u\ a)\ c')$$

(This has been called an 'τ-invariant' or 'mongruence' [42] elsewhere.) Still, we do not have any useful applications of an unfold to a *Generic* type after a foldl.

7.2 Related work: Back to basics

Some of the ideas presented here were inspired by the work of Hutton and Meijer [43]. They studied *representation changers*, consisting of one abstraction function followed by the converse of another abstraction function. Their representation changers are analogous to our metamorphisms, with the function corresponding to the fold and the converse of a function to the unfold: in a relational setting, an unfold is just the converse of a fold, and so our metamorphisms could be seen as a special case of representation changers in which both abstraction functions are folds. We feel that restricting attention to the special case of folds and unfolds is worthwhile, because we can capitalize on their universal properties; without this restriction, one has to resort to reasoning from first principles.

Hutton and Meijer illustrate with two examples: a carry-save incrementer and radix conversion. The carry-save representation of numbers is redundant, us-

ing the redundancy to avoid carry propagation. Incrementing such a number can certainly be seen as a change of representation, and indeed their resulting program is a simple unfold. However, the problem is more than just a change of representation: the point of the exercise is to preserve some of the input representation in the output, and it isn't immediately clear how to fit that requirement into our pattern of folding to an abstract value and independently unfolding to a different representation. Their radix conversion problem is similar to ours, but their resulting algorithm is not streaming: all of the input must be read before any of the output is produced.

Acknowledgements

The material on arithmetic coding in Section 6.2, and indeed the idea of streaming algorithms in the first place, came out of joint work with Richard Bird [8] and Barney Stratford. The principles behind reversing the order of evaluation and defunctionalization presented in Section 4.2 have been known for a long time [44,20], but the presentation used here is due to Geraint Jones.

We are grateful to members of the *Algebra of Programming* research group at Oxford and of *IFIP Working Group 2.1*, the participants in the *Datatype-Generic Programming* project, and the anonymous *Mathematics of Program Construction* referees for their helpful suggestions regarding this work. The observation in Section 4.4 that *fstream* can be expressed in terms of *stream* is due to Barney Stratford.

References

[1] T. Hagino, A typed lambda calculus with categorical type constructors, in: D. H. Pitt, A. Poigné, D. E. Rydeheard (Eds.), Category Theory and Computer Science, Vol. 283 of Lecture Notes in Computer Science, Springer-Verlag, 1987, pp. 140–157.

[2] G. Malcolm, Data structures and program transformation, Science of Computer Programming 14 (1990) 255–279.

[3] R. Bird, O. de Moor, The Algebra of Programming, Prentice-Hall, 1996.

[4] L. H. Sullivan, The tall office building artistically considered, Lippincott's Magazine, Mar. 1896.

[5] E. Meijer, M. Fokkinga, R. Paterson, Functional programming with bananas, lenses, envelopes and barbed wire, in: J. Hughes (Ed.), Functional Programming Languages and Computer Architecture, Vol. 523 of Lecture Notes in Computer Science, Springer-Verlag, 1991, pp. 124–144.

[6] D. Swierstra, O. de Moor, Virtual data structures, in: B. Möller, H. Partsch, S. Schumann (Eds.), IFIP TC2/WG2.1 State-of-the-Art Report on Formal Program Development, Vol. 755 of Lecture Notes in Computer Science, Springer-Verlag, 1993, pp. 355–371.

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

[8] R. Bird, J. Gibbons, Arithmetic coding with folds and unfolds, in: J. Jeuring, S. P. Jones (Eds.), Advanced Functional Programming 4, Vol. 2638 of Lecture Notes in Computer Science, Springer-Verlag, 2003, pp. 1–26.

[9] J. Gibbons, Streaming representation-changers, in: D. Kozen (Ed.), Mathematics of Program Construction, Vol. 3125 of Lecture Notes in Computer Science, 2004, pp. 142–168.

[10] E. Gamma, R. Helm, R. Johnson, J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995.

[11] J. Hughes, Why functional programming matters, Computer Journal 32 (2) (1989) 98–107, also in [45].

[12] S. Peyton Jones, The Haskell 98 Language and Libraries: The Revised Report, Cambridge University Press, 2003.

[13] J. Gibbons, G. Jones, The under-appreciated unfold, in: Proceedings of the Third ACM SIGPLAN International Conference on Functional Programming, Baltimore, Maryland, 1998, pp. 273–279.

[14] J. Gibbons, Origami programming, in: J. Gibbons, O. de Moor (Eds.), The Fun of Programming, Cornerstones in Computing, Palgrave, 2003, pp. 41–60.

[15] J. Gibbons, Calculating functional programs, in: R. Backhouse, R. Crole, J. Gibbons (Eds.), Algebraic and Coalgebraic Methods in the Mathematics of Program Construction, Vol. 2297 of Lecture Notes in Computer Science, Springer-Verlag, 2002, pp. 148–203.

[16] M. A. Jackson, Principles of Program Design, Academic Press, 1975.

[17] C. A. R. Hoare, Quicksort, Computer Journal 5 (1962) 10–15.

[18] L. Augusteijn, Sorting morphisms, in: S. D. Swierstra, P. R. Henriques, J. N. Oliveira (Eds.), Advanced Functional Programming, Vol. 1608 of Lecture Notes in Computer Science, 1998, pp. 1–27.

[19] R. S. Bird, P. L. Wadler, An Introduction to Functional Programming, Prentice-Hall, 1988.

[20] J. C. Reynolds, Definitional interpreters for higher-order programming languages, Higher Order and Symbolic Computing 11 (4) (1998) 363–397, reprinted from the Proceedings of the 25th ACM National Conference, 1972.

[21] V. Vene, T. Uustalu, Functional programming with apomorphisms (corecursion), Proc. of the Estonian Academy of Sciences: Physics, Mathematics 47 (3) (1998) 147–161, 9th Nordic Workshop on Programming Theory.

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

[23] R. S. Bird, Introduction to Functional Programming Using Haskell, Prentice-Hall, 1998.

[24] J. Gibbons, G. Hutton, Proof methods for corecursive programs, Fundamenta Informatica.

[25] M. Beeler, R. W. Gosper, R. Schroeppel, Hakmem, AIM 239, MIT (Feb. 1972).

[26] B. Gosper, Continued fraction arithmetic, unpublished manuscript (1981).

[27] S. Peyton Jones, Arbitrary precision arithmetic using continued fractions, INDRA Working Paper 1530, Dept of CS, University College, London (Jan. 1984).

[28] J. Vuillemin, Exact real arithmetic with continued fractions, IEEE Transactions on Computers 39 (8) (1990) 1087–1105.

[29] D. Lester, Vuillemin's exact real arithmetic, in: R. Heldal, C. K. Holst, P. Wadler (Eds.), Glasgow Functional Programming Workshop, 1991, pp. 225–238.

[30] D. Lester, Effective continued fractions, in: Proceedings of the Fifteenth IEEE Arithmetic Conference, 2001, pp. 163–172.

[31] P. J. Potts, Exact real arithmetic using Möbius transformations, Ph.D. thesis, Imperial College, London (Jul. 1998).

[32] J. Gibbons, Unbounded spigot algorithms for the digits of π, submitted for publication (Aug. 2004).

[33] S. Rabinowitz, S. Wagon, A spigot algorithm for the digits of π, American Mathematical Monthly 102 (3) (1995) 195–203.

[34] I. H. Witten, R. M. Neal, J. G. Cleary, Arithmetic coding for data compression, Communications of the ACM 30 (6) (1987) 520–540.

[35] A. Edalat, P. J. Potts, A new representation for exact real numbers, Electronical Notes in Theoretical Computer Science 6 (1997) 13 pp.

[36] D. E. Knuth, The Art of Computer Programming, Volume 1: Fundamental Algorithms, Addison-Wesley, 1968.

[37] D. E. Knuth, The Art of Computer Programming, Volume 2: Seminumerical Algorithms, Addison-Wesley, 1969.

[38] J. Gibbons, Algebras for tree algorithms, D. Phil. thesis, Programming Research Group, Oxford University, available as Technical Monograph PRG-94. ISBN 0-902928-72-4 (1991).

[39] J. Gibbons, Polytypic downwards accumulations, in: J. Jeuring (Ed.), Proceedings of Mathematics of Program Construction, Vol. 1422 of Lecture Notes in Computer Science, Springer-Verlag, Marstrand, Sweden, 1998, pp. 207–233.

[40] J. Gibbons, Generic downwards accumulations, Science of Computer Programming 37 (2000) 37–65.

[41] A. Pardo, Generic accumulations, in: J. Gibbons, J. Jeuring (Eds.), Generic Programming, Kluwer Academic Publishers, 2003, pp. 49–78, proceedings of the IFIP TC2 Working Conference on Generic Programming, Schloß Dagstuhl, July 2002. ISBN 1-4020-7374-7.

[42] B. Jacobs, Mongruences and cofree coalgebras, in: Algebraic Methodology and Software Technology, Vol. 936 of Lecture Notes in Computer Science, 1995, pp. 245–260.

[43] G. Hutton, E. Meijer, Back to basics: Deriving representation changers functionally, Journal of Functional Programming 6 (1) (1996) 181–188.

[44] E. Boiten, The many disguises of accumulation, Tech. Rep. 91-26, Department of Informatics, University of Nijmegen (Dec. 1991).

[45] D. A. Turner (Ed.), Research Topics in Functional Programming, University of Texas at Austin, Addison-Wesley, 1990.

A Appendix: Haskell code

In the body of the paper, we have presented fragments of 'idealized Haskell'. To avoid obscuring the discussion with notation, we have used typographic effects in formatting the code, and have elided some details (such as coercions between Haskell's different numeric types, and declaration of type class contexts for ad-hoc polymorphism). For absolute clarity, we include the raw Haskell code here.

A.1 Code from Section 2

```
> data Tree a = Empty | Node (a, Tree a, Tree a)

> foldt :: (Maybe (a,b,b) -> b) -> Tree a -> b
> foldt f Empty = f Nothing
> foldt f (Node (a,t,u)) = f (Just (a, foldt f t, foldt f u))

> unfoldt :: (b -> Maybe (a,b,b)) -> b -> Tree a
> unfoldt f b = case f b of
```

```
>   Nothing -> Empty
>   Just (a, b1, b2) -> Node (a, unfoldt f b1, unfoldt f b2)
```

A.2 Code from Section 3.1

```
> reformat :: Int -> [[a]] -> [[a]]
> reformat n = writeLines n . readLines

> readLines :: [[a]] -> [a]
> readLines = foldr (++) []
> writeLines :: Int -> [a] -> [[a]]
> writeLines n = unfoldr (split n)
> split n x = if null x then Nothing else Just (splitAt n x)
```

A.3 Code from Section 3.2

```
> radixConvert :: (Integer,Integer) -> [Integer] -> [Integer]
> radixConvert (b,c) = toBase c . fromBase b

> fromBase :: Integer -> [Integer] -> Rational
> fromBase b = foldr (step b) 0
> step b n x = (x + fromInteger n) / fromInteger b

> toBase :: Integer -> Rational -> [Integer]
> toBase b = unfoldr (next b)
> next b 0 = Nothing
> next b x = Just (floor y, y - i2r (floor y))
>   where y = i2r b * x

> i2r :: Integer -> Rational
> i2r = fromInteger
```

A.4 Code from Section 3.3

```
> quicksort :: Ord a => [a] -> [a]
> quicksort = foldt join . unfoldt partition
```

```
> partition :: Ord a => [a] -> Maybe (a, [a], [a])
> partition [] = Nothing
> partition (a:x) = Just (a, filter (<=a) x, filter (>a) x)

> join :: Maybe (a, [a], [a]) -> [a]
> join Nothing = []
> join (Just (a,x,y)) = x ++ [a] ++ y

> heapsort :: Ord a => [a] -> [a]
> heapsort = unfoldr splitMin . foldr insert Empty

> insert :: Ord a => a -> Tree a -> Tree a
> insert a t = merge (Node (a,Empty,Empty), t)

> splitMin :: Ord a => Tree a -> Maybe (a, Tree a)
> splitMin Empty = Nothing
> splitMin (Node (a,t,u)) = Just (a, merge (t,u))

> merge :: Ord a => (Tree a, Tree a) -> Tree a
> merge (t, Empty) = t
> merge (Empty, u) = u
> merge (t, u)
>   | minElem t < minElem u = glue (t,u)
>   | otherwise             = glue (u,t)
>   where glue (Node (a,t,u), v) = Node (a, u, merge (t,v))
>         minElem (Node (a,t,u)) = a
```

A.5 Code from Section 4.1

```
> stream :: (c -> Maybe (b,c)) -> (c -> a -> c) -> c -> [a] -> [b]
> stream f g c x =
>   case f c of
>     Just (b, c') -> b : stream f g c' x
>     Nothing -> case x of
>       (a:x') -> stream f g (g c a) x'
>       []     -> []
```

A.6 Code from Section 4.2

```
> fromBase' :: Integer -> [Integer] -> Rational
> fromBase' b = app . foldl (astle b) ident where
>   astle b (n,d) m = (m + b * n,  b * d)
>   ident = (0,1)
>   app (n,d) = step d n 0
```

A.7 Code from Section 4.4

```
> guard :: (b -> Bool) -> (b -> Maybe a) -> (b -> Maybe a)
> guard p f b = if p b then f b else Nothing

> fstream :: (c -> Maybe (b,c)) -> (c -> a -> c) -> (c -> [b]) ->
>            c -> [a] -> [b]
> fstream f g h c x =
>   case f c of
>     Just (b, c') -> b : fstream f g h c' x
>     Nothing -> case x of
>       (a:x') -> fstream f g h (g c a) x'
>       []     -> h c

> apol :: (b -> Maybe (a,b)) -> (b -> [a]) -> b -> [a]
> apol f h b =
>   case f b of
>     Just (a,b') -> a : apol f h b'
>     Nothing     -> h b

> fstream' f g h c x = stream f' g' (Left c) (map Just x ++ [Nothing])
>   where
>     f' (Left c)  = case f c of Just (b,c') -> Just (b, Left c')
>                                Nothing     -> Nothing
>     f' (Right y) = case y of b:y' -> Just (b, Right y')
>                              nil  -> Nothing
>     g' (Left c) (Just a) = Left (g c a)
>     g' (Left c) Nothing  = Right (h c)
```

A.8 Code from Section 4.6

```
> radixConvert' :: (Integer,Integer) -> [Integer] -> [Integer]
> radixConvert' (b,c) = fstream snextapp astle (unfoldr nextapp) (0,1)
>   where
>     nextapp (0,r) = Nothing
>     nextapp (n,r) = Just (u, (n - i2r u * r / i2r c, r / i2r c))
>                       where u = floor (n * i2r c / r)
>     snextapp = guard safe nextapp
>     safe (n,r) = floor (n * i2r c / r) == floor ((n+1) * i2r c / r)
>     astle (n,d) m = (i2r m + i2r b * n,  i2r b * d)
```

A.9 Code from Section 5.1

```
> data ExtRat = Finite Rational | Infinite deriving (Eq,Ord)

> lift2 :: (Rational->Rational->Rational) -> (ExtRat->ExtRat->ExtRat)
> lift2 f (Finite x) (Finite y) = Finite (f x y)
> lift2 f _ _ = Infinite

> lift1 :: (Rational->Rational) -> (ExtRat->ExtRat)
> lift1 f (Finite x) = Finite (f x)
> lift1 f Infinite = error "Infinity"

> instance Num ExtRat where
>   (+) = lift2 (+)
>   (-) = lift2 (-)
>   (*) = lift2 (*)
>   signum = lift1 signum
>   abs = lift1 abs
>   fromInteger n = Finite (fromInteger n)

> instance Real ExtRat where
>   toRational (Finite x) = x
>   toRational Infinite = error "Infinity"

> instance Fractional ExtRat where
>   recip (Finite 0) = Infinite
>   recip (Finite x) = Finite (recip x)
>   recip Infinite = Finite 0
>   fromRational x = Finite x
```

```
> instance RealFrac ExtRat where
>   properFraction (Finite x) = (n, Finite y)
>     where (n,y) = properFraction x

> instance Show ExtRat where
>   show (Finite x) = show x
>   show Infinite = "oo"

> i2e :: Integer -> ExtRat
> i2e = fromInteger

> type CF = [Integer]

> toCF :: ExtRat -> CF
> toCF = unfoldr get where
>   get Infinite = Nothing
>   get (Finite x) = Just (y, recip (Finite (x - i2r y)))
>     where y = floor x

> fromCF :: CF -> ExtRat
> fromCF = foldr put Infinite
>   where put n y = fromInteger n + recip y

> fromCF' :: CF -> ExtRat
> fromCF' = apph . foldl astleh identh

> fromCFi :: CF -> [ExtRat]
> fromCFi = map apph . scanl astleh identh

> type Homog = (Integer,Integer,Integer,Integer)

> astleh :: Homog -> Integer -> Homog
> astleh (q,r,s,t) n = (n*q+r, q, n*s+t, s)

> identh :: Homog
> identh = (1,0,0,1)

> apph :: Homog -> ExtRat
> apph (q,r,s,t) = i2e q / i2e s

> absh :: Homog -> Integer -> ExtRat
> absh (q,r,s,t) n = fromInteger (n*q+r) / fromInteger (n*s+t)
```

A.10 Code from Section 5.2

```
> emit :: Homog -> Integer -> Homog
> emit (q,r,s,t) n = (s, t, q-n*s, r-n*t)

> rfc :: Homog -> CF -> CF
> rfc h = unfoldr geth . foldl astleh h

> geth :: Homog -> Maybe (Integer, Homog)
> geth (q,r,s,t)
>   | s /= 0    = Just (n, emit (q,r,s,t) n)
>   | otherwise = Nothing
>   where n = q `div` s

> gets :: Homog -> Maybe (Integer, Homog)
> gets = guard same geth
>   where same (q,r,s,t) = s*t>0 && (q+r) `div` (s+t) == q `div` s

> rfc' :: Homog -> CF -> CF
> rfc' h = fstream gets astleh (unfoldr geth) h

> rf :: Homog -> CF -> CF
> rf h [] = rfc h []
> rf h (n:x) = rfc (astleh h n) x
```

A.11 Code from Section 6.1

```
> pi = stream prod mm identh lfts where
>   lfts = [ (k, 4*k+2, 0, 2*k+1) | k <- [1..] ]
>   prod h
>     | floor (absh h 4) == n = Just (n, mm (10, -10*n, 0, 1) h)
>     | otherwise             = Nothing
>     where n = floor (absh h 3)
>   mm (q,r,s,t) (q',r',s',t')
>     = (q*q'+r*s',q*r'+r*t',s*q'+t*s',s*r'+t*t')
```

A.12 Code from Section 6.2

```
> type Interval = (Rational, Rational)

> unit :: Interval
> unit = (0,1)

> m `within` (l,r) = l<=m && m<r

> narrow :: Interval -> Interval -> Interval
> narrow (l,r) (p,q) = (l+(r-l)*p, l+(r-l)*q)

> model :: Char -> Interval
> model c = maybe undefined id (lookup c table)
> table :: [ (Char, Interval) ]
> table = zip chars ints where
>     chars = "abcde"
>     weights = [4,3,2,3,5] :: [Rational]
>     cumul = scanl (+) 0 (map (/ sum weights) weights)
>     ints = zip cumul (tail cumul)

> encode, encode' :: [Char] -> [Bool]
> encode = unfoldr nextBit . foldl narrow unit . map model
> encode' = stream nextBit narrow unit . map model
> nextBit (l,r)
>   | r <= 1/2 = Just (False, narrow (0,2) (l,r))
>   | 1/2 <= l = Just (True, narrow (-1,1) (l,r))
>   | otherwise = Nothing

> decode :: [Bool] -> [Char]
> decode = unfoldr nextChar . centre . foldl narrow unit . map follow
> follow b = if b then (1/2,1) else (0,1/2)
> centre (l,r) = (l+r)/2
> nextChar x = lookup True (invert x table)
> invert x t = [ (x `within` (p,q),(c, (x-p)/(q-p))) | (c,(p,q)) <- t ]
```

A.13 Code from Section 7.1

```
> data Functor f => Generic f a = Gen (Maybe (a, f (Generic f a)))
```