Polymonadic programming

Mike Hicks[2], Gavin Bierman[1], Nataliya Guts[2], Daan Leijen[1], and Nik Swamy[1]

[1] Microsoft Research
[2] University of Maryland, College Park

Abstract. Monads are a popular tool for the working functional programmer to structure effectful computations. This paper presents *polymonads*, a generalization of monads. Polymonads give the familiar monadic bind the more general type $\forall a, b.\ \mathsf{L}\ a \rightarrow (a \rightarrow \mathsf{M}\ b) \rightarrow \mathsf{N}\ b$, to compose computations with three different kinds of effects, rather than just one. Polymonads subsume monads and parameterized monads, and can express other constructions, including precise type-and-effect systems and information flow tracking; more generally, polymonads correspond to Tate's *productoid* semantic model. We show how to equip a core language (called $\lambda_{\underline{\mathrm{PM}}}$) with syntactic support for programming with polymonads. Type inference and elaboration in $\lambda_{\underline{\mathrm{PM}}}$ allows programmers to write polymonadic code directly in an ML-like syntax—our algorithms compute principal types and produce elaborated programs wherein the binds appear explicitly. Furthermore, we prove that the elaboration is *coherent*: no matter which (type-correct) binds are chosen, the elaborated program's semantics will be the same. Pleasingly, the inferred types are easy to read: the polymonad laws justify (sometimes dramatic) simplifications, but with no effect on a type's generality.

1 Introduction

Since the time that Moggi first connected them to effectful computation [19], *monads* have proven to be a surprisingly versatile computational structure. Perhaps best known as the foundation of Haskell's support for state, I/O, and other effects, monads have also been used to structure APIs for libraries that implement a wide range of programming tasks, including parsers [12], probabilistic computations [23], and functional reactivity [7, 4].

Monads (and morphisms between them) are not a panacea, however, and so researchers have proposed various extensions. Examples include Wadler and Thiemann's [28] indexed monad for typing effectful computations; Filliâtre's generalized monads [9]; Atkey's parameterized monad [3], which has been used to encode disciplines like regions [16] and session types [22]; Devriese and Piessens' [6] monad-like encodings for information flow controls; and many others. Oftentimes these extensions are needed to prove stronger properties about computations, for instance to prove the absence of information leaks or memory errors.

Unfortunately, these extensions do not enjoy the same status as monads in terms of language support. For example, the conveniences that Haskell provides for monadic programs (e.g., the `do` notation combined with type-class inference) do not apply to these extensions. One might imagine adding specialized support for each of these extensions on a case-by-case basis, but a unifying construction into which all of them, including normal monads, fit is clearly preferable.

This paper proposes just such a unifying construction, making several contributions. Our first contribution is the definition of a *polymonad*, a new way to structure effectful computations. Polymonads give the familiar monadic bind (having type $\forall a, b.\ \mathsf{M}\ a \to (a \to \mathsf{M}\ b) \to \mathsf{M}\ b$) the more general type $\forall a, b.\ \mathsf{L}\ a \to (a \to \mathsf{M}\ b) \to \mathsf{N}\ b$. That is, a polymonadic bind can compose computations with three different types to a monadic bind's one. Section 2 defines polymonads formally, along with the *polymonad laws*, which we prove are a generalization of the monad and morphism laws. To precisely characterize their expressiveness, we prove that polymonads correspond to Tate's *productoids* [26] (Theorem 1), a recent semantic model general enough to capture most known effect systems, including all the constructions listed above.[3]

Whereas Tate's interest is in semantically modeling sequential compositions of effectful computations, our interest is in supporting practical programming in a higher-order language. Our second contribution is the definition of $\lambda\underline{\text{PM}}$ (Section 3), an ML-like programming language well-suited to programming with polymonads. We work out several examples in $\lambda\underline{\text{PM}}$, including novel polymonadic constructions for stateful information flow tracking, contextual type and effect systems [20], and session types.

Our examples are made practical by $\lambda\underline{\text{PM}}$'s support for type inference and elaboration, which allows programs to be written in a familiar ML-like notation while making no mention of the bind operators. Enabling this feature, our third contribution (Section 4) is an instantiation of Jones' theory of qualified types [13] to $\lambda\underline{\text{PM}}$. In a manner similar to Haskell's type class inference, we show that type inference for $\lambda\underline{\text{PM}}$ computes *principal types* (Theorem 2). Our inference algorithm is equipped with an elaboration phase, which translates source terms by inserting binds where needed. We prove that elaboration is *coherent* (Theorem 3), meaning that when inference produces constraints that could have several solutions, when these solutions are applied to the elaborated terms the results will have equivalent semantics, thanks to the polymonad laws. This property allows us to do better than Haskell, which does not take such laws into account, and so needlessly rejects programs it thinks might be ambiguous. Moreover, as we show in Section 5, the polymonad laws allow us to dramatically simplify types, making them far easier to read without compromising their generality. A prototype implementation of $\lambda\underline{\text{PM}}$ is available from the first author's web page and has been used to check all the examples in the paper.

Put together, our work lays the foundation for providing practical support for advanced monadic programming idioms in typed, functional langaues.

[3] In fact, we discovered the same model concurrently with Tate and independently of him, though our presentation here has benefited from conversations with him.

2 Polymonads

We begin by defining polymonads formally. We prove that a polymonad generalizes a collection of monads and morphisms among those monads. We also establish a correspondence between polymonads and productoids, placing our work on a semantic foundation that is known to be extremely general.

Definition 1. *A* ***polymonad*** $(\mathcal{M}, \Sigma)$ *consists of (1) a collection $\mathcal{M}$ of unary type constructors, with a distinguished element* $\mathsf{Id} \in \mathcal{M}$*, such that* $\mathsf{Id}\ \tau = \tau$*, and (2) a collection, Σ, of* bind *operators such that the laws below hold, where* $(\mathsf{M}, \mathsf{N}) \rhd \mathsf{P} \triangleq \forall \mathsf{a}\ \mathsf{b}.\ \mathsf{M}\ \mathsf{a} \to (\mathsf{a} \to \mathsf{N}\ \mathsf{b}) \to \mathsf{P}\ \mathsf{b}$*.*

For all $\mathsf{M}, \mathsf{N}, \mathsf{P}, \mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}, \mathsf{U} \in \mathcal{M}$.

(Functor) $\exists \mathsf{b}.\mathsf{b}{:}(\mathsf{M}, \mathsf{Id}) \rhd \mathsf{M} \in \Sigma$ *and* $\mathsf{b}\ m\ (\lambda \mathsf{y}.\mathsf{y}) = m$

(Paired morphisms) $\exists \mathsf{b}_1{:}(\mathsf{M}, \mathsf{Id}) \rhd \mathsf{N} \in \Sigma \iff \exists \mathsf{b}_2{:}(\mathsf{Id}, \mathsf{M}) \rhd \mathsf{N} \in \Sigma$ and
$\forall \mathsf{b}_1{:}(\mathsf{M}, \mathsf{Id}) \rhd \mathsf{N}, \mathsf{b}_2{:}(\mathsf{Id}, \mathsf{M}) \rhd \mathsf{N}.\mathsf{b}_1\ (f\ v)\ (\lambda \mathsf{y}.\mathsf{y}) = \mathsf{b}_2\ v\ f$

(Associativity) $\forall \mathsf{b}_1, \mathsf{b}_2, \mathsf{b}_3, \mathsf{b}_4.$*If*
$\{\mathsf{b}_1{:}(\mathsf{M}, \mathsf{N}) \rhd \mathsf{P}, \mathsf{b}_2{:}(\mathsf{P}, \mathsf{R}) \rhd \mathsf{T}, \mathsf{b}_3{:}(\mathsf{M}, \mathsf{S}) \rhd \mathsf{T}, \mathsf{b}_4{:}(\mathsf{N}, \mathsf{R}) \rhd \mathsf{S}\} \subseteq \Sigma$
then $\mathsf{b}_2\ (\mathsf{b}_1\ m\ f)\ g = \mathsf{b}_3\ m\ (\lambda x.\mathsf{b}_4\ (f\ x)\ g)$

(Diamond) $\exists \mathsf{V}_1, \mathsf{b}_1, \mathsf{b}_2.\{\mathsf{b}_1{:}(\mathsf{M}, \mathsf{N}) \rhd \mathsf{V}_1, \mathsf{b}_2{:}(\mathsf{V}_1, \mathsf{P}) \rhd \mathsf{T}\} \subseteq \Sigma \iff$
$\exists \mathsf{V}_2, \mathsf{b}_3, \mathsf{b}_4.\{\mathsf{b}_3{:}(\mathsf{N}, \mathsf{P}) \rhd \mathsf{V}_2, \mathsf{b}_4{:}(\mathsf{M}, \mathsf{V}_2) \rhd \mathsf{T}\} \subseteq \Sigma$

(Closure) *If* $\exists \mathsf{b}_1, \mathsf{b}_2, \mathsf{b}_3, \mathsf{b}_4.$
$\{\mathsf{b}_1{:}(\mathsf{M}, \mathsf{N}) \rhd \mathsf{P}, \mathsf{b}_2{:}(\mathsf{S}{,}\mathsf{Id}) \rhd \mathsf{M}, \mathsf{b}_3{:}(\mathsf{T}{,}\mathsf{Id}) \rhd \mathsf{N}, \mathsf{b}_4{:}(\mathsf{P}{,}\mathsf{Id}) \rhd \mathsf{U}\} \subseteq \Sigma$
then $\exists \mathsf{b}.\mathsf{b}{:}(\mathsf{S}, \mathsf{T}) \rhd \mathsf{U} \in \Sigma$

Definition 1 may look a little austere, but there is a simple refactoring that recovers the structure of functors and monad morphisms from a polymonad. Given $(\mathcal{M}, \Sigma)$, we can easily construct the following sets:

$$\begin{array}{lll}
\text{(Maps)} & M & = \{(\lambda f m.\mathsf{bind}\ m\ f) \colon (\mathsf{a} \to \mathsf{b}) \to \mathsf{M}\ \mathsf{a} \to \mathsf{M}\ \mathsf{b} \mid \mathsf{bind} \colon (\mathsf{M}, \mathsf{Id}) \rhd \mathsf{M} \in \Sigma\} \\
\text{(Units)} & U & = \{(\lambda x.\mathsf{bind}\ x\ (\lambda y.y)) \colon \mathsf{a} \to \mathsf{M}\ \mathsf{a} \mid \mathsf{bind} \colon (\mathsf{Id}, \mathsf{Id}) \rhd \mathsf{M} \in \Sigma\} \\
\text{(Lifts)} & L & = \{(\lambda x.\mathsf{bind}\ x\ (\lambda y.y)) \colon \mathsf{M}\ \mathsf{a} \to \mathsf{N}\ \mathsf{a} \mid \mathsf{bind} \colon (\mathsf{M}, \rhd \mathsf{N} \in \Sigma\}
\end{array}$$

It is fairly easy to show that the above structure satisfies generalizations of the familiar laws for monads and monad morphisms. For example, one can prove $\mathsf{bind}\ (\mathsf{unit}\ e)\ f = f\ e$, and $\mathsf{lift}\ (\mathsf{unit}_1\ e) = \mathsf{unit}_2\ e$ for all suitably typed $\mathsf{unit}_1, \mathsf{unit}_2 \in U$, $\mathsf{lift} \in L$ and $\mathsf{bind} \in \Sigma$.

With these intuitions in mind, one can see that the **Functor** law ensures that each $\mathsf{M} \in \Sigma$ has a map in M, as expected for monads. From the construction of L, one can see that a bind $(\mathsf{M}, \mathsf{Id}) \rhd \mathsf{N}$ is just a morphism from M to N. Since this comes up quite often, we write $\mathsf{M} \hookrightarrow \mathsf{N}$ as a shorthand for $(\mathsf{M}, \mathsf{Id}) \rhd \mathsf{N}$. The **Paired morphisms** law amounts to a coherence condition that all morphisms can be re-expressed as binds. The **Associativity** law is the familiar associativity law for monads generalized for both our more liberal typing for bind operators and for the fact that we have a *collection* of binds rather than a single bind. The **Diamond** law essentially guarantees a coherence property for associativity, namely that it is always possible to complete an application of **Associativity**. The **Closure** law ensures closure under composition of monad morphisms with binds, also for coherence.

Signatures$(\mathcal{M}, \Sigma)$:	k-ary constructors	$\mathcal{M}$	$::= \cdot \mid M/k, \mathcal{M}$
	ground constructor	M	$::= M\ \overline{\tau}$
	bind set	Σ	$::= \cdot \mid \mathsf{b}{:}s, \Sigma$
	bind specifications	s	$::= \forall \bar{a}.\Phi \Rightarrow (\mathsf{M}_1, \mathsf{M}_2) \rhd \mathsf{M}_3$
	theory constraints	Φ	
Terms:	values	v	$::= x \mid c \mid \lambda x.e$
	expressions	e	$::= v \mid e_1\ e_2 \mid \mathsf{let}\ x = e_1\ \mathsf{in}\ e_2$
			$\mid\ \mathsf{if}\ e\ \mathsf{then}\ e_1\ \mathsf{else}\ e_2 \mid \mathsf{letrec}\ f = v\ \mathsf{in}\ e$
Types:	monadic types	m	$::= \mathsf{M} \mid \rho$
	value types	τ	$::= a \mid T\ \overline{\tau} \mid \tau_1 \to m\ \tau_2$
	type schemes	σ	$::= \forall \bar{a}\bar{\rho}.P \Rightarrow \tau$
	bag of binds	P	$::= \cdot \mid \pi, P$
	bind type	π	$::= (m_1, m_2) \rhd m_3$

Fig. 1. $\lambda_{\underline{\mathrm{PM}}}$: Syntax for signatures, types, and terms

It is easy to prove that every collection of monads and monad morphisms is also a polymonad. In fact, in Appendix A, we prove a stronger result that relates polymonads to Tate's *productoids* [26].

Theorem 1. *Every polymonad gives rise to a productoid, and every productoid that contains an* Id *element and whose joins are closed with respect to the lifts, is a polymonad.*

Tate developed productoids as a categorical foundation for effectful computation. He demonstrates the expressive power of productoids by showing how they subsume other proposed extensions to monads [28, 8, 3]. This theorem shows polymonads can be soundly interpreted using productoids. Strictly speaking, productoids are more expressive than polymonads, since they do not, in general, need to have an Id element, and only satisfy a slightly weaker form of our **Closure** condition. However, these restrictions are mild, and certainly in categories that are Cartesian closed, these conditions are trivially met for all productoids. Thus, for programming purposes, polymonads and productoids have exactly the same expressive power. The development of the rest of this paper shows, for the first time, how to harness this expressive power in a higher-order programming language, tackling the problem of type inference, elaborating a program while inserting binds, and proving elaboration coherent.

3 Programming with polymonads

This section presents $\lambda_{\underline{\mathrm{PM}}}$, an ML-like language for programming with polymonads, along with several examples of its use. Figure 1 presents $\lambda_{\underline{\mathrm{PM}}}$'s syntax.

Polymonadic signatures. A $\lambda_{\underline{\mathrm{PM}}}$ *polymonadic signature* $(\mathcal{M}, \Sigma)$ amends Definition 1 in two ways. Firstly, each element M of $\mathcal{M}$ may be *type-indexed*—we write M/k to indicate that M is a $(k + 1)$-ary type constructor (we sometimes omit k for brevity). For example, constructor $W/1$ could represent an effectful computation so that $W\ \epsilon\ \tau$ characterizes computations of type τ that have effect ϵ. We write M to denote *ground constructors*, which are monadic constructors applied to all their type indexes; e.g., $W\ \epsilon$ is ground. Secondly, a bind set Σ is not specified intensionally as a set, but rather extensionally using a language of *theory constraints* Φ. In particular, Σ is a list of mappings $\mathsf{b}{:}s$ where s contains a triple $(\mathsf{M}_1, \mathsf{M}_2) \rhd \mathsf{M}_3$ along with constraints Φ, which determine how the triple's constructors may be instantiated. For example, a mapping $\mathsf{sube} : \forall \varepsilon_1, \varepsilon_2.\, \varepsilon_1 \subseteq \varepsilon_2 \Rightarrow (W\ \varepsilon_1, \mathsf{Id}) \rhd W\, \varepsilon_2$ specifies the set of binds involving type indexes $\varepsilon_1, \varepsilon_2$ such that the theory constraint $\varepsilon_1 \subseteq \varepsilon_2$ is satisfied.

$\lambda\underline{\text{PM}}$'s type system is parametric in the choice of theory constraints Φ, which allows us to encode a variety of prior monad-like systems with $\lambda\underline{\text{PM}}$. To interpret a particular set of constraints, $\lambda\underline{\text{PM}}$ requires a theory entailment relation $\vDash$. Elements of this relation, written $\Sigma \vDash \pi \rightsquigarrow \mathsf{b}; \theta$, state that there exists $\mathsf{b}{:}\forall\bar{a}.\Phi \Rightarrow (\mathsf{M}_1, \mathsf{M}_2) \rhd \mathsf{M}_3$ in Σ and a substitution θ' such that $\theta\pi = \theta'(\mathsf{M}_1, \mathsf{M}_2) \rhd \mathsf{M}_3$, and the constraints $\theta'\Phi$ are satisfiable. Here, θ is a substitution for the free (non-constant) variables in π, while θ' is an instantiation of the abstracted variables in the bind specification. The interpretation of Σ is thus the set $\{\mathsf{b}{:}\pi \mid \Sigma \vDash \pi \rightsquigarrow \mathsf{b}; \cdot\}$. Signature $(\mathcal{M}, \Sigma)$ is a polymonad if this set satisfies the polymonad laws (where each ground constructor is treated distinctly).

Our intention is that type indices are *phantom*, meaning that they are used as a type-level representation of some property of the polymonad's current state, but a polymonadic bind's implementation does not depend on them. For example, we would expect that binds treat objects of type $W\, \varepsilon\, \tau$ uniformly, for all ε; different values of ε could statically prevent unsafe operations like double-frees or dangling pointer dereferences. If an object has different states that would affect the semantics of binds, the programmer can use different constructors M for each state (rather than different type indices and the same constructor).

Terms and types. $\lambda\underline{\text{PM}}$'s term language is standard. $\lambda\underline{\text{PM}}$ programs do not explicitly reference binds, but are written in *direct style*, with implicit conversions between computations of type $m\ \tau$ and their τ-typed results. Type inference determines the bind operations to insert (or abstract) to type check a program.

To make inference feasible, we rely crucially on $\lambda\underline{\text{PM}}$'s call-by-value structure. Following our prior work on monadic programming for ML [25], we restrict the shape of types assignable to a $\lambda\underline{\text{PM}}$ program by separating value types τ from the types of polymonadic computations $m\ \tau$. Here, metavariable m may be either a ground constructor M or a polymonadic type variables ρ. The co-domain of every function is required to be a computation type $m\ \tau$, although pure functions can be typed $\tau \to \tau'$, which is a synonym for $\tau \to \mathsf{Id}\ \tau'$. We also include types $T\ \bar{\tau}$ for fully applied type constructors, e.g., $\mathsf{list}\ \mathit{int}$.

Programs can also be given type schemes σ that are polymorphic in their polymonads, e.g., $\forall a, b, \rho.(a \to \rho\, b) \to a \to \rho\, b$. Here, the variable a ranges over value types τ, while ρ ranges over ground constructors M. Type schemes may also be qualified by a set P of bind constraints π. For example, $\forall \rho.(\rho, \mathsf{Id}) \rhd \mathsf{M} \Rightarrow (\mathit{int} \to \rho\ \mathit{int}) \to \mathsf{M}\ \mathit{int}$ is the type of a function that abstracts over a bind

Signature:

$$
\begin{array}{lcll}
\mathcal{M} & = & IST/2 & \\
\Phi & ::= & l_1 \sqsubseteq l_2 \mid \Phi_1, \Phi_2 & \\
\Sigma & = & \mathsf{bId} : & \mathsf{Id} \hookrightarrow \mathsf{Id}, \\
 & & \mathsf{unitIST} : & \forall p, l.\, \mathsf{Id} \hookrightarrow IST\ p\ l, \\
 & & \mathsf{mapIST} : & \forall p_1, l_1, p_2, l_2.\, p_2 \sqsubseteq p_1, l_1 \sqsubseteq l_2 \Rightarrow \\
 & & & IST\ p_1\ l_1 \hookrightarrow IST\ p_2\ l_2, \\
 & & \mathsf{appIST} : & \forall p_1, l_1, p_2, l_2.\, p_2 \sqsubseteq p_1, l_1 \sqsubseteq l_2 \Rightarrow \\
 & & & (\mathsf{Id}, IST\ \mathsf{p}_1\ \mathsf{l}_1) \rhd IST\ \mathsf{p}_2\ \mathsf{l}_2, \\
 & & \mathsf{bIST} : & \forall p_1, l_1, p_2, l_2, p_3, l_3. \\
 & & & l_1 \sqsubseteq p_2, l_1 \sqsubseteq l_3, l_2 \sqsubseteq l_3, \\
 & & & p_3 \sqsubseteq p_1, p_3 \sqsubseteq p_2 \Rightarrow \\
 & & & (IST\ \mathsf{p}_1\ \mathsf{l}_1, IST\ \mathsf{p}_2\ \mathsf{l}_2) \rhd IST\ \mathsf{p}_3\ \mathsf{l}_3
\end{array}
$$

Types and auxiliary functions:

$$
\begin{array}{ll}
\tau : & \ldots \mid intref\ \tau \mid L \mid H \\
\mathsf{read} : & \forall l.\, intref\ l \to IST\ H\ l\ int \\
\mathsf{write} : & \forall l.\, intref\ l \to int \to IST\ l\ L\ ()
\end{array}
$$

Example program:

```
let add_interest = λsavings. λinterest.
 let currinterest = read interest in
 if currinterest > 0 then
   let currbalance = read savings in
   let newbalance =
      currbalance + currinterest in
   write savings newbalance
 else ()
```

Fig. 2. Polymonad IST, implementing stateful information flow control

having shape $(\rho, \mathsf{Id}) \rhd \mathsf{M}$. Notice that π triples may contain polymonadic type variables ρ while specification triples $s \in \Sigma$ may not. Moreover, Φ constraints never appear in σ, which is thus entirely independent of the choice of the theory.

3.1 Polymonadic information flow controls

Polymonads are appealing because they can express many interesting constructions as we now show.

Figure 2 presents a polymonad IST, which implements *stateful* information flow tracking [6, 24, 18, 5, 1]. The idea is that some program values are secret and some are public, and no information about the former should be learned by observing the latter—a property called noninterference [10]. In the setting of IST, we are worried about leaks via the heap.

Heap-resident storage cells are given type $intref\ l$ where l is the secrecy label of the referenced cell. Labels $l \in \{L, H\}$ form a lattice with order $L \sqsubset H$. A program is acceptable if data labeled H cannot flow, directly or indirectly, to computations or storage cells labeled L. In our polymonad implementation, L and H are just types T (but only ever serve as indexes), and the lattice ordering is implemented by theory constraints $l_1 \sqsubseteq l_2$ for $l_1, l_2 \in \{L, H\}$.

The polymonadic constructor $IST/2$ uses secrecy labels for its type indexes. A computation with type $IST\ p\ l\ \tau$ potentially writes to references labeled p and returns a τ-result that has security label l; we call p the *write label* and l the *output label*. Function read reads a storage cell, producing a $IST\ H\ l\ int$ computation—the second type index l matches that of l-labeled storage cell. Function write writes a storage cell, producing a $IST\ l\ L\ ()$ computation—the first type index l matches the label of the written-to storage cell. H is the most permissive write label and so is used for the first index of read, while L is the most permissive output label and so is used for the second index of write.

Aside from the identity bind bId, implemented as reverse apply, there are four kinds of binds. Unit unitIST $p\ l$ lifts a normal term into an IST computation. Bind mapIST $p\ l$ lifts a computation into a more permissive context (i.e., p_2 and

l_2 are at least as permissive as l_1 and l_2), and $\mathsf{applST}\ p\ l$ does likewise, and are implemented using mapIST as follows: $\mathsf{applST}\ p\ l = \lambda x.\lambda f.\mathsf{mapIST}\ p\ l\ (f\ x)\ (\lambda x.x)$. Finally, bind bIST composes a computation $IST\ p_1\ l_1\ \alpha$ with a function $\alpha \to IST\ p_2\ l_2\ \beta$. The constraints ensure safe information flow: $l_1 \sqsubseteq p_2$ prevents the second computation from leaking information about its l_1-secure α-typed argument into a reference cell that is less than l_1-secure. Dually, the constraints $l_1 \sqsubseteq l_3$ and $l_2 \sqsubseteq l_3$ ensure that the β-typed result of the composed computation is at least as secure as the results of each component. The final constraints $p_3 \sqsubseteq p_1$ and $p_3 \sqsubseteq p_2$ ensure that the write label of the composed computation is a lower bound of the labels of each component.

Proving $(\mathcal{M}, \Sigma)$ satisfies the polymonad laws is straightforward. The functor and paired morphism laws are easy to prove. The diamond law is more tedious: we must consider all possible pairs of binds that compose. This reasoning involves consideration of the theory constraints as implementing a lattice, and so would work for any lattice of labels, not just H and L. In all, there were ten cases to consider. We prove the associativity law for the same ten cases. This proof is straightforward as the implementation of IST ignores the indexes: read, write and various binds are just as in a normal state monad, while the indexes serve only to prevent illegal flows. Finally, proving closure is relatively straightforward—we start with each possible bind shape and then consider correctly-shaped flows into its components; in all there were eleven cases.

Example. The lower right of Figure 2 shows an example use of IST. The add_interest function takes two reference cells, savings and interest, and modifies the former by adding to it the latter if it is non-negative. Notice that expressions of type $IST\ p\ l\ \tau$ are used as if they merely had type τ—see the branch on currinterest, for example. The program is rewritten during type inference to insert, or abstract, the necessary binds so that the program type checks. This process results in the following type for add_interest:[4]

$$\forall \rho_6, \rho_{27}, a_1, a_2.P \Rightarrow intref\ a_1 \to intref\ a_2 \to \rho_{27}\ ()$$
$$\text{where } P = (\mathsf{Id}, \mathsf{Id}) \rhd \rho_6, (IST\ H\ \mathsf{a}_1, IST\ \mathsf{a}_1\ L) \rhd \rho_6, (IST\ H\ \mathsf{a}_2, \rho_6) \rhd \rho_{27}$$

The rewritten version of add_interest starts with a sequence of λ abstractions, one for each of the bind constraints in P. If we imagine these are numbered b1 ... b3, e.g., where b1 is a bind with type $(\mathsf{Id}, \mathsf{Id}) \rhd \rho_6$, then the term looks as follows (notation ... denotes code elided for simplicity):

```
λsavings. λinterest. b3 (read interest)
  (λ currinterest. if currinterest > 0 then (b2 ...) else (b1 () (λ z. z)))
```

In a program that calls add_interest, the bind constraints will be solved, and actual implementations of these binds will be passed in for each of b_i (using a kind of dictionary-passing style as with Haskell type classes).

Looking at the type of add_interest we can see how the constraints prevent improper information flows. In particular, if we tried to call add_interest with $a_1 = L$ and $a_2 = H$, then the last two constraints become $(IST\ H\ L, IST\ L\ L) \rhd$

[4] This and other example types were generated by our prototype implementation.

$$
\begin{array}{ll}
\mathcal{M} = \mathit{CE}/3 \\
\Sigma\ = \mathsf{bId} : (\mathsf{Id}, \mathsf{Id}) \rhd \mathsf{Id}, \\
\quad \mathsf{unitce} : (\mathsf{Id}, \mathsf{Id}) \rhd \mathit{CE}\ \top\ \emptyset\ \top \\
\quad \mathsf{appce} : \forall \alpha_1, \alpha_2, \epsilon_1, \epsilon_2, \omega_1, \omega_2. \\
\qquad (\alpha_2 \subseteq \alpha_1, \epsilon_1 \subseteq \epsilon_2, \omega_2 \subseteq \omega_1) \Rightarrow \\
\qquad (\mathsf{Id}, \mathit{CE}\ \alpha_1\ \epsilon_1\ \omega_1) \rhd \mathit{CE}\ \alpha_2\ \epsilon_2\ \omega_2 \\
\quad \mathsf{mapce} : \forall \alpha_1, \alpha_2, \epsilon_1, \epsilon_2, \omega_1, \omega_2. \\
\qquad (\alpha_2 \subseteq \alpha_1, \epsilon_1 \subseteq \epsilon_2, \omega_2 \subseteq \omega_1) \Rightarrow \\
\qquad (\mathit{CE}\ \alpha_1\ \epsilon_1\ \omega_1, \mathsf{Id}) \rhd \mathit{CE}\ \alpha_2\ \epsilon_2\ \omega_2 \\
\quad \mathsf{bindce} : \forall \alpha_1, \epsilon_1, \omega_1, \alpha_2, \epsilon_2, \omega_2, \epsilon_3. \\
\qquad \epsilon_2 \cup \omega_2 = \omega_1, \epsilon_1 \cup \alpha_1 = \alpha_2, \epsilon_1 \cup \epsilon_2 = \epsilon_3) \Rightarrow \\
\qquad (\mathit{CE}\ \alpha_1\ \epsilon_1\ \omega_1, \mathit{CE}\ \alpha_2\ \epsilon_2\ \omega_2) \rhd \mathit{CE}\ \alpha_1\ \epsilon_3\ \omega_2
\end{array}
$$

Types and theory constraints:

$$
\begin{array}{ll}
\tau & ::= \ldots \mid \{A_1\} \ldots \{A_n\} \mid \emptyset \mid \top \mid \tau_1 \cup \tau_2 \\
\Phi & ::= \tau \subseteq \tau' \mid \tau = \tau' \mid \Phi, \Phi
\end{array}
$$

Auxiliary functions:

$$
\begin{array}{l}
\mathsf{read} :\ \forall \alpha, \omega, r.\ \mathit{intref}\ r \rightarrow \mathit{CE}\ \alpha\ r\ \omega\ \mathit{int} \\
\mathsf{write} : \forall \alpha, \omega, r.\ \mathit{intref}\ r \rightarrow \mathit{int} \rightarrow \mathit{CE}\ \alpha\ r\ \omega\ ()
\end{array}
$$

Fig. 3. Polymonad expressing contextual type and effect systems

$\rho_6, (\mathit{IST}\ H\ H, \rho_6) \rhd \rho_{27}$, and so we must instantiate ρ_6 and ρ_{27} in a way allowed by the signature in Figure 2. While we can legally instantiate $\rho_6 = \mathit{IST}\ L\ l_3$ for any l_3 to solve the second constraint, there is then no possible instantiation of ρ_{27} that can solve the third constraint. After substituting for ρ_6, this constraint has the form $(\mathit{IST}\ H\ H, \mathit{IST}\ L\ l_3) \rhd \rho_{27}$, but this form is unacceptable because the H output of the first computation could be leaked by the L side effect of the second computation. On the other hand, all other instantiations of a_1 and a_2 (e.g., $a_1 = H$ and $a_2 = L$ to correspond to a secret savings account but a public interest rate) do have solutions and do not leak information.

The type given above for add_interest is not its principal type, but an *improved* one. As it turns out, the principal type is basically unreadable, with 19 bind constraints! Fortunately, Section 5 shows how some basic rules can greatly simplify types without reducing their applicability, as has been done above. Moreover, our coherence result (given in the next section) assures that the corresponding changes to the elaborated term do not depend on the particular simplifications: the polymonad laws ensure all such elaborations will have the same semantics.

3.2 Contextual type and effect systems

Wadler and Thiemann [28] showed how a monadic-style construct can be used to model type and effect systems. Polymonads can model standard effect systems, but more interestingly can be used to model *contextual effects* [20], which augment traditional effects with the notion of *prior* and *future* effects of an expression within a broader context. As an example, suppose we are using a language that partitions memory into *regions* $R_1, \ldots, R_n$ and reads/writes of references into region R have effect $\{R\}$. Then in the context of the program read r_1; read r_2, where r_1 points into region R_1 and r_2 points into region R_2, the contextual effect of the subexpression read r_1 would be the triple $[\emptyset; \{R_1\}; \{R_2\}]$: the prior effect is empty, the present effect is $\{R_1\}$, and the future effect is $\{R_2\}$.

Figure 3 models contextual effects as the polymonad $\mathit{CE}\ \alpha\ \epsilon\ \omega\ \tau$, for the type of a computation with prior, present, and future effects α, ϵ, and ω, respectively. Indices are sets of atomic effects $\{A_1\} \ldots \{A_n\}$, with $\emptyset$ the empty effect, $\top$ the effect set that includes all other effects, and $\cup$ the union of two effects. We also

$$\begin{array}{ll}
\mathcal{M} = \mathsf{Id}, A/2 & \textit{Types:} \\
\Sigma \;= \mathsf{bId} : (\mathsf{Id}, \mathsf{Id}) \rhd \mathsf{Id}, & \tau ::= \cdots \mid \mathit{send}\,\tau_1\,\tau_2 \mid \mathit{recv}\,\tau_1\,\tau_2 \mid \mathit{end} \\
\quad\;\; \mathsf{mapA} : \forall p, r.\ (A\ p\ r, \mathsf{Id}) \rhd A\ p\ r, & \\
\quad\;\; \mathsf{appA} : \forall p, r.\ (\mathsf{Id}, A\ p\ r) \rhd A\ p\ r, & \textit{Auxiliary functions:} \\
\quad\;\; \mathsf{unitA} : \forall p.\ (\mathsf{Id}, \mathsf{Id}) \rhd A\ p\ p, & \mathsf{send}\ : \forall a, q.\, a\ \to A\,(\mathit{send}\,q\,a)\,q\,() \\
\quad\;\; \mathsf{bindA} : \forall p, q, r.\ (A\,p\,q,\ A\,q\,r) \rhd A\,p\,r & \mathsf{recv}\ : \forall a, q.\,()\ \to A\,(\mathit{recv}\,q\,a)\,q\,a
\end{array}$$

Fig. 4. Parameterized monad for session types, expressed as a polymonad

introduce theory constraints for subset relations and extensional equality on sets, with the obvious interpretation. As an example source of effects, we include read and write functions on references into region sets r. The bind unitce ascribes a pure computation as having an empty effect and any prior and future effects. The binds appce and mapce express that it is safe to consider an additional effect for the current computation (the ϵs are covariant), and fewer effects for the prior and future computations (αs and ωs are contravariant). Finally, bindce composes two computations such that the future effect of the first computation includes the effect of the second one, provided that the prior effect of the second computation includes the first computation; the effect of the composition includes both effects, while the prior effect is the same as before the first computation, and the future effect is the same as after the second computation.

3.3 Parameterized monads, and session types

Finally, we show $\lambda\underline{\text{PM}}$ can express Atkey's parameterized monad [3], which has been used to encode disciplines like regions [16] and session types [22]. The type constructor $A\ p\ q\ \tau$ can be thought of (informally) as the type of a computation producing a τ-typed result, with a pre-condition p and a post-condition q.

As a concrete example, Figure 4 gives a polymonadic expression of Pucella and Tov's notion of session types [22]. The type $A\,p\,q\,\tau$ represents a computation involved in a two-party session which starts in protocol state p and completes in state q, returning a value of type τ. The key element of the signature Σ is the bindA, which permits composing two computations where the first's post-condition matches the second's precondition. We use the type index $\mathit{send}\,q\,\tau$ to denote a protocol state that requires a message of type τ to be sent, and then transitions to q. Similarly, the type index $\mathit{recv}\,r\,\tau$ denotes the protocol state in which once a message of type τ is received, the protocol transitions to r. We also use the index *end* to denote the protocol end state. The signatures of two primitive operations for sending and receiving messages capture this behavior.

As an example, the following $\lambda\underline{\text{PM}}$ program implements one side of a simple protocol that sends a message x, waits for an integer reply y, and returns y+1.

$$\textbf{let}\ \mathsf{go} = \lambda \mathsf{x}.\ \textbf{let}\ \_ = \mathsf{send}\ \mathsf{x}\ \textbf{in}\ \mathsf{incr}\ (\mathsf{recv}\ ())$$
$$\text{Simplified type: } \forall a, b, q, \rho.\,(A\,(\mathit{send}\ \ a\,b)\,a,\ A\,(\mathit{recv}\ \ q\ \mathit{int})\,q)) \rhd \rho \Rightarrow\ (b \to \rho\ \mathit{int})$$

There are no specific theory constraints for session types: constraints simply arise by unification and are solved as usual when instantiating the final program (e.g., to call go 0).

$$\boxed{P \models P'} \qquad \frac{\forall \pi \in P'.\pi \in P \vee \pi \in \Sigma}{P \models P'}\ \text{(TS-Entail)}$$

$$\boxed{P \models \sigma \geqslant \tau \ \rightsquigarrow \mathsf{f}} \qquad \frac{\theta = [\bar{\tau}/\bar{a}][\bar{m}/\bar{\rho}] \qquad P \models \theta P_1}{P \models (\forall \bar{a}\bar{\rho}.\, P_1 \Rightarrow \tau) \geqslant \theta\tau \ \rightsquigarrow \mathsf{app}(\theta P_1)}\ \text{(TS-Inst)}$$

$$\boxed{P \,|\, \Gamma \vdash v : \tau \ \rightsquigarrow \mathsf{e}} \qquad \frac{v \in \{x, c\} \qquad P \models \Gamma(v) \geqslant \tau \ \rightsquigarrow \mathsf{f}}{P \,|\, \Gamma \vdash v : \tau \ \rightsquigarrow \mathsf{f}\, v}\ \text{(TS-XC)}$$

$$\frac{P \,|\, \Gamma, x{:}\tau_1 \vdash e : m\ \tau_2 \ \rightsquigarrow \mathsf{e}}{P \,|\, \Gamma \vdash \lambda x.e : \tau_1 \to m\ \tau_2 \ \rightsquigarrow \lambda x.\mathsf{e}}\ \text{(TS-Lam)}$$

$$\boxed{P \,|\, \Gamma \vdash e : m\ \tau \ \rightsquigarrow \mathsf{e}} \qquad \frac{P \,|\, \Gamma \vdash v : \tau \ \rightsquigarrow \mathsf{e}}{P, \mathsf{Id} \hookrightarrow m \,|\, \Gamma \vdash v : m\ \tau \ \rightsquigarrow \mathsf{b}_{\mathsf{Id},\mathsf{Id},m}\ \mathsf{e}\ (\lambda x.x)}\ \text{(TS-V)}$$

$$\frac{\begin{array}{c} P_1 \,|\, \Gamma, x{:}\tau \vdash v : \tau \ \rightsquigarrow \mathsf{e}_1 \qquad (\sigma, \mathsf{e}_2) = \mathit{Gen}(\Gamma, P_1 \Rightarrow \tau, \mathsf{e}_1) \\ P \,|\, \Gamma, x{:}\sigma \vdash e : m\ \tau' \ \rightsquigarrow \mathsf{e}_3 \end{array}}{P \,|\, \Gamma \vdash \mathsf{letrec}\ x{=}v\ \mathsf{in}\ e : m\ \tau' \ \rightsquigarrow \mathsf{letrec}\ x{=}\mathsf{e}_2\ \mathsf{in}\ \mathsf{e}_3}\ \text{(TS-Rec)}$$

$$\frac{\begin{array}{c} P_1 \,|\, \Gamma \vdash v : \tau \ \rightsquigarrow \mathsf{e}_1 \qquad (\sigma, \mathsf{e}_2) = \mathit{Gen}(\Gamma, P_1 \Rightarrow \tau, \mathsf{e}_1) \\ P \,|\, \Gamma, x{:}\sigma \vdash e : m\ \tau' \ \rightsquigarrow \mathsf{e}_3 \end{array}}{P \,|\, \Gamma \vdash \mathsf{let}\ x{=}v\ \mathsf{in}\ e : m\ \tau' \ \rightsquigarrow \mathsf{let}\ x{=}\mathsf{e}_2\ \mathsf{in}\ \mathsf{e}_3}\ \text{(TS-Let)}$$

$$\frac{\begin{array}{cc} P \,|\, \Gamma \vdash e_1 : m_1\ \tau_1 \ \rightsquigarrow \mathsf{e}_1 & P \,|\, \Gamma, x{:}\tau_1 \vdash e_2 : m_2\ \tau_2 \ \rightsquigarrow \mathsf{e}_2 \\ e_1 \neq v & P \models (m_1, m_2) \rhd m_3 \end{array}}{P \,|\, \Gamma \vdash \mathsf{let}\ x{=}e_1\ \mathsf{in}\ e_2 : m_3\ \tau_2 \ \rightsquigarrow \mathsf{b}_{m_1,m_2,m_3}\ \mathsf{e}_1\ (\lambda x.\, \mathsf{e}_2)}\ \text{(TS-Do)}$$

$$\frac{\begin{array}{cc} P \,|\, \Gamma \vdash e_1 : m_1\ (\tau_2 \to m_3\ \tau) \ \rightsquigarrow \mathsf{e}_1 & P \,|\, \Gamma \vdash e_2 : m_2\ \tau_2 \ \rightsquigarrow \mathsf{e}_2 \\ P \models (m_1, m_4) \rhd m_5 & P \models (m_2, m_3) \rhd m_4 \end{array}}{P \,|\, \Gamma \vdash e_1\ e_2 : m_5\ \tau \ \rightsquigarrow \mathsf{b}_{m_1,m_4,m_5}\ \mathsf{e}_1\ (\mathsf{b}_{m_2,m_3,m_4}\ \mathsf{e}_2)}\ \text{(TS-App)}$$

$$\frac{\begin{array}{cc} P \,|\, \Gamma \vdash e_1 : m_1\ \mathsf{bool} \ \rightsquigarrow \mathsf{e}_1 & P \,|\, \Gamma \vdash e_2 : m_2\ \tau \ \rightsquigarrow \mathsf{e}_2 \\ P \,|\, \Gamma \vdash e_3 : m_3\ \tau \ \rightsquigarrow \mathsf{e}_3 & P \models m_2 \hookrightarrow m, m_3 \hookrightarrow m, (m_1, m) \rhd m' \end{array}}{\begin{array}{c} P \,|\, \Gamma \vdash \mathsf{if}\ e_1\ \mathsf{then}\ e_2\ \mathsf{else}\ e_3 : m'\ \tau \\ \rightsquigarrow \mathsf{b}_{m_1,m,m'}\ \mathsf{e}_1\ (\lambda b.\, \mathsf{if}\ b\ \mathsf{then}\ \mathsf{b}_{m_2,\mathsf{Id},m}\ \mathsf{e}_2\ (\lambda x.x)\ \mathsf{else}\ \mathsf{b}_{m_3,\mathsf{Id},m}\ \mathsf{e}_3\ (\lambda x.x)) \end{array}}\ \text{(TS-If)}$$

$$\begin{array}{ll} \mathit{Gen}(\Gamma, P \Rightarrow \tau, \mathsf{e}) & = (\forall(\mathsf{ftv}(P \Rightarrow \tau) \setminus \mathsf{ftv}(\Gamma)).P \Rightarrow \tau,\ \mathsf{abs}(P, \mathsf{e})) \\ \mathsf{abs}(((m_1, m_2) \rhd m_3, P), \mathsf{e}) & = \lambda \mathsf{b}_{m_1,m_2,m_3}\,.\, \mathsf{abs}(P, \mathsf{e}) \\ \mathsf{abs}(\cdot, \mathsf{e}) & = \mathsf{e} \\ \mathsf{app}(P, (m_1, m_2) \rhd m_3)) & = \lambda f.\, \mathsf{app}(P)\ (f\ \mathsf{b}_{m_1,m_2,m_3}\ ) \\ \mathsf{app}(P, \cdot) & = \lambda x.\, x \end{array}$$

Fig. 5. Syntax-directed type rules for $\lambda\underline{\text{PM}}$, where Σ is an implicit parameter.

4 Coherent type inference for $\lambda\underline{\mathbf{pm}}$

This section defines our declarative type system for $\lambda\underline{\text{PM}}$ and proves that type inference produces principal types, and that elaborated programs are coherent.

Figure 5 gives a syntax-directed type system, organized into two main judgments. The value-typing judgment $P \,|\, \Gamma \vdash v : \tau \;\leadsto \mathsf{e}$ types a value v in an environment Γ (binding variables x and constants c to type schemes) at the type τ, provided the constraints P are satisfiable. Moreover, it *elaborates* the value v into a lambda term e that explicitly contains binds, lifts, and evidence passing (as shown in Section 3.1). However, note that the elaboration is independent and we can read just the typing rules by igoring the elaborated terms. The expression-typing judgment $P \,|\, \Gamma \vdash e : m\ \tau \;\leadsto \mathsf{e}$ is similar, except that it yields a computation type. Constraint satisfiability $P \models P'$, defined in the figure, states that P' is satisfiable under the hypothesis P if $P' \subseteq P \cup \Sigma$ where we consider $\pi \in \Sigma$ if and only if $\Sigma \vDash \pi \leadsto \mathsf{b}; \cdot$ (for some b).

The rule (TS-XC) types a variable or constant at an instance of its type scheme in the environment. The instance relation for type schemes $P \models \sigma \geq \tau \;\leadsto \mathsf{f}$ is standard: it instantiates the bound variables, and checks that the abstracted constraints are entailed by the hypothesis P. The elaborated f term supplies the instantiated evidence using the app form. The rule (TS-Lam) is straightforward where the bound variable is given a value type and the body a computation type.

The rule (TS-V) allows a value $v : \tau$ to be used as an expression by lifting it to a computation type $m\ \tau$, so long as there exists a morphism (or unit) from the Id functor to m. The elaborated term uses $\mathsf{b}_{\mathsf{Id},\mathsf{Id},m}$ to lift explicitly to monad m. Note that for evidence we make up names ($\mathsf{b}_{\mathsf{Id},\mathsf{Id},m}$) based on the constraint ($\mathsf{Id} \hookrightarrow m$). This simplifies our presentation but an implementation would name each constraint explicitly [15]. We use the name $\mathsf{b}_{m_1,\mathsf{Id},m_2}$ for morphism constraints $m_1 \hookrightarrow m_2$, and use b_{m_1,m_2,m_3} for general bind constraints $(m_1, m_2) \rhd m_3$.

(TS-Rec) types a recursive let-binding by typing the definition v at the same (mono-)type as the letrec-bound variable f. When typing the body e, we generalize the type of f using a standard generalization function $Gen(\Gamma, P \Rightarrow \tau,\ \mathsf{e})$, which closes the type relative to Γ by generalizing over its free type variables. However, in constrast to regular generalization, we return both a generalized type, as well as an elaboration of e that takes all generalized constraints as explicit evidence parameters (as defined by rule abs). (TS-Let) is similar, although somewhat simpler since there is no recursion involved.

(TS-Do) is best understood by looking at its elaboration: since we are in a call-by-value setting, we interpret a **let**-binding as forcing and sequencing two computations using a single bind where e_1 is typed monomorphically.

(TS-App) is similar to (TS-Do), where, again, since we use call-by-value, in the elaboration we sequence the function and its argument using two bind operators, and then apply the function. (TS-If) is also similar, since we sequence the expression e in the guard with the branches. As usual, we require the branches to have the same type. This is achieved by generating morphism constraints, $m_2 \hookrightarrow m$ and $m_3 \hookrightarrow m$ to coerce the type of each branch to a functor m before sequencing it with the guard expression.

$$
\begin{aligned}
&[\![x]\!]^\star &&= x\\
&[\![c]\!]^\star &&= c\\
&[\![\lambda x.e]\!]^\star &&= \lambda x.[\![e]\!]\\
\\
&[\![v]\!] &&= \texttt{ret}\ [\![v]\!]^\star\\
&[\![e_1\ e_2]\!] &&= \mathsf{app}\ [\![e_1]\!]\ [\![e_2]\!]\\
&[\![\mathsf{let}\ x = v\ \mathsf{in}\ e]\!] &&= \mathsf{let}\ x = [\![v]\!]^\star\ \mathsf{in}\ [\![e]\!]\\
&[\![\mathsf{let}\ x = e_1\ \mathsf{in}\ e_2]\!] &&= [\![e_1]\!]\ \texttt{>>=}\ [\![\lambda x.e_2]\!]^\star \qquad (\text{when } e_1 \neq v)\\
&[\![\mathsf{if}\ e_1\ \mathsf{then}\ e_2\ \mathsf{else}\ e_3]\!] &&= \texttt{cond}\ [\![e_1]\!]\ \lambda().[\![e_2]\!]\ \lambda().[\![e_3]\!]\\
&[\![\mathsf{letrec}\ f = v\ \mathsf{in}\ e]\!] &&= \texttt{letrec} f = [\![v]\!]^\star\ \texttt{in}\ [\![e]\!]
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{ret} &&: \forall \alpha\rho.\, (\mathsf{Id} \hookrightarrow \rho) \Rightarrow \alpha \to \rho\ \alpha\\
&\mathsf{do} &&: \forall \alpha\beta\rho_1\rho_2\rho.\, ((\rho_1,\rho_2) \rhd \rho) \Rightarrow \rho_1\ \alpha \to (\alpha \to \rho_2\ \beta) \to \rho\ \beta\\
&\mathsf{app} &&: \forall \alpha\beta\rho_1\rho_2\rho_3\rho_4\rho.\, ((\rho_1,\rho_4) \rhd \rho, (\rho_2,\rho_3) \rhd \rho_4) \Rightarrow \rho_1\ (\alpha \to \rho_3\ \beta) \to \rho_2\ \alpha \to \rho\ \beta\\
&\mathsf{cond} &&: \forall \alpha\rho_1\rho_2\rho_3\rho\rho'.\, (\rho_2 \hookrightarrow \rho, \rho_3 \hookrightarrow \rho, (\rho_1,\rho) \rhd \rho')\\
& &&\quad \Rightarrow \rho_1\ \mathsf{bool} \to (() \to \rho_2\ \alpha) \to (() \to \rho_3\ \alpha) \to \rho'\ \alpha
\end{aligned}
$$

Fig. 6. Type inference for $\lambda_{\underline{\mathrm{PM}}}$ via elaboration to OML

4.1 Principal types

The type rules admit principal types, and there exists an efficient type inference algorithm that finds such types. The way we show this is by a translation of polymonadic terms (and types) to terms (and types) in OML [13] and prove this translation is sound and complete: a polymonadic term is well-typed if and only if its translated OML term has an equivalent type. OML's type inference algorithm is known to enjoy principal types, so a corollary of our translation is that principal types exist for our system too.

We encode terms in our language into OML as shown in Figure 6. We rely on four primitive OML terms that force the typing of the terms to generate the same constraints as our type system does: ret for lifting a pure term, do for typing a do-binding, app for typing an application, and cond for conditionals. Using these primitives, we encode values and expressions of our system into OML.

We write $P \,|\, \Gamma \vdash_{\mathrm{OML}} e : \tau$ for a derivation in the syntax directed inference system of OML (cf. Jones [13], Fig. 4).

Theorem 2 (Encoding to OML is sound and complete).
Soundness: *Whenever* $P \,|\, \Gamma \vdash v : \tau$ *we have* $P \,|\, \Gamma \vdash_{\mathrm{OML}} [\![v]\!]^\star : \tau$. *Similarly, whenever* $P \,|\, \Gamma \vdash e : m\ \tau$ *then we have* $P \,|\, \Gamma \vdash_{\mathrm{OML}} [\![e]\!] : m\ \tau$.
Completeness: *Whenever* $P \,|\, \Gamma \vdash_{\mathrm{OML}} [\![v]\!]^\star : \tau$, *then we have* $P \,|\, \Gamma \vdash v : \tau$. *Similarly, whenever* $P \,|\, \Gamma \vdash_{\mathrm{OML}} [\![e]\!] : m\ \tau$, *then we have* $P \,|\, \Gamma \vdash e : m\ \tau$.

The proof is by straightforward induction on the typing derivation of the term. It is important to note that our system uses the same instantiation and generalization relations as OML which is required for the induction argument. Moreover, the constraint entailment over bind constraints also satisfies the monotonicity, transitivity and closure under substitution properties required by OML. As a corollary of the above properties, our system admits principal types via the general-purpose OML type inference algorithm.

4.2 Ambiguity

Seeing the previous OML translation, one might think we could directly translate our programs into Haskell since Haskell uses OML style type inference. Unfortunately, in practice, Haskell would reject many useful programs. In particular, Haskell rejects as ambiguous any term whose type $\forall\bar{\alpha}.P \Rightarrow \tau$ includes a variable α that occurs free in P but not in τ;[5] we call such type variables *open*. Haskell, in its generality, must reject such terms since the instantiation of an open variable can have operational effect, while at the same time, since the variable does not appear in τ, the instantiation for it can never be uniquely determined by the context in which the term is used. A common example is the term $\mathsf{show}\ .\ \mathsf{read}$ with the type $(\mathsf{Show\ a}, \mathsf{Read\ a}) \Rightarrow \mathsf{String} \rightarrow \mathsf{String}$, where a is open. Depending on the instantiation of a, the term may parse and show integers, or doubles, etc.

Rejecting all types that contain open variables works well for type classes, but it would be unacceptable for $\lambda_{\underline{\text{PM}}}$. Many simple terms have principal types with open variables. For example, the term $\lambda f.\lambda x.f\ x$ has type $\forall ab\rho_1\rho_2\rho_3.\ ((\mathsf{Id},\rho_1) \rhd \rho_2, (\mathsf{Id},\rho_2) \rhd\ \rho_3) \Rightarrow\ (a \rightarrow \rho_1\ b) \rightarrow \alpha \rightarrow \rho_3\ b$ where type variable ρ_2 is open.

A major contribution of this paper is that we show that for our specific bind constraints, we can relax this rule and solve much more aggressively. In particular, by appealing to the polymonadic laws, we can prove that programs with open type variables in bind constraints are indeed unambigiuous. Even if there are many possible instantiations, the semantics of each instantiation is equivalent. This *coherence* result is at the essence of making programming with polymonads practical and the next section treats this in detail.

4.3 Coherence

The main result of this section (Theorem 3) establishes that for a certain class of polymonads, the ambiguity check of OML can be weakened to accept more programs while still ensuring that programs are coherent. Thus, for this class of polymonads, programmers can reliably view our syntax-directed system as a specification without being concerned with the details of how the type inference algorithm is implemented or how programs are elaborated.

The proof of Theorem 3 is a little technical—the following roadmap summarizes the structure of the development.

– We define the class of *principal* polymonads for which unambiguous typing derivations are coherent. All polymonads that we know of are principal.
– Given $P \,|\, \Gamma \vdash e : t \rightsquigarrow \mathsf{e}$ (with $t \in \{\tau, m\ \tau\}$), the predicate $\mathsf{unambiguous}(P, \Gamma, t)$ characterizes when the derivation is unambiguous. This notion requires interpreting P as a graph G_P, and ensuring (roughly) that all open variables in P have non-zero in/out-degree in G_P.
– A *solution* S to a constraint graph with respect to a polymonad $(\mathcal{M}, \Sigma)$ is an assignment of ground polymonad constructors $\mathsf{M} \in \mathcal{M}$ to the variables in

[5] The actual ambiguity rule in Haskell is more involved due to functional dependencies and type families but that does not affect our results.

the graph such that each instantiated constraint is present in Σ. We give an equivalence relation on solutions such that $S_1 \cong S_2$ if they differ only on the assignment to open variables in a manner where the composition of binds still compute the same function according to the polymonad laws.
- Finally, given $P \,|\, \Gamma \vdash e : t \leadsto \mathsf{e}$ and $\mathsf{unambiguous}(P, \Gamma, t)$, we prove that all solutions to P that agree on the free variables of Γ and t are in the same equivalence class.

While Theorem 3 enables our type system to be used in practice, this result is not the most powerful theorem one can imagine. Ideally, one might like a theorem of the form $P \,|\, \Gamma \vdash e : t \leadsto \mathsf{e}$ and $P' \,|\, \Gamma \vdash e : t \leadsto \mathsf{e}'$ implies $\mathsf{e} \cong \mathsf{e}'$, given that both P and P' are satisfiable. However, a result of this form is out of our reach, at present. There are at least two difficulties. First, a coherence result of this form is unknown for qualified type systems in a call-by-value setting. In an unpublished paper, Jones [14] proves a coherence result for OML, but his techique only applies to call-by-name programs. Jones also does not consider reasoning about coherence based on an equational theory for the evidence functions, i.e., the binds in our case. So, proving the ideal coherence theorem would require both generalizing Jones' approach to call-by-value and then extending it with support for equational reasoning about evidence. In the meantime, Theorem 3 provides good assurance and lays the foundation for future work in this direction.

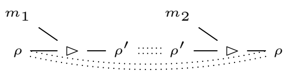
Defining and analyzing principality. We introduce a notion of principal polymonads that corresponds to Tate's principalled productoids. Informally, in a principal polymonad, if there is more than one way to combine pairs of computations in the set F (e.g., $(\mathsf{M}, \mathsf{M}') \rhd \mathsf{M}_1$ and $(\mathsf{M}, \mathsf{M}') \rhd \mathsf{M}_2$), then there must be a "best" way to combine them. This best way is called the principal join of F, and all other ways to combine the functors are related to the principal join by morphisms. All the polymonadic libraries we have encountered so far are principal polymonads. It is worth emphasizing that principality does not correspond to functional dependency—it is perfectly reasonable to combine M and M' in multiple ways, and indeed, for applications like sub-effecting, this expressiveness is important. We only require that there be an ordering among the choices. In the definition below, we take $\downarrow\!\mathcal{M}$ to be set of ground instances of all constructors in $\mathcal{M}$.

Definition 2 (Principal polymonad). *A polymonad $(\mathcal{M}, \Sigma)$ is a* principal polymonad *if and only if for any set $F \subseteq \downarrow\!\mathcal{M}^2$, and any $\{\mathsf{M}_1, \mathsf{M}_2\} \subseteq \downarrow\!\mathcal{M}$ such that $\{(\mathsf{M}, \mathsf{M}') \rhd \mathsf{M}_1 \mid (\mathsf{M}, \mathsf{M}') \in F\} \subseteq \Sigma$ and $\{(\mathsf{M}, \mathsf{M}') \rhd \mathsf{M}_2 \mid (\mathsf{M}, \mathsf{M}') \in F\} \subseteq \Sigma$, then there exists $\mathsf{M} \in \downarrow\!\mathcal{M}$ such that $\{\mathsf{M} \hookrightarrow \mathsf{M}_1, \mathsf{M} \hookrightarrow \mathsf{M}_2\} \subseteq \Sigma$, and $\{(\mathsf{M}, \mathsf{M}') \rhd \mathsf{M} \mid (\mathsf{M}, \mathsf{M}') \in F\} \subseteq \Sigma$. We call M the principal join of F and write it as $\bigsqcup F$*

Definition 3 (Graph-view of a constraint-bag P). *A graph-view $G_P = (V, A, E_\rhd, E_{eq})$ of a constraint-bag P is a graph consisting of a set of vertices V, a vertex assignment $A : V \to m$, a set of directed edges $E_\rhd$, and a set of undirected edges E_{eq}, where:*

- $V = \{\pi.0, \pi.1, \pi.2 \mid \pi \in P\}$, *i.e., each constraint contributes three vertices.*
- $A(\pi.i) = m_i$ *when* $\pi = (m_0, m_1) \rhd m_2$, *for all* $\pi.i \in V$
- $E_{\rhd} = \{(\pi.0, \pi.2), (\pi.1, \pi.2) \mid \pi \in P\}$
- $E_{eq} = \{(v, v') \mid v, v' \in V \ \wedge \ v \neq v' \wedge \exists \rho. \rho = A(v) = A(v')\}$

Notation We use v in this section to stand for a graph vertex, rather than a value in a program. We also make use of a pictorial notation for graph views, distinguishing the two flavors of edges in a graph. Each constraint $\pi \in P$ induces two edges in $E_{\rhd}$. These edges are drawn with solid lines, with a triangle for orientation. Unification constraints arise from correlated variable occurrences in multiple constraints—we depict these with double dotted lines. For example, the pair of constraints $(m_1, \rho) \rhd \rho', (m_2, \rho') \rhd \rho$ contributes four unification edges, two for ρ and two for ρ'. We show its graph view alongside.



Unification constraints reflect the dataflow in a program. Referring back to Figure 5, in a principal derivation using (TS-App), correlated occurrences of unification variables for m_4 in the constraints indicate how the two binds operators compose. The following definition captures this dataflow and shows how to interpret the composition of bind constraints using unification edges as a lambda term (in the expected way).[6]

Definition 4 (Functional view of a flow edge). *Given a constraint graph* $G = (V, A, E_{\rhd}, E_{eq})$, *an edge* $\eta = (\pi.2, \pi'.i) \in E_{eq}$, *where* $i \in \{0, 1\}$ *and* $\pi \neq \pi'$ *is called a* flow edge. *The flow edge* η *has a functional interpretation* $F_G(\eta)$ *defined as follows:*

If $i = 0$, $F_G(\eta) = \lambda(x{:}A(\pi.0)\ a)\ (y{:}a \rightarrow A(\pi.1)\ b)\ (z{:}b \rightarrow A(\pi'.1)\ c).$
$\quad\quad \mathsf{bind}_{A(\pi'.0),A(\pi'.1),A(\pi'.2)}(\mathsf{bind}_{A(\pi.0),A(\pi.1),A(\pi.2)}\ x\ y)\ z$

If $i = 1$, $F_G(\eta) = \lambda(x{:}A(\pi'.0)\ a)\ (y{:}a \rightarrow A(\pi.0)\ b)\ (z{:}b \rightarrow A(\pi.1)\ c).$
$\quad\quad \mathsf{bind}_{A(\pi'.0),A(\pi'.1),A(\pi'.2)}\ x\ (\lambda a.\mathsf{bind}_{A(\pi.0),A(\pi.1),A(\pi.2)}\ (y\ a)\ z)$

We can now define our ambiguity check—a graph is unambiguous if it contains a sub-graph that has no cyclic dataflows, and where open variables only occur as intermediate variables in a sequence of binds.

Definition 5 (Unambiguous constraints). *Given* $G_P = (V, A, E_{\rhd}, E_{eq})$, *the predicate* $\mathsf{unambiguous}(P, \Gamma, t)$ *holds if and only if there exists* $E'_{eq} \subseteq E_{eq}$, *such that in the graph* $G' = (V, A, E_{\rhd}, E'_{eq})$ *all of the following are true.*

1. *For all* $\pi \in P$, *there is no path from* $\pi.2$ *to* $\pi.0$ *or* $\pi.1$.
2. *For all* $v \in V$, *if* $A(v) \in \mathsf{ftv}(P) \setminus \mathsf{ftv}(\Gamma, t)$, *then there exists a flow edge that connects to* v.

We call G' *a* core *of* G_P.

[6] Note, for the purposes of our coherence argument, unification constraints between value-type variables a are irrelevant. Such variables may occur in two kinds of contexts. First, they may constraint some value type in the program, but these do not depend on the solutions to polymonadic constraints. Second, they may constrain some index of a polymonadic constructor; but, as mentioned previously, these indices are phantom and do not influence the semantics of elaborated terms.

Definition 6 (Solution to a constraint graph). *For a polymonadic signature $(\mathcal{M}, \Sigma)$, a solution to a constraint graph $G = (V, A, E_{\rhd}, E_{eq})$, is a vertex assignment $S : V \to \mathcal{M}$ such that all of the following are true.*

1. *For all $v \in V$, if $A(v) \in \mathcal{M}$ then $S(v) = A(v)$*
2. *For all $(v_1, v_2) \in E_{eq}$, $S(v_1) = S(v_2)$.*
3. *For all $\{(\pi.0, \pi.2), (\pi.1, \pi.2)\} \subseteq E_{\rhd}$, $(S(\pi.0), S(\pi.1)) \rhd S(\pi.2) \in \Sigma$.*

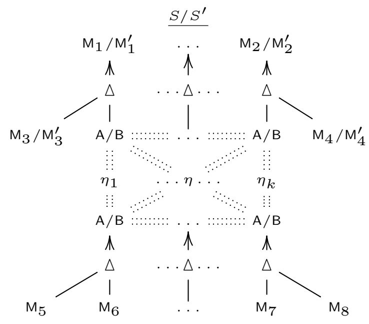
We say that two solutions S_1 and S_2 to G agree on ρ *if for all vertices $v \in V$ such that $A(v) = \rho$, $S_1(v) = S_2(v)$.*

Now we define $\cong_R$, a notion of equivalence of two solutions which captures the idea that the differences in the solutions are only to the internal open variables while not impacting the overall function computed by the binds in a constraint. It is easy to check that $\cong_R$ is an equivalence relation.

Definition 7 (Equivalence of solutions). *Given a polymonad $(\mathcal{M}, \Sigma)$ and constraint graph $G = (V, A, E_{\rhd}, E_{eq})$, two solutions S_1 and S_2 to G are equivalent with respect to a set of variables R (denoted $S_1 \cong_R S_2$) if and only if S_1 and S_2 agree on all $\rho \in R$ and for each vertex $v \in V$ such that $S_1(v) \neq S_2(v)$ for all flow edges η incident on v, $F_{G_1}(\eta) = F_{G_2}(\eta)$, where $G_i = (V, S_i, E_{\rhd}, E_{eq})$.*

Theorem 3 (Coherence). *For all principal polymonads, derivations $P|\Gamma \vdash e : t \rightsquigarrow \mathsf{e}$ such that $\mathsf{unambiguous}(P, \Gamma, t)$, and for any two solutions S and S' to G_P that agree on $R = FTV(\Gamma, t)$, we have $S \cong_R S'$.*

(Sketch; full version in appendix) The main idea is to show that all solutions in the core of G_P are in the same equivalence class (the solutions to the core include S and S'). The proof proceeds by induction on the number of vertices at which S and S' differ. For the main induction step, we take vertices in topological order, considering the least (in the order) set of vertices Q, all related by unification constraints, and whose assignment in S is A and in S' is B, for some $\mathsf{A} \neq \mathsf{B}$. The vertices in Q are shown in the graph alongside, all connected to each other by double dotted lines (unification constraints), and their neighborhood is shown as well. Since vertices are considered in topological order, all the vertices below Q in the graph have the same assignment in S and in S'. We build solutions S_1 and S_1' from S and S' respectively, that instead assign the principal join $\mathsf{J} = \bigsqcup\{(\mathsf{M}_5, \mathsf{M}_6), \ldots, (\mathsf{M}_7, \mathsf{M}_8)\}$ to the vertices in Q, where $S_1 \cong_R S_1'$ by the induction hypothesis. Finaly, we prove $S \cong_R S_1$ and $S' \cong_R S_1'$ by showing that the functional interpretation of each of the flow edges η_i are equal according to the polymonad laws, and conclude $S \cong_R S'$ by transitivity.



$$\text{S-}\Uparrow\ \frac{\pi = (\mathsf{Id},\mathsf{m}) \rhd \rho \ \vee\ \pi = (\mathsf{m},\mathsf{Id}) \rhd \rho \quad \rho \in \bar\rho \quad \mathit{flowsFrom}_{P,P'}\ \rho \neq \{\} \quad \mathit{flowsTo}_{P,P'}\ \rho = \{\}}{P, \pi, P' \xrightarrow{\text{simplify}(\bar\rho)} \rho \mapsto m}
\qquad
\text{S-}\Downarrow\ \frac{\pi = (\mathsf{Id},\rho) \rhd \mathsf{m} \ \vee\ \pi = (\rho,\mathsf{Id}) \rhd \mathsf{m} \quad \rho \in \bar\rho \quad \mathit{flowsFrom}_{P,P'}\ \rho = \{\} \quad \mathit{flowsTo}_{P,P'}\ \rho \neq \{\}}{P, \pi, P' \xrightarrow{\text{simplify}(\bar\rho)} \rho \mapsto m}$$

$$\text{S-}\sqcup\ \frac{F = \mathit{flowsTo}_P\ \rho \quad m \in F \Rightarrow m = \mathsf{M} \quad \text{for some } \mathsf{M}}{P \xrightarrow{\text{simplify}(\bar\rho)} \rho \mapsto \bigsqcup F}
\qquad
\frac{P \xrightarrow{\text{simplify}(\bar\rho)} \theta \quad \theta P \xrightarrow{\text{simplify}(\bar\rho)} \theta'}{P \xrightarrow{\text{simplify}(\bar\rho)} \theta'\theta}
\qquad
\frac{}{P \xrightarrow{\text{simplify}(\bar\rho)} \cdot}$$

$$\text{where}\ \begin{array}{ll} \mathit{flowsTo}_P\ \rho & = \{ (m_1, m_2) \mid (\mathsf{m}_1, \mathsf{m}_2) \rhd \rho \in P \} \\ \mathit{flowsFrom}_P\ \rho & = \{ m \mid \exists m'.\ \pi \in P \ \wedge\ (\pi = (\rho, \mathsf{m}') \rhd \mathsf{m} \ \vee\ \pi = (\mathsf{m}', \rho) \rhd \mathsf{m}) \} \end{array}$$

Fig. 7. Eliminating open variables in constraints

5 Simplification and solving

Before running a program, we must solve the constraints produced during type inference, and apply the appropriate evidence for these constraints in the elaborated program. We also perform *simplification* on constraints prior to generalization to make types easier to read, but without compromising their utility.

A simple syntactic transformation on constraints can make inferred types easier to read. For example, we can hide duplicate constraints, identity morphisms (which are trivially satisfiable), and constraints that are entailed by the signature. More substantially, we can find instantiations for open variables in a constraint set before generalizing a type (and at the top-level, before running a program). To do this, we introduce below a modified version of (TS-Let) (from Figure 5); a similar modification is possible for (TS-Rec).

$$\frac{P_1 \mid \Gamma \vdash v : \tau \rightsquigarrow \mathsf{e}_1 \qquad \bar\rho, \bar a = \mathsf{ftv}(P_1 \Rightarrow \tau) \setminus \mathsf{ftv}(\Gamma) \qquad P_1 \xrightarrow{\text{simplify}(\bar\rho \setminus \mathsf{ftv}(\tau))} \theta \qquad (\sigma, \mathsf{e}_2) = \mathit{Gen}(\Gamma, \theta P_1 \Rightarrow \tau, \mathsf{e}_1) \qquad P \mid \Gamma, x{:}\sigma \vdash e : m\ \tau' \rightsquigarrow \mathsf{e}_3}{P \mid \Gamma \vdash \mathsf{let}\ x{=}v\ \mathsf{in}\ e : m\ \tau' \rightsquigarrow \mathsf{let}\ x{=}\mathsf{e}_2\ \mathsf{in}\ \mathsf{e}_3}$$

This rule employs the judgment $P \xrightarrow{\text{simplify}(\bar\rho)} \theta$, defined in Figure 7, to simplify constraints by eliminating some open variables in P (via the substitution θ) before type generalization. There are three main rules in the judgment (S-$\Uparrow$), (S-$\Downarrow$) and (S-$\sqcup$), while the last two simply take the transitive closure.

Rule (S-$\Uparrow$) solves monad variable ρ with monad m if we have a constraint $\pi = (\mathsf{Id}, \mathsf{m}) \rhd \rho$, where the only edges directed inwards to ρ are from Id and m, although there may be many out-edges from ρ. (The case where $\pi = (\mathsf{m}, \mathsf{Id}) \rhd \rho$ is symmetric.) Such a constraint can always be solved without loss of generality using an identity morphism, which, by the polymonad laws is guaranteed to exist. Moreover, by the closure law, any solution that chooses $\rho = m'$, for some $m' \neq m$ could just as well have chosen $\rho = m$. Thus, this rule does not impact solvability of the costraints. Rule S-$\Downarrow$ follows similar reasoning in the reverse direction. Finally, we the rule (S-$\sqcup$) exploits the properties of a principal polymonad. Here we have a variable ρ such that all its in-edges are from pairs of ground constructors M_i, so we can simply apply the join function to compute a solution for ρ. For a principal polymonad, if such a solution exists, this simplification does not impact solvability of the rest of the constraint graph.

Example. Recall the information flow example we gave in Section 3.1, in Figure 2. Its principal type is the following, which is hardly readable:

$$\begin{array}{l} \forall \bar{\rho}_i, a_1, a_2.P_0 \Rightarrow \mathit{intref}\ a_1 \rightarrow \mathit{intref}\ a_2 \rightarrow \rho_{27}\ () \\ \text{where } P_0 = (\mathsf{Id}, \rho_3) \rhd \rho_2, (\mathsf{Id}, \mathit{IST}\ H\ \mathsf{a}_2) \rhd \rho_3, (\rho_{26}, \mathsf{Id}) \rhd \rho_4, (\mathsf{Id}, \mathsf{Id}) \rhd \rho_4, \\ \qquad (\rho_8, \rho_4) \rhd \rho_6, (\mathsf{Id}, \rho_9) \rhd \rho_8, (\mathsf{Id}, \mathsf{Id}) \rhd \rho_9, (\rho_{11}, \rho_{25}) \rhd \rho_{26}, \\ \qquad (\mathsf{Id}, \rho_{12}) \rhd \rho_{11}, (\mathsf{Id}, \mathit{IST}\ H\ \mathsf{a}_1) \rhd \rho_{12}, (\rho_{17}, \rho_{23}) \rhd \rho_{25}, (\rho_{14}, \rho_{18}) \rhd \rho_{17}, \\ \qquad (\mathsf{Id}, \mathsf{Id}) \rhd \rho_{18}, (\mathsf{Id}, \rho_{15}) \rhd \rho_{14}, (\mathsf{Id}, \mathsf{Id}) \rhd \rho_{15}, (\rho_{20}, \rho_{24}) \rhd \rho_{23}, \\ \qquad (\mathsf{Id}, \mathit{IST}\ \mathsf{a}_1\ L) \rhd \rho_{24}, (\mathsf{Id}, \rho_{21}) \rhd \rho_{20}, (\mathsf{Id}, \mathsf{Id}) \rhd \rho_{21}. \end{array}$$

After applying (S-$\Uparrow$) and (S-$\Downarrow$) several times, and then hiding redundant constraints, we simplify P_0 to P which contains only three constraints. If we had fixed a_1 and a_2 (the labels of the function parameters) to H and L, respectively, we could do even better. The three constraints would be $(\mathit{IST}\ H\ L, \rho_6) \rhd \rho_{27}, (\mathsf{Id}, \mathsf{Id}) \rhd \rho_6, (\mathit{IST}\ H\ H, \mathit{IST}\ H\ L) \rhd \rho_6$. Then, applying (S-$\sqcup$) to ρ_6 we would get $\rho_6 \mapsto \mathit{IST}\ H\ H$, which when applied to the other constraints leaves only $(\mathit{IST}\ H\ L, \mathit{IST}\ H\ H) \rhd \rho_{27}$, which cannot be simplified further, since ρ_{27} appears in the result type.

Pleasingly, this process yields a simpler type that can be used in the same contexts as the original principal type, so we are not compromising the generality of the code by simplifying its type.

Lemma 1 (Simplification improves types). *For a principal polymonad, given σ and σ' where σ is $\forall \bar{a}\bar{\rho}.P \Rightarrow \tau$ and σ' is an* improvement *of σ, having form $\forall \bar{a'}\bar{\rho'}.\theta P \Rightarrow \tau$ where $P \xrightarrow{\mathit{simplify}(\bar{\rho})} \theta$ and $\bar{a'}\bar{\rho'} = (\bar{a}\bar{\rho}) - \mathit{dom}(\theta)$. Then for all $P'', \Gamma, x, e, m, \tau$, if $P'' \,|\, \Gamma, x{:}\sigma \vdash e : m\,\tau$ such that $\models P''$ then there exists some P''' such that $P''' \,|\, \Gamma, x{:}\sigma' \vdash e : m\,\tau$ and $\models P'''$.*

Note that our $\xrightarrow{\text{simplify}(\bar{\rho})}$ relation is non-deterministic in the way it picks constraints to analyze, and also in the order in which rules are applied. In practice, for an acyclic constraint graph, one could consider nodes in the graph in topological order and, say, apply (S-$\sqcup$) first, since, if it succeeds, it eliminates a variable. For principal polymonads and acyclic constraint graphs, this process would always terminate.

However, if unification constraints induce cycles in the constraint graph, simply computing joins as solutions to internal variables may not work. This should not come as a surprise. In general, finding solutions to arbitrary polymonadic constraints is undecidable, since, in the limit, they can be used to encode the correctness of programs with general recursion. Nevertheless, simple heuristics such as unrolling cycles in the constraint graph a few times may provide good mileage, as would the use of domain-specific solvers for particular polymonads, and such approaches are justified by our coherence proof.

6 Related work and conclusions

This paper has presented *polymonads*, a generalization of monads and morphisms, which, by virtue of their relationship to Tate's *productoids*, are extremely powerful, subsuming monads, parameterized monads, and several other interesting constructions. Thanks to supporting algorithms for (principal) type inference, (provably coherent) elaboration, and (generality-preserving) simplification (none of which Tate considers), this power comes with strong supports for the programmer. Like monads before them, we believe polymonads can become a useful and important element in the functional programmer's toolkit.

Constructions resembling polymonads have already begun to creep into languages like Haskell. Notably, Kmett's `Control.Monad.Parameterized` Haskell package [17] provides a type class for bind-like operators that have a signature resembling our $(m_1, m_2) \rhd m_3$. One key limitation is that Kmett's binds must be *functionally dependent*: m_3 must be a function of m_1 and m_2. As such, it is not possible to program morphisms between different constructors, i.e., the pair of binds $(m_1, \mathsf{Id}) \rhd m_2$ and $(m_1, \mathsf{Id}) \rhd m_3$ would be forbidden, so there would be no way to convert from m_1 to m_2 and from m_1 to m_3 in the same program. Kmett also does not permit polymorphic units—he requires units into Id, which may later be lifted. But this only works for first-order code before running afoul of Haskell's ambiguity restriction. Polymonads do not have either limitation. Kmett does not discuss laws that should govern the proper use of non-uniform binds. As such, our work provides the formal basis to design and reason about libraries that functional programmers have already begun developing.

While polymonads subsume a wide range of prior monad-like constructions, and indeed can express any system of *producer effects* [26], as might be expected, other researchers have explored generalizing monadic effects along other dimensions that are incomparable to polymonads. For example, Altenkirch et al. [2] consider *relative monads* that are not endofunctors; each polymonad constructor must be an endofunctor. Uustalu and Vene [27] suggest structuring computations comonadically, particularly to work with context-dependent computations. This suggests a loose connection with our encoding of contextual effects as a polymonad, and raises the possibility of a "copolymonad", something we leave for the future. Still other generalizations include reasoning about effects equationally using Lawvere theories [21] or with arrows [11]—while each of these generalize monadic constructions, they appear incomparable in expressiveness to polymonads. A common framework to unify all these treatments of effects remains an active area of research—polymonads are a useful addition to the discourse, covering at least one large area of the vast design space.

References

1. M. Abadi, A. Banerjee, N. Heintze, and J. Riecke. A core calculus of dependency. In *POPL*, 1999.
2. T. Altenkirch, J. Chapman, and T. Uustalu. Monads need not be endofunctors. In *Proceedings of the 13th international conference on Foundations of Software Science and Computational Structures*, FOSSACS'10, 2010.

3. R. Atkey. Parameterised notions of computation. *Journal of Functional Programming*, 19(3 & 4):335–376, 2009.
4. G. Cooper and S. Krishnamurthi. Embedding dynamic dataflow in a call-by-value language. In *ESOP*, 2006.
5. K. Crary, A. Kliger, and F. Pfenning. A monadic analysis of information flow security with mutable state. *J. Funct. Program.*, 15(02):249–291, 2005.
6. D. Devriese and F. Piessens. Information flow enforcement in monadic libraries. In *TLDI*, 2011.
7. C. Elliott and P. Hudak. Functional reactive animation. In *ICFP*, 1997.
8. A. Filinski. Representing layered monads. In *POPL*, pages 175–188, 1999.
9. J.-C. Filliâtre. A theory of monads parameterized by effects, 1999.
10. J. Goguen and J. Meseguer. Security policy and security models. In *Symposium on Security and Privacy*, 1982.
11. J. Hughes. Generalising monads to arrows. *Sci. Comput. Program.*, 37(1-3), May 2000.
12. G. Hutton and E. Meijer. Monadic Parsing in Haskell. *J. Funct. Program.*, 8(4):437–444, 1998.
13. M. P. Jones. A theory of qualified types. In *ESOP*, 1992.
14. M. P. Jones. Coherence for qualified types. Technical Report YALEU/DCS/RR-989, Yale University, Sept. 1993.
15. M. P. Jones. Simplifying and Improving Qualified Types. Technical Report YALEU/DCS/RR-1040, Yale University, June 1994.
16. O. Kiselyov and C. Shan. Lightweight monadic regions. In *Haskell Symposium*, 2008.
17. E. Kmett. `Control.Monad.Parameterized` package. On `Hackage` repository, 2012.
18. P. Li and S. Zdancewic. Encoding information flow in Haskell. In *CSFW*, 2006.
19. E. Moggi. Computational lambda-calculus and monads. In *LICS*, 1989.
20. I. Neamtiu, M. Hicks, J. S. Foster, and P. Pratikakis. Contextual effects for version-consistent dynamic software updating and safe concurrent programming. In *POPL*, 2008.
21. G. D. Plotkin and J. Power. Semantics for algebraic operations. *Electr. Notes Theor. Comput. Sci.*, 45, 2001.
22. R. Pucella and J. Tov. Haskell session types with (almost) no class. In *Haskell Symposium*, 2008.
23. N. Ramsey and A. Pfeffer. Stochastic lambda calculus and monads of probability distributions. In *POPL*, 2002.
24. A. Russo, K. Claessen, and J. Hughes. A library for light-weight information-flow security in Haskell. In *Haskell*, 2008.
25. N. Swamy, N. Guts, D. Leijen, and M. Hicks. Lightweight monadic programming in ML. In *ICFP*, 2011.
26. R. Tate. The sequential semantics of producer effect systems. In *POPL*, 2013.
27. T. Uustalu and V. Vene. Comonadic notions of computation. *Electron. Notes Theor. Comput. Sci.*, 203(5), June 2008.
28. P. Wadler and P. Thiemann. The marriage of effects and monads. *ACM Trans. Comput. Logic*, 4:1–32, 2003.

Appendix

A Polymonads are productoids and vice versa

Given a polymonad $(\mathcal{M}, \Sigma)$, we can construct a 4-tuple $(\mathcal{M}, U, L, B)$ as follows:

(Units) $U = \{(\lambda x.\mathsf{bind}\ x\ (\lambda y.y))\colon \mathsf{a} \to \mathsf{M\ a} \mid \mathsf{bind}\colon (\mathsf{Id}, \mathsf{Id}) \rhd \mathsf{M} \in \Sigma\}$,
(Lifts) $L = \{(\lambda x.\mathsf{bind}\ x\ (\lambda y.y))\colon \mathsf{M\ a} \to \mathsf{N\ a} \mid \mathsf{bind}\colon \mathsf{M} \hookrightarrow \mathsf{N} \in \Sigma\}$,
(Binds) The set $B = \Sigma - \{\mathsf{bind} \mid \mathsf{bind}\colon (\mathsf{Id}, \mathsf{Id}) \rhd \mathsf{M}$ or $\mathsf{bind}\colon (\mathsf{M}, \mathsf{Id}) \rhd \mathsf{N} \in \Sigma\}$.

It is fairly easy to show that the above structure satisfies generalizations of the familiar laws for monads and monad morphisms.

Theorem 4. *Given a polymonad $(\mathcal{M}, \Sigma)$, the induced 4-tuple $(\mathcal{M}, U, L, B)$ satisfies the following properties.*

(Left unit) $\forall \mathsf{unit} \in U, \mathsf{bind} \in B.$ *if* $\mathsf{unit}\colon \forall \mathsf{a}.\ \mathsf{a} \to \mathsf{M\ a}$ *and* $\mathsf{bind}\colon(\mathsf{M}, \mathsf{N}) \rhd \mathsf{N}$ *then* $\mathsf{bind}\ (\mathsf{unit}\ e)\ f = f(e)$ *where* $e\colon \tau$ *and* $f\colon \tau \to \mathsf{N}\ \tau'$.
(Right unit) $\forall \mathsf{unit} \in U, \mathsf{bind} \in B.$ *if* $\mathsf{unit}\colon \forall \mathsf{a}.\ \mathsf{a} \to \mathsf{N\ a}$ *and* $\mathsf{bind}\colon(\mathsf{M}, \mathsf{N}) \rhd \mathsf{M}$ *then* $\mathsf{bind}\ m\ (\mathsf{unit}) = m$ *where* $m\colon \mathsf{M}\ \tau$.
(Associativity) $\forall \mathsf{bind}_1, \mathsf{bind}_2, \mathsf{bind}_3, \mathsf{bind}_4 \in B.$ *if* $\mathsf{bind}_1 : (\mathsf{M}, \mathsf{N}) \rhd \mathsf{P}$, $\mathsf{bind}_2 : (\mathsf{P}, \mathsf{R}) \rhd \mathsf{T}$, $\mathsf{bind}_3 : (\mathsf{M}, \mathsf{S}) \rhd \mathsf{T}$, *and* $\mathsf{bind}_4 : (\mathsf{N}, \mathsf{R}) \rhd \mathsf{S}$ *then* $\mathsf{bind}_2\ (\mathsf{bind}_1\ m\ f)\ g = \mathsf{bind}_3\ m\ (\lambda x.\mathsf{bind}_4\ (f\ x)\ g)$ *where* $m\colon \mathsf{M}\ \tau$, $f\colon \tau \to \mathsf{N}\ \tau'$ *and* $g\colon \tau' \to \mathsf{R}\ \tau''$
(Morphism 1) $\forall \mathsf{unit}_1, \mathsf{unit}_2 \in U, \mathsf{lift} \in L.$ *if* $\mathsf{unit}_1\colon \forall \mathsf{a}.\ \mathsf{a} \to \mathsf{M\ a}$, $\mathsf{unit}_2\colon \forall \mathsf{a}.\ \mathsf{a} \to \mathsf{N\ a}$ *and* $\mathsf{lift}\colon \forall \mathsf{a}.\ \mathsf{M\ a} \to \mathsf{N\ a}$ *then* $\mathsf{lift}\ (\mathsf{unit}_1\ e) = \mathsf{unit}_2\ e$ *where* $e\colon \tau$.
(Morphism 2) $\forall \mathsf{bind}_1, \mathsf{bind}_2 \in B, \mathsf{lift}_1, \mathsf{lift}_2, \mathsf{lift}_3 \in L.$ *if* $\mathsf{bind}_1 : (\mathsf{M}, \mathsf{P}) \rhd \mathsf{S}$, $\mathsf{bind}_2 : (\mathsf{N}, \mathsf{Q}) \rhd \mathsf{T}$, $\mathsf{lift}_1\colon \forall \mathsf{a}.\ \mathsf{M\ a} \to \mathsf{N\ a}$, $\mathsf{lift}_2\colon \forall \mathsf{a}.\ \mathsf{P\ a} \to \mathsf{Q\ a}$ *and* $\mathsf{lift}_3\colon \forall \mathsf{a}.\ \mathsf{S\ a} \to \mathsf{T\ a}$ *then* $\mathsf{lift}_3\ (\mathsf{bind}_1\ m\ f) = \mathsf{bind}_2\ (\mathsf{lift}_1\ m)\ (\lambda x.\mathsf{lift}_2\ (f\ x))$ *where* $m\colon \mathsf{M}\ \tau$ *and* $f\colon \tau \to \mathsf{P}\ \tau'$.

Now we show how this definition can be used to relate polymonads to Tate's *productoids* [26]. The definition of a productoid is driven by an underlying algebraic structure: the effectoid [26, Theorem 1]. An effectoid $(E, U, \leq, \mapsto)$ is a set E, with an identified subset $U \subseteq E$ and relations $\leq \subseteq E \times E$ and $(\_;\_) \mapsto \_ \subseteq E \times E \times E$, that satisfies six monoid-like conditions. It is possible to show that a polymonad directly induces an effectoid structure and hence a productoid.

Lemma 2. *Given a polymonad $(\mathcal{M}, U, L, B)$ we can define an effectoid $(E, U, \leq, (\_;\_) \mapsto \_)$ as follows.*

$$E = \mathcal{M} \qquad\qquad U = \{\mathsf{M} \mid \mathsf{unit}\colon \mathsf{a} \to \mathsf{M\ a} \in U\}$$
$$\leq\ = \{(\mathsf{M}, \mathsf{N}) \mid \mathsf{lift}\colon \mathsf{M\ a} \to \mathsf{N\ a} \in L\} \qquad (\_;\_) \mapsto \_\ = \{(\mathsf{M}, \mathsf{N}, \mathsf{P}) \mid (\mathsf{M}, \mathsf{N}) \rhd \mathsf{P} \in B\}$$

Lemma 3. *Every polymonad gives rise to a productoid.*

Proof. We have shown that a polymonad gives rise to an effectoid. Given an effectoid $(E, U, \leq, (\_;\_) \mapsto \_)$ a productoid is defined as a collection of functors indexed by the collection E, and three collections of natural transformations indexed by the three relations. These functors and natural transformations are required to satisfy five addition properties [26, Theorem 2]. The five properties are the five properties of Theorem 4, so the proof is immediate.

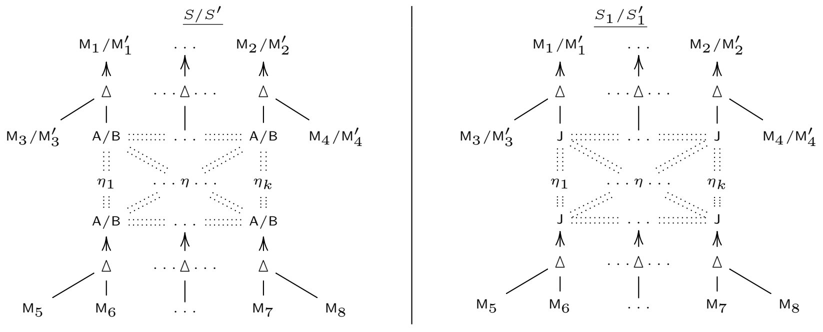


Fig. 8. Constraint graphs used to illustrate the proof of coherence (Theorem 3)

Interestingly, we can identify conditions where the opposite direction also holds.

Lemma 4. *A productoid* $(\mathbf{C}, \{F_e \colon \mathbf{C} \to \mathbf{C}\}_{e \in E}, \{\eta \colon 1 \Rightarrow F_e\}_{e \in U}, \{\mu \colon F_{e_1} \circ F_{e_2} \Rightarrow F_{e_3}\}_{(e_1;e_2) \mapsto e_3}, \{\sigma \colon F_{e_1} \Rightarrow F_{e_2}\}_{e_1 \leq e_2})$ *that in addition satisfies the following conditions gives rise to a polymonad.*

1. $\mathsf{Id} \in E$ *and* $F_{\mathsf{Id}} = 1$

2. $\mathsf{Id} \in U$

3. For all $e \in E$, $(e;\mathsf{Id}) \mapsto e$

4. For all $e \in E, \mu \colon F_e \circ 1 \Rightarrow F_e = \mathsf{Id}$

5. $(e_1;e_2) \mapsto e \ \wedge \ e_1' \leq e_1 \ \Rightarrow \ (e_1';e_2) \mapsto e$

6. $(e_1;e_2) \mapsto e \ \wedge \ e_2' \leq e_2 \ \Rightarrow \ (e_1;e_2') \mapsto e$

These additional conditions are fairly mild: (1)-(4) simply ensure that the Id element is interpreted as the identity functor. Conditions (5)-(6) are also quite straightforward; certainly if the category is cartesian closed then the extra natural transformations are always defined.

B Coherence of solutions

Lemma 5 (Solutions to a core). *For a polymonad* $(\mathcal{M}, \Sigma)$*, and a constraint graph* G *with a core* G'*, the set of all solutions* $\mathcal{S}'$ *to* G' *includes all the solutions* $\mathcal{S}$ *of* G*.*

Proof. (Sketch) This is easy to see, since G' *differs from* G *only in that it includes fewer unification constraints. So, all solutions to* G *are also solutions to* G'*.*

Theorem 5 (Coherence). *For all principal polymonads, derivations* $P|\Gamma \vdash e : t \rightsquigarrow \mathsf{e}$ *such that* $\mathsf{unambiguous}(P, \Gamma, t)$*, and for any two solutions* S *and* S' *to* G_P *that agree on* $R = FTV(\Gamma, t)$*, we have* $S \cong_R S'$*.*

Proof. We consider the set $\mathcal{S}$ *of all solutions to the core of* G_P *that agree on* $FTV(\Gamma, t)$*, and prove that all these solutions are in the same equivalence class. By Lemma 5,* $\{S, S'\} \subseteq \mathcal{S}$*, establishing our goal.*

Let $G = (V, A, E_{\rhd}, E_{eq})$ *be a core of* G_P *and let* S *and* S' *be arbitrary elements of* $\mathcal{S}$*.* S *and* S' *may only differ on the open variables of* P*. Since* G *is unambiguous, the nodes associated with these variables all have non-zero in- and out-degree. Let* $U_{S,S'} = \{v \mid v \in V \ \wedge \ S(v) \neq S'(v)\}$*; the proof proceeds by induction on the the size of* U*.*

Base case $|U_{S,S'} \models 0$***:*** *Trivial, since we have $S(v) = S'(v)$, for all v.*
Induction step $|U_{S,S'} = i|$***:*** *From the induction hypothesis: All solutions S_1 and S'_1 such that $|U_{S_1,S'_1}| < i$, we have $S_1 \cong S'_1$.*

Topologically sort G, such that all vertices in the same connected component following edges in E_{eq} have the same index, and each vertex v is assigned an index greater than the index of all vertices v' such that (v, v') is an edge in $E_{\rhd}$. That is, "leaf" nodes have the highest indices.

Pick a vertex v with the maximal index, such that $S(v) = \mathsf{A}$ and $S'(v) = \mathsf{B}$, for $\mathsf{A} \neq \mathsf{B}$, and let I be the set of vertices reachable from v via unification edges. Since both S and S' are solutions, there must exist an open variable ρ such that $A(v) = \rho$, and since G is a core, there must be some non-empty set of flow edges incident on v.

Thus, the neighborhood of v in the graphs G, under assignment S and S' has a shape as shown in graph at left in Figure 8. All the nodes in I are shown connected by double dotted lines—they each have assignment A/B in S/S'. Since all the nodes in I have an index greater than the index of any variable that differs among S and S', all their immediate predecessors have identical assignments in the two solutions (i.e,. $\mathsf{M}_5, \ldots, \mathsf{M}_8$). However, the other assignments may defer, (e.g., the top-left node could be assigned M_1 in S and M'_1 in S', etc.) Each flow-edge $\{\eta_1, \ldots, \eta_k\}$ incident upon one of the nodes with the same index as v is also labeled.

Now, since we have a principal polymonad, there exists a principal join of $\{(\mathsf{M}_5, \mathsf{M}_6), \ldots, (\mathsf{M}_7, \mathsf{M}_8)\}$—call it J. Consider the assigment S_1 (resp. S'_1) that differs from S_1 (resp. S'_1) only by assigning J to each vertex in I instead of A (resp. B).

We first show that S_1 (resp. S'_1) is a solution and that $S \cong S_1$ (resp. $S' \cong S'_1$). Then, we note that since S_1 and S'_1 agree on all the vertices in I, $|U_{S_1,S'_1}| < i$, so we apply the induction hypothesis to show that $S_1 \cong S'_1$ and conclude with transitivity of $\cong$.

To show that S_1 (resp S'_1) is a solution, since J is a join of $\mathsf{M}_5, \mathsf{M}_6, \ldots$, then $\{(\mathsf{M}_5, \mathsf{M}_6) \rhd \mathsf{J}, \ldots, (\mathsf{M}_7, \mathsf{M}_8) \rhd \mathsf{J}\}$ all exist, as well as $\mathsf{J} \hookrightarrow \mathsf{A}$ (resp. $\mathsf{J} \hookrightarrow \mathsf{B}$). By the Closure property, for every $(\mathsf{M}, \mathsf{A}) \rhd \mathsf{M}'$ (resp. B) there also exists $(\mathsf{M}, \mathsf{J}) \rhd \mathsf{M}'$. Thus, the assignment of J to I is valid for a solution.

To show that $S \cong S_1$ (resp. $S' \cong S'_1$), we have to show that $F_S(\eta_i) = F_{S_1}(\eta_i)$ (resp. $F_{S'}(\eta_i) = F_{S'_1}(\eta_i)$), for all i. Taking η_k as a representative case (the other cases are similar), we need to show the identity below, which is an immediate corollary of Associativity 1 and 2 (resp. for $\mathsf{B}, \mathsf{M}'_4, \mathsf{M}'_2$).

$$\mathsf{bind}_{\mathsf{A},\mathsf{M}_4,\mathsf{M}_2}(\mathsf{bind}_{\mathsf{M}_7,\mathsf{M}_8,\mathsf{A}}\ x\ y)\ z = \mathsf{bind}_{\mathsf{J},\mathsf{M}_4,\mathsf{M}_2}(\mathsf{bind}_{\mathsf{M}_7,\mathsf{M}_8,\mathsf{J}}\ x\ y)\ z$$