First-class subtypes

Jeremy Yallop  Stephen Dolan
University of Cambridge
{jeremy.yallop,stephen.dolan}@cl.cam.ac.uk

1. Introduction

One purpose of the ML module system is to hide equalities between types exposed in an interface and types used in the implementation [10]. Generalized algebraic data types (GADTs) [13], a more recent addition, make these type equalities first class. GADTs attach first class equalities to data; equalities may be hidden by polymorphism or modular abstraction, and later revealed by scrutinising the data.

While GADTs are frequently useful, type equality is sometimes too strong a property. For example, here is a function that prints arrays by calling the `print` method of each element:

```
let print_array = Array.iter (fun o → print o#name)
```

To call `name`, `print_array` does not need to know the full element type: it needs only to know that there is a method `name` returning `string`. OCaml gives `print_array` a row type, indicating that the element type may have other methods:

```
val print_array : <name: string; ..> array → unit
```

But rows are sometimes too inflexible. Given two arrays `a`, `b`, of different element types, unification will fail:

```
List.iter print_array [a; b]
```

This note introduces an interface based on a type constructor `sub` with a coercion operator `>:` that supports passing information around a program, making it possible to combine iterations over arrays whose element types belong to the same subtyping hierarchy.

```
type 'a arr = Arr : 'x array * ('x, 'a) sub

let aiter f (Arr (a,sub)) = Array.iter (fun s → f (s >: sub)) a

List.iter print_array [Arr (a,refl); Arr (b,refl)]
```

2. First-class subtypes defined

Subtypes ala Liskov & Wing The first ingredient in a representation of subtyping proofs is a definition of subtyping. Here is Liskov and Wing’s characterization [11]:

Let $\phi(x)$ be a property provable about objects $x$ of type $T$. Then $\phi(y)$ should be true for objects $y$ of type $S$ where $S$ is a subtype of $T$.

For instance, properties of a record type $r$ should also hold for a widening of $r$, since the extra fields can be ignored.

Subtypes ala Curry & Howard The Curry-Howard correspondence turns Liskov and Wing’s characterization of subtyping into an executable program.

With a propositions-as-types perspective [15], a property provable about objects of type $T$ is represented as a type $\phi(T)$ involving $T$, and a proof of that property is a value $\phi(S)$ of that type. Liskov and Wing’s proposition that $S$ is a subtype of $T$ corresponds to the following (poly)type:

$$\forall \phi. \phi(T) \to \phi(S)$$

Two points deserve note. First, the characterization involves properties of all objects of a particular type, not properties


\[\text{module type POS = sig type } +'a t \text{ end}\]
\[\text{module type NEG = sig type } -'a t \text{ end}\]
\[\text{module Id = struct type } 'a t = 'a t \text{ end}\]
\[\text{module Compose}_+ (F:POS)(G:POS) = struct type } 'a t = 'a F.t G.t \text{ end}\]
\[\text{module Compose}_- (F:NEG)(G:NEG) = struct type } 'a t = 'a F.t G.t \text{ end}\]

\text{Fig. 1. Positive and negative contexts}

\text{val refl : }('a, 'a) sub
\text{val lift : } (F:POS) \to (a, 'b) sub \to ('a F.t, 'b F.t) sub
\text{val lift : } ('a, 'b) sub \to 'b

\text{Fig. 2. First-class subtypes: minimal interface}

of individual objects, which would need dependent types. Second, a “property about objects” is a context that consumes an object; $\phi$ therefore ranges over negative contexts.

Contexts and variance Fig. 1 defines OCaml signatures, POS and NEG, of positive and negative type contexts. The - indicates that the parameter can only appear in negative (contravariant) positions in instantiations of the signature. The Id module and Compose functors represent the identity context and the composition of two contexts. Each composition of variance in the argument contexts requires a separate Compose (but see §3 for a generalization).

Encoding subtypes Fig. 2 defines an interface to subtype witnesses. A value of type $(s, t)$ sub serves as evidence that $s$ is a subtype of $t$. There are two ways to construct such evidence. First, `refl` represents the fact that every type is a subtype of itself. Second, `lift` represents the fact that subtyping lifts through covariant contexts, which are passed as implicit arguments [17]. The single destructor, `>:`, which mimics OCaml’s built-in coercion operator `->`, supports converting a value of type $s$ to a supertype $t$.

This small interface suffices as a basis for many useful subtyping-related functions. For example, the transitivity of subtyping is represented by a function of the following type:

```
val trans : ('a, 'b) sub → ('b, 'c) sub → ('a, 'c) sub
```

and may be defined as follows:

```
let trans (type a b) (x : (a, b) sub) y =
  let module M = struct type +'c t = (a,'c) sub end
  in x >: lift (M y)
```

and other operations, such as a function to lift through negative contexts, can be defined similarly.

Using the variance of sub, `refl` can be used to define a witness for any subtyping fact in the environment. For example, in OCaml the object type `<m:int >`, with one method $n$, is a subtype of the type $< x$ of objects with no methods. This fact can be turned into a sub value by coercing `refl`, either by lowering the contravariant parameter:

```
(refl : (< < > > ) sub ) sub (m:int) + > sub
```

or by raising the covariant parameter:

```
(refl : (m:int) sub ) sub : (m:int) + > sub
```

The resulting value can be passed freely through abstraction boundaries that conceal the types involved, eventually being used to coerce a value of type `<m:int>` to its supertype `< >`.

\[\text{Fig. 2. First-class subtypes: minimal interface}\]
The generality of the interface in Fig. 2 places constraints on the implementation. Most notably, since lift can transport subtyping evidence through any positive context, coercion must pass values through unexamined. For example, lift might be used to build a value of type (a list, b list) from a value of type (a, b):

```
Figure 3. First-class subtypes via negative contexts
```

```
let module N = struct type t = 'a → b end in
f (N) id x
```

but applying f cannot involve list traversal, since the subtyping interface says nothing about list structure. A polymorphic interface thus ensures an efficient implementation.

**Three implementations of subtyping** Several implementations of Fig. 2 are possible. Fig. 3 gives an implementation based on negative contexts that directly follows Liskov & Wing’s definition. A value of type (s, t) sub is a proof that t can be replaced with s in any negative context; operationally it must be the identity, as discussed above, and so the two constructors lift and refl both correspond to the identity function. Fig. 4 gives a similar but simpler implementation, based on positive contexts. Fig. 5 takes an alternative view, based on initiality: (a, 't) sub is the smallest contra/co-variant binary type constructor equipped with an inhabitant refl: (a, 't) sub, from which there is a mapping into any other such type constructor with refl. Despite the different starting points, the three implementations are interdefinable.

All three implementations use the modular implicit extension to OCaml — not for implicit instantiation of arguments, but because modular implicit supports higher-kind quantification with propagation of variance information. Other approaches to higher-kind polymorphism could perhaps be used instead.

**From subtyping to equality** The variance annotations (+, -) in Figs. 1–5, constrain the instantiation of each type constructor. Without these annotations each representation of subtyping collapses to a representation of equality for which additional properties such as symmetry become derivable. For example, stripping the variance annotations turns Fig. 1 into the standard Leibniz encoding [6, 2, 5, 16, 18], and Fig. 5 into a Church encoding of the equality GADT [1].

### 3. First-class subtypes: further examples

Some of OCaml’s built-in type constructors, such as ref and array, are invariant, and so interact poorly with subtyping. The situation may be improved by decomposing each invariant type parameter into a co/contra-variant pair [4], e.g. writing type (+'r,'v) ref rather than type 'a ref. With this decomposition, abstraction and first-class subtypes may be combined to selectively expose capabilities to different parts of a program: a function may be given the ability to write to a reference, write at a particular subtype, or not write at all (and similarly for read). Other types of capability (e.g. for file descriptors) can be written similarly.

A second class of examples arises from selective abstraction, where an abstract type comes with a proof of a property about that type. For example, here is a module that exports a type t along with a proof that t is a subtype of int:

```
module N: DIAG = sig type t (* t is a subtype of int *)
val refl : (t,int) sub
```

Outside the module, values of type t can be coerced to int, but not vice versa. This approach supports a style similar to refinement types, in which abstraction boundaries distinguish values of a type for which some additional predicate has been shown to hold. OCaml’s private types [9] provide direct language support for this feature, but first-class subtypes allow more flexibility: for example, they allow some of the methods of an object type to be hidden from the exposed interface, and also support the dual of private types (called invisible types [12]), and zero cost-coercions [3], where coercions in both directions are available, but actual type equality is not exposed.

Dual to abstraction, combining first-class subtypes with OCaml’s first-class polymorphism encodes bounded quantification. For example, the type ∀ α ≤ t. α → t might be written like this: type t = { f : 'a → t; α → t }.

Finally, first-class subtypes can express proofs of variance. For example, the covariance of list might be represented by a value of type ('a,'b) sub → ('a,'b) list. Subtyping over proofs of variance, we might build a Compose functor that generalizes Compose+ and Compose- (Fig. 1).

### 4. Limitations and further work

The encodings given here are useful for exploratory work, for demonstrating soundness, and for showcasing OCaml’s expressivity. However, direct language support would make first-class subtypes more usable. Scherer and Rényi [12] discuss design issues and related work (e.g. [8, 14]).

The encodings suffer from some awkwardness, since contexts must be applied explicitly, unlike the equalities revealed by pattern matching with GADTs, which the type checker applies implicitly. With language support for sub-type witnesses, coercions would still be explicit, but constraints in scope could be implicitly lifted through contexts.

First-class subtypes might be added to OCaml by extending the interpretation of the existing notation for GADTs, so that indexes with variance annotations carry subtyping constraints rather than equality constraints. For example, the sub type (Fig. 2) might be defined as follows:

```
type ('a,'b) sub = Sub : ('a,'a) sub
```

so that when a value of type (a,b) sub matches a pattern x in the type checker knows that a is a subtype of b.

Extending an ML dialect centred around subtyping, such as MLsub [7], might prove even more fruitful.
References


