Chapter 1

Polymorphism

In programming language theory, polymorphism (where poly means “many” and morph “shape”) refers to the mechanism by which the same piece of code can be reused for different types of objects. C++ templates are a good example of a language construct for polymorphism: the same C++ template can be instantiated to different classes which operate on different types of objects all in a uniform way. The recent version of Java (J2SE 5.0) also supports generics which provides polymorphism in a similar way to C++ templates.

There are two kinds of polymorphism: parametric polymorphism and ad hoc polymorphism. Parametric polymorphism enables us to write a piece of code that operates on all types of objects all in a uniform way. Such a piece of code provides a high degree of generality by accepting all types of objects, but cannot exploit specific properties of different types of objects. Ad hoc polymorphism, in contrast, allows a piece of code to exhibit different behavior depending on the type of objects it operates on. The operator + of SML is an example of ad hoc polymorphism: both int * int -> int and real * real -> real are valid types for +, which manipulates integers and floating point numbers differently. In this chapter, we restrict ourselves to parametric polymorphism.

We begin with System F, an extension of the untyped λ-calculus with polymorphic types. Despite its syntactic simplicity and rich expressivity, System F is not a good framework for practical functional languages because the problem of assigning a polymorphic type (of System F) to an expression in the untyped λ-calculus is undecidable (i.e., there is no algorithm for solving the problem for all input expressions). We will then take an excursion to the predicate polymorphic λ-calculus, another extension of the untyped λ-calculus with polymorphic types which is a sublanguage of System F and is thus less expressive than System F. (Interestingly it uses slightly more complex syntax.) Our study of polymorphism will culminate in the formulation of the polymorphic type system of SML, called let-polymorphism, which is a variant of the type system of the predicate polymorphic λ-calculus. Hence the study of System F is our first step toward the polymorphic type system of SML!

1.1 System F

Consider a λ-abstraction λx. x of the untyped λ-calculus. We wish to extend the definition of the untyped λ-calculus so that we can assign a type to λx. x. Assigning a type to λx. x involves two tasks: binding variable x to a type and deciding the type of the resultant expression.

In the case of the simply typed λ-calculus, we have to choose a specific type for variable x, say, bool. Then the resultant λ-abstraction λx: bool. x has type bool → bool. Ideally, however, we do not want to stipulate a specific type for x because λx. x is an identity function that works for any type. For example, λx. x is an identity function for an integer type int, but once the type of λx. x is fixed as bool → bool, we cannot use it for integers. Hence a better answer would be to bind x to an “any type” α and assign type α → α to λx:α. x.

Now every variable in the untyped λ-calculus is assigned an “any type,” and there arises a need to distinguish between different “any types.” As an example, consider a λ-abstraction λx. λy. (x, y) where

1Generics in the Java language does not fully support parametric polymorphism: it accepts only Java objects (of class Object), and does not accept primitive types such as int.
(x, y) denotes a pair of x and y. Since both x and y may assume an “any type,” we could assign the same “any type” to x and y as follows:

\[ \lambda x: \alpha. \lambda y: \alpha. (x, y) : \alpha \to \alpha \to \alpha \times \alpha \]

Although it is fine to assign the same “any type” to both x and y, it does not give the most general type for \( \lambda x. \lambda y. (x, y) \) because x and y do not have to assume the same type in general. Instead we need to assign a different “any type” \( \beta \) to y so that x and y remain independent of each other:

\[ \lambda x: \alpha. \lambda y: \beta. (x, y) : \alpha \to \beta \to \alpha \times \beta \]

Since each variable in a given expression may need a fresh “any type,” we introduce a new construct \( \Lambda \alpha. e \), called a type abstraction, for declaring \( \alpha \) as a fresh “any type,” or a fresh type variable. That is, a type abstraction \( \Lambda \alpha. e \) declares a type variable \( \alpha \) for use in expression \( e \); we may rename \( \alpha \) in a way analogous to \( \alpha \)-conversions on \( \lambda \)-abstractions. If \( e \) has type \( A \), then \( \Lambda \alpha. e \) is assigned a polymorphic type \( \forall \alpha. A \) which reads “for all \( \alpha \), \( A \).” Note that \( A \) in \( \forall \alpha. A \) may use \( \alpha \) (e.g., \( A = \alpha \to \alpha \)). Then \( \Lambda x. \lambda y. (x, y) \) is converted to the following expression:

\[ \Lambda \alpha. \Lambda \beta. \lambda x: \alpha. \lambda y: \beta. (x, y) : \forall \alpha. \forall \beta. \alpha \to \alpha \times \beta \]

→ has a higher operator precedence than \( \forall \), so \( \forall \alpha. A \to B \) is equal to \( \forall \alpha. (A \to B) \), not \( (\forall \alpha. A) \to B \). Hence \( \forall \alpha. \forall \beta. \alpha \to \beta \to \alpha \times \beta \) is equal to \( \forall \alpha. \forall \beta. (\alpha \to \beta) \to \alpha \times \beta \).

Back to the example of the identity function which is now written as \( \Lambda \alpha. \lambda x: \alpha. x \) of type \( \forall \alpha. \alpha \to \alpha \), let us apply it to a boolean truth \( \text{true} \) with \( \text{true} \) having type \( \text{bool} \). But the \( \text{true} \) of type \( \text{bool} \) is not a boolean true in type \( \text{bool} \). Hence we need to convert \( \Lambda \alpha. \lambda x: \alpha. x \) to an identity function \( \lambda x: \text{bool}. x \) by instantiating \( \alpha \) to a specific type \( \text{bool} \). To this end, we introduce a new construct \( e [\alpha] \), called a type application, such that \( (\Lambda \alpha. e) [\alpha] \) reduces to \( [\alpha]/e \) which substitutes \( A \) for \( \alpha \) in \( e \):

\[(\Lambda \alpha. e) [\alpha] \mapsto [\alpha]/e\]

(Thus the only difference of a type application \( e [\alpha] \) from an ordinary application \( e_1 \, e_2 \) is that a type application substitutes a type for a type variable instead of an expression for an ordinary variable.) Then \( (\Lambda \alpha. \lambda x: \alpha. x) [\text{bool}] \) reduces to an identity function specialized for type \( \text{bool} \), and an ordinary application \( (\Lambda \alpha. \lambda x: \alpha. x) [\text{true}] \) true finishes the job.

System F is essentially an extension of the untyped \( \lambda \)-calculus with type abstractions and type applications. Although variables in \( \lambda \)-abstractions are always annotated with their types, we do not consider System F as an extension of the simply typed \( \lambda \)-calculus because System F does not have to assume base types. The abstract syntax for System F is as follows:

\[
\begin{align*}
\text{type} & : A ::= \text{A} \mid \alpha \mid \forall \alpha. A \\
\text{expression} & : e ::= x \mid \lambda x: \text{A}. e \mid e \, e \mid \Lambda \alpha. e \mid e [\alpha] \\
\text{value} & : v ::= \lambda x: \text{A}. e \mid \Lambda \alpha. e
\end{align*}
\]

The reduction rules for type applications are analogous to those for ordinary applications except that there is no reduction rule for types:

\[
\begin{align*}
\text{e} & \mapsto e' \\
\text{e} [\alpha] & \mapsto e' [\alpha] \\
\text{Tlam} & : (\Lambda \alpha. e) [\alpha] \mapsto [\alpha]/e
\end{align*}
\]

\([\alpha]/e\) substitutes \( A \) for \( \alpha \) in \( e \); we omit its pedantic definition here. For ordinary applications, we reuse the reductions rules for the simply typed \( \lambda \)-calculus.

There are two important observations to make about the abstract syntax. First the syntax for type applications implies that type variables may be instantiated to all kinds of types, including even polymorphic types. For example, the identity function \( \Lambda \alpha. \lambda x: \alpha. x \) may be applied to its own type \( \forall \alpha. \alpha \to \alpha \)!

Such flexibility in type applications is the source of rich expressivity of System F, but on the other hand, it also makes System F a poor choice as a framework for practical functional languages. Second we define a type abstraction \( \Lambda \alpha. e \) as a value, even though it appears to be computationally equivalent to \( e \).
As an example, let us write a function compose for composing two functions. One approach is to require that all type variables in the type of compose be instantiated before producing a λ-abstraction:

\[
\text{compose} : \forall \alpha. \forall \beta. \forall \gamma. (\alpha \to \beta) \to (\beta \to \gamma) \to (\alpha \to \gamma)
\]

Alternatively we may require that only the first two type variables \( \alpha \) and \( \beta \) be instantiated before producing a λ-abstraction which returns a type abstraction expecting the third type variable \( \gamma \):

\[
\text{compose} = \Lambda \alpha. \Lambda \beta. \Lambda \gamma. \lambda f : \alpha \to \beta. \lambda g : \beta \to \gamma. \lambda x : \alpha. g (f x)
\]

As for the type system, System F is not a straightforward extension of the simply typed \( \lambda \)-calculus because of the inclusion of type variables. In the simply typed \( \lambda \)-calculus, \( x : A \) qualifies as a valid type binding regardless of type \( A \) and the order of type bindings in a typing context does not matter by Proposition ???. In System F, \( x : A \) may not qualify as a valid type binding if type \( A \) contains type variables. For example, without type abstractions declaring type variables \( \alpha \) and \( \beta \), we may not use \( \alpha \to \beta \) as a type and hence \( x : \alpha \to \beta \) is not a valid type binding. This observation leads to the conclusion that in System F, a typing context consists not only of type bindings but also of a new form of declarations for indicating which type variables are valid and which are not; moreover the order of elements in a typing context now does matter because of type variables.

We define a typing context as an ordered set of type bindings and type declarations; a type declaration \( \alpha \) type declares \( \alpha \) as a type, or equivalently, \( \alpha \) as a valid type variable:

\[
\text{tying context} \quad \Gamma ::= \emptyset \mid \Gamma, x : A \mid \Gamma, \alpha \text{ type}
\]

We simplify the presentation by assuming that variables and type variables in a typing context are all distinct. We consider a type variable \( \alpha \) as valid only if its type declaration appears to its left. For example, \( \Gamma_1, \alpha \text{ type}, x : \alpha \to \alpha \) is a valid typing context because \( \alpha \) in \( x : \alpha \to \alpha \) has been declared as a type variable in \( \alpha \) type (provided that \( \Gamma_1 \) is also a valid typing context). \( \Gamma_1, x : \alpha \to \alpha, \alpha \) type is, however, not a valid typing context because \( \alpha \) is used in \( x : \alpha \to \alpha \) before it is declared as a type variable in \( \alpha \) type.

The type system of System F uses two forms of judgments: a typing judgment \( \Gamma \vdash e : A \) whose meaning is the same as in the simply typed \( \lambda \)-calculus, and a type judgment \( \Gamma \vdash A \) type which means that \( A \) is a valid type with respect to typing context \( \Gamma \). We need type judgments because the definition of syntactic category type in the abstract syntax is incapable of differentiating valid type variables from invalid ones. We refer to an inference rule deducing a type judgment as a type rule.

The type system of System F uses the following rules:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Gamma \vdash A \text{ type} ) ( \Gamma \vdash B \text{ type} ) ( \text{Ty} \to ) ( \alpha \text{ type} \in \Gamma ) ( \Gamma \vdash \alpha \text{ type} )</td>
<td>( \Gamma \vdash A \to B \text{ type} )</td>
</tr>
<tr>
<td>( \Gamma \vdash x : A ) ( \text{Var} ) ( \Gamma, x : A \vdash e : B ) ( \text{\textit{\text{\textbackslash -}\text{l}}\text{\textit{\text{\textbackslash -}}} ) ( \Gamma \vdash e : A \to B ) ( \Gamma \vdash e' : A )</td>
<td>( \text{\textit{\text{\textbackslash -}}\text{E} )</td>
</tr>
<tr>
<td>( \Gamma, \alpha \text{ type} \vdash e : A ) ( \forall ) ( \Gamma \vdash \alpha e : \forall \alpha . A ) ( \forall )</td>
<td>( \Gamma \vdash e : \forall \alpha . B ) ( \Gamma \vdash A \text{ type} ) ( \forall \text{E} )</td>
</tr>
<tr>
<td>( \Gamma \vdash e [A] : [A/\alpha ]B ) ( \forall \text{E} )</td>
<td></td>
</tr>
</tbody>
</table>

A proof of a type judgment \( \Gamma \vdash A \) type does not use type bindings in \( \Gamma \). In the rule \( \text{\textit{\text{\textbackslash -}}\text{l}} \), the typing context \( \Gamma, x : A \) assumes that \( A \) is a valid type with respect to \( \Gamma \). Hence the rule \( \text{\textit{\text{\textbackslash -}}\text{l}} \) does not need a separate premise \( \Gamma \vdash A \) type. The rule \( \forall \text{\textit{\text{\textbackslash -}}} \), called the \( \forall \) Introduction rule, introduces a polymorphic type \( \forall \alpha . A \) from the judgment in the premise. The rule \( \forall \text{E} \), called the \( \forall \) Elimination rule, eliminates a polymorphic type \( \forall \alpha . B \) by substituting a valid type \( A \) for variable \( \alpha \). Note that the typing rule \( \forall \text{E} \) uses a substitution of a type into another type whereas the reduction rule \( \text{Tapp} \) uses a substitution of a type into an expression.

A type judgment \( \Gamma \vdash A \) type is also an example of a hypothetical judgment which deduces a “judgment” \( A \) type using each “judgment” \( A_i \) type in \( \Gamma \) as a hypothesis.

June 16, 2007 3
As an example of a typing derivation, let us find the type of an identity function specialized for type bool; we assume that $\Gamma \vdash \text{bool}$ type holds for any typing context $\Gamma$ (see the type rule $\text{TyBool}$ below):

\[
\frac{\text{Var}}{\alpha \text{ type}, x : \alpha \vdash x : \alpha} \quad \frac{\text{Var}}{\alpha \text{ type} \vdash \lambda x : \alpha. x : \alpha \to \alpha} \quad \frac{\text{Var} \quad \forall \gamma : \alpha. \lambda x : \alpha. x : \forall \alpha. \alpha \to \alpha} {\vdash \lambda \alpha. \lambda x : \alpha. x : \forall \alpha. \alpha \to \alpha}
\]

Since $[\text{bool}/\alpha](\alpha \to \alpha)$ is equal to $\text{bool} \to \text{bool}$, the type application has type $\text{bool} \to \text{bool}$.

The proof of type safety of System F needs three substitution lemmas as there are three kinds of substitutions: $[\alpha/A]B$ for the rule $\forall E$, $[\alpha/A]e$ for the rule $\text{Tapp}$, and $[e'/x]e$ for the rule $\text{App}$. We write $[A/\alpha]\Gamma$ for substituting $A$ for $\alpha$ in all type bindings in $\Gamma$.

Lemma 1.1 (Type substitution into types).
If $\Gamma \vdash A$ type and $\Gamma$, $\alpha$ type, $\Gamma' \vdash B$ type, then $\Gamma, [A/\alpha]\Gamma' \vdash [A/\alpha]B$ type.

Lemma 1.2 (Type substitution into expressions).
If $\Gamma \vdash A$ type and $\Gamma$, $\alpha$ type, $\Gamma' \vdash e : B$, then $\Gamma, [A/\alpha]\Gamma' \vdash [A/\alpha]e : [A/\alpha]B$.

Lemma 1.3 (Expression substitution).
If $\Gamma \vdash e : A$ and $\Gamma, x : A, \Gamma' \vdash e' : C$, then $\Gamma, \Gamma' \vdash [e/x]e' : C$.

In Lemmas 1.1 and 1.2, we have to substitute $A$ into $\Gamma'$, which may contain types involving $\alpha$. In Lemma 1.2, we have to substitute $A$ into $e$ and $B$, both of which may contain types involving $\alpha$. Lemma 1.3 reflects the fact that typing contexts are ordered sets.

The proof of type safety of System F is similar to the proof for the simply typed $\lambda$-calculus. We need to extend the canonical forms lemma (Lemma ??) and the inversion lemma (Lemma ??):

Lemma 1.4 (Canonical forms).
If $v$ is a value of type $\forall \alpha. A$, then $v$ is a type abstraction $\lambda \alpha. e$.

Lemma 1.5 (Inversion). Suppose $\Gamma \vdash e : C$.
If $e = \lambda \alpha. e'$, then $C = \forall \alpha. A$ and $\Gamma$, $\alpha$ type $\vdash e' : A$.

Theorem 1.6 (Progress). If $\cdot \vdash e : A$ for some type $A$, then either $e$ is a value or there exists $e'$ such that $e \leftrightarrow e'$.

Theorem 1.7 (Type preservation). If $\Gamma \vdash e : A$ and $e \leftrightarrow e'$, then $\Gamma \vdash e' : A$.

1.2 Type reconstruction

The type systems of the simply typed $\lambda$-calculus and System F require that all variables in $\lambda$-abstractions be annotated with their types. While it certainly simplifies the proof of type safety (and the study of type-theoretic properties in general), such a requirement on variables is not a good idea when it comes to designing practical functional languages. One reason is that annotating all variables with their types does not always improve code readability. On the contrary, excessive type annotations often reduce code readability! For example, one would write an SML function adding two integers as $\text{fn} \ x \Rightarrow \text{fn} \ y \Rightarrow x + y$, which is no less readable than a fully type-annotated function $\text{fn} \ x : \text{int} \Rightarrow \text{fn} \ y : \text{int} \Rightarrow x + y$. A more important reason is that in many cases, types of variables can be inferred, or reconstructed, from the context. For example, the presence of $+$ in $\text{fn} \ x \Rightarrow \text{fn} \ y \Rightarrow x + y$ gives enough information to decide a unique type $\text{int}$ for both $x$ and $y$. Thus we wish to eliminate such a requirement on variables, so as to provide programmers with more flexibility in type annotations, by developing a type reconstruction algorithm which automatically infers types for variables.

In the case of System F, the goal of type reconstruction is to convert an expression $e$ in the untyped $\lambda$-calculus to a well-typed expression $e'$ in System F such that erasing type annotations (including type abstractions and type applications) in $e'$ yields the original expression $e$. That is, by reconstructing
types for all variables in $e$, we obtain a new well-typed expression $e'$ in System F. Formally we define an erasure function $\text{erase}(\cdot)$ which takes an expression in System F and erases all type annotations in it:

\[
\begin{align*}
\text{erase}(x) &= x \\
\text{erase}(\lambda x: A. e) &= \lambda x. \text{erase}(e) \\
\text{erase}(e_1 e_2) &= \text{erase}(e_1) \text{erase}(e_2) \\
\text{erase}(\lambda a. e) &= \text{erase}(e) \\
\text{erase}(e[A]) &= \text{erase}(e)
\end{align*}
\]

The erasure function respects the reduction rules for System F in the following sense:

**Proposition 1.8.** If $e \mapsto e'$ holds in System F, then $\text{erase}(e) \mapsto^* \text{erase}(e')$ holds in the untyped $\lambda$-calculus.

The problem of type reconstruction is then to convert an expression $e$ in the untyped $\lambda$-calculus to a well-typed expression $e'$ in System F such that $\text{erase}(e') = e$. We say that an expression $e$ in the untyped $\lambda$-calculus is typable in System F if there exists such a well-typed expression $e'$.

As an example, let us consider an untyped $\lambda$-abstraction $\lambda x. x x$. It is not typable in the simply typed $\lambda$-calculus because the first $x$ in $x x$ must have a type strictly larger than the second $x$, which is impossible. It is, however, typable in System F because we can replace the first $x$ in $x x$ by a type application. Specifically $\lambda x: \forall \alpha. \alpha \to \alpha. x [\forall \alpha. \alpha \to \alpha] x$ is a well-typed expression in System F which erases to $\lambda x. x x$:

\[
\begin{array}{c}
\frac{\alpha \text{ type } \in e : \forall \alpha. \alpha \to \alpha, \alpha \text{ type } TyVar \text{ Var } x : \forall \alpha. \alpha \to \alpha \text{ type } TyVar \text{ Var } x : \forall \alpha. \alpha \to \alpha \text{ type }}{e : \forall \alpha. \alpha \to \alpha \to (\forall \alpha. \alpha \to \alpha) - \text{Var}} \\
\frac{\alpha \text{ type } \in e : \forall \alpha. \alpha \to \alpha, \alpha \text{ type } TyVar \text{ Var } x : \forall \alpha. \alpha \to \alpha \text{ type } TyVar \text{ Var } x : \forall \alpha. \alpha \to \alpha \text{ type } TyVar}{x : \forall \alpha. \alpha \to \alpha \to (\forall \alpha. \alpha \to \alpha) - \text{Var}} \\
\frac{x : \forall \alpha. \alpha \to \alpha \to (\forall \alpha. \alpha \to \alpha) - \text{Var}}{x : \forall \alpha. \alpha \to \alpha \to (\forall \alpha. \alpha \to \alpha) - \text{Var}}
\end{array}
\]

The proof of $x : \forall \alpha. \alpha \to \alpha \to (\forall \alpha. \alpha \to \alpha)$ is shown below:

(Proposition does not use the type binding $x : \forall \alpha. \alpha \to \alpha$.) Hence a type reconstruction algorithm for System F, if any, would convert $\lambda x. x x$ to $\lambda x: \forall \alpha. \alpha \to \alpha. x [\forall \alpha. \alpha \to \alpha] x$.

It turns out that not every expression in the untyped $\lambda$-calculus is typable in System F. For example, $\omega = (\lambda x. x x) (\lambda x. x x)$ is not typable: there is no well-typed expression in System F that erases to $\omega$. The proof exploits the normalization property of System F which states that the reduction of a well-typed expression in System F always terminates. Thus a type reconstruction algorithm for System F first decides if a given expression $e$ is typable or not in System F; if $e$ is typable, the algorithm yields a corresponding expression in System F.

Unfortunately the problem of type reconstruction in System F is undecidable: there is no algorithm for deciding whether a given expression in the untyped $\lambda$-calculus is typable or not in System F. Our plan now is to find a compromise between rich expressivity and decidability of type reconstruction — we wish to identify a sublanguage of System F that supports polymorphic types and also has a decidable type construction algorithm. Section 1.4 presents such a sublanguage, called the predicative polymorphic $\lambda$-calculus, which is extended to the polymorphic type system of SML in Section 1.5.

### 1.3 Programming in System F

We have seen in Section ?? how to encode common datatypes in the untyped $\lambda$-calculus. While these expressions correctly encode their respective datatypes, unavailability of a type system makes it difficult to express the intuition behind the encoding of each datatype. Besides it is often tedious and even unreliable to check the correctness of an encoding without recourse to a type system.

In this section, we rewrite these untyped expressions into well-typed expressions in System F. A direct definition of a datatype in terms of types in System F provides the intuition behind its encoding,
and availability of type annotations within expressions makes it easy to check the correctness of the encoding.

Let us begin with base types bool and nat for Church booleans and numerals, respectively. The intuition behind Church booleans is that a boolean value chooses one of two different options. The following definition of the base type bool is based on the decision to assign the same type α to both options:

$$\text{bool} \, = \, \forall \alpha. \alpha \to \alpha \to \alpha$$

Then boolean values true and false, both of type bool, are encoded as follows:

$$\text{true} \, = \, \Lambda \alpha. \lambda t: \alpha. \lambda f: \alpha. \, t$$
$$\text{false} \, = \, \Lambda \alpha. \lambda t: \alpha. \lambda f: \alpha. \, f$$

The intuition behind Church numerals is that a Church numeral \(\hat{n}\) takes a function \(f\) and returns another function \(f^n\) which applies \(f\) exactly \(n\) times. In order for \(f^n\) to be well-typed, its argument type and return type must be identical. Hence we define the base type nat in System F as follows:

$$\text{nat} \, = \, \forall \alpha. (\alpha \to \alpha) \to (\alpha \to \alpha)$$

Then a zero of type nat and a successor function succ of type \(\text{nat} \to \text{nat}\) are encoded as follows:

$$\text{zero} \, = \, \Lambda \alpha. \lambda f: \alpha \to \alpha. \lambda x: \alpha. \, x$$
$$\text{succ} \, = \, \lambda n: \text{nat}. \Lambda \alpha. \lambda f: \alpha \to \alpha. \lambda x: \alpha. \, (n \llbracket \alpha \rrbracket) \, f \, x$$

The definition of a product type \(A \times B\) in System F exploits the fact that in essence, a value of type \(A \times B\) contains a value of type \(A\) and another value of type \(B\). If we think of \(A \to B \to \alpha\) as a type for a function taking two arguments of types \(A\) and \(B\) and returning a value of type \(\alpha\), a value of type \(A \times B\) contains everything necessary for applying such a function, which is expressed in the following definition of \(A \times B\):

$$A \times B \, = \, \forall \alpha. (A \to B \to \alpha) \to \alpha$$

Pairs and projections are encoded as follows; note that without type annotations, these expressions degenerate to pairs and projections for the untyped \(\lambda\)-calculus given in Section ??:

$$\text{pair} \, : \, \forall \alpha. \forall \beta. \alpha \to B \to \beta \to \alpha \times \beta = \Lambda \alpha. \Lambda \beta. \lambda x: \alpha. \lambda y: \beta. \Lambda \gamma. \lambda f: \alpha \to \beta \to \gamma. \, \lambda x: \alpha. \, \lambda y: \beta. \, f \, x \, y$$
$$\text{fst} \, : \, \forall \alpha. \forall \beta. \alpha \times \beta \to \alpha = \Lambda \alpha. \Lambda \beta. \lambda p: \alpha \times \beta. \lambda p \, [\alpha]. \, \lambda x: \alpha. \, \lambda y: \beta. \, x$$
$$\text{snd} \, : \, \forall \alpha. \forall \beta. \alpha \times \beta \to \beta = \Lambda \alpha. \Lambda \beta. \lambda p: \alpha \times \beta. \lambda p \, [\beta]. \, \lambda x: \alpha. \, \lambda y: \beta. \, y$$

The type unit is a general product type with no element and is thus defined as \(\forall \alpha. \alpha \to \alpha\) which is obtained by removing \(A\) and \(B\) from the definition of \(A \times B\). The encoding of a unit () is obtained by removing \(x\) and \(y\) from the encoding of pair:

$$() \, : \, \text{unit} \, = \, \Lambda \alpha. \lambda x: \alpha. \, x$$

The definition of a sum type \(A + B\) in System F reminds us of the typing rule +E for sum types: given a function \(f\) of type \(A \to \alpha\) and another function \(g\) of type \(B \to \alpha\), a value \(v\) of type \(A + B\) applies the right function (either \(f\) or \(g\)) to the value contained in \(v\):

$$A + B \, = \, \forall \alpha. (A \to \alpha) \to (B \to \alpha) \to \alpha$$

Injections and case expressions are translations of the typing rules +I_L, +I_R, and +E:

$$\text{inl} \, : \, \forall \alpha. \forall \beta. \alpha \to \alpha + \beta$$
$$\text{inr} \, : \, \forall \alpha. \forall \beta. \beta \to \alpha + \beta$$
$$\text{case} \, : \, \forall \alpha. \forall \beta. \forall \gamma. \alpha + \beta \to (\alpha \to \gamma) \to (\beta \to \gamma) \to \gamma$$

Exercise 1.9. Encode inl, inr, and case in System F.

The type void is a general sum type with no element and is thus defined as \(\forall \alpha. \alpha\) which is obtained by removing \(A\) and \(B\) from the definition of \(A + B\). Needless to say, there is no expression of type void in System F. (Why?)

---

3We may also interpret nat as \(\forall \alpha. (\alpha \to \alpha) \to \alpha \to \alpha\) such that a Church numeral \(\hat{n}\) takes a successor function succ of type \(\alpha \to \alpha\) and a zero zero of type \(\alpha\) to return succ^n 0 of type \(\alpha\).
1.4 Predicative polymorphic $\lambda$-calculus

This section presents the predicative polymorphic $\lambda$-calculus which is a sublanguage of System F with a decidable type construction algorithm. It is still not a good framework for practical functional languages because polymorphic types are virtually useless! Nevertheless it helps us a lot to motivate the development of let-polymorphism, the most popular polymorphic type system found in modern functional languages.

The key observation is that undecidability of type reconstruction in System F is traced back to the self-referential nature of polymorphic types: we augment the set of types with new elements called type variables and polymorphic types, but the syntax for type applications allows type variables to range over not only existing types (such as function types) but also these new elements which include polymorphic types themselves. That is, there is no restriction on type $A$ in a type application $e[A]$ where type $A$, which is to be substituted for a type variable, can be not only a function type but also another polymorphic type.

The predicative polymorphic $\lambda$-calculus recovers decidability of type reconstruction by prohibiting type variables from ranging over polymorphic types. We stratify types into two kinds: monotypes which include all kinds of types: polynomials and polytypes. A typing judgment $\Gamma \vdash e : U$ now accepts only a monotype $\Gamma$, $\vdash x : A$ where $U$ is another polytype (see the typing rule $\forall E$ below).

As in System F, the type system of the predicative polymorphic $\lambda$-calculus uses two forms of judgments: a typing judgment $\Gamma \vdash e : U$ and a type judgment $\Gamma \vdash A$ type. The difference is that $\Gamma \vdash A$ type now checks if a given type is a valid monotype. That is, we do not use a type judgment $\Gamma \vdash U$ type (which is actually unnecessary because every polytype is written in prenex form anyway). Thus the system system uses the following rules; note that the rule $\forall E$ from System F is gone:

<table>
<thead>
<tr>
<th>Monotype</th>
<th>Polymorphic Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A ::= A \rightarrow \alpha$</td>
<td>$\forall \alpha. A$</td>
</tr>
<tr>
<td>$\alpha \text{ type } \in \Gamma$</td>
<td>$\forall \alpha. e : \forall \alpha. U$</td>
</tr>
<tr>
<td>$\forall \alpha. \text{ type } \in \Gamma$</td>
<td>$\forall \alpha. e : [\alpha] : [A/\alpha]U$</td>
</tr>
</tbody>
</table>

Unfortunately the use of a monotype $A$ in a $\lambda$-abstraction $\lambda x : A. e$ defeats the purpose of introducing polymorphic types into the type system: even though we can now write an expression of a polymorphic type $U$, we can never instantiate type variables in $U$ more than once! Suppose, for example, that we wish to apply a polymorphic identity function $id = \Lambda \alpha. \lambda x : \alpha. x$ to two different types, say, bool and int. In the untyped $\lambda$-calculus, we would bind a variable $f$ to an identity function and then apply $f$ twice:

$(\lambda f. \text{pair} (f \text{ true}) (f \text{ false})) (\lambda x. x)$

In the predicative polymorphic $\lambda$-calculus, it is impossible to reuse $id$ more than once in this way, since $f$ must be given a monotype while $id$ has a polytype.
We require that let of type bool identity function to two different (mono)types let for declaring variables of polytypes. A \(\lambda\)-Let-polymorphism extends the predicative polymorphic \(\lambda\)-calculus with a new construct that enables us to use a polymorphic expression *polymorphically* in the sense that type variables in it can be instantiated more than once. The new construct preserves decidability of type reconstruction, so let-polymorphism is a good compromise between expressivity and decidability of type reconstruction.

### 1.5 Let-polymorphism

Let-polymorphism extends the predicative polymorphic \(\lambda\)-calculus with a new construct, called a let-binding, for declaring variables of polytypes. A let-binding let \(x:U = e\) binds \(x\) to a polymorphic expression \(e\) of type \(U\) and allows multiple occurrences of \(x\) in \(e\). With a let-binding, we can apply a polymorphic identity function to two different (mono)types bool and int and as follows:

\[
\text{let } f: \forall \alpha. \alpha \rightarrow \alpha. \text{ let } x:U = f \text{ in } \text{type variable}
\]

Since variables can now assume polytypes, we use type bindings of the form \(x : U\) instead of \(x : A\). We require that let-bindings themselves be of monotypes:

| expression | \(e\) ::= \(\cdots\) | let \(x : U = e\) in \(e\) |
| typing context | \(\Gamma\) ::= \(\cdots\) | \(\Gamma, x : U \mid \Gamma, \alpha\) type |

\[\frac{x : U \in \Gamma}{\Gamma \vdash x : U} \quad \frac{\Gamma \vdash e : U}{\Gamma \vdash \text{let } x : U = e} \quad \text{let} \]

The reduction of a let-binding let \(x : U = e\) in \(e\) proceeds by substituting \(e\) for \(x\) in \(e\):

\[
\text{let } x : U = e\text{ in } e' \mapsto [e/x]e'
\]

Depending on the reduction strategy, we may choose to fully evaluate \(e\) before performing the substitution.

Although let \(x : U = e\) in \(e\) reduces to the same expression that an application \((\lambda x : A. e)\) \(e\) reduces to, it is not syntactic sugar for \((\lambda x : A. e)\) \(e\): when \(e\) has a polytype \(U\), let \(x : U = e\) in \(e\) may typecheck by the rule Let, but in general, \((\lambda x : A. e)\) \(e\) does not typecheck because monotype \(A\) does not match the type of \(e\). Therefore, in order to use a polymorphic expression polymorphically, we must bind it to a variable using a let-binding instead of a \(\lambda\)-abstraction.

Then why do we not just allow a \(\lambda\)-abstraction \(\lambda x : U. e\) binding \(x\) to a polytype (which would degenerate let \(x : U = e\) in \(e\) into syntactic sugar)? The reason is that with an additional assumption that \(e\) may have a polytype (e.g., \(\lambda x : U. x\)), such a \(\lambda\)-abstraction collapses the distinction between monotypes and polytypes. That is, polytypes constitute types of System F:

\[
\text{monotype } A ::= U \rightarrow U \mid \alpha \\
\text{polytype } U ::= A \mid \forall \alpha. U \\
\text{type } U ::= U \rightarrow U \mid \alpha \mid \forall \alpha. U
\]

We may construe a let-binding as a restricted use of a \(\lambda\)-abstraction \(\lambda x : U. e\) (binding \(x\) to a polytype) such that it never stands alone as a first-class object and must be applied to a polymorphic expression immediately. At the cost of flexibility in applying such \(\lambda\)-abstractions, let-polymorphism retains decidability of type reconstruction without destroying the distinction between monotypes and polytypes and also without sacrificing too much expressivity. After all, we can still enjoy both polymorphism and decidability of type reconstruction, which is why let-polymorphism is so popular among mainstream functional languages.
1.6 Implicit polymorphism

The polymorphic type systems considered so far are all “explicit” in that polymorphic types are introduced explicitly by type abstractions and that type variables are instantiated explicitly by type applications. An explicit polymorphic type system has the property that every well-typed polymorphic expression has a unique polymorphic type.

The type system of SML uses a different approach to polymorphism: it makes no use of type abstractions and that type variables are instantiated explicitly by type applications. That is, polymorphic types arise “implicitly” from lack of type annotations in λ-abstractions.

As an example, consider an identity function \( \lambda x. x \). It can be assigned such types as \( \text{bool} \to \text{bool} \), \( \text{int} \to \text{int} \), \( \text{(int} \to \text{int}) \to (\text{int} \to \text{int}) \), and so on. These types are all distinct, but are subsumed by the same polytype \( \forall \alpha. \alpha \to \alpha \) in the sense that they are results of instantiating \( \alpha \) in \( \forall \alpha. \alpha \to \alpha \). We refer to \( \forall \alpha. \alpha \to \alpha \) as the principal type of \( \lambda x. x \), which may be thought of as the most general type for \( \lambda x. x \), as opposed to specific types such as \( \text{bool} \to \text{bool} \) and \( \text{int} \to \text{int} \). The type reconstruction algorithm of SML infers a unique principal type for every well-typed expression. Below we discuss the type system of SML and defer details of the type reconstruction algorithm to Section 1.8.

In essence, the type system of SML uses let-polymorphism without type annotations (in λ-abstractions and let-bindings), type abstractions, and type applications:

\[
\begin{align*}
\text{monotype} & : A ::= A \to A | \alpha \\
\text{polytype} & : U ::= A | \forall \alpha. U \\
\text{expression} & : e ::= x | \lambda x. e | e \ e | \text{let } x = e \text{ in } e \\
\text{value} & : v ::= \lambda x. e \\
\text{typing context} & : \Gamma ::= \cdot | \Gamma, x : U | \Gamma, \alpha \text{ type}
\end{align*}
\]

We use a new typing judgment \( \Gamma \vdash e : U \) to express that untyped expression \( e \) is typable with a polytype \( U \). The intuition (which will be made clear in Theorem 1.11) is that if \( \Gamma \vdash e : U \) holds where \( e \) is untyped, there exists a typed expression \( e' \) such that \( \Gamma \vdash e' : U \) and \( e' \) erases to \( e \) by the following erasure function:

\[
\begin{align*}
\text{erase}(x) & = x \\
\text{erase}(\lambda x. A. e) & = \lambda x. \text{erase}(e) \\
\text{erase}(e_1, e_2) & = \text{erase}(e_1) \text{ erase}(e_2) \\
\text{erase}(\Lambda \alpha. e) & = \text{erase}(e) \\
\text{erase}(e[[\alpha]]) & = \text{erase}(e) \\
\text{erase}(\text{let } x : U = e \text{ in } e') & = \text{let } x = \text{erase}(e) \text{ in erase}(e')
\end{align*}
\]

That is, if \( \Gamma \vdash e : U \) holds, \( e \) has its counterpart in let-polymorphism in Section 1.5. The rules for the typing judgment \( \Gamma \vdash e : U \) are given as follows:

\[
\begin{array}{c}
\Gamma \vdash e : U \quad \Gamma, x : U \vdash e' : A \\
\Gamma \vdash \text{let } x = e \text{ in } e' : A \\
\Gamma \vdash \alpha \text{ type} : U \\
\Gamma \vdash e : \forall \alpha. U \\
\Gamma \vdash A \text{ type} : U
\end{array}
\]

\[
\begin{array}{c}
\Gamma \vdash e : U \\
\Gamma, x : U \vdash e' : A \\
\Gamma \vdash A \to B \\
\Gamma \vdash e' : B \\
\Gamma \vdash e : A \\
\Gamma \vdash e : B \\
\Gamma \vdash e : \forall \alpha. U \\
\Gamma \vdash \forall \alpha. A \text{ type} : \alpha U
\end{array}
\]

Note that unlike in the predicative polymorphic λ-calculus, the rule \(-\text{I}\) allows us to assign any monotype \( A \) to variable \( x \) as long as expression \( e \) is assigned a valid monotype \( B \). Hence, for example, the same λ-abstraction \( \lambda x. x \) can now be assigned different monotypes such as \( \text{bool} \to \text{bool} \), \( \text{int} \to \text{int} \), and \( \alpha \to \alpha \). The rules Gen and Spec correspond to the rules \(-\forall\) and \(-\forall\) in the predicative polymorphic λ-calculus (but not in System F because \( A \) in the rule Spec is required to be a monotype).

In the rule Gen (for generalizing a type), expression \( e \) in the conclusion plays the role of a type abstraction. That is, we can think of \( e \) in the conclusion as \( \text{erase}(\Lambda \alpha. e) \). As an example, let us assign a polytype to the polymorphic identity function \( \lambda x. x \):

\[
\Gamma \vdash \lambda x. x : ?
\]
Intuitively $\lambda x. x$ has type $\alpha \rightarrow \alpha$ for an “any type” $\alpha$, so we first assign a monotype $\alpha \rightarrow \alpha$ under the assumption that $\alpha$ is a valid type variable:

\[
\begin{array}{c}
\Gamma, \alpha \text{ type}, x : \alpha \rightarrow x : \alpha \\
\Gamma, \alpha \text{ type} > \lambda x. x : \alpha \rightarrow \alpha \rightarrow I
\end{array}
\]

Note that $\lambda x. x$ has not been assigned a polytype yet. Also note that $\Gamma > \lambda x. x : \alpha \rightarrow \alpha$ cannot be a valid typing derivation because $\alpha$ is a fresh type variable which is not declared in $\Gamma$. Assigning a polytype $\forall \alpha. \alpha \rightarrow \alpha$ to $\lambda x. x$ is accomplished by the rule Gen:

\[
\begin{array}{c}
\Gamma, \alpha \text{ type}, x : \alpha \rightarrow x : \alpha \\
\Gamma, \alpha \text{ type} > \lambda x. x : \alpha \rightarrow \alpha \\
\Gamma > \lambda x. x : \forall \alpha. \alpha \rightarrow \alpha \rightarrow \text{Gen}
\end{array}
\]

As an example of using two type variables, we assign a polytype $\forall \alpha. \forall \beta. \beta \rightarrow (\alpha \times \beta)$ to $\lambda x. \lambda y. (x, y)$ as follows (where we assume that product types are available):

\[
\begin{array}{c}
\Gamma, \alpha \text{ type}, \beta \text{ type}, x : \alpha \rightarrow y : \beta \rightarrow x : \alpha \\
\Gamma, \alpha \text{ type}, \beta \text{ type}, x : \alpha \rightarrow y : \beta \rightarrow y : \beta \\
\Gamma, \alpha \text{ type}, \beta \text{ type} > \lambda x. \lambda y. (x, y) : \beta \rightarrow (\alpha \times \beta) \\
\Gamma > \lambda x. \lambda y. (x, y) : \forall \alpha. \forall \beta. \beta \rightarrow (\alpha \times \beta) \rightarrow \text{Gen}
\end{array}
\]

Generalizing the example, we can assign a polytype to an expression $e$ in two steps. First we introduce as many fresh type variables as necessary to assign a monotype $A$ to $e$. Then we keep applying the rule Gen to convert, or generalize, $A$ to a polytype $U$. If $A$ uses fresh type variables $\alpha_1, \alpha_2, \ldots, \alpha_n$, then $U$ is given as $\forall \alpha_1. \forall \alpha_2. \ldots \forall \alpha_n. A$:

\[
\begin{array}{c}
\Gamma, \alpha_1 \text{ type}, \alpha_2 \text{ type}, \ldots, \alpha_n \text{ type} > e : A \\
\Gamma, \alpha_1 \text{ type}, \alpha_2 \text{ type}, \ldots, e : \forall \alpha_n. A \rightarrow \text{Gen} \\
\vdots \\
\Gamma, \alpha_1 \text{ type}, \alpha_2 \text{ type} > e : \forall \alpha_2. \ldots \forall \alpha_n. A \rightarrow \text{Gen} \\
\Gamma > e : \forall \alpha_1. \forall \alpha_2. \ldots \forall \alpha_n. A \rightarrow \text{Gen}
\end{array}
\]

In the rule Spec (for specializing a type), expression $e$ in the conclusion plays the role of a type application. That is, we can think of $e$ in the conclusion as $\text{erase}(e [A])$. Thus, by applying the rule Spec repeatedly, we can convert, or specialize, any polytype into a monotype.

A typical use of the rule Spec is to specialize the polytype of a variable introduced in a let-binding (in which case expression $e$ in the rule Spec is a variable). Specifically a let-binding $let x = e \text{ in } f \text{ true, } f \text{ false}$ binds variable $x$ to a polymorphic expression $e$ and uses $x$ monomorphically within $e'$ after specializing the type of $x$ to monotypes by the rule Spec. For example, the following typing derivation for $let f = \lambda x. x \text{ in } (f \text{ true, } f \text{ false})$ applies the rule Spec to variable $f$ twice, where we abbreviate $\Gamma, f : \forall \alpha. \alpha \rightarrow \alpha$ as $\Gamma'$:

\[
\begin{array}{c}
\Gamma' > f : \forall \alpha. \alpha \rightarrow \alpha \\
\Gamma' > f : \text{bool} \rightarrow \text{bool} \rightarrow \text{Spec} \\
\Gamma' > f : \text{true} : \text{bool} \rightarrow \text{True} \\
\Gamma' > f : \text{false} : \text{bool} \rightarrow \text{Int} \\
\Gamma' > (f \text{ true, } f \text{ false}) : \text{bool} \times \text{int} \rightarrow \text{Let}
\end{array}
\]

Note that in typechecking $(f \text{ true, } f \text{ false})$, it is mandatory to specialize the type of $f$ to a monotype $\text{bool} \rightarrow \text{bool}$ or $\text{int} \rightarrow \text{int}$, since an application $f e$ typechecks by the rule $\rightarrow \text{E}$ only if $f$ is assigned a monotype.
If expression $e$ in the rule Spec is not a variable, the typing derivation of the premise $\Gamma \vdash e : \forall \alpha.U$ must end with an application of the rule Gen or another application of the rule Spec. In such a case, we can eventually locate an application of the rule Gen that is immediately followed by an application of the rule Spec:

$$
\frac{
\Gamma, \alpha \text{ type} \vdash e : U \\
\Gamma \vdash e : \forall \alpha.U
}{
\Gamma \vdash e : [A/\alpha]U
}
$$

Here we introduce a type variable $\alpha$ only to instantiate it to a concrete monotype $A$ immediately, which implies that such a typing derivation is redundant and can be removed. For example, when typechecking $(\lambda x. x)$ true, there is no need to take a detour by first assigning a polytype $\forall \alpha. \alpha \rightarrow \alpha$ to $\lambda x. x$ and then instantiating $\alpha$ to bool. Instead it suffices to assign a monotype $\text{bool} \rightarrow \text{bool}$ directly because $\lambda x. x$ is eventually applied to an argument of type bool:

$$
\frac{
\Gamma, \alpha \text{ type} \vdash x : \alpha \text{ type} \vdash \lambda x. x : \alpha \rightarrow \alpha
}{
\Gamma \vdash \lambda x. x : \text{bool} \rightarrow \text{bool}
}
$$

This observation suggests that it is unnecessary to specialize the type of an expression that is not a variable — we only need to apply the rule Spec to polymorphic variables introduced in let-bindings.

The implicit polymorphic type system of SML is connected with let-polymorphism in Section 1.5 via the following theorems:

**Theorem 1.10.** If $\Gamma \vdash e : U$, then $\Gamma \vdash \text{erase}(e) : U$.

**Theorem 1.11.** If $\Gamma \vdash e : U$, then there exists a typed expression $e'$ such that $\Gamma \vdash e' : U$ and $\text{erase}(e') = e$.

### 1.7 Value restriction

The type system presented in the previous section is sound only if it does not interact with computational effects such as mutable references and input/output. To see the problem, consider the following expression where we assume constructs for integers, booleans, and mutable references:

```plaintext
let x = ref (\lambda y, y) in
let _ = x := \lambda y, y + 1 in
(\lambda x) true
```

We can assign a polytype $\forall \alpha. \text{ref} (\alpha \rightarrow \alpha)$ to $\text{ref} (\lambda y, y)$ as follows (where we ignore store typing contexts):

$$
\frac{
\Gamma, \alpha \text{ type} \vdash x : \alpha \text{ type} \vdash \lambda y, y : \alpha \rightarrow \alpha
}{
\Gamma \vdash \lambda y, y : \forall \alpha. \text{ref} (\alpha \rightarrow \alpha)
}
$$

By the rule Spec, then, we can assign either $\text{ref} (\text{int} \rightarrow \text{int})$ or $\text{ref} (\text{bool} \rightarrow \text{bool})$ to variable $x$. Now both expressions $x := \lambda y, y + 1$ and $(\lambda x) \text{true}$ are well-typed, but the reduction of $(\lambda x) \text{true}$ must not succeed because it ends up adding a boolean truth and an integer 1!

In order to avoid the problem arising from the interaction between polymorphism and computational effects, the type system of SML imposes a requirement, called value restriction, that expression $e$ in the rule Gen be a syntactic value:

$$
\frac{
\Gamma, \alpha \text{ type} \vdash v : U
}{
\Gamma \vdash v : \forall \alpha.U
}
$$

The idea is to exploit the fact that computational effects cannot interfere with (polymorphic) values, whose evaluation terminates immediately. Now, for example, $\text{ref} (\lambda y, y)$ cannot be assigned a polytype because $\text{ref} (\lambda y, y)$ is not a value and thus its type cannot be generalized by the rule Gen.
As a consequence of value restriction, variable \( x \) in a let-binding \( \text{let } x = e \text{ in } e' \) can be assigned a polytype only if expression \( e \) is a value. If \( e \) is not a value, \( x \) must be used monomorphically within expression \( e' \), even if \( e \) itself does not specify a unique monotype. This means that we may have to analyze \( e' \) in order to decide the monotype to be assigned to \( x \). As an example, consider the following expression:

\[
\text{let } x = (\lambda y. y) (\lambda z. z) \text{ in } x \text{ true}
\]

As \((\lambda y, y) (\lambda z. z)\) is not a value, variable \( x \) must be assigned a monotype. \((\lambda y, y) (\lambda z. z)\), however, does not specify a unique monotype for \( x \); it only specifies that the type of \( x \) must be of the form \( A \rightarrow A \) for some monotype \( A \). Fortunately the application \( x \text{ true} \) fixes such a monotype \( A \) as \text{bool} and \( x \) is assigned a unique monotype \text{bool} \rightarrow \text{bool}. The following expression, in contrast, is ill-typed because variable \( x \) is used polymorphically:

\[
\text{let } x = (\lambda y. y) (\lambda z. z) \text{ in } (x \text{ true}, x \ 1)
\]

The problem here is that \( x \) needs to be assigned two monotypes \text{bool} \rightarrow \text{bool} and \text{int} \rightarrow \text{int} simultaneously, which is clearly out of the question.

### 1.8 Type reconstruction algorithm

This section presents a type reconstruction algorithm for the type system with implicit polymorphism in Section 1.6. Given an untyped expression \( e \), the goal is to infer a polytype \( U \) such that \( \cdot \triangleright e : U \) holds. In addition to being a valid type for \( e, U \) also needs to be the most general type for \( e \) in the sense that every valid type for \( e \) can be obtained a special case of \( U \) by instantiating some type variables in \( U \). Given \( \lambda x. x \) as input, for example, the algorithm returns the most general polytype \( \forall \alpha. \alpha \rightarrow \alpha \) instead of a specific monotype such as \text{bool} \rightarrow \text{bool}.

Typically the algorithm creates (perhaps a lot of) temporary type variables before finding the most general type of a given expression. We design the algorithm in such a way that all these temporary type variables are valid (simply because there is no reason to create invalid ones). As a result, we no longer need a type declaration \( \alpha \) type in the rule \text{Gen} (because \( \alpha \) is assumed to be a valid type variable) and a type judgment \( \Gamma \vdash A \) type in the rule \text{Spec} (because \( A \) is a valid type if all type variables in it are valid). Accordingly a typing context now consists only of type bindings:

\[
\text{typing context } \quad \Gamma := \cdot | \Gamma, x : U
\]

With the assumption that all type variables are valid, the rules \text{Gen} and \text{Spec} are revised as follows:

\[
\begin{array}{c|c|c}
\text{Gen} & \text{Spec} \\
\hline
\Gamma \triangleright e : U \quad \alpha \notin \text{ftv}(\Gamma) & \Gamma \triangleright e : \forall \alpha. U \\
\hline
\end{array}
\]

Here \( \text{ftv}(\Gamma) \) denotes the set of free type variables in \( \Gamma \); \( \text{ftv}(U) \) denotes the set of free type variables in \( U \):

\[
\begin{align*}
\text{ftv}(\cdot) &= \emptyset \\
\text{ftv}(\Gamma, x : U) &= \text{ftv}(\Gamma) \cup \text{ftv}(U) \\
\text{ftv}(\alpha) &= \{ \alpha \} \\
\text{ftv}(\forall \alpha. U) &= \text{ftv}(U) - \{ \alpha \}
\end{align*}
\]

In the rule \text{Gen}, the condition \( \alpha \notin \text{ftv}(\Gamma) \) checks that \( \alpha \) is a fresh type variable. If \( \Gamma \) contains a type binding \( x : U \) where \( \alpha \) is already in use as a free type variable in \( U \), \( \alpha \) cannot be regarded as a fresh type variable and generalizing \( U \) to \( \forall \alpha. U \) is not justified. In the following example, \( \alpha \rightarrow \alpha \) may generalize to \( \forall \alpha. \alpha \rightarrow \alpha \), assigning the desired polytype to the polymorphic identity function \( \lambda x. x \), because \( \alpha \) is not in use in the empty typing context:

\[
\begin{array}{c|c|c}
\text{Var} & \text{l} & \text{Gen} \\
\hline
\cdot \triangleright x : \alpha & \cdot \triangleright \lambda x. x : \alpha \rightarrow \alpha \\
\end{array}
\]

12

June 16, 2007
If $\alpha$ is already in use as a free type variable, however, such a generalization results in assigning a wrong type to $\lambda x. x$:

$$\begin{array}{c}
\frac{x : \alpha, y : \beta \vdash x : \alpha}{\frac{x : \alpha \vdash \lambda y. x : \alpha \rightarrow \alpha}{x : \alpha \rightarrow \lambda y. x : \forall \alpha. \alpha \rightarrow \alpha}} \\
\frac{x : \alpha \vdash \lambda y. x : \forall \alpha. \alpha \rightarrow \alpha}{\frac{x : \alpha \vdash \lambda y. x : \forall \beta. \beta \rightarrow \alpha}{x : \alpha \rightarrow \forall \beta. \beta \rightarrow \alpha}}
\end{array}$$

Here variable $y$ is unrelated to variable $x$, yet is assigned the same type in the premise of the rule $\rightarrow \text{l}$. A correct typing derivation assigns a fresh type variable to $y$ to reflect the fact that $x$ and $y$ are unrelated:

$$\begin{array}{c}
\frac{x : \alpha, y : \beta \vdash x : \alpha}{\frac{x : \alpha \vdash \lambda y. x : \alpha \rightarrow \alpha}{x : \alpha \rightarrow \lambda y. x : \forall \alpha. \alpha \rightarrow \alpha}} \\
\frac{x : \alpha \vdash \lambda y. x : \forall \alpha. \alpha \rightarrow \alpha}{\frac{x : \alpha \vdash \lambda y. x : \forall \beta. \beta \rightarrow \alpha}{x : \alpha \rightarrow \forall \beta. \beta \rightarrow \alpha}}
\end{array}$$

As an example of applying the rule $\text{Spec}$, here is a typing derivation assigning a monotype $\text{bool} \rightarrow \text{bool}$ to $\lambda x. x$ by instantiating a type variable:

$$\begin{array}{c}
\frac{x : \alpha \vdash \lambda x. x : \forall \alpha. \alpha \rightarrow \alpha}{\frac{x : \alpha \vdash \lambda x. x : \forall \text{bool} \rightarrow \text{bool}}{x : \alpha \rightarrow \lambda x. x : \forall \text{bool} \rightarrow \text{bool}}}
\end{array}$$

For the above example, the rule $\text{Spec}$ is unnecessary because we can directly assign $\text{bool}$ to variable $x$:

$$\begin{array}{c}
\frac{x : \text{bool} \vdash x : \text{bool}}{\frac{x : \alpha \vdash \lambda x. x : \forall \alpha. \alpha \rightarrow \alpha}{x : \alpha \rightarrow \lambda x. x : \forall \text{bool} \rightarrow \text{bool}}}
\end{array}$$

As we have seen in Section 1.6, however, the rule $\text{Spec}$ is indispensable for specializing the type of a variable introduced in a let-binding. In the following example, the same type variable $\alpha$ in $\forall \alpha. \alpha \rightarrow \alpha$ is instantiated to two different types $\beta \rightarrow \beta$ and $\beta$ by the rule $\text{Spec}$:

$$\begin{array}{c}
\frac{x : \alpha \vdash \lambda x. x : \forall \alpha. \alpha \rightarrow \alpha}{\frac{x : \alpha \vdash \lambda x. x : \forall \beta. \beta \rightarrow \alpha}{\frac{x : \alpha \vdash \lambda x. x : \forall \beta. \beta \rightarrow \alpha}{x : \alpha \rightarrow \lambda x. x : \forall \beta. \beta \rightarrow \alpha}}}
\end{array}$$

Exercise 1.12. What is wrong with the following typing derivation?

$$\begin{array}{c}
\frac{x : \forall \alpha. \alpha \rightarrow \alpha \vdash x : \forall \alpha. \alpha \rightarrow \alpha}{\frac{x : \forall \alpha. \alpha \rightarrow \alpha \vdash x : (\forall \alpha. \alpha \rightarrow \alpha) \rightarrow (\forall \alpha. \alpha \rightarrow \alpha)}{\frac{x : \forall \alpha. \alpha \rightarrow \alpha \vdash x : \forall \alpha. \alpha \rightarrow \alpha}{x : \forall \alpha. \alpha \rightarrow \alpha \vdash \lambda x. x : \forall \alpha. \alpha \rightarrow \alpha}}}
\end{array}$$

The type reconstruction algorithm, called $\mathcal{W}$, takes a typing context $\Gamma$ and an expression $e$ as input, and returns a pair of a type substitution $S$ and a monotype $A$ as output:

$$\mathcal{W}(\Gamma, e) = (S, A)$$

A type substitution is a mapping from type variables to monotypes. Note that it is not a mapping to polytypes because type variables range only over monotypes:

$$\text{type substitution } \\ S ::= \text{id} \mid \{A/\alpha\} \mid S \circ S$$

$id$ is an identity type substitution which changes no type variable. $\{A/\alpha\}$ is a singleton type substitution which maps $\alpha$ to $A$. $S_1 \circ S_2$ is a composition of $S_1$ and $S_2$ which applies first $S_2$ and then $S_1$. $id$ is the identity for the composition operator $\circ$, i.e., $id \circ S = S \circ id = S$. As $\circ$ is associative, we write $S_1 \circ S_2 \circ S_3$ for $S_1 \circ (S_2 \circ S_3) = (S_1 \circ S_2) \circ S_3$.

An application of a type substitution to a polytype $U$ is formally defined as follows:

June 16, 2007
The specification of the algorithm \( \mathcal{W} \) is concisely stated in its soundness theorem:

**Theorem 1.13 (Soundness of \( \mathcal{W} \)).** If \( \mathcal{W}(\Gamma, e) = (S, A) \), then \( S \vdash e : A \).

Given a typing context \( \Gamma \) and an expression \( e \), the algorithm analyzes \( e \) to build a type substitution \( S \) mapping free type variables in \( \Gamma \) so that \( e \) typechecks with a monotype \( A \). An invariant is that \( S \) has no effect on \( A \), i.e., \( S \cdot A = A \), since \( A \) obtained after applying \( S \) to free type variables in \( \Gamma \). Here are a few examples:

- \( \mathcal{W}(x : \alpha, x + 0) = (\{\text{int}/\alpha\}, \text{int}) \) where we assume a base type int.
  When the algorithm starts, \( x \) has been assigned a yet unknown monotype \( \alpha \). In the course of analyzing \( x + 0 \), the algorithm discovers that \( x \) must be assigned type int, in which case \( x + 0 \) is also assigned type int. Thus the algorithm returns a type substitution \( \{\text{int}/\alpha\} \) along with int as the type of \( x + 0 \).

- \( \mathcal{W}(\lambda x. x + 0) = (\{\text{int}/\alpha\}, \text{int} \rightarrow \text{int}) \) where we assume a base type int.
  When it starts to analyze the \( \lambda \)-abstraction, the algorithm creates a fresh type variable, say \( \alpha \), for variable \( x \) because nothing is known about \( x \) yet. In the course of analyzing the body \( x + 0 \), the algorithm discovers that \( \alpha \) must be identified with int, in which case the type of the \( \lambda \)-abstraction becomes \( \text{int} \rightarrow \text{int} \). Hence the algorithm returns a type substitution \( \{\text{int}/\alpha\} \) (which is not used afterwards) with \( \text{int} \rightarrow \text{int} \) as the type of \( \lambda x. x + 0 \).

- \( \mathcal{W}(\cdot, \lambda x. x) = (\text{id}, \alpha \rightarrow \alpha) \)
  When it starts to analyze the \( \lambda \)-abstraction, the algorithm creates a fresh type variable, say \( \alpha \), for variable \( x \) because nothing is known about \( x \) yet. The body \( x \), however, provides no information on the type of \( x \), either, and the algorithm ends up returning \( \alpha \rightarrow \alpha \) as a possible type of \( \lambda x. x \).

**Exercise 1.14.** What is the result of \( \mathcal{W}(y : \beta, (\lambda x. x) y) \) if the algorithm \( \mathcal{W} \) creates a temporary type variable \( \alpha \) for variable \( x \)? Is the result unique?

Figure 1.1 shows the pseudocode of the algorithm \( \mathcal{W} \). We write \( \vec{\alpha} \) for a sequence of distinct type variables \( \alpha_1, \alpha_2, \ldots, \alpha_n \). Then \( \forall \vec{\alpha}. A \) stands for \( \forall \alpha_1, \forall \alpha_2, \ldots, \forall \alpha_n. A \), and \( \{\beta/\vec{\alpha}\} \) stands for \( \{\beta/\alpha_1\} \circ \cdots \circ \{\beta/\alpha_2\} \circ \{\beta/\alpha_1\} \). We write \( \Gamma + x : U \) for \( \Gamma - \{x : U'\}, x : U \) if \( x : U' \in \Gamma \), and for \( \Gamma, x : U \) if \( \Gamma \) contains no type binding for variable \( x \).

The first case \( \mathcal{W}(\Gamma, x) \) summarizes the result of applying the rule Spec to \( \forall \vec{\alpha}. A \) as many times as the length of \( \vec{\alpha} \). Note that \( \{A/\alpha\} U \) is written as \( \{A/\alpha\} \cdot U \) in the following typing derivation.

\[
\begin{align*}
\forall \alpha_1, \forall \alpha_2, \ldots, \forall \alpha_n. A & \in \Gamma \\
\Gamma \vdash x : \forall \alpha_1, \forall \alpha_2, \ldots, \forall \alpha_n. A & \text{ Var} \\
\Gamma \vdash x : \{\beta_1/\alpha_1\} \cdot \forall \alpha_2, \ldots, \forall \alpha_n. A & \text{ Spec} \\
\Gamma \vdash x : \{\beta_2/\alpha_2\} \circ (\beta_1/\alpha_1) \cdots \forall \alpha_n. A & \text{ Spec} \\
\Gamma \vdash x : \{\beta_n/\alpha_n\} \circ \cdots \circ (\beta_2/\alpha_2) \circ (\beta_1/\alpha_1) \cdot A & \text{ Spec}
\end{align*}
\]
The second case $W(\Gamma, \lambda x. e)$ creates a fresh type variable $\alpha$ to be assigned to variable $x$.

The case $W(\Gamma, e_1 e_2)$ uses an auxiliary function $\text{Unify}(E)$ where $E$ is a set of type equations between monotypes:

$$E ::= \cdot \mid E, A = A$$

$\text{Unify}(E)$ attempts to calculate a type substitution that unifies two types $A$ and $A'$ in each type equation $A = A'$ in $E$. If no such type substitution exists, $\text{Unify}(E)$ fails. Figure 1.2 shows the definition of $\text{Unify}(E)$. We write $S \cdot E$ for the result of applying type substitution $S$ to every type in $E$:

$$S \cdot (E, A = A') = S \cdot E, S \cdot A = S \cdot A'$$

The specification of the function $\text{Unify}$ is stated as follows:

**Proposition 1.15.** If $\text{Unify}(A_1 = A'_1, \cdots, A_n = A'_n) = S$, then $S \cdot A_i = S \cdot A_i$ for $i = 1, \cdots, n$.

Here are a few examples of $\text{Unify}(E)$ where we assume a base type int:

1. $\text{Unify}(\alpha \rightarrow \text{int} \rightarrow \alpha) = \text{fail}$
2. $\text{Unify}(\alpha \rightarrow \alpha \rightarrow \alpha) = \text{fail}$
3. $\text{Unify}(\alpha \rightarrow \alpha \rightarrow \text{int}) = \{\text{int}/\alpha\}$
4. $\text{Unify}(\alpha \rightarrow \beta = \alpha \rightarrow \text{int}) = \{\text{int}/\beta\}$
5. $\text{Unify}(\alpha \rightarrow \beta = \beta \rightarrow \alpha) = \{\beta/\alpha\}$ or $\{\alpha/\beta\}$
6. $\text{Unify}(\alpha \rightarrow \beta = \beta \rightarrow \alpha, \alpha = \text{int}) = \{\text{int}/\beta\} \circ \{\text{int}/\alpha\}$

In cases (1) and (2), the unification fails because both $\text{int} \rightarrow \alpha$ and $\alpha \rightarrow \alpha$ contain $\alpha$ as a free type variable, but are strictly larger than $\alpha$. In case (5), either $\beta/\alpha$ or $\alpha/\beta$ successfully unifies $\alpha \rightarrow \beta$ and $\beta \rightarrow \alpha$. Case (6) uses an additional assumption $\text{Unify}(E, \text{int} = \text{int}) = \text{Unify}(E)$:

$$\text{Unify}(\alpha \rightarrow \beta = \beta \rightarrow \alpha, \alpha = \text{int}) = \text{Unify}(\text{int}/\alpha) \circ (\alpha \rightarrow \beta \rightarrow \beta \rightarrow \alpha) \circ \text{int}/\alpha)$$

$$= \text{Unify}(\text{int}/\alpha) \circ (\alpha \rightarrow \beta \rightarrow \text{int} \circ \text{int}/\alpha)$$

$$= \text{Unify}(\text{int} \circ \beta = \beta \rightarrow \text{int} \circ \text{int}/\alpha)$$

$$= \text{Unify}(\text{int} = \text{int} \circ \text{int}/\alpha)$$

$$= \text{id} \circ \text{int}/\alpha$$

$$= \{\text{int}/\beta\} \circ \{\text{int}/\alpha\}$$
The case $W(\Gamma, \text{let } x = e_1 \text{ in } e_2)$ uses another auxiliary function $\text{Gen}_\Gamma(A)$ which generalizes monotype $A$ to a polytype after taking into account free type variables in typing context $\Gamma$:

$$\text{Gen}_\Gamma(A) = \forall \alpha_1. \forall \alpha_2. \cdots \forall \alpha_n. A$$

where $\alpha_i \notin \text{ftv}(\Gamma)$ and $\alpha_i \in \text{ftv}(A)$ for $i = 1, \cdots, n$.

That is, if $\alpha \in \text{ftv}(A)$ is in $\text{ftv}(\Gamma)$, $\alpha \in A$ is not interpreted as “any type” with respect to $\Gamma$. Note that $\text{Gen}_\Gamma(A) = \forall \alpha_1. \forall \alpha_2. \cdots \forall \alpha_n. A$ is equivalent to applying the rule Gen exactly $n$ times as follows:

$$\frac{\Gamma \vdash e : A \quad \alpha_n \notin \text{ftv}(\Gamma)}{\Gamma \vdash e : \forall \alpha_n. A} \quad \text{Gen}$$

$$\frac{\Gamma \vdash e : \forall \alpha_{n-1}. A \quad \alpha_{n-1} \notin \text{ftv}(\Gamma)}{\Gamma \vdash e : \forall \alpha_{n-2}. \forall \alpha_n. A} \quad \text{Gen}$$

Here are a few examples of $\text{Gen}_\Gamma(A)$:

$$\text{Gen}_{(\alpha \to \alpha)} = \forall \alpha. \alpha \to \alpha$$

$$\text{Gen}_{x: \alpha}(\alpha \to \alpha) = \alpha \to \alpha$$

$$\text{Gen}_{x: \alpha}(\alpha \to \beta) = \forall \beta. \alpha \to \beta$$

$$\text{Gen}_{x: \alpha, y: \beta}(\alpha \to \beta) = \alpha \to \beta$$

Given an expression $e$, the algorithm $W$ returns a monotype $A$ which may contain free type variables. If we wish to obtain the most general polytype for $e$, it suffices to generalize $A$ with respect to the given typing context. Specifically, if $W(\Gamma, e) = (S, A)$ holds, Theorem 1.13 justifies $S \cdot \Gamma \vdash e : A$, which in turn justifies $S \cdot \Gamma \vdash e : \text{Gen}_{S \cdot \Gamma}(A)$. Hence we may take $\text{Gen}_{S \cdot \Gamma}(A)$ as the most general type for $e$ under typing context $\Gamma$, although we do not formally prove this property (called the completeness of $W$) here.