Type Inference

Viktor Kunčak

Type inference

In a statically typed language, when a program expression type checks, we can usually assign a type to each of its parts.

- for some parts the type is given directly: types of symbols (propagated from declarations using type environment Γ)
- for other parts, the type is **inferred** so that the entire expression type checks

We can consider different type inference algorithms, even for the same type system.

Two examples:

- bottom up propagation of types
- Hindley-Milner constraint-based type inference

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose    > 1)      { print(s      )      }
  else { print(".")    })
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose : Int > 1)      { print(s : String)      }
  else { print(".")    })
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose : Int  > 1) : Bool  { print(s : String)      }
  else { print(".")    })
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose : Int  > 1) : Bool  { print(s : String) : Unit  }
  else { print(".")    })
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose : Int  > 1) : Bool  { print(s : String ) : Unit  }
  else { print(".") : Unit })
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int)      = {
  (if (verbose : Int  > 1) : Bool  { print(s : String ) : Unit  }
  else { print(".") : Unit }) : Unit
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int): Unit = {
  (if (verbose: Int > 1): Bool { print(s: String): Unit }
  else { print("."): Unit}): Unit
}
```

Bottom up inference example

Require types of parameters to be declared
Apply type system rules to compute the types of a tree node from the types of children

Gives recursive algorithm to compute types for all tree nodes (starting from leaves)

```
def message(s: String, verbose: Int): Unit = {
  (if (verbose: Int > 1): Bool { print(s: String): Unit }
  else { print("."): Unit}): Unit
}
```

Inferred types for sub-expressions and established that the program type checks.

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s   , verbose   )    =
  (if (verbose    > 1)    { print(s   )    }
  else { print(".")   })
```

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer ($\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2$)

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer ($\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2$)
- generate constraints (equalities) between variables, according to type system rules

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = Int$, $Int = Int$, $\tau_c = Bool$ (from _ > _)

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)
  - $\tau_c = \mathit{Bool}$, $\tau_2 = \tau_1$, $\tau_0 = \tau_1$ (from **if**)

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)
  - $\tau_c = \mathit{Bool}$, $\tau_2 = \tau_1$, $\tau_0 = \tau_1$ (from **if**)
- solve constraints. Here, eliminate variables using "defining" equations:

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)
  - $\tau_c = \mathit{Bool}$, $\tau_2 = \tau_1$, $\tau_0 = \tau_1$ (from **if**)
- solve constraints. Here, eliminate variables using "defining" equations:
  - $\tau_v = \mathit{Int}$, $\tau_c = \mathit{Bool}$, $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$, $\tau_2 = \mathit{Unit}$

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)
  - $\tau_c = \mathit{Bool}$, $\tau_2 = \tau_1$, $\tau_0 = \tau_1$ (from **if**)
- solve constraints. Here, eliminate variables using "defining" equations:
  - $\tau_v = \mathit{Int}$, $\tau_c = \mathit{Bool}$, $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$, $\tau_2 = \mathit{Unit}$
    substituting τ_1 in $\tau_0 = \tau_1$ gives: $\tau_0 = \mathit{Unit}$

Constraint-based inference example (primitive types)

- start with a program (may or may not have type annotations)

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- assign type variables to denote types we need to infer $(\tau_s, \tau_v, \tau_0, \tau_c, \tau_1, \tau_2)$
- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = \mathit{Int}$, $\mathit{Int} = \mathit{Int}$, $\tau_c = \mathit{Bool}$ (from _ > _)
  - $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$ (from first print)
  - $\mathit{String} = \mathit{String}$, $\tau_2 = \mathit{Unit}$ (from second print)
  - $\tau_c = \mathit{Bool}$, $\tau_2 = \tau_1$, $\tau_0 = \tau_1$ (from **if**)
- solve constraints. Here, eliminate variables using "defining" equations:
  - $\tau_v = \mathit{Int}$, $\tau_c = \mathit{Bool}$, $\tau_s = \mathit{String}$, $\tau_1 = \mathit{Unit}$, $\tau_2 = \mathit{Unit}$
    substituting τ_1 in $\tau_0 = \tau_1$ gives: $\tau_0 = \mathit{Unit}$
- insert the inferred types into the syntax tree

From a type system rule to a constraint

From a type system rule to a constraint

```
def message(s:τs, verbose:τv):τ0 =
  (if (verbose:τv > 1):τc { print(s:τs):τ1 }
  else { print("."):τ2}):τ0
```

- generate constraints (equalities) between variables, according to type system rules
  - $\tau_v = Int$, $Int = Int$, $\tau_c = Bool$ (from _ > _)
  - $\tau_s = String$, $\tau_1 = Unit$ (from first print)
  - $String = String$, $\tau_2 = Unit$ (from second print)
  - $\boxed{\tau_c = Bool,\ \tau_2 = \tau_1,\ \tau_0 = \tau_1}$ (from **if**)

$$\frac{\vdash b : Bool \quad \vdash t_1 : \tau \quad \vdash t_2 : \tau}{\vdash (\textbf{if}\ (b)\ t_1\ \textbf{else}\ t_2) : \tau} \rightsquigarrow \boxed{\frac{\vdash b : \tau_c \quad \vdash t_1 : \tau_1 \quad \vdash t_2 : \tau_2}{\vdash (\textbf{if}\ (b)\ t_1\ \textbf{else}\ t_2) : \tau_0} \tau_c = Bool, \tau_2 = \tau_1, \tau_0 = \tau_1}$$

Hindley-Milner type inference overview

Part of type systems of languages such as Haskell, ML, ocaml.
Supports not only primitive types, but also generic structured types such as:
$Function[\tau_A, \tau_B], Pair[\tau_A, \tau_B], List[\tau_A]$.
Type inference:

1. Use **type variables** (e.g. τ_v, τ_s) to denote unknown types
2. Use type checking rules to derive **constraints** among type variables (e.g., arguments have expected types)
3. Use a **unification algorithm** to solve the constraints
   $List[\tau_A] = List[Int], \quad Pair[Int, \tau_B] = Pair[\tau_A, Bool]$

Programs can often be as concise as in a dynamically typed language.

Type inference still catches meaningless programs: if the equations have no solution so the compiler reports a type error.

Small language with tuples and functions

Types are:

1. primitive types: Int, Bool, String, Unit
2. type constructors:
   - Pair[A,B] or (A,B) denotes set of pairs
   - Function[A,B] or $A \Rightarrow B$ denotes functions from A to B

Abstract syntax of types:

$$t := Int \mid Bool \mid String \mid Unit \mid (t_1, t_2) \mid (t_1 \Rightarrow t_2)$$

Terms include pairs and anonymous functions (x denotes variables, c literals):

$$t := x \mid c \mid f(t_1, \ldots, t_n) \mid \textbf{if } (t) \ t_1 \textbf{ else } t_2 \mid (t_1, t_2) \mid (x \Rightarrow t)$$

Primitives P1,P2 for pair components, if $t = (x, y)$ then $P1(t) = x$, $P2(t) = y$.
We write them as in Scala: $t.\_1 = P1(t)$ and $t.\_2 = P2(t)$
For values and types, (x, y, z) is shorthand for $(x, (y, z))$

Type rules

Rule for **if**:

$$\frac{\Gamma \vdash b : \mathit{Bool} \quad \Gamma \vdash t_1 : \tau \quad \Gamma \vdash t_2 : \tau}{\Gamma \vdash (\textbf{if}\ (b)\ t_1\ \textbf{else}\ t_2) : \tau}$$

Rules for variables:

$$\frac{\Gamma(x) = \tau}{\Gamma \vdash x : \tau}$$

Rules for constants:

$$\frac{}{\text{"..."} : \mathit{String}} \qquad \frac{}{\mathit{true} : \mathit{Bool}} \qquad \frac{}{\mathit{false} : \mathit{Bool}}$$

Rules for pairs

$$\frac{\Gamma \vdash t_1 : \tau_1 \quad \Gamma \vdash t_2 : \tau_2}{\Gamma \vdash (t_1, t_2) : (\tau_1, \tau_2)}$$

If the first component t_1 has type τ_1 and the second component t_2 has type τ_2 then the pair (t_1, t_2) has the type (τ_1, τ_2).

$$\frac{\Gamma \vdash t : (\tau_1, \tau_2)}{\Gamma \vdash t.\_1 : \tau_1}$$

$$\frac{\Gamma \vdash t : (\tau_1, \tau_2)}{\Gamma \vdash t.\_2 : \tau_2}$$

Functions of one argument

$$\frac{\Gamma \vdash f : \tau \Rightarrow \tau_0 \quad \Gamma \vdash t : \tau}{\Gamma \vdash f(t) : \tau_0}$$

Functions of one argument

$$\frac{\Gamma \vdash f : \tau \Rightarrow \tau_0 \quad \Gamma \vdash t : \tau}{\Gamma \vdash f(t) : \tau_0}$$

Why give rule for only one argument?

Functions of one argument

$$\frac{\Gamma \vdash f : \tau \Rightarrow \tau_0 \quad \Gamma \vdash t : \tau}{\Gamma \vdash f(t) : \tau_0}$$

Why give rule for only one argument?

Note that τ can be a tuple $(\tau_1, \ldots, \tau_n)$, so we can derive:

$$\frac{\dfrac{\Gamma \vdash t_1 : \tau_1 \ \ldots \ \Gamma \vdash t_n : \tau_n \quad \Gamma \vdash f : (\tau_1, \ldots, \tau_n) \Rightarrow \tau_0}{\Gamma \vdash (t_1, \ldots, t_n) : (\tau_1, \ldots, \tau_n) \quad \Gamma \vdash f : (\tau_1, \ldots, \tau_n) \Rightarrow \tau_0}}{\Gamma \vdash f((t_1, \ldots, t_n)) : \tau_0}$$

Rules for anonymous function

$$\frac{\Gamma[x := \tau_1] \vdash t : \tau_2}{\Gamma \vdash (x \Rightarrow t) : (\tau_1 \Rightarrow \tau_2)}$$

What does this rule say?

Rules for anonymous function

$$\frac{\Gamma[x := \tau_1] \vdash t : \tau_2}{\Gamma \vdash (x \Rightarrow t) : (\tau_1 \Rightarrow \tau_2)}$$

What does this rule say?

Anonymous function $x \Rightarrow t$ (maps value x to t), has a function type $\tau_1 \Rightarrow \tau_2$, where

Rules for anonymous function

$$\frac{\Gamma[x := \tau_1] \vdash t : \tau_2}{\Gamma \vdash (x \Rightarrow t) : (\tau_1 \Rightarrow \tau_2)}$$

What does this rule say?

Anonymous function $x \Rightarrow t$ (maps value x to t), has a function type $\tau_1 \Rightarrow \tau_2$, where τ_1 is the type of x and τ_2 is the type of t.

Rules for anonymous function

$$\frac{\Gamma[x := \tau_1] \vdash t : \tau_2}{\Gamma \vdash (x \Rightarrow t) : (\tau_1 \Rightarrow \tau_2)}$$

What does this rule say?
Anonymous function $x \Rightarrow t$ (maps value x to t), has a function type $\tau_1 \Rightarrow \tau_2$, where τ_1 is the type of x and τ_2 is the type of t.

Within t, there may be uses of x, which has some type τ_1.
This is why Γ is extended with binding of x to τ_1 when type checking t.

Example for type inference

Program without type annotations:

```
def translatorFactory(dx, dy) = {
  p ⇒ (p._1 + dx, p._2 + dy) // returns anonymous function
}
def upTranslator = translatorFactory(0, 100)
def test = upTranslator((3, 5)) // computes (3, 105)
```

Type inference can find types that correspond to this annotated program:

Example for type inference

Program without type annotations:

```
def translatorFactory(dx, dy) = {
  p ⇒ (p._1 + dx, p._2 + dy) // returns anonymous function
}
def upTranslator = translatorFactory(0, 100)
def test = upTranslator((3, 5)) // computes (3, 105)
```

Type inference can find types that correspond to this annotated program:

```
def translatorFactory(dx: Int, dy: Int): (Int,Int) ⇒ (Int,Int) = {
  p ⇒ (p._1 + dx, p._2 + dy) }
def upTranslator : (Int,Int) ⇒ (Int,Int) = translatorFactory(0, 100)
def test: (Int,Int) = upTranslator((3, 5))
```

Are the suggested types in this example correct?

```
def translatorFactory(dx: Int, dy: Int): (Int,Int) ⇒ (Int,Int) = {
  p ⇒ (p._1 + dx, p._2 + dy) }
def upTranslator : (Int,Int) ⇒ (Int,Int) = translatorFactory(0, 100)
def test: (Int,Int) = upTranslator((3, 5))
```

$$\Gamma \vdash p \Rightarrow (p.\_1 + dx, p.\_2 + dy) \quad : \quad (Int, Int) \Rightarrow (Int, Int)$$

From bottom up to Hindley-Milner type inference

```
def translatorFactory(dx: Int, dy: Int): (Int,Int) ⇒ (Int,Int) = {
  p ⇒ (p._1 + dx, p._2 + dy) }
def upTranslator : (Int,Int) ⇒ (Int,Int) = translatorFactory(0, 100)
def test: (Int,Int) = upTranslator((3, 5))
```

Example steps in type checking the body. Let $\Gamma' = \Gamma[p := (Int, Int)]$

$$\dfrac{\dfrac{\dfrac{\dfrac{\Gamma' \vdash p.\_1 : Int \quad \Gamma' \vdash dx : Int}{\Gamma' \vdash (p.\_1 + dx) : Int \quad \ldots}}{\Gamma' \vdash (p.\_1 + dx, p.\_2 + dy) : (Int, Int)}}{\Gamma \vdash p \Rightarrow (p.\_1 + dx, p.\_2 + dy) \;\; : \;\; (Int, Int) \Rightarrow (Int, Int)}}$$

How did type inference discover $dx : Int$? We construct the derivation tree keeping type of dx symbolic until some derivation step tells us what it must be. Here, $+$ expects two integers in $p.\_1 + dx$

Generating Constraints During Type Inference

```
def translatorFactory(dx, dy) = {
  p ⇒ (p._1 + dx, p._2 + dy)
}
```

Let $\Gamma_1 = \Gamma[p := \tau_p]$ where τ_p is to be determined later

$$\frac{\frac{\frac{\frac{\Gamma_1 \vdash p : \tau_p \quad \tau_p = (\tau_3, \tau_4)}{\Gamma_1 \vdash p.\_1 : \tau_3 \quad \Gamma_1 \vdash dx : \tau_{dx} \quad \Gamma_1 \vdash + : (Int, Int) \rightarrow Int}}{\Gamma_1 \vdash p.\_1 + dx \ : \ \tau_1 \quad \tau_3 = Int,\ \tau_{dx} = Int,\ \tau_1 = Int}}{\Gamma_1 \vdash (p.\_1 + dx, p.\_2 + dy) : \tau_r \quad \tau_r = (\tau_1, \tau_2)}}{\Gamma \vdash (p \Rightarrow (p.\_1 + dx, p.\_2 + dy)) : \tau_{fun} \quad \tau_{fun} = \tau_p \Rightarrow \tau_r}$$

Generating Constraints During Type Inference

```
def translatorFactory(dx, dy) = {
  p ⇒ (p._1 + dx, p._2 + dy)
}
```

Let $\Gamma_1 = \Gamma[p := \tau_p]$ where τ_p is to be determined later

$$\frac{\dfrac{\dfrac{\dfrac{\Gamma_1 \vdash p : \tau_p \quad \tau_p = (\tau_3, \tau_4)}{\Gamma_1 \vdash p.\_1 : \tau_3 \quad \Gamma_1 \vdash dx : \tau_{dx} \quad \Gamma_1 \vdash + : (Int, Int) \to Int}}{\Gamma_1 \vdash p.\_1 + dx \ : \ \tau_1 \quad \tau_3 = Int,\ \tau_{dx} = Int,\ \tau_1 = Int}}{\Gamma_1 \vdash (p.\_1 + dx, p.\_2 + dy) : \tau_r \quad \tau_r = (\tau_1, \tau_2)}}{\Gamma \vdash (p \Rightarrow (p.\_1 + dx, p.\_2 + dy)) : \tau_{fun} \quad \tau_{fun} = \tau_p \Rightarrow \tau_r}$$

Analogously, for the second component of the pair, we derive $\tau_2 = Int$, $\tau_4 = Int$ on other branches of the derivation tree.
From these constraints it follows $\tau_p = (Int, Int)$, $\tau_r = (Int, Int)$ and

$$\tau_{fun} \quad = \quad (Int, Int) \Rightarrow (Int, Int)$$

Constraints

Generate fresh type variable for (in principle) each AST node. Collect these constraints:

AST node	node with type vars	constraint
$f(t)$	$(f : \tau_f)(t : \tau) : \tau_0$	$\tau_f = (\tau \Rightarrow \tau_0)$
$x \Rightarrow t$	$((x : \tau_x) \Rightarrow (t : \tau_t)) : \tau_{fun}$	$\tau_{fun} = (\tau_x \Rightarrow \tau_t)$ (x, τ_x) added to Γ' for t
(t_1, t_2)	$(t_1 : \tau_1, t_2 : \tau_2) : \tau$	$\tau = (\tau_1, \tau_2)$
$t.\_1$	$(t : \tau).\_1 : \tau_1$	$\tau = (\tau_1, \tau_2)$ τ_2 is a fresh type variable
$t.\_2$	$(t : \tau).\_2 : \tau_2$	$\tau = (\tau_1, \tau_2)$ τ_1 is a fresh type variable
x	$x : \tau_x$	$\Gamma(x) = \tau_x$
false	*false* $: \tau$	$\tau = Bool$
true	*true* $: \tau$	$\tau = Bool$
k	$k : \tau$	$\tau = Int$
"..."	"..." $: \tau$	$\tau = String$
if $(b : \tau_b)$ $t_1 : \tau_1$ **else** $t_2 : \tau_2$	$) : \tau$	$\tau = \tau_1, \tau = \tau_2, \tau_b = Bool$

Summary of type inference

1. Introduce type variable for each tree node
2. For each tree node use type rules to derive constraints among the type variables
3. Solve the resulting set of equations on type variables

Solving equations on simple types: unification (as in Prolog)

Types in equations have the following syntax:

$$t := \tau \mid \mathit{Int} \mid \mathit{Bool} \mid \mathit{String} \mid \mathit{Unit} \mid (t_1, t_2) \mid (t_1 \Rightarrow t_2)$$

We assume that

- primitive types are disjoint and distinct from pairs and functions
- pairs and functions are always distinct
- two pairs are equal iff their corresponding component types are equal
- two functions are equal iff their argument and result types are equal

Idea: eliminate variables, decompose pair and function equalities.
Algorithm works for any *term algebra* (algebra of syntactic terms)

- Pair[A,B] and Function[A,B] are two distinct binary term constructors
- Int, Bool, String are distinct nullary constructors

Analogy: Solving Equations over Non-negative Integers

Use Gaussian elimination to solve the system of equations:

$$\begin{aligned} x+y+z &= 5 \\ x+2y+z &= 6 \\ 2x+y+2z &= 5 \end{aligned}$$

For example, we can express x and substitute:

$$\begin{aligned} x &= 5-y-z \\ (5-y-z)+2y+z &= 6 \\ 2(5-y-z)+y+z &= 5 \end{aligned} \quad \text{i.e.,} \quad \begin{aligned} x &= 5-y-z \\ y &= 1 \\ y+z &= 5 \end{aligned}$$

Here, $y=1, z=4, x=0$ is unique solution.
There are systems with infinitely many solutions.
There are systems with no solutions.
Over non-negative integers, $x = x+y+1$ has no solutions.

Unification Algorithm

Applies the following rules as long as they change the current set of equations:
(Let x denote a type variable and T a type term.)

Orient: Replace $T = x$ with $x = T$ when x is not a type variable

Delete useless: Remove $T = T$ (both sides syntactically identical)

Eliminate: Given $x = T$ where T does not contain x, replace x with T in all remaining equations

Occurs check: Given $x = T$ where T properly contains x, report clash (no solutions)

Decompose pairs: Replace $(T_1, T_2) = (T_1', T_2')$ with two equations:
$T_1 = T_1'$ and $T_2 = T_2'$.

Decompose functions: Replace $(T_1 \Rightarrow T_2) = (T_1' \Rightarrow T_2')$ with:
$T_1 = T_1'$ and $T_2 = T_2'$.

Decomposition clash (remaining cases): Given equality where two sides start with different constructors report clash (no solution).

Examples: $(T_1, T_2) = (T_1' \Rightarrow T_2')$, $Int = (T_1, T_2)$, $Int = Bool$, $(T_1 \Rightarrow T_2) = String$

Franz Baader, Wayne Snyder: Unification Theory, In Handbook of Automated Reasoning, Chapter 8, Volume 1, MIT Press 2001.

Properties of Unification

Algorithm always terminates.
Running time is linear given the right data structures and with lazy substitution of variables.

If it reports clash it means that equations have no solution (there exist no annotations that make program type check).

Otherwise, the equations have one or more solutions. Note that a variable that appears on left of equation does not appear on the right (else the eliminate rule would apply). Call a variable that only appears on the right a *parameter*.
If there are no parameters, there is exactly one solution. Otherwise, for each assignment of types to parameters we obtain a solution. Moreover, all solutions are obtained by instantiating parameters.
Therefore, the result of the unification algorithm describes all possible ways to assign simple types to the program.

Use the algorithm to infer the type of rightNest

```
def rightNest(t) = {
  (t._1._1, (t._1._2, t._2))
}
def test1 = rightNest(((1, 2), 3)) // computes (1,(2,3))
```

Type variable for each sub-expression (same τ_1 for same expression, to keep it short)

$$\begin{array}{l}\big(\ ((t:\tau).\_1:\tau_1).\_1:\tau_2,\\ \ \ (((t:\tau).\_1:\tau_1).\_2:\tau_3,(t:\tau).\_2:\tau_4):\tau_5\ \big):\tau_6\end{array}$$

$$\left|\begin{array}{l}\tau=(\tau_1,\tau_{10})\\ \tau_1=(\tau_2,\tau_{20})\\ \tau=(\tau_1,\tau_{30})\\ \tau_1=(\tau_{40},\tau_3)\\ \tau=(\tau_{50},\tau_4)\\ \tau_5=(\tau_3,\tau_4)\\ \tau_6=(\tau_2,\tau_5)\end{array}\right| \Rightarrow \left|\begin{array}{l}\tau=(\tau_1,\tau_{10})\\ \tau_1=(\tau_2,\tau_{20})\\ (\tau_1,\tau_{10})=(\tau_1,\tau_{30})\\ \tau_1=(\tau_{40},\tau_3)\\ (\tau_1,\tau_{10})=(\tau_{50},\tau_4)\\ \tau_5=(\tau_3,\tau_4)\\ \tau_6=(\tau_2,\tau_5)\end{array}\right| \Rightarrow \left|\begin{array}{l}\tau=(\tau_1,\tau_{10})\\ \tau_1=(\tau_2,\tau_{20})\\ \tau_1=\tau_1,\tau_{10}=\tau_{30}\\ \tau_1=(\tau_{40},\tau_3)\\ (\tau_1,\tau_{10})=(\tau_{50},\tau_4)\\ \tau_5=(\tau_3,\tau_4)\\ \tau_6=(\tau_2,\tau_5)\end{array}\right| \Rightarrow$$

Applying Unification Rules Some More

$$
\left|\begin{array}{l}
\tau = (\tau_1, \tau_{10}) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_{10} = \tau_{30} \\
\tau_1 = (\tau_{40}, \tau_3) \\
(\tau_1, \tau_{10}) = (\tau_{50}, \tau_4) \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
\left|\begin{array}{l}
\tau = (\tau_1, \tau_{10}) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_{10} = \tau_{30} \\
\tau_1 = (\tau_{40}, \tau_3) \\
\tau_1 = \tau_{50}, \tau_{10} = \tau_4 \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
\left|\begin{array}{l}
\tau = (\tau_1, \tau_4) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_4 = \tau_{30} \\
\tau_1 = (\tau_{40}, \tau_3) \\
\tau_1 = \tau_{50}, \tau_{10} = \tau_4 \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
$$

$$
\left|\begin{array}{l}
\tau = (\tau_1, \tau_4) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_{30} = \tau_4 \\
\tau_1 = (\tau_{40}, \tau_3) \\
\tau_{50} = \tau_1, \tau_{10} = \tau_4 \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
\left|\begin{array}{l}
\tau = ((\tau_2, \tau_{20}), \tau_4) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_{30} = \tau_4 \\
(\tau_2, \tau_{20}) = (\tau_{40}, \tau_3) \\
\tau_{50} = (\tau_2, \tau_{20}), \tau_{10} = \tau_4 \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
\left|\begin{array}{l}
\tau = ((\tau_2, \tau_{20}), \tau_4) \\
\tau_1 = (\tau_2, \tau_{20}) \\
\tau_{30} = \tau_4 \\
\tau_2 = \tau_{40}, \tau_{20} = \tau_3 \\
\tau_{50} = (\tau_2, \tau_{20}), \tau_{10} = \tau_4 \\
\tau_5 = (\tau_3, \tau_4) \\
\tau_6 = (\tau_2, \tau_5)
\end{array}\right|
\Rightarrow
$$

And More

$$\left|\begin{array}{l}\tau = ((\tau_2, \tau_3), \tau_4)\\ \tau_1 = (\tau_2, \tau_3)\\ \tau_{30} = \tau_4\\ \tau_2 = \tau_{40}, \tau_{20} = \tau_3\\ \tau_{50} = (\tau_2, \tau_3), \tau_{10} = \tau_4\\ \tau_5 = (\tau_3, \tau_4)\\ \tau_6 = (\tau_2, \tau_5)\end{array}\right| \Rightarrow \left|\begin{array}{l}\tau = ((\tau_2, \tau_3), \tau_4)\\ \tau_1 = (\tau_2, \tau_3)\\ \tau_{30} = \tau_4\\ \tau_{40} = \tau_2, \tau_{20} = \tau_3\\ \tau_{50} = (\tau_2, \tau_3), \tau_{10} = \tau_4\\ \tau_5 = (\tau_3, \tau_4)\\ \tau_6 = (\tau_2, \tau_5)\end{array}\right| \Rightarrow \left|\begin{array}{l}\tau = ((\tau_2, \tau_3), \tau_4)\\ \tau_1 = (\tau_2, \tau_3)\\ \tau_{30} = \tau_4\\ \tau_{40} = \tau_2, \tau_{20} = \tau_3\\ \tau_{50} = (\tau_2, \tau_3), \tau_{10} = \tau_4\\ \tau_5 = (\tau_3, \tau_4)\\ \tau_6 = (\tau_2, (\tau_3, \tau_4))\end{array}\right|$$

No more rule applies. Variables on

- right-hand sides: τ_2, τ_3, τ_4
- left-hand sides: all others

The argument type is $\tau = ((\tau_2, \tau_3), \tau_4)$
The result type is $\tau_6 = (\tau_2, (\tau_3, \tau_4))$
So, rightNest has type $((\tau_2, \tau_3), \tau_4) \to (\tau_2, (\tau_3, \tau_4))$
The types τ_2, τ_3, τ_4 can be picked arbitrarily—there are infinitely many solutions.

Adding Constraints for Function Call

We have:

$$\text{rightNest} : ((\tau_2, \tau_3), \tau_4) \Rightarrow (\tau_2, (\tau_3, \tau_4))$$

Given a call rightNest(((1, 2), 3)), we add constraints equivalent to

$$(\tau_2, \tau_3), \tau_4) = ((\mathit{Int}, \mathit{Int}), \mathit{Int})$$

Thus we conclude $\tau_2 = \mathit{Int}, \tau_3 = \mathit{Int}, \tau_4 = \mathit{Int}$. Given that

$$\mathit{rightNest}(((1,2),3)) : (\tau_2, (\tau_3, \tau_4))$$

we conclude

$$\mathit{rightNest}(((1,2),3)) : (\mathit{Int}, (\mathit{Int}, \mathit{Int}))$$

What happens in this case?

```
def rightNest(t) = {
  (t._1._1, (t._1._2, t._2))
}
def test1 = rightNest(((1, 2), 3))
def test2 = rightNest((false , true), false)
```

$$(\tau_2, \tau_3), \tau_4) = ((Int, Int), Int) \quad \text{because of test1}$$
$$(\tau_2, \tau_3), \tau_4) = ((Bool, Bool), Bool) \quad \text{because of test2}$$

which implies $Int = Bool$ and is contradictory.
Program fails to type check because the argument type of t becomes equal to both Int and Bool, which is inconsistent.

This is a pity, because we could copy rightNest into rightNest2 with the same body as rightNest, then call rightNest2((false, true), false), and everything would work. But the new program executes the same as old.

More flexibility through generalization

```
def rightNest(t) = {
  (t._1._1, (t._1._2, t._2))
}
def test1 = rightNest(((1, 2), 3))
def test2 = rightNest((false , true), false)
```

After completing the inference for rightNest, first generalize its free type variables into a variable schema:

$$\forall a, b, c.\ ((a, b), c)) \rightarrow (a, (b, c))$$

Then, each time we use the function, replace quantified variables with fresh variables.
Use in test1:

$$((a_1, b_1), c_1)) \rightarrow (a_1, (b_1, c_1))$$

$a_1 = Int,\ b_1 = Int,\ c_1 = Int$
Use in test2:

$$((a_2, b_2), c_2)) \rightarrow (a_2, (b_2, c_2))$$

$a_2 = Bool,\ b_2 = Bool,\ c_2 = Bool$

More flexibility through generalization

```
def rightNest(t) = {
  (t._1._1, (t._1._2, t._2))
}
def test1 = rightNest(((1, 2), 3))
def test2 = rightNest((false , true), false)
```

With this new approach, the program type checks and its types are inferred as follows:

```
def rightNest[A,B,C](t : ((A, B), C)) : (A, (B, C)) = {
  (t._1._1, (t._1._2, t._2))
}

def test1 : (Int, (Int, Int)) =
  rightNest[Int, Int, Int](((1, 2), 3))

def test2 : (Bool, (Bool, Bool))=
  rightNest[Bool,Bool, Bool]((false , true), false)
```