1 Reduction Strategies

In general there may be many possible β-reductions that can be performed on a given λ-term. How do we choose which one to perform next? Does it matter?

A specification that tells which of the possible β-reductions to perform next is called a *reduction strategy*. The λ-calculus does not specify a reduction strategy; it is *nondeterministic*. A reduction strategy is needed in real programming languages to resolve the nondeterminism.

Two common reduction strategies for the λ-calculus are *normal order* and *applicative order*. Under the normal order reduction strategy, the leftmost-outermost redex is always the next to be reduced. By leftmost-outermost, we mean that if e_1 and e_2 are redexes in a term and e_1 is a subterm of e_2, then e_1 will not be reduced next; and among those redexes that are not subterms of other redexes, which are all pairwise incomparable with respect to the subterm relation, the leftmost one is chosen for reduction. It is known that if a term has a normal form at all, then normal order reduction will converge to it.

The applicative order reduction strategy is similar, except that the leftmost-innermost redex is chosen. That is, if e_1 and e_2 are redexes in a term and e_1 is a subterm of e_2, then e_2 will not be reduced next; and among those redexes that do not contain other redexes as subterms, which are all pairwise incomparable with respect to the subterm relation, the leftmost one is chosen for reduction.

In real functional programming languages, reductions inside the body of a λ-abstraction are usually not performed (although optimizing compilers may do so in some instances). If we restrict the normal order and applicative order strategies so as not to perform reductions inside the body of λ-abstractions, we obtain strategies known as *call-by-name* (CBN) and *call-by-value* (CBV), respectively.

Most functional programming languages use CBV, with the notable exception of Haskell. Let us define a *value* to be a λ-term for which no β-reductions are possible, given our chosen reduction strategy. For example, $\lambda x.\, x$ would always be a value, whereas $(\lambda x.\, x)\, 1$ would most likely not be.

Under CBV, functions may only be called on values; that is, the arguments must be fully evaluated. Thus the β-reduction step $(\lambda x.\, e_1)\, e_2 \xrightarrow{1} e_1\{e_2/x\}$ only applies if e_2 is a value. Here is an example of a CBV evaluation sequence, where we consider 3 and inc (the successor function) to be primitive constants.

$$(\lambda x.\, \mathsf{inc}\, x)\, ((\lambda y.\, \mathsf{inc}\, y)\, 3) \xrightarrow{1} (\lambda x.\, \mathsf{inc}\, x)\, (\mathsf{inc}\, 3) \xrightarrow{1} (\lambda x.\, \mathsf{inc}\, x)\, 4 \xrightarrow{1} \mathsf{inc}\, 4 \xrightarrow{1} 5.$$

An alternative strategy is CBN. Under CBN, we defer evaluation of arguments until as late as possible, applying reductions from left to right within the expression. Here is the same term evaluated under CBN.

$$(\lambda x.\, \mathsf{inc}\, x)\, ((\lambda y.\, \mathsf{inc}\, y)\, 3) \xrightarrow{1} \mathsf{inc}\, ((\lambda y.\, \mathsf{inc}\, y)\, 3) \xrightarrow{1} \mathsf{inc}\, (\mathsf{inc}\, 3) \xrightarrow{1} \mathsf{inc}\, 4 \xrightarrow{1} 5.$$

This is the preferred strategy of the language Haskell. Another way to view this is as a form of *lazy evaluation*; the arguments to a function are not evaluated until they are actually needed.

2 Structured Operational Semantics (SOS)

Let us formalize CBV for the pure λ-calculus. First, we restrict our attention to *closed* λ-terms (those with no free variables). Then the values of the language are simply the closed λ-abstractions:

$$v \;::=\; \lambda x.\, e$$

The use of this BNF definition specifies that the metavariable v stands for a value; in this case, a closed λ-abstraction.

Next, we can write *inference rules* to specify when reductions are allowed:

$$\frac{}{(\lambda x.\, e)\, v \xrightarrow{1} e\{v/x\}} \qquad \frac{e_1 \xrightarrow{1} e_1'}{e_1 e_2 \xrightarrow{1} e_1' e_2} \qquad \frac{e \xrightarrow{1} e'}{v\, e \xrightarrow{1} v\, e'} \tag{1}$$

This is a simple operational semantics for a programming language based on the λ-calculus. An operational semantics is a language semantics that describes how to run the program. This can be done through informal human-language text, as in the Java Language Specification [1], or through more formal rules, as we have done here.

The leftmost rule of (1) is just β-reduction. But by the use of the metavariable v for the argument of the function, we have indicated that the rule may only be applied when the argument is a value. The second rule says that $e_1 e_2$ reduces to $e_1' e_2$ in one step provided e_1 reduces to e_1' in one step. The rightmost rule says that $v\, e$ reduces to $v\, e'$ in one step provided e reduces to e' in one step and v is already reduced.

Rules of the form (1) are known as a Structural Operational Semantics (SOS). They define evaluation as the result of applying the rules to transform the expression. The rules are typically inductive on the structure of the expression being evaluated.

As defined above, CBV evaluation is *deterministic*: there is at most one evaluation rule that applies in any situation (we will prove this later).

This kind of operational semantics is known as a *small-step* semantics because it describes only one step at a time. An alternative is a *big-step* semantics that describes the entire evaluation of the program to a final value.

We will see other kinds of semantics later in the course, such as *axiomatic semantics*, which describes the behavior of a program in terms of the observable properties of the input and output states, and *denotational semantics*, which associates a mathematical object with each each program (its *extension*).

CBN has slightly simpler rules:

$$\frac{}{(\lambda x.\, e_1)\, e_2 \xrightarrow{1} e_1\{e_2/x\}} \qquad \frac{e_0 \xrightarrow{1} e_0'}{e_0 e_1 \xrightarrow{1} e_0' e_1}$$

We don't need the rule for evaluating the right-hand side of an application because β-reductions are performed immediately once the left-hand side is a value.

What happens if we try using Ω as a parameter? It depends on the evaluation strategy. Consider

$$(\lambda x.\, \lambda y.\, y)\, \Omega$$

Using the CBV evaluation strategy, we must first reduce Ω. This puts the evaluator into an infinite loop. On the other hand, CBN reduces the term above to $\lambda y.\, y$. CBN has an important property: CBN will not loop infinitely unless every other semantics would also loop infinitely, yet it agrees with CBV whenever CBV terminates successfully.

2.1 Other Reduction Strategies

As mentioned above, in *normal order*, the leftmost-outermost redex is reduced first. This is closely related to CBN evaluation, but also allows reductions in the body of a λ-term. Like CBN, it finds a value if one exists,

albeit not necessarily in the most efficient way. Call-by-value (CBV) is correspondingly related to *applicative order*, where the argument to a function must be reduced to a value before the function is applied.

In the programming language C, the order of evaluation of arguments is not defined by the language; it is implementation-specific. Because of this and the fact that C has side effects, C is not confluent. For example, the value of the expression $(x = 1) + x$ is 2 if the left operand of $+$ is evaluated first, $x + 1$ if the right operand is evaluated first. This makes writing correct C programs more challenging!

The absence of confluence in concurrent imperative languages is one reason that concurrent programming is difficult. In the λ-calculus, confluence guarantees that reductions can be done in parallel without fear of changing the result.

3 Term Equivalence

When should two terms be considered equal? This question is not as simple as it may seem. The strictest definition of equality is syntactic identity, but this is not very interesting or useful. For example, it seems clear that $\lambda x.\, x$ and $\lambda y.\, y$ should be considered equal, as the parameter name is inconsequential. So we might declare two terms equal if they are syntactically identical modulo α-renaming. This is a reasonable definition if we wish to regard λ-terms as *intensional* objects.

As *extensional* objects, however, it does not go far enough. Ideally, we would like to consider two terms equal if they represent the same function. The terms $\lambda x.\, x$ and $\lambda y.\, y$ certainly represent the same function (the identity), but there are others; for example, $\lambda x.\, (\lambda y.\, y)\, x$. So terms do not have to be α-equivalent to represent the same function.

It would be nice if we could just say that two terms are equivalent if they give equivalent results on equivalent inputs. Unfortunately, this is a circular statement, so it doesn't define anything! It is not even clear that there is a "right" definition.

Another complication is undecidability. Any reasonable notion of extensional equivalence is likely to be undecidable due to the relationship between the λ-calculus and Turing machines. If we could test equivalence, then we could test equivalence with Ω, which is tantamount to solving the halting problem.

A commonly used notion of intensional equivalence is *β-equivalence*. Recall that two λ-terms e_1 and e_2 are called *α-equivalent* (notation: $e_1 =_\alpha e_2$) if one can be derived from the other simply by renaming bound variables. Let us call e_1 and e_2 *β-equivalent* and write $e_1 =_\beta e_2$ if either (i) they have a common normal form up to α-equivalence, or (ii) neither has a normal form; that is, either (i) they both converge to α-equivalent normal forms under some sequence of β-reductions, or (ii) neither converges under any sequence of β-reductions. By confluence, $=_\beta$ is an equivalence relation. This notion of equivalence is useful for some compiler optimizations and for checking type equality in some advanced type systems. Unfortunately, it would not work for reduction strategies like CBN and CBV, which do not allow reductions inside the bodies of λ-abstractions.

3.1 Contexts and Observational Equivalence

Another useful notion of equivalence is *observational equivalence*. Intuitively, two terms are observationally equivalent if they behave indistinguishably in any possible context. But what do we mean by "behave indistinguishably," and what is a "context"?

For simplicity, let us assume that we are working with an evaluation strategy such as CBV or CBN that is *deterministic*, which means that there is at most one next β-reduction that can be performed. We say that

a term e *terminates* or *converges* if there is a finite sequence of reductions

$$e \;\to\; e' \;\to\; e'' \;\to\; \cdots \;\to\; v$$

leading to a value v. We write $e \Downarrow v$ when this happens, and we write $e \Downarrow$ when $e \Downarrow v$ for some v. The other possibility is that it keeps on reducing forever without ever arriving at a value. When this happens, we say that e *diverges* and write $e \Uparrow$. Because we have assumed that we are using a deterministic evaluation strategy, exactly one of these two cases will occur.

With CBN or CBV, there are infinitely many divergent terms. One example is $\Omega = (\lambda x.\, xx)\,(\lambda x.\, xx)$. We might consider all divergent terms equivalent, since none of them produce a value.

While we may not have a precise definition of extensional equivalence yet, we can postulate a desirable property: two equivalent terms, when placed in the same context, should either both diverge or both converge and give indistinguishable values. Here a *context* is any term $C[\cdot]$ with a single occurrence of a distinguished special variable, called the *hole*, and $C[e]$ denotes the context $C[\cdot]$ with the hole replaced by the term e. This notion of equivalence is called *observational equivalence*.

More formally, suppose we already have a notion of equivalence $\equiv$ on values. Then we will say that two terms are *observationally equivalent* (with respect to $\equiv$) and write $e_1 \equiv_{\text{obs}} e_2$ if for all contexts $C[\cdot]$, either

- $C[e_1] \Uparrow$ and $C[e_2] \Uparrow$; or
- $C[e_1] \Downarrow v_1$, $C[e_2] \Downarrow v_2$, and $v_1 \equiv v_2$.

In other words, either both $C[e_1]$ and $C[e_2]$ diverge, or both converge and produce $\equiv$-equivalent values.

Note that on values themselves, equivalence is not necessarily the same as observational equivalence. Certainly two values that are observationally equivalent are equivalent in the sense of $\equiv$, because we could put them in the trivial context consisting of just the hole. However, the converse is not true in general. Is it possible to have $\equiv_{\text{obs}}$ and $\equiv$ coincide on values? In other words, does there exist a *fixpoint* of the transformation $\equiv \;\mapsto\; \equiv_{\text{obs}}$? If so, is it unique? Even if not, is there a reasonable choice for the definition of extensional equivalence?

The answers to these questions lie in the following facts, none of which are difficult to prove. We leave them as exercises.

Lemma 4.1. *Let $\equiv$ be an arbitrary equivalence relation on values and let $\equiv_{\text{obs}}$ be the relation of observational equivalence on terms determined by $\equiv$.*

(i) *The relation $\equiv_{\text{obs}}$ is an equivalence relation on terms.*

(ii) *Restricted to values, $\equiv_{\text{obs}}$* refines *$\equiv$; that is, viewed as sets of ordered pairs, $\equiv_{\text{obs}}$ restricted to values is a subset of $\equiv$. Thus for any values v_1 and v_2, if $v_1 \equiv_{\text{obs}} v_2$, then $v_1 \equiv v_2$.*

(iii) *The transformation $\equiv \;\mapsto\; \equiv_{\text{obs}}$ is* monotone *with respect to the refinement relation. That is, if $\equiv^1$ refines $\equiv^2$, then $\equiv^1_{\text{obs}}$ refines $\equiv^2_{\text{obs}}$.*

It turns out that there can be several fixpoints of the transformation $\equiv \;\mapsto\; \equiv_{\text{obs}}$, depending on the reduction strategy. For CBV and CBN, there is a *coarsest* one that is refined by every other fixpoint: define

$$e_1 \equiv_{\Downarrow} e_2 \quad \triangleq \quad \text{for all contexts } C[\cdot],\ C[e_1] \Downarrow \text{ iff } C[e_2] \Downarrow.$$

Theorem 4.2. *For CBV and CBN, the relation $\equiv_{\Downarrow}$ is a fixpoint of the transformation $\equiv \;\mapsto\; \equiv_{\text{obs}}$; that is, $\equiv_{\Downarrow} = (\equiv_{\Downarrow})_{\text{obs}}$. Moreover, it is the coarsest such fixpoint.*

The relation $\equiv_{\Downarrow}$ may be a reasonable candidate for extensional equivalence. By definition, to check that e_1 and e_2 are observationally equivalent, it is enough to check that e_1 and e_2 both converge or both diverge in any context; it is unnecessary to compare the resulting values in the case of convergence. This is because if the values are not equivalent, one can devise a context in which one converges and the other diverges.

4 β-Equivalence

Returning to our notion of β-equivalence, we have

Theorem 4.3. *Allowing arbitrary β-reductions, the following are equivalent:*

(i) $e_1\ (=_\alpha)_{\text{obs}}\ e_2$.

(ii) $e_1\ (=_\beta)_{\text{obs}}\ e_2$.

(iii) *For all contexts $C[\cdot]$, $C[e_1] =_\beta C[e_2]$.*

Proof. The equivalence of (i) and (ii) is immediate from the definition, since $=_\alpha$ and $=_\beta$ agree on values. The equivalence of (ii) and (iii) is simply the definition of $(=_\beta)_{\text{obs}}$. □

Theorem 4.4. *$=_\beta$ is a fixpoint of the monotone map $\equiv \mapsto \equiv_{\text{obs}}$ on values.*

Proof. We know that $=_\alpha$ and $=_\beta$ agree on values, and it follows from Theorem 4.3 that $(=_\alpha)_{\text{obs}}$ and $(=_\beta)_{\text{obs}}$ agree on values as well. From Lemma 4.1(ii), we have that $(=_\alpha)_{\text{obs}}$ refines $=_\alpha$. Finally, if $v_1 =_\alpha v_2$, then for all contexts $C[\cdot]$, $C[v_1] =_\alpha C[v_2]$, therefore $C[v_1] =_\beta C[v_2]$ since we can α-convert between v_1 and v_2 at any time. By Theorem 4.3, $v_1\ (=_\alpha)_{\text{obs}}\ v_2$. Thus all four relations $=_\alpha$, $=_\beta$, $(=_\alpha)_{\text{obs}}$, and $(=_\beta)_{\text{obs}}$ agree on values. □

References

[1] James Gosling, Bill Joy, Jr. Guy L. Steele, and Gilad Bracha. *The Java Language Specification*. Prentice Hall, 3rd edition, 2005.