

Symbolic Execution Semantics for WHILE Language

Arie Gurfinkel

January 23, 2018

1 Operational Semantics

Let L be the set of program variables. A *concrete state* q (also called an *environment*) is a map from program variables L to integers. We write $[\]$ for an empty state, i.e., a state in which no variables are defined. We write $[x := u, y := v]$ for a state in which x has value u and y has value v . Given a state q , a variable x and a value n , we write $q(x)$ for the value of x in q , and $q[x := n]$ for a state obtained from q by replacing the value of x by n . Formally,

$$q[x := v](w) = \begin{cases} v & \text{if } x = w \\ q(w) & \text{otherwise} \end{cases} \quad (1)$$

Operational semantics for the WHILE language is shown in Fig. 1. The judgement has the form $\langle s, q \rangle \Downarrow q'$, where s is a statement, q and q' are the input and output states, respectively. All inference rules in Fig. 1 are deterministic – each input state has only one legal output state. The only exception is the rule for the **havoc** statement.

2 Symbolic Execution Semantics

A *symbolic state* (or symbolic environment) q is a map from program variables to symbolic expressions. A *path condition* is a formula over symbolic expressions. Symbolic execution semantics of the WHILE language are shown in Fig. 2. A judgement in symbolic execution has a form $\langle s, q, pc \rangle \Downarrow q', pc'$, where s is a statement, q and q' are the input and output symbolic environments, respectively, and pc and pc' are input and output path conditions, respectively. Note that since WHILE language does not have inputs, symbolic expressions are introduced into the state by the **havoc** statement.

Unlike the concrete operational semantics in Fig. 1, the rules of symbolic semantics are non-deterministic. A statement might have several legal executions in a given symbolic state q . For example, both branches of an if-statement can be executed if both the condition b and its negation $\neg b$ are consistent with the current path condition pc .

$$\begin{array}{c}
\overline{\langle \mathbf{skip}, q \rangle \Downarrow q} \\
\frac{\langle s_1, q \rangle \Downarrow q'' \quad \langle s_2, q'' \rangle \Downarrow q'}{\langle s_1 ; s_2, q \rangle \Downarrow q'} \\
\frac{\langle b, q \rangle \Downarrow \mathbf{true} \quad \langle s_1, q \rangle \Downarrow q'}{\langle \mathbf{if } b \mathbf{ then } s_1 \mathbf{ else } s_2, q \rangle \Downarrow q'} \\
\frac{\langle b, q \rangle \Downarrow \mathbf{false}}{\langle \mathbf{while } b \mathbf{ do } s, q \rangle \Downarrow q} \\
\overline{\langle \mathbf{havoc } x, q \rangle \Downarrow q[x := n]}
\end{array}
\qquad
\begin{array}{c}
\overline{\langle \mathbf{print_state}, q \rangle \Downarrow q} \\
\frac{\langle e, q \rangle \Downarrow n}{\langle x := e, q \rangle \Downarrow q[x := n]} \\
\frac{\langle b, q \rangle \Downarrow \mathbf{false} \quad \langle s_2, q \rangle \Downarrow q'}{\langle \mathbf{if } b \mathbf{ then } s_1 \mathbf{ else } s_2, q \rangle \Downarrow q'} \\
\frac{\langle b, q \rangle \Downarrow \mathbf{true} \quad \langle s ; \mathbf{while } b \mathbf{ do } s, q \rangle \Downarrow q'}{\langle \mathbf{while } b \mathbf{ do } s, q \rangle \Downarrow q'}
\end{array}$$

Figure 1: Operational semantics for the statements of the WHILE language.

3 Operational Semantics of Concolic Execution

In concolic execution, the set of program variables L is partitioned into symbolic, $Sym(L)$, and concrete, $Con(L)$, variables. It is possible that all variables are symbolic, i.e., $Con(L) = \emptyset$ or that all variables are concrete. For a given variable a , we write $Sym(a)$ and $Con(a)$ to indicate that a is symbolic or concrete, respectively.

A state of concolic execution is a triple $q = \langle c, s, pc \rangle$, where c is a concrete state, s a symbolic environment, and pc is a formula called the *path condition*. Given a state $q = \langle c, s, pc \rangle$, we write $con(q)$ for c , $sym(q)$ for s , and $pc(q)$ for pc . Symbolic environment, path condition, and concrete state are as in symbolic and concrete execution, respectively. However, concrete state also has a value for every symbolic variable. We call those concrete variables *concrete shadows*. That is, if b is a symbolic variable and q a concolic state, then $sym(q)(b)$ is the symbolic value of b , and $con(q)(b)$ is the value of the concrete shadow of b . Given two concrete states c_1 and c_2 , we write $c_1 \equiv_{con} c_2$ to indicate that that they are identical on the concrete variables:

$$c_1 \equiv_{con} c_2 \iff \forall a \in Con(L) \cdot c_1(a) = c_2(a)$$

Given a concrete state c and a symbolic state $\langle s, pc \rangle$, we write $c \models \langle s, pc \rangle$ to indicate that the concrete state c is contained in the symbolic state.

The semantics of expressions is as usual with variables evaluated based on their kind: concrete variables are evaluated over $con(q)$ and symbolic over $sym(q)$:

$$\frac{con(a) \quad \langle a, con(q) \rangle \Downarrow v}{\langle a, q \rangle \Downarrow v} \quad \frac{sym(a) \quad \langle a, sym(q) \rangle \Downarrow v}{\langle a, q \rangle \Downarrow v}$$

$$\frac{}{\langle \mathbf{skip}, q, pc \rangle \Downarrow q, pc}$$

$$\frac{}{\langle \mathbf{print_state}, q, pc \rangle \Downarrow q, pc}$$

$$\frac{\langle s_1, q, pc \rangle \Downarrow q'', pc'' \quad \langle s_2, q'', pc'' \rangle \Downarrow q', pc'}{\langle s_1 ; s_2, q, pc \rangle \Downarrow q', pc'}$$

$$\frac{\langle e, q \rangle \Downarrow v}{\langle x := e, q, pc \rangle \Downarrow q[x := v], pc}$$

$$\frac{\langle b, q \rangle \Downarrow v \quad pc \wedge v \text{ is SAT} \quad \langle s_1, q, pc \wedge v \rangle \Downarrow q', pc'}{\langle \mathbf{if } b \text{ then } s_1 \text{ else } s_2, q, pc \rangle \Downarrow q', pc'}$$

$$\frac{\langle b, q \rangle \Downarrow v \quad pc \wedge \neg v \text{ is SAT} \quad \langle s_2, q, pc \wedge \neg v \rangle \Downarrow q', pc'}{\langle \mathbf{if } b \text{ then } s_1 \text{ else } s_2, q, pc \rangle \Downarrow q', pc'}$$

$$\frac{\langle b, q \rangle \Downarrow v \quad pc \wedge \neg v \text{ is SAT}}{\langle \mathbf{while } b \text{ do } s, q, pc \rangle \Downarrow q, pc \wedge \neg v}$$

$$\frac{\langle b, q \rangle \Downarrow v \quad pc \wedge v \text{ is SAT} \quad \langle s ; \mathbf{while } b \text{ do } s, q, pc \wedge v \rangle \Downarrow q', pc'}{\langle \mathbf{while } b \text{ do } s, q, pc \rangle \Downarrow q', pc'}$$

V is a fresh symbolic constant

$$\frac{}{\langle \mathbf{havoc } x, q, pc \rangle \Downarrow q[x := V]}$$

Figure 2: Symbolic Execution Semantics of the WHILE language.

For most statements, the semantics is extended by applying both symbolic and concrete operational semantics in parallel. The last pre-condition ensures that the concrete and symbolic states are chosen consistently.

$$\frac{\langle s, con(q) \rangle \Downarrow c \quad \langle s, sym(q), pc(q) \rangle \Downarrow s', pc' \quad c \models \langle s', pc' \rangle}{\langle s, q \rangle \Downarrow \langle c, s', pc' \rangle}$$

Assignment of values to concrete variables is limited to concrete values only.

$$\frac{\langle e, q \rangle \Downarrow n \quad con(x) \quad n \in \mathbb{Z}}{\langle x := e, q \rangle \Downarrow q[x := n]}$$

Thus, it is not possible to assign symbolic variables (or symbolic expressions) to concrete variables. If necessary, assigning symbolic values to concrete variables

can be done either by treating all variables as symbolic (i.e., $Con(L) = \emptyset$), or concretizing symbolic state before assignment (which we describe later on).

Assignment to symbolic variables also assigns to their concrete shadows.

At if-statement, concolic execution can chose to switch to the branch that is not consistent with current concrete state, as long as the concrete state can be adjusted. We only show one of the cases:

$$\frac{\langle b, q \rangle \Downarrow v \quad pc(q) \wedge \neg v \text{ is SAT} \quad c \models \langle sym(q), pc(q) \wedge \neg v \rangle \quad c \equiv_{con} con(q) \quad \langle s_2, \langle c, sym(q), pc(q) \wedge \neg v \rangle \rangle \Downarrow q'}{\langle \text{if } b \text{ then } s_1 \text{ else } s_2, q \rangle \Downarrow q'}$$

In this case, according to the concrete state the branch condition b is true. At the same time, according to symbolic state, the negation of the branch condition $\neg b$ and the current path condition are satisfiable. The concolic execution can proceed according to the else-branch, but the concolic state needs to be first updated such that:

1. the concrete portion is updated to be consistent with the negation of the branch condition
2. the path condition is extended with the negation of the path condition.

Note that it is only possible to take the else-branch if the condition value can be controlled by the symbolic part of the state. Branch conditions that do not depend on the input (such as iterations of the loops) can only be resolved one way (i.e., either true or false).

Finally, concolic execution semantics provide *concretization* step that allows to turn symbolic variables (or values) to their concrete values in the current concrete state. The effect of concretization is captured by the so called *concretization constraints* in the path condition:

$$\frac{\langle x, sym(q) \rangle \Downarrow v \quad sym(x) \quad \langle x, con(q) \rangle \Downarrow n \quad \langle s, \langle con(q), sym(q)[x := n], pc(q) \wedge v = n \rangle \rangle \Downarrow q'}{\langle s, q \rangle \Downarrow q'}$$

The rule says that if x is a symbolic variable with symbolic value v and it is currently shadowed concretely by a concrete value n , then we can update the symbolic value of x to n as long as we also update the path condition with $v = n$ to reflect the concretization step.