Symbolic Execution(Working title)

Aarhus Universitet

Søren Baadsgaard

May 31, 2019

Abstract

Contents

Chapter 1

Introduction

Chapter 2

Motivation

In this chapter we will present the motivation for this project, by considering a motivating example that illustrates the usefulness of symbolic execution as a software testing technique.

2.1 A motivating example

Consider a company that sells a product with a unit price 2. If the revenue of an order is greater than or equal 16, a discount of 10 is applied. The following program that takes integer inputs *units* and *cost* computes the total revenue based on this pricing scheme.

```
procedure ComputeRevenue($units, cost$)
    $revenue := 2 \cdot units$
    if $revenue \geq 16$ then
        $revenue := revenue - 10$
        assert $revenue \geq cost$
    end if
    return $revenue$
end procedure
```

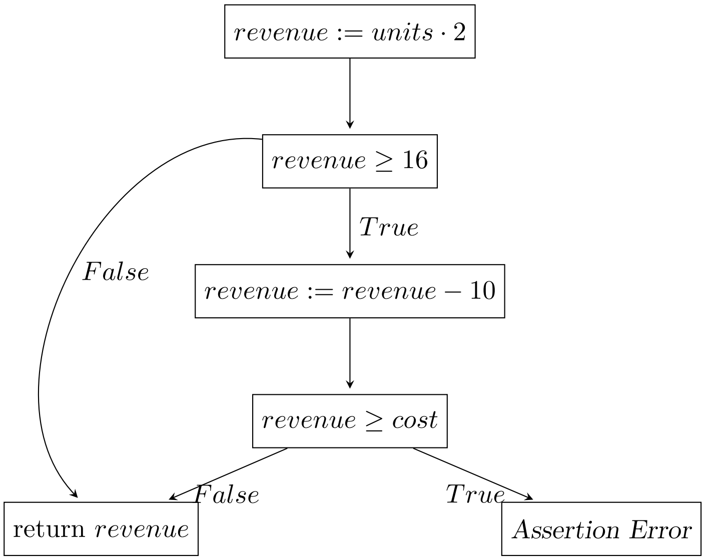

Figure 2.1: Control-flow graph for ComputeRevenue

After applying the discount, we assert that $revenue \geq cost$ since we do not wish to sell at a loss. We would like to know if this program ever fails due an assertion error, so we have to figure out if there exist integer inputs for which the program reaches the *Assertion Error* node in the control-flow graph. We might try to run the program on different input values, e.g ($units = 8, cost = 5$), ($units = 7, cost = 10$). These input values does not cause the program to fail, but we are still not convinced that it wont fail for some other input values. By observing the program for some time, we realize that the input must satisfy the following two constraints to fail:

$$units \cdot 2 \geq 16$$
$$units \cdot 2 < cost$$

which is the case e.g for ($units = 8, cost = 7$). This realization was not immediately obvious, and for more complex programs, answering the same question is even more difficult. The key insight is that the conditional statements dictates which execution path the program will follow. In this report we will present *symbolic execution*, which is a technique to systematically explore different execution paths and generate concrete input values that will follow these same paths.

Chapter 3

Principles of symbolic execution

In this chapter we will cover the theory behind symbolic execution. We will start by describing what it means to *symbolically* execute a program and how we deal branching. We will also explain the connection between a symbolic execution of a program, and a concrete execution. We shall restrict our focus to programs that takes integer values as input and allows us to do arithmetic operations on such values. In the end we will cover the challenges and limitations of symbolic execution that arises when these restrictions are lifted.

3.1 Symbolic executing of a program

During a normal execution of a program, input values consists of integers. During a symbolic execution we replace concrete values by symbols e.g α and β, that acts as placeholders for actual integers. We will refer to symbols and arithmetic expressions over these as *symbolic values.* The program environment consists of variables that can reference both concrete and symbolic values. [1] [4].
To illustrate this, we consider the following program that takes parameters a, b and c and computes their sum. The computation may seem unnecessarily complicated, but we do it this way to clearly illustrate the relationship between concrete and symbolic values:

```
**procedure** ComputeSum($a, b, c$)
    $x := a + b$
    $y := b + c$
    $z := x + y - b$
    **return** $z$
**end procedure**
```

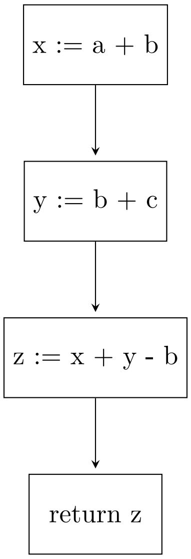

Figure 3.1: Control-flow graph for ComputeSum

Lets consider running the program with concrete values $a = 2, b = 3$ and $c = 4$. We then get the following execution: First we assign $a + b = 5$ to the variable x. Then we assign $b + c = 7$ to the variable y. Next we assign $x + y - b = 9$ to variable z and finally we return $z = 9$, which is indeed the sum of 2, 3 and 4.
Let us now run the program with symbolic input values α, β and γ for a, b and c respectively.

We then get the following execution: First we assign $\alpha + \beta$ to x. We then assign $\beta + \gamma$ to y. Next we assign $(\alpha + \beta) + (\beta + \gamma) - \beta$ to z.Finally we return $z = \alpha + \beta + \gamma$. We can conclude that the program correctly computes the sum of a, b and c, for any possible value of these.

3.2 Execution paths and path constraints

The program that we considered in the previous section contains no conditional statements, which means it only has a single possible execution path. In general, a program with conditional statements $s_1, s_2, \ldots, s_n$ with conditions $q_1, q_2, \ldots, q_n$, will have several execution paths that are uniquely determined by the value of these conditions. In symbolic execution, we model this by introducing a *path-constraint* for each execution path. The *path-constraint* is a list of boolean expressions $[q_1, q_2, \ldots, q_k]$ over the symbolic values, corresponding to conditions from the conditional statements along the path. At the start of an execution, the *path-constraint* only contains the expression *true*, since we have not encountered any conditional statements. to continue execution along a path, $q_1 \wedge \ldots \wedge q_k$ must be *satisfiable*. To be *satisfiable*, there must exist an assignment of integers to the symbols, such that the conjunction of the conditions evaluates to true. For example, $q = (2 \cdot \alpha > \beta) \wedge (\alpha < \beta)$ is satisfiable, because we can select $\alpha = 10$ and $\beta = 15$ in which case q evaluates to *true*. On the other hand $q' = (2 \cdot \alpha < 4) \wedge (alpha > 4)$ is clearly not satisfiable since the first condition stipulates that $\alpha < 2$ and the second condition stipulates that $\alpha > 4$.

Whenever we reach a conditional statement with condition q_k, we consider the two following expressions:

1. $q_1 \wedge q_2 \wedge \ldots \wedge q_k$
2. $q_2 \wedge q_2 \wedge \ldots \neg q_k$

This gives a number of possible scenarios:

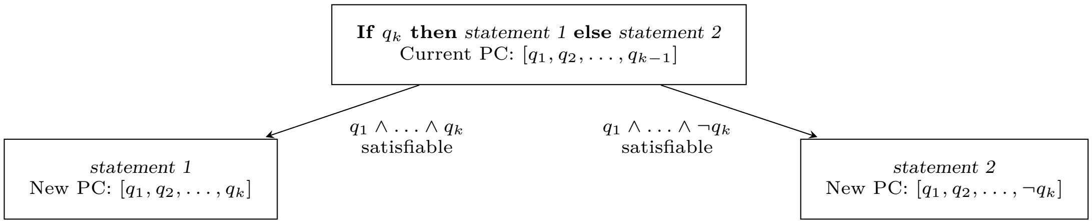
If q_k **then** *statement 1* **else** *statement 2*
Current PC: $[q_1, q_2, \ldots, q_{k-1}]$
$q_1 \wedge \ldots \wedge q_k$ satisfiable
$q_1 \wedge \ldots \wedge \neg q_k$ satisfiable
statement 1
New PC: $[q_1, q_2, \ldots, q_k]$
statement 2
New PC: $[q_1, q_2, \ldots, \neg q_k]$

Figure 3.2: Abstract overview of the symbolic execution of an *if*-statement, which potentially leads to two new execution paths, each with a new *path-constraint* .

- **Only the first expression is satisfiable**: Execution continues with a new *path-constraint* $[q_1, q_2, \ldots, q_k]$, along the path corresponding to q_k evaluating to to *true*.
- **Only the second expression is satisfiable**: Execution continues with a new *path-constraint* $[q_1, q_2, \ldots, \neg q_k]$, along the path corresponding to q_k evaluating to to *false*.
- **Both expressions are satisfiable**: In this case, the execution can continue along two paths; one corresponding to the condition being *false* and one being *true*. At this point we *fork* the execution by considering two different executions of the remaining part of the program. Both executions start with the same environment and *path-constraints* that are equal up to the final condition. One will have q_k as the final condition and the other will have $\neg q_k$. These two executions will continue along different execution paths that differ from this conditional statement and onward [4].

To illustrate this, we consider the program from the motivating example, that takes input parameters *units* and *costs*:

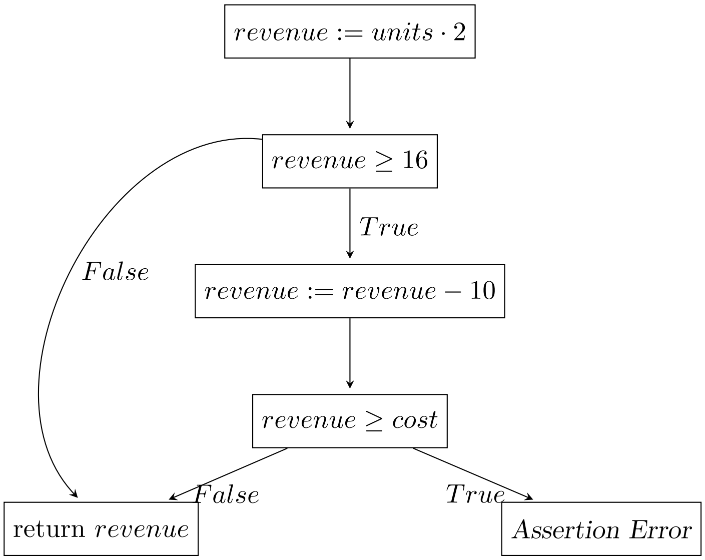

Figure 3.3: Control-flow graph for ComputeRevenue

We assign symbolic values α and β to *units* and *cost* respectively, and get the following symbolic execution:

First we assign $2{\cdot}\alpha$ to *revenue*. We then reach an *if*-statement with condition $\alpha \cdot 2 \geq 16$. To proceed, we need to check the satisfiability of the following two expressions:

1. $true \wedge (\alpha \cdot 2 \geq 16)$
2. $true \wedge \neg(\alpha \cdot 2 \geq 16)$.

Since both these expressions are satisfiable, we need to fork. We continue execution with a new *path-constraint* $[true, (\alpha \cdot 2 \geq 16)]$, along the path corresponding to the condition evaluating to *true*. We also start a new execution with the same environment and a *path-constraint* equal to $[true, \neg(\alpha \cdot 2 \geq 16)]$. This execution will continue along the path corresponding to the condition evaluating to *false*, and it immediately reaches the return statement and returns $\alpha \cdot 2$. The first execution assigns $2 \cdot \alpha - 10$ to *revenue* and then reach an *assert*-statement with condition $2 \cdot \alpha - 10 \geq \beta$. We consider the following expressions:

1. $true \wedge (\alpha \cdot 2 \geq 16) \wedge (((2 \cdot \alpha) - 10) \geq \beta)$
2. $true \wedge (\alpha \cdot 2 \geq 16) \wedge \neg(((2 \cdot \alpha) - 10) \geq \beta)$

Both of these expressions are satisfiable, so we fork again. In the end we have discovered all three possible execution paths with the following *path-constraints*:

1. $true \wedge \neg(\alpha \cdot 2 \geq 16)$
2. $true \wedge (\alpha \cdot 2 \geq 16) \wedge (((2 \cdot \alpha) - 10) \geq \beta)$
3. $true \wedge (\alpha \cdot 2 \geq 16) \wedge \neg(((2 \cdot \alpha) - 10) \geq \beta)$.

The first two *path-constraints* corresponds to the two different paths that leads to the return ment, where the first one returns $2 \cdot \alpha$ and the second one returns $2 \cdot \alpha - 10$. Inputs that satisfy these, does not result in a crash. The final *path-constraint* corresponds to the path that leads to the *Assertion Error*, so we can conclude that all input values that satisfy these constraints, will result in a program crash.

3.3 Constraint solving

As we just described, a symbolic execution of a program results in one or more *path-constraints* corresponding to each possible execution path. We know that each of these *path-constraints* are satisfiable, so we can solve the *path-constraint* by finding an assignment of concrete values to the symbols, that causes it to evaluate to true. Consider for example

$$true \wedge (\alpha \cdot 2 \geq 16) \wedge (2 \cdot \alpha - 10 \geq \beta)$$

which is one of the three resulting *path-constraints* from the motivating example. From the first condition we get that $\alpha \geq 8$, so we can select $\alpha = 8$. The second condition then gives us that $6 \geq \beta$, so we can select $\beta = 6$. If we do a concrete execution of the program with $units = 8$ and $cost = 6$, it will follow the execution path that correspond to this *path-constraint* . We can do the same for the remaining two *path-constraints*, and in the end we will have a pair of concrete input values for each possible execution path. So symbolic execution not only allows us to explore all possible execution paths, it also allows to generate a small set of concrete input values that cover all these paths.

3.4 Limitations and challenges of symbolic execution

So far we have only considered symbolic execution of programs with a small number of execution paths. Furthermore, the constraints placed on the input symbols have all been expressions. In this section we will cover the challenges that arise when we consider more general programs.

3.4.1 The number of possible execution paths

Since each conditional statement in a given program can result in two different execution paths, the total number of paths to be explored is potentially exponential in the number of conditional statements. For this reason, the running time of the symbolic execution quickly gets out of hands if we explore all paths. The challenge gets even greater if the program contains a looping statement. In this case, the number of execution paths is potentially infinite [1]. We illustrate this by considering the following program that computes a^b for integers a and b, with symbolic values α and β for a and b:

```
**procedure** ComputePow($a, b$)
    $r := 1$
    $i := 1$
    **while** $i \leq b$ **do**
        $r := r \cdot a$
        $i := i + 1$
    **end while**
    **return** $r$
**end procedure**
```

This program contains a *while*-statement with condition $i \leq b$. The $k'th$ time we reach this statement we will consider the following two expressions:

1. $true \wedge (1 \leq \beta) \wedge (2 \leq \beta) \wedge \ldots \wedge (k-1 \leq \beta)$
2. $true \wedge (1 \leq \beta) \wedge (1 \leq \beta) \wedge \ldots \wedge \neg(k-1 \leq \beta)$.

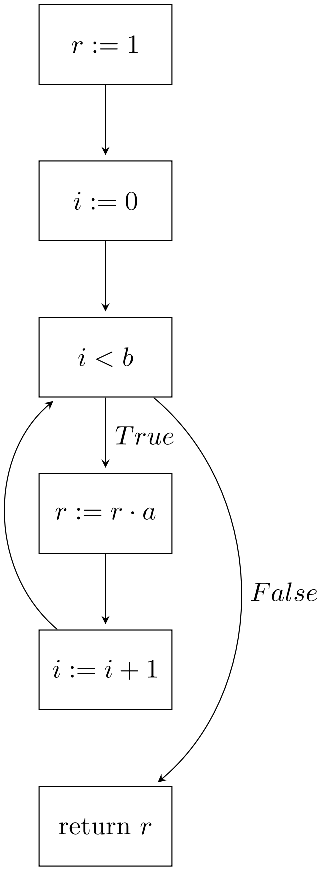

Figure 3.4: Control-flow graph for ComputePow

Both of these expressions are satisfiable, so we fork the execution. This is the case for any $k > 0$, which means that the number of possible execution paths is infinite. If we insist on exploring all paths, the symbolic execution will simply continue for ever. To avoid this, we can include some other termination criteria. As an example, we could have limit on the number of times we allow the execution to fork, and as soon as this limit is reached we simply ignore any further execution paths.

3.4.2 Deciding satisfiability of ***path-constraints***

A key component of symbolic execution, is deciding if a *path-constraint* is satisfiable, in which case the corresponding execution path is eligible for exploration. Consider the following *path-constraint* from the motivating example:

$$\mathit{true} \wedge (\alpha \cdot 2 \geq 16) \wedge \neg(2 \cdot \alpha - 10 < \beta).$$

To decide if this is satisfiable or not, we must determine if there exist an assignment of integer values to α and β such that the formula evaluates to *true*. We notice that the formula is a conjunction of linear inequalities. We can assign these to variables q_1 and q_2 and get

$$\begin{aligned} q_1 &= (\alpha \cdot 2 \geq 16) \\ q_2 &= (2 \cdot \alpha - 10 < \beta) \end{aligned}$$

The formula would then be $\mathit{true} \wedge q_1 \wedge \neg q_2$, where q_1 and q_2 can have values *true* or *false* depending on whether or not the linear inequality holds for some integer values of α and β. The question then becomes twofold: Does there exist an assignment of *true* and *false* to q_1 and q_2 such that the formula evaluates to true? And if so, does this assignment lead to a system of linear inequalities that is satisfiable? In this example, we can assign *true* to q_1 and *false* to q_2, which gives the following system of linear inequalities:

$$\begin{gathered} \alpha \cdot 2 \geq 16 \\ 2{\cdot}\alpha - \beta \geq 10 \end{gathered}$$

where we gathered the constant terms on the left hand side, and the symbols the right hand side. From the first equation we get that $\alpha \geq 8$so we select $\alpha = 8$. From the second equation we then get that $\beta \leq 6$, so we select $\beta = 6$ and this gives us a satisfying assignment for the path constraint. Since deciding satsifiability of a *path-constraint* is such an critical part of symbolic execution, it is important to know if we can always correctly *yes* or *no*. In the next section we will study this question more closely, and we will see that it depends on what sort of constraints we place on the input values.

The SMT problem

The example we just gave, is an instance of the *Satisfiability Modulo Theories*(SMT) problem. To understand SMT, we first consider the *Boolean Satisfiability*(SAT) problem. In this problem we are given a logical formula over boolean variables $q_1, q_2, \ldots, q_n$. We want to decide if there exists an assignment of truth values to each variable such that the formula evaluates to true. for example, $(q_1 \wedge \neg q_2)$ is a *yes*-instance of this problem, since $q_1 = \mathit{true}$ and $q_2 = false$ causes the formula to evaluate to *true*. On the other hand $q_1 \wedge q_2 \wedge \neg q_2$ is a *no*-instance since the formula is clearly false no matter which values we assign. This problem is *decidable*, meaning that we can always correctly answer *yes* or *no* to whether or not a given formula is satisfiable. However it is also NP-complete, which means that the best known method for solving this problem takes worst-case exponential time. SMT is then an extention of this problem. We now let the boolean variables $q_1, q_2, \ldots, q_n$ represent expressions from some theory such as the *theory of integer linear arithmetic(LIA)* [2]. In LIA, an expression e is defined as

$$\begin{array}{rcl} e \in \textsc{Expression} & := & p \bowtie p \\ \bowtie \in \textsc{Constraint} & := & \leq \ | \ \geq \ | \ = \\ p \in \textsc{Polynomial} & := & a \mid a \cdot x \mid p \circ p \\ \circ \in \textsc{Operator} & := & + \mid - \\ a \in \textsc{Coefficient} & := & \mathbb{Z} \\ x \in \textsc{Variable} & := & \mathbb{Z} \end{array}$$

We now want to decide if the formula is satisfiable with respect to the theory. To be satisfiable w.r.t LIA, we require that there exist an assignment of truth values to the boolean variables such that the formula evaluates to *true*, and that for such an assignment, there exists an assignment of integer values to the variables in the underlying expressions such that all the constraints are satisfied. In the previous section we showed that

$$\mathit{true} \land (\alpha \cdot 2 \geq 16) \land \neg(2 \cdot \alpha - 10 < \beta).$$

is a *yes*-instance of the SMT problem, since we could let q_1 represent $(\alpha \cdot 2 \geq 16)$ and q_2 represent $(2 \cdot \alpha - 10 < \beta)$. We could then assign $q_1 = \mathit{true}$ and $q_2 = false$. For this assignment of truth variables, we could choose $\alpha = 8$ and $\beta = 6$ in which case all the constraints hold. On the other hand

$$(2 \cdot \alpha < 4) \land (\alpha > 4)$$

is clearly a *no*-instance. If we let q_1 represent $(2 \cdot \alpha < 4)$ and q_2 represent $(\alpha > 4)$, the formula becomes $q_1 \land q_2$. The only assignment of truth variables that satisfy this is $q_1 = \mathit{true}$ and $q_2 = \mathit{true}$. But it is clear that there does not exist an integer value for α such that both the constraints hold. LIA is a *decidable* theory, meaning that we can always correctly answer *yes* or *no* to an instance of the SMT problem, when it is with respect to LIA. Given an SMT formula, we can construct an Integer Linear Program with the expressions from the formula as constraints. This program will we feasible if and only if the SMT formula is satsifiable w.r.t LIA. We can check the feasibility of the Integer linear Program by using the *branch-and-bound* algorithm. If the program is feasible, this will also give us a satisfying assignment.

This means that as long as we only consider programs with linear expressions as conditions, we can always decide the satisfiability of a *path-constraint* .

Undecidable theories

Let us consider the following extension of the definition of expressions in LIA:

$$\begin{array}{rcl}
e \in \textsc{Expression} & := & p \bowtie p \\
\bowtie\, \in \textsc{Constraint} & := & \leq \;\mid\; \geq \;\mid\; = \\
p \in \textsc{Polynomial} & := & a \mid a \cdot x \mid p \circ p \\
\circ \in \textsc{Operator} & := & + \mid - \mid \cdot \\
a \in \textsc{Coefficient} & := & \mathbb{Z} \\
x \in \textsc{Variable} & := & \mathbb{Z}
\end{array}$$

This definition allows for expressions with nonlinear terms such as $5 \cdot x - 10 \leq 3 \cdot y^3$. Such expressions belong to the *Theory of Nonlinear Integer Arithmetic*(NLIA). This theory is *undecidable*, meaning that there does not exist an algorithm that always halt with a correct answer to the question of whether or not there exists an assignment of integer values to the variables such that all constraints are satisfied. This means that if we try to do symbolic execution on a program that has conditional statements with such nonlinear expressions in the condition, we might get stuck trying to decide the satisfiability of a path constraint forever. To avoid this, we can include some other termination criteria, such as a time limit, at which point we output *unknown* instead of *yes* or *no*. This leaves us with two options. We can continue the symbolic execution and assume that the *path-constraint* is in fact satisfiable. In this case, we might end up exploring execution paths that are not reachable for any concrete input values. The other option is to assume that the *path-constraint* is unsatisfiable. In this case we do not explore unreachable execution paths, but we can no longer be sure that we explore all reachable paths.

Chapter 4

Principles of Concolic execution

In this chapter we will introduce *concolic execution*, which is a technique that combines concrete and symbolic execution to explore possible execution paths. We start by describing how the technique works, and then look at the advantages that it offers compared to only using symbolic execution. As in the previous chapter, we restrict our focus to programs that takes integer values as input, and performs arithmetic operations and comparisons on these.

4.1 Concolic execution of a program

During a symbolic execution of a program, we replace the inputs of the program with symbols that acts as placeholders for concrete integer value. The program environment maps variables to *symbolic values*, which can either be integers, symbols or arithmetic expressions over these two. During a concolic execution of a program, we maintain two environments. One is a concrete environment M_c, which maps variables to concrete integer values. The other is a symbolic environment M_s which maps variables to symbolic values. At the beginning of the execution, the concrete environment is initialized with a random integer value for each input. The symbolic environment is initialized with symbols for each input. Finally we initialize an empty *path-constraint* . The program is then executed both concretely and symbolically, and at the end of this execution we determine a new set of concrete input values that will cause the concrete execution to follow a different path. In the next iteration we initialize the concrete environment with these values and the program is executed concretely and symbolically again. We continue doing this until all execution paths have been explored, or some other predefined termination criteria is met [3]. We will now describe what we mean by executing the program both concretely and symbolically. Specifically, we will explain how we maintain our environments, how we handle conditional statements and how we generate the next set of concrete input values.

4.1.1 Maintaining the environments

If we reach an assignment statement *v* := *e*, for some variable *v* and expression *e*, we evaluate *e* concretely and update our concrete environment. We also evaluate *e* symbolically and update our symbolic environment.

4.1.2 Conditional statements

Whenever we reach a conditional statement with condition q, we evaluate q concretely and chose a path accordingly. At the same time we evaluate q symbolically and get some constraint c. If the concrete value of q is true, we add c to a *path-constraint* . If the concrete value of q is false, we add $\neg c$ to the path constraint. This way track which choices of paths we have made during an iteration.

4.1.3 Generating input values for next iteration

At the end of an iteration we will have a *path-constraint* that, for each encountered conditional statement, describes what choice of path we made. To generate the input for the next iteration, we construct a new *path-constraint* by negating the condition of the last conditional statement where we have not explored the other path. We then solve the constraints from this new *path-constraint* to obtain the next set of input values. If some input values are not constrained by this *path-constraint* , we keep the current concrete value for the next iteration.

To illustrate concolic execution, we consider the program from the motivating example:
First we initialize the two environments. We let

$$M_c = \{units = 27,\ cost = 34\}$$

where 27 and 34 is chosen randomly. We let

$$M_s = \{units = \beta,\ cost = \alpha\}$$

where α and β are symbols. The first statement is an assignment statement, so we get two new environments:

$$M_c = \{units = 27,\ cost = 34,\ revenue = 54\}$$
$$M_s = \{units = \alpha,\ cost = \beta,\ revenue = 2 \cdot \alpha\}$$

Next, we reach an *if* statement with condition $revenue \geq 16$. Since $M_c(revenue) = 54$ we follow the *then* branch. We update the *path-constraint* to $[(2 \cdot \alpha \geq 16)]$. Next, we reach another *assign* statement, so we get the following environments:

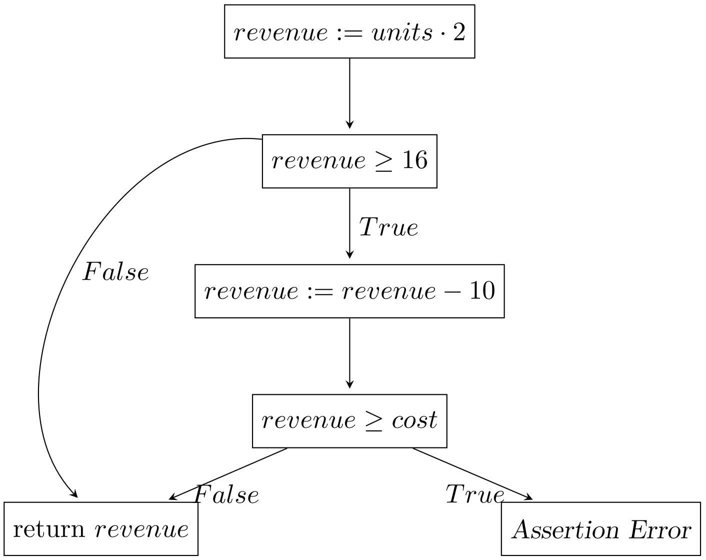

Figure 4.1: Control-flow graph for ComputeRevenue

$$M_c = \{units = 27,\ cost = 34,\ revenue = 44\}$$
$$M_s = \{units = \alpha,\ cost = \beta,\ revenue = 2 \cdot \alpha - 10\}$$

Next, we reach an *assert* statement which condition $revenue \geq cost$. Since $M_c(revenue) = 44$ the assertion succeeds. We update the *path-constraint to* $[(2 \cdot \alpha \geq 16), ((2 \cdot \alpha - 10) \geq \beta)]$. Finally we return 44, which finishes one execution path.
To discover a new *path-constraint* , we make a new *path-constraint* by negating the final condition of the current *path-constraint* , so we get $[(2 \cdot \alpha \geq 16), \neg((2 \cdot \alpha - 10) \geq \beta)]$. To get the next set of concrete input values, M_c we solve the following system of constraints

$$2 \cdot \alpha \geq 16$$
$$(2 \cdot \alpha - 10) < \beta.$$

This gives us e.g $\alpha = 8$ and $\beta = 7$. We then execute the program with $units = 8$ and $cost = 7$. This execution will follow the same path until we reach the *assert*-statement. This time the execution results in an error due to the assertion being violated, and we have now explored all possible paths from the *assert*-statement. To generate the next input values, we negate the condition from the first *if*-statement and get $[\neg(2 \cdot \alpha \geq 16)]$. From this we get e.g $\alpha = 5$. Since there are no constraints on the value of β, we keep the previous value. We now execute the program with $units = 5$ and $cost = 7$. This execution will not follow the *then* branch in the *if*-statement, so we immediately return *revenue* which is 10. At this point we have explored all possible branches from each conditional statement, so we have explored all execution paths.

4.2 Handling undecidable *path-constraints*

In the previous chapter we described how symbolic execution was limited by the ability to decide satisfiability of a *path-constraint* . For example, if we are executing a program with inputs x and y and encounter a non linear condition $5 \cdot x - 10 \leq 3 \cdot y^3$, we might fail to decide the satisfiability of the two branches. In this case we can assume that the paths are not satisfiable and ignore them, or assume that the paths are satisfiable and continue. In the first case, we potentially miss a large number of interesting paths, and in the second case we can no longer guarantee that we only explore feasible paths.
The same issue may arise in concolic execution when generating the input values for the next iteration, but we are not left with same options as in symbolic execution. Since we always have access to both a concrete and a symbolic environment, we can avoid having non linear conditions in the *path-constraint* . Non linear conditions comes from arithmetic operations involving multiplication or division, where both operands contain symbols. In this case we can instead choose to evaluate the expression with the concrete environment[3]. Consider the condition $5 \cdot x - 10 \leq 3 \cdot y^3$ again. Assume that y have concrete value 2, we then evaluate the right-hand side and get 24. The condition then becomes $5 \cdot x - 10 \leq 24$, which is within the theory of integer linear arithmetic, which we know we can solve. We can then decide satisfiability of the path with this more restrictive constraint. This allows us to continue exploring more paths without also exploring infeasible paths. The cost of this strategy is the loss of completeness. If we decide to negate $5 \cdot x - 10 \leq 24$ and get $5 \cdot x - 10 > 24$, it is clear that solving both these constraints is not sufficient to say that all paths from the statement with condition $5 \cdot x - 10 \leq 3 \cdot y^3$ is explored. To guarantee this we would have to check for an infinite number of possible values for y.

Chapter 5

Introducing the language *SIMPL*

In chapter 5 and 6,

In this chapter we will introduce *SIMPL* which is a small programming language which consists of top-level functions, expressions and statements. We start by describing the syntax of the language, and then we describe a concrete interpreter for the language.

5.1 Syntax of *SIMPL*

SIMPL consists of expressions and statements. Expressions evaluates to values and does not change the control flow of the program. A statement evaluates to a value and a possibly updated variable environment. They may also change the control flow of the program through conditional statements.

5.1.1 Expressions

Expressions consists of concrete values which can be integers$\langle I, \rangle$, booleans$\langle B \rangle$, and a special *unit* value. Furthermore they consist of variables referenced by identifiers. Finally they consist of arithmetic operations on integers and comparison of integers

$$
\begin{aligned}
\mathit{int} \in \textsc{Integer} &:= \quad 0 \mid 1 \mid -1 \mid \ldots \\
\mathit{bool} \in \textsc{Boolean} &:= \quad \mathit{True} \mid \mathit{False} \\
\mathit{val} \in \textsc{ConcreteValue} &:= \quad \mathit{int} \mid \mathit{bool} \mid \mathit{unit} \\
\mathit{id} \in \textsc{Identifier} &:= \quad a \mid b \mid c \mid \ldots \\
\mathit{exp} \in \textsc{Expression} &:= \quad \mathit{val} \mid \mathit{id} \\
&\mid \mathit{exp} + \mathit{exp} \mid \mathit{exp} - \mathit{exp} \mid \mathit{exp} * \mathit{exp} \mid \mathit{exp}/\mathit{exp} \\
&\mid \mathit{exp} < \mathit{exp} \mid \mathit{exp} > \mathit{exp} \mid \mathit{exp} \leq \mathit{exp} \\
&\mid \mathit{exp} \geq \mathit{exp} \mid \mathit{exp} == \mathit{exp}
\end{aligned}
$$

5.1.2 Statements

Statements in *SIMPL* consists of assignments to values. Variables are implicitly declared, so assigning to a variable that does not already exist, will create the variable.

$$\mathit{smt} \in \textsc{Statement} := \quad \mathit{id} = \mathit{exp}$$

The value of an assignment statement is the value of the expression on the right hand side. This expression must evaluate to either an integer value or a boolean value. *SIMPL* supports three different conditional statements; *if*-statements, *while*-statements and *while*-statements.

$$
\begin{aligned}
\mathit{stm} \in \textsc{Statement} :=& \quad \textbf{if } \mathit{exp} \textbf{ then } \mathit{stm} \textbf{ else } \mathit{stm} \\
&\mid \textbf{while } \mathit{exp} \textbf{ do } \mathit{stm} \\
&\mid \textbf{assert } \mathit{exp}
\end{aligned}
$$

The condition expression must evaluate to a boolean value. The value of an *if*-statement is the value of the statement that gets evaluated depending on the value of the condition. The value of a *while*-statement is *unit*. The value of an *assert*-statement is *unit* if the condition evaluates to *True*, otherwise the execution ends with an error. Finally a statement can simply be an expression or a sequence of expressions.

$$\mathit{stm} \in \textsc{Statement} := \quad \mathit{exp} \mid \mathit{stm}\ \mathit{stm}$$

5.1.3 Functions and programs

SIMPL supports top-level functions that must be defined at the beginning of the program. A function definition consists of an identifier, followed by a parameter list with zero or more identifiers. The function body consists of one or more statements. A function call is an expression consisting of an identifier matching one of the defined functions followed by a argument list with zero or more expresssions.

$$\begin{aligned} f \in \textsc{Function} &:= \quad id\ (id^*)\ \{stm\} \\ call \in \textsc{Expression} &:= \quad id\ (exp^*) \end{aligned}$$

Since expressions only evaluates to a value, it follows that function calls does not have any side effects.

Finally we define the syntax of a *SIMPL* program to be one or more function declarations, followed by a call to one of these functions.

$$prog \in \textsc{Program} := \quad f\ f^*\ call$$

5.2 Concrete interpreter for *SIMPL*

To interpret *SIMPL* we define the three following functions

```
class ConcreteInterpreter {
  def interpProg(p: Prog): Result[ConcreteValue, String]

  def interpExp(p: Prog, e: Exp, env: HashMap[Id, ConcreteValue]):
      Result[ConcreteValue, String]

  def interpStm(p: Prog, s: Stm, env: HashMap[Id, ConcreteValue]):
      Result[(ConcreteValue, HashMap[Id, ConcreteValue]), String]
}
```

Figure 5.1: High level overview of the interpreter implementation in *Scala*.

where *interpProg* acts as the entry function. It takes a program p of type *Prog(funcs: Map[Id, FDecl], fCall: CallExp)* and calls *interpExp* with p, *fCall* and a fresh environment. We use an immutable map of type *Map[Id, ConcreteValue]* to represent our environment. From the signature of *interpExp* we see that the interpretation of an expression depends on the current environment and the program which holds the top level functions, and that it results in a concrete value. The final function is *interpStm* whose signature tells us that the interpretation of a statement also depends on the current environment and the program, and that results in both a concrete value and a potentially updated environment.

5.2.1 Error handling

We wish to keep our implementation purely functional, so we need to avoid throwing exceptions whenever we encounter an error.

Instead we define a trait *Result[+V, +E]*, that can either be *Ok(v: V)*, or *Error(e: E)*, for some types *V* and *E*.

```
  trait Result[+V, +E]
  case class Ok[V](v: V) extends Result[V, Nothing]
  case class Error[E](e: E) extends Result[Nothing, E]
```

We let *InterpExp* have return value *Result[ConcreteValue, String]*, meaning we return *Ok(v: ConcreteValue)* if we do not encounter any errors, and *Error(e: String)* if we do. In this case *e* will be a message that describes what sort of error we encountered. *InterpStm* has return value *Result[(ConcreteValue, HashMap[Id, String]), String]* since we also return a potentially updated environment.

Map and flatMap

We define three functions *map* and *flatMap* and traverse:

```
def map[T](f: V => T): Result[T, E] = this match {
case Ok(v) => Ok(f(v))
case Error(e) => Error(e)
}

def flatMap[EE >: E, T](f: V => Result[T, EE]): Result[T, EE]
= this match {
  case Ok(v) => f(v)
  case Error(e) => Error(e)
}

def traverse[V, W, E](vs: List[V])(f: V => Result[W, E]):
 Result[List[W], E] = vs match {
  case Nil => Ok(Nil)
  case hd::tl => f(hd).flatMap( w => traverse(tl)(f)
.map(w :: _))
}
```

For some result *r*, *map* allows us to apply a function $f : V \to T$ to *v* if $r =$ *Ok(v)*. Otherwise we simply return *r*. *flatMap* has the same functionality except that we apply a function that also returns a *Result*. This way we avoid nesting like *Ok(Ok(v))*. These two functions allows us to handle errors seamlessly. If, for example we wish to interpret an arithmetic expression *AExp(e1, e2, Add())*, we simply do

```
interpExp(p, e1, env).flatMap(
  v1 => interpExp(p, e2, env).map(v2 => v1.v + v2.v)
)
```

If we encounter an error during the interpretation of *e1* this error is returned immediately. If not we continue by interpreting *e2*. If we encounter an error here, we return this error, otherwise we continue and return the sum of the two expressions. Here we assume that *v1* and *v2* are of type *IntValue(v: Int)*. The full implementation also includes checking that both expressions evaluate to the proper types.

Finally we define a function *traverse*, that takes a list of type *V*, a function $f : V \rightarrow Result[W, E]$ and returns a *Result[List[W], E]*. We use this function to interpret functions calls *CallExp(Id, List[Exp])*. To do this, we need to evaluate the argument list, and if we do not encounter any errors, construct a local environment and evaluate the function body. We could simply map *interpExp* on to the argument list, and check for each element in the resulting list if it is an error. This would require two passes over the list. Instead, *traverse* applies f to each element in the list, and if it results in an error, we immediately return this error. Otherwise we prepend the resulting value onto the result of traversing the remaining list. We only pass over the list once, and we immediately return the first error encountered.

Chapter 6

Symbolic execution of *SIMPL*

In this chapter we will describe a symbolic interpreter for *SIMPL* . We start by extending the grammar to include symbolic values. Next, we describe the implementation of *path-constraints* and the functions for interpreting expressions and statements.

6.1 Extension of grammar

The original language have three types of values, namely integers, booleans and the unit value. Symbolic execution requires that we extend the language to include symbolic counterparts of these. We therefore introduce a symbolic value which can either be a symbolic integer, a symbolic boolean or the unit value.

$$val \in \textsc{SymbolicValue} := \quad int_{sym} \mid bool_{sym} \mid unit$$

A symbolic integer can either be a concrete integer, a symbol or an arithmetic expression over these two:

$$\begin{aligned} int \in \textsc{Integer} &:= \quad 0 \mid 1 \mid -1 \mid \ldots \\ sym \in \textsc{Symbol} &:= \quad a \mid b \mid c \mid \ldots \\ int_{sym} \in \textsc{SymbolicInteger} &:= \quad int \mid sym \mid int_{sym} + int_{sym} \\ &\quad \mid int_{sym} - int_{sym} \mid int_{sym} * int_{sym} \\ &\quad \mid int_{sym}/int_{sym} \end{aligned}$$

A symbolic boolean can either be $True$, $False$ or an symbolic boolean expression, which is a comparison of two symbolic integers. Finally a symbolic boolean can be the negation of a symbolic boolean. This extension is needed to be able to represent the two different *path-constraints* that might arise from a conditional statement.

$$\begin{aligned} bool \in \textsc{Boolean} &:= \quad True \mid False \\ bool_{sym} \in \textsc{SymbolicBoolean} &:= \quad int_{sym} < int_{sym} \mid int_{sym} > int_{sym} \\ &\mid int_{sym} \leq int_{sym} \mid int_{sym} \geq int_{sym} \\ &\mid int_{sym} == int_{sym} \mid !\ int_{sym} \end{aligned}$$

6.2 Path-constraints

To represent a *path-constraint* , we implement the following classes:

```
case class PathConstraint(conds: List[SymbolicBool], ps: PathStatus) {
     def :+(b: SymbolicBool): PathConstraint =
         PathConstraint(this.conds :+ b, this.ps)
}
sealed trait PathResult
case class Certain() extends PathResult
case class Unknown() extends PathResult
```

The definition of *PathConstraint* consists of two elements. *conds* is a list of symbolic booleans from the conditional statements met so far. The *PathStatus* tells us whether or not we can guarantee that the path constraint is in fact satisfiable. This is necessary because we allow for nonlinear constraints, which our SMT solver may fail to solve. In the case of a failure, we still explore the path but the status of the *path-constraint* will be *Unknown*, meaning that we cannot guarantee that there exists concrete input values that follow this path.

6.3 Interpretation of expressions

To interpret an expression, we define the following function

```
def interpExp(p: Prog, e: Exp, env: HashMap[Id, SymbolicValue],
  pc: PathConstraint, forks: Int): List[ExpRes]

case class ExpRes ExpRes(res: Result[SymbolicValue, String],
  pc: PathConstraint)
```

This version of *interpExp* is an extension of the version from the concrete interpreter. It still depends current environment, which is now of type *Map[Id, SymbolicValue]* and the program itself. However, it now also depends on a *path-constraint* and integer *forks* which is the current number of times the execution has been forked. The return type has also changed to *List[ExpRes]* where *ExpRes* is pair consisting of a symbolic value, which is the value of the expression, and a *path-constraint* . This tells us that the symbolic interpretation of an expression does necessarily result in a single symbolic value. In fact it will result in in a list of symbolic values, one for each possible execution path. This also effects how we interpret the different types of expressions. Most of them consists of one or more sub-expressions, each of which may result in several different values, so we have to consider all possible combinations of results.

6.3.1 Arithmetic and boolean expressions

Consider an arithmetic expression *AExp(e1: Exp, e2: Exp, op: AOp)*. When we recursively interpret e_1 and e_2, we get two lists L_{e_1}, L_{e_2} of possible results. We need to take the cartesian $L_{e_1} \times L_{e_2}$ of the lists, and for each pair of results, we have to evaluate the arithmetic expression.

To do this we use a *for-comprehension* which given two lists, iterate over each ordered pair of elements from the two lists. For each pair, we *flatMap* over the two results, and if no errors are encountered, we check that both expressions evaluates to integers and compute the result. Boolean expressions are interpreted in a similar fashion.

6.3.2 Function calls

Consider a *Call*-expression *CallExp(id: Id, args: List[Exp])*. First, we must check that the function is defined and that the formal and actual argument list does not differ in length. If both of these checks out, we map *interpExp* on to *args*. This gives a list of lists $[L_{e_1}, L_{e_2}, \ldots, L_{e_n}]$, where the $i'th$ list contains the possible results of interpreting the $i'th$ expression in *args*. We take the cartesian product $L_{e_1} \times L_{e_2} \times \ldots \times L_{e_n}$, which gives us all possible argument lists. For each of these lists, we zip it with formal argument list. This gives a list of the following type *List[(ExpRes, Id)]*. We attempt to build a local environment by folding over this list, starting with the original environment that was passed with the call to *interpExp*. If we encounter an error, either from the interpretation of the expressions in *args* or from an expression evaluating to the unit value, the result of the function call with this argument list will be that error. Otherwise, for each pair of result and id, we make a new environment with the appropriate binding added. We then interpret the statement in the function body with the local environment, and the result of the function call with this argument list, will be the value of the statement.

6.4 Interpreting statements

To interpret statements, we define the following function:

```
interpStm(p: Prog,
       stm: Stm,
       env: HashMap[Id, SymbolicValue],
       pc: PathConstraint,
       forks: Int): List[StmRes]

case class StmRes(res: Result[SymbolicValue, String],
            env: HashMap[Id, SymbolicValue],
            pc: PathConstraint)
```

The signature is similar to *interpExp*, except that the return type now also includes a possibly updated environment.

6.4.1 Assignment statements

Given an *Assignment*-statement *AssignStm(v: Var, e: Exp)*, we interpret the expression *e*, and get a list of results for each possible execution path. For each of these results, we return the value of the expression and an updated environment. If the expression resulted in an error, or a unit value, we instead return error and the original environment.

6.4.2 Conditional statements

We define the following classes to represent each type of conditional statement:

```
  case class IfStm(cond: Exp, thenStm: Stm, elseStm: Stm)
  case class WhileStm(cond: Exp, doStm: Stm)
  case class AssertStm(cond: Exp)
```

For each possible result of interpreting the condition expression we do the following: If it is either *True()* or *False()*, we follow the appropriate execution path, with the original values of *pc* and *forks*.
If the result is a symbolic boolean expression *b*, we have two potential paths to follow. We must check the satisfiability of *pc :+ b* and *pc :+ Not(b)*. We do this by calling *(checkSat(pc, b)* and *checkSat(pc, Not(b))*. If *checkSat* reports *SATISFIABLE*, the given path is eligible for exploration. If it reports *UNSATISFIABLE*, we can safely ignore the path. If it reports *UNKNOWN*, we still explore the path, but the new *path-constraint* will have its status set to *UNKNOWN()*. If both calls to *checksat* return either *SATISFIABLE* or *UNKNOWN*, we explore both paths, and return a list containing the results of the first path, followed by the results of the second path.

Checking satisfiability

To check the satisfiability of a path, we define the following function

```
def checkSat(pc: PathConstraint, b: SymbolicBool): z3.Status
```

which takes a *path-constraint* and a symbolic boolean b, and checks the satisfiability of $b_1 \wedge b_2 \wedge \ldots \wedge b_k \wedge b$, where $b_1, b_2, \ldots, b_k$ are the symbolic booleans in *pc.conds*. To do this, we use the *Java* implementation of the *Z3* SMT solver.

Uppper bound on number of forks

The interpretation of an *If*-statement and a *While*-statement starts with checking whether $forks > maxForks$, in which case we immediately return an Error. We include this check to prevent the interpreter to run forever on programs with an infinite number of execution paths.

6.4.3 Sequence statements

We define the following class to represent a sequence statement:

```
  case class SeqStm(s1: Stm, s2: Stm)
```

For each possible result from interpreting *s1*, we interpret *s2* with the new environment. If the interpretation of *s1* results in an error, this will be the result of the sequence expression. otherwise it will be the result of the interpretation of *s2*.

6.4.4 Expression statements

We define the following class to represent an *expression*-statement:

```
case class ExpStm(e: Exp)
```

To interpret this, we interpret e. For each result of this, the result of the statement, will be the value of the expression and the original environment.

Chapter 7

Conclusion

Appendix A

Source code

Appendix B

Figures

Bibliography

[1] Cristian Cadar and Koushik Sen. Symbolic execution for software testing: Three decades later. *Communications of the ACM*, 56:82–90, 02 2013. doi: 10.1145/2408776.2408795.

[2] Leonardo De Moura and Nikolaj Bjørner. Satisfiability modulo theories: Introduction and applications. *Commun. ACM*, 54(9):69–77, September 2011. ISSN 0001-0782. doi: 10.1145/1995376.1995394. URL `http://doi.acm.org/10.1145/1995376.1995394`.

[3] Patrice Godefroid, Nils Klarlund, and Koushik Sen. Dart: Directed automated random testing. *SIGPLAN Not.*, 40(6):213–223, June 2005. ISSN 0362-1340. doi: 10.1145/1064978.1065036. URL `http://doi.acm.org/10.1145/1064978.1065036`.

[4] James C. King. Symbolic execution and program testing. *Commun. ACM*, 19(7):385–394, July 1976. ISSN 0001-0782. doi: 10.1145/360248.360252. URL `http://doi.acm.org/10.1145/360248.360252`.