Symbolic Execution(Working title)

Aarhus Universitet

Søren Baadsgaard

March 25, 2019

Abstract

Contents

Chapter 1

Introduction

Chapter 2

Motivation

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 explain the connection between a symbolic execution of a program, and a concrete one.

3.1 Symbolically executing a program

When we execute a program symbolically, we use symbolic values as input data to the program, instead of concrete values. Instead of referencing concrete values, variables will reference expressions over the symbolic values, and therefore the return value of a program will also be such expressions [1].

To illustrate this, we consider the following program, that takes parameters a, b, c and computes the sum:

```
x = a + b
    |
y = b + c
    |
z = x + y - b
    |
return z
```

If we run this program on with concrete values, say $a = 2, b = 3, c = 4$, we would get the following execution: First we assign $a + b = 5$ to the variable x. Then we assign $b + c = 7$ to the variable $y\cdot$ 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 which symbolic input values α, β and γ for a, b and c respectively.

We would then get the following execution: We assign $\alpha + \beta$ to x. We then assign $\beta + \gamma$ to y. Finally we assign $(\alpha + \beta) + (\beta + \gamma) - \beta$ to z and 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 Path constraints and constraint solving.

In the previous section we gave an example of a symbolic execution of a program that computes the sum of three integers. This program contains no conditional statements, so it will follow the same execution path no matter what input we run it with, and we do not place any constraints on the input. In general, a program will follow different execution paths, depending on the outcome of any conditional statements along the path.

To encapsulate this, we introduce a *path-constraint* which is a list of boolean expressions over the symbolic values. To follow a given path, the conjuction of these expressions must be *satisfiable*. To be *satisfiable*, there must exist an assignment of concrete values, to the symbolic ones so that the conjunction of the boolean expressions evaluates to true.

The program is an extremely simple case, in which the program will behave exactly the same for any possible input values. In reality however, most program languages have some way of allowing branching to happen, and with this feature, we cannot guarantee that the program will behave exactly the same for all inputs. If our program contains an expression like *If* $x > 2$ *Then ... Else ...* , the program will behave differently depending on whether or nor $x > 2$. To encapsulate this, we introduce a *path-constraint (PC)*, which is a boolean expression that will contain all properties that the input must satisfy to follow the given path. At the beginning of the execution, the PC will be initialized with the value *true* as no assumptions have been made yet (If we introduce pre-conditions on the input, the PC will of course contain these). At any *if-expression* with condition q, during the execution we will look at the following to expressions

1. $PC \wedge q$
2. $PC \wedge \neg q$.

To encapsulate this, we introduce a *path-constraint* which is a list of boolean expressions $\{q_1, q_2, \ldots, q_n\}$ over the symbolic values. To follow a given path, $q_1 \wedge \ldots \wedge q_n$ must be *satisfiable*. To be *satisfiable*, there must exist an assignment of concrete values, to the symbolic ones so that the conjunction of the expressions evaluates to true. For example, $q = (2 \cdot \alpha > \beta) \wedge (\alpha < \beta)$ is satisfiable, because we can choose $\alpha = 10$ and $\beta = 15$ in which case q evaluates to *true*.
Whenever we reach a conditional statement with condition q, we consider the two following expressions:

1. $pc \wedge q$
2. $pc \wedge \neg q$

where *pc* is the conjunction of all the expressions currently contained in the *path-constraint*.
This gives a number of possible scenarios:

- **Only the first expression is satisfiable**: We add q to the *path-constraint*, and we continue the execution along the same path.
- **If only the second expression is satisfiable**: We add $\not q$ to the *path-constraint*, and we continue along the same path.
- **Both expressions are satisfiable**: In this case, the execution can follow two paths; one where we assume q and one where we assume $\not q$. At this point we *fork* we fork the execution by considering two different executions of remaining program. Both executions start with the same shared variable state, and *path-constraints* that is equal up to the final element. One execution will have q as the final element and the other will have $\not q$. These two executions will now follow two different execution paths that differ from this conditional statement and onward.

We illustrate this with the following example:

```
Fun pow(a, b)
  var r = 1
  var i = 0
  while (i < b)
    r = r*a
    i = i + 1
  return r
```

If we assign $a = \alpha$ and $b = \beta$, we get the following execution:

- PC is initialized to *true*
- $r \leftarrow 1$
- $i \leftarrow 0$
- We hit a branching point, so we check if $true \wedge (0 < \beta)$ and $true \wedge \neg(0 < \beta)$ are satisfiable. Since they both are, we must fork:
- **Case** $\neg(0 < \beta)$: $PC' \leftarrow true \wedge \neg(0 < \beta)$. The program returns 1. So we can conclude that the program returns 1 when $\beta \leq 0$.
- **Case** $(0 < \beta)$: $PC \leftarrow true \wedge (0 < \beta)$.

  $r \leftarrow 1 \cdot a$

  $i \leftarrow 0 + 1$
- We hit a branching point again, so we check if $true \wedge (0 < \beta) \wedge (1 < \beta)$ and $true \wedge (0 < \beta) \wedge \neg(1 < \beta)$ are satisfiable. Since both they are, we fork again:

- **Case** $\neg(1 < \beta)$: $PC' \leftarrow true \wedge (0 < \beta) \wedge \neg(1 < \beta)$. The program returns α. So we can conclude that the program returns α when $\beta = 1$.
- **Case** $1 < \beta$: $PC \leftarrow true \wedge (0 < \beta) \wedge (1 < \beta)$.

  $r \leftarrow a * a$

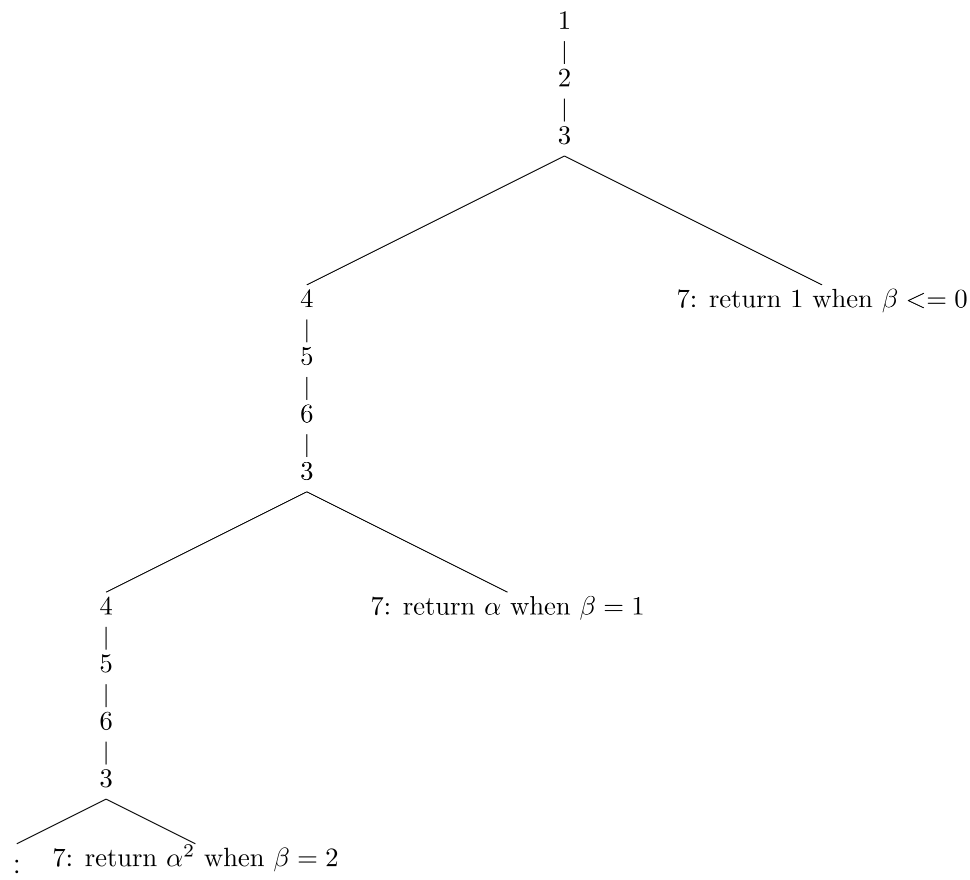
  $i \leftarrow 1 + 1$
- We hit a branching point ...

An important property of the *path-constraint* is that it can never become identically false. To see why this is the case, we have to look at the possible updates of PC. At the start of an execution, it will be initialized with the value *true*. At any branching point, it will be updated with exactly one of the expressions $PC \wedge q$ and $PC \wedge \neg q$, and only if the given expression is satisfiable. So PC will never end up looking like $\ldots \wedge q \wedge \ldots \wedge \neg q \wedge \ldots$ for some condition q. What this means is that when the program terminates at the end of some execution path, PC will be a satisfiable formula over the symbolic values, which means that we can solve the constraints and derive a set of concrete values which will follow the exact same path if we execute the program normally.

3.3 Limitations of symbolic execution

3.3.1 Infinite execution trees

As demonstrated in the last example in the previous sections, a symbolic execution can easily become infinite as soon as we introduce branching and some looping structure. This is further illustrated if we consider an execution tree for a program. In [2], they are described by enumerating each statement, and let each node in the tree, correspond to an execution of one of the command. The edges going out from a node corresponds to the transition from one statement to the next. From this we can see that non forking statements will only have a single edge outgoing, while forking statements will have two. The forking edges will then correspond to splitting up the execution and following a different path, with a different *path-constraint*. As an example, we can look at the execution tree for the program *pow*, that computes a^b.



As we can see, this tree is infinite, since $b = \beta$ and β can be arbitrarily large. In this case, our symbolic execution would run forever if we insisted on exhaustively exploring all possible paths. This illustrates one of the limitations of symbolic execution. For programs with infinite execution trees, we simply cannot exhaust all possible inputs, so we have to restrict our testing to exploring only a finite number of paths.

(Maybe write something about induction over trees, and finite trees allowing for exhaustive search)

3.3.2 the ability(or inability) to decide whether a given path is feasible

(Write something about restrictions on SMT-solvers e.g SAT being NP Complete and some theories may be undecidable)

Chapter 4

Basic symbolic execution for the *SImPL* language

4.1 description

In this chapter we will describe the process of implementing symbolic execution for a simple imperative language called ***SImPL*** .

4.2 Introducing the *SImPL* language

SImPL (**Simple Imperative Programming Language**) is a small imperative programming language, designed to highlight the interesting use cases of symbolic execution. The language supports two types, namely the set integers $\mathbb{N}$ and the boolean values $true$ and $false$. ***SImPL*** supports basic variables that can be assigned integer values, as well as basic branching functionality through an **If - Then - Else** statement. Furthermore it allows for looping through a **While - Do** statement. It also supports top-level functions and the use of recursion.

We will describe the language formally, by the following Context Free Grammar:

$\langle int \rangle$ ::= 0 | 1 | -1 | 2 | -2 | . . .

$\langle Id \rangle$ ::= a | b | c | . . .

$\langle exp \rangle$ ::= $\langle aexp \rangle$ | $\langle bexp \rangle$ | $\langle nil \rangle$

$\langle nil \rangle$ ::= ()

$\langle bexp \rangle$ ::= True | False
 | $\langle aexp \rangle > \langle aexp \rangle$
 | $\langle aexp \rangle == \langle aexp \rangle$

$\langle aexp \rangle$ ::= $\langle int \rangle$ | $\langle id \rangle$
 | $\langle aexp \rangle + \langle aexp \rangle$ | $\langle aexp \rangle$ - $\langle aexp \rangle$
 | $\langle aexp \rangle \cdot \langle aexp \rangle$ | $\langle aexp \rangle$ / $\langle aexp \rangle$
 | $\langle cexp \rangle$

$\langle cexp \rangle$ ::= $\langle Id \rangle$ ($\langle aexp \rangle$*) # Call expression

$\langle stm \rangle$ ::= $\langle exp \rangle$
 | $\langle Id \rangle = \langle aexp \rangle$
 | $\langle stm \rangle$ $\langle stm \rangle$
 | if $\langle exp \rangle$ then $\langle stm \rangle$ else $\langle stm \rangle$
 | while $\langle exp \rangle$ do $\langle stm \rangle$

$\langle fdecl \rangle$ ::= $\langle Id \rangle$ ($\langle Id \rangle$*) $\langle fbody \rangle$

$\langle fbody \rangle$ ::= $\langle stm \rangle$
 | $\langle fdecl \rangle$ $\langle fbody \rangle$

$\langle prog \rangle$ ::= $\langle fdecl \rangle$* $\langle stm \rangle$

Expressions

SImPL supports two different types of expressions, arithmetic expressions and boolean expressions. **arithmetic expressions** consists of integers, variables referencing integers, or the usual binary operations on these two. We also consider function calls an arithmetic expression, and therefore functions must return integer values. **boolean expressions** consists of the boolean values $true$ and $false$, as well as comparisons of arithmetic expressions.

Statements

Statements consists of assigning integer values to variables, *if-then-else* statements for branching and a *while-do* statement for looping. Finally a statement can simply an expression, as well as a compound statement to allow for more than one statement to be executed.

Function declarations

Function declarations consists of an identifier, followed by a list of zero or more identifiers for parameters, and finally a function body which is simply a statement. Functions does not have any side effects, so any variables declared in the function body will be considered local. Furthermore, any mutations of globally defined variables will only exist in the scope of that particular function.

programs

We consider a program to be zero of more top-level function declarations, as well a statement. The statement will act as the starting point when executing a a program.

4.2.1 Interpreting ***SImPL***

In order to work with ***SImPL*** , we have build a simple interpreter using the *Scala* programming language. To keep track of our program state, we define a map

$$env : \langle Id \rangle \to \mathbb{N} \tag{4.1}$$

that maps variable names to integer values. When interpreting a statement or an expression, this map will be passed along. Whenever we interpret a function call, we make a copy of the current environment to which we add the call-parameters. This copy is then passed to the interpretation of the function body, in order to ensure that functions are side-effect free.

A program is represented as an object $Prog$ which carries a map

$$funcs : \langle Id \rangle \to \langle fdecl \rangle$$

from function names to top-level function declarations, as well as a root statement. To interpret the program we simply traverse the AST starting with the root statement.

return values

In order to handle return types, we extend the implementation of the grammar with a non-terminal

$\langle value \rangle$ ::= IntValue | BoolValue | Unit.

Arithmetic expressions will always return an $IntValue$ and Boolean expressions will always return a $BoolValue$. $Unit$ is a special return value which is reserved for *while*-statements.

4.2.2 Symbolic interpreter for *SImPL*

To be able to do symbolic execution of a program written in ***SImPL***, we must extend our implementation to allow for symbolic values to exists. To do this we will add an extra non-terminal to our grammar which will represent symbolic values. Our grammar will then look like

$\langle sym \rangle ::= \alpha \mid \beta \mid \gamma \mid \ldots$

$\langle int \rangle ::= 0 \mid 1 \mid -1 \mid \ldots$

$\vdots$

$\langle aexp \rangle ::= \langle sym \rangle \mid \langle int \rangle \mid \langle id \rangle \mid \ldots$

$\vdots$

Determining feasible paths

In order to determine which execution paths are feasible, we use the **Java** implementation of the *SMT-solver* **Z3**.

Return values

We extend our return values with the terminal $SymValue$ which contains expressions of the type *Expr* from **Z3**, over integers and symbolic values.

Representing the path constraint

To represent the *path-constraint* we implement a class $PathConstraint$ which contains a boolean formula $f = BoolExpr \wedge BoolExpr \wedge ...$ where each expression of type *BoolExpr* is a condition that the input values must satisfy.

Representing the program state

We must extend the capabilities of our environment, so that it does not only map to Integer values, but instead to arbitrary expressions over integers and symbolic values. Therefore we define environment as a map

$$m : \langle Id \rangle \rightarrow SymValue$$

from variable identifiers to values of type $SymValue$.

Execution strategy

The first execution strategy that we implement is a naive approach, where all feasible paths will be explored in *Depth-first* order, starting with the *else*-branch. Note that our definition of *feasible* is any path that we can determine to be satisfiable. This means that *path-constraints* that **Z3** cannot determine the satisfiability of, will be regarded as infeasible, and ignored. This strategy is sufficient for small *finite* programs, but scales badly to programs with large recursion trees, and it runs forever on programs with infinite recursion trees.

Chapter 5

Further extensions

Chapter 6

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] 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`.