Just Fuzz It: Solving Floating-point Constraints Using Coverage-guided Fuzzing

Daniel Liew
dan@su-root.co.uk
Imperial College London
United Kingdom

Cristian Cadar
c.cadar@imperial.ac.uk
Imperial College London
United Kingdom

Alastair F. Donaldson
afd@imperial.ac.uk
Imperial College London
United Kingdom

J. Ryan Stinnett
jryans@gmail.com
Mozilla
United States

ABSTRACT
We investigate the use of coverage-guided fuzzing as a means of proving satisfiability of SMT formulas over finite variable domains, with specific application to floating-point constraints. We show how an SMT formula can be encoded as a program containing a location that is reachable if and only if the program’s input corresponds to a satisfying assignment to the formula. A coverage-guided fuzzer can then be used to search for an input that reaches the location, yielding a satisfying assignment. We have implemented this idea in a tool, Just Fuzz-it Solver (JFS), and we present a large experimental evaluation showing that JFS is both competitive with and complementary to state-of-the-art SMT solvers with respect to solving floating-point constraints, and that the coverage-guided approach of JFS provides significant benefit over naive fuzzing in the floating-point domain. Applied in a portfolio manner, the JFS approach thus has the potential to complement traditional SMT solvers for program analysis tasks that involve reasoning about floating-point constraints.

CCS CONCEPTS
• Theory of computation → Constraint and logic programming
• Software and its engineering → Software testing and debugging.

KEYWORDS
Constraint solving, feedback-directed fuzzing

1 INTRODUCTION
Satisfiability modulo theories (SMT) solvers have made tremendous progress over the last decade [25] and now underpin many important software engineering tools, including symbolic execution engines (e.g. [14, 28, 29, 50, 57]), program verifiers (e.g. [18, 39]) and program synthesis frameworks (e.g. [31, 32]).

Despite these advances, SMT solvers often exhibit limited scalability on large problems [48], and solving can be challenging for certain underlying theories. Scalable SMT solving in the theory of floating-point arithmetic is a particular challenge, and the subject of a lot of recent and ongoing work [3, 11, 35, 42].

An unrelated technology, coverage-guided mutation-based fuzzing, is widely used to automatically find inputs to a system under test (SUT) that expose crashes and potentially exploitable undefined behaviours [40, 45]. For an SUT that has been instrumented to record coverage information, a coverage-guided mutation-based fuzzer takes an initial corpus of inputs and uses genetic algorithms to synthesise further inputs by mutating and combining elements of the corpus. Inputs that cover new parts of the SUT are added to the corpus, under the hypothesis that via further manipulation they may yield inputs that provide even more coverage, and that aiming for high coverage is a good strategy for triggering bugs.

In this paper, we present an in-depth investigation into the potential for coverage-guided mutation-based fuzzing to be used to solve SMT formulas. Our idea is to transform an SMT formula into a program whose input corresponds to an assignment to the free variables of the formula, containing a statement, target, that is reachable if and only if the input corresponds to a satisfying assignment. A coverage-guided fuzzer aims to find inputs that maximise coverage, so when applied to this program it will search relentlessly for an input that reaches target, i.e. for a satisfying assignment to the formula. Our hypothesis is that this technique may sometimes be able to rapidly find satisfying assignments for formulas that are challenging for general-purpose solvers. The method we propose does not intend to help in proving unsatisfiability of formulas.

We present JFS (Just Fuzz it Solver), a prototype constraint solver based on coverage-guided mutation-based fuzzing. JFS is sound: a SAT result can be trusted. However, it is incomplete: JFS could time out and, as discussed above, unsatisfiability cannot be proven. We envision JFS would be run in parallel with a complete solver to form a portfolio solver. JFS was inspired by the limited scalability we have observed for state-of-the-art SMT solvers with respect to floating-point constraints, and currently supports the combination of boolean, bitvector, and floating-point theories, but our idea of SMT solving via coverage-guided fuzzing should be straightforward to adapt to any SMT theory over finite-domain variables.

We present a large experimental evaluation comparing JFS with seven floating-point-capable SMT solvers, over a set of 1344 benchmarks from three different SMT-COMP [58] suites. Our evaluation aims to answer the following research questions:

RQ1 To what extent is coverage-guided mutation-based fuzzing superior to naive random input generation for SMT solving?
RQ2 To what extent can JFS be accelerated via “smart seeds” derived from the formula under analysis and/or the associated SMT theory?

RQ3 How does the execution time of JFS compare with state-of-the-art SMT solvers when applied to satisfiable formulas over Boolean, bitvector, and floating-point variables?

Our main finding is that JFS is competitive with state-of-the-art solvers such as MathSAT5 and Z3 on floating-point constraints, complementing these solvers both in terms of number of solved benchmarks and execution time. By contrast, it is uncompetitive on bitvector-only constraints. In terms of design features, we found coverage-guided mutation-based fuzzing superior to naive random input generation, and the use of smart seeds to be beneficial.

We qualify the relative success of JFS with respect to floating-point constraints by acknowledging that solver support for floating point, as well as available evaluation benchmarks, are relatively recent, while traditional solvers are very mature for bitvector-only constraints, and the benchmarks available for this theory are known to be challenging.

In summary, our main contributions are:

1. The idea of leveraging coverage-guided mutation-based fuzzing to find satisfying assignments to SMT formulas (illustrated concretely via a worked example in §3);
2. JFS, a sound, incomplete solver for floating-point and bitvector constraints based on this idea (§4);
3. An evaluation comparing JFS with seven floating-point-capable SMT solvers over 1344 SMT-COMP benchmarks, addressing the above research questions (§5).

After covering relevant background (§2), we give an overview of JFS (§3), discuss the design and implementation of the approach and tool (§4), and present a detailed experimental evaluation (§5). We then discuss related work (§6), and ideas for future research directions (§7). Throughout the paper we discuss the limitations of JFS and threats to the validity of our approach.

2 BACKGROUND

We provide relevant background on coverage-guided fuzzing (§2.1) and some brief notes on floating-point arithmetic (§2.2).

2.1 Coverage-guided mutation-based fuzzing

Mutation-based fuzzing starts with a set of existing seed inputs, known to already exercise the SUT in some depth, and generates further inputs by mutating and combining seeds. Intuitively, the resulting inputs are much more likely to exercise the SUT in interesting ways compared with inputs generated in a purely random fashion. If code coverage data for the SUT can be obtained, through compile-time or binary instrumentation, a fuzzer can operate in a coverage-guided manner. Code covered by an input can be used as a proxy for measuring how interesting that input is, with an input that covers new code being deemed interesting.

Coverage-guided mutation-based fuzzing combines these ideas: starting from an initial corpus, SUT inputs are generated via mutation. An input that covers new code is added to the corpus to be considered as a seed for future mutation. Typical mutations include making small changes to an input in isolation, and performing “crossover”, where multiple inputs are combined into one. This approach is essentially an evolutionary algorithm [33] where an input is considered fit if it covers new code. An evolutionary algorithm used in this context is part of a broader research area known as search-based test case generation [2].

Two popular coverage-guided mutation-based fuzzers, AFL [45] and LibFuzzer [40] (on which JFS is based) have found numerous bugs in real-world software [41, 44].

2.2 Floating-point arithmetic

A motivating use case for our work is SMT formulas that contain constraints over floating-point variables. We recap here a few terms and concepts that will be used later on.

Single- and double-precision floating-point numbers are represented in SMT-LIB by the $\text{float32}$/$\text{float64}$ types, which correspond to the IEEE-754 binary32/binary64 types [34]. The semantics of most floating-point operations match the process of performing the operation with real number semantics then rounding the result to a nearby floating-point number. Several rounding modes can be used, including rounding to the nearest floating-point number with ties favouring an even binary representation (RNE), and rounding towards positive infinity (RTP). The set of floating-point bit patterns includes special patterns to represent infinities, as well as “not a number” (NaN), which handles the results of computations for which no numerical representation makes sense (such as 0/0).

3 OVERVIEW OF JFS

In brief, JFS uses the following method to find a satisfying assignment to a formula $Q$ presented as a conjunction of constraints:

A program $P$ is constructed such that:

- $P$ takes a sequence of variables as input, with each variable corresponding to a free variable in $Q$.
- $P$ contains a sequence of constraint branches, one per constraint in $Q$, each of which is an if statement whose condition corresponds exactly to the associated constraint.
- $P$ contains a target statement that returns 1 if and only if all the true branches of the constraint branches are traversed.

JFS then passes the program $P$ to a coverage-guided mutation-based fuzzer, which repeatedly runs $P$ with different inputs until an input that reaches the target is found (corresponding to a satisfying assignment to $Q$), or the fuzzer reaches a given time limit. The intuition behind applying a coverage-guided fuzzer is that it will relentlessly try to generate inputs that cover new code. In particular, the program location that returns 1 is a target for the fuzzer.

As an illustration of this idea, consider the example constraints in Listing 1, shown in SMT-LIBv2.5 format [8]. Free variables $a$ and $b$, of type $\text{float64}$ (see §2.2), are declared on lines 1 and 2 respectively. On lines 3 and 4, variables $\text{div}_\text{r}_\text{rne}$ and $\text{div}_\text{r}_\text{rtp}$ are defined to be the division of a by b using the rounding to nearest, ties to even (RNE) and rounding toward positive infinity (RTP) rounding modes, respectively.

The satisfiability problem captured by the example is the conjunction of the constraints specified in the five assert statements. The

\[1\text{Any formula can be transformed to an equisatisfiable formula in conjunctive form, e.g. by using the linear-time Tseytin transformation [65].}\]
first four constraints state that none of \( a, b, \text{div\_rne} \) and \( \text{div\_rtp} \) are NaN; the last states that \( \text{div\_rne} \) is not equal to \( \text{div\_rtp} \).

These constraints are satisfiable. Using C++ hexfloat notation, one satisfying assignment has a set at \( 0x\text{.}410815d750e65p+1022 \) \((\approx 5.65235 \times 10^{10})\) and \( b \) to \( 0x\text{.}821c1b000e7cp+28 \) \((\approx 2.70648 \times 10^{8})\). Dividing \( a \) by \( b \) rounding to nearest (ties to even) yields \( 0x\text{.}000000040000p+1022 \) \((\approx 2.088452 \times 10^{37})\) and rounding toward positive infinity results in \( 0x\text{.}000000000000p+1022 \) \((\approx 2.088453 \times 10^{37})\).  

A possible translation of these constraints into a C++ program is shown in Listing 2, where the guard of each if statement corresponds to a constraint. The fuzzer will repeatedly call FuzzOneInput (line 1), each time passing an input of \( \text{size} \) bytes via the data buffer. If 1 is returned, the input corresponds to a satisfying assignment, otherwise the fuzzer proceeds to try another input.

The program first constructs the free variables from the input buffer data. Variables \( a \) and \( b \) correspond directly to the free variables \( a \) and \( b \) in Listing 1 and are constructed on lines 2 and 3 from the data buffer using bits 0 to 63 (\( a \)), and bits 64 to 127 (\( b \)).

An if statement checks whether \( a \) is NaN (line 4), encoding the constraint on line 5 of Listing 1. Whether \( b \) is NaN is handled analogously (line 5). Variable \( a\_b\_rne \) corresponds to the \( \text{div\_rne} \) macro on line 3 of Listing 1, and is set to the result of calling \( \text{div\_rne}(a, b) \) (line 6). This performs floating-point division rounding the result to the nearest value (ties to even). The assignment to \( a\_b\_rtp \) is analogous, with rounding towards positive infinity.

The checks for whether \( a\_b\_rne \) and \( a\_b\_rtp \) and NaN are handled similarly to the checks for whether \( a \) and \( b \) are NaN (lines 8 and 9). The comparison of \( a\_b\_rne \) and \( a\_b\_rtp \) (line 10) corresponds to the constraint on line 9 of Listing 1.

Finally, on line 11 the function returns 1, which tells the fuzzer that a satisfying assignment has been found. Note that this line is only reachable if all previously evaluated constraints were true.

There are multiple ways of encoding constraints as a program. Listing 2 uses the fail-fast encoding, discussed further in §4.3.

4 DESIGN AND IMPLEMENTATION OF JFS

JFS is written in C++11 and builds on several existing projects: the constraint language and API of Z3 [24] is used for in-memory constraint representation, allowing reuse of Z3’s parser and constraint simplification tactics; Clang and LLVM are used to compile generated C++ code [36]; and the coverage-guided mutation-based fuzzer LibFuzzer [40] is used to fuzz the resulting binary.

JFS accepts an SMT-LIBv2 [8] formula consisting of a conjunction of top-level constraints. Program analysis tools—such as those based on dynamic symbolic execution [15], but not only [66]—generate such conjunctions directly, and as mentioned in §3, any formula can be transformed to an equisatisfiable formula in conjunctive form, e.g. by using the linear-time Tseytin transformation [65].

The design of JFS in principle supports finding satisfying assignments to any theory using finite data types. Our current implementation supports combinations of the Core (i.e. Boolean), FixedSizeBitVectors, and FloatingPoint SMT-LIBv2 theories, over Float32 and Float64 floating-point variables, and bitvector variables of arbitrary widths up to 64 bits.

We now discuss practical issues related to the design of JFS, covering simplification of formulas pre-fuzzing (§4.1); the mapping of formula variables to the program input buffer (§4.2); choices for how to encode the formula as a program (§4.3); and the injection of “smart seeds” to guide the fuzzer (§4.4). We also briefly discuss JFS’s runtime library (§4.5).

4.1 Formula simplification

To make the C++ program that JFS will ultimately generate more friendly to LibFuzzer, JFS first applies the simplification passes detailed in Table 1, in order, to the input formula. The table indicates which passes were already available via calls into the Z3 library, vs. which we implemented using the Z3 API. These passes represent various cheap ways to simplify formulas that we observed to be useful useful during early prototyping of JFS. We remark briefly on the And hoisting pass: JFS uses Z3 to parse constraints, and parsing always returns a single conjunct; the AND hoisting pass simply splits this into independent conjuncts.

If after simplification the formula is syntactically equivalent to false, JFS immediately reports UNSAT without invoking LibFuzzer.

4.2 Input buffer preparation

Having simplified the formula, JFS must decide how to represent free variables of the formula in the program’s input buffer.

First, an equality extraction pass is used to partition the free variables and constants appearing in the formula into equivalence classes based on syntactic equalities, such that members of the same equivalence class are guaranteed to be constrained to be equal. Each resulting class contains at most one constant: if multiple distinct constants were constrained to be equal, JFS would have reported the formula as trivially UNSAT after formula simplification (§4.1).

Each equivalence class is then considered. If a class contains a constant \( c \) then there is no need to reserve space for variable of the class in the input buffer: each variable is declared in the
We have experimented with two ways to encode an SMT formula as a program: \textit{fail-fast} and \textit{try-all}.

### 4.3 Program encodings

We have experimented with two ways to encode an SMT formula as a program: \textit{fail-fast} and \textit{try-all}.

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Description</th>
<th>Already in Z3?</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{And} hoisting</td>
<td>Separates the constraint ((a \land b)) into two separate constraints</td>
<td>No</td>
</tr>
<tr>
<td>Constant propagation</td>
<td>Apply Z3’s \texttt{propagate-values} tactic to propagate constants</td>
<td>Yes</td>
</tr>
<tr>
<td>Duplicate constraint elimination</td>
<td>Removes duplicate constraints from the constraint set</td>
<td>No</td>
</tr>
<tr>
<td>Expression simplification</td>
<td>Invokes Z3’s \texttt{propagate-values} tactic, which performs e.g. constant folding</td>
<td>Yes</td>
</tr>
<tr>
<td>Simplify contradictions</td>
<td>Replaces ((a \land \neg a)) with \texttt{false}</td>
<td>No</td>
</tr>
<tr>
<td>True elimination</td>
<td>Removes constraints of the form \texttt{true} from the constraint set</td>
<td>No</td>
</tr>
</tbody>
</table>

### Program and initialization

In the \textit{fail-fast} encoding, the input buffer is allocated to represent the common value of all free variables in the class, where \(k\) is the width of the associated data type (e.g. \(k = 32\) for \texttt{Float32} variables). The variables are all declared locally in the program and initialized via the same \(k\) bits of the input buffer.

This process is illustrated by the formula of Listing 3 (where \(\_ + \text{zero} 11 53\) denotes the 64-bit positive zero constant) and the associated program in Listing 4. The equivalence classes are \(\{a, b, c\}\) and \(\{d, 0.0\}\). As a result, the input buffer data requires 8 bytes in order to store the double-precision value common to \(a, b\) and \(c\). The \texttt{makeFloatFromData} function initializes \(a\) via this buffer, and the value of \(a\) is then copied into \(b\) and \(c\). Variable \(d\) does not require associated space in the buffer; it is initialized with the constant value 0.0. Because this process fully accounts for equality constraints between variables and constants, such constraints do not need to be modelled in the control flow of the generated program.

Equality extraction both reduces the size of the input buffer, and alleviates LibFuzzer from the onerous task of guessing equality between certain sets of variables.

The input buffer is tightly packed, so that the chunks of data associated with variables need not be aligned to word or even byte boundaries. Chunks are ordered by the order they appear while traversing the input formula. This makes the order deterministic (useful for reproducibility) but arbitrary. Non-aligned accesses make reading from the buffer sub-optimal, but avoids padding bits that have no impact on program behaviour. Such bits would be detrimental to LibFuzzer as it would waste time attempting to mutate those bits to increase coverage. With additional engineering effort we could adapt JFS to make LibFuzzer aware of padding bits and instruct it not to mutate them, allowing the performance benefits associated with better alignment.

### 4.4 Smart seeds

As discussed in §2.1, a coverage-guided mutation-based fuzzer relies on a corpus of initial seed inputs, which in the case of JFS are initial...
valuations of the input buffer. We have experimented with two modes for selecting seeds.

In naive seeds mode, JFS generates two seeds: a buffer of all zeros and a buffer of all ones, which at least provide LibFuzzer’s crossover mutator with a pair of diverse inputs to work with.

In smart seeds mode, seeds are generated as follows. For each distinct data type associated with a free variable (e.g. Float64, bv32, etc.), we construct a set consisting of (1) special values for that type, such as positive/negative zero, infinities, and NaN bit patterns for floating-point types (see §2.2), and bit patterns encoding 0, 1 and -1 for bitvector types; and (2) values of constants of the given type that appear in the input formula. We then construct a seed by randomly sampling from the space of possible input permutations that can be generated from these sets. The number of seeds selected is configurable and set to 100 by default.

Our hypotheses for why smart seeds may be valuable are that (1) special values are often important for particular data types (e.g. a floating-point formula that looks unsatisfiable on first sight often turns out to be satisfiable due to the subtle semantics of NaN values), and (2) the satisfiability of constraints is more likely to depend on values equal or similar to values appearing in the formula than on arbitrary values (with mutations of seeds being likely to yield said similar values). We evaluate the benefits of smart seeds experimentally in §5.

4.5 Runtime library

The program that JFS generates calls into a runtime library that implements the semantics of relevant FloatingPoint and BitVector types from the SMT-LIBv2 standard, handling rounding modes that are natively supported by the x86_64 architecture (all modes except round to nearest, ties away from zero).

5 EVALUATION

We now turn to the evaluation of JFS, comparing it against seven state-of-the-art SMT solvers that support solving floating-point constraints. We discuss the benchmark selection process (§5.1), the solvers and how we configured them (§5.2), and our experimental setup (§5.3). We then present the results of the experiments (§5.4), in the context of the research questions identified in §1. We have made the source code of JFS and all our data sets publicly available.2 3

5.1 Benchmark selection

Table 2 summarises the QF_FP, QF_BVFP and QF_BV SMT-LIB suites from which we have drawn benchmarks for our experiments. A subset of these suites are used in SMT-COMP, the annual SMT solver competition. All benchmarks are quantifier-free (QF), beyond which the suites are built over floating-point (QF_FP), bitvector (QF_BV), and a combination of bitvector and floating-point (QF_BVFP) types.

For each suite, the Suite column provides a reference to the git repository and SHA-1 hash associated with the version of the suite that we used. The SAT and UNSAT columns under Unpruned show the total number of benchmarks in each suite either already labelled SAT or UNSAT, or that were unlabelled but could be classified empirically as SAT or UNSAT by either MathSAT5 or Z3 within 900 seconds seconds on our test platform. The UNKNOWN and Total columns show the number of benchmarks that remained unlabelled, and the total number of benchmarks, respectively.

Since JFS is not designed to prove unsatisfiability, we pruned all benchmarks labelled UNSAT. We also pruned all benchmarks for which the pre-processing steps performed by JFS (§4.1) reduced the benchmark to contain only constants. We believe it was important to remove such trivial benchmarks to focus our evaluation on the effectiveness of fuzzing for constraint solving, rather than the effectiveness of these well-known pre-processing steps. The pruned benchmarks are summarised under Non-trivial, no UNSAT in Table 2. Notice that many SAT benchmarks were found to be trivial, including the vast majority of the QF_FP suite.

The large numbers of remaining QF_BVFP and QF_BV benchmarks would require prohibitive computation resources for our experiments. Therefore, in a final step, we sampled a subset of these benchmarks. To make sure we include benchmarks of varying difficulty, we performed stratified random sampling [4] based on the performance of both MathSAT5 and Z3. That is, for each benchmark suite, we computed two histograms (one for MathSAT5 and one for Z3) of solver execution time with five-second-wide bins. To select a benchmark, first a histogram is selected (round-robin), then a histogram bin is selected (random), and then a benchmark is selected from that bin (random). This process was repeated until the desired number of benchmarks were selected. We selected 5% of the benchmarks from each of the pruned QF_BVFP and QF_BV suites, using the pruned QF_FP suite in its entirety. Details of the final benchmark subsets are summarised under Final subsets in Table 2, which we refer to as QF_BVFP_fs, QF_FP_fs and QF_BV_fs, respectively (where fs stands for "final subset").

5.2 Solver configurations

We compare JFS against seven state-of-the-art constraint solvers for floating-point constraints. For each solver, Table 3 summarizes the version (v) or revision (r) used, and the main technique on which the solver is based. We also include a synthetic portfolio solver (JFS+MathSAT5) to aid discussions of using JFS in a portfolio setting. JFS+MathSAT5 models a complete portfolio solver that runs both JFS-LF-SS and MathSAT5 in parallel and returns the answer from which ever solver answers first. It is synthetic because solving time is computed as the minimum of the solving times of existing runs of JFS-LF-SS and MathSAT5. JFS-LF-SS and MathSAT5 are combined because as we shall see, they are the best performing JFS configuration and solver for QF_FP respectively.

We acknowledge that some of these solvers are capable of proving UNSAT as well as SAT, while JFS is only capable of proving SAT. This might appear to give JFS an advantage, but we are not aware of any way to configure those solvers to only focus on SAT, hence we believe there is no fairer way of performing the comparison.

At the time experiments were run, XSat had not been officially released; we use a version of the solver uploaded to STAR-EXEC for the 2017 SMT-COMP competition.

The CORAL, goSAT and XSat solvers do not support bitvector reasoning, thus we can only apply them to the QF_FP_fs benchmark suite. Instead of the SMT-LIBv2.5 format, CORAL uses its own...
constraint language that only supports a subset of the semantics of the QF_FP theory. To allow a best-effort comparison with CORAL, we have implemented a tool to convert SMT-LIBv2.5 constraints into this language.

We run each solver using its default configuration, edited if necessary to enable floating-point reasoning and to enforce SMT-LIB compliance. Exceptions are CORAL, which we run using options suggested by the developers as we were unsure how to best invoke the solver, and MathSAT5, which comes with a file describing preferred options for each benchmark suite. (smtcomp2015_main.txt).

We run CORAL in two distinct modes: alternating variable method (CORAL-AVM) and particle swarm optimisation (CORAL-PSO). We run JFS in three modes: using LibFuzzer with naive seeds (JFS-LF-NS), using LibFuzzer with smart seeds (JFS-LF-SS), and using purely naive random input generation, i.e. without LibFuzzer (JFS-NR). In all cases the fail-fast encoding is used (see §4.3).

Where solvers support setting a random seed, we use a fixed value to try to enable reproducible results.

5.3 Experimental setup

We ran the 11 configurations (eight solvers, with CORAL in two and JFS in three configurations respectively) on a machine with two Intel® Xeon® E5-2450 v2 CPUs (8 physical cores each) with 256GiB of RAM running Ubuntu 16.04 LTS. Each solver was run five times per benchmark with a timeout of 900 seconds per run and with a fixed random seed (if supported). The repeat runs of a solver are used to compute average execution time and observe non-deterministic behaviour. To allow experiments to complete within a reasonable time-frame, each solver was executed in parallel over the set of benchmarks, with at most 13 benchmarks running in parallel.

Each time a solver is run on a benchmark we record a result label. If solver reports UNSAT, crashes, or hits the memory or time limit, the result is labelled as UNSAT. If the solver reports SAT (UNSAT) and that matches the expected satisfiability of the benchmark or the expected satisfiability is UNKNOWN then the result is labelled as SAT (UNSAT). If the solver reports SAT (UNSAT) and the expected satisfiability of the benchmark is UNSAT (SAT) then the result is labelled as WRONG.

We combined results labels for repeat runs of a solver on a benchmark as follows: if at least one label is SAT (UNSAT) and all labels are either SAT (UNSAT) or UNKNOWN, the combined label is SAT (UNSAT). If at least one label is WRONG or the labels include a mixture of SAT and UNSAT, the combined label is WRONG. Otherwise, in the case where all labels are UNKNOWN, the combined label is UNKNOWN.

To combine the execution times (wall clock time), the arithmetic mean and confidence intervals (99.9%) are computed. Mean execution times are only considered distinguishable between solvers if their confidence intervals do not overlap.

5.4 Results

We now present and discuss our experimental results, relating them to the research questions identified in §1. In §5.4.1 we address RQ1 and RQ2 by comparing different JFS configurations. Then in §5.4.2 we compare the overall best JFS configuration found in §5.4.1 against other solvers in order to address RQ3.

To visualise solver performance we use quantile plots (e.g. Figure 1). Each curve on a plot corresponds to a solver configuration. A curve is plotted by computing a score for each run on a benchmark (1 for correct, −1 for wrong, and 0 for unknown), sorting correct results by solver execution time and then plotting accumulating score against solver execution time. An extra leftmost point is then added to the curve and all other points are offset along the x-axis by this value. The x value of this point is the sum of negative scores.

The resulting quantile plot has the following properties: (1) the x-value of the leftmost point on a curve indicates the number of incorrect solver answers (e.g. a value of −5 on the x-axis indicates the solver incorrectly reported satisfiability on five benchmarks); therefore, the x-value of the leftmost points can be compared between curves (ranked by least number of wrong answers); (2) the x-value of the right-most point on a curve is the difference between the number of correct vs. incorrect solver answers; therefore, the x-value of the rightmost points can be compared between curves (ranked by total solver score); (3) the total execution time of a solver on correctly solved benchmarks is equal to the area under the curve. We cannot compare the points with the same y-value between curves because the points do not necessarily refer to the same benchmark. However, we can compare the general shapes of curves. The quantile plots that follow are best viewed in colour.

5.4.1 JFS configuration comparison. We compare JFS in three different configurations JFS-LF-NS, JFS-LF-SS, JFS-NR (§5.2) on the three benchmark suites.
to determine which configuration was fastest, either because the solver execution time confidence intervals overlapped or because none of the solvers reported SAT. It is expected that JFS-NR might sometimes be faster because it has lower overhead than the other configurations (e.g., no coverage instrumentation, no seeds to read).

For \( \text{QF}_\text{FP} \), Figure 2 shows that JFS-LF-NS solved the most benchmarks (685), followed by JFS-LF-SS (684) and finally JFS-NR (656). We can see that the shape of the curves for the LibFuzzer configurations are similar, suggesting little difference in overall performance between them. However, the naive random configuration is clearly worse. These results provide positive support for RQ1, and are inconclusive with respect to RQ2.

Quantile plots do not tell the complete story. Tables 4 and 5 show JFS-LF-SS similarity, complementarity, and limitations for the \( \text{QF}_\text{FP} \) and \( \text{QF}_\text{BVFP} \) benchmarks, compared to the other JFS configurations. In each table, the "Both" column states the number of benchmarks shown to be satisfiable by both JFS-LF-SS and the other JFS configuration. The "Only LF-SS (Only other)" column shows the number of benchmarks that were shown to be satisfiable by JFS-LF-SS (other configuration) and not by the other configuration (JFS-LF-SS). The "Neither" column shows the number of benchmarks that were shown to be satisfiable by neither JFS-LF-SS nor the other configuration. Each row of the table corresponds to the other solver (specified by the "Solver" column). The "All above" row has a special meaning and is a combination of all the above results. For the "All above" row: the "Both" table cell is the union of all benchmarks that both JFS-LF-SS and another JFS configuration managed to solve (i.e., it is a union of intersections, not an intersection of intersections); the "Only LF-SS" table cell is the number of benchmarks found satisfiable by JFS-LF-SS and none of the other configurations; the "Only other" table cell is the union of all benchmarks found satisfiable by another configuration and not JFS-LF-SS; and the "Neither" table cell is the number of benchmarks not found satisfiable by any JFS configuration.

For \( \text{QF}_\text{FP} \), Table 4 shows that JFS-LF-SS and JFS-LF-NS are quite similar (67.50% of the benchmarks solved by both and 27.50% by neither); perhaps unsurprising given that they only differ in the seeds fed to LibFuzzer. By contrast, JFS-NR is less similar, with 15.0% of benchmarks solved only by JFS-LF-SS.

In terms of complementarity, JFS-LF-SS always solved benchmarks that the other configurations did not. Although the converse is true (other configurations solving benchmarks that JFS-LF-SS did not) it is less frequent. Looking at limitations, 26.88% of the benchmarks were not solved by any JFS configuration.

For \( \text{QF}_\text{BVFP} \), while the quantile plot of Figure 2 suggests that JFS-LF-NS performs slightly better than JFS-LF-SS due to the number of benchmark solved, Table 5 shows that there are two benchmarks that only JFS-LF-SS solved and three that only JFS-LF-NS solved, showing that neither configuration is strictly superior to the other on this benchmark suite. Regarding limitations, the JFS

Table 4: JFS-LF-SS vs. other JFS configurations over \( \text{QF}_\text{FP} \).

<table>
<thead>
<tr>
<th>Solver</th>
<th>Both</th>
<th>Only LF-SS</th>
<th>Only other</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>JFS-LF-NS</td>
<td>108 (67.50%)</td>
<td>6 (3.75%)</td>
<td>2 (1.25%)</td>
<td>44 (27.50%)</td>
</tr>
<tr>
<td>JFS-NR</td>
<td>90 (56.25%)</td>
<td>24 (15.0%)</td>
<td>1 (0.62%)</td>
<td>45 (28.12%)</td>
</tr>
<tr>
<td>All above</td>
<td>108 (67.50%)</td>
<td>6 (3.75%)</td>
<td>3 (1.88%)</td>
<td>43 (26.88%)</td>
</tr>
</tbody>
</table>

---

Figure 1: Comparing JFS configurations over \( \text{QF}_\text{FP} \).

Figure 2: Comparing JFS configurations over \( \text{QF}_\text{BVFP} \).
Table 5: JFS-LF-SS vs. other JFS configurations over QF_BVFP_f1.

<table>
<thead>
<tr>
<th>Solver</th>
<th>Both</th>
<th>Only LF-SS</th>
<th>Only other</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>JFS-LF-NS</td>
<td>682 (97.15%)</td>
<td>2 (0.28%)</td>
<td>3 (0.43%)</td>
<td>15 (2.14%)</td>
</tr>
<tr>
<td>JFS-NR</td>
<td>655 (93.30%)</td>
<td>29 (4.13%)</td>
<td>1 (0.14%)</td>
<td>17 (2.42%)</td>
</tr>
<tr>
<td>All above</td>
<td>682 (97.15%)</td>
<td>2 (0.28%)</td>
<td>3 (0.43%)</td>
<td>15 (2.14%)</td>
</tr>
</tbody>
</table>

configurations performed collectively well on this suite, with only 2.14% not solved by any JFS configuration.

Overall, for formulas involving floating-point constraints, the results of this subsection show that using coverage-guided fuzzing over naive random input generation offers benefit, supporting RQ1. The results also partially support RQ2 in this domain, showing that smart seeds improve the performance of JFS over QF_FP_f1. While the performance results for JFS-LF-SS and JFS-LF-NS over QF_BVFP_f1 do not reveal a clear winner, we use JFS-LF-SS as the JFS configuration for comparison against other solvers in §5.4.2 due to its superior performance on the QF_FP_f1 suite.

5.4.2 JFS compared with other solvers. We now address RQ3 by comparing the JFS-LF-SS configuration of JFS against seven solvers on the QF_FP_f1 benchmarks and four on the QF_BVFP_f1 and QF_BV_f benchmarks.

Comparison over QF_FP_f1. The quantile plot of Figure 3 summarises performance results for the eight non-portfolio solvers plus JFS+MathSAT5 over QF_FP_f1 benchmarks. The leftmost points for XSat and COLIBRI indicate that they solved 34 and 5 wrong answers, respectively. In all cases this was due to UNSAT being reported for a SAT-labelled benchmark.

In terms of the number of benchmarks found to be satisfiable, JFS+MathSAT5 was the most successful (126) followed by MathSAT5 (125), JFS (114), CVC4 (110), COLIBRI (104), Z3 (102), goSAT (91), XSat (69), CORAL-PSO (60), and finally CORAL-AVM (31). Even though JFS does not rank the highest by number of benchmarks solved, we can see from the shape of the curves that JFS’s total solving time is significantly smaller than MathSAT5’s which solved the most benchmarks out of the non-portfolio solvers. The JFS+MathSAT5 synthetic portfolio solver illustrates that a portfolio combination of JFS-LF-SS and MathSAT5 would perform well because it would solve the most benchmarks and in less time on average.

Table 6 shows JFS’s capability, complementarity, and limitations for the QF_FP_f1 benchmarks. The columns and special All above row have the same meaning as discussed for Table 4 in §5.4.1. Table 6 shows great deal of similarity (Both column) between MathSAT5 and JFS, followed by COLIBRI and CVC4, and then Z3. The similarity with the other search-based solvers (CORAL-AVM, CORAL-PSO, goSAT, and XSat) is somewhat lower.

JFS complements every other non-portfolio solver, i.e. there is at least one benchmark that JFS can solve and the other solver cannot. However, every benchmark solved by JFS can be solved by at least one other solver. For the search-based solvers (CORAL, goSAT, JFS, and XSat) JFS finds many benchmarks to be satisfiable that the other solver does not. This shows that out of all the search-based solvers, JFS is the most competitive on the QF_FP_f1 benchmark suite. In terms of limitations, every solver except CORAL-AVM finds some benchmarks to be satisfiable that JFS does not (i.e. most solvers are able to complement JFS). There are also some benchmarks that neither JFS, nor another solver manage to show as satisfiable.

JFS is also complementary in terms of execution time. Figures 4 and 5 show scatter plots comparing the execution time of JFS against MathSAT5 and CVC4 respectively. We show MathSAT5 and CVC4 here because these are the solvers that found the highest number of benchmarks to be satisfiable that JFS did not. On these plots, each point represents a benchmark. A diagonal line (y = x) is drawn, upon which a benchmark would lie if both solvers solved the benchmark in an identical amount of time. Points that appear below the diagonal are cases where JFS was faster, and points above the line are cases where the other solver was faster. The number of points where this is the case (and where confidence intervals do not overlap) are shown on the figures along with annotation indicating how many points are cases where both solvers reached a timeout. These plots only show cases where both solvers either reported SAT or reached a timeout because it does not make sense to compare execution times if one of the solvers crashed. The plots show that the solvers are highly complementary, with JFS being faster for 86 benchmarks in each case, while MathSAT5 was faster for 27 benchmarks and CVC4 for 24.

In relation to RQ3, these results show that JFS is very competitive with other solvers on the QF_FP_f1 benchmarks and is able to complement every solver considered.

Table 6: JFS compared to other solvers over QF_FP_f1.

<table>
<thead>
<tr>
<th>Solver</th>
<th>Both</th>
<th>Only JFS</th>
<th>Only other</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLIBRI</td>
<td>98 (61.25%)</td>
<td>16 (10.00%)</td>
<td>6 (3.75%)</td>
<td>40 (25.00%)</td>
</tr>
<tr>
<td>CORAL-AVM</td>
<td>31 (19.38%)</td>
<td>83 (51.88%)</td>
<td>0 (0.00%)</td>
<td>46 (28.75%)</td>
</tr>
<tr>
<td>CORAL-PSO</td>
<td>59 (36.88%)</td>
<td>55 (34.38%)</td>
<td>1 (0.62%)</td>
<td>45 (28.12%)</td>
</tr>
<tr>
<td>CVC4</td>
<td>98 (61.25%)</td>
<td>16 (10.00%)</td>
<td>12 (7.58%)</td>
<td>34 (21.25%)</td>
</tr>
<tr>
<td>goSAT</td>
<td>86 (53.75%)</td>
<td>28 (17.50%)</td>
<td>5 (3.12%)</td>
<td>41 (25.62%)</td>
</tr>
<tr>
<td>MathSAT5</td>
<td>113 (70.62%)</td>
<td>1 (0.00%)</td>
<td>12 (7.50%)</td>
<td>34 (21.25%)</td>
</tr>
<tr>
<td>XSat</td>
<td>62 (38.75%)</td>
<td>52 (32.50%)</td>
<td>7 (4.38%)</td>
<td>39 (24.38%)</td>
</tr>
<tr>
<td>Z3</td>
<td>96 (60.00%)</td>
<td>18 (11.25%)</td>
<td>6 (3.75%)</td>
<td>40 (25.00%)</td>
</tr>
<tr>
<td>All above</td>
<td>114 (71.25%)</td>
<td>0 (0.00%)</td>
<td>21 (13.12%)</td>
<td>25 (15.62%)</td>
</tr>
</tbody>
</table>
JFS: Solving FP Constraints using Coverage-guided Fuzzing

![Figure 4](image1.png)

Figure 4: Scatter plot comparing the execution time of MathSAT5 and JFS on the QF_FP\_fs benchmark.

![Figure 5](image2.png)

Figure 5: Scatter plot comparing the execution time of CVC4 and JFS on the QF_FP\_fs benchmark.

Comparison over QF_BVFP\_fs. Figure 6 shows a quantile plot comparing JFS against JFS+MathSAT5 and the other three non-portfolio solvers that support the QF_BVFP\_fs suite. The plot shows that the none of the solvers reports incorrect answers and that they all report a similar number of benchmarks as satisfiable. JFS+MathSAT5, CVC4, MathSAT5 and Z3 report 699 benchmarks as satisfiable, followed by JFS with 684, and COLIBRI with 666. The figure also shows that for every solver, over 600 benchmarks were solved in under a second. This suggests that the benchmark suite (despite our best efforts during stratified sampling) is not well balanced in terms of difficulty and may not accurately reflect the kind of constraints that might be encountered in practice. Table 7 shows the similarity, complementarity, and limitations of JFS on this benchmark compared to other non-portfolio solvers. The table shows a high degree of similarity between the solvers and that JFS is only able to complement COLIBRI. Every solver is able to solve benchmarks that JFS is unable to solve. However, if we make scatter plots comparing the execution time of JFS with that of other solvers, we find in each case a significant number of benchmarks where JFS solves the constraints faster (56 faster than CVC4, 27 faster than COLIBRI, 55 faster than MathSAT5, and 69 faster than Z3). We omit these plots for brevity but they look very similar to Figures 4 and 5.

![Figure 6](image3.png)

Figure 6: Quantile plot comparing the performance of solvers on the QF_BVFP\_fs benchmarks.

Table 7: JFS compared to other solvers over QF_BVFP\_fs.

<table>
<thead>
<tr>
<th>Solver</th>
<th>Both</th>
<th>Only JFS</th>
<th>Only other</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLIBRI</td>
<td>661 (94.16%)</td>
<td>23 (3.28%)</td>
<td>5 (0.71%)</td>
<td>13 (1.85%)</td>
</tr>
<tr>
<td>CVC4</td>
<td>684 (97.44%)</td>
<td>0 (0.00%)</td>
<td>15 (2.14%)</td>
<td>3 (0.43%)</td>
</tr>
<tr>
<td>MathSAT5</td>
<td>684 (97.44%)</td>
<td>0 (0.00%)</td>
<td>15 (2.14%)</td>
<td>3 (0.43%)</td>
</tr>
<tr>
<td>Z3</td>
<td>684 (97.44%)</td>
<td>0 (0.00%)</td>
<td>15 (2.14%)</td>
<td>3 (0.43%)</td>
</tr>
<tr>
<td>All above</td>
<td>684 (97.44%)</td>
<td>0 (0.00%)</td>
<td>15 (2.14%)</td>
<td>3 (0.43%)</td>
</tr>
</tbody>
</table>

With reference to RQ3, these results show that JFS is competitive over QF_BVFP\_fs, complementing COLIBRI in the number of benchmarks solved, and all other solvers in terms of execution time. However, as discussed the results across all solvers suggest that QF_BVFP\_fs may not be an especially challenging suite.

Comparison over QF_BV\_fs. JFS is not competitive on the QF_BV\_fs suite, finding only 22 benchmarks satisfiable, compared to e.g. 419 for Z3 and 344 for MathSAT5. (We omit the associated quantile plot for space reasons.) However, for each solver except CVC4, there are always cases where JFS is able to solve some benchmarks faster.

We suspect two main reasons for the poor performance of JFS on the bitvector-only theory, compared to the theories involving floating point. First, floating-point constraints result in much more complex circuits, which often blow-up the underlying SAT solvers used by state-of-the-art SMT solvers. As a result, a more lightweight approach like the one used by JFS is competitive on these theories.

Second, bitvector solvers have been available for over a decade, which has allowed a set of difficult and challenging benchmarks to be developed over a long period of time. These benchmarks likely evolved in difficulty as bitvector solvers gradually increased their capability. On the other hand, solvers for floating-point constraints are comparatively new and have had much less time to develop. As a consequence, the available floating-point benchmarks are a reflection of the relatively immature floating-point constraint solvers currently available.

It is also worth drawing an analogy with coverage-guided fuzzers applied to bug finding (their usual domain). These fuzzers are typically good at finding shallow bugs, and can only excel at finding...
deep bugs with a large amount of compute time, good seeds, or
domain-specific knowledge. It could be the case that the floating-
point benchmarks currently available in SMT-LIB are the equivalent
of shallow programs, where bugs are easy for a fuzzer to find.

In summary, with respect to RQ3: the results across all three
benchmark suites show that JFS is highly competitive on two suites
(both involving floating point), and uncompetitive on the bitvector
benchmark suite.

6 RELATED WORK
There is a large body of existing work that seeks to improve solv-
ing floating-point constraints. The CORAL [62] and FloPSy [35]
solvers apply meta-heuristic search techniques to try to find satis-
fying assignments to floating-point constraints. Like JFS, these
methods are incomplete because they can only show satisfiability.
All solvers construct a fitness function which they attempt to max-
imise. JFS’s fitness function is coarse—the number of new branches
covered—in contrast to CORAL’s and FloPSy’s fitness functions,
which gradually change as candidate solutions get closer to a satis-
fying assignment. Despite the coarseness of JFS’s fitness function,
our results show that JFS performs better overall than CORAL, both
in terms of the number of benchmarks it can show to be satisfiable,
and in execution time. We could not easily compare with FloPSy
due to its tight integration with Pex [64], the symbolic execution
tool it is designed to work with.

CORAL supports using an interval solver to improve the quality
of its initial candidate inputs. It’s likely we could apply a similar
approach in JFS to generate higher quality seeds for the fuzzer.

The goSAT [9] and XSat [26] solvers both reformulate finding a
satisfying assignment as a mathematical optimisation problem and
apply existing mathematical optimisation algorithms to try to find a
global minimum. This is similar in spirit to JFS, FloPSy and CORAL
in that the functions that goSAT and XSat seek to minimise are
essentially fitness functions. The difference is in the algorithms used
to perform the search. Like JFS, this strategy is incomplete. Again,
despite JFS’s coarser fitness functions, the experimental evaluation
found JFS to perform better on those benchmarks.

CVC4 [6], MathSAT5 [17], SONOLAR [51] and Z3 [24] solve
floating-point constraints by transforming floating-point opera-
tions into bitvector circuits and then bit-blasting these into a SAT
problem. This problem is then solved using a SAT solver. Unlike
JFS, these solvers are complete, but they can end up generating very
large SAT problems, which are difficult to solve. Like JFS, these
solvers support a combination of the bitvector and floating-point
SMT-LIBv2.5 theories. Our comparison with CVC4, MathSAT5 and
Z3 indicates that the approaches are complementary, particularly
for the floating-point benchmarks, suggesting these solvers would
likely benefit from incorporating a JFS-style search-based strategy
with their existing strategies, to form a portfolio solver. We did not
compare JFS with SONOLAR, but given that its design is similar
to that of SAT based solvers, we do not expect such experiments
to change our main conclusions. A prior study comparing SAT-
based solving with random and heuristic solvers also found that a
portfolio approach performs best [63].

COLIBRI [13] and FPCS [46] use interval solving as a complete
method for solving floating-point constraints. As for the comparison
with SAT based solvers, our comparison with COLIBRI showed
complementarity, suggesting that these solvers could also benefit
from incorporating a search-based strategy. We did not compare
against FPCS because it is not publicly available.

REALIZER [38] tries to solve floating-point constraints by trans-
forming (in an equisatisfiable manner) floating-point constraints
into constraints over reals, using Z3 as a back-end to solve these
constraints. REALIZER’s strategy is particularly suitable for working
with constraints that check the accuracy of floating-point ex-
pressions compared to their real counterparts. JFS cannot do this
because it cannot handle constraints over reals. We have not yet
had time to compare JFS with REALIZER.

More generally, floating-point constraint solving has gathered a
lot of attention from the research community, with several tools
based on symbolic execution, model checking, abstract interpreta-
tion, etc. using it to perform test-case generation, precision tuning,
verification, equivalence checking, peephole optimizations, branch
instability assessment, etc. involving floating-point code [1, 3, 5, 10–
12, 16, 19–23, 27, 30, 35, 37, 42, 43, 47, 49, 52–56].

7 CONCLUSION
We have investigated using coverage-guided mutation-based fuzzing
to prove satisfiability of SMT formulas over finite variable domains,
and floating-point constraints in particular, via a prototype solver.
JFS. Our main experimental findings are that in the domain of
floating-point constraints, solving via coverage-guided fuzzing out-
performs solving via naive fuzzing, and performance can be further
improved by generating initial seeds in a smart manner; JFS is highly
competitive with and complementary to all solvers we compared
with in the floating-point domain; and JFS is much less effective
when applied to the domain of bitvectors. Our synthetic portfolio
solving results indicate that JFS’s complementary nature would
make it a useful component in a portfolio solver.

In future work, we would like to better understand the properties
of benchmarks that dictate whether JFS performs well, with a view
to developing heuristics to help decide when it would be beneficial
to apply JFS. A first step in this direction would be to use model
counting solvers to understand whether suitability for solving via
fuzzing relates to number of solutions. A practical problem here is
that model counting suffers from limited scalability.

Regarding our smart seeds, smarter seeds could be generated
based on domain-specific knowledge about the context in which JFS
is being used. For example, if JFS were integrated with a symbolic
execution engine, seeds encoding knowledge about feasible paths
(and thus feasible inputs) could be communicated from the symbolic
execution engine to JFS. We also envisage several improvements
to the fuzzing component of JFS: designing mutators tailored to
the context of SMT formulas would likely be beneficial; the fuzzer
could be made aware of data flow, using information about the bytes
that caused a constraint to become satisfied to guide mutations;
and candidates for mutation could be prioritised according to the
number of constraints they satisfy, which we hypothesise would
lead to faster synthesis of satisfying assignments.

Acknowledgements. This research was generously sponsored by
the UK EPSRC through grants EP/N007166/1, EP/P010040/1 and
EP/R006865/1.


