Automatically Testing Implementations of Numerical Abstract Domains

Alexandra Bugariu
Department of Computer Science, ETH Zurich
Switzerland
alexandra.bugariu@inf.ethz.ch

Valentin Wüstholz
Department of Computer Science, ETH Zurich
Switzerland
wuestholz@gmail.com

Maria Christakis
MPI-SWS
Germany
maria@mpi-sws.org

Peter Müller
Department of Computer Science, ETH Zurich
Switzerland
peter.mueller@inf.ethz.ch

ABSTRACT

Static program analyses are routinely applied as the basis of code optimizations and to detect safety and security issues in software systems. For their results to be reliable, static analyses should be sound (i.e., should not produce false negatives) and precise (i.e., should report a low number of false positives). Even though it is possible to prove properties of the design of a static analysis, ensuring soundness and precision for its implementation is challenging. Complex algorithms and sophisticated optimizations make static analyzers difficult to implement and test.

In this paper, we present an automatic technique to test, among other properties, the soundness and precision of abstract domains, the core of all static analyzers based on abstract interpretation. In order to cover a wide range of test data and input states, we construct inputs by applying sequences of abstract-domain operations to representative domain elements, and vary the operations through gray-box fuzzing. We use mathematical properties of abstract domains as test oracles. Our experimental evaluation demonstrates the effectiveness of our approach. We detected several previously unknown soundness and precision errors in widely-used abstract domains. Our experiments also show that our approach is more effective than dynamic symbolic execution and than fuzzing the test inputs directly.

CCS CONCEPTS

• Software and its engineering → Software testing and debugging.

KEYWORDS

soundness testing, precision testing, abstract interpretation

1 INTRODUCTION

Static program analyses compute semantic properties of programs, which are the basis for program optimizations and for detecting program errors and security vulnerabilities. Since most properties of programs are undecidable, static analyses approximate the set of possible program behaviors. For its results to be reliable, a static analysis should be sound and precise. A sound analysis considers each possible program behavior, that is, computes an over-approximation of all possible behaviors. Consequently, sound analyses do not produce false negatives. A precise analysis computes a tight approximation to minimize the number of false positives.

Many static analyses are based on the abstract-interpretation framework [20]. In this framework, abstractions of the program state are represented by elements of abstract domains; for instance, numerical abstract domains may track intervals of possible values for numerical variables or constraints between them. The semantics of program operations is represented by abstract transformers, which specify the effect of an operation on the abstract state.

Even though it is possible to prove properties of the design of a static analysis, ensuring soundness and precision for its implementation is challenging. In fact, implementations of abstract domains are often complex and highly optimized in order to maximize performance and scalability [46]. Errors in these implementations likely affect the usefulness of all static analyzers that build on them.

Imagine a static analysis that abstracts numerical variables to intervals of possible values. For instance, the abstract value [0, 5] for an integer variable x expresses that, in each program execution, the concrete (actual) value of x satisfies 0 ≤ x ≤ 5. The example in Fig. 1 illustrates a potential soundness error due to arithmetic overflow. Without prior knowledge about parameter p, its abstract value is [INT_MIN, INT_MAX], and consequently, the abstract value of a after the assignment is [INT_MIN + 1, INT_MAX + 1]. If this abstract domain is implemented using bounded integers, a naive implementation of the addition operator will lead to an overflow and...
produce $[\text{INT\_MIN} + 1, \text{INT\_MIN}]$. This empty interval indicates that the code after the assignment is unreachable, which is unsound. In particular, this unsound result might mask program errors and security vulnerabilities in subsequent code.

In this paper, we assess the effectiveness of existing automatic testing techniques in finding, among other errors, soundness and precision issues in widely-used implementations of abstract domains. We generate test inputs using a novel combination of existing ideas: starting from a pool of pre-selected values, we apply abstract-domain operations to create representative domain elements, and vary the operations by employing an off-the-shelf gray-box fuzzing, such as AFL [7] or LibFuzzer [6], to maximize coverage. As in earlier work by Midtgaard and Møller [40], we use mathematical properties of abstract domains as test oracles. However, we target real-world implementations of complex abstract domains (e.g., APRON’s Octagons domain [41] and ELINA’s Polyhedra domain [46]), and we extend the set of tested properties by including more precision properties and by approximating termination properties.

Our evaluation on several abstract domains of the APRON [31] and ELINA [46] libraries shows that our combination effectively detects soundness and precision problems in complex, mature implementations. In particular, we show that it is more effective than purely relying on existing test case generation techniques, such as gray-box fuzzing and dynamic symbolic execution (DSE, also known as cocolic testing) [14, 28].

Our main contributions are the following:

1. We present a novel combination of automatic test case generation techniques to detect, among other errors, soundness and precision issues in implementations of abstract domains.
2. We demonstrate that our technique effectively finds both seeded and real errors in widely-used implementations of numerical abstract domains.
3. We show that our technique tests abstract domains more effectively than standard DSE or gray-box fuzzing.

Our experience is useful for developers of abstract domains to ensure soundness and precision of their implementations. It is also useful for developers of static analyzers to assess the quality of available abstract-domain libraries. Even if the design of an abstract domain intentionally sacrifices soundness in favor of other qualities [16, 37], it is still important to check the implementation for unintentional unsoundness caused by implementation errors.

### Outline
The next section summarizes some background on abstract interpreters. Sect. 3 gives an overview of our approach, and Sect. 4 explains the technical details. In Sect. 5, we present our experimental evaluation on real-world implementations of abstract domains. We discuss related work in Sect. 6 and conclude in Sect. 7.

## 2 BACKGROUND

Abstract interpretation [20] is a theoretical framework for expressing static analyses, used by many industrial analyzers, such as Aostre [12] and Clousot [26]. It relies on abstract domains to represent abstractions of concrete program states, and on abstract transformers to model the behavior of program instructions, such as assignments and conditionals.

Abstract domains are often reused across many different program analyses. For example, most static analyzers employ numerical domains, which are, therefore, the focus of this paper. Widely-used numerical domains include Intervals [19], which capture the range of values for each variable, Octagons [41], which can also express simple relations between two variables, Polyhedra [23], which can capture linear inequalities between arbitrarily many variables, and Zonotopes [29], which express affine relations.

Most abstract domains are represented by complete lattices of the form $(\hat{A}, \subseteq, \perp, \top, \sqcup, \sqcap)$. $\hat{A}$ denotes the set of abstract elements $\hat{x}$, which are partially ordered by inclusion $\subseteq$. Each abstract element represents a set of constraints, i.e., mathematical relations between variables and constants. The bottom element $\perp$ is the least element of the lattice; it represents the empty set of concrete states and corresponds to an unsatisfiable set of constraints. The top element $\top$ is the greatest element and represents the universal set of concrete states; that is, all variables are unconstrained.

$\sqcup$ and $\sqcap$ are the join and meet operators, which are used to obtain the union, respectively the intersection, of two abstract elements. Additionally, an abstract domain whose lattice has an infinite height requires a widening operator ($\triangledown$) to ensure that the analysis eventually reaches a fixed point. Some domains (such as Intervals and Octagons) also support a narrowing operator ($\triangle$), which can improve the precision of the analysis [22].

Abstract transformers are typically specific to a given analysis and programming language, but some transformers are universal building blocks for many analyses. These include assign to represent an assignment, cond to assume a condition to hold, and project to eliminate any previous information about one of the variables (e.g., when a new value of a variable is read from a file). Here, we focus on these transformers because they are the most complex to implement [47] (and thus the most error prone). A generalization of our approach to other transformers is mostly straightforward.

## 3 OVERVIEW

Since implementations of abstract domains are often used as libraries by many different program analyses, we apply a unit testing approach. Compared to system testing, unit testing allows us to specify generic test oracles, which are independent of a specific abstract domain or static analysis. Moreover, it facilitates the generation of test data because abstract-domain elements can be generated much more easily than input programs for a whole analyzer [40]. In this section, we give an overview of the three main ingredients of our automatic unit test approach: test oracles, test input generation, and test drivers. Details will be presented in Sect. 4.

### 3.1 Test Oracles

The abstract interpretation framework prescribes a number of properties of the domain operators and abstract transformers (collectively referred to as domain operations in the rest of the paper), which are required for soundness. For instance, if the abstract element $\hat{x}$ (capturing the pre-state of an assignment) is reachable

```plaintext
int foo(int p) {
    int a := p + 1;
    ...
}
```

Figure 1: Potential soundness error due to overflow.
We use these general soundness conditions as test oracles. We also identify a number of general precision conditions, whose violation indicates that the result of a domain operation is over-approximated more than necessary. For this purpose, we compare the result of a domain operation to the best transformer, that is, the most precise result representable in a given abstract domain. For instance, the result of intersecting top with an element should be equal to the element itself:
\[ T \cap \bar{x} = \bar{x} \]  
(b)
Moreover, we check that widening and narrowing converge within a given number of iterations, which ensures the termination of any fixed-point computation in which they are used.
Note that all the properties considered in this paper are defined under the assumption that the analyzed programs do not raise exceptions, otherwise the behavior of the static analyzer depends on the semantics of the programming language. For example in C, division by zero causes undefined behavior. When classifying the properties into soundness and precision (as described in Sect. 4), we assume that the abstract domain does not model error states.

3.2 Input Data
Abstract domains often use sophisticated data structures to optimize performance. For instance, elements of the Polyhedra domain are typically represented using both matrices and vectors of floating-point numbers. In our experiments, we observed that the standard test case generation techniques do not work well for complex abstract domains. In particular, fuzzing failed to detect subtle interactions between domain operations, and DSE did not effectively explore real-world implementations that make heavy use of floating-point arithmetic and libraries. We, therefore, use a combination of test case generation techniques to create a pool of domain elements, which serve as test inputs to domain operations. The pool is populated in two steps.

Step 1 creates an initial pool by combining boundary and random testing. Each element of a numerical domain can be constructed from numerical constraints. For instance, an element of the Intervals domain, which maps program variables \( v_i \) to their possible ranges, is created from the constraint \( k_l \leq v_i \leq k_u \) (the constants \( k_l \) and \( k_u \) are the lower and upper bounds). We generate the numerical constants randomly or choose them from a pre-defined set of boundary values that are more likely to expose bugs, such as off-by-one errors and arithmetic overflows. If the Intervals domain is implemented using machine integers, these boundary values

\[
\begin{align*}
\text{INT} & \Rightarrow 0, \text{INT} - \text{MAX} \\
\text{INT} - \text{MAX} & \Rightarrow 0, \text{INT} - \text{MIN} \\
\text{INT} - \text{MIN} & \Rightarrow 0, \text{INT} - \text{MAX} \\
\text{INT} - \text{MAX} & \Rightarrow 0, \text{INT} - \text{MIN}
\end{align*}
\]

The initial pool also contains the extreme elements \( T \) and \( \bot \).
Selecting inputs from this pool (together with a suitable expression \( e \)) as arguments to the assign transformer will likely detect the possible unsoundness illustrated in Fig. 1. The initial pool is likely to contain an element mapping the variable \( p \to [0, \text{INT} - \text{MAX}] \) since 0 and \( \text{INT} - \text{MAX} \) are pre-defined boundary values; moreover, we are likely to obtain an expression \( e \) of the from \( p + k \) for a positive constant \( k \) (see Sect. 4 for details). Evaluating the assign transformer on these inputs violates soundness property (a) defined above.

The initial pool allows us to test individual domain operations. However, this approach is insufficient in two situations. First, some implementations rely on complex consistency conditions on their data structures. For Polyhedra, for instance, the two internal representations must be kept consistent. If a faulty operation violates this invariant, the effect can be often observed only when applying a subsequent operation; it is thus necessary to perform at least two consecutive operations to detect the bug.

Second, there are certain soundness or precision properties of individual operations that cannot be checked by generic, domain-independent oracles. For instance, the assign transformer for Octagons should, in some cases, apply a so-called closure operation; failing to do so or using an imprecise closure may lead to loss of precision in subsequent operations, such as inclusion or equality tests [41]. A test oracle that directly detects a missing closure would be specific to Octagons and, thus, not reusable. A more generic way to detect this problem is by intersecting the result of the assign transformer with top. The expected output of the intersection is the same as the result of the assign transformer itself. However, if the transformer does not apply the closure when expected, the domain-independent property (b) may fail due to imprecision.

The above situations can both be tested by executing at least two consecutive domain operations before checking the oracle. Therefore, step 2 of our input generation expands the pool of domain elements by iteratively applying a domain operation to existing pool elements (and possibly other arguments) and adding the result to the pool. By repeating this process, we increase the likelihood of constructing elements that need to be built incrementally with several domain operations. As a result, we are also more likely to detect bugs that manifest themselves only in consecutive operations.

3.3 Test Drivers and Exploration
For each property under test, we manually write a test driver that is parametric in: (1) the operations and their arguments used in step 2 to populate the pool of domain elements, and (2) the arguments of the property under test.

Fig. 2 shows, in pseudo code, our test driver for checking soundness property (a) for the assign transformer. The driver takes as arguments a sequence of operations (ops), which are used in step 2 of the pool creation, a sequence of indices into the pool (elems), a sequence of expressions (exprs), and a sequence of program variables (vars). The latter three provide arguments to the domain operations. The driver also takes as arguments an index into the pool (oelem), an expression (oexpr), and a variable (ovar), which are inputs to the property under test (or oracle) (lines 17–21).

The test driver initializes the pool of abstract elements by creating valid inputs for the currently tested domain, such as Intervals or Octagons (step 1, line 4). Then, it extends the pool by applying domain operations (step 2, lines 7–14). Each iteration obtains the next operation and its arguments from the parameters of the test driver, applies it, and adds the result to the pool. Finally, the test driver asserts property (a) after computing the result of the assign operation. For simplicity, we omit optimizations (for instance, our implementation initializes the pool only once, before all the drivers start operations).
We have discussed an example in the previous section (property 35).

Since convergence is a termination property, which cannot be tested, ensure that the static analysis of loops and recursion terminates.

We use the state-of-the-art gray-box fuzzer LibFuzzer [6] to provide the arguments to the test driver, i.e., to choose the operations used in step 2 together with their arguments, and the parameters for the property under test. Gray-box fuzzers use a lightweight code instrumentation to generate inputs that are likely to execute previously uncovered code. In our implementation, all parameters of the test driver are encoded into one array, which is created by the fuzzer.

4 TECHNICAL SOLUTION

In this section, we provide the technical details of our approach.

4.1 Test Oracles

Our test oracles are based on domain-independent mathematical properties of abstract operations. In the following, we give an overview of these properties and explain how they are checked by the test drivers.

Properties. Based on the abstract-interpretation literature [20, 21] and earlier work on testing static analyzers [40], we identified 46 properties that need to be satisfied by domain operations to ensure soundness, precision, and termination. Fig. 3 provides an overview.

The soundness properties are required to ensure that an abstract-domain element over-approximates the concrete state it represents. We have discussed an example in the previous section (property 35).

To deal with the undecidability of most semantic program properties, static analyses over-approximate the set of concrete program behaviors and then infer or check properties on this abstraction. Since they are intended to compute an approximation, one cannot expect the operations of an abstract domain to be precise w.r.t. the concrete program execution. Therefore, our precision properties check that the domain operations do not lead to unnecessary information loss, that is, they compare the result of an operation to the

most precise representable result as obtained by applying the best transformer (see Sect. 3).

For instance, computing the join of two intervals [1, 1] and [3, 3] yields [1, 3], which loses the information that the variable is different from 2. Despite this inevitable information loss, a join operator should fulfill the soundness property (33 and 46) requiring widening and narrowing to eventually reach a fixed point, which is necessary to ensure that the static analysis of loops and recursion terminates.

The convergence properties (33 and 46) require widening and narrowing to eventually reach a fixed point, which is necessary to ensure that the static analysis of loops and recursion terminates. Since convergence is a termination property, which cannot be tested, we instead check the stronger property that a fixed point is reached within a given number of iterations, as we explain below.

Executable oracles. We manually construct a test driver for each of the properties in Fig. 3. This driver selects suitable inputs for
we use a slightly different test driver that applies widening on
As a consequence of this monotonicity precondition, property 31
(1) program variables, (2) expressions over them (for the
Testing the properties from Fig. 3 requires three kinds of input data:
variables is configurable; we use eight variables in our experiments.
Octagons widening reaches a fixed point within
represented in the domain under test. For instance, for an octagon
coefficients, which are chosen by the fuzzer to increase the likelihood
For the numerical domains considered in this paper, these expres-
non-trivial partitions $\{v_i \in \mathbb{Z} \mid i < k\}$; that is, by changing the last parameter of
for variable $\bar{x}$ and checks that the $\sqsubseteq$ operator yields true when applied to $d$ and $d$.
Translating properties into executable oracles is straightforward
for most soundness and precision properties, but slightly more in-
volved for convergence properties. Fig. 4 shows a fragment of our
test driver for checking whether the Octagons widening converges
after $k$ iterations [41]. The driver computes an increasing chain of
x elements (as checked by property 29), each obtained by widen-
ing the previous element with an arbitrary element $y$. Widening
converges if the $x$-chain becomes stable, here, within $k$ steps. We
observed that $k = 100$ is a sufficient upper bound for all our tested
analyzers, because widening converges much faster in practice.
Note that, for most abstract domains (such as Intervals, Octagons,
etc.), widening can be applied to arbitrary elements. There are,
however, some exceptions; for instance, the Polyhedra widening
requires monotonicity of its operands (that is, $x \sqsubseteq y$). In such cases,
we use a slightly different test driver that applies widening on $x$
and the join of $x$ and $y$; that is, by changing the last parameter of
apply on line 6 of Fig. 4 to $x \sqcup y$, because $x \sqsubseteq x \sqcup y$ (see property 8).
As a consequence of this monotonicity precondition, property 31
needs to hold for the Polyhedra domain only for $\bar{x} = \bot$.

4.2 Input Data
Testing the properties from Fig. 3 requires three kinds of input data:
(1) program variables, (2) expressions over them (for the assign
and cond transformers), and (3) abstract-domain elements that contain
constraints over these variables. We construct this data as follows.
Program variables. All our test cases operate on a set of pre-
defined integer variables. The number of variables must be suffi-
ciently small to keep the memory consumption and execution time
of the test cases low. On the other hand, abstract-domain optimiza-
tions, such as decomposition, require enough variables to obtain
non-trivial partitions [46]. In our implementation, the number of
variables is configurable; we use eight variables in our experiments.
Expressions. Testing assignments requires numerical expressions.
For the numerical domains considered in this paper, these expres-
sions are linear sums over the program variables with integer coef-
ficients, which are chosen by the fuzzer to increase the likelihood
of constructing suitable expressions. Note that, to test precision
properties, we also need to generate expressions that are not rep-
resentable in the domain under test. For instance, for an octagon
{\bar{v}_0 \geq 0 \land \bar{v}_1 \geq 0}, the assignment $\bar{v}_2 := \bar{v}_0 + 2\bar{v}_1$ is expected to
produce {\bar{v}_0 \geq 0 \land \bar{v}_1 \geq 0 \land \bar{v}_2 \geq 0} (and not $\bot$), even though
the assigned expression is not octagonal. For the cond transformer, we
obtain boolean expressions by comparing the linear sums to zero.

Domain elements. As explained in Sect. 3, we create a pool of
abstract-domain elements (such as intervals or octagons) in two
steps: Step 1 populates the pool by constructing elements using
a combination of boundary and random testing, whereas step 2
expands it by applying domain operations to existing pool elements.
Besides $\top$ and $\bot$, step 1 also creates simple domain elements
that contain only one constraint on the pre-defined program variables.
More complex domain elements are constructed in step 2. Tab. 1
shows the structure of simple domain elements for common nu-
merical domains. These elements can be constructed by choosing
values for the constants $k$ (e.g., the bounds of an interval) and
the coefficients $c$. By default, we pick them from a small set of pre-
defined values (e.g., boundary values such as 0 or $\infty$) and a small set
of arbitrary values. The use of pre-defined values is optional and
can be configured in the test drivers (see Sect. 5).
These sets of values depend on both the domain under test and
its implementation. For instance, octagonal coefficients must be in
$\{-1, 0, 1\}$, whereas polyhedral ones are arbitrary integers. Moreover,
different implementations represent numbers differently; e.g., we
use the pre-defined values \{$\text{INT_MIN}$, 0, $\text{INT_MAX}$\} for integer
intervals if the implementation uses machine integers, $(-\infty, 0, \infty)$
for arbitrary-precision integers, and additionally $\text{NaN}$ for floating-
point representations. Even though we focus on integer program
variables here, internal floating-point computations may lead to
rounding errors, as we observed in our experiments (see Sect. 5).
The values in both sets are not chosen by the fuzzer. However,
the fuzzer can still control the pool of domain elements by selecting
suitable operations in step 2. This step constructs more diverse
domain elements, usually with more complex constraints, by
applying domain operations to the existing elements. Step 2 makes
use of all domain operators that yield domain elements ($\sqcup$, $\sqcap$, $\forall$)
as well as the abstract transformers assign and project. For simplicity,
we omit narrowing, which is not supported by all domains, and
conditionals. Using all these domain operations not only allows
us to detect errors in their implementation (such as the missing
closure in assignments that we discussed in Sect. 3), but it also
efficiently generates a diverse set of valid domain elements. While
it is theoretically possible to create a new element by generating
an arbitrary set of constraints, such an approach would often pro-
duce unsatisfiable conjunctions of constraints, represented by
the already considered $\bot$ element.

Figure 4: Fragment of a test driver for checking whether the
Octagons widening reaches a fixed point within $k$ steps.

![Figure 4: Fragment of a test driver for checking whether the Octagons widening reaches a fixed point within $k$ steps.](image-url)

Table 1: Structure of abstract elements with one constraint
for commonly used numerical domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Element with one constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervals</td>
<td>${k_1 \leq v_1 \leq k_u}$</td>
</tr>
<tr>
<td>Zonotopes</td>
<td>${v_1 = \frac{x_1}{k_1} + \frac{x_2}{k_2} + \ldots + \frac{x_n}{k_n}}$</td>
</tr>
<tr>
<td>Octagons</td>
<td>${v_1, v_3, \ldots, \forall k} \sqsubseteq \bot$</td>
</tr>
</tbody>
</table>

$v$: variables, $c$: coefficients, $\varepsilon \in \{\geq, \leq\}$
5 EXPERIMENTAL EVALUATION

To evaluate the effectiveness of our approach, we apply it to two complex libraries for numerical analysis, namely APRON [31] and ELINA [46]. We were able to find errors in several of the implemented domains. Most of them are already confirmed and fixed.

APRON is a mature library, extensively used in many academic and industrial static analyzers, such as Astére [12] and PAGAI [30], as well as in the CPAchecker verification platform [11]. ELINA is a recent library that uses highly-optimized algorithms based on online decomposition to achieve significant speedups [46, 47]. These algorithms are difficult to implement correctly.

In our experiments, we consider three variants of APRON with different internal representations for numerical values (Sects. 5.1 and 5.2), and two versions of ELINA (Sect. 5.3). We also evaluate different configurations of our technique (Sect. 5.4), and compare it to pure fuzzing and DSE (Sect. 5.5).

Experimental setup. Since the tested domains have different complexity (i.e., the implementation of Polyhedra is significantly slower), we estimated the maximum execution time required to test each property approximately one million times for Intervals, Zonotopes, and Octagons, and half a million times for Polyhedra (see Tab. 2). The values are smaller for ELINA than for APRON because ELINA’s code is highly optimized. Intervals and Zonotopes were not considered for ELINA, as they were not part of the tested artifacts [3, 4]. All the experiments were performed on a 3.3 GHz Intel Xeon E5-4627 v2 CPU with 236 GB memory and RAID6 HDD.

5.1 APRON Double and Rll

APRON supports different internal representations for numerical values. For instance, the Double representation uses floating-point numbers, while Rll uses an approximation of rational numbers based on two 64-bit integers for the numerator and denominator. Compared to APRON MPQ, which uses arbitrary-precision rationals (see Sect. 5.2), these representations offer better performance, but may lose precision and cause termination [1]. Intervals and Octagons support Double, while Rll is available for Polyhedra.

Our experiments indeed uncovered soundness, precision, and termination problems in several domains of the latest version of APRON (0.9.10), as shown in Tab. 3. Here, the three versions of Rll refer to different test configurations that we discuss later; the domain implementation is always the same. The third column presents the total number of properties from Fig. 3 that we attempted to test for each domain. We tested only the first 41 properties for Intervals and Polyhedra since narrowing is not implemented for Intervals in APRON and not mathematically defined for Polyhedra. The reported violations in the fourth column are not necessarily all caused by different bugs. Nevertheless, observing multiple violations caused by the same bug can provide additional information for error localization. We used a configuration that initializes the pool with 32 elements (step 1 of the pool population), includes LONG_MIN and LONG_MAX as boundary values, and applies 16 operations to generate more complex domain elements (step 2). This corresponds to configuration C2 from Tab. 6, which we discuss in Sect. 5.4, together with measurements on the testing time.

Intervals and Octagons. All our generic properties hold for Intervals, the simplest domain we tested. Nonetheless, we indirectly found imprecisions for $\leq$ and $\leq$, by testing APRON’s efficient implementation of Zonotopes [27] (discussed in Sect. 5.2), which uses the Interval operations. These issues were confirmed and fixed.

For Octagons, three precision properties (17, 22, and 23) are violated because the equality test gives imprecise results due to rounding errors. In Fig. 3, we show an example input generated by our approach that violates property 17. oct represents a call to the Octagons constructor. The root cause of the imprecision is the underlying double representation. For octs, LONG_MAX cannot be precisely represented as a double value. The resulting rounding error gives approximate results in the subsequent computations and makes the assertion on line 5 fail.

Polyhedra. With the same test configuration (C2), all the tests fail to terminate for Polyhedra Rll (Rll v1 in Tab. 3). The problem is that APRON enters infinite loops during step 1 of the pool construction because of unhandled arithmetic overflows. The bug can be seen when constructing at least two consecutive domain elements, as in Fig. 6. The second constructor call enters an infinite loop.

This bug causes all test drivers to time out before they even reach the test oracle. To work around it and look for additional bugs, we replaced LONG_MIN and LONG_MAX by INT_MIN and INT_MAX as pre-defined values in the pool construction. In this case (Rll v2 in Tab. 3), step 1 of the pool construction succeeds, but for 21 test drivers, step 2 times out. The remaining 20 test drivers lead to violations of soundness and precision properties. The root cause of all these failures is unhandled arithmetic overflows in various operators. Dropping pre-defined values entirely (Rll v3 in

Table 3: Results for APRON Double and Rll. The third column reports how many of the properties from Fig. 3 were applicable for each tested domain. The first value in the fourth column shows the number of violated soundness properties, the second represents precision properties, and the third indicates how often errors in the domain implementation caused the test driver to crash or time out.

<table>
<thead>
<tr>
<th>Variant</th>
<th>Domain</th>
<th>#Properties Tested</th>
<th>#Properties Violated</th>
<th>Causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td>Intervals</td>
<td>41</td>
<td>0 / 0 / 0</td>
<td>–</td>
</tr>
<tr>
<td>Double</td>
<td>Octagons</td>
<td>46</td>
<td>0 / 3 / 0</td>
<td>rounding</td>
</tr>
<tr>
<td>Rll v1</td>
<td>Polyhedra</td>
<td>41</td>
<td>0 / 0 / 41</td>
<td>overflow</td>
</tr>
<tr>
<td>Rll v2</td>
<td>Polyhedra</td>
<td>41</td>
<td>5 / 15 / 21</td>
<td>overflow</td>
</tr>
<tr>
<td>Rll v3</td>
<td>Polyhedra</td>
<td>41</td>
<td>5 / 19 / 4</td>
<td>overflow</td>
</tr>
</tbody>
</table>

Figure 5: Input violating property 17 for Octagons.

Table 2: Maximum execution time per test driver.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Maximum execution time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRON</td>
<td>ELINA</td>
</tr>
<tr>
<td>Intervals</td>
<td>750</td>
</tr>
<tr>
<td>Zonotopes</td>
<td>2'400</td>
</tr>
<tr>
<td>Octagons</td>
<td>1'900</td>
</tr>
<tr>
<td>Polyhedra</td>
<td>17'700</td>
</tr>
</tbody>
</table>
of soundness or precision properties, or through crashes and assertion failures in APRON’s internal consistency checks. In total, we were able to find 14 out of the 17 seeded bugs. In the following paragraphs, we present two bugs that we detected and explain why three of the seeded errors could not be found.

A seeded bug in Intervals uses a slightly modified version of an unsound definition for the widening operator [38]. This definition uses \( \leq \) instead of \( > \) to compare two bounds, which leads to a violation of soundness property 30. Our tests reveal this problem.

A seeded bug in Octagons removes the call of closure in one special case of the assignment transformer. As explained in Sect. 3, we detect this bug by generating an assignment during step 2 of the pool construction, followed by a meet, which violates four of our precision properties because the equality check becomes imprecise.

While our approach found the vast majority of the seeded bugs, there were three it did not detect. (1) One seeded bug makes the Octagons closure less precise. Detecting this problem would require additional, octagon-specific precision properties for closure [41]. This can be easily done, but our focus here is on domain-independent properties. (2) Another seeded bug affects the precision of widening for Intervals in a way that does not violate the properties from Fig. 3. Detecting it would again require additional, domain-specific properties. (3) The last undetected bug changes the assignment operator for Polyhedra to act like project, making it trivially sound, but imprecise. This bug can be easily found if the input elements are polyhedra of dimension 1 (practically intervals). Our default configuration excludes such elements; adjusting the range of dimensions to include the value 1 leads to a violation of property 37, revealing the imprecision.

5.3 ELINA

To evaluate how effective our technique is on highly optimized implementations, we applied it to test ELINA. We used the same configuration as for APRON (C2) on the following abstract domains:

- \( EP1 \): Polyhedra with decomposition [46]
- \( EOD \): Octagons with decomposition [45]
- \( EO \): Octagons without decomposition [45]
- \( EP2 \): more recent version (including bug fixes) of \( EP1 \)
- \( EP3 \): decomposed Polyhedra with further optimizations [47]

\( EP1 \) is the implementation in the artifact [3] from POPL’17 [46], and the other variants are part of the artifact [4] from POPL’18 [47].

Note that all ELINA domains are based on floating-point numbers (not on the slower arbitrary-precision rationals). This design decision may compromise precision to achieve high performance.

Initially, the pool construction for \( EP1 \) failed for all the tests due to corner cases like \( LONG\_MIN \). This step was also not successful for \( EP2 \) and \( EP3 \) if the polyhedra had \( LONG\_MAX \) coefficients (a similar issue as for APRON Double, see Sect. 5.1). To
We now assess the impact of these parameters on its effectiveness. We report the abstract domain in which it was inserted, the violated properties, and the execution time until each configuration detects the corresponding test inputs in this experiment (only the first 41 properties are tested, as shown in Tab. 5). For both variants of Octagons, like for APRON Double (see Sect. 5.1), properties 17, 22, and 23 were violated. These imprecisions are caused by rounding errors when performing closure and, implicitly, in the equality tests. We reported the issues and they were confirmed. However, obtaining very precise results with finite-precision representations is challenging and the developer is still working on a fix.

### 5.4 Different Configurations

Our technique relies on three main configurable parameters: (1) the size of the initial pool in step 1 of the input generation (see Sect. 3), (2) the number of operations applied in step 2, and (3) whether pre-defined values are used to construct elements and expressions. We now assess the impact of these parameters on its effectiveness.

For this experiment, we used the three versions of APRON MPQ with the bugs seeded by the experts and the configurations shown in Tab. 6. The results are presented in Tab. 7. For each seeded bug, we report the abstract domain in which it was inserted, the violated properties, and the execution time until each configuration detects the violation. As shown in the table, configurations C2, C5, and C6 all find the maximum number of violations, and implicitly all the bugs, within the time limits defined in Tab. 2. However, C2 does so significantly faster than C5 and slightly faster than C6.

#### Initial pool size.

When the initial pool includes just $\top$ and $\bot$ (as in C1), no violations are detected for Intervals in comparison to C2, that is, 4 bugs are missed. On the other hand, a very large initial pool (C3) increases the execution time without detecting all violations. We attribute this to the fact that the size of the initial pool directly influences the number of possible arguments to the subsequent operations in step 2 and to the test oracle; exploring all of them takes more time without necessarily being more effective.

#### Number of operations.

Without using step 2 of the pool construction (C4), some bugs and property violations are missed (see Sect. 3 for an example). However, for our technique to be effective, the number of operations should not be too large since it increases the execution time. In Tab. 7, C5 is usually slower than C2 even though both configurations find the same number of violations.

#### Pre-defined values.

For APRON MPQ, which uses arbitrary precision rationals, both C2 and C6 find all the violations, C2 being slightly faster. Pre-defined values are particularly important for testing abstract-domain implementations based on fixed-precision numbers such as APRON Double; as our experiments show (see Sect. 5.1 and Sect. 5.3), these implementations may suffer from arithmetic overflows and rounding errors.

### 5.5 Fuzzing and Dynamic Symbolic Execution

In this section, we compare our technique to pure gray-box fuzzing and DSE. In particular, we use LibFuzzer [6] and KLEE [13], a state-of-the-art DSE engine, to generate inputs for the test oracles corresponding to the properties from Fig. 3. For a fair comparison, we write alternative test drivers that do not create and populate a pool of abstract elements. Instead, we allow the tools to directly generate the coefficients and constants of up to 50 constraints per element. The test oracles are the same as in our test drivers.

#### Gray-box fuzzing.

We ran LibFuzzer (with default options) on the three APRON MPQ versions with seeded bugs, using the same time limits as in Tab. 2, and compared the results to our C2 configuration. LibFuzzer detected 45 out of the 58 property violations that our approach found. It revealed these violations generally faster than our approach for Intervals and Polyhedra, but slower for Octagons. A manual inspection of the generated counterexamples shows that our technique produces significantly simpler and more readable test inputs, which was very useful in debugging the detected issues.

Tab. 8 provides additional details for the bugs seeded by Expert 3. The last two columns show the time it takes to detect a property violation for our technique (with C2) and for LibFuzzer, respectively. LibFuzzer missed one of the Octagon bugs. Moreover, LibFuzzer’s results for Polyhedra suggest that the implementation of join is...
We grouped the violations by the seeded bugs they reveal (dashed lines) and by the expert who seeded the bugs (solid lines). We excluded two seeded bugs in Polyhedra that caused assertions failures in APRON for all configurations, and thereby masked the other bugs seeded by Expert 2.

### Threats to Validity

5.6 Threats to Validity.

We identified two threats to the validity of our experiments. Our comparisons to alternative approaches heavily rely on floating-point arithmetic, which KLEE does not handle very well.

## Dynamic symbolic execution.

Since KLEE does not try to explore all execution paths in external libraries and APRON makes heavy use of libraries, we performed the comparison on the EOD and EO ELINA domains, using KLEE’s default options. Our alternative test drivers use ELINA’s functions directly, not its APRON interface, to avoid external library calls. With the latest version of KLEE (1.4.0), all but 1 test throw an error for both tested domains, because KLEE is not able to model malloc instructions with symbolic sizes [5].

To overcome this limitation, we extended KLEE with an option for specifying the upper bound for the symbolic size; we used 8192 in our experiments. For each tested domain, our technique detected 3 violated properties (see Tab. 5), but KLEE was not able to detect any, even with a time limit of 17'700s (the 25-fold of the time limit used for our technique). We believe this is due to the fact that ELINA heavily relies on floating-point arithmetic, which KLEE does not handle very well.

### Test generation tool.

Our comparisons to alternative approaches focus on one fuzzer and one DSE tool. Since we chose mature, state-of-the-art tools, we believe that our results are representative. Similarly, we did not use alternative gray-box fuzzers when evaluating our own approach. Since most fuzzers make no assumptions about the code under test, we do not expect to see significantly different results for other fuzzers.
Random initialization. Our pool initialization step chooses some of the coefficients and constants randomly. To ensure that our results are deterministic, all test drivers use the same, pre-defined random seed.

6 RELATED WORK

Our approach is the first to systematically test a wide range of soundness, precision, and convergence properties on complex abstract domains. It combines several existing test case generation techniques in a novel way. On the one hand, we derive executable test oracles [9] to check high-level, mathematical properties of abstract-domain implementations. On the other hand, we incorporate ideas from boundary and random testing (step 1 of the pool population) and from feedback-directed random testing [42] (step 2) to obtain inputs for the test oracles. A key difference with the latter is that, in our case, the fuzzer controls which elements are added to the pool, by providing the operations and their arguments.

Testing static analyzers with random programs. One way to test static analyzers is by randomly generating input programs [24]. This approach is particularly effective in testing the robustness of analyzers, that is, for detecting which input programs make the analyzers crash. To also test soundness properties, Cuq et al. instrument the code of the analyzer under test with assertions about inferred values or relations between program variables. These assertions are then checked against concrete executions. In contrast, our technique generates input data systematically, and does not require any modifications to the implementation of the tested analyzers.

Analysis testing and delta debugging. Similarly to Cuq et al.’s work, Andreasen et al. [8] compare concrete executions to abstract domain elements to detect soundness and precision problems. They use delta debugging to reduce the size of the input programs in order to report the errors concisely. In contrast, we propose a technique for automatically generating the input domain elements; our approach starts with simple elements and applies a small number of operations, generating small counterexamples by construction.

Systematically testing lattice properties. Midtgaard and Møller [40] focus on quickchecking [18] basic lattice properties of abstract interpreters. Our technique was inspired by their work; it extends the set of tested properties and, as our evaluation shows, is effective on widely-used and highly-optimized abstract domain implementations. A comprehensive experimental comparison with their tool was not possible, as the authors provide constructors and helper functions for generating ordered pairs of elements only for Intervals, but not for the complex numerical domains that we consider. Without them, many of the properties cannot be tested, as the randomly generated inputs very likely do not satisfy the preconditions. For this reason, we expect that their technique cannot significantly outperform gray-box fuzzing, which we showed to be less effective than our approach (see Tab. 8). We solve the problem of generating ordered inputs by applying domain operations to the existing pool elements (e.g., the result of a join over-approximates its operands).

Formally verified static analyzers. Interactive theorem provers such as Coq [2] have been used to verify the soundness of the design of static analyses (e.g., in the context of type systems [25, 44]). However, the proofs do not typically provide any guarantees about the actual implementation of the analyses, and, thus, could still benefit from automated testing techniques like ours. The Verasco [32] project extracts executable code from verified Coq formalizations of several abstract domains. This approach produces implementations that are correct by construction, but is not yet practical for complex, highly-optimized implementations. Recent work by Madsen and Lhoták [39] uses symbolic evaluation (based on symbolic executions and SMT solvers) to verify the correctness, i.e., safety and soundness, of abstract-domain implementations. If the problem is undecidable, it relies on a quickchecking approach inspired by [40]. Their evaluation does not consider complex numerical domains.

Unsoundness in static analyzers. Even for analyzers that are unsound by design [10, 15–17, 37], our technique is useful to detect unintentional sources of unsoundness and imprecision (e.g., caused by implementation errors).

Testing compilers and program analyzers. Abstract domains are one of many components that are used in modern compilers and program analyzers. Besides efforts in proving properties about such components (e.g., CompCert [36] and Verasco [32]), there is a significant body of work on using testing techniques to detect issues in compilers [34, 35, 48, 51], and recently, in DSE engines [33].

7 CONCLUSION

We have presented an automated testing technique for detecting soundness, precision, and convergence errors in abstract-domain implementations, which are crucial components of many static analyzers. We have evaluated our approach on several complex, real-world abstract domains from two widely-used libraries for numerical analysis and demonstrated its effectiveness in finding both seeded and previously unknown errors.

Even though our evaluation focuses on numerical abstract domains, we believe that the high-level ideas of our technique also apply to other domains such as string or heap domains. Such domains require different techniques to construct domain elements as well as suitable parameters for assignments and conditionals. Moreover, one needs to assess whether fuzzers can effectively explore the search space of non-numerical domain implementations. We leave these generalizations as future work.

There are several lessons to be learned from this work. First, an automated testing technique offers a pragmatic and effective solution for uncovering issues in static analyzers that would be difficult to find using manual testing. Second, off-the-shelf testing tools are less effective for complex, highly-optimized domain implementations than a well-designed combination of techniques. We believe that these observations carry over to other application areas, such as machine-learning frameworks; exploring those is future work.

ACKNOWLEDGMENTS

We would like to thank the developers of APRON, Khalil Ghorbal and Antoine Miné, and of ELINA, Gagandeep Singh, for their help and support. We are also grateful to Jérôme Dohrau, Gagandeep Singh, and Caterina Urban for seeding bugs for our evaluation, and to our anonymous reviewers for their helpful comments. Maria Christakis’s work was supported in part by a Facebook Faculty Research Award.