Trapezoidal Generalization over Linear Constraints

David Greve
david.greve@rockwellcollins.com

Andrew Gacek
andrew.gacek@rockwellcollins.com

We are developing a model-based fuzzing framework that employs mathematical models of system behavior to guide the fuzzing process. Whereas traditional fuzzing frameworks generate tests randomly, a model-based framework can deduce tests from a behavioral model using a constraint solver. Because the state space being explored by the fuzzer is often large, the rapid generation of test vectors is crucial. The need to generate tests quickly, however, is antithetical to the use of a constraint solver. Our solution to this problem is to use the constraint solver to generate an initial solution, to generalize that solution relative to the system model, and then to perform rapid, repeated, randomized sampling of the generalized solution space to generate fuzzing tests. Crucial to the success of this endeavor is a generalization procedure with reasonable size and performance costs that produces generalized solution spaces that can be sampled efficiently. This paper describes a generalization technique for logical formulae expressed in terms of Boolean combinations of linear constraints that meets the unique performance requirements of model-based fuzzing. The technique represents generalizations using *trapezoidal* solution sets consisting of ordered, hierarchical conjunctions of linear constraints that are more expressive than simple intervals but are more efficient to manipulate and sample than generic polytopes. Supporting materials contain an ACL2 proof that verifies the correctness of a low-level implementation of the generalization algorithm against a specification of generalization correctness. Finally a post-processing procedure is described that results in a restricted trapezoidal solution that can be sampled (solved) rapidly and efficiently without backtracking, even for integer domains. While informal correctness arguments are provided, a formal proof of the correctness of the restriction algorithm remains as future work.

1 Motivation

Fuzzing is a form of robustness testing in which random, invalid or unusual inputs are applied while monitoring the overall health of the system. Model-based fuzzing is a fuzzing technique that employs a mathematical model of system behavior to guide the fuzzing process and explore behaviors that would be difficult to reach by chance. Whereas many fuzzing frameworks generate tests randomly, a model-based framework can deduce tests from a behavioral model using a constraint solver. Because the state space being explored by the fuzzer is generally large, the rapid generation of test vectors is crucial. Unfortunately, the need to generate tests quickly is antithetical to the use of a constraint solver. Our solution to this problem is to use a constraint solver to generate an initial solution and then to generalize that solution relative to the constraint. Test generation in our model-based fuzzing framework, therefore, consists of repeated, randomized sampling of the generalized solution space. Generalization is crucial to the performance of our framework because it allows us to decouple constraint solving (slow) from test generation (fast).

The fuzzing algorithm, outlined in Algorithm 1, begins by identifying an appropriate logical constraint, selected according to some testing heuristic. A constraint solver then attempts to generate a

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under DARPA/AFRL Contract FA8750-16-C-0218. The views, opinions and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Approved for Public Release, Distribution Unlimited

Submitted to:
ACL2 2018

Algorithm 1: Conceptual Model-Based Fuzzing Algorithm

```
while True do
    L ⟵ nextConstraint()
    v⃗, SAT ⟵ solve(L)
    if SAT then
        T ⟵ generalize(L, v⃗)
        T′, σ ⟵ restrict(T, v⃗)
        for A While do
            w⃗ ⟵ sample(T′, σ)
            fuzzTarget(w⃗)
        end
    end
end
```

variable assignment that satisfies the constraint. If it succeeds, the solution is generalized relative to the constraint. The generalization is then restricted so that it can be sampled efficiently. The restricted generalization is then repeatedly sampled to generate random vectors (known to satisfy the constraint) that are then applied to the fuzzing target. This process is repeated for the duration of the fuzzing session.

Trapezoidal generalization is uniquely suited for use in model-based fuzzing. The trapezoidal generalization process is reasonably efficient, both in terms of speed and the size of the final representation. Unlike interval generalization, trapezoidal generalization is capable of representing many linear model features exactly, allowing sampled tests to better target relevant model behaviors. Finally, unlike generic polytope generalization, trapezoidal generalization supports efficient sampling of the solution space, enabling rapid test generation and addressing the performance requirements of model-based fuzzing.

This paper provides details on our trapezoidal generalization, restriction, and sampling algorithms and the supporting material includes an ACL2 proof of the correctness of a low-level implementation of generalization. Section 2 describes the trapezoidal data structure and show how it can be used to generalize solutions relative to logical formulae expressed as Boolean combinations of linear constraints. It includes a formal specification of generalization correctness and an outline of crucial aspects of the correctness proof for our generalization algorithm. Section 3 describes sampling and provides details of the restriction process that refines a trapezoid so that it can be rapidly and efficiently sampled (solved) without backtracking, even for integer domains. While informal correctness arguments are provided, a formal proof of the correctness of the restriction algorithm remains as future work. Section 4 provides background on our implementation and formalization of trapezoidal generalization and Section 5 compares our experience using both trapezoidal and interval generalization and provides an overview of related work in the field.

1.1 Problem Syntax

Let $x_1, \ldots, x_n$ be variables of mixed integer and rational types. Each variable has an associated numeric *dimension*, denoted by the variable subscript, that establishes a complete ordering among the variables. Numeric expressions are represented as polynomials (P) over variables having the form $c_n x_n + \cdots c_1 x_1 + c_0$ where each c_i is a rational constant. We specifically restrict our attention to linear, rational, multivariate polynomials.

$$P \equiv c_n x_n + \cdots c_1 x_1 + c_0$$

The largest k for which c_k is non-zero is called the dimension of a polynomial. We sometimes write P_k to explicitly identify the dimension of a polynomial, $\dim(P_k) = k$. Note that constants have dimension 0.

A vector is a sequence of values ordered by dimension. Given a vector $\vec{v} = v_n, \ldots, v_1$, we write $E[\vec{v}]$ to denote the evaluation of an expression E where each x_i in that expression is replaced by v_i. We assume that every variable has an associated value in the vector and consider only *consistent vectors*, vectors where each dimension's value is consistent with it's corresponding variable's type, either rational or integral. Constants are self-evaluating and the evaluation operation distributes over the standard operations in the expected ways.

The atoms in our formulae are linear constraints. A linear constraint (L) is an equality or inequality over polynomials. We will often write $\prec$ to stand for either $<$ or $\leq$ and $\succ$ to stand for either $>$ or $\geq$.

$$L \equiv P_i = P_j \mid P_i > P_j \mid P_i \geq P_j \mid P_i < P_j \mid P_i \leq P_j$$

Logical formulae (F) are defined over linear constraints using conjunction, disjunction, and negation.

$$F \equiv F \wedge F \mid F \vee F \mid \neg F \mid L$$

We define a variable *bound* (B) as a linear constraint on a single variable. The dimension of a variable bound is the dimension of the bound variable. We say that a variable bound is *normalized* if the dimension of the variable is greater than the dimension of the bounding polynomial, $n > m$.

$$B_n \equiv (x_n \prec P_m) \mid (x_n \succ P_m) \mid (x_n = P_m)$$

We define a *trapezoid* (T) as a (possibly empty) set of variable bounds. A trapezoid can be interpreted logically as the conjunction of all of its variable bounds or geometrically as the intersection of the volumes defined by the planes bounding each variable. We say that a vector *satisfies* a trapezoid if it is a consistent vector and the logical interpretation of the trapezoid evaluates to True at the vector. A trapezoid is *satisfiable* if a consistent vector exists for which it evaluates to True. Equivalently, we can say that a trapezoid *contains* a vector if the point defined by that vector resides inside of the volumes defined by its geometric interpretation. A satisfiable trapezoid contains at least one point. Below is a grammar for trapezoids expressed in terms of the intersection of variable bounds.

$$T \equiv B \cap T \mid B \mid \varnothing$$

In addition the above syntactic restrictions, we also require that the constituent bounds of a trapezoid satisfy the following three properties:

1. All variable bounds must be normalized
2. The set of bounds must be simultaneously satisfiable as witnessed by a consistent reference vector $\vec{v}$
3. Each x_n is either unbound or is bound by either a single equality or at most one upper bound and at most one lower bound.

We call this data structure a trapezoid because, for every dimension greater than one, the bounds on the variable of that dimension are, in general, linear relations expressed in terms of the preceding variable. Holding all other variables constant, such linear bounds describe trapezoidal regions; hence the name.

The data structure we use to represent generalizations is a *region*. A region is defined as either a trapezoid or the complement of a trapezoid:

$$R \equiv \{T\} \mid {}^{\sim}\{T\}$$

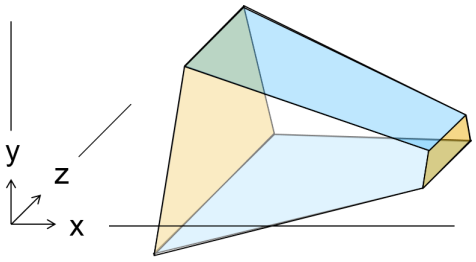


Figure 1: 3-d Trapezoidal Volume

2 Generalization

An ACL2 proof that a low-level implementation of the following generalization procedure satisfies our notion of generalization correctness is provided in the supporting materials accompanying this paper. The following presentation provides an overview of the generalization process but limits discussion of correctness to observations concerning key steps in this proof along with references to the supporting materials.

2.1 Generalization Correctness

Given a formula F and a consistent reference vector $\vec{v} = v_n, \ldots, v_1$, the objective is to compute a region R which correctly generalizes $\vec{v}$ with respect to F. We say that a generalization is correct if it satisfies the following properties:

- **Invariant 1** $F[\vec{v}] = R[\vec{v}]$
- **Invariant 2.a** If $F[\vec{v}]$, then $\forall \vec{w}.\ R[\vec{w}] \Rightarrow F[\vec{w}]$
- **Invariant 2.b** If not $F[\vec{v}]$, then $\forall \vec{w}.\ F[\vec{w}] \Rightarrow R[\vec{w}]$

The first invariant ensures that the reference vector is contained in the generalization iff the vector satisfies the original formula. The second two invariants ensure that the generalized result is a *conservative under-approximation* of the original formula. Together they guarantee that the state space that contains the reference vector in the generalized result is a subset of the state space that contains the reference vector in the original formula.

2.2 Generalization Process

The generalization process involves transforming a reference vector and a logical formula into a trapezoidal region. Generalizing relative to linear formula (Section 2.6) begins by transforming linear relations into regions (Section 2.3) and it proceeds by alternately intersecting and complementing the resulting regions (Section 2.5). The process of intersecting regions may further involve generalizing the intersection of sets of linear bounds into trapezoids (Section 2.4). The rules for each of these operations is presented in a bottom-up fashion.

2.3 Generalizing Linear Relations

We use $\langle\langle L\rangle\rangle \xrightarrow{\vec{v}} R$ to represent the generalization of a linear relation (L) into a trapezoidal region (R), a process which may involve several steps. First, arbitrary linear relations between two polynomials can be expressed as linear relations between a polynomial and zero.

$$\langle\langle P_i \prec P_j\rangle\rangle \xrightarrow{\vec{v}} \langle\langle P_i - P_j \prec 0\rangle\rangle \quad \langle\langle P_i \succ P_j\rangle\rangle \xrightarrow{\vec{v}} \langle\langle P_j - P_i \prec 0\rangle\rangle \quad \langle\langle P_i = P_j\rangle\rangle \xrightarrow{\vec{v}} \langle\langle P_i - P_j = 0\rangle\rangle$$

Constant linear relations normalize into either the true (empty) trapezoid or its complement.

$$\langle\langle c_0 \prec 0\rangle\rangle \xrightarrow{\vec{v}} \{\} \quad \text{if } c_0 \prec 0 \qquad \langle\langle c_0 = 0\rangle\rangle \xrightarrow{\vec{v}} \{\} \quad \text{if } c_0 = 0$$
$$\langle\langle c_0 \prec 0\rangle\rangle \xrightarrow{\vec{v}} {}^{\sim}\{\} \quad \text{if } {}^{\neg}(c_0 \prec 0) \qquad \langle\langle c_0 = 0\rangle\rangle \xrightarrow{\vec{v}} {}^{\sim}\{\} \quad \text{if } {}^{\neg}(c_0 = 0)$$

Non-constant linear relations between polynomials and zero are normalized so that they relate the variable with the largest dimension to a polynomial consisting only of smaller variables and constants.

$$\begin{aligned}
\langle\langle P_n < 0\rangle\rangle &\xrightarrow{\vec{v}} \langle\langle x_n < -(P_n/c_n - x_n)\rangle\rangle && \text{if } c_n > 0\\
\langle\langle P_n < 0\rangle\rangle &\xrightarrow{\vec{v}} \langle\langle -(P_n/c_n - x_n) < x_n\rangle\rangle && \text{if } c_n < 0\\
\langle\langle P_n \leq 0\rangle\rangle &\xrightarrow{\vec{v}} \langle\langle x_n \leq -(P_n/c_n - x_n)\rangle\rangle && \text{if } c_n > 0\\
\langle\langle P_n \leq 0\rangle\rangle &\xrightarrow{\vec{v}} \langle\langle -(P_n/c_n - x_n) \leq x_n\rangle\rangle && \text{if } c_n < 0\\
\langle\langle P_n = 0\rangle\rangle &\xrightarrow{\vec{v}} \langle\langle x_n = -(P_n/c_n - x_n)\rangle\rangle
\end{aligned}$$

Normalized linear relations that are true at the reference vector simply become singleton trapezoidal regions.

$$\begin{aligned}
\langle\langle x_n \prec P\rangle\rangle &\xrightarrow{\vec{v}} \{(x_n \prec P)\} && \text{if } x_n[\vec{v}] \prec P[\vec{v}]\\
\langle\langle P \prec x_n\rangle\rangle &\xrightarrow{\vec{v}} \{(P \prec x_n)\} && \text{if } P[\vec{v}] \prec x_n[\vec{v}]\\
\langle\langle x_n = P\rangle\rangle &\xrightarrow{\vec{v}} \{(x_n = P)\} && \text{if } x_n[\vec{v}] = P[\vec{v}]
\end{aligned}$$

A normalized linear relation that evaluates to false at the reference vector, however, is expressed as a negated trapezoidal region containing a single linear relation that is true at the reference vector.

$$\begin{aligned}
\langle\langle x_n < P\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(P \leq x_n)\} && \text{if } {}^{\neg}(x_n[\vec{v}] < P[\vec{v}])\\
\langle\langle x_n \leq P\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(P < x_n)\} && \text{if } {}^{\neg}(x_n[\vec{v}] \leq P[\vec{v}])\\
\langle\langle P < x_n\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(x_n \leq P)\} && \text{if } {}^{\neg}(x_n[\vec{v}] > P[\vec{v}])\\
\langle\langle P \leq x_n\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(x_n < P)\} && \text{if } {}^{\neg}(x_n[\vec{v}] \geq P[\vec{v}])\\
\langle\langle x_n = P\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(P < x_n)\} && \text{if } (x_n[\vec{v}] > P[\vec{v}]) \qquad \text{(neq.1)}\\
\langle\langle x_n = P\rangle\rangle &\xrightarrow{\vec{v}} {}^{\sim}\{(x_n < P)\} && \text{if } (x_n[\vec{v}] < P[\vec{v}]) \qquad \text{(neq.2)}
\end{aligned}$$

Normalizing linear relations in this way ensures that every variable bound contained in the body of a trapezoid is true at the reference vector. It also ensures that regions that contain the reference vector simplify into simple trapezoids while those that do not are expressed as the complement of a trapezoid that does.

With the exception of rules neq.1 and neq.2, the rules for generalizing linear relations ensure that the generalization is equal to the original formula (i.e. $\forall w.\ F[\vec{w}] = L[\vec{w}]$), trivially satisfying our correctness invariants. Crucially, rules neq.1 and neq.2 do satisfy Invariants 1 and 2.b (and, trivially, 2.a) as expressed in Section 2.1, ensuring that, in all cases, our generalization of linear constraints is correct. See the functions `normalize-equal-0` and `normalize-gt-0` and their associated lemmas in the file `top.lisp` of the supporting materials.

2.4 Generalizing Bound Intersections

Let C be a set of normalized bounds known to be satisfiable as witnessed by the consistent reference vector $\vec{v}$. The relation $C \xrightarrow{\vec{v}}{}^{*} T$ represents the fixed-point of intersecting and reducing the various bounds contained in C into a trapezoidal region. The following collection of rules govern the individual intersections of the members of such a set of bounds to produce a trapezoid. The rules of the rewrite system operate on an un-ordered set of bounds, allowing for arbitrary reordering of the bounds. The key to this process are rules for eliminating multiple bounds on the same variable. The rules for eliminating multiple upper bounds are shown below. The rules for multiple lower bounds are symmetric.

$$(x_n < P) \cap (x_n < Q) \xrightarrow{\vec{v}} (x_n < P) \cap \langle\langle P \leq Q \rangle\rangle \qquad \text{if } P[\vec{v}] \leq Q[\vec{v}] \qquad \text{(lt-int.1)}$$

$$(x_n < P) \cap (x_n \leq Q) \xrightarrow{\vec{v}} (x_n < P) \cap \langle\langle P \leq Q \rangle\rangle \qquad \text{if } P[\vec{v}] \leq Q[\vec{v}] \qquad \text{(lt-int.2)}$$

$$(x_n \leq P) \cap (x_n < Q) \xrightarrow{\vec{v}} (x_n \leq P) \cap \langle\langle P < Q \rangle\rangle \qquad \text{if } P[\vec{v}] < Q[\vec{v}] \qquad \text{(lt-int.3)}$$

$$(x_n \leq P) \cap (x_n \leq Q) \xrightarrow{\vec{v}} (x_n \leq P) \cap \langle\langle P \leq Q \rangle\rangle \qquad \text{if } P[\vec{v}] \leq Q[\vec{v}] \qquad \text{(lt-int.4)}$$

Note that while $\langle\langle P \prec Q \rangle\rangle$ normalizes the linear relation as described in Section 2.3 this bound will never evaluate to false due to the rules' side-conditions. Consequently the result will always be a simple trapezoid which is either empty (if both bounds are constant) or contains a single normalized bound on a variable whose dimension is less than that of x_n. Similar observations apply to the rules for simplifying bounds involving equality, as shown below. This process and it's correctness is captured by the function `andTrue-variableBound-variableBound` and its associated lemmas (generated by the `def::trueAnd` macro) in the file `poly-proofs.lisp` of the supporting materials.

$$(x_n = P) \cap (x_n \prec Q) \xrightarrow{\vec{v}} (x_n = P) \cap \langle\langle P \prec Q \rangle\rangle \qquad \text{(eq-int.1)}$$

$$(x_n = P) \cap (Q \prec x_n) \xrightarrow{\vec{v}} (x_n = P) \cap \langle\langle Q \prec P \rangle\rangle \qquad \text{(eq-int.2)}$$

$$(x_n = P) \cap (x_n = Q) \xrightarrow{\vec{v}} (x_n = P) \cap \langle\langle P = Q \rangle\rangle \qquad \text{(eq-int.3)}$$

Lemma 2.1. *The intersection rules are terminating.*

Proof. Define a measure on the conjunction of bounds to be the sum of the dimension of each bound. Each rule of the intersection rewrite system reduces this measure. For example, consider rule lt-int.1. We

know $dim(x) > dim(P)$ and $dim(x) > dim(Q)$, and thus $dim(x) > dim(\langle P \leq Q \rangle)$ and in particular $dim(x < Q) > dim(\langle P \leq Q \rangle)$. Since the measure is always non-negative, the intersection rules are terminating. See the `def::total` event for the function `intersect` in the file `intersection.lisp` of the supporting materials for a proof of termination for ordered (not un-ordered, see Section 3.0.1) bounds. □

Lemma 2.2. *When run to completion, the intersection rules result in a trapezoid.*

Proof. The intersection rules apply to a set of bounds that are normalized and satisfiable. To qualify as a trapezoid the only missing requirement is that there may be multiple upper, lower, and equality bounds on each variables. For multiple upper bounds, note that the 4 rules above cover all cases. Although rule lt-int.3 does not cover the case when $P[\vec{v}] = Q[\vec{v}]$, that case will in fact be covered by rule lt-int.2 since then $Q[\vec{v}] \leq P[\vec{v}]$. The case of multiple lower bounds is symmetric. Therefore, the rewrite rules will eliminate all multiple and lower and upper bounds so that the result is a trapezoid. See lemma `trapezoid-p-intersect` in the file `intersection.lisp` of the supporting materials. □

While, in general, the the fixed-point resulting from intersecting and reducing a set of bounds will differ from the original set, each rule that we apply to perform this normalization satisfies Invariants 1 and 2.a (and trivially 2.b) as expressed in Section 2.1, ensuring the correctness of this process.

2.5 Generalizing Region Intersections and Complements

The relation $R^a \cap R^b \overset{\vec{v}}{\rightarrowtail} R^c$ indicates that region R^c is a generalized intersection of regions R^a and R^b relative to $\vec{v}$ and ${}^{\sim}R^a \overset{\vec{v}}{\rightarrowtail} R^b$ to indicate that R^b is the generalized complement of R^a. We characterize the behavior of region intersection and complement with the following set of rules:

$$\frac{}{{}^{\sim}{}^{\sim}\{T\} \overset{\vec{v}}{\rightarrowtail} \{T\}}\ \text{COMP} \qquad \frac{T^a \cap T^b \overset{\vec{v}}{\longrightarrow}{}^* T^c}{\{T^a\} \cap \{T^b\} \overset{\vec{v}}{\rightarrowtail} \{T^c\}}\ \text{TINT}$$

$$\frac{}{{}^{\sim}\{T\} \cap R \overset{\vec{v}}{\rightarrowtail} {}^{\sim}\{T\}}\ \text{CINT.1} \qquad \frac{}{R \cap {}^{\sim}\{T\} \overset{\vec{v}}{\rightarrowtail} {}^{\sim}\{T\}}\ \text{CINT.2}$$

While perhaps surprising, the rules CINT.1 and CINT.2 not only simplify the region intersection process, they are also both consistent with and essential for preserving generalization correctness (specifically, Invariant 2.b) as described in Section 2.1. See the functions `and-regions` and `not-region` and their associated lemmas in the file `top.lisp` from the supporting materials.

2.6 Generalizing Linear Formulae

We write $F \overset{\vec{v}}{\Rightarrow} R$ to mean that region R generalizes $\vec{v}$ with respect to formula F. In general, conjunction generalizes into region intersection, negation generalizes into region complement, and disjunction generalizes into a complement of the intersection of the complements of the generalized arguments. The generalization relation is defined inductively over the structure of the formula F.

$$\frac{\langle\langle L \rangle\rangle \overset{\vec{v}}{\longrightarrow} R}{L \overset{\vec{v}}{\Rightarrow} R}\ \text{GEN-ATOM} \qquad \frac{F_1 \overset{\vec{v}}{\Rightarrow} R_1 \quad F_2 \overset{\vec{v}}{\Rightarrow} R_2}{F_1 \wedge F_2 \overset{\vec{v}}{\Rightarrow} R_1 \cap R_2}\ \text{GEN-AND}$$

$$\frac{F \overset{\vec{v}}{\Rightarrow} R}{\neg F \overset{\vec{v}}{\Rightarrow} {\sim}R}\ \text{GEN-NOT} \qquad \frac{F_1 \overset{\vec{v}}{\Rightarrow} R_1 \quad F_2 \overset{\vec{v}}{\Rightarrow} R_2}{F_1 \vee F_2 \overset{\vec{v}}{\Rightarrow} {\sim}({\sim}R_1 \cap {\sim}R_2)}\ \text{GEN-OR}$$

This procedure can be shown to satisfy our correctness invariants by induction over the structure of a linear formula and by appealing to the correctness of the generalization of linear relations and region intersection. See the definition `generalize-ineq` and the theorems `inv1-generalize-ineq` and `inv2-generalize-ineq` in the file `top.lisp` of the supporting materials.

3 Sampling

Sampling is the process of identifying randomized, concrete, type-consistent variable assignments (vectors) that satisfy all of the variable bounds in a trapezoid [1]. We call this process sampling, rather than solving, to emphasize the fact that it is intended to be randomized, simple, and efficient. Indeed, the structure of a trapezoid lends itself to a simple and efficient sampling algorithm. The trapezoidal structure ensures that the variable with the smallest dimension is bounded only by constants[2]. Sampling of the smallest dimension variable therefore involves simply choosing a value consistent with its type and constant bounds. Because each normalized variable bound is expressed strictly in terms of variables of smaller dimension, sampling the variables by proceeding from the smallest to the largest dimension guarantees that the bounds on each subsequent variable will evaluate to constants relative to variables already assigned. Sampling each variable from the smallest to the largest dimension, therefore, involves nothing more than choosing a value for each variable consistent with type and constant bounds computed based on the previous variable assignments.

While the above algorithm is certainly simple and efficient, the trapezoidal data structure as defined is not sufficient to guarantee that every variable assignment that satisfies the first N bounds will satisfy the $N+1^{th}$ bound. A given set of variable assignments may result in an upper bound that is smaller than a lower bound or, more subtly, an integer variable may be constrained by bounds that do not permit an integral solution. While backtracking and attempting different sets of variable assignments is possible in such cases, it can also be very inefficient.

To address this concern we present an overview of a post-processing step for trapezoids, performed before sampling, that results in a restricted trapezoid with properties that ensure that the simple and efficient sampling algorithm presented above will never encounter a bound that cannot be satisfied (even for integer variables) and thus never needs to backtrack. While the following procedure has not been mechanically verified, we sketch selected aspects of the informal proof with the long term objective of formalizing and verifying its correctness in ACL2.

3.0.1 Ordered Trapezoids

To better explain our post-processing approach we first introduce a more refined representation for trapezoids, beginning with the concept of variable *intervals*. A variable interval is either a single variable bound or a pair of bounds consisting of an upper and lower bound on the same variable. The dimension of an interval is simply the dimension of the bound variable. Variable intervals allow us to gather all of the linear bounds relevant to each variable in one place.

[1] We focus on trapezoids because solving a trapezoid complement is trivial. The complement of a trapezoid is merely a disjunction of negated linear bounds. An assignment that satisfies any one of the negated bounds satisfies the trapezoid complement.

[2] Technically, variables may also be unbound. In such case arbitrary value assignments suffice.

$$I_n \equiv B_n \mid (P^L \prec x_n) \cap (x_n \prec P^U)$$

Variable intervals can be used to construct ordered trapezoids. An ordered trapezoid is simply a sequence of variable intervals organized in descending order based on dimension, terminating in the empty trapezoid. Note that these new syntactic constructs do not change in any way the nature of trapezoids other than to impose additional ordering constraints on their representation.

$$\begin{aligned} T_n &\equiv I_n \cap T_m \text{ where } n > m \\ T_0 &\equiv \varnothing \end{aligned}$$

Using ordered trapezoids, we can define the process for sampling a trapezoid, $\varepsilon(T)$ to produce a consistent vector that satisfies the trapezoid. Note that $\varepsilon(I_n[\vec{w}])$ represents a random choice of a value consistent in type with variable n and satisfying the interval I_n evaluated at the vector $\vec{w}$ (assuming such a value exists), $\{\}$ represents an empty vector, and $\vec{w}[i] = x$ represents an update of vector $\vec{w}$ such that dimension i is equal to the value x.

$$\frac{}{\varepsilon(\varnothing) \rightarrow \{\}} \qquad \frac{\varepsilon(T_m) \rightarrow \vec{w}}{\varepsilon(I_n \cap T_m) \rightarrow \vec{w}[n] = \varepsilon(I_n[\vec{w}])}$$

3.1 Restriction

The purpose of post-processing is to restrict the domain of values that can be assigned to selected variables in the trapezoid to ensure that the sampling process can proceed without backtracking. We call the trapezoidal post-processing step *restriction*. The restriction process takes as input a trapezoid, a change of basis (σ, initially an identity function), and a consistent reference vector. The output of the restriction process is a new (restricted) trapezoid and a change of basis that maps vectors sampled from the restricted trapezoid back into the vector space of the original trapezoid.

$$T, \sigma, \vec{v} \xrightarrow{R} T', \sigma'$$

The restriction process is specified as a single pass over the trapezoid intervals, starting with the largest interval and proceeding to the smallest. Each interval in the trapezoid is restricted by applying the steps summarized below.

- **Bound Fixing (BF)** Ensures that linear bounds are expressed only in terms of $\geq$ and $\leq$.
- **Integer Equality (IE)** Ensures that equalities involving integer variables are always satisfiable.
- **Interval Restriction (IR)** Ensures that intervals are always satisfiable.

Each interval in the trapezoid may suggest a domain restriction, in the form of a set of linear constraints, and a change of basis in the form of a linear transformation. Domain restrictions are intersected with the remaining trapezoid before it, too, is restricted. Changes of base are accumulated and applied to the current interval, the remaining trapezoid, and the reference vector. When complete, the final restricted trapezoid and the final change of basis are returned.

$$\frac{}{\varnothing, \sigma, \vec{v} \xrightarrow{R} \varnothing, \sigma} \text{ RESTRICTION-BASE}$$

$$\frac{\begin{array}{c} I_n \cap T_m, \sigma, \vec{v} \xrightarrow{BG} I'_n \cap T'_m, \sigma', \vec{v}' \\ I'_n \cap T'_m, \sigma', \vec{v}' \xrightarrow{IE} I''_n \cap T''_m, \sigma'', \vec{v}'' \\ I''_n \cap T''_m, \sigma'', \vec{v}'' \xrightarrow{IR} I'''_n \cap T'''_m, \sigma''', \vec{v}''' \\ T'''_m, \sigma''', \vec{v}''' \xrightarrow{R} T''''_m, \sigma'''' \end{array}}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{R} I'''_n \cap T''''_m, \sigma'''} \text{ RESTRICTION-STEP}$$

The restricted trapezoid can then be sampled using our simple sampling algorithm and the resulting vectors can be mapped back into the original vector space using the change of basis in such a way that $\sigma'[\varepsilon(T')]$ is a type consistent vector and $T[\sigma'[\varepsilon(T')]]$ is true.

The following discussion assumes that all integer valued variables are bounded only by other integer valued variables. This can be ensured if, say, the dimension of each rational variable is always greater than the dimension of every integer variable.

3.2 Bound Fixing (BF)

Bound fixing transforms exclusive inequalities on integer variables into inclusive inequalities and tightens any remaining integer bounds.

$$\frac{BF(I_n) \to I'_n}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{BF} I'_n \cap T_m, \sigma, \vec{v}} \text{ BOUND-FIXING}$$

We omit a more detailed description of this operation for space considerations.

3.3 Integer Equality (IE)

Any integer variable equality bound appearing in a trapezoid can be expressed as:

$$x_n = (N_k * x_k + \ldots + N_0)/D_n$$

where all x_i and N_i are integer valued and D_n is a positive integer, representing the least common denominator of the bounding polynomial coefficients. Given $a, b \in \mathbb{Z}$ we write $(a|b)$, called "a divides b", if and only if there exists $c \in \mathbb{Z}$ such that $b = a * c$. A divisibility constraint is an expression of the form $(m|E)$ where $m \in \mathbb{Z}$ and $E \in \mathscr{E}$. A vector $\vec{v}$ is called a *solution* to a divisibility constraint $(m|E)$ if and only if $(m|E[\vec{v}])$. A divisibility constraint is called solvable if and only if it has at least one solution. To ensure that x_n has an integer solution, $\vec{w}$ must be chosen so that $(D_n|(N_k * x_k + \ldots + N_0)[\vec{w}])$. Of course if D_n is 1, this is trivial and no additional work is required.

Given a non-trivial, solvable divisibility constraint $(m|E)$ we want to find an efficient way to enumerate all solutions. Towards this end, we introduce the concept of a *Trapezoidal Change of Basis* (TCOB): a variable substitution that preserves dimension.

Definition 3.1 (Trapezoidal Change of Basis)**.** *A map $\sigma : \mathscr{E} \to \mathscr{E}$ is called an trapezoidal change of basis, if*

1. $\sigma(c_0 + c_1x_1 + \cdots + c_nx_n) = c_0 + c_1\sigma(x_1) + \cdots + c_n\sigma(x_n)$*, i.e.* σ *is a homomorphism, and*
2. $\dim(\sigma(x_i)) = \dim(x_i)$*, for all* $i = 1, \ldots, n$.

For a solvable divisibility constraint $(m|E)$, we propose to construct a TCOB σ such that 1) every vector is a solution to the divisibility constraint $(m|\sigma(E))$ and 2) each solution to $(m|E)$ is represented by a solution to $(m|\sigma(E))$.

We define the notion of *span* to clarify when exactly a TCOB preserves a solution:

Definition 3.2. *Let σ be a trapezoidal change of basis defined by $\sigma(x_i) = E_i$ for each $i = 1, \ldots, n$. Let $\vec{v}$ and $\vec{\eta}$ be vectors. We say $\vec{v} = \sigma[\vec{\eta}]$ if and only if $\vec{v} = (E_1[\vec{\eta}], \ldots, E_n[\vec{\eta}])$. We define* $\text{span}(\sigma) = \{\vec{v} : \exists\vec{\eta}.\ \vec{v} = \sigma[\vec{\eta}]\}$.

Lemma 3.3. *For every $E \in \mathscr{E}$ we have $E[\sigma[\vec{\eta}]] = \sigma(E)[\vec{\eta}]$.*

Proof. By induction on the structure of E. □

Lemma 3.4. *Let $(m|E)$ be a solvable divisibility constraint. Then there exists a trapezoidal change of basis σ such that*

1. $\sigma(E) = mE'$ *for some E', and*
2. $\forall\vec{v}.\ (m|E[\vec{v}]) \Longrightarrow \vec{v} \in \text{span}(\sigma)$.

Note that first condition implies that every vector is a solution to $(m|\sigma(E))$.

Proof. Induction on n, the number of dimensions. In the base case $n = 0$ for which case $E = c$ where $c \in \mathbb{Z}$. Take σ to be the identity map on $\mathscr{E}$. Since $(m|c)$ is solvable we have $c = md$ for some $d \in \mathbb{Z}$. Thus $\sigma(E) = md$, satisfying property 1. The second property is trivially true since there is only a single (empty) vector of dimension 0.

Now consider the case for $n > 0$ dimensions. Let $E = F + cx$ where $dim(F) < n$, $c \in \mathbb{Z}$, and $dim(x) = n$. Let $g = \gcd(c, m)$ and let c' and m' be such that $c = gc'$ and $m = gm'$. Since $(m|E)$ is solvable, $(g|E)$ must also be solvable. Furthermore, we know $(g|c)$ so $(g|F)$ is solvable. By the inductive hypothesis, let σ' be a TCOB with

1. $\sigma'(F) = gF'$ for some F', and
2. $\forall\vec{v}.\ (g|F[\vec{v}]) \Longrightarrow \vec{v} \in \text{span}(\sigma')$.

Then, let σ be an extension of σ' with $\sigma(x) = m'x - \text{invmod}(c', m')F'$. Here $\text{invmod}(c', m')$ is the multiplicative inverse of c' modulo m' which is well-defined since $\gcd(c', m') = 1$. Note also that

$dim(\sigma(x)) = dim(x)$. Using σ we have,

$$\begin{aligned}
\sigma(E) &= gF' + c\sigma(x) \\
&= gF' + c(m'x - \text{invmod}(c', m')F') \\
&= cm'x + gF' - c\,\text{invmod}(c', m')F' \\
&= cm'x + gF' - gc'\,\text{invmod}(c', m')F' \\
&= cm'x + g(1 - c'\,\text{invmod}(c', m'))F' \\
&= cm'x + g(1 - (1 + m'k))F' && \text{for some } k \in \mathbb{Z} \\
&= cm'x + gm'kF' \\
&= gc'm'x + gm'kF' \\
&= gm'(c'x + kF') \\
&= m(c'x + kF')
\end{aligned}$$

So that σ satisfies property 1.

To prove property 2 on σ, let $\vec{v}$ be a vector such that $(m|E[\vec{v}])$. Then we have $(m|(F[\vec{v}] + c\vec{v}_x))$ where $\vec{v}_x$ is the x-component of $\vec{v}$. This implies $(g|F[\vec{v}])$ and so by the inductive hypothesis we have $\vec{v}' \in span(\sigma')$ where $\vec{v}'$ is $\vec{v}$ without the final x-component. This means there is a vector $\vec{\eta}'$ such that $\vec{v}' = \sigma[\vec{\eta}']$. To show $\vec{v} \in \text{span}(\sigma)$ we want to extend $\vec{\eta}'$ with an x-component, $\vec{\eta}_x$, such that $\vec{v} = \sigma[\vec{\eta}]$. Such a $\vec{\eta}_x$ exist if and only if we can satisfy $\vec{v}_x = m'\vec{\eta}_x - \text{invmod}(c', m')F'[\vec{\eta}']$. This equation will have an integral solution for $\vec{\eta}_x$ if and only if

$$\begin{aligned}
\vec{v}_x &\equiv -\text{invmod}(c', m')F'[\vec{\eta}'] \pmod{m'} && \iff \\
\vec{v}_x + \text{invmod}(c', m')F'[\vec{\eta}'] &\equiv 0 \pmod{m'} && \iff \\
c'\vec{v}_x + c'\,\text{invmod}(c', m')F'[\vec{\eta}'] &\equiv 0 \pmod{m'} && \iff \\
c'\vec{v}_x + F'[\vec{\eta}'] &\equiv 0 \pmod{m'} && \iff \\
gc'\vec{v}_x + gF'[\vec{\eta}'] &\equiv 0 \pmod{gm'} && \iff \\
c\vec{v}_x + \sigma(F)[\vec{\eta}'] &\equiv 0 \pmod{m} && \iff \\
c\vec{v}_x + F[\vec{v}'] &\equiv 0 \pmod{m}
\end{aligned}$$

Note that $F[\vec{v}'] = F[\vec{v}]$ since $dim(F) < n$. Thus we only need to show $(m|(c\vec{v}_x + F[\vec{v}]))$ which is equivalent to $(m|E[\vec{v}])$. This is true by assumption. Thus property 2 also holds on σ. □

The proof of Lemma 3.4 contains an algorithm for recursively computing a TCOB σ for any solvable divisibility constraint $(m|E)$. We demonstrate this algorithm in the following example.

Example 3.5. *Consider the divisibility constraint* $(2|(1+x+y))$ *where* $dim(x) = 1$ *and* $dim(y) = 2$*. The algorithm implicit in the proof of Lemma 3.4 produces the TCOB* $\sigma(x) = x$, $\sigma(y) = 2y - x - 1$*. Letting* $E = 1 + x + y$ *we have* $\sigma(E) = 2y$ *so that* $(2|\sigma(E))$ *is true for all vectors.*

Given a constraint $x_n = (N_k * x_k + \ldots + N_0)/D_n$ we have shown that we can compute a TCOB σ so that $(D_n|(N_k * x_k + \ldots + N_0))$ is always satisfied, allowing us to always compute an integral value for x_n. We can translate any such satisfying solution $(\vec{\eta})$ into a solution of our original trapezoid via $\vec{v} = \sigma[\vec{\eta}]$. It is easy to show that σ preserves all constraints on trapezoids. Therefore, we can repeatably apply this procedure starting from the highest dimensions and working our way down, eliminating all non-trivial integral constraints along the way. Our final step is to update our reference vector to reflect the change

of base, $\vec{v}' = \sigma^{-1}[\vec{v}] = \{\vec{w} : \vec{v} = \sigma[\vec{w}]\}$. We know that such a vector exists because the span of σ includes all of the vectors that satisfied the original constraint, including the original reference vector.

$$\frac{\begin{array}{l} I_n \equiv (x_n = P/D_n) \\ x_n \in \mathbb{Z} \quad (D_n|P) \rightarrow \sigma_n \quad I'_n = (x_n = \sigma_n(P)/D_n) \\ \sigma' = \sigma_n(\sigma) \quad \vec{v}' = \sigma_n^{-1}(\vec{v}) \quad T'_m = \sigma_n(T_m) \end{array}}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{IE} I'_n \cap T'_m, \sigma', \vec{v}'} \text{ INTEGER-EQUALITY}$$

3.4 Interval Restriction

While double bounded intervals are known to be satisfiable at the reference vector, there is no guarantee that they are satisfiable under arbitrary variable assignments. To ensure satisfiability, we need to ensure that, for any valid variable assignment, the upper bound will not be less than the lower bound and, for integer variables, that an integer solution exists between the upper and lower bound.

For all rational variables and for integer variables when $lcd(P^L) = 1$ or $lcd(P^U) = 1$ we simply intersect the domain restriction $\langle\langle P^L \leq P^U \rangle\rangle$ with the remaining trapezoid. This ensures that, for any variable assignment that satisfies the newly restricted trapezoid, the lower bound will not be greater than the upper bound. For integer variables this restriction is sufficient because either P^L or P^U will be an integer valued polynomial and, as a result, when $P^L \leq P^U$ is true it always contains an integer solution.

$$\frac{\begin{array}{l} I_n = (P^L \leq x_n) \cap (x_n \leq P^U) \\ (\ x_n \in \mathbb{Q} \ \vee \ lcd(P^L) = 1 \ \vee \ lcd(P^U) = 1\ ) \\ \langle\langle P^L \leq P^U \rangle\rangle \cap T_m \xrightarrow{\vec{v}}{}^* T'_m \end{array}}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{IR} I_n \cap T'_m, \sigma, \vec{v}} \text{ INTERVAL-RESTRICTION-1}$$

When $lcd(P^L) > 1$ and $lcd(P^U) > 1$ the constraint $P^L \leq P^U$ is not sufficient to ensure an integer solution for x because P^L and P^U are not guaranteed to be integer valued for every variable assignment. The constraint $P^L + 1 \leq P^U$, on the other hand, does always contain an integer solution. If this constraint contains the reference vector (ie: $P^L[\vec{v}] + 1 \leq P^U[\vec{v}]$) then it is conservative to use this constraint as a domain restriction.

$$\frac{\begin{array}{l} I_n = (P^L \leq x_n) \cap (x_n \leq P^U) \\ x_n \in \mathbb{Z} \\ lcd(P^L) > 1 \\ lcd(P^U) > 1 \\ P^L[\vec{v}] + 1 \leq P^U[\vec{v}] \\ \langle\langle P^L + 1 \leq P^U \rangle\rangle \cap T_m \xrightarrow{\vec{v}}{}^* T'_m \end{array}}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{IR} I_n \cap T'_m, \sigma, \vec{v}} \text{ INTERVAL-RESTRICTION-2}$$

If $lcd(P^L) > 1$, $lcd(P^U) > 1$, and $P^L[\vec{v}] + 1 > P^U[\vec{v}]$ we perform a change of base to ensure that either the upper or lower bound is always integral. Here we consider the change of base for the lower bound

with the understanding that the case of the upper bound is symmetrical. We first adjust (tighten) the lower bound in such a way that it is equal to $x[\vec{v}]$ when evaluated at the reference vector. Note that, because $P^L[\vec{v}]+1 > P^U[\vec{v}]$, there is only 1 integer solution in this interval and $x_n[\vec{v}]$ is it. We compute a change of base for P'^L to ensure that it is always integral. We then restrict the domain of T_m to ensure that the resulting upper bound is never less than our new lower bound.

$$\frac{\begin{array}{l} I_n = (P^L \leq x_n) \cap (x_n \leq P^U) \\ x_n \in \mathbb{Z} \quad lcd(P^L) > 1 \quad lcd(P^U) > 1 \\ P'^L = P^L + (x_n[\vec{v}] - P^L[\vec{v}]) \\ D^L = lcd(P'^L) \\ (D^L | P'^L) \rightarrow \sigma_n \\ P'^U = \sigma_n(P^U) \quad P''^L = \sigma_n(P'^L)/D^L \\ I'_n = (P''^L \leq x_n) \cap (x_n \leq P'^U) \\ \sigma' = \sigma_n(\sigma) \quad \vec{v}' = \sigma_n^{-1}(\vec{v}) \quad T'_m = \sigma_n(T_m) \\ \left\langle \left\langle P''^L \leq P'^U \right\rangle \right\rangle \cap T'_m \xrightarrow{v'} {}^* T''_m \end{array}}{I_n \cap T_m, \sigma, \vec{v} \xrightarrow{IR} I'_n \cap T''_m, \sigma', \vec{v}'} \text{ INTERVAL-RESTRICTION-3}$$

4 Implementation and ACL2 Formalization

FuzzM[8] is a model-based fuzzing framework, implemented primarily in Java, that employs Lustre[10] as a modeling and specification language and leverages counterexamples produced by the the JKind[4] model checker to drive the fuzzing process. The FuzzM infrastructure includes support for trapezoidal generalization of JKind counterexamples relative to Lustre models involving linear constraints, integer division and remainder, and uninterpreted functions. Our implementation of trapezoidal generalization requires approximately 5K lines of Java code.

Early versions of the fuzzing framework leveraged an interval generalization capability provided by JKind. At the time we observed that, while the results of interval generalization are trivial to sample, it does not generalize linear model features well. Trapezoidal generalization was envisioned as an extension of interval generalization that would preserve the property of efficient sampling while improving support for linear constraints. In architecting the generalization procedure, however, it soon became clear that we didn't know what it meant for our generalizer to be correct. To address this issue we developed a formalization of generalization correctness in ACL2. In a rump session of the 2017 ACL2 workshop we presented our formalization of generalization correctness, a high-level specification for our implementation, and a proof that our implementation was correct[7]. We also reported that, in proving that our implementation was correct, we had identified an error in our initial high-level specification. Our inability to correctly specify even the high-level behavior of our system without mechanical assistance further motivated us to formalize and verify the correctness of our low-level implementation as well. Much of our low-level ACL2 specification has been transliterated from the Java implementation to ensure that it captures our essential design decisions. While the proofs of low-level correctness were useful in ensuring the absence of off-by-one errors and the like, perhaps the primary benefit of this activity was in the formalization process itself. Just as formalizing our high-level algorithms helped to clarify our software

architecture, formalizing the low-level behavior of our algorithms has helped us both to clarify and to simplify the resulting implementation.

We now have a proof that the low-level implementation of our trapezoidal generalization procedure over linear constraints satisfies our notion of generalization correctness. At the heart of our formalization is a library for manipulating, reasoning about, and evaluating linear relations over rational polynomials. On top of this library we introduce a notion of normalized variable bounds upon which we define the trapezoidal data structure and a predicate that recognizes ordered trapezoids and regions. The proof that our low-level implementation satisfies our high level specification was largely an exercise in proving that each low-level function preserved our set of correctness invariants. While the majority of this proof effort was routine, two specific challenges were addressed with specialized ACL2 libraries.

First, to prove that functions operating on normalized variable bounds and trapezoidal regions preserve our variable ordering invariants we employ an arcane feature of the *nary*[5] library that, effectively, enables us to treat subset relations as equivalence relations in appropriate contexts, rewriting subset terms into their super-sets. This feature allows us to specify function contracts using rewrite rules, rather than back-chaining rules, to establish ordering relations among the variables appearing in the function's results. In the supporting materials see `set-upper-bound-equiv` and `set-upper-bound-ctx` defined by the `defequiv+` event in the file `poly.lisp` along with the congruence rule `set-upper-bound-append` and the family of driver rules exemplified by `>-all-upper-bound-backchaining`. See also the nary equivalence relation `set-upper-bound-equiv` as used in `set-upper-bound-equiv-all-bound-list-variables-restrict` to specify a subset rewrite for specific variables returned by the function `intersect` in the file `intersect.lisp` and compare it with the more traditional `subset-p` rule found just above in the lemma `trapezoid-p-restrict`.

Second, because the termination of the function for performing top-level trapezoidal intersection is a bit tricky, in that it is doubly recursive, reflexive, and it relies on the variable ordering property of trapezoids, we introduced its definition using the *def::ung*[9, 6] macro. This allowed us to postpone its termination proof until we had established that the intersection function preserved the variable ordering. We then used *def::total* to prove that the function does, in fact, terminate on well-formed trapezoids. These events are found in the file `intersect.lisp` in the supporting material.

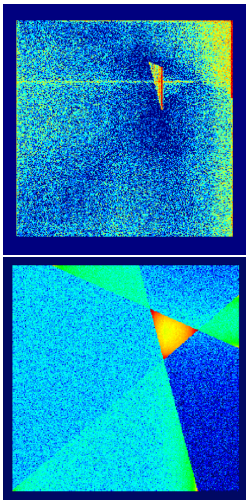
In future work we plan to verify that our method for trapezoidal restriction, as described in Section 3, is correct. Specifically we hope to show that restricted trapezoids are valid generalizations and that they can be sampled to produce satisfying variable assignments without backtracking, even for integer variables. We also hope to extend our infrastructure to allow us to verify that our approaches for generalizing uninterpreted function instances and integer division and remainder, which we do not discuss in this paper, are sound.

5 Related Work

We have compared the performance of our trapezoidal generalization technique with the interval generalization capability provided by the JKind model checker[4]. In all but the smallest models, trapezoidal generalization was competitive with interval generalization in terms of generalization speed. For large models, interval generalizations are actually more expensive to compute than trapezoidal generalizations. This is likely because trapezoidal generalizations are computed in a single symbolic simulation of a model whereas the interval generalization technique requires multiple simulations to compute. Within our fuzzing framework the sampling rates for trapezoidal generalizations are also competitive with interval generalization. Tests in our framework suggest that interval sampling rarely provides better than

twice the overall performance of trapezoidal sampling, and is typically less than 50 percent faster[3]. In terms of absolute performance, our deployed system routinely exhibits effective trapezoidal sampling rates on the order of 100's of thousands of generalized variables per second.

Trapezoidal generalization also clearly outperforms interval generalization in terms of preserving linear model features. On the right are two heat maps generated by capturing thousands of samples from generalized solutions to a number of properties involving arbitrary Boolean combinations of three linear constraints. The upper heat map employed interval generalization and the lower trapezoidal generalization. The three linear constraints involved in the various properties are clearly visible in the lower heat map, whereas the upper heat map is dominated by interval generalization artifacts.



Welp[12] discusses a polytope generalization which is more expressive than trapezoidal generalization but is also more expensive to manipulate. Operations on polytopes can be exponential in both time and space. The worst case size of our representation is quadratic in the number of variables. The intersection operation (on ordered trapezoids) is worst case cubic and a naive implementation of post-processing is worst case quartic for integer domains. Sampling a polytope is at least as hard as linear programming and for integers it is NP-complete. After post-processing, sampling a trapezoidal generalization requires only a quadratic number of evaluations. Trapezoidal generalization is thus more expressive than interval generalization but more efficient than polytope generalization, both from a computational point of view as well as from the perspective of sampling.

Symbolic generalization techniques are also used in abstract interpretation[2]. Because efficiency is an overriding concern the abstractions commonly used therein are often quite limited[3]. While more complex and expressive representations are typically less efficient, the octagon domain[11] is actually quite comparable to trapezoids in terms of both computational and representational complexity. Efficient sampling of such abstract domains, however, is generally not a concern. Also, the objective of generalization in abstract interpretation is typically verification; generalization for this purpose is common[1]. When used for verification, generalization should provide a conservative over-approximations of variable domains. Our specification, on the other hand, calls for a conservative under-approximation of such domains. As a result, a generalization technique designed for one domain would be unsound in the other.

6 Conclusion

We have presented a generalization technique for logical formulae described in terms of Boolean combinations of linear constraints that is optimized for use with model-based fuzzing. The generalization technique employs trapezoidal solution sets that can be computed with reasonable space and time bounds and then sampled efficiently. We developed a formal specification of generalization correctness and used it to verify the correctness of key aspects of our generalization algorithm using ACL2. Finally we described a post-processing procedure that produces restricted trapezoidal solutions that can be rapidly and efficiently sampled without backtracking, even for integer domains. The formal verification of this post-processing step is expected to be addressed in future work.

[3]Measurements are based on overall test vector generation rates, not generalization sampling rates in isolation

References

[1] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu & Helmut Veith (2003): *Counterexample-guided Abstraction Refinement for Symbolic Model Checking*. *J. ACM* 50(5), pp. 752–794, doi:10.1145/876638.876643. Available at `http://doi.acm.org/10.1145/876638.876643`.

[2] Patrick Cousot & Radhia Cousot (1977): *Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints*. In: *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, POPL '77, ACM, New York, NY, USA, pp. 238–252, doi:10.1145/512950.512973. Available at `http://doi.acm.org/10.1145/512950.512973`.

[3] Patrick Cousot & Radhia Cousot (1977): *Static Determination of Dynamic Properties of Generalized Type Unions*. *SIGPLAN Not.* 12(3), pp. 77–94, doi:10.1145/390017.808314. Available at `http://doi.acm.org/10.1145/390017.808314`.

[4] Andrew Gacek, John Backes, Mike Whalen, Lucas G. Wagner & Elaheh Ghassabani (2018): *The JKind Model Checker*. In: *Computer Aided Verification - 30th International Conference, CAV 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17, 2018, Proceedings, Part II*, pp. 20–27, doi:10.1007/978-3-319-96142-2_3. Available at `https://doi.org/10.1007/978-3-319-96142-2_3`.

[5] David Greve (2006): *Parameterized Congruences in ACL2*. In: *Proceedings of the Sixth International Workshop on the ACL2 Theorem Prover and Its Applications*, ACL2 '06, ACM, New York, NY, USA, pp. 28–34, doi:10.1145/1217975.1217981. Available at `http://doi.acm.org/10.1145/1217975.1217981`.

[6] David Greve (2009): *Assuming Termination*. In: *Proceedings of the Eighth International Workshop on the ACL2 Theorem Prover and Its Applications*, ACL2 '09, ACM, New York, NY, USA, pp. 114–122, doi:10.1145/1637837.1637856. Available at `http://doi.acm.org/10.1145/1637837.1637856`.

[7] David Greve (2017): *Generalization Correctness*. `http://www.cs.utexas.edu/users/moore/acl2/workshop-2017/slides-rump/greve-ACL2-2017.pdf`.

[8] David Greve (2018): *FuzzM: Fuzzing with Models*. `https://github.com/collins-research/FuzzM`.

[9] David Greve & Konrad Slind (2013): *A Step-Indexing Approach to Partial Functions*. In Ruben Gamboa & Jared Davis, editors: Proceedings International Workshop on the *ACL2 Theorem Prover and its Applications*, Laramie, Wyoming, USA , May 30-31, 2013, *Electronic Proceedings in Theoretical Computer Science* 114, Open Publishing Association, pp. 42–53, doi:10.4204/EPTCS.114.4.

[10] N. Halbwachs, P. Caspi, P. Raymond & D. Pilaud (1991): *The synchronous data flow programming language LUSTRE*. *Proceedings of the IEEE* 79(9), pp. 1305–1320, doi:10.1109/5.97300.

[11] Antoine Miné (2006): *The Octagon Abstract Domain*. *Higher Order Symbol. Comput.* 19(1), pp. 31–100, doi:10.1007/s10990-006-8609-1. Available at `http://dx.doi.org/10.1007/s10990-006-8609-1`.

[12] Tobias Welp & Andreas Kuehlmann (2013): *QF_BV Model Checking with Property Directed Reachability*. In: *Proceedings of the Conference on Design, Automation and Test in Europe*, DATE '13, EDA Consortium, San Jose, CA, USA, pp. 791–796, doi:10.7873/DATE.2013.168. Available at `http://dl.acm.org/citation.cfm?id=2485288.2485480`.