Data-Driven Inference of Representation Invariants

Anders Miltner  
Princeton University  
amiltner@cs.princeton.edu

Saswat Padhi  
UCLA  
padhi@cs.ucla.edu

Todd Millstein  
UCLA  
todd@cs.ucla.edu

David Walker  
Princeton University  
dpw@cs.princeton.edu

Abstract

A representation invariant is a property that holds of all values of abstract type produced by a module. Representation invariants play important roles in software engineering and program verification. In this paper, we develop a counterexample-driven algorithm for inferring a representation invariant that is sufficient to imply a desired specification for a module. The key novelty is a type-directed notion of visible inductiveness, which ensures that the algorithm makes progress toward its goal as it alternates between weakening and strengthening candidate invariants. The algorithm is parameterized by an example-based synthesis engine and a verifier, and we prove that it is sound and complete for first-order modules over finite types, assuming that the synthesizer and verifier are as well. We implement these ideas in a tool called HANOI, which synthesizes representation invariants for recursive data types. HANOI not only handles invariants for first-order code, but higher-order code as well. In its back end, HANOI uses an enumerative synthesizer called Myth and an enumerative testing tool as a verifier. Because HANOI uses testing for verification, it is not sound, though our empirical evaluation shows that it is successful on the benchmarks we investigated.

1 Introduction

A representation invariant is a property that holds of all values of abstract type produced by a module. For instance, a module that implements a set using a list might maintain a no duplicates or is sorted invariant over the lists. Module implementers can rely on the invariant for correctness and efficiency and must ensure that it is maintained by each function in the module. Making representation invariants explicit has a number of software engineering benefits: they can be used as documentation, dynamically checked as contracts [9, 16], and used for automated testing [3, 6].

Representation invariants also play a key role in modular verification of software components. Consider a module that implements sets; its specification φ might demand that (lookup (insert s i) i) return true for all sets s and items i. A standard way to prove such a specification [1] is in two steps: 1) prove that a predicate I is a representation invariant of the module; and 2) prove that I is stronger than φ, i.e., all module states that satisfy I also satisfy φ. In other words, modular verification can be reduced to the problem of synthesizing a sufficient representation invariant.

In this paper, we develop an approach to automatically infer a sufficient representation invariant given a pure, functional module and a specification. To our knowledge, the only prior work to tackle this problem [15] builds candidate invariants out of a fixed set of atomic predicates and provides no correctness guarantees. We address both of these limitations through a form of counterexample-guided inductive synthesis (CEGIS) [27, §5], which consists of an interaction between two black-box components: (1) a synthesizer that generates a candidate invariant consistent with given sets of positive and negative examples, (2) a verifier that either proves that a candidate is a sufficient representation invariant or produces a counterexample, which becomes a new example for the synthesizer.

Our approach is inspired by recent work in data-driven inference of inductive invariants in other settings [7, 20, 24, 30]. As in that work, a key challenge is how to handle inductiveness counterexamples, pairs of module states (s, s′) such that s satisfies the candidate invariant, some module operation transforms state s to state s′, but s′ does not satisfy the candidate invariant. The problem is that there are two ways to resolve such counterexamples and it is not clear which is correct: treat s as a new negative example or treat s′ as a new positive example.

We observe that if s is a constructible state of the module, meaning that it is reachable by a sequence of module operations, then s′ must be as well. Therefore, any representation invariant will include both states, so we must treat s′ as a new positive example. Based on this observation, we define a candidate invariant I to be visibly inductive on a module relative to a set S of known constructible states if every module operation produces a state satisfying I when invoked from a state in S. For each candidate invariant, we first iteratively weaken it until it is visibly inductive relative to the current set of known constructible states, in the process adding new states to this set, and only then do we consider
other inductiveness violations. Intuitively, this approach eagerly identifies and exploits inductiveness counterexamples for which no “guessing” is required.

We have formalized a general notion of inductiveness as a type-indexed logical relation, of which both our notion of visible inductiveness and the traditional notion of (full) inductiveness are special cases. We have also formalized our overall algorithm using this notion. We have proven the algorithm sound and complete when the module contains first-order code and the implementation of the abstract type is a finite domain, provided the underlying synthesizer and verifier are also sound and complete.

We have implemented our algorithm in OCaml and call the resulting tool Hanoi. To instantiate the verification component of the system, we use Myth [19], a type- and example-directed synthesis engine. Myth is capable of synthesizing invariants over recursive data types in many cases, so it is a good fit for tackling proofs about modules that implement recursive data types, which are the focus of our benchmarks. To instantiate the verification component, we use a form of enumerative test generation. Despite the unsoundness of this underlying verifier, our experimental results show that Hanoi still infers sufficient representation invariants in practice. Such likely representation invariants can be used by module implementers and verifiers for many purposes.

We have also implemented extensions that allow Hanoi to be used with higher-order code. Here, the main challenge comes in how to extract counterexamples from higher-order arguments. It turns out that our first-order scheme for extracting counterexamples is essentially an application of a first-order contract that guards and logs values passing through the first-order interface. The solution to counterexample-extraction from higher-order code then is to implement higher-order contracts [9] that guard and log values across this higher-order interface.

To evaluate our tools, we constructed a benchmark suite that includes 28 different modules, including a variety of modules over lists and trees, many drawn from Coq libraries and books [1]. We find that Hanoi is able to synthesize 22 of these invariants within 30 minutes.

To summarize, the main contributions of this work are:

- An algorithm for automated synthesis of representation invariants, parameterized by a verifier and synthesizer.
- A formalization of the algorithm and the key notion of visible inductiveness, over a first-order type theory. We prove soundness and completeness in the case of finite domains, if the given verifier and synthesizer are sound and complete.
- An extension of the algorithm capable of extracting counterexamples from higher-order interfaces.
- Implementation, optimization and evaluation of a tool called Hanoi that synthesizes invariants over recursive

Figure 1. A module that implements SET using lists.

data types, using an unsound, enumerative testing engine for verification.

2 A Motivating Example

In this section, we give a high-level overview of Hanoi using an example. Consider the interface SET:

```
module type SET = sig
  type t
  val empty : t
  val insert : t -> int -> t
  val delete : t -> int -> t
  val lookup : t -> int -> bool
end
```

The interface declares an abstract type t and a number of functions that operate over t. Figure 1 shows a module ListSet that implements the SET interface, using int list as the concrete type.

We study the problem of verifying that ListSet satisfies some standard properties of sets. An example specification follows.

\[(\phi \ s) \iff \forall i : \text{int}. \\
\neg (\text{lookup empty } i) \\
\land (\text{lookup } (\text{insert } s \ i) \ i) \land \neg (\text{lookup } (\text{delete } s \ i) \ i)\]

Note that this specification does not hold for arbitrary integer lists. For example, \(\text{lookup } (\text{delete } [1;1] \ 1)\) returns true. Nonetheless, the ListSet module is a correct implementation of the SET interface, because the specification holds for all values of the abstract type t that the module can actually construct. Such values are usually called the representations of the abstract type t. To emphasize that such values can be constructed by execution of module operations, we say a value $v$ is constructible at type $\tau$ whenever a client with access to the module can produce $v$ at the type $\tau$. 
\( V_s = \) a set of known constructible values \( I = \) a candidate invariant

\[
\begin{align*}
I_s & = \{ [] \} \\
V_s & = \{ [1; 3], [1; 1] \}
\end{align*}
\]

**Figure 2.** A sufficient representation invariant \( I_s \) implies the spec and is an overapproximation of the set of representations \( R \) of the module’s abstract type.

A standard approach \cite{1} to prove that a module implementation satisfies such a specification is to identify a **sufficient representation invariant**. In our example, such an invariant for ListSet is a predicate \( I_s : (\text{int list} \rightarrow \text{bool}) \) that is

- sufficient for \( \phi \), i.e. \( \forall s : \text{int list}. (I_s s) \implies (\phi s) \), and
- whenever operations of ListSet module are supplied with argument values of abstract type that satisfy the invariant, they produce values of abstract type that satisfy the invariant, i.e., the module is **inductive** with respect to the invariant.

In other words, \( I_s \) contains all integer lists that are representations of type \( t \), and is contained in the set of integer lists that satisfy \( \phi \). Figure 2 shows this relationship pictorially.

For ListSet, the predicate demanding an integer list has no duplicates is a sufficient representation invariant for \( \phi \). Our tool Hanoi automatically generates that invariant:

\[
\begin{align*}
1 & \text{let rec } I_s : \text{int list} \rightarrow \text{bool} = \text{function} \\
2 & \mid [] \rightarrow \text{true} \\
3 & \mid \text{hd :: tl} \rightarrow (\text{not (lookup tl hd)}) \land (I_s \text{ tl})
\end{align*}
\]

### 2.1 Overview of Hanoi

Given a module, an interface and a specification, Hanoi employs a form of **counterexample-guided inductive synthesis** (CEGIS) \cite{27, 6} to infer a sufficient representation invariant. Specifically, we use a generate-and-check approach that iterates between two black-box components: (1) a synthesizer Synth that generates a candidate invariant, which is a predicate that separates a set \( V_s \) of positive and a set \( V_\neg \) of negative examples, and (2) a verifier Verify that checks if a candidate invariant satisfies the desired properties and otherwise generates a counterexample. CEGIS has been successfully applied to other forms of invariant inference \cite{7, 20, 24, 30}.

We illustrate Hanoi and its key challenges via our running example. Initially the \( V_s \) and \( V_\neg \) sets are empty, so suppose that Synth generates the candidate invariant \( \text{fun } _{-} \rightarrow \text{true} \). This invariant is inductive, but not sufficient. Hence Verify will provide a counterexample, for instance \( [1; 1] \), which is an integer list that satisfies the candidate invariant but not the specification \( \phi \) (see \( z \) in Figure 2). As the final invariant must imply \( \phi \), \( [1; 1] \) is added to \( V_s \), which forces Synth to choose a stronger candidate invariant in the next round.

The main challenge in using this approach is the need to handle counterexamples to inductiveness. For instance, suppose that at some point during the algorithm we have \( V_s = \{ [1; 3], [1; 1] \} \) and \( V_\neg = \{ [1; 1] \} \), and suppose Synth generates the following candidate invariant:

\[
\begin{align*}
1 & \text{let } I : \text{int list} \rightarrow \text{bool} = \text{function} \\
2 & \mid [] \rightarrow \text{true} \\
3 & \mid \text{hd :: _} \rightarrow \text{hd} \not= 1
\end{align*}
\]

This candidate is not inductive over the ListSet module. For instance, \( [\emptyset] \) satisfies the candidate, but \( \text{insert} [\emptyset] 1 = [1; \emptyset] \) does not. Hence the pair \( ([\emptyset], [1; \emptyset]) \) constitutes an inductiveness counterexample (see \( x, y \) in Figure 2). Resolving such an inductiveness counterexample requires ensuring that both either \( x \) and \( y \) satisfy the candidate invariant or that neither does. This leads to two possibilities, and the problem is that it’s unclear which one is correct:

- add \( [\emptyset] \) to \( V_\neg \) so that it will be excluded from the next candidate invariant
- add \( [1; \emptyset] \) to \( V_s \) so that it will be included in the next candidate invariant

However, observe that if \( (x, y) \) is an inductiveness counterexample and \( x \) is known to be constructible, then \( y \) is constructible as well. Since a representation invariant must include all constructible integer lists, there is no choice to make in this case: we must add \( y \) to \( V_s \).

Based on this observation, our algorithm maintains the property that all elements of \( V_s \) both satisfy the current candidate invariant and are known to be constructible. To verify inductiveness of a candidate invariant, we first check a property that we call **visible inductiveness**, which formally requires that there are no inductiveness counterexamples \( (x, y) \) such that \( x \in V_s \). If such a counterexample exists, we add \( y \) to \( V_s \), ask Synth for a new candidate invariant, and re-check visible inductiveness on the updated \( V_s \).

In our running example, the candidate invariant \( I \) shown above is not visibly inductive, so Verify would produce a counterexample, for instance \( [\emptyset], [11] \). Unlike the case for the counterexample \( ([\emptyset], [1; \emptyset]) \) shown earlier, by construction the first element of this new pair is in \( V_s \), so we know that we must add \( [11] \) to \( V_s \). We then re-check visible inductiveness, continuing in this way until the candidate invariant is visibly inductive on \( V_s \).

At that point, we check full inductiveness. Because \( I \) is visibly inductive with respect to \( V_s \), any counterexample to full inductiveness is a pair \( (x, y) \) where \( x \) is not in \( V_s \). In this case, in order to maintain the invariant that \( V_s \) only contains constructible values we resolve the counterexample by adding \( x \) to \( V_s \). So in general, the elements of \( V_s \) all falsify the current candidate invariant, but they may or may not be constructible. With this new negative example, Synth will produce a stronger candidate invariant. We then restart the
process all over again, first weakening this new invariant to be visibly inductive and then strengthening it to be inductive.

In §3.4 we show that despite this interplay between weakening and strengthening, HANOI is sound and complete over finite domains if Verify and Synth are sound and complete. That is, if a sufficient representation invariant exists then HANOI will produce one.

The question of how to handle inductiveness counterexamples arises in prior work on data-driven invariant inference. Some of this work also observes that if \( x \) is constructible in an inductiveness counterexample \( (x, y) \), then so is \( y \) [24, 30]. However, those approaches only leverage this observation opportunistically, when a counterexample to full inductiveness happens to satisfy it. In contrast, we define the notion of visible inductiveness and use this notion to eagerly weaken a candidate invariant until no such counterexamples exist. We demonstrate empirically in Section 5 that our eager search for visible inductiveness counterexamples provides performance benefits. We also prove a completeness result for our approach in the context of finite domains, which those prior approaches lack. To our knowledge, the only prior CEGIS-based approaches to inductive invariant inference that have mains in the presence of these extensions, assuming the verifier and synthesizers are as well.

3 The Inference Algorithm

In this section, we describe our algorithm formally and characterize its key properties.

3.1 Preliminaries

Our programming language is a first-order variant of the simply-typed lambda calculus with functions, pairs, a base type \( (\beta) \) and a single designated abstract type \( (\alpha) \). The syntax of 0-order types \((\sigma)\), 1st-order types \((\tau)\), values \( (v) \) and expressions \( (e) \) are provided below.

\[
\begin{align*}
(0\text{-types}) \quad \sigma & \coloneqq \beta | \alpha | (\sigma \circ \sigma) \\
(1\text{-types}) \quad \tau & \coloneqq \sigma | \sigma \to \tau | (\tau \circ \tau) \\
(values) \quad v & \coloneqq \mathsf{w} | \mathsf{v}_1, \mathsf{v}_2 | (\lambda x : \sigma. e) \\
(expressions) \quad e & \coloneqq x | v | (\pi_1 e) | (e_1 e_2)
\end{align*}
\]

We use \( x \) for value variables and \( w \) for constants of the base type \( \beta \). The expression \((\pi_1 e)\) is the \( 1 \)-th projection from the pair \( e \). We write \( \Gamma \vdash e : \tau \) to indicate that an expression \( e \) has type \( \tau \) in the context \( \Gamma \), which maps variables to their types. We write \( \vdash e : \tau \) when \( \Gamma \) is empty, as will be the case in most of this work. We write \( \Gamma \vdash \alpha \rightarrow \tau \) to substitute \( \tau \) for \( \alpha \) in \( \tau \). Finally, and we use \( \mathsf{e} \parallel v \) to indicate that \( e \) evaluates to \( v \). We refer the reader to Pierce [21] for the details.

We assume a module defines a single abstract type \((\alpha)\), which is declared in its interface. A module interface \((\mathsf{F} = \exists \alpha. \tau_m)\) is a pair of a name \((\alpha)\) for the abstract data type and a signature \(\tau_m\) that specifies the types for operations over the abstract type. A module implementation \(\mathsf{M} = \langle \tau_c, \nu_m \rangle\) is the classic existential package of a concrete type \(\tau_c\) and a value \(\nu_m\) containing operations over the type \(\tau_c\). We say a module \(\langle \tau_c, \nu_m \rangle\) implements an interface \(\exists \alpha. \tau_m\) when it is well-typed as per the usual rules for existential introduction [21, §24], i.e. \(\vdash \nu_m : \tau_m[\alpha \mapsto \tau_c] \).

In addition to an interface, we also assume the existence of a target specification \(\varphi\), which captures the desired correctness criteria for a module implementation. These specifications are universal properties of the values of the abstract type; we formalize them as polymorphic functions over the
module operations, \(\text{i.e.}, \varphi : \forall \alpha . (\tau_m \rightarrow \alpha \rightarrow \text{bool})\). We saw an example specification for integer sets in §2.

The values of an abstract type \(\alpha\) are simply the values that are constructible at type \(\alpha\) through the module interface. Below we define the notion of a \(\tau\)-constructible value and then use it to define when a module satisfies a specification.

**Definition 3.1** (\(\tau\)-Constructible Value: \(\mathcal{C}_{M}[v; \tau]\)). A value \(v\) is \(\tau\)-constructible using \(M\), denoted \(\mathcal{C}_{M}[v; \tau]\), iff there exists a function \(f : \forall \alpha . (\tau_m \rightarrow \tau)\) such that \((f[\tau_\alpha] \mid v_m) \downarrow v\).

**Definition 3.2** (Specification Satisfaction: \(\mathcal{M} : F \models \phi\)). A module \(M\) with interface \(F\) is said to satisfy a given specification \(\phi\), denoted \(\mathcal{M} : F \models \phi\), iff every \(\alpha\)-constructible value satisfies \(\phi\), i.e. \(\forall v : \tau_\alpha. \mathcal{C}_{M}[v; \alpha] \Rightarrow (\phi[v_\alpha] \mid v_m)\).

### 3.2 Representation Invariants

Loosely speaking, a representation invariant is a property that is preserved by operations over the abstract type of a module. As such, we say that a representation invariant is a fully inductive property of a module. The first part of Figure 3 defines a relation that we call conditional inductiveness, which is a generalization of both full inductiveness and the notion of visible inductiveness described earlier. Specifically, the relation \(v : \tau \uparrow^P \text{Valid}\) may be read as “value \(v\) is conditionally inductive at type \(\tau\) with respect to properties \(P\) and \(Q\)”.

**Full inductiveness.** When \(P\) and \(Q\) are the same property \(I\) (i.e., \(P = Q = I\)), these rules correspond to the standard logical relation over closed values for System F [26], but where there is exactly one free type variable \((\alpha)\) and that type variable is associated with the concrete type \(\tau_\alpha\) and the unary relation \(I\). Values of the abstract type \(\alpha\) are in the relation if they satisfy \(I\) (rule I-A). Products satisfy the relation if their components satisfy the relation (rule I-Prod). Functions satisfy the relation if they take arguments in the relation to results in the relation (rule I-Fun).

The following corollary of Reynolds’ theory of parametricity [22] says that if \(I\) is a representation invariant then all \(\alpha\)-constructible values satisfy it.

**Corollary 3.3.** \(v_m : \tau_m \uparrow^f \text{Valid} \Rightarrow (\forall v : \tau_\alpha. \mathcal{C}_{M}[v; \alpha] \Rightarrow (I \mid v))\)

Therefore, to prove that a module meets a specification it is enough to identify a sufficient representation invariant.

**Definition 3.4** (Sufficient Predicate: \(\text{Suf}_M^\phi[p]\)). A predicate \(p : (\tau_\alpha \rightarrow \text{bool})\) is sufficient for proving that \(M\) satisfies \(\phi\), denoted \(\text{Suf}_M^\phi[p]\), iff \(\forall v : \tau_\alpha.(p \mid v) \Rightarrow (\phi[v_\alpha] \mid v_m)\).

**Definition 3.5** (Sufficient Representation Invariant). A predicate \(I : (\tau_\alpha \rightarrow \text{bool})\) is called a sufficient representation invariant for a module \(M\) with respect to a specification \(\phi\), denoted \(\mathcal{M} : F \models_I \phi\), iff \(\text{Suf}_M^\phi[I] \land v_m : \tau_m \uparrow^f \text{Valid}\).

**Theorem 3.6.** If a sufficient representation invariant exists, then the module satisfies the specification, i.e.

\[
(\exists I : (\tau_\alpha \rightarrow \text{bool}) \cdot \mathcal{M} : F \models_I \phi) \\
\Rightarrow (\forall v : \tau_\alpha. \mathcal{C}_{M}[v; \alpha] \Rightarrow (\phi[v_\alpha] \mid v_m)\).
\]

**Proof.** Follows from Corollary 3.3, Definition 3.4, and Definition 3.5. \(\square\)

**Conditional inductiveness.** When \(P\) and \(Q\) are not the same, conditional inductiveness informally requires that if the client supplies values of abstract type satisfying \(P\) then the module will produce values of abstract type satisfying \(Q\). When conditional inductiveness is used to check visible inductiveness in our algorithm, \(P\) will be the set \(V_+\) of examples that are known to be \(\alpha\)-constructible by the module and \(Q\) will be a candidate representation invariant. The most interesting rule when \(P\) and \(Q\) are different is the I-Fun-CEx rule for functions. Specifically, notice the inversion of \(P\) and \(Q\) in the negative position: If the argument is a value of abstract type \(\alpha\) it must satisfy \(P\), not \(Q\). In other words, this element of the formalism codifies the intuition that if the client supplies values that satisfy \(P\) then the module will supply values that satisfy \(Q\).

**Counterexamples.** Normally logical relations are only used to prove that an invariant is inductive. However, we additionally require counterexamples from failed inductiveness checks, to drive our CEGIS-based invariant inference algorithm. The second section of Figure 3 provides the logic for refuting conditional inductiveness and generating counterexamples. This judgement has the form \(v : \tau \uparrow^P \text{CEx} \langle S, V \rangle\), which can be read as “value \(v\) is not conditionally inductive at type \(\tau\) with respect to properties \(P\) and \(Q\), with inductiveness counterexample witnesses \(S\) and \(V\)”.

Here the set \(S\) contains values that satisfy \(P\), the set \(V\) contains values that falsify \(Q\), and intuitively the values in \(V\) can be computed using module operations given inputs from \(S\).

As an example, consider the rule for values of abstract type (rule I-A-CEx). Here, a value \(v\) of type \(\alpha\) is not conditionally inductive if it falsifies \(Q\). The counterexample produced includes \(v\) in the set \(V\) (and returns the empty set \(S\)), and hence satisfies the judgemental invariant explained above. As another example, a function is not conditionally inductive (rule I-Fun-CEx) if there is an argument \(v_1\) in the relation that causes the function to produce a result \(v_2\) that is not in the relation. In that case, the function \(\| v \|_\alpha\) is used to collect all values of type \(\alpha\) in \(v_1\) to put in the returned set \(S\), since they are the inputs that led to the result \(v_2\).

The completeness of our algorithm for inferring sufficient representation invariants depends critically on these rules for generating counterexamples. In particular, values in \(S\) are added to the set \(V_+\) of negative examples in \(V\) are added
3.3 The Inference Algorithm

The invariant synthesis algorithm is parameterized by a verifier Verify and a synthesizer Synth. A call Verify returns Valid when \( P \in \) is true on all inputs of type \( r_c \). Otherwise, it returns a counterexample \( v \) to the predicate. A call Synth \( V_r V_v \) returns a predicate \( P \) that returns true on the positive examples \( V_p \) and false on the negative ones \( V_n \). \( V_r \) and \( V_v \) should not overlap; if they do then Synth will fail.

Figure 4 presents our invariant inference algorithm. To execute the algorithm, a user invokes Hanoi (line 32) with empty sets for \( V_r \) and \( V_v \) respectively. Hanoi first generates a candidate invariant \( I \) using Synth, given the current \( V_r \) and \( V_v \) sets. It then attempts to produce a candidate invariant that is visibly inductive relative to \( V_r \). That is the role of the call to ClosedPositives (line 36). That function simply calls CondInductive, which uses the inference rules in Figure 3. In the implementation these rules are executed through interaction with the verifier Verify.

Since everything in \( V_r \) is known to be constructible, the set \( V \) of values that violate the candidate invariant must also be constructible. Therefore, those values are returned from ClosedPositives, and they are added to \( V_r \) via a recursive call to Hanoi (line 37). This forces future candidate invariants produced by Synth to return true on elements in \( V \). Note that each time \( V_r \) is augmented, \( V_v \) is reset to the empty set, so the next synthesized invariant will be the constant true function, which is trivially visibly inductive. While we maintain the invariant that the positive examples are constructible and so must be included in the final invariant, negative examples are simply values that violate the current candidate invariant (but may in fact be constructible).

Once the candidate invariant \( I \) is visibly inductive with respect to \( V_r \), Hanoi checks for sufficiency and full inductiveness by calling NoNegatives at line 39. The NoNegatives procedure interacts with Verify to check sufficiency and calls CondInductive to check full inductiveness. If either of these checks fail, NoNegatives will return counterexample values that satisfy the current invariant — either a sufficiency
violation or the set \( S \) from an inductiveness counterexample. Because \( I \) is visibly inductive, \( N \setminus V_i \) can only be empty if there is a sufficiency violation. In that case, we have found a constructible violation of the specification \( \phi \), so Hanoi terminates and provides this counterexample. If \( N \setminus V_i \) is non-empty, Hanoi adds all of these values to \( V_i \), and Synth will generate a stronger candidate invariant in the next iteration. If a constructible counterexample is added to \( V_i \), it will eventually be generated by \text{ClosedPositives} and moved to \( V_i \). If both checks in \text{NoNegatives} succeed, then we have found a sufficient representation invariant and it is returned.

### 3.4 Soundness and Completeness

We say that Verify is sound if \((\text{Verify } P) = \text{Valid} \) implies \( \forall v : \tau_c. ( (p v) \downarrow \text{true} ) \). Further, Verify is said to be complete if \((\text{Verify } (p v)) \downarrow \text{false} \). Likewise, we say that Synth is sound if for all sets \( V_i \) and \( V_i \) of \( \tau_c \) values, \((\text{Synth } V_i \downarrow V_i) \) implies \( \forall v \in V_i. ( (p v) \downarrow \text{true} \) and \( \forall v \in V_i. ( (p v) \downarrow \text{false} \). Further, Synth is said to be complete if for all sets \( V_i \) and \( V_i \) of \( \tau_c \) values, whenever there exists a predicate \( P : \tau_c \rightarrow \text{bool} \) such that \( \forall v^+ \in V_i. ( (p v^+) \downarrow \text{true} \) and \( \forall v^- \in V_i. ( (p v^-) \downarrow \text{false} \), Synth always returns some predicate \( P' \).

**Definition 3.7 (Soundness).** An inference system for representation invariants is said to be sound iff whenever the system generates a predicate \( I \), it is indeed a sufficient representation invariant, i.e. \( M : F \models_I \phi \).

**Definition 3.8 (Completeness).** An inference system for representation invariants is said to be complete iff whenever there exists a sufficient representation invariant \( I \) such that \( M : F \models_I \phi \), the system always generates (terminates with) some predicate \( I : (\tau_c \rightarrow \text{bool}) \).

**Theorem 3.9.** If Verify is sound, then Hanoi \( \emptyset \emptyset \) is sound.

**Theorem 3.10.** If Verify and Synth are both sound and complete, and \( \tau_c \) is a finite domain, then Hanoi \( \emptyset \emptyset \) is complete.

The proofs can be found in the appendix of the full version of this paper [? ]. The soundness of Hanoi is straightforward and follows from the fact that an invariant is only returned if it is both sufficient and inductive. The completeness argument for finite domains is much more involved. As mentioned earlier, it depends on several properties of the rules for generating counterexamples in Figure 3. Further, we must prove that the Hanoi algorithm always terminates. Notice that the size of the set \( V_i \) monotonically increases during the algorithm. While \( V_i \) is reset to empty on some recursive calls, this is only done when \( V_i \) is augmented. Hence the following is a rank function that is bounded from below and decreases lexicographically with each recursive Hanoi call, where \([\tau_c]\) denotes the number of values of type \( \tau_c \):

\[
\mathcal{R}(V_i, V_i) = \left( [\tau_c] - |V_i|, [\tau_c] - |V_i| \right)
\]

### 4 Implementation

This section describes a variety of additional aspects of our ~5 KLOC OCaml implementation of Hanoi.

#### 4.1 The Programming Language

We have implemented a pure, simply-typed, call-by-value functional language with recursive data types. Numbers are
implemented as a recursive data type, where a number is either 0 or the successor of a number. Each program includes a prelude that may contain data type declarations and functions over those data types. A program also contains a single module declaring an abstract type together with operations over that abstract type. Finally, a program includes a universally quantified specification that defines the intended behavior of the module in terms of its operations.

### 4.2 Tackling Higher-Order Functions

While the theory presented in the previous section only supports first-order terms, our implementation allows modules to include arbitrary higher-order functions. As mentioned earlier, the key extension required is the ability to extract counterexample values of the abstract type from functions. Here we discuss how our implementation does that.

First, consider a natural extension to the SET interface from §2 to include a map function.

```plaintext
module type HOSET = sig
end

val fold : (int -> t -> t) -> t -> t -> t
let rec fold f a s =
| hd :: tl -> f hd (fold f a tl)

include SET
include HOSET
val map : (int -> int) -> t -> t
| [] -> a
```

Notice that while map is a higher-order function, the type of the higher-order argument does not involve any occurrences of the abstract type \( t \). The same is true of iter, zip, and many other variants. Consequently, if, during invariant inference, \( (\text{map } f \, v) \) fails an inductiveness check on some candidate representation invariant, the counterexample values that represent the “reason” for this failure will never come from \( f \). More generally, when the abstract type does not appear in a higher type \( \tau \), the value with type \( \tau \) cannot contain counterexamples. Our implementation therefore simply ignores such higher-order values when extracting counterexamples, just as it ignores ordinary base types such as \( \text{int} \).

Now consider a further extension that includes a fold.

```plaintext
module type FSET = sig
end

val fold : (int -> t -> t) -> t -> t -> t

let rec fold f a s =
| [] -> a
| hd :: tl -> f hd (fold f a tl)
```

Here, \( \text{fold} \) contains a function argument with a type including \( t \). The fold might be implemented as follows.

```plaintext
let rec fold f a s =
| [] -> a
| hd :: tl -> f hd (fold f a tl)
```

Specifically, the implementation supplies a value to the client when it calls a function argument, and the client supplies a value to the module when it returns from such a function. Fortunately, a mechanism already exists for tracking such boundary crossings in the general case: the higher-order contracts of Findler and Felleisen [9].

Therefore, our implementation extracts counterexamples through higher-order contract checking. The first-order case is straightforward. For example, when the type is \( t \to t \), we generate a contract \( P \to Q \) to check that arguments satisfy \( P \) and results satisfy \( Q \), and we log situations where \( P \) is satisfied by an argument but \( Q \) is violated by the result. This is a direct implementation of the rule \( \text{I-Fun-CEx} \) in Figure 3.

For a type such as \( (\text{int} \to \text{t} \to \text{t}) \to \text{t} \to \text{t} \to \text{t} \), we simply extend the idea, giving rise to the following contract.

\[
(\text{any_int} \to Q \to P) \to P \to P \to Q
\]

As per usual, all negative positions must satisfy \( P \) and the positive ones \( Q \). Then contract checking is used to identify runs that satisfy all of the \( P \) checks but fail a \( Q \) check. In that case, if \( S \) is the set of values that satisfy \( P \) and \( v \) is the value that violates \( Q \), then the extracted inductiveness counterexample is \( (S, \{v\}) \).

With this extension, \( \text{Hanoi} \) is trivially sound, for the same reason that the first-order algorithm is sound (the algorithm checks for soundness just before termination). We conjecture that \( \text{Hanoi} \) with this extension is also complete for finite domains but have not proven it. However, in the next section, we demonstrate empirically that our implementation can infer representation invariants in the presence of higher-order functions.

### 4.3 Verifier and Synthesizer

To implement \textbf{Verify}, we use a size-bounded enumerative tester, which is unsound but effective in practice. To validate a predicate with a single quantifier, we test the predicate on data structures, from smallest to largest, until either 3000 data structures have been processed, or the data structure has over 30 AST nodes, whichever comes first. To validate predicates with two or more quantifiers, we instantiate each quantifier with the smallest 3000 data structures with under 15 AST nodes. We further limit the total number of data structures processed to 30000. These restrictions limit the total amount of time spent in the verifier at the cost of soundness guarantees.

To implement \textbf{Synth}, we use \textsc{Myth} [19], adapting it slightly in two ways. First, we modified it to return a set of candidate invariants, instead of just one. Doing so permits the caching of synthesis results described earlier. Second, we had to manage \textsc{Myth}’s requirement for \textit{trace completeness}. Trace completeness requires that whenever we provide an input-output example \((x, y)\) for a recursive data type, we also provide input-output examples for each subvalue of \( x \). We generate input-output pairs for \textsc{Myth} by pairing each
To accelerate invariant inference, we have implemented two

To handle trace completeness, we first identify all subvalues

\( \lambda \)
sizer proposes

called with

\( v \)

noi

shown in Figure 5(a). In this example,

Consider the trace of

invariant is used instead of a freshly synthesized one.

ants satisfy the input/output example set. If one does, that

solutions away, we store them for future synthesis calls. When

given a set of input/output examples, before making a call to

MYTH, we check if any of the previously synthesized invariants satisfy the input/output example set. If one does, that

invariant is used instead of a freshly synthesized one.

\( \lambda x. true \)
as a candidate invariant and then

verification subsequently provides the negative counterexample, \( v_2 \), which then becomes the only negative example in the next attempt at synthesis. This loop of proposing new invariants, and adding their negative counterexamples to the negative example set continues until \( \lambda x. e_3 \) is proposed, which provides the positive counterexample, \( v_5 \).

Next, according to the unoptimized algorithm, one should begin a run with \( \{ v_1, v_2 \} \) as positive examples and no negative examples—see Figure 5(b) for a partial trace of this subsequent execution. Suppose that \( v_5 \) satisfies the first two synthesized invariants from the original run. In that case, those invariants will simply be synthesized again as the first two candidates of this new run, as shown in the figure. To

avoid this recomputation, we cache these traces of synthesis and verification calls. When we receive a new positive example, we run it on the synthesized invariants in the trace, as shown in Figure 6. Because \( \lambda x. true \) and \( \lambda x. e_1 \) both return \( true \) on \( v_5 \), we can skip the entirety of the trace shown in Figure 5(b) and begin by synthesizing from the testbed \( V_4 = \{ v_1, v_5 \} \) and \( V_5 = \{ v_2, v_3 \} \).

5 Experimental Results

We aim to answer the following research questions:

1. Can we infer sufficient representation invariants in practice?
2. What are the primary performance factors?
3. What effect on performance do our optimizations have?
4. How does our algorithm compare with prior work?

5.1 Benchmark Suite

We evaluate Hanoi on a total of 28 verification problems, most of which require reasoning over list or tree structures. We categorize them into the following four groups.

- **VFA** (5): Four modules from Verified Functional Algorithms (VFA) [1] that have interfaces and specifications over those interfaces, including tree- and list-based implementations of lookup tables and priority queues. We also experimented with a second version of priority queues that excludes the merge function.
- **VFAExt** (3): Three VFA modules with additional function(s) and corresponding specifications from the Coq [29] standard library.
- **Coq** (14): Five tree- and list-based implementations of data structures from the Coq [29] standard library. One additional problem for each of the five by introducing
additional binary functions. Four more problems by extending interfaces with higher-order functions.

- **OTHER (6):** Six additional benchmarks of our own creation requiring reasoning over lists, natural numbers, monads and other basic data structures.

### 5.2 Experimental Setup

All experiments were performed on a 2.5 GHz Intel Core i7 processor with 16 GB of 1600 MHz DDR3 RAM running macOS Mojave. We ran each benchmark 10 times with a timeout of 30 minutes and report the average time. If any of the 10 runs time out then we consider the benchmark as a whole to have timed out.

### 5.3 Inferred Invariants

Figure 7 presents our results. Overall, Hanoi terminated with an invariant on 22 out of 28 benchmarks within the timeout bound. The second column shows the sizes of the inferred invariants, in terms of their abstract syntax trees.

Though our verifier is unsound, there was no effect on the reliability of the system on our benchmark suite: 22 of the 22 inferred invariants are correct. Further, some of them are quite sophisticated. For example, we synthesize a heap invariant, which requires that the the elements of each node’s subtrees is smaller than that node’s label. We synthesize invariants over lists including “max element first,” “no duplicates,” and “ordered.” If we allow the system to exceed the 30 min threshold, the system will infer a binary search tree invariant as well.

In seven of the cases above, we run into a limitation of the Myth synthesizer rather than our algorithm: Myth cannot synthesize functions that require recursive “helper” functions. To bypass this restriction, we added a true_maximum function (that finds the maximum element of a tree) to our tree-heaps benchmark and a min_max_tree function (that finds the minimum and maximum elements of trees) to our bst and red-black-tree benchmarks. We added a * next to the names of benchmarks that we altered by providing a helper function in this way (see Figure 7).

### 5.4 Primary Performance Factors

When benchmarks complete within the 30 minute bound, most of the time is spent in verification. Indeed, for all but two of the terminating benchmarks, the total time spent synthesizing is under two seconds.

Three factors affect verification times significantly: (1) the strength of the specification, (2) the complexity of the underlying data structure, and (3) the presence of higher order functions. First, it takes our verifier longer to validate a true fact than to find a counterexample to a false one (validation requires enumeration of all tests; in contrast, the moment a counterexample is found, the enumeration is short-circuited). Many candidate invariants imply weak specifications but are not inductive. Hence weak specifications, ironically, are quite costly, because many candidate invariants wind up being sufficient (incuring a significant verification expense each time), only to be thrown away later when it turns out they are not inductive. Second, relatively simple data structures, like natural numbers and lists with numbers as their elements, take less time to verify than more complex data structures, like trees, tries, and lists with more complex elements. Third, like other complicated data types, the use of higher-order functions increases verification time. There are many ways to build a function, so enumeratively verifying a higher-order function requires searching through many possible functions.

However, the story is different for the complex benchmarks that do not complete within 30 minutes. When we ran Hanoi on our bst set benchmark, it completed in 78.4 minutes. Unlike the prior benchmarks, the majority of the time (65%) was spent in synthesis. Moreover, 30% of the total time was spent on the synthesis call that generated the final invariant. This indicates that Hanoi is currently not gated by the verifier, but by the synthesizer. Indeed, the implementation of bst set that includes binary functions like union and intersection is actually much faster than that without union and intersection, terminating within our 30 minute timeout. Adding these functions makes the verification harder, but Myth can use them to generate simpler invariants. Adding helper functions that permit simpler invariants also makes our implementation of a bst table verifiable in under 30 minutes.

Due to these limitations, we believe that a smarter synthesizer would be able to find more invariants. To this end, we built a prototype synthesizer that can generate more complex types of functions. This synthesizer has similarities to Myth as it is type-and-example directed and enumerative. However, where Myth can only synthesize simple recursive functions, this alternate synthesizer can synthesize folds, letting our synthesizer generate functions that require accumulators. Our synthesizer performs comparably to Myth, synthesizing invariants for the 20 benchmarks Myth that can solve an average of 11% slower. However, our synthesizer is also able to find the invariant for a binary heap (/vfa/tree-:::pqueuee) without requiring helper functions or functions defined in the module in 185.4 seconds (55.8 seconds for /vfa/tree-:::pqueuee:::binfoode), while Myth fails.

### 5.5 Comparisons

Figure 8 summaries the results of running of Hanoi, Hainol without optimizations, and our implementations of prior approaches adapted to our setting.

**Impact of Optimizations.** The modes Hanoi_SRC and Hanoi_CLC tested the impacts of our optimizations described in §4.4. Hanoi_SRC runs the benchmarks with synthesis result...
caching turned off. **Hanoi**-**CLC** runs the benchmarks with counterexample list caching turned off.

Removing synthesis result caching does not have a large impact on the majority of benchmarks as the majority of our benchmarks spend relatively little time in synthesis. However, more complex benchmarks are able to enjoy the benefits of this optimization.

Counterexample list caching has significant impact on complex benchmarks as they have more synthesis and verification calls. The synthesizer requires more input/output examples to synthesize the correct invariant on complex benchmarks, so saving time reconstructing the negative examples via counterexample list caching has great impact.

**Comparison to AStr.** The AStr mode simulates the LOOP-InvGen algorithm [20], a related data-driven system for inferring loop invariants. When running AStr, if a candidate invariant $I_1$ is sufficient to prove the specification, but is not inductive, the algorithm attempts to synthesize a new predicate $I_2$ such that $\forall m : \tau_m \rightarrow I_1 \land I_2 \text{ valid.}$ In that case, $I_1 \land I_2$ is considered the new candidate invariant. This process continues until either the conjoined invariants are inductive, or they are overly strong so a new positive counterexample is found, at which point the whole process restarts. **Hanoi** outperforms AStr on all the benchmarks and solves 3 more benchmarks within 30 minutes. The main downside of AStr is that it can only add new positive examples in order to weaken the candidate invariant after it has obviously over-strengthened. **Hanoi**, in contrast, uses visible inductiveness checks to eagerly weaken in a directed manner.

**Comparison to LA.** LA mode simulates the **LINEARARBITRARY algorithm** [30], which is used in a data-driven CHC solver. There are two differences from **Hanoi**. First, LA tries to satisfy individual inductiveness constraints, generated for each function in the module, one at a time rather than all at once. Second, rather than eagerly searching for visible inductiveness violations, only full inductiveness counterexamples

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Time (s)</th>
<th>TVT (s)</th>
<th>TVC</th>
<th>MVT (s)</th>
<th>TST (s)</th>
<th>TSC</th>
<th>MST (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/coq/bst::&lt;set&gt;</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>97</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
</tr>
<tr>
<td>/coq/bst::&lt;set&gt;+binfuncs</td>
<td>15</td>
<td>42.0</td>
<td>41.8</td>
<td>11</td>
<td>3.80</td>
<td>0.2</td>
<td>3</td>
<td>0.07</td>
</tr>
<tr>
<td>/coq/rbtree::&lt;set&gt;</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>97</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
</tr>
<tr>
<td>/coq/rbtree::&lt;set&gt;+hofs*</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>103</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
</tr>
<tr>
<td>/coq/maxfirst-list::&lt;heap&gt;</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>103</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
</tr>
<tr>
<td>/coq/maxfirst-list::&lt;heap&gt;+binfuncs</td>
<td>35</td>
<td>6.2</td>
<td>5.8</td>
<td>16</td>
<td>0.36</td>
<td>0.3</td>
<td>5</td>
<td>0.06</td>
</tr>
<tr>
<td>/coq/sorted-list::&lt;set&gt;</td>
<td>49</td>
<td>22.9</td>
<td>22.0</td>
<td>21</td>
<td>1.05</td>
<td>0.7</td>
<td>7</td>
<td>0.10</td>
</tr>
<tr>
<td>/coq/sorted-list::&lt;set&gt;+binfuncs</td>
<td>49</td>
<td>17.3</td>
<td>16.0</td>
<td>21</td>
<td>0.76</td>
<td>1.2</td>
<td>8</td>
<td>0.15</td>
</tr>
<tr>
<td>/coq/sorted-list::&lt;set&gt;+hofs</td>
<td>49</td>
<td>101.3</td>
<td>99.6</td>
<td>21</td>
<td>4.74</td>
<td>1.6</td>
<td>8</td>
<td>0.20</td>
</tr>
<tr>
<td>/coq/unique-list::&lt;set&gt;</td>
<td>35</td>
<td>13.2</td>
<td>12.9</td>
<td>20</td>
<td>0.65</td>
<td>0.3</td>
<td>6</td>
<td>0.05</td>
</tr>
<tr>
<td>/coq/unique-list::&lt;set&gt;+binfuncs</td>
<td>15</td>
<td>15.7</td>
<td>15.6</td>
<td>8</td>
<td>1.95</td>
<td>0.1</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>/coq/unique-list::&lt;set&gt;+hofs</td>
<td>17</td>
<td>81.7</td>
<td>81.4</td>
<td>8</td>
<td>10.18</td>
<td>0.2</td>
<td>2</td>
<td>0.10</td>
</tr>
<tr>
<td>/other/cache</td>
<td>29</td>
<td>1.3</td>
<td>1.0</td>
<td>16</td>
<td>0.06</td>
<td>0.2</td>
<td>5</td>
<td>0.04</td>
</tr>
<tr>
<td>/other/listlike-tree</td>
<td>53</td>
<td>9.0</td>
<td>8.8</td>
<td>14</td>
<td>0.63</td>
<td>0.1</td>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>/other/nat-nat-option::&lt;range&gt;</td>
<td>23</td>
<td>1.6</td>
<td>1.5</td>
<td>12</td>
<td>0.12</td>
<td>0.1</td>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>/other/rational</td>
<td>28</td>
<td>8.6</td>
<td>7.8</td>
<td>8</td>
<td>0.97</td>
<td>0.7</td>
<td>2</td>
<td>0.35</td>
</tr>
<tr>
<td>/other/sized-list</td>
<td>45</td>
<td>15.4</td>
<td>15.0</td>
<td>20</td>
<td>0.75</td>
<td>0.3</td>
<td>6</td>
<td>0.05</td>
</tr>
<tr>
<td>/other/stutter-list</td>
<td>49</td>
<td>6.9</td>
<td>5.6</td>
<td>20</td>
<td>0.28</td>
<td>1.1</td>
<td>7</td>
<td>0.16</td>
</tr>
<tr>
<td>/vfa-extended/assoc-list::&lt;table&gt;</td>
<td>4</td>
<td>2.6</td>
<td>2.5</td>
<td>3</td>
<td>0.83</td>
<td>0.0</td>
<td>0</td>
<td>undef</td>
</tr>
<tr>
<td>/vfa-extended/bst::&lt;table&gt;</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>103</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
<td>t/o</td>
</tr>
<tr>
<td>/vfa-extended/trie::&lt;table&gt;</td>
<td>4</td>
<td>15.5</td>
<td>15.4</td>
<td>3</td>
<td>5.13</td>
<td>0.0</td>
<td>undef</td>
<td></td>
</tr>
<tr>
<td>/vfa/assoc-list::&lt;table&gt;</td>
<td>4</td>
<td>1.9</td>
<td>1.9</td>
<td>3</td>
<td>0.63</td>
<td>0.0</td>
<td>undef</td>
<td></td>
</tr>
<tr>
<td>/vfa/bst::&lt;table&gt;</td>
<td>4</td>
<td>12.9</td>
<td>12.8</td>
<td>3</td>
<td>4.27</td>
<td>0.0</td>
<td>undef</td>
<td></td>
</tr>
<tr>
<td>/vfa/trie::&lt;priqueue*&gt;</td>
<td>47</td>
<td>65.7</td>
<td>53.6</td>
<td>54</td>
<td>0.99</td>
<td>9.6</td>
<td>15</td>
<td>0.64</td>
</tr>
<tr>
<td>/vfa/tree::&lt;priqueue&gt;+binfuncs*</td>
<td>47</td>
<td>79.4</td>
<td>64.5</td>
<td>54</td>
<td>1.19</td>
<td>12.4</td>
<td>15</td>
<td>0.83</td>
</tr>
<tr>
<td>/vfa/trie::&lt;table&gt;</td>
<td>4</td>
<td>17.7</td>
<td>17.6</td>
<td>3</td>
<td>5.87</td>
<td>0.0</td>
<td>0</td>
<td>undef</td>
</tr>
</tbody>
</table>

Figure 7. Information from running Hanoi on our benchmark suite. **Name** is the name of the benchmark. **Size** is the size of the inferred invariant. **Time** is the time to run the benchmark from start to end. **TVT** is the total time spent verifying. **TVC** is the total number of verification calls. **MVT** is the average time per single verification call. **TST** is the total time spent synthesizing. **TSC** is the total number of synthesis calls. **MST** is the average time for a single synthesis call. Benchmarks marked with a * were provided an additional function to enable synthesis by **Myth**.
are obtained. However, if a full inductiveness counterexample happens to also be a visible inductiveness counterexample then it is treated accordingly. Hanoi outperforms LA on all the benchmarks and is able to solve 4 more benchmarks within 30 minutes. While Hanoi checks eagerly for positive counterexamples, LA finds them nondeterministically. Without performing the guided search through visible inductiveness checks, the algorithm sometimes gets “stuck” in holes of negative counterexamples. While the algorithm does seem to emerge from these holes eventually, it takes time.

Comparison to OneShot. The OneShot mode uses “one shot learning” rather than an interactive CEGIS algorithm. The OneShot algorithm runs the specification over the smallest 30 elements of the concrete implementation type, tagging each element as either positive or negative. Doing so generates sets $V_1$ and $V_2$, which may be supplied to the synthesizer. Whatever invariant synthesized is returned as the result. (This algorithm only works when the specification quantifies over a single element of the abstract type, which is true for all but 7 of our benchmarks.) Running the OneShot algorithm fails on all but one of our benchmarks, coq/unique-list-set, and does so for a variety of reasons. On some benchmarks, Myth times out, indicating that the given synthesis problem was too hard, and Myth needed to be provided fewer examples to find the right invariant. On some benchmarks, Myth returns the wrong invariant, indicating that the synthesis problem was underspecified, and too few examples were given. Merely choosing some fixed number of examples to build the invariant with is insufficient, that fixed number is too high for some benchmarks, and too low for others.

6 Related Work

Inferring Representation Invariants. To our knowledge, the only prior work that attempts to automatically infer representation invariants for data structures is the Deryaft system by Malik et al. [15], which targets Java classes. There are three key differences between systems. First, Deryaft requires a fixed set of predicates (e.g., sortedness) as an input; the invariants generated are conjunctions of these predicates. In contrast, Hanoi1 can learn new predicates from a general grammar of programs. Second, the conjunction of predicates that Deryaft produces consists of those predicates that hold on a fixed set of examples (generated from test executions). There is no guarantee the final invariant is inductive. In contrast, Hanoi1 can learn new predicates from a general grammar of programs. Second, the conjunction of unknown predicates in the final invariant can be and do contain unknown predicates, whereas there is only one in ours.

Solving constrained Horn clauses. Recently, several tools have been developed to infer predicates that satisfy a given set of constrained Horn clauses (CHCs). Inferring representation invariants can be seen as a special case of CHC solving, since all of our inductiveness constraints are Horn clauses (e.g., $(I_1 s_1) \land (I_2 s_2) \Rightarrow (I_2 \, \text{union} \, s_1 \, s_2)$). CHCs can include multiple unknown predicates in their inference problem, whereas there is only one in ours.

However, existing CHC solvers do not support inductive predicates, which is necessary to handle representation invariants over recursive data types. Several solvers support only arithmetic constraints [7, 8, 30], while others support arrays or bit vectors [5, 11, 13] as well. To our knowledge, Eldarica [11] is the only CHC solver that supports algebraic data types. However, Eldarica only computes recursion-free solutions [11, Sec III C] and therefore cannot express the sortedness or no-duplicates properties, for instance.

Though they only handle arithmetic constraints, two of these solvers employ a data-driven technique that is similar to our approach, iterating between a synthesizer and a verifier [7, 30]. Like our work, the approach of Zhu et al. [30] leverages the observation that handling inductiveness counterexamples $(x, y)$ is easy when we know that $x$ is constructible. However, their approach simply checks if a counterexample to full inductiveness happens to have this property, while we exhaustively iterate through these counterexamples until a candidate invariant is visibly inductive. Intuitively, our approach minimizes the number of inductiveness counterexamples that must be treated heuristically.
We show that this difference results in a significant performance improvement, and we have proven a completeness result for finite domains, while the approach of Zhu et al. [30] lacks a completeness result. The approach of Ezudheen et al. [7] does have a completeness result, and it applies to the infinite domain of integers. However, they achieve this guarantee through the use of a specialized synthesizer designed to handle inductiveness counterexamples directly, while our approach can use any off-the-shelf synthesizer. There is also no obvious analogue to our analysis of higher-order programs in this context.

**Inferring inductive invariants.** There have been many techniques developed to infer individual inductive invariants for program verification, for example an inductive invariant for a loop or for a system transition relation. As discussed in §2.2, module functions may consume or produce multiple arguments or results of the abstract type, which results in a more general class of inductiveness counterexamples, whereas loops and transition relations, when viewed as functions, consume and produce exactly one “state” (the analogue of an abstract value in our setting).

**Hanoi** is similar in structure to several data-driven invariant synthesis engines [2, 7, 8, 10, 14, 17, 18, 20, 24, 25]. We experimentally compared our algorithm to our implementations of the most closely related ones, adapted to the context of representation invariant synthesis (§5.5). Broadly, the technical distinctions are similar to those described above for data-driven CHC solvers. In particular: (1) our development of visible inductiveness is novel; (2) aside from one tool [10] that depends upon a special-purpose synthesizer to handles inductiveness counterexamples directly, none are proven complete; (3) they cannot process higher-order programs; and (4) to our knowledge, none infer recursive invariants.

The ic3 algorithm for SAT-based model checking employs a notion of relative inductiveness [4], which is closely related to our notions of conditional and visible inductiveness. Formally, relative inductiveness is the special case of our conditional inductiveness relation $\nu : \tau \gg P \tau$ where $P$ has the form $P' \land Q$ and $\tau = \alpha \rightarrow \alpha$. The ic3 algorithm uses relative inductiveness to incrementally produce an inductive invariant, by iteratively identifying a state $s$ that leads to a property violation and then conjoining an inductive strengthening of $\neg s$ to the candidate invariant. Our notion of visible inductiveness is also used to incrementally produce an inductive invariant, but it works in the opposite direction: we iteratively identify constructible values of the abstract type and use them to weaken the candidate invariant. This approach is a natural fit for our data-driven setting.

**Automatic data structure verification.** The Leon framework [28] can automatically verify correctness of data structure implementations, but to do so, a user must manually define an abstraction function, which plays a similar role to a representation invariant. Namely, the abstraction function is a partial function mapping an element of the concrete type to an element of the abstract type.

There are many techniques for proving properties of heap-based data structures, including shape analysis [23] and liquid types [12]. These techniques can prove and/or infer sophisticated invariants, often of imperative code. However, they are designed to tackle a different problem and do not infer the inductive representation invariants that are needed to prove modules correct.

### 7 Conclusion

We present a novel algorithm for synthesis of representation invariants. Our key insight is that it is possible to drive progress of the algorithm towards its goal not by eagerly searching for fully inductive invariants, but rather by searching first for visibly inductive invariants. We have proven that our algorithm is sound and complete, given a sound and complete verifier and synthesizer, for a first-order type theory with finite types. We also explain how to extend the algorithm to modules containing higher-order functions, which involves using contracts to validate and collect objects that cross the module boundary. We evaluate our algorithm on 28 benchmarks and find that we are able to synthesize 22 of the invariants within 30 minutes (and most of those in under a minute). Our algorithm is defined independently of the black-box verifier and synthesizers; as research in verifier and synthesizer technologies improve, so too will the capabilities of our overall system. While the tool is currently fully automated, we view this as a step towards an interactive approach to helping users of proof assistants produce correct representation invariants.

### Acknowledgments

Thanks to our shepherd James Bornholt and the other reviewers for their helpful comments. This work is supported in part by NSF grant CCF-1837129.

### References


