Learning Formulas in Finite Variable Logics

PAUL KROGMEIER, University of Illinois, Urbana-Champaign, USA
P. MADHUSUDAN, University of Illinois, Urbana-Champaign, USA

We consider grammar-restricted exact learning of formulas and terms in finite variable logics. We propose a novel and versatile automata-theoretic technique for solving such problems. We first show results for learning formulas that classify a set of positively- and negatively-labeled structures. We give algorithms for realizability and synthesis of such formulas along with upper and lower bounds. We also establish positive results using our technique for other logics and variants of the learning problem, including first-order logic with least fixed point definitions, higher-order logics, and synthesis of queries and terms with recursively-defined functions.

CCS Concepts: • Theory of computation → Tree languages; • Computing methodologies → Machine learning approaches.

Additional Key Words and Phrases: exact learning, learning formulas, tree automata, version space algebra, program synthesis, interpretable learning

ACM Reference Format:
Paul Krogmeier and P. Madhusudan. 2022. Learning Formulas in Finite Variable Logics. Proc. ACM Program. Lang. 6, POPL, Article 10 (January 2022), 39 pages. https://doi.org/10.1145/3498671

1 INTRODUCTION

Learning symbolically representable concepts from data is an important emerging area of research. Symbolic expressions, such as logical formulas or programs, can be easily analyzed and interpreted, which aids downstream applications (e.g., analyzing a large system that has a classifier as a component) and makes them easier to communicate to both humans and computers.

In this paper, we embark on a foundational study of exact learning of logical formulas. For a logic $L$, we study the separability problem: given a set of positively- and negatively-labeled finite structures, we want to learn a sentence $\varphi$ in $L$ that is true on the positive structures and false on the negative structures. Separability consists of two related problems. First, the realizability problem: a decision problem that asks whether such a sentence exists, and second, the synthesis problem, which asks to construct a sentence if one exists. For logics that contain infinitely many semantically inequivalent formulas, including most of the logics considered in this paper, the realizability problem itself is not trivial.

In a learning context, one is often interested in how well a learned artifact generalizes to unseen inputs. In practice, most learning algorithms typically attempt only to minimize loss in accuracy on a set of training samples [Mitchell 1997]. Exact learning asks for a perfect classifier with respect to the training samples. Two common strategies to mitigate overfitting are (1) to only consider classifiers from a restricted hypothesis class $\mathcal{H}$ and (2) to prefer simple concepts over complex ones.
The problems we study here reflect (1) by considering exact learning with respect to *grammars*. Problem instances are equipped with a grammar $G$ that defines a subset $L(G)$ of logical expressions in $L$ to which classifiers must belong. The problems reflect (2) by requiring a synthesizer to construct small (perhaps the smallest) formulas that separate sample structures.

We describe a very general technique for solving the realizability and synthesis problems for *several* logics with finitely many variables. In particular, a main contribution of our work is to solve realizability and synthesis for $\text{FO}(k)$, a version of first-order logic with $k$ variables. This logic allows for an arbitrary number and nesting depth of quantifiers. Although the number of variables is bounded, it is possible to *reuse* variables. For instance, consider finite graphs. Given two constants $s$ and $t$ and any $n \in \mathbb{N}$, the property that $t$ is reachable from $s$ using at most $n$ edges is expressible using just two variables, and thus $\text{FO}(k)$ with $k = 2$ contains an infinite set of inequivalent formulas.

We prove that for every $k \in \mathbb{N}$, the realizability problem for $\text{FO}(k)$ over a grammar is decidable. That is, given a (tree) grammar $G$ defining a subclass of $\text{FO}(k)$ and sets of positively- and negatively-labeled structures $\text{Pos}$ and $\text{Neg}$, it is decidable to check whether there is a sentence $\varphi \in L(G)$ that is true on all structures in $\text{Pos}$ and false on all structures in $\text{Neg}$. We give an algorithm to synthesize such a sentence if one exists. Notice that since structures are finite, the sentence can also be converted to a program operating over structures that realizes the classifier effectively.

**Automata over Parse Trees for Realizability and Synthesis.** Our primary technique for solving exact learning problems of this kind is based on automata over finite trees. Intuitively, given finite (disjoint) sets of positive and negative structures, we need to search through an infinite set of formulas that adhere to the grammar $G$ in order to find a separating formula. We use tree automata working over *formula parse trees* and show that the set of all separating formulas for $\text{Pos}$ and $\text{Neg}$ that adhere to $G$ forms a regular class of trees. Building the tree automaton and checking emptiness gives us a decision procedure for realizability. Algorithms for tree automaton emptiness are used to solve the synthesis problem. Furthermore, the algorithms can be adapted to find the smallest trees that are accepted by the tree automaton, hence giving us small formulas as separators.

The key idea is to show that, given a single structure $A$, the set of parse trees for all sentences adhering to the grammar $G$ that are true (or false) in $A$ is a regular language. Given $A$, we can define an automaton that interprets an input formula (rather, its parse tree) on $A$ and checks whether $A$ satisfies the formula. This evaluation follows the usual semantics of the logic in question, which is typically defined *recursively*, and hence can be evaluated bottom-up in the structure of the formula. In general, this requires simulating the semantics of formulas for each assignment to free variables over the structure $A$. The bounded variable restriction is therefore crucial to ensure that the automaton only needs an amount of state that depends on the size of the structure but is *independent* of the size of the formula.

We can then construct an automaton that captures precisely the set of separating formulas for the given labeled structures. We do this by constructing automata for (a) the set of all formulas that are true on positive structures, (b) the set of all formulas that are false on negative structures, and (c) the intersection of the automata from (a) and (b). We further intersect (c) with an automaton that accepts formulas allowed by the grammar $G$. Checking emptiness of this final automaton solves realizability, and we can construct an accepted formula if nonempty.

**Query and Term Synthesis.** We study two related learning problems in addition to the separability problem. We study *query* synthesis, where we are given a grammar $G$ and a finite set of structures, each accompanied by an *answer set* of $r$-tuples from the domain of the structure. The query synthesis problem is to find a query, namely, a first-order formula $\varphi \in L(G)$ with $r$ free variables,
such that the sets of tuples that satisfy $\varphi$ in each structure are precisely the given answer sets. We also study the problem of term synthesis, which is closer in spirit to program synthesis from input-output examples (using logic as a programming language). In this problem we are given a set of input structures and a grammar $G$, and each structure interprets a set of constants, e.g., $in_1, \ldots, in_d$ and $out$, as a particular input-output example. The term synthesis problem is to find a closed first-order term $t$ such that $t$ evaluates to $out$ in each structure. The grammar can ensure that $out$ is not used in $t$. Again, we note that such first-order queries and terms can also be realized effectively as programs that operate over structures: a program for a query instantiates the free variables with all $r$-tuples, evaluates the formula, and returns those that satisfy it. Similarly, a term can be converted to a program that recursively evaluates its subterms and returns an element of the structure. We give adaptations of the automata-theoretic technique to solve both the query and term synthesis problems for the logic $\text{FO}(k)$.

Learning Algorithms for First-Order Logic with Least Fixed Points. A second contribution of this work is showing that a bounded variable version of first-order logic with least fixed points also has decidable separator realizability and synthesis. Least fixed point definitions add the power of recursion to first-order logic, resulting in a more expressive logic that, for instance, can describe transitive closure of relations and the semantics of languages like Datalog. Furthermore, over finite ordered structures, first-order logic with least fixed points captures the class $\mathcal{P}$ of all functions computable in polynomial time. Consequently, learned formulas in this logic can be realized as polynomial-time programs.

In this case we use two-way alternating tree automata to succinctly encode the semantics of expressions with least fixed point definitions over a given structure. Intuitively, we need to use recursion to evaluate a recursively-defined relation, and in each recursive call we need to read the definition of the relation once more. This capability is elegantly provided by two-way automata, and alternation gives a way to compositionally send copies of the automaton to check various subformulas effectively. Two-way alternating automata can be converted to one-way nondeterministic automata (with an exponential increase in states), and emptiness checking for the resulting automata gives the algorithm we seek for separator realizability and synthesis. We also solve the term synthesis problem for a logic with least fixed points, a problem which resembles functional program synthesis.

Further Results. A remarkable aspect of the automata-theoretic approach is that it provides algorithms for synthesis in many settings. The constructions smoothly extend to virtually any logic where the bounded variable restriction yields a formula evaluation strategy with a memory requirement that is independent of the size of the formula. In §9 we discuss decidable realizability and synthesis results for other settings where related problems have been studied, including languages with mutual recursion, e.g., Datalog [Albarghouthi et al. 2017; Evans and Grefenstette 2018] and inductive logic programming [Cropper et al. 2020; Muggleton and de Raedt 1994; Muggleton et al. 2014]. We also discuss how the technique extends to higher-order logics over finite models.

Complexity. For each logic we consider, we present the upper bound that the automaton construction yields in terms of various parameters: the number of variables $k$, the maximum size of each structure, the number of structures, and the size of the grammar $G$. A sample of upper bounds is given in Table 1 in §4. We also prove lower bounds for the logic $\text{FO}(k)$, arguing that the complexity of the upper bounds on certain parameters is indeed tight. In particular, we show that for fixed $k$, separator realizability is EXPTIME-hard. This matches the upper bound, and also proves matching lower bounds for more expressive logics.
In summary, our work provides an extremely general tree automata-theoretic technique that yields effective solutions for the problems of learning separators/queries/terms for several finite-variable logics. Our contributions here are theoretical. We establish decidability for several exact learning problems over different logics, and we give algorithms based on tree automata and some matching lower bounds. We believe and hope that this work will inform the design of practical algorithms for these problems. In particular, our automata constructions and the “bottom up” fixed point procedure for checking automaton emptiness gives a design framework for such algorithms. Due to the relatively high worst-case complexity of synthesis, practical algorithms will need to adapt to application domains and cater to restricted logics and languages that admit more efficient synthesis, potentially using heuristics, space-efficient data structures, and fast search (e.g., BDDs, SAT solvers, etc.) (see [Bloem et al. 2012; Wang et al. 2017b, 2018] for examples).

Details for proofs and automata constructions can be found in the appendix.

2 EXAMPLES

We begin with some examples to illustrate instances of the exact learning problems considered here. The first two examples explore the separability problem for first-order logic and the subtlety around reusing variables in bounded variable logics. The third example illustrates learning formulas with least fixed point definitions, and the fourth example illustrates term synthesis.

2.1 Example 1: Learning Formulas, Significance of Grammar, and Unrealizability

Consider the problem of finding a separating first-order sentence for the structures depicted in Figure 1 using a vocabulary that includes the binary edge relation $E$ and constants $s$ and $t$.

![Figure 1](image-url)

Fig. 1. Find a sentence in first-order logic that is true for $+$ structures and false for $-$ structures.

One possible solution asserts that every node that is adjacent to $s$ is adjacent to a node that is adjacent to $t$. In first-order logic:

$$\forall x. (E(s, x) \rightarrow \exists y. (E(x, y) \land E(y, t)))$$

If the grammar $G$ allowed, say, all first-order logic formulas with two variables, the formula above would indeed be a solution. If instead the grammar allowed only conjunction as a Boolean connective, then the formula above is not a separator. If the grammar allowed only one variable, or allowed two variables but disallowed universal quantification, then there is no separator.

In fact, for a grammar that only allows conjunction as Boolean connective, there is no separator for the structures above. Proof gist. For a contradiction: suppose there is a separator $\varphi$ that does not use negation or disjunction. Then $\varphi$ has a positive matrix (inner formula has no negations), and its standard conversion to an equivalent formula in prenex form, $\text{prenex}(\varphi)$, still has a positive matrix and will be a separator (though it may have many more variables). Since $\text{prenex}(\varphi)$ has a positive matrix, any graph that satisfies it will continue to satisfy it if we add any number of edges. Since the leftmost negative structure adds a single edge to the leftmost positive structure, we have a contradiction. Thus there is no separator.
Note that while algorithms can search for separators in $G$, the problem of declaring that there is no separator is a nontrivial problem, especially for arbitrary grammars. The algorithms we seek will terminate and declare unrealizability when separators do not exist (as in the above example).

### 2.2 Example 2: Reuse of Variables and Infinite Semantic Concept Space

Consider the problem of finding a separator that uses only three variables for the labeled structures in Figure 2. One possible solution is $\sqrt{\sum_{i=1}^{7} \text{path}_i(s, t)}$, where $\text{path}_i(x, y)$ holds for elements $x, y$ if there is a (directed) $E$-path of length $i$ from $x$ to $y$. Note that, by reusing variables, the formula $\text{path}_i(x, y)$ can be defined using only 3 variables for any $i \in \mathbb{N}$:

$$\text{path}_i(x, y) \leftrightarrow E(x, y)$$

$$\text{path}_{i+1}(x, y) \leftrightarrow \exists z. (E(x, z) \land \exists x. (x = z \land \text{path}_i(x, y))) \quad i > 0$$

This example shows that even when we restrict the number of variables, there is an infinite number of logically inequivalent sentences in first-order logic (e.g., $\text{path}_i$ for any $i > 0$). But note that this is not true if we bound the number of quantifiers or bound the depth of quantifiers. This infinite semantic space of concepts is what makes declaring unrealizability a nontrivial problem. As we will see, our technique works for finite variable logics in general, despite the fact that they admit infinitely-many inequivalent formulas.

![Fig. 2. Find a sentence in first-order logic that is true on + structures and false on − structures.](image)

### 2.3 Example 3: Least Fixed Points and Recursive Definitions

Notice that the separating sentence from Figure 2 has a size that depends on the sizes of the input structures, and it fails to capture the notion of a path of unbounded length. Consider the problem in Figure 3. A separating concept is all nodes can reach some cycle. By augmenting first-order logic with (least fixed point) recursive definitions, this concept can be expressed using the following recursive definition for $\text{reach}$ (which captures reachability using at least one edge):

$$\varphi := \text{let } \text{reach}(x, y) =_{lfp} E(x, y) \lor \exists z. (E(x, z) \land \text{reach}(z, y))$$

$$\text{in } \forall x. \exists y. (\text{reach}(x, y) \land \text{reach}(y, y))$$

![Fig. 3. Find a sentence in first-order logic with least fixed point definitions that is true on + structures and false on − structures and that does not depend on the sizes of the structures.](image)
Recursive definitions dramatically increase the expressivity of first-order logic. As studied in finite model theory, such logics encompass (over structures equipped with a linear order on the domain) all polynomial-time computable properties, i.e., the class $\mathcal{P}$ [Immerman 1982; Libkin 2004; Vardi 1982]. We consider learning in a finite-variable version of first-order logic with least fixed points that captures all properties computable in time $n^k$ using $O(k)$ variables. We note here a connection to the problem of synthesizing programs that are syntactically restricted in order to guarantee a specific implicit complexity [Dal Lago 2012], say polynomial time; we leave an exploration of this connection to future work.

2.4 Example 4: Term Synthesis and Program Synthesis

In addition to the separability problem illustrated in the previous examples, we also study the problem of term synthesis. In the term synthesis problem we aim to synthesize a term that evaluates to a specific element of the domain for each structure in a set of input structures. Specifically, given a set of structures, each with an interpretation for constants $i_{n_1}, \ldots, i_{n_d}$ and $\text{out}$, we want to construct a term $t$ in the language of a grammar $G$ such that $t$ has the same interpretation as $\text{out}$ in each structure. (Note the structures are not labeled in this problem.) The term synthesis problem, especially in the presence of recursive function definitions, resembles functional program synthesis.

Consider the problem of merging two sorted lists. We can model this setting with structures that represent finite prefixes of an abstract datatype for lists over a linearly ordered finite domain $\{a_1, a_2, \ldots, a_n\}$ (with ordering $<$). Figure 4 (top) depicts a portion of one such structure and its operations. We can model input-output tuples for the desired merge operation using an interpretation of constants $i_{n_1}, i_{n_2}$, and $\text{out}$, as depicted in the bottom of the figure.

![Figure 4](image-url)

**Table:**

| Example | $in_1$ | $in_2$ | $\text{out}$ |
|---------|--------|--------|-------------|
| $A_1$   | $\text{cons}(a_4, \text{nil})$ | $\text{cons}(a_2, \text{cons}(a_3, \text{nil}))$ | $\text{cons}(a_2, \text{cons}(a_3, \text{cons}(a_4, \text{nil})))$ |
| $A_2$   | $\text{cons}(a_1, \text{cons}(a_4, \text{nil}))$ | $\text{cons}(a_3, \text{nil})$ | $\text{cons}(a_2, \text{cons}(a_3, \text{cons}(a_4, \text{nil})))$ |

Fig. 4. (Top) Partial picture of a structure $A_1$ that encodes a finite prefix of a datatype for lists over an ordered domain, with terms bounded to depth 3. (Bottom) Input-output examples for merge. The goal is to find a closed term in first-order logic with least fixed point relations and recursive functions that evaluates to $\text{out}$ on structures $A_1$ (top) and $A_2$ (not shown).

One possible solution is the following term $t$ that defines a recursive function $\text{merge}$ and applies it to the inputs $i_{n_1}, i_{n_2}$:

$$
t := \text{let } \text{merge}(x, y) =_{\text{ifp}} \text{ite}(x = \text{nil}, y, \text{ite}(y = \text{nil}, x, (\text{ite}(\text{head}(y) > \text{head}(x), \text{cons}(\text{head}(x), \text{merge}(\text{tail}(x), y)), \text{cons}(\text{head}(y), \text{merge}(x, \text{tail}(y)))))))
\text{in } \text{merge}(i_{n_1}, i_{n_2})$$
\[ \varphi ::= R(\bar{t}) \mid \varphi \lor \varphi \mid \varphi \land \varphi \mid \neg \varphi \mid \exists x. \varphi \mid \forall x. \varphi \quad t ::= x \mid c \mid f(\bar{t}) \mid \text{ite}(\varphi, t, t') \]

Fig. 5. Grammar for first-order logic with if-then-else terms, denoted FO.

3 BACKGROUND

We begin with some preliminary notions from logic as well as the concepts of term, tree, and regular tree grammar, which we will need for our definitions of various automata.

3.1 Logic

3.1.1 Structures and Signatures. A first-order signature, or simply signature, is a set \( \tau \) of sets of relation symbols \( \{R_1, R_2, \ldots\} \), function symbols \( \{f_1, f_2, \ldots\} \), and constant symbols \( \{c_1, c_2, \ldots\} \). Each symbol \( s \) has an associated arity, denoted \( \text{arity}(s) \in \mathbb{N} \). The meaning of symbols in a signature depends on a \( \tau \)-structure, which is a tuple \( A = \langle \text{dom}(A), R^A_1, \ldots, R^A_{\bar{t}}, f^A_1, \ldots, f^A_{\bar{t}}, c^A_1, \ldots, c^A_{\bar{t}} \rangle \). The domain \( \text{dom}(A) \) is a set, each \( R^A_i \) is a relation on the domain, i.e., \( R^A_i \subseteq \text{dom}(A)^{\text{arity}(R_i)} \), each \( f^A_j \) is a total function on the domain, i.e., \( f^A_j : \text{dom}(A)^{\text{arity}(f_j)} \rightarrow \text{dom}(A) \), and each constant \( c \) denotes an element \( c^A \in \text{dom}(A) \). For simplicity, we model constants as nullary functions. Each problem addressed in this work involves finite structures, i.e., those for which \( |\text{dom}(A)| \in \mathbb{N} \). Thus structure will always mean finite structure. We omit \( \tau \) and write structure whenever \( \tau \) can be understood from context or is unimportant. We use \( A \) to denote an arbitrary structure.

3.1.2 First-Order Logic. Though the technique presented in this work is highly versatile, we will focus the majority of our presentation on variants and extensions of first-order logic. As a starting point, we consider first-order logic extended with an if-then-else term. Syntax for this logic, denoted FO, is given in Figure 5. The semantics of the usual FO formulas and terms is standard. We denote the interpretation of a term \( t \) in a structure \( A \) and variable assignment \( \gamma \) as \( t^A_{\gamma} \). The interpretation of the if-then-else term in \( A, \gamma \), is:

\[
\text{ite}(\varphi, t_1, t_2)^{A_{\gamma}} = \begin{cases} t_1^{A_{\gamma}} & A, \gamma \models \varphi \\ t_2^{A_{\gamma}} & \text{otherwise} \end{cases}
\]

We use \( \mathcal{L} \) to refer to an arbitrary logic, and occasionally, if we want to emphasize the signature we write \( \mathcal{L}(\tau) \) for a logic \( \mathcal{L} \) over \( \tau \). See [Enderton 2001] for syntax, semantics, and basic results in first-order logic.

3.2 Terms and Trees

Rather than working with strings, it will be simpler to instead consider logical formulas as finite ordered ranked trees, sometimes called terms. Intuitively, to build terms we use symbols from a finite ranked alphabet, that is, a set of symbols with corresponding arities. We use \( T_{\Sigma}(X) \) to denote the set of terms over a ranked alphabet \( \Sigma \) augmented with nullary symbols \( X \) (with \( X \) disjoint from \( \Sigma \)). When \( X = \emptyset \) we just write \( T_{\Sigma} \).

It will be convenient to also use the language of ordered trees. An ordered tree \( \rho \) over a label set \( W \) is a partial function \( \rho : \mathbb{N}^* \rightarrow W \) defined on \( \text{Nodes}(\rho) \subseteq \mathbb{N}^* \), a prefix-closed set of positions containing a root \( \epsilon \in \text{Nodes}(\rho) \). In this view, terms are simply ordered trees whose labels respect ranks, that is, ordered trees subject to the following requirement: if \( a \in \Sigma \) with \( \text{arity}(a) = n \) and for some \( x \in \text{Nodes}(\rho) \) we have \( \rho(x) = a \), then \( \{j \in \mathbb{N} \mid x \cdot j \in \text{Nodes}(\rho)\} = \{1, \ldots, n\} \). We will refer to ordered (ranked) trees as simply trees to avoid confusion with the usual syntactic category of logical terms.
3.3 Finite Variable Logics to Trees

When can the formulas of a logic be represented as trees over a finite alphabet? We probably must have a finite signature, as well as syntax formation rules that take a finite number of subformulas. For any logic \( \mathcal{L} \) and signature \( \tau \) that meets these requirements, if we bound the number of variables that can appear in any formula, then we can define a finite ranked alphabet \( \Sigma_{\mathcal{L}(\tau)} \) such that any formula \( \varphi \in \mathcal{L}(\tau) \) has at least one corresponding tree \( t \in T_{\mathcal{L}(\tau)} \). For example, consider a variant of FO restricted to the \( k \) variables in \( V = \{x_1, \ldots, x_k\} \), which we denote FO\((k)\). With superscripts for arity, the ranked alphabet looks as follows:

\[
\Sigma_{\text{FO}(k)} = \left\{ \text{arity}(R) \mid R \in \tau \right\} \cup \left\{ \text{arity}(f) \mid f \in \tau \right\} \cup \left\{ \exists^1, \forall^1, x^0 \mid x \in V \right\}
\]

We sometimes drop the subscript for the underlying logic and just use \( \Sigma \) to refer to finite ranked alphabets of this kind.

3.4 Regular Tree Grammars

A regular tree grammar (RTG) is a tuple \( G = \langle \{S\}, \Sigma, S, P \rangle \), where \( N \) is a finite set of nonterminal symbols, \( \Sigma \) is a finite ranked alphabet, \( S \in N \) is the axiom nonterminal, and \( P \) is a set of rewrite rules of the form \( B \to t \), where \( B \in N \) and \( t \in T_\Sigma(N) \). The language \( L(G) \) of \( G \) is the set of trees \( \{t \in T_\Sigma \mid S \Rightarrow^* t\} \), where \( t \Rightarrow t' \) holds whenever there is a context \( C \) and tree \( t'' \in T_\Sigma(N) \) such that \( t = C[B], t' = C[t''] \) and \( B \to t'' \in P \). These are standard notions; see [Comon et al. 2007; Grädel et al. 2002] for details.

Given a logic \( \mathcal{L} \), we consider RTGs over \( \mathcal{L} \). If \( \Sigma_{\mathcal{L}} \) is a finite ranked alphabet for \( \mathcal{L} \), then an RTG over \( \mathcal{L} \) is of the form \( G = \langle N, \Sigma_{\mathcal{L}}, S, P \rangle \) for some \( N, S, P \). In Figure 6 we give an example RTG over FO\((k)\)(graph) with \( k = 2 \), i.e., first-order logic with variables \( V = \{x, y\} \) over a signature graph consisting of a single binary relation symbol \( E \). When we refer to sentences or formulas in the remainder of this paper we mean their corresponding trees in a suitable RTG, and when we refer to a grammar we mean an RTG. When there are multiple ways to represent a given formula as a tree, we pick one arbitrarily.

3.5 Alternating Tree Automata

As we will show, the formalism of alternating tree automata yields an elegant technique for evaluating expressions on a fixed structure. We summarize the relevant ideas from automata theory.

An alternating tree automaton (ATA) over \( k \)-ary trees is a tuple \( \mathcal{A} = \langle Q, \Sigma, I, \delta \rangle \), where \( Q \) is a finite set of states, \( \Sigma \) is a finite ranked alphabet, \( I \subseteq Q \) is a set of initial states, and the transition function has the form \( \delta : Q \times \Sigma \to \mathcal{B}^+(Q \times \{1, \ldots, k\}) \), where \( \mathcal{B}^+(X) \) denotes the set of positive propositional formulas over atoms from a set \( X \). For any \( (q, a) \in Q \times \Sigma \) we require \( \delta(q, a) \in \{
\forall x \}, \{E(x, y) \}, \{E(x, x) \}
\}

\[
S \to \forall(S, S) \mid \land(S, S) \mid \neg(S) \mid \exists x(S) \mid \forall x(S) \mid \exists y(S) \mid \forall y(S) \mid E(x, x) \mid E(x, y) \mid E(y, x) \mid E(y, y)
\]

Fig. 6. (Left) Production rules from the set \( P \) for a regular tree grammar \( G = \langle \{S\}, \Sigma, S, P \rangle \), where \( \mathcal{L} \) is FO\((k)\)(graph) with \( k = 2 \), and (Right) a tree in \( L(G) \) for the sentence \( \forall x.\exists y. (E(x, y) \lor E(y, x)) \).

\footnote{We overload notation, using FO\((\tau)\) for FO over signature \( \tau \) and FO\((k)\) for FO with \( k \) variables and an unspecified signature. If the two notations are both needed at once we put the signature last, e.g., FO\((k)(\tau)\) is FO with \( k \) variables over \( \tau \).}
In this section we define three exact learning problems that are parameterized by a logic $L$. The first problem, $L$-separaror realizability and synthesis, is defined in Problem 1. Given positively- and negatively-labeled structures and a grammar $G$ over $L$, the problem is to synthesize a sentence in $G$ such that all positive structures make the sentence true and all negative structures make it false, or declare no such sentence exists. Sometimes we refer to this as the separability problem.

The second problem involves synthesizing $r$-ary queries. A $r$-ary query for a logic $L$ is a formula $\varphi(x_1, \ldots, x_r) \in L$ that has exactly $r$ distinct free variables, all first-order. The answer set for a $r$-ary query $\varphi$ in a structure $A$ is the precise set of tuples $\text{Ans} = \{ \bar{a} \in \text{dom}(A)^r \mid A \models \varphi(\bar{a}) \}$ that make the query true in the structure. For example, consider a “family relationships” domain...
Table 1. Summary of main results for a fixed signature and fixed arities of relations and functions.

| Problem                        | Parameters | Time complexity | Combined complexity (fixed variables) |
|--------------------------------|------------|----------------|---------------------------------------|
| FO separability               | $m$ input structures | $O\left(2^{\text{poly}(m^k)}|G|\right)$ | EXPTIME-complete in $mn + |G|$ |
| FO queries                    | $n$ max structure size $k$ first-order variables |                      |                                       |
| FO term synthesis             | FO parameters and $k'$ relation variables | $O\left(2^{\text{poly}(mn^k')}|G|\right)$ |                                       |
| FO-LFP separability           | FO parameters and $k'$ relation variables |                      |                                       |
| FO-LFP-FUN term synthesis     | FO parameters and $k_1$ relation variables $k_2$ function variables | $O\left(2^{\text{poly}(mn^k(k_1+k_2))}|G|\right)$ |                                       |

**Problem 1: $L$-separator realizability and synthesis**

**Input:** $\langle Pos = \{A_1, \ldots, A_{m_1}\}, Neg = \{B_1, \ldots, B_{m_2}\}, G \rangle$ where $A_i, B_j$ are $r$-structures $G$ an RTG over $\Sigma_L$

**Output:** $\phi \in L(G)$ s.t. for all $A_i \in Pos$, $A_i \models \phi$, and for all $B_j \in Neg$, $B_j \not\models \phi$

Or "No" if no such $\phi$ exists

with two structures $A_1$ and $A_2$. In $A_1$ there are domain elements Sue and Bob and the relationships Mother(Sue,Bob), and in $A_2$ there are elements Maria, Tom, and Anne and the relationships Mother(Maria, Tom) and Mother(Maria, Anne). Suppose the answer sets are $Ans_1 = \{\text{Sue}\}$ and $Ans_2 = \{\text{Maria}\}$. Then one possible solution is the query $\phi(x) := \exists y.\text{Mother}(x, y)$.

We call this second problem $L$-query realizability and synthesis, which is defined formally in Problem 2. Given a grammar $G$ over $L$ and a set of pairs, where each pair is a structure and an answer set, synthesize a query $\phi$ in $G$ such that $\phi$ precisely defines the given answer set in each structure, or declare no such $\phi$ exists.

**Problem 2: $L$-query realizability and synthesis**

**Input:** $\langle \{(A_1, Ans_1), \ldots, (A_m, Ans_m)\}, G \rangle$ where $A_i$ are $r$-structures $Ans_i \subseteq \text{dom}(A_i)'$ $G$ an RTG over $\Sigma_L$

**Output:** $\phi(x_1, \ldots, x_r) \in L(G)$ s.t. $\{\bar{a} \in \text{dom}(A_i)' \mid A_i \models \phi(\bar{a})\} = Ans_i$ for all $i \in [m]$

Or "No" if no such $\phi$ exists

The third problem, $L$-term synthesis, is defined in Problem 3. The input is a grammar $G$ and a set of (unlabeled) structures $\{A_1, \ldots, A_m\}$. Each structure $A_i$ interprets constants $in_1, \ldots, in_d$ and $out$, where $out$ is the target element in the domain of each structure. The goal is to synthesize a term $t$ from $G$ (which precludes using $out$) such that $A_i \models (t = out)$ for each $i$.

In Table 1 we highlight our main results for these three problems instantiated with various logics. Note that the upper bounds on time complexity assume a fixed signature, and in particular, fixed arities of symbols. The remainder of the paper lays out our general automata-theoretic solution.
Problem 3: \( L \)-term synthesis

**Input:** \( \{A_1, \ldots, A_m\}, G \) where
- \( A_i \) are \( \tau \)-structures
- \( G \) an RTG over \( \Sigma_L \)

**Output:** \( t \in L(G) \) s.t. \( A_i \models (t = \text{out}) \) for all \( i \in [m] \)
- Or "No" if no such \( t \) exists

## 5 SOLVING REALIZABILITY AND SYNTHESIS FOR FIRST-ORDER LOGIC

In this section we describe our general technique by instantiating it on the separability problem for the logic \( \text{FO}(k) \) (§5.1). We then show how to adapt the solution to solve query synthesis for the same logic (§5.2). Term synthesis is covered in §7.

### 5.1 Separator Realizability and Synthesis in First-Order Logic

Consider separability for \( \text{FO}(k) \) over an arbitrary signature. We are given a grammar \( G \) and sets of positive and negative structures \( \text{Pos} \) and \( \text{Neg} \). The main idea is to build an alternating tree automaton that accepts the parse trees of all formulas that separate \( \text{Pos} \) and \( \text{Neg} \). This automaton itself is constructed as the product of automata \( \mathcal{A}_M \), one for each structure \( M \in \text{Pos} \cup \text{Neg} \). If \( M \in \text{Pos} \), then \( \mathcal{A}_M \) accepts all formulas that are true on \( M \). If \( M \in \text{Neg} \), then \( \mathcal{A}_M \) accepts all formulas that are false on \( M \). Clearly, the intersection of these automata gives the desired automaton \( \mathcal{A}_\text{\#} \) that accepts formulas which separate \( \text{Pos} \) and \( \text{Neg} \). In §5.1.1, we give the main construction of \( \mathcal{A}_M \) for each \( M \in \text{Pos} \cup \text{Neg} \), which involves evaluating a given input formula on a fixed structure \( M \).

We build another tree automaton \( \mathcal{A}_G \) (§5.1.2) that accepts precisely the formulas from \( G \), and finally we construct an automaton accepting the intersection of languages for \( \mathcal{A}_\text{\#} \) and \( \mathcal{A}_G \) (§5.1.3). Checking emptiness of this automaton solves the realizability problem and, when the language is nonempty, finding a member of the language solves the synthesis problem.

#### 5.1.1 Automaton for Evaluating First-Order Logic Formulas

We now show how to construct a tree automaton that accepts the set of sentences in \( \text{FO}(k) \) that are true in a given structure \( A \).

For clarity, we present an automaton for a slightly simpler version of \( \text{FO}(k) \) over an arbitrary relational signature \( \tau \) and without an if-then-else term (thus the only terms are variables). The ranked alphabet for this simplification over \( \tau = \langle R_1, \ldots, R_s \rangle \) is:

\[
\Sigma'_\text{FO}(k) = \left\{ R_i(\bar{x})^0 \mid R_i \in \tau, \bar{x} \in \text{arity}(R_i) \right\} \cup \{\land^2, \lor^2, \neg^1\} \cup \{\forall x^1, \exists x^1 \mid x \in V\}
\]

Note that each atomic formula over variables \( V \) becomes a nullary symbol. Handling the full gamut of terms in \( \text{FO} \) from Figure 5 is straightforward but tedious, so we omit the details. Following the simpler construction, we give the high-level idea for the full version.

Fix a \( \tau \)-structure \( A \) with \( |\text{dom}(A)| = n \). We define an ATA \( \mathcal{A}_A = \langle Q, \Sigma'_\text{FO}(k), I, \delta \rangle \) whose language is the set of trees over \( \Sigma'_\text{FO}(k) \) corresponding to sentences that are true in the structure \( A \). Each component is discussed below.

**States.** The states of \( \mathcal{A}_A \) are partial assignments from variables \( V = \{x_1, \ldots, x_k\} \) to the domain \( \text{dom}(A) \). We denote the set of partial assignments by \( \text{Assign} \) \( \definedas V \rightarrow \text{dom}(A) \), and we use \( \gamma \) to range over \( \text{Assign} \). States of the automaton keep track of assignments that accrue when the automaton reads quantification symbols. The crucial idea is that for each syntax formation rule, we can express the conditions under which the formula is true with the current assignment as a positive Boolean formula over assignments and subformulas. The only hiccup is that the automaton needs to keep track of whether or not a formula should be satisfied or not satisfied, which is dictated by
occurrences of negation. We need a single bit for this, and so the state space increases by a factor of two. A state of \( \gamma \in \text{Assign} \) can be marked \( \bar{\gamma} \) to indicate that under assignment \( \gamma \) the input formula should not be true. We use \( \text{Dual}(X) \triangleq \{ x, \bar{x} \mid x \in X \} \) to denote a set \( X \) together with marked copies of its elements. With this notation, the state set for our automaton is \( Q \triangleq \text{Dual}(\text{Assign}) \), and \( |Q| = O(n^k) \). For a given \( \gamma \) we abuse notation and treat \( \gamma \) as a set of variable-binding pairs, for instance, \( \{ x \mapsto a_1, y \mapsto a_2 \} \) and \( \emptyset \) denote assignments in this way. We use \( \gamma[x \mapsto a] \) to denote the assignment that is identical to \( \gamma \) except it maps \( x \) to \( a \). We write \( \gamma(\bar{x}) \downarrow \) to denote that \( \gamma \) is defined on each \( x_1 \in \bar{x} \) and \( \gamma(\bar{x}) \) to denote the tuple of elements obtained by applying \( \gamma \) to \( \bar{x} \).

**Initial states.** There is only one initial state, namely, the one that assigns no variables: \( I = \{ \emptyset \} \).

**Transitions.** To define the transition function, for each assignment \( \gamma \in \text{Assign} \) and each symbol \( a \in \Sigma'_{\text{FO}(k)} \), we give a propositional formula that naturally mimics the semantics of first-order logic. The intuition is that, from a state \( \gamma \in \text{Assign} \), the automaton accepts every formula that is true in the structure \( A \) when free variables are interpreted according to \( \gamma \). For \( \gamma \in \text{Assign} \) and \( x, \bar{x} \) ranging over \( V \), the transitions are as follows:

\[
\delta(\gamma, \wedge) = (\gamma, 1) \wedge (\gamma, 2) \quad \delta(\bar{\gamma}, \wedge) = (\bar{\gamma}, 1) \vee (\bar{\gamma}, 2) \\
\delta(\gamma, \vee) = (\gamma, 1) \vee (\gamma, 2) \quad \delta(\bar{\gamma}, \vee) = (\bar{\gamma}, 1) \wedge (\bar{\gamma}, 2) \\
\delta(\gamma, \forall x) = \bigwedge_{a \in \text{dom}(A)} (\gamma[x \mapsto a], 1) \quad \delta(\bar{\gamma}, \forall x) = \bigvee_{a \in \text{dom}(A)} (\bar{\gamma}', 1), \quad \gamma' = \gamma[x \mapsto a] \\
\delta(\gamma, \exists x) = \bigvee_{a \in \text{dom}(A)} (\gamma[x \mapsto a], 1) \quad \delta(\bar{\gamma}, \exists x) = \bigwedge_{a \in \text{dom}(A)} (\bar{\gamma}', 1), \quad \gamma' = \gamma[x \mapsto a] \\
\delta(\gamma, R(\bar{x})) = \begin{cases} \text{True} & \gamma(\bar{x}) \downarrow, \ A, \gamma \models R(\bar{x}) \\ \text{False} & \text{otherwise} \end{cases} \quad \delta(\bar{\gamma}, R(\bar{x})) = \begin{cases} \text{True} & \gamma(\bar{x}) \downarrow, \ A, \gamma \not\models R(\bar{x}) \\ \text{False} & \text{otherwise} \end{cases} \\
\delta(\gamma, \neg) = (\bar{\gamma}, 1) \quad \delta(\bar{\gamma}, \neg) = (\gamma, 1)
\]

Note that for all \( (q, a) \in Q \times \Sigma'_{\text{FO}(k)} \) we have \( \delta(q, a) \in \mathcal{B}^+(Q \times \{ \text{arity}(a) \}) \), where for nullary symbols we can take \([0] = \emptyset \).

**Lemma 1.** \( \mathcal{A}_A \) accepts any sentence \( \varphi \) over \( \Sigma'_{\text{FO}(k)} \) that is true in \( A \).

**Proof Sketch.** A simple induction shows that for each assignment \( \gamma \) (resp. \( \bar{\gamma} \)), the language of \( \mathcal{A}_A \) from that state is precisely the set of formulas that are true (resp. false) in \( A \) under \( \gamma \), i.e.,

\[
L(\mathcal{A}_A, \gamma) = \{ \varphi(\bar{x}) \mid \gamma(\bar{x}) \downarrow, \ A, \gamma \models \varphi(\bar{x}) \} \quad (\text{resp. } \gamma(\bar{x}) \downarrow, \ A, \gamma \not\models \varphi(\bar{x}))
\]

If we fix an ordering on variables, then we can identify a state \( \gamma \) in the obvious way with a tuple of domain elements \( \bar{\gamma} \). Then the language of the automaton at \( \gamma \) coincides with the notion of *logical type* for the pair \( (A, \bar{\gamma}) \) [Libkin 2004]. One consequence of this generality is that the automaton can be seamlessly adapted to solve the query problem for \( \text{FO}(k) \), as we will see in §5.2.2.

### 5.1.2 Grammar Automaton.
As the name suggests, the language of an RTG is regular, and so it is the language of some tree automaton. Given \( G = \langle N, \Sigma, S, P \rangle \) an RTG, a (nondeterministic) tree automaton for it is simple to define. We let \( \mathcal{A}_G = \langle N, \Sigma, S, \delta \rangle \), where for each \( B \rightarrow f(B_1, \ldots, B_{\text{arity}(f)}) \in P \) we set \( \delta(B, f) = \bigwedge_i (B_i, i) \). Observe that we have not used the full power of alternation: the transition function puts at most one condition on any given child. Thus the automaton is already nondeterministic, which keeps the size of the final automaton (§5.1.3) linear in the size of the grammar. Notice also that we have made a simplifying assumption about the form of rules in \( P \), since the right-hand side could contain subterms that are not nonterminal symbols.
5.1.3 Decision Procedure. The decision procedure for realizability and synthesis is as follows. We define the automaton $\mathcal{A}_A$ as described above for each structure $A \in Pos$, as well as the automaton $\mathcal{A}_G$ for the grammar $G$. For each structure $B \in Neg$ we define $\mathcal{A}_B$ in the same way as for positive structures, with one tweak: instead of initial states $I = \{\emptyset\}$ we have $I = \{\tilde{\emptyset}\}$. We take the product of the structure automata to get:

$$\mathcal{A}_\cap = \bigtimes_{M \in Pos \cup Neg} \mathcal{A}_M$$

with number of states $O(mn^k)$, where $m = |Pos \cup Neg|$, $n = \max_{M \in Pos \cup Neg} |\text{dom}(M)|$, and $k$ is the number of variables. There is an exponential increase in states to convert $\mathcal{A}_\cap$ to a nondeterministic automaton $\mathcal{A}'_\cap$ with $L(\mathcal{A}'_\cap) = L(\mathcal{A}_\cap)$ [Comon et al. 2007; Grädel et al. 2002]. Finally, we take the product of $\mathcal{A}'_\cap$ and $\mathcal{A}_G$ to get $\mathcal{A} = \mathcal{A}'_\cap \times \mathcal{A}_G$, with $L(\mathcal{A}) = L(\mathcal{A}'_\cap) \cap L(\mathcal{A}_G)$, and furthermore, $L(\mathcal{A}) \neq \emptyset$ if and only if there is a sentence $\varphi \in L(G)$ that separates the input structures. We solve realizability by checking emptiness of $\mathcal{A}$ in time linear in its size, which is $O(2^{\text{poly}(mn^k)} |G|)$, and the emptiness checking algorithm can construct a (small) tree if nonempty.

This construction can be easily extended to give us the following theorem for realizability and synthesis in the full logic $FO(k)$ that includes if-then-else and function terms.

**Theorem 2.** $FO(k)$-separator realizability and synthesis is decidable in EXPTIME for a fixed signature and fixed $k \in \mathbb{N}$.

The construction for the full gamut of terms is straightforward. It can be accomplished by not only keeping partial assignments but also states that encode the currently expected domain element for a term under evaluation by the automaton. We give more details for how this extension works when we discuss term synthesis in §7.

5.2 Query Realizability and Synthesis

As noted, the automaton $\mathcal{A}_A$ (§5.1.1) is more general than an acceptor of sentences, and it can be easily modified as follows to solve query synthesis for $FO(k)$ with no increase in complexity.

For a pair $(A, Ans)$ of a structure and an answer set, with $Ans \subseteq \text{dom}(A)'$, we define an ATA $\mathcal{A}_A = (Q, \Sigma, I, \delta)$ whose language is the set of all formulas $\varphi(y_1, \ldots, y_r) \in FO(k)$, with $r \leq k$, whose answer set in $A$ is $Ans$. We describe each component below. For simplicity, we work with a fixed permutation of $r$ distinct variables $\bar{y} \in V'$, where $V = \{x_1, \ldots, x_k\}$.

**States.** The set of states is unchanged: $Q := \text{Dual}(\text{Assign})$.

**Transitions.** The transition function $\delta$ is unchanged, with the exception of the following transitions for the initial state $q_1 = \emptyset$.

**Initial states.** $I = \{q_i\}$. The main idea is to (a) require that the automaton reads and accepts the input formula from all states (partial variable assignments) that correspond to tuples in the answer set $Ans$, and (b) reject from states corresponding to the complement of $Ans$. Let $S(\bar{y}) \subseteq Q$ be the set of assignments defined only on $\bar{y}$, and let $S(Ans) \subseteq S(\bar{y})$ be the subset of assignments that map $\bar{y}$ to a member of the answer set. For any $a \in \Sigma'_{FO(k)}$, the transition out of $q_i$ is given by:

$$\delta(q_i, a) = \left( \bigwedge_{y \in S(Ans)} \delta(y, a) \right) \land \left( \bigwedge_{y \in S(\bar{y}) \setminus S(Ans)} \delta(\bar{y}, a) \right)$$

The following theorem follows easily from the proof of Lemma 1.
**Theorem 3.** $\text{FO}(k)$-query realizability and synthesis is decidable in \text{EXPTIME} for a fixed signature and fixed $k \in \mathbb{N}$.

### 6 REALIZABILITY AND SYNTHESIS WITH LEAST FIXED POINT DEFINITIONS

In this section we study separability for logics with least fixed point operators. In particular, we choose a logic with a finite set of relation variables, each of which can be defined recursively. These relation variables, though finite in number, can be redefined any number of times (similar to reusing variables, as we saw in §2.2). Further, as we will discuss in §9.1, the ideas presented in this section can be extended to handle mutually recursive definitions.

Note that the ability to define a relation is valuable independently of recursion: with definitions, we can require a formula to be synthesized and then used in multiple distinct places. For example, we may want to express: there exist $x$ and $y$ that are related in some unknown way, and further, all things related in that way also share a property $\psi$. In logic, this amounts to a separator of the form:

$$\exists x. \exists y. \phi(x, y) \land (\forall x. \forall y. \phi(x, y) \rightarrow \psi(x, y))$$

Notice that $\phi$ appears twice, which we cannot express with a regular tree grammar. However, with relation variables and definitions we can ask to synthesize a formula $\phi$ in a template as follows:

let $R(x, y) = \phi(x, y)$ in $\exists x. \exists y. R(x, y) \land (\forall x. \forall y. R(x, y) \rightarrow \psi(x, y))$

Following the semantics of our logic with least fixed point definitions, we will see how the automata-theoretic approach extends neatly to accommodate both definitions and recursion by moving from alternating tree automata to two-way tree automata.

#### 6.1 First-Order Logic with Least Fixed Points

Here we describe $\text{FO-LFP}$, which is an extension of FO with recursively-defined relations with least fixed point semantics. The syntax is given in Figure 7. Formulas in $\text{FO-LFP}$ can define relations using a set $\{P_1, P_2, \ldots\}$ of symbols disjoint from the signature. Such symbols are interpreted as least fixed points of the set operators induced by their definitions. Note that, although not shown in Figure 7, we require all relations to be defined before they are used.

Recall the definition for reachability from Figure 3:

$$\phi := \text{let } \text{reach}(x, y) =_{\text{lfp}} (E(x, y) \lor \exists z. E(x, z) \land \text{reach}(z, y)) \text{ in } \phi'(\text{reach})$$

For a fixed structure $A$, the meaning of $\text{reach}(x, y)$ is obtained by first interpreting the definition $\psi(x, y, \text{reach}) := E(x, y) \lor \exists z. E(x, z) \land \text{reach}(z, y)$ as a monotonic function $F_\psi : 2^X \rightarrow 2^X$ over the lattice defined by the subset relation on $2^X$, where $X = \text{dom}(A)^2$. Formally, for $Y \subseteq X$,

$$F_\psi(Y) \triangleq \{(a_1, a_2) \in X \mid \psi(a_1/x, a_2/y, Y/\text{reach})\},$$

where by $(Y/\text{reach})$ we mean that $\text{reach}$ is interpreted as the relation $Y$ in $\psi$ (similarly for $a_1/x$ and $a_2/y$). Then for any $a, a' \in \text{dom}(A)$, $\text{reach}(a, a')$ holds if and only if $(a, a') \in \text{lfp}(F_\psi)$, where $\text{lfp}(F_\psi)$ denotes the least fixed point of $F_\psi$. More generally, definitions in $\text{FO-LFP}$ are interpreted as follows in a structure $A$:

$$A \models \text{let } P(\bar{x}) =_{\text{lfp}} \psi(\bar{x}, P) \text{ in } \phi(P) \iff A \models \phi(\text{lfp}(F_\psi)/P)$$

(1)

Note that the least fixed point $\text{lfp}(F_\psi)$ may not exist for an arbitrary formula $\psi$. It turns out, however, that a simple syntactic restriction can ensure existence of least fixed points. Technically, we require all occurrences of $P$ in $\psi$ to occur under an even number of negations. This restriction can be enforced by the grammar, and we will not mention it further.
A tree is accepted by a two-way automaton if there is a run for which every branch reaches a final state. We define a two-way tree automaton as a variant of a nondeterministic automaton in exponential time. When reading an occurrence of a definable relation symbol, any node can navigate an input tree in both directions (from a node to its children or to its parent), thus making all parts of a tree accessible from any node. When reading an occurrence of a definable relation symbol \( P \), the automaton can navigate to the corresponding definition, which is elsewhere in the tree, and read it. This same capacity to move up and down in the tree gives us an elegant way to describe the evaluation of recursive definitions, which must be read multiple times to compute.

### 6.2.2 Automaton for Evaluating Formulas with Recursive Definitions

For simplicity, we again describe a construction for the simpler variant of FO-LFP without functions and if-then-else terms. The ranked alphabet for this simplification looks as follows:

\[
\Sigma'_{\text{FO-LFP}(k,k')} = \left\{ \text{let } P(x)^2, P(x)^0 \mid P \in \{P_1, \ldots, P_{k'}\}, x \in V^{\text{arity}(P)} \right\} \cup \Sigma'_{\text{FO}(k)}
\]

Let us assume each symbol \( P_i \) has \( \text{arity}(P_i) = r \). For a fixed structure \( A \) with \( |\text{dom}(A)| = n \), we define a two-way tree automaton \( \mathcal{A}_A = \langle Q, \Sigma'_{\text{FO-LFP}(k,k')}', I, \delta, F \rangle \) whose language is the set of sentences in (simplified) FO-LFP(k, k') that are true in A. We discuss each component next.
**States.** In addition to assignments, each state keeps track of information that enables the automaton to evaluate recursively-defined relations. This includes (1) whether the automaton is going up to find a definition or down to evaluate a formula, (2) a counter value from \( \text{Count} \triangleq \{0, \ldots, n'\} \) that tracks the stage of the current least fixed point computation, and (3) the current definition being evaluated (if any), i.e., a member of the set \( \text{Defn} \triangleq \{\bot, P_1, \ldots, P_k\} \). Finally, some states have a tuple of domain elements from \( \text{Val} \triangleq \text{dom}(A)^r \) rather than a partial assignment, which is used to pass values to the body of a definition whenever a defined relation is used. Note that the distinction between a tuple and an assignment takes care of part (1) above. Similar to partial assignments, the tuples are marked to indicate the automaton’s mode of operation: checking a formula is either true (verifying) or false (falsifying). Combining the above, we have

\[
Q := \text{Dual} (\text{Assign}) \times \text{Count} \times \text{Defn} \cup \text{Dual} (\text{Val}) \times \text{Count} \times \text{Defn} \cup \{q_f\},
\]

where \( q_f \) is distinct from all other states and \( F = \{q_f\} \).

The states \( Q \) can be divided into two categories: up and down. The up states correspond to checking membership in a defined relation. In an up state \( \langle \text{val}, \text{count}, \text{defn} \rangle \in \text{Dual} (\text{Val}) \times \text{Count} \times \text{Defn} \subseteq Q \), the automaton navigates up on the input tree to find the definition for \( \text{defn} \), and it carries a tuple \( \text{val} \) of domain elements that it should check for membership in the defined relation. In a down state \( \langle \text{assign}, \text{count}, \text{defn} \rangle \in \text{Dual} (\text{Assign}) \times \text{Count} \times \text{Defn} \subseteq Q \), the automaton evaluates a formula under the variable assignment \( \text{assign} \) while navigating down in the input tree.

**Initial states.** There is one initial state containing an empty assignment, a counter at 0, and the current definition set to \( \bot \), i.e., \( I = \{(\emptyset, 0, \bot)\} \).

**Transitions.** The transitions for symbols shared with FO are similar to the earlier construction (§5.1.1). The novelty is to define transitions for definitions and occurrences of defined relations \( P_i(\bar{x}) \). For intuition, consider the increasing sequence of “approximations” for a relation defined by a formula \( \psi \). The least fixed point for the operator \( \text{Dual} \) (see §6.1) can be computed in \( n' \) steps by iteratively applying \( \text{Dual} \), starting from \( \emptyset \), giving us the sequence

\[
\emptyset \subseteq F_{\psi}(\emptyset) \subseteq \cdots \subseteq F_{\psi}^i(\emptyset) = F_{\psi}^{i+1}(\emptyset),
\]

where \( i \leq n' \) (follows from monotonicity of \( F_{\psi} \)). When the automaton reads a defined relation \( P_i(\bar{x}) \) in a state \( \langle \gamma, j, P_i \rangle \), it will attempt to verify that \( \gamma(\bar{x}) = \bar{a} \in F_{\psi}^j(\emptyset) \). Similarly, in a state \( \langle \gamma, j, P_i \rangle \) it will attempt to verify that \( \gamma(\bar{x}) = \bar{a} \notin F_{\psi}^j(\emptyset) \).

Presenting all of the many transitions would obscure the main ideas, so we give only a description of interesting ones here; a full account can be found in the appendix. We focus on four cases: (1) reading a defined symbol, (2) finding a definition, (3) reading a definition, and (4) a variation on (1) where the defined symbol being read is not the current definition. Below, we use \( \text{assign} \in \text{Dual} (\text{Assign}), \gamma \in \text{Assign}, \text{val} \in \text{Dual} (\text{Val}), v \in \text{Val}, j \in \text{Count}, \text{and} \ P_i, P_j \in \text{Defn}. \)

1. **Reading a defined symbol.** Suppose the automaton is reading “\( P_i(\bar{x}) \)” in a down state \( \langle \text{assign}, j, P_i \rangle \). Thus it is currently reading the definition for \( P_i \) (call the definition \( \psi \)) and has encountered a use of \( P_i \) in the form \( P_i(\bar{x}) \). Suppose \( j = 0 \). If \( \text{assign} = \gamma \), then the automaton is verifying and it must verify that \( v = \gamma(\bar{x}) \) is in the \( j \)th stage of the least fixed point computation for the definition of \( P_i \). The transition for this case is False, since \( v \notin \emptyset = F_{\psi}^0(\emptyset) \). Otherwise, if \( \text{assign} = \gamma \), then the automaton is falsifying and must verify that \( v = \gamma(\bar{x}) \) is not in the \( j \)th stage. So the transition for this case is True, since \( v \notin \emptyset \). If \( j > 0 \), in both cases the transition forces the automaton to navigate up to the definition and evaluate it. It does this by changing to the up state \( \langle v, j, P_i \rangle \), if verifying, and to the up state \( \langle \bar{v}, j, P_i \rangle \), if falsifying.
(2) **Finding a definition.** Suppose step (1) has just occurred, and thus the automaton is looking for a definition of \( P_i \). The automaton continues moving up on the input tree until it encounters a symbol that marks the definition of \( P_i \), i.e., a symbol of the form “let \( P_i(\bar{\cdot}) \)”. Note: if the definition does not exist in the tree, then the automaton continues to the root, at which point it can make no valid transition and the tree is rejected.

(3) **Reading a definition.** Suppose step (2) has just occurred and the automaton is in state \( \langle \text{val}, j, P_i \rangle \) reading “let \( P_i(\bar{\cdot}) \)”. The automaton decrements the counter and proceeds to evaluate the definition by moving into the child tree corresponding to the definition of \( P_i \). It enters a down state with assignment \( \gamma = \bigcup_i \{ x_i \mapsto v_i \} \) that maps variables in \( \bar{\cdot} \) to the passed values in \( \text{val} \). If the automaton is verifying (\( \text{val} = v \)), then the new state is \( \langle \gamma, j - 1, P_i \rangle \). Otherwise, the automaton is falsifying (\( \text{val} = \bar{\cdot} \)) and the new state is \( \langle \bar{\cdot}, j - 1, P_i \rangle \).

(4) **Reading a new defined symbol.** Suppose the automaton is reading “\( P_i(\bar{\cdot}) \)” in a down state \( \langle \text{assign}, \text{count}, P_j \rangle \), where \( P_j \neq P_i \). Thus it has encountered a use of a defined relation \( P_i \) that it is not currently reading. Rather than checking the value of \( \text{count} \), as in case (1) above, the automaton continues on to case (2) with the current definition set to \( P_i \) and with a fresh counter initialized at \( \text{count} = n' \).

**Acceptance.** Recall that the automaton accepts a tree \( t \) if it has a run on \( t \) where every branch reaches the final state \( q_f \). We note that the full set of transitions ensures that the only way to reach the final state \( q_f \) is via True transitions.

**Lemma 4.** \( A_A \) accepts any sentence \( \varphi \in \Sigma'_\text{FO-LFP}(k,k') \) that is true in \( A \).

### 6.2.3 Decision Procedure.

We define automata \( A_A \) for each input structure \( A \) as described in the construction, with the proviso that negative structures have initial states \( I = \{ \langle \emptyset, 0, \bot \rangle \} \). We take the product of these automata as before and convert the resulting automaton to a one-way nondeterministic automaton without alternation by adapting the technique of [Vardi 1998]. We further take the product of the nondeterministic automaton with the grammar automaton \( A_G \). Checking emptiness of the final automaton, which has size \( O(\text{poly}(mn^k k')(G)) \), gives us the decision procedure for realizability and synthesis. Again, it is straightforward to adapt the construction to full FO-LFP \( (k,k') \) with if-then-else and function terms, giving us:

**Theorem 5.** FO-LFP \( (k,k') \)-separator realizability and synthesis is decidable in EXPTIME for a fixed signature and fixed \( k, k' \in \mathbb{N} \).

### 7 TERM SYNTHESIS

In this section we show that it is possible to use the same general approach to **synthesize terms**. We examine the term synthesis problem for a logic similar to FO-LFP with recursively-defined functions (adaptations for other logics are similar). Much of the construction is similar to the construction for FO-LFP (§6). We give the main idea by showing how the evaluation automaton’s state space changes, and we describe at a high level some new transitions related to terms.

#### 7.1 First-Order Logic with Least Fixed Points and Recursive Functions

We consider term synthesis for a logic with least fixed point relations and recursively-defined functions. Recall the list merge example from Figure 4. Given a set of (unlabeled) structures that each interpret the constants \( in_1, \ldots, in_d, \) and \( out \) as an input-output example for some functional relationship (e.g., \( \text{merge}(in_1, in_2) = out \)), the goal is to decide whether there is a term \( t \) that evaluates
We now sketch the main idea for defining an automaton that accepts all terms which evaluate to a

\[ \overline{u_1} \overline{u_2} \overline{u_3} \overline{u_4} \overline{u_5} \overline{u_6} \overline{u_7} \overline{u_8} \overline{u_9} \overline{u_{10}} \]

function \( f \) of the chain \( \overline{u_1} \overline{u_2} \overline{u_3} \overline{u_4} \overline{u_5} \overline{u_6} \overline{u_7} \overline{u_8} \overline{u_9} \overline{u_{10}} \) has finite height since all structures here are finite. Suppose \( A \) is a two-way automaton

\[ \psi := \exists x. \psi \mid \forall x. \psi \mid R(\overline{t}) \mid P(\overline{t}) \]

\[ \text{let } P(x) = \text{If } \psi \text{ in } \varphi \]

\[ \text{let } g(x) = \text{If } t \text{ in } \varphi \]

\[ \text{let } P(x) = \text{If } \psi \text{ in } \varphi \]

\[ t := x \mid c \mid f(\overline{t}) \mid \text{ite}(\psi, t, t') \]

Fig. 8. Syntax for FO-LFP-FUN with \( \varphi \) the starting nonterminal. The language is similar to FO-LFP, but it yields terms rather than formulas and adds (recursively) definable functions \( g \).

to \( \text{out} \) in each structure, and to synthesize one if it exists. We explore this problem for the language \( \text{FO-LFP-FUN} \) (syntax in Figure 8), which is similar to FO-LFP and, additionally, has recursively-definable functions and yields \textit{terms} rather than formulas.

The semantics for \( \text{FO-LFP-FUN} \) coincides with FO-LFP on shared features. The only novelty is in how we define the semantics of recursive functions and, in particular, how we account for functions that may not be total on the domain of the structure \( A \). We choose to interpret them as \textit{partial functions} on \( \text{dom}(A) \) and to interpret formulas in a 3-valued logic. Note that in this setting it is also simple and convenient to allow input structures to interpret function symbols from the signature as partial functions. For instance, in the \textit{merge} example from Figure 4, we might prefer to specify \textit{head} as a partial function that is only defined on elements that denote lists. We give here a high-level description of the semantics for recursive functions; details can be found in the appendix.

\textbf{Semantics for Recursive Functions.} A defined function \( g \) with \( \text{arity}(g) = d \) is interpreted as a partial function \( g^A : \text{dom}(A)^d \rightarrow \text{dom}(A) \), which is a member of the bottomed partial order \( O = (\text{dom}(A)^d, \subseteq, \bot) \), where \( \bot \) is undefined everywhere and \( f \subseteq f' \) holds if for all \( \overline{a} \in \text{dom}(A)^d \), whenever \( f(\overline{a}) \) is defined, then \( f'(\overline{a}) \) is defined and \( f(\overline{a}) = f'(\overline{a}) \). This partial order has finite height since all structures here are finite. Suppose \( g \) is defined recursively using a term \( t(x_1, \ldots, x_d, g) \). We associate a monotone function \( F_t : O \rightarrow O \) to the defining term \( t \), and let \( g^A \) be the least fixed point of \( F_t \), which can easily be shown to exist and to be equal to the stable point of the chain \( \bot \subseteq F_t(\bot) \subseteq \cdots \subseteq F_t^i(\bot) = F_t^{i+1}(\bot) \), with \( i \leq |\text{dom}(A)^d| \). The syntax and semantics for terms in FO-LFP-FUN guarantee monotonicity of the function \( F_t \) for any term \( t \). It follows that least fixed points exist for each definition. In general, however, care is needed to ensure that the least fixed point is total on \( \text{dom}(A)^d \), and whether or not this is so depends on the definition and the structure \( A \).

7.2 Automaton for Evaluating Terms with Recursive Functions

We now sketch the main idea for defining an automaton that accepts all terms which evaluate to a given domain element \( a \) in a structure \( A \). We consider the language \( \text{FO-LFP-FUN}(k, k') \), which restricts FO-LFP-FUN to \( k \) first-order variables and \( k' \) definable relations and functions from \( P = \{ P_1, \ldots, P_d \} \) and \( F = \{ g_1, \ldots, g_k \} \), respectively, with \( k' = k_1 + k_2 \). A ranked alphabet for \( \Sigma_{\text{FO-LFP-FUN}(k, k')} \) extends \( \Sigma_{\text{FO-LFP-FUN}(k', k')} \) in an obvious way. Note that here we consider input trees representing arbitrarily deep logical terms, in contrast to our earlier simplifications.

Fix a structure \( A \) of size \( n = |\text{dom}(A)| \) and fix a domain element \( a \in \text{dom}(A) \). We want to define a two-way automaton \( \mathcal{A}_A = (Q_{\text{TERM}}, \Sigma_{\text{FO-LFP-FUN}(k, k')}, I, \delta, F) \) that accepts the set of closed terms \( t \in \text{FO-LFP-FUN}(k, k') \) such that \( t^A = a \). For simplicity, assume \( r \text{-ary functions and relations only} \).
The sets $\text{Assign} \triangleq V \rightarrow \text{dom}(A)$, $\text{Val} \triangleq \text{dom}(A)'$, $\text{Dual}(X) \triangleq \{x, \bar{x} \mid x \in X\}$, $\text{Count} \triangleq \{0, \ldots, n'\}$, and $\text{Defn} \triangleq \{\bot, P_1, \ldots, g_{k_j}\}$ serve the same purposes as before. Notice that the automaton has a new category of states, namely, those for evaluating terms. The transitions related to formulas are very similar to those for FO-LFP. Below, we give three representative cases for the transition function $\delta$. Let $\gamma \in \text{Assign}$, $j \in \text{Count}$, and $g, \text{defn} \in \text{Defn}$.

Reading variables. The automaton is reading “$x$” and verifying that the input tree evaluates to $a \in \text{dom}(A)$. It only needs to check that the variable $x$ is mapped to $a$ in the current assignment $\gamma$:

$$\delta(\langle \gamma, j, \text{defn}, a, x \rangle) = \begin{cases} \text{True} & \gamma(x) \downarrow, \gamma(x) = a \\ \text{False} & \text{otherwise} \end{cases}$$

Reading if-then-else terms. The automaton is reading “ite” and verifying that the input tree evaluates to $a \in \text{dom}(A)$. It must either (1) verify the condition formula (first child) and verify that the term in the “then” branch (second child) evaluates to $a$ or (2) falsify the condition formula and verify that the term in the “else” branch (third child) evaluates to $a$:

$$\delta(\langle \gamma, j, \text{defn}, a, \text{ite} \rangle) = (\langle \gamma, j, \text{defn}, a, 1 \rangle \land (\langle \gamma, j, \text{defn}, a, 2 \rangle \lor (\langle \gamma, j, \text{defn}, 1 \rangle \land (\langle \gamma, j, \text{defn}, a, 3 \rangle)

Reading a defined function $g$. The automaton is reading “$g$” and the current definition is set to $g'$, with $g' \neq g$, analogous to case (1) from §6.2.2. The automaton “guesses and checks” that the argument terms for $g$ evaluate to $\bar{a}$ and ascends to the definition of $g$ with a fresh counter set to $n'$ to verify that $g$ evaluates to $a$ when applied to $\bar{a}$:

$$\delta(\langle \gamma, j, g', a, g \rangle) = \bigvee_{a \in \text{Val}} \left( (\langle \bar{a}, n', g, a, -1 \rangle \land \bigwedge_{i \in [r]} (\langle \gamma, j, g', a_i, i \rangle) \right) \quad (g' \neq g)$$

The rest of the transitions follow along these lines and are similar to those for FO-LFP. There is a single initial state with $I = \{\langle \emptyset, 0, \bot, a \rangle\}$, and the acceptance condition is again reachability with $F = \{q_f\}$. Similar reasoning to that for the FO-LFP construction can be used to show:

**Theorem 6.** FO-LFP-FUN($k, k'$)-term synthesis is decidable in \textsc{Exptime} for a fixed signature $\tau$ and fixed $k, k' \in \mathbb{N}$.

The idea sketched here can be easily added to earlier constructions without increasing complexity in order to solve synthesis for logics with the full gamut of terms and, in particular, to solve term synthesis for FO($k$) in the same complexity as the separability problem.

8 LOWER BOUNDS

Here we present lower bounds arguing the upper bound complexity we obtain on certain parameters of the problem is indeed tight. Given the number of logics and variants (and problems for separators, queries, and terms), we focus on lower bounds only for FO($k$); of course, these also give lower bounds for more expressive languages and variants.

Proc. ACM Program. Lang., Vol. 6, No. POPL, Article 10. Publication date: January 2022.
8.1 A Lower Bound for Separability in FO(k)

The upper bound for separability in FO(k) from §5 is linear in the size of the grammar and exponential in $mn^k$, where $m$ is the number of input structures and $n$ is the maximum size of any input structure. Hence, for a fixed $k$, the algorithm we propose is exponential time in the size of the input. We show a matching lower bound (this can be adapted for queries and terms as well).

**Theorem 7.** FO(k)-separator realizability is EXPTIME-hard for any fixed $k > 4$.

**Proof.** The reduction is from the word acceptance problem for an alternating polynomial space Turing machine. Given an alternating Turing machine $M$ and an input $w$, with $M$ using space $s$ that is polynomial in $|w|$, the reduction yields $s$ positively labeled first-order $r$-structures $A_1, \ldots, A_s$, with $|\text{dom}(A_j)| = O(s)$, and a grammar $G$ of size polynomial in $|\langle M, w \rangle|$. The signature $r$ depends only on $M$. Each structure consists of two parts: (1) a cycle of length $s$ and (2) a gadget encoding the transition relation for $M$ along with unique constants for tape symbols from the machine’s tape alphabet $\Gamma$. (We use constants that encode the machine head and state, i.e., $\Gamma' = \Gamma \times Q \cup \Gamma$.)

Let us consider the purposes of the *structures* and the *grammar*, which are complementary. The structures can be viewed as distinct copies of the machine $M$ that are used to verify that a computation tree for $M$ on $w$, encoded in a sentence from the grammar, obeys the transition relation for each of the $s$ tape cells. The grammar can be viewed as a skeleton of computation trees for $M$ on input $w$. A particular sentence $\varphi$ from the grammar $G$ encodes a computation tree of potentially exponential depth, and it asserts many things about the structure on which it is interpreted. When understood together across all structures, the truth of the assertions is equivalent to the computation tree being an accepting computation tree, i.e., that successive configurations follow the transition relation and the final configuration is accepting. The main trick is to emulate in the grammar the generation of each successive machine configuration in a way that allows checking the transition relation. This is not entirely straightforward because $s$ tape symbols cannot all be stored at once in $k$ variables ($k$ is fixed). Here is some intuition for how we accomplish this.

![Fig. 9. Structure $A_j$ tracks the window centered on the $i$th cell and has a special cycle element (the dark node) of distance $i$ from the start of the cycle, denoted $\ast$. As the symbols of a configuration are produced, the variables $\tilde{y}$ are equated with the current tape window if the cycle pointer is equal to the dark element.](image)

Each of the $s$ structures is made to track a distinct window of three contiguous tape cells (as well as the previous contents for the window). The grammar uses a polynomial-sized gadget of nonterminals to iteratively produce the tape cell contents of a given configuration. Refer to Figure 9 in the following for a picture of how this gadget works over each structure. In each iteration, the grammar moves a $\text{ptr}$ variable along the cycle of size $s$. If $\text{ptr}$ is equal to a special element, filled dark in Figure 9, then the grammar requires the current tape cell’s contents (and neighbors) to be stored in variables by asserting an equality. Each structure is made to track a unique window by differently interpreting the distance between a starting node, denoted $\ast$, and the special dark node. The grammar $G$ ensures the following invariant holds for all sentences $\varphi \in L(G)$: if we evaluate $\varphi$ in $A_i$, then upon evaluating the subformula of $\varphi$ that picks a symbol for cell $i+1$, window $i$ of the previous configuration is stored in variables $x_1, x_2, x_3$ and window $i$ of the current configuration

Proc. ACM Program. Lang., Vol. 6, No. POPL, Article 10. Publication date: January 2022.
is stored in variables $y_1, y_2, y_3$. The grammar checks that successive windows obey the transition relation by asserting the relation $\delta(\text{choice}, \hat{x}, \hat{y})$, where $\delta$ encodes the transition relation for $M$ and $\text{choice} \in \{0, 1\}$ encodes which of two transitions for the alternating machine is being verified.

More details can be found in the appendix. \hfill \Box

**Theorem 8.** $\text{FO}(k)$-query realizability is EXPTIME-hard for fixed $k > 4$.

**Proof.** Reduction from $\text{FO}(k)$-separator realizability. Positive structures have a full query answer set and negative structures have an empty query answer set. \hfill \Box

**Theorem 9.** $\text{FO-LFP}(k, k')$-separator realizability is EXPTIME-hard for fixed $k, k' \in \mathbb{N}$ with $k > 4$.

**Proof.** Reduction from $\text{FO}(k)$-separator realizability. \hfill \Box

### 8.2 More Lower Bounds and Open Problems

We can now ask whether separator realizability for $\text{FO}(k)$ is decidable in polynomial time if there is only one structure. With only one structure (positively labeled, say) the problem of separability may seem odd, but checking whether there is any sentence in the grammar $G$ that is true on the single structure is actually a nontrivial problem; indeed, the grammar is quite powerful.

We can show a general reduction from separator realizability for multiple structures to realizability for a single structure (but over a different grammar and for $k' = k + 1$). Given a set of positive structures $\text{Pos}$, negative structures $\text{Neg}$, and a grammar $G$ (over $\Sigma_{\text{FO}(k)}$), we can reduce the realizability problem to a new realizability problem over a single positive structure $M$ and a grammar $G'$ (over $\Sigma_{\text{FO}(k+1)}$). The idea is that $M$ has (a) copies of all the structures in $\text{Pos}$ and $\text{Neg}$, (b) a set of elements $i$, one for each $i \in [\text{Pos} \sqcup \text{Neg}]$, which represent structure identifiers, (c) unary relations $\text{Id}$ and $P$, where $\text{Id}$ holds for the set of structure identifiers $i$ and $P$ holds for the set of identifiers for structures in $\text{Pos}$, and (d) a binary relation $\text{Owns}$ that associates each $i$ with the elements of the copy of structure $i$ in $M$. The grammar $G'$ is designed to generate only formulas of the form $\forall i. \text{Id}(i) \rightarrow (P(i) \leftrightarrow \alpha'(i))$. The formula $\alpha'$ is obtained by taking a formula $\alpha$ admitted by $G$ and relativizing the quantification so that it is restricted to those elements that are associated to $i$. For example, $\forall x. \beta(x)$ is relativized to $\forall x. \text{Owns}(i, x) \rightarrow \beta(x)$ and $\exists x. \beta(x)$ is relativized to $\exists x. \text{Owns}(i, x) \land \beta(x)$. A formula $\varphi' \in L(G')$ is true in $M$ if and only if there is a formula $\varphi \in L(G)$ that is a separator for $\text{Pos}$ and $\text{Neg}$. Note that the formulas in $G'$ use one extra variable (namely $i$).

This reduction combined with Theorem 7 shows the following:

**Theorem 10.** For any fixed $k > 5$, given a single structure $M$ and a RTG $G$ over $\text{FO}(k)$, checking whether there is a formula in $L(G)$ that is true in $M$ is EXPTIME-complete.

**Open Problems.** Our algorithm has exponential dependence on the number of structures $m$. We do not know whether algorithms polynomial in $m$ can be achieved. More precisely, we do not know if separator realizability can be achieved in time $O(f(m, n, k) \cdot g(n, k))$, where $f$ is a polynomial function and $g$ is an arbitrary function. Learning algorithms that scale linearly or polynomially with the number of data samples are clearly desirable.

Interestingly, if there is no grammar restriction, i.e., we look for a separator in $\text{FO}(k)$, then such an algorithm is indeed possible. This follows from a suggestion by Victor Vianu [Vianu 2020]. The algorithm works on the basis of $\text{FO}(k)$-types [Libkin 2004], which capture equivalence classes of finite structures that cannot be distinguished from each other by any $\text{FO}(k)$ formula. Of crucial importance is the fact that these equivalence classes of structures can be defined by an $\text{FO}(k)$ formula, which can be effectively computed for a given structure. Consequently, we can independently compute the defining formula, denoted $\text{type}(A_i)$, for each $A_i \in \text{Pos}$ and then form the disjunction...
\[ \psi := \bigvee_i \text{type}(A_i). \] If \( \psi \) holds for any structure in \( \text{Neg} \) then there can be no separator. Otherwise \( \psi \) is a separator. This procedure works in time polynomial in the number of structures.

However, we do not see any way to adapt the above procedure to arbitrary grammars. Furthermore, it has a disadvantage as an algorithm for learning— it yields very large formulas that essentially overfit the positive samples. In contrast, the automata-theoretic method can find the smallest formulas.

There are, of course, many lower bound problems that are open for different logics and variants, and each of them has many parameters (\( |G|, m, n, k, k', k_1, k_2 \), as well as the arities of symbols). One can ask several parameterized complexity \([\text{Flum and Grohe 2006}]\) lower bound questions for each of our problems, and we leave this to future work. In particular, one key question involves the parameter \( k \) (which we have assumed is fixed in most of our treatment): is the double exponential dependence on \( k \) tight?

## 9 FURTHER RESULTS AND DISCUSSION

In this section we discuss how the technique illustrated in §5, §6, and §7, can be adapted to solve problems in two other settings: logic programming and second- and higher-order logics. We also remark on the generality of the approach and give a connection to Ehrenfeucht-Fraïssé games.

### 9.1 Mutual Recursion and Logic Programming

Recall that our treatment of FO-LFP from §6 did not include mutually-recursive definitions. In fact, mutual recursion can be handled with a modest increase in the number of automaton states. Consider a variant of FO-LFP that allows blocks of defined relations, in which all relations in a single block can refer to each other in their definitions, like the following:

\[
\begin{align*}
\text{let } & \left\{ P_1(x_1, x_2) \equiv_{\text{LFP}} \varphi_1(x_1, x_2, P_1, P_2) \right. \\
& \left. P_2(x_1, x_2) \equiv_{\text{LFP}} \varphi_2(x_1, x_2, P_1, P_2) \right\} \text{ in } \varphi(P_1, P_2)
\end{align*}
\]

The semantics for blocks of mutually-recursive definitions can be defined in terms of a simultaneous fixed point. For the example above, we can define functions \( F_1, F_2 : 2^X \times 2^X \to 2^X \), where \( X = \text{dom}(A)^2 \), and for \( X_1, X_2 \subseteq 2^X \):

\[
F_i(X_1, X_2) \triangleq \{ \bar{a} \in X \mid A \models \varphi_i(\bar{a}/\bar{x}, X_1/X_1, X_2/X_2) \} \quad i \in \{1, 2\}
\]

We can interpret the relations \( P_1 \) and \( P_2 \) as the components of the simultaneous least fixed point of the system of equations above; see [Fritz 2002] for more on simultaneous fixed points.

**Evaluating Mutually-Recursive Definitions.** In the spirit of our technique, we ask how an automaton can check membership for a relation defined by mutual recursion using state bounded by the structure. The same ideas carry over from §6 with a modification. As before, all tuples can be associated with the stage at which they enter the (now) simultaneous fixed point computation, and the automaton can use counters to check membership at a given stage. However, the number of stages grows exponentially in the number of relations in a block of definitions (which we can assume is bounded by the number of definable symbols \( k' \)). The automaton state must now include a product of counters, one for each definable symbol. Other than this change to the states, the construction that handles mutual recursion in FO-LFP remains essentially the same.

**Logic Programming.** With mutually-recursive definitions, our technique can be used to solve Datalog synthesis problems; this is not surprising since Datalog is logically similar to standard first-order logics with least fixed points (in fact, it corresponds to an existential fragment \( \exists \text{LFP} \) of first-order logic with least fixed points that only allows negation on atomic relations from the
signature and disallows universal quantification [Libkin 2004]). See the appendix for more details on a Datalog synthesis problem that our technique can solve. (We note for a fixed number of variables and definable relations, the space of Datalog programs is finite and thus decidability is not theoretically interesting.) We can also model problems from inductive logic programming (ILP) [Muggleton and de Raedt 1994], e.g., learning from entailment over bounded variable horn-clause programs, by encoding background knowledge (a set of definite horn clauses) in the grammar.

9.2 Second-Order Logic
The approach naturally extends to second-order logic (SO) (see, e.g., [Libkin 2004] for syntax and semantics). We state here a result for relational SO\((k, k')\), a version of SO restricted to \(k\) first-order variables and \(k'\) second-order relation variables. An alternating one-way automaton \(A\) can evaluate SO\((k, k')\) formulas on a fixed structure \(A\) by keeping track of an assignment to \(k\) first-order variables and an assignment to \(k'\) second-order relation variables of maximum arity \(n\) (the relation variables map to sets of \(r\)-tuples) using a state space of size \(O(2^{(k' n') n^k})\). The decision procedure follows the same lines as before.

**Theorem 11.** SO\((k, k')\)-separator realizability and synthesis is decidable in 2EXPTIME for a fixed signature and fixed \(k, k' \in \mathbb{N}\).

The same idea sketched above easily extends to logics with variables over higher-order functions.

9.3 Discussion
We believe the tree automata-theoretic approach proposed in this work is extremely versatile. The crux is to build automata that, when reading the parse tree of an expression, can evaluate it on a fixed structure using finitely many states. This typically is true if there is a way to recursively evaluate the semantics of expressions using memory that depends on the size of the structure but not on the size of expressions. Bounding the number of variables is one way to achieve this.

We claim our technique applies to any logic or language for which (a) the semantics of expressions can be described locally in the parse tree in terms of the semantics for subexpressions and (b) evaluating the semantics at each node of the parse tree requires memory that is bounded by a function of the structure size (and not the formula size). The fact that logics with definitions and recursion can be captured with two-way tree automata shows that they also meet these conditions, since the automaton can be converted to a deterministic bottom-up automaton. We leave formalizing this claim, proving it, and finding further instantiations of the technique to future work.

Finally, we note that the separability problems considered here can be viewed from the perspective of Ehrenfeucht-Fraïssé games [Ehrenfeucht 1961; Fraïssé 1953], which are typically used to show formulas in a logic can or cannot distinguish between two structures. The separability problem instead asks whether a set of positive structures can be separated from a set of negative structures using formulas that conform to a given grammar. Hence the game in our setting is one that is specific to the given grammar and furthermore forces the players to play simultaneously on all the structures. We leave further investigation of this relationship to future work.

10 RELATED WORK
*Program Synthesis from Examples.* Learning logical formulas is closely related to program synthesis, and especially, program synthesis from examples (as opposed to deductive approaches from specifications [Manna and Waldinger 1980]). Synthesis from examples, or *programming by examples*
(PBE), has been active in recent years and has seen successes in practice (e.g., [Polozov and Gulwani 2015]). In PBE, the goal is to synthesize a program consistent with a set of input-output examples; several domains have been explored, e.g., synthesis of database queries from examples [Shen et al. 2014; Thakkar et al. 2021; Wang et al. 2017a] and from analysis of database-backed application code [Cheung et al. 2013], synthesis of data completion scripts [Wang et al. 2017c], data structure transformations [Feser et al. 2015], and typed functional programs [Osera and Zdancewic 2015; Polikarpova et al. 2016]. A common approach to PBE involves version space algebra [Mitchell 1982], where the idea is to capture the set of all programs that work on each example in a compact representation and then intersect the sets for each example to represent programs consistent with all examples (e.g., see [Gulwani 2011]). Our approach essentially uses tree automata as a version space algebra to capture all logical expressions that satisfy some criterion over input structures.

Synthesis with Grammar. Using grammar to constrain the hypothesis space follows a line of work in program synthesis that uses syntactic biases like partial programs, e.g. [Solar-Lezama et al. 2006], and more broadly, syntax-guided synthesis (SyGuS) for logics (typically logics supported by SMT theories) [Alur et al. 2015]. In a SyGuS problem, one is given a grammar from which to synthesize a logical expression (similar to the setting in this paper) as well as a specification in the form of a universally-quantified formula that refers to a placeholder \( e \), which must be valid when the synthesized expression is plugged in for \( e \). The separability problem can in fact be formulated as a SyGuS problem, though SyGuS divisions and tools only support synthesis of quantifier-free formulas. There is a large body of work exploring program synthesis and syntax-guided synthesis for quantifier-free logics that focuses on practical and scalable techniques, and for the most part does not offer any guarantee of completeness. When grammars admit infinitely many expressions, these solvers cannot report unrealizability, which is in general undecidable [Caulfield et al. 2015].

The ability to decide realizability and synthesis is a crucial difference in our work.

Decidable Realizability and Synthesis. For systems and programs that have finite state spaces, the realizability problem has been extremely well studied and a rich class of specifications for such systems is known to admit decidable realizability. The crux of the techniques used in this domain rely on tree automata that work on infinite trees and infinite games played on finite graphs (while our work uses tree automata on finite trees). This problem was first proposed by Church [Church 1960], and a rich theory of realizability/synthesis has emerged [Buchi and Landweber 1969; Grädel et al. 2002; Kupferman et al. 2000, 2010; Madhusudan and Thiagarajan 2001; Pnueli and Rosner 1989, 1990; Rabin 1972]. The key idea is to encode the branching behavior of a reactive system using an infinite tree and build automata that accept systems (trees) whose behaviors satisfy a specification.

Our work is technically closer to the approach in [Madhusudan 2011], which studies synthesizing imperative reactive programs over a finite number of variables ranging over finite domains with logical specifications (e.g., linear temporal logic). The decidability of realizability/synthesis is proved using tree automata that work on finite trees (parse trees of programs), similar to the work presented here. Unlike our work, the tree automata have infinitary acceptance conditions in order to capture properties of infinite executions of programs. Other differences include (1) our work interprets logical expressions over unbounded structures, and (2) the specification for synthesis is not a logical formula, but rather a set of labeled structures. Intuitively, we trade the power of logical specifications in [Madhusudan 2011] and replace it with a finite set of structures in order to synthesize over unbounded domains. Though the constructions in our work are too large to implement naively, the core idea to use tree automata on parse trees of expressions for synthesis has been made practical in some recent work, e.g., for string and matrix transformations [Wang et al. 2017b] and string encoders/decoders and comparators [Wang et al. 2018].
Decidability results for synthesis of expressions over unbounded data domains are uncommon, though there are some recent results for restricted classes of programs and models of computation, e.g., synthesizing finite-state transducers [Khalimov et al. 2018] and synthesizing a restricted class of imperative programs [Krogmeier et al. 2020]. In [Krogmeier et al. 2020], the authors study the problem of synthesizing uninterpreted imperative programs from a given grammar, where programs come with assertions that must be satisfied for any interpretation of function and relation symbols over any domain, possibly infinite. For the restricted subclass of coherent programs [Mathur et al. 2019], there is a decision procedure based on tree automaton emptiness, and, similar to our work, the solution uses tree automata working over parse trees. There is also recent work giving sound techniques for proving unrealizability of SyGuS problems [Hu et al. 2019], and, more recently, a decision procedure for SyGuS problems over linear integer arithmetic with conditionals over finitely-many examples [Hu et al. 2020].

**Learning Logical Formulas.** In [Koenig et al. 2020], the authors study a separability problem for first-order logic formulas with bounded quantifier depth. In contrast to the problems we consider in this work, bounding the quantifier depth makes the search space finite up to logical equivalence, enabling a reduction to and from SAT. There is also work on the decidability of learning separators from labeled examples for various description logics [Funk et al. 2019; Jung et al. 2020]. There, separation problems are studied in the presence of an ontology, which is a finite set of logical sentences. The presence of ontologies makes the problem different from our work; adapting our general synthesis approach to the world of description logics remains future work. There is also prior work studying the complexity of learning logical concepts by characterizing the VC-dimension of logical hypothesis classes [Grohe and Turán 2004], work on parameterized complexity for logical separation problems in the PAC model [van Bergerem et al. 2021], learning MSO-definable concepts on strings [Grohe et al. 2017] and concepts definable in first-order logic with counting [van Bergerem 2019], learning temporal logic formulas from examples [Neider and Gavran 2018], and learning quantified invariants for arrays [Garg et al. 2015].

**Inductive Logic Programming.** In inductive logic programming (ILP) [Muggleton and de Raedt 1994], the goal is to learn a logic program from data, typically positive and negative examples of a target relation. ILP systems can learn from a small number of examples and with background knowledge (e.g., a set of horn clauses), and some systems are able to invent new predicates and learn programs with recursion [Cropper et al. 2020]. Typically, ILP systems learn Prolog programs, but recent work has explored learning in restricted hypothesis spaces for logic programs, e.g., Datalog [Albarghouthi et al. 2017; Evans and Grefenstette 2018] and answer set programming [Law et al. 2014]. As discussed in §9.1, our approach can be used to model some forms of ILP by encoding background knowledge in the grammar, and it seems possible that aspects of metarules [Muggleton et al. 2014] can also be achieved with our technique; exploring connections to ILP is an interesting avenue for future work.

## 11 Conclusion

We have argued for a very general tree automata-theoretic approach to learning logical formulas and, more generally, any expression which can be evaluated using state dependent on a background structure but independent of the expression size. This is the case for the finite variable logics studied in this work, as well as higher-order logics and logics with fixed point operators over finite structures. Precisely characterizing the power of this approach is an interesting direction for future work, and so too are the lower bounds and parameterized complexity questions we leave open.

What is nice about the tree automaton-based approach advocated here is that various infinite concept spaces constrained by a grammar can be seen to have finitely-many equivalence classes
modulo example structures. Indeed, the states of the (minimal) automaton correspond to equivalence classes of formulas from the grammar that are equivalent with respect to the given input structures. Effective emptiness checking algorithms for tree automata show that we only need to keep a single representative from each equivalence class to solve synthesis. Exploring practical algorithms for restricted grammars and classes of structures, including learning in the presence of background theories (such as arithmetic, used say for counting), are intriguing directions for future work.

ACKNOWLEDGMENTS

We thank Victor Vianu for discussions and for suggesting the connection to logical types. This work was supported in part by a Discovery Partner’s Institute (DPI) science team seed grant and a research grant from Amazon.

REFERENCES

Aws Albarghouthi, Paraschos Koutris, Mayur Naik, and Calvin Smith. 2017. Constraint-Based Synthesis of Datalog Programs. In Principles and Practice of Constraint Programming, J. Christopher Beck (Ed.). Springer International Publishing, Cham, 689–706. https://doi.org/10.1007/978-3-319-66158-2_44

Rajeev Alur, Rastislav Bodik, Eric Dallal, Dana Fisman, Pranav Garg, Garavit Juniwal, Hadas Kress-Gazit, P. Madhusudan, Milo M. K. Martin, Mukund Raghothaman, Shambwadiya Saha, Sanjit A. Seshia, Rishabh Singh, Armando Solar-Lezama, Emina Torlak, and Abhishek Udupa. 2015. Syntax-Guided Synthesis. In Dependable Software Systems Engineering. NATO Science for Peace and Security Series, D: Information and Communication Security, Vol. 40. IOS Press, 1–25.

Roderick Bloem, Barbara Jobstmann, Nir Piterman, Amir Pnueli, and Yaniv Sa’ar. 2012. Synthesis of Reactive(1) designs. J. Comput. System Sci. 78, 3 (2012), 911–938. https://doi.org/10.1016/j.jcss.2011.08.007

J. Richard Buchi and Lawrence H. Landweber. 1969. Solving Sequential Conditions by Finite-State Strategies. Trans. Amer. Math. Soc. 138 (1969), 295–311. http://www.jstor.org/stable/1994916

Benjamin Caulfield, Markus N. Rabe, Sanjit A. Seshia, and Stavros Tripakis. 2015. What’s Decidable about Syntax-Guided Synthesis? CoRR abs/1510.08393 (2015). http://arxiv.org/abs/1510.08393

Alvin Cheung, Armando Solar-Lezama, and Samuel Madden. 2013. Optimizing Database-Backed Applications with Query Synthesis. In Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation (Seattle, Washington, USA) (PLDI ’13). Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/2491956.2462180

Alonzo Church. 1960. Application of recursive arithmetic to the problem of circuit synthesis. Summaries of talks presented at the Summer Institute for Symbolic Logic Cornell University, 1957, 2nd edn., Journal of Symbolic Logic 28, 4 (1960), 30–50. 3a–45a.

H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. 2007. Tree Automata Techniques and Applications. Available on: http://www.grappa.univ-lille3.fr/tata. release October, 12th 2007.

Andrew Cropper, Sebastijan Dumancic, and Stephen H. Muggleton. 2020. Turning 30: New Ideas in Inductive Logic Programming. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, Christian Bessiere (Ed.). ijcai.org, 4833–4839. https://doi.org/10.24963/ijcai.2020/673

Ugo Dal Lago. 2012. A Short Introduction to Implicit Computational Complexity. Springer Berlin Heidelberg, Berlin, Heidelberg, 89–109. https://doi.org/10.1007/978-3-642-31485-8_3

Andrzej Ehrenfeucht. 1961. An application of games to the completeness problem for formalized theories. Fundamenta Mathematicae 49, 2 (1961), 129–141. http://eudml.org/doc/213582

Herbert B. Enderton. 2001. Chapter 2 - First-Order Logic. In A Mathematical Introduction to Logic (second edition ed.). Herbert B. Enderton (Ed.). Academic Press, Boston, 67–181. https://doi.org/10.1016/B978-0-08-049646-7.50008-4

Richard Evans and Edward Grefenstette. 2018. Learning Explanatory Rules from Noisy Data. J. Artif. Int. Res. 61, 1 (Jan. 2018), 1–64. https://dl.acm.org/doi/10.5555/3241691.3241692

John K. Feser, Swarat Chaudhuri, and Isil Dillig. 2015. Synthesizing Data Structure Transformations from Input-Output Examples. In Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (Portland, OR, USA) (PLDI ’15). Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/2737924.2737977

J. Flum and M. Grohe. 2006. Parameterized Complexity Theory (Texts in Theoretical Computer Science. An EATCS Series). Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/3-540-29953-X

Proc. ACM Program. Lang., Vol. 6, No. POPL, Article 10. Publication date: January 2022.
10:28 P. Krogmeier, P. Madhusudan

https://doi.org/10.4230/LIPIcs.CSL.2011.428
P. Madhusudan and P. S. Thiagarajan. 2001. Distributed Controller Synthesis for Local Specifications. In ICALP (Lecture Notes in Computer Science, Vol. 2076). Springer, 396–407. https://doi.org/10.1007/3-540-48224-5_33
Zohar Manna and Richard Waldinger. 1980. A Deductive Approach to Program Synthesis. ACM Trans. Program. Lang. Syst. 2, 1 (Jan. 1980), 90–121. https://doi.org/10.1145/357084.357090
Umang Mathur, P. Madhusudan, and Mahesh Viswanathan. 2019. Decidable Verification of Uninterpreted Programs. Proc. ACM Program. Lang. 3, POPL, Article 46 (Jan. 2019), 29 pages. https://doi.org/10.1145/3290359
Tom M. Mitchell. 1982. Generalization as search. Artificial Intelligence 18, 2 (1982), 203–226. https://doi.org/10.1016/0004-3702(82)90040-6
Thomas M. Mitchell. 1997. Machine Learning (1 ed.). McGraw-Hill, Inc., USA.
Stephen Muggleton and Luc de Raedt. 1994. Inductive Logic Programming: Theory and methods. The Journal of Logic Programming 19-20 (1994), 629–679. https://doi.org/10.1016/0743-1066(94)90035-3
Stephen H. Muggleton, Dianhuan Lin, Niels Pahlavi, and Alireza Tamaddoni-Nezhad. 2014. Meta-interpretive learning: application to grammatical inference. Machine Learning 94, 1 (01 Jan 2014), 25–49. https://doi.org/10.1007/s10994-013-5358-3
Daniel Neider and Ivan Gavran. 2018. Learning Linear Temporal Properties. In 2018 Formal Methods in Computer Aided Design (FMCAD). 1–10. https://doi.org/10.23919/FMCAD.2018.8603016
Peter-Michael Osera and Steve Zdancewic. 2015. Type-and-Example-Directed Program Synthesis. In Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (Portland, OR, USA) (PLDI ’15). Association for Computing Machinery, New York, NY, USA, 619–630. https://doi.org/10.1145/2737924.2738007
Amir Pnueli and Roni Rosner. 1989. On the Synthesis of a Reactive Module. In POPL. ACM Press, 179–190. https://doi.org/10.1145/75277.75293
Amir Pnueli and Roni Rosner. 1990. Distributed Reactive Systems Are Hard to Synthesize. In FOCS. IEEE Computer Society, 746–757. https://doi.org/10.1109/SFCS.1990.89597
Nadia Polikarpova, Ivan Kuraj, and Armando Solar-Lezama. 2016. Program Synthesis from Polymorphic Refinement Types. In Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation (Santa Barbara, CA, USA) (PLDI ’16). Association for Computing Machinery, New York, NY, USA, 522–538. https://doi.org/10.1145/2908080.2908093
Oleksandr Polozov and Sumit Gulwani. 2015. FlashMeta: A Framework for Inductive Program Synthesis. In Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (Pittsburgh, PA, USA) (OOPSLA 2015). Association for Computing Machinery, New York, NY, USA, 107–126. https://doi.org/10.1145/2814270.2814310
Michael Oser Rabin. 1972. Automata on Infinite Objects and Church’s Problem. American Mathematical Society, Boston, MA, USA. https://dl.acm.org/doi/10.5555/540412
Yanyan Shen, Kaushik Chakrabarti, Surajit Chaudhuri, Bolin Ding, and Lev Novik. 2014. Discovering Queries Based on Example Tuples. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (Snowbird, Utah, USA) (SIGMOD ’14). Association for Computing Machinery, New York, NY, USA, 493–504. https://doi.org/10.1145/2588555.2596664
Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Sanjit Seshia, and Vijay Saraswat. 2006. Combinatorial Sketching for Finite Programs. In Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (San Jose, California, USA) (ASPLOS XII). Association for Computing Machinery, New York, NY, USA, 404–415. https://doi.org/10.1145/1168577.1168907
Aalok Thakkar, Aaditya Naik, Nathaniel Sands, Rajeev Alur, Mayur Naik, and Mukund Raghothaman. 2021. Example-Guided Synthesis of Relational Queries. In Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (Virtual, Canada) (PLDI 2021). Association for Computing Machinery, New York, NY, USA, 1110–1125. https://doi.org/10.1145/3453483.3454098
Steffen van Bergerem. 2019. Learning Concepts Definable in First-Order Logic with Counting. In 2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS). 1–13. https://doi.org/10.1109/LICS.2019.8785811
Steffen van Bergerem, Martin Grohe, and Martin Ritzert. 2021. On the Parameterized Complexity of Learning Logic. CoRR abs/2102.12201 (2021). arXiv:2102.12201
Moshe Y. Vardi. 1982. The Complexity of Relational Query Languages (Extended Abstract). In Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing (San Francisco, California, USA) (STOC ’82). Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/800070.802186
Moshe Y. Vardi. 1998. Reasoning about the past with two-way automata. In Automata, Languages and Programming, Kim G. Larsen, Sven Skyum, and Glynn Winskel (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 628–641. https://doi.org/10.1007/BFb0550590
Victor Vianu. 2020. Personal communication.

Proc. ACM Program. Lang., Vol. 6, No. POPL, Article 10. Publication date: January 2022.
Chenglong Wang, Alvin Cheung, and Rastislav Bodik. 2017a. Synthesizing Highly Expressive SQL Queries from Input-Output Examples. In Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation (Barcelona, Spain) (PLDI 2017). Association for Computing Machinery, New York, NY, USA, 452–466. https://doi.org/10.1145/3062341.3062365

Xinyu Wang, Isil Dillig, and Rishabh Singh. 2017b. Program Synthesis Using Abstraction Refinement. Proc. ACM Program. Lang. 2, POPL, Article 63 (Dec. 2017), 30 pages. https://doi.org/10.1145/3158151

Xinyu Wang, Isil Dillig, and Rishabh Singh. 2017c. Synthesis of Data Completion Scripts Using Finite Tree Automata. Proc. ACM Program. Lang. 1, OOPSLA, Article 62 (Oct. 2017), 26 pages. https://doi.org/10.1145/3133886

Yuepeng Wang, Xinyu Wang, and Isil Dillig. 2018. Relational Program Synthesis. Proc. ACM Program. Lang. 2, OOPSLA, Article 155 (Oct. 2018), 27 pages. https://doi.org/10.1145/3276525
A TWO-WAY TREE AUTOMATA

A.0.1 Two-way tree automata. Two-way (alternating) tree automata over $\kappa$-ary trees generalize ATAs by enabling them to transition not only down into children but also up to the parent on the input tree. This extra capability is formally captured by generalizing the transition function from the form $\delta : Q \times \Sigma \rightarrow \mathbb{B}^+(Q \times \{1, \ldots, \kappa\})$ to the form $\delta : Q \times \Sigma \rightarrow \mathbb{B}^+(Q \times \{-1, \ldots, \kappa\})$, where $-1$ means ascending to a parent node and 0 means staying at the current node. In the context of two-way automata only, we use $[\kappa]$ to denote the set $\{-1, \ldots, \kappa\}$ for $\kappa \in \mathbb{N}$. For each $(q, a) \in Q \times \Sigma$ we require $\delta(q, a) \in \mathbb{B}^+(Q \times \text{arity}(a))$. For $x \in \mathbb{N}^*$, $j \in \mathbb{N}$, let $(x \cdot j) \cdot -1 = x$ and $(x \cdot j) \cdot 0 = x \cdot j$ and let $e \cdot -1$ be undefined.

A two-way tree automaton is a tuple $\mathcal{A} = \langle Q, \Sigma, I, \delta, F \rangle$, where the only difference with ATAs is the acceptance condition $F \subseteq Q$ and the definition of $\delta$ as described above. The notion of a run in the two-way case is identical to the definition for one-way alternating automata from §3.5, modulo the difference in $\delta$. Nevertheless, we repeat the definition here and encourage the reader to observe that the nodes of a run may be labeled by positions in the input tree that go both up and down. A run for a two-way tree automaton $\mathcal{A} = \langle Q, \Sigma, I, \delta, F \rangle$ on a tree $t \in T_\Sigma$ is an ordered tree $\rho$ over $Q \times \text{Nodes}(t)$ satisfying the following two conditions:

- $\rho(e) = (q_i, e)$ for some state $q_i \in I$
- Let $p \in \text{Nodes}(\rho)$. If $\rho(p) = (q, x)$ with $t(x) = a$, then there is a subset $S = \{(q_1, i_1), \ldots, (q_l, i_l)\} \subseteq Q \times \text{arity}(a)$ such that $S \models \delta(q, a)$ and $\rho(p \cdot j) = (q_j, x \cdot i_j)$ for $1 \leq j \leq l$.

Observe that a run for a two-way automaton may be infinite on a finite input tree. We want only finite runs to be accepting, rather than simply requiring the existence of a run, as we did for ATAs. This corresponds to a reachability acceptance condition, wherein a run is accepting if every branch reaches some state $q_f \in F$, and a two-way automaton accepts a tree if it has an accepting run on it.

Note that the two-way tree automata we use here are no more expressive than alternating tree automata, and there are algorithms to convert a two-way automaton to a one-way automaton [Vardi 1998], and thus membership and emptiness are decidable. We next sketch this conversion.

B TWO-WAY TREE AUTOMATA TO ONE-WAY (§6.2.2)

We can convert the two-way automaton into a language-equivalent nondeterministic automaton by adapting the technique of [Vardi 1998]. The key idea is to view the membership problem of the two-way automaton as a finite reachability game, where a Protagonist is trying to show the automaton has an accepting run and the Antagonist is trying to refute this. Game positions are pairs of automaton states and positions on the input tree. From the position $(q, x) \in Q \times \text{Nodes}(t)$, with $t(x) = a \in \Sigma$, the Protagonist picks a move $s \subseteq Q \times \text{arity}(a)$ and the Antagonist responds by picking $s' \in S$, with the game continuing from the position indicated by $s$. Play begins from the state $(q_i, e)$. The Protagonist wins if she has a winning strategy to reach game positions of the form $(q_f, x)$ for $x \in \text{Nodes}(t)$, otherwise the Antagonist wins. These notions are standard for finite reachability games, and we refer the reader to [Fritz 2002] for details.

We can define a top-down nondeterministic automaton (without alternation) that reads trees annotated with winning strategy information. Since the game is a finite reachability game, it is determined with memoryless strategies (see [Grädel et al. 2002]). That is, each position in the game is won by a single player using a strategy that does not depend on the preceding history of plays. It follows that the strategy annotations can be represented with a finite alphabet. There are now three related challenges: (1) the strategy must be verified to comply with the transition function, (2) all plays that could result from the strategy must be winning, and (3) these must both be accomplished.
in a single downward pass over the annotated input tree. The final nondeterministic automaton over (unannotated) logical formulas is obtained by projecting out the annotation, appealing to closure of tree regular languages under homomorphisms.

The strategy annotation for a given node $x \in \text{Nodes}(t)$ in an input tree $t$ must decide which states to go into and in what directions on $t$ to go from each possible game position at $x$. The strategy annotation associates to each node $x$ of $t$ a member of $2^{Q\times|\text{arity}(t(x))| \times Q}$. At a game position $(q, x)$, the strategy for the Protagonist corresponds to a subset $X = \{(d, q') \mid (q, d, q') \subseteq 2^{Q\times|\text{arity}(t(x))| \times Q}$, elements of which indicate a direction on the input tree and a next state. Thus the final nondeterministic automaton has number of states exponential in the number of states for the two-way automaton, and this conversion can be done in time exponential in the size of the two-way automaton.

C FO-LFP AUTOMATON TRANSITIONS (§6.2.2)

Recall we have fixed a structure $A$ and we are defining the transitions for a two-way automaton $\mathcal{A}_A$ that accepts the set of FO-LFP($k, k'$) formulas that are true in $A$. Recall the state space for this automaton is defined as:

$$Q := \text{Dual}(\text{Assign}) \times \text{Count} \times \text{Defn} \cup \text{Dual}(\text{Val}) \times \text{Count} \times \text{Defn} \cup \{q_f\},$$

In the rules below, $y$ ranges over $\text{Assign} := V \rightarrow \text{dom}(A)$, $d$ ranges over $\text{Dual}(\text{Val})$, $a$ ranges over $\text{Val}$, the counter value $j$ ranges over $\text{Count} := \{0, \ldots, n'\}$, and $Y$ ranges over $\text{Defn} := \{\bot, P_1, \ldots, P_{k'}\}$. All transitions not covered below are False.

- $\delta((y, j, Y), \land) = ((y, j, Y), 1) \land ((y, j, Y), 2)$
- $\delta((y, j, Y), \lor) = ((y, j, Y), 1) \lor ((y, j, Y), 2)$
- $\delta((y, j, Y), \forall x) = \land_{a \in \text{dom}(A)}((y[x \mapsto a], j, Y), 1)$
- $\delta((y, j, Y), \exists x) = \lor_{a \in \text{dom}(A)}((y[x \mapsto a], j, Y), 1)$
- $\delta((y, j, Y), R(x)) = \begin{cases} \text{True} & y(x) \downarrow \text{ and } A, y \models R(x) \\ \text{False} & \text{otherwise} \end{cases}$
- $\delta((\tilde{y}, j, Y), \neg) = ((\tilde{y}, j, Y), 1)$
- $\delta((\tilde{y}, j, Y), \land) = ((\tilde{y}, j, Y), 1) \land ((\tilde{y}, j, Y), 2)$
- $\delta((\tilde{y}, j, Y), \lor) = ((\tilde{y}, j, Y), 1) \lor ((\tilde{y}, j, Y), 2)$
- $\delta((\tilde{y}, j, Y), \forall x) = \land_{a \in \text{dom}(A)}((\tilde{y}[x \mapsto a], j, Y), 1)$
- $\delta((\tilde{y}, j, Y), \exists x) = \lor_{a \in \text{dom}(A)}((\tilde{y}[x \mapsto a], j, Y), 1)$
- $\delta((\tilde{y}, j, Y), R(x)) = \begin{cases} \text{True} & y(x) \downarrow \text{ and } A, y \not\models R(x) \\ \text{False} & \text{otherwise} \end{cases}$
- $\delta((d, j, Y), \land) = ((d, j, Y), -1)$
- $\delta((d, j, Y), \lor) = ((d, j, Y), -1)$
- $\delta((d, j, Y), \forall x) = ((d, j, Y), -1)$
- $\delta((d, j, Y), \exists x) = ((d, j, Y), -1)$
- $\delta((d, j, Y), \neg) = ((d, j, Y), -1)$
- $\delta((y, j, Y), \text{let } P_i(\tilde{x}) = ((y, j, Y), 2)$
- $\delta((y, j, Y), P_i(\tilde{x}) = ((a, n', P_i), -1) \text{ for } Y \neq P_i, y(\tilde{x}) \downarrow, \text{ and } a = y(\tilde{x})$
- $\delta((\tilde{y}, j, Y), P_i(\tilde{x}) = ((\tilde{a}, n', P_i), -1) \text{ for } Y \neq P_i, y(\tilde{x}) \downarrow, \text{ and } a = y(\tilde{x})$
- $\delta((d, j, Y), \text{let } P_i(\tilde{x}) = ((d, j, Y), -1) \text{ for } Y \neq P_i$
- $\delta((a, j, P_i), \text{let } P_i(\tilde{x}) = ((y, j - 1, P_i), 1) \text{ for } j > 0 \text{ and } y = \cup_i \{x_i \mapsto a_i\}$
- $\delta((\tilde{y}, 0, P_i), P_i(\tilde{x}) = \text{True}$. 

Proc. ACM Program. Lang., Vol. 6, No. POPL, Article 10. Publication date: January 2022.
while it checks that (for every run) some path does not reach the final state. If the count reaches

The finite height of FO

10:32 P. Krogmeier, P. Madhusudan

The semantics for recursively-defined functions is as follows. Each defined function is witnessed within the first stages of the computation. For

assign corresponding to the defined term

license to stop when

Here we give more details from §

obvious that such a witness can be finite itself. However, any tuple must enter the computation from a down state on input trees. The key for correctness is that, when the automaton is entering a definition body assigning for each definition. However, in general, care is needed to ensure that the least fixed point is total

with regard to verifying true membership in a defined relation, we can show by induction that the automaton always has a run that reaches the final state q_f along each branch. For non-membership, on the other hand, the witness is that there is no run witnessing the opposite, i.e., that for every run there is some branch that never reaches the final state. It may not be immediately obvious that such a witness can be finite itself. However, any tuple must enter the computation within n^r iterations. Thus it is sufficient to force the automaton to count down from this bound while it checks that (for every run) some path does not reach the final state. If the count reaches zero then there can be no run witnessing membership.

It is more tedious to describe the language of the automaton at every state (as was simple to do for the automaton without definitions and least fixed points), given that it moves up and down on input trees. The key for correctness is that, when the automaton is entering a definition body from a down state \langle assign, j, defn \rangle, it checks that the membership of the tuple of domain elements corresponding to assign is witnessed within the first j stages of the least fixed point computation for the definition. The dual case for checking non-membership in the least fixed point is similar: the automaton checks that membership is not witnessed in the first j stages of the computation. The finite height of n^r of the powerset lattice for definitions of arity r and structure size n gives the license to stop when j is 0 and either allow the automaton to accept (if checking non-membership) or reject (if checking membership).

D TERM SYNTHESIS DETAILS

Here we give more details from §7: a ranked alphabet for the language FO-LFP-FUN, the semantics for FO-LFP-FUN, and the automaton transitions for reading terms in FO-LFP-FUN.

D.1 Ranked Alphabet for FO-LFP-FUN(k, k')

A ranked alphabet for FO-LFP-FUN(k, k') with definable functions from F = \{g_1, \ldots, g_k\}:

\[ \Sigma_{\text{FO-LFP-FUN}(k, k')} = \left\{ \text{let } g(\bar{x})^2, \mid g \in F, \bar{x} \in \text{arity}(g) \right\} \cup \left\{ g^{\text{arity}(g)} \right\} \cup \Sigma_{\text{FO-LFP}(k, k' - k_2)} \]

D.2 Semantics for FO-LFP-FUN

The semantics for recursively-defined functions is as follows. Each defined function g of arity d is interpreted as a partial function g^A : dom(A)^d \rightharpoonup dom(A), which is a member of the bottomed partial order \[ \hat{O} = \langle \text{dom}(A)^d \rightharpoonup \text{dom}(a), \sqsubseteq, \bot \rangle, \] where \bot is undefined everywhere and \forall \in \text{arity} holds if for all \bar{a} \in \text{dom}(A)^d, whenever \bar{f}(\bar{a}) \downarrow, then \bar{f}'(\bar{a}) \downarrow and \bar{f}(\bar{a}) = \bar{f}'(\bar{a}). This partial order has finite height since we work with finite structures. Now, suppose a function g of arity d is defined recursively using a term t(x_1, \ldots, x_d, g). We associate a monotone function F_t : \hat{O} \rightharpoonup \hat{O} to the defining term t, and let g^A be the least fixed point of F_t, which can easily be shown to exist and to be equal to the stable point of the finite chain \bot \sqsubseteq F_t(\bot) \sqsubseteq \cdots \sqsubseteq F^{i}_{t}(\bot) = F^{i+1}_{t}(\bot), with i \leq |\text{dom}(A)^d|). The syntax and semantics for terms in FO-LFP-FUN guarantees monotonicity of the function F_t for any term t, though we do not prove it here. It follows that least fixed points exist for each definition. However, in general, care is needed to ensure that the least fixed point is total on \text{dom}(A), and whether or not this is so depends on the definition and the structure A. Terms and formulas for FO-LFP-FUN are interpreted in a 3-valued logic with interpretation functions
\[\varphi\]_{A,Y,D} and \[t\]_{A,Y,D}, where \(A\) is a finite structure, \(Y\) is a partial variable assignment, and \(D\) maps definable symbols to the terms or formulas that define them. The semantics is given in Figure 10.

Partial functions disrupt the usual semantics of first-order logic because we want to define entailment even when there are undefined terms. As mentioned, we handle this with a 3-valued logic over \{True, False, \bot\}, where \(\bot\) means undefined. Undefinedness propagates across the various formation rules of the logic as one might expect. If the classical truth value of a formula cannot be determined based on the values of its subformulas, then it has an undefined value. For example, the formula \(\varphi_0 \land \varphi_1\) is undefined in a structure if \(\varphi_1\) has value 1 and \(\varphi_1 \neg 1\) is undefined (\(i \in \{0, 1\}\)). (See Figure 10 for the semantics.)

Aside: We note that another possibility for the semantics, which would allow us to avoid a 3-valued logic, is to interpret recursive functions as total functions on a lattice, which could work as follows. Given an input structure \(A\), the automaton works over a modified structure \(Lat(A)\), which equips \(A\) with a lattice structure by introducing new elements \(\bot, \top\) and putting \(x \leq_{Lat(A)} \top\) and \(\bot \leq x\) for all \(x \in dom(A)\). Recursively-defined functions are then interpreted over a partial order of functions \(F = \langle Lat(A) \to Lat(A), \subseteq\rangle\), with \(f_1 \subseteq f_2\) if \(f_1(x) \leq_{Lat(A)} f_2(x)\) for all \(x \in Lat(A)\) and \(f_1, f_2 \in F\). If a definable function symbol \(g\) is defined using term \(t\), we can associate a function \(\mathcal{F}_t : F \to F\) and interpret \(g\) as the least fixed point of \(\mathcal{F}_t\). The semantics of formulas is modified so that quantification is defined only over elements in \(dom(A)\). Furthermore, partial functions (e.g., \text{car} from the merge example) can be modeled in \(Lat(A)\) as total functions that use \(\bot\) in the obvious way to model undefinedness, and each function \(f\) can be extended to have \(f(\top) = \top\).

D.3 Automaton for Evaluating Terms with Recursive Functions

Fix a structure \(A\) and a domain element \(a \in dom(A)\). We want to define a two-way automaton \(\mathcal{A} = \langle Q, \Sigma, F, I, \delta, F \rangle\), that accepts the set of terms \(t \in FO-LFP-FUN(k, k')\) such that \([t]_{A,\varnothing,\varnothing} = a\). For simplicity we assume \(r\)-ary functions and relations only. Similar to the construction for FO-LFP. The state space, denoted \(Q_{TERM}\), includes the states for FO-LFP except now it also includes states that encode the element which the current term should evaluate to, as follows:

\[
Q_{TERM} := EvalForm \cup EvalTerm \cup \{q_f\}
\]

\[
EvalForm := (Dual(Assign) \cup Dual(Val)) \times Count \times Defn
\]

\[
EvalTerm := (Assign \cup Val) \times Count \times Defn \times dom(A)
\]

where \(Assign = V \rightarrow dom(A)\) is the set of partial variable assignments, \(Val = dom(A)^r\) is a set of values that are passed up to definitions of functions and relations, \(Dual(X) \triangleq \{x, \bar{x} \mid x \in X\}\) helps us define sets of symbols equipped with “dual” marked copies of their members, and \(Count = \{0, \ldots, n'\}\) and \(Defn = \{\bot, P_1, \ldots, g_{ki}\}\) serve the same purposes as before. Below, we give the main transitions that allow the automaton to evaluate terms. We use \(\star\) as a wildcard to range over any element from a set, determined by context. Here we denote \(\{1, \ldots, r\}\) with \([r]\).

Reading variables. The automaton simply checks that the variable is mapped to \(a\) in the current assignment:

\[
\delta(\langle y, \star_1, \star_2, a \rangle, x) = \begin{cases} 
True & y(x) = a \\
False & \text{otherwise}
\end{cases}
\]

Reading if-then-else terms. The automaton must either (i) verify the condition formula in the first child and verify that the term in the “then” branch (second child) evaluates to \(a\) or (ii) falsify the condition formula and verify that the term in the “else” branch (third child) evaluates to \(a\):

\[
\delta(\langle y, \star_1, \star_2, a \rangle, \text{ite}) = (\langle y, \star_1, \star_2, 1 \rangle \land (\langle y, \star_1, \star_2, a \rangle, 2) \lor (\langle y, \star_1, \star_2, 1 \rangle \land (\langle y, \star_1, \star_2, 3 \rangle, 3)
\]
where they are defined (similar to a closure). We write

\[ R(I)_{A,Y,D} = \begin{cases} 1 & \text{if } I_{A,Y,D} \downarrow \text{ and } I_{A,Y,D} \in R^A \\ 0 & \text{otherwise} \end{cases} \]

\[ \neg \psi_{A,Y,D} = \begin{cases} 1 & \text{if } \neg \psi_{A,Y,D} = 0 \\ 0 & \text{otherwise} \end{cases} \]

\[ \varphi_1 \land \varphi_2_{A,Y,D} = \begin{cases} 1 & \text{if } \varphi_1_{A,Y,D} = 1 \text{ and } \varphi_2_{A,Y,D} = 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ \varphi_1 \lor \varphi_2_{A,Y,D} = \begin{cases} 1 & \text{if } \varphi_1_{A,Y,D} = 1 \text{ or } \varphi_2_{A,Y,D} = 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ \forall x. \varphi_{A,Y,D} = \begin{cases} 1 & \text{if } \varphi_{A,Y[D[x\rightarrow a],D]} = 1 \text{ all } a \in A \\ 0 & \text{some } a \in A \\ \bot & \text{otherwise} \end{cases} \]

\[ \exists x. \varphi_{A,Y,D} = \begin{cases} 1 & \text{if } \varphi_{A,Y[D[x\rightarrow a],D]} = 0 \text{ all } a \in A \\ 0 & \text{some } a \in A \\ \bot & \text{otherwise} \end{cases} \]

\[ \text{let } P(\bar{x}) \equiv_{lp} \psi \text{ in } \varphi_{A,Y,D} = \varphi_{A,Y,D[D \rightarrow (\psi,D)]} \]

\[ P(I)_{A,Y,D} = \begin{cases} 1 & \text{if } I_{A,Y,D} \downarrow \text{ and } I_{A,Y,D} \in \text{lfp}(D(P)) \\ 0 & \text{otherwise} \end{cases} \]

Formulas

Terms

\[ x_{A,Y,D} = y(x) \]

\[ c_{A,Y,D} = c^A \]

\[ f(I)_{A,Y,D} = \begin{cases} f^A(I_{A,Y,D}) & \text{if } f^A(I_{A,Y,D}) \downarrow \text{ and } I_{A,Y,D} \in R^A \\ \bot & \text{otherwise} \end{cases} \]

\[ \text{ite}(\varphi, t_1, t_2)_{A,Y,D} = \begin{cases} t_1_{A,Y,D} & \text{if } \varphi_{A,Y,D} = 1 \\ t_2_{A,Y,D} & \text{if } \varphi_{A,Y,D} = 0 \\ \bot & \text{otherwise} \end{cases} \]

\[ \text{let } g(\bar{x}) \equiv_{lp} t \text{ in } t'_{A,Y,D} = \begin{cases} t'_{A,Y,D[D \rightarrow (t,D)]} & \text{if } \text{lfp}(D(g))(I_{A,Y,D}) \downarrow \text{ and } I_{A,Y,D} \in \text{lfp}(D(P)) \\ \bot & \text{otherwise} \end{cases} \]

Fig. 10. Semantics for FO-LFP-FUN. \( A \) is a finite structure, \( y \) is a variable assignment, \( D \) is a map from definable symbols to their defining terms or formulas, as well as the environment of definitions at the point where they are defined (similar to a closure). We write \( [I] \) to denote the tuple of interpretations of the terms in \( I \), and \( [I] \downarrow \) to denote that an interpretation of a term is defined.

**Reading function symbols.** The automaton must verify that the \( r \) argument terms evaluate to some tuple \( \langle a_1, \ldots, a_r \rangle \in \text{dom}(A)^r \) such that \( f^A(a_1, \ldots, a_r) = a \):

\[ \delta(\langle y, *, 1, *, 2, a \rangle, f) = \bigvee_{a \in \text{dom}(A)^r} \left( \bigwedge_{i \in [r]} \left( (\langle y, *, 1, *, 2, a_i \rangle, i) \right) \right) \]

**Reading relation symbols.** The automaton must verify that the \( r \) argument terms evaluate to some tuple \( \langle a_1, \ldots, a_r \rangle \in R^A \subseteq \text{dom}(A)^r \):

\[ \delta(\langle y, *, 1, *, 2, a \rangle, R) = \bigvee_{a \in R^A} \left( \bigwedge_{i \in [r]} \left( (\langle y, *, 1, *, 2, a_i \rangle, i) \right) \right) \]
**Reading a defined function** $g$. The automaton “guesses and verifies” that the argument terms for $g$ evaluate to $\bar{a}$ and ascends to the definition of $g$ to verify that it evaluates to $\bar{a}$ when applied to $\bar{a}$. Depending on the current definition it either decrements or resets the counter.

$$
\delta((\gamma, j, g', a), g) = \bigvee_{\bar{a} \in \text{dom}(A)^r} \left( \bigwedge_{i \in [r]} ((\gamma, j, g', a_i), i) \right) \land ((\bar{a}, n', g, a), -1) \quad \text{for } g' \neq g
$$

$$
\delta((\gamma, j, g, a), g) = \bigvee_{\bar{a} \in \text{dom}(A)^r} \left( \bigwedge_{i \in [r]} ((\gamma, j, g, a_i), i) \right) \land ((\bar{a}, j - 1, g, a), -1) \quad \text{for } j > 0
$$

$$
\delta((\gamma, 0, g, a), g) = \text{False}
$$

**Reading a defined relation** $P$. The automaton “guesses and verifies” that the argument terms for $P$ evaluate to $\bar{a}$ and ascends to the definition of $P$ to verify that $\bar{a}$ is a member of the defined relation. Depending on the current definition it either decrements or resets the counter.

$$
\delta((\gamma, j, P_1), P_1) = \bigvee_{\bar{a} \in \text{dom}(A)^r} \left( \bigwedge_{i \in [r]} ((\gamma, j, P_1, a_i), i) \right) \land ((\bar{a}, n', P_1), -1) \quad \text{for } P_1 \neq P_1
$$

$$
\delta((\gamma, j, P_1), P_1) = \bigvee_{\bar{a} \in \text{dom}(A)^r} \left( \bigwedge_{i \in [r]} ((\gamma, j, P_1, a_i), i) \right) \land ((\bar{a}, j - 1, P_1), -1) \quad \text{for } j > 0
$$

$$
\delta((\gamma, 0, P_1), P_1) = \text{False}
$$

The rest of the transitions are similar in spirit to those for FO-LFP. Initial states: $I = \{\langle \varnothing, 0, \bot, \bar{a} \rangle\}$. The acceptance condition is again reachability, with $F = \{q_f\}$, and an analysis of the automaton size is similar.

### E DETAILS FROM LOWER BOUNDS (§8)

We describe the important parts of the grammar $G$ for the proof of Theorem 7. A given sentence in the grammar can be viewed as a tree of constraints. Branches of the tree assert various things in order to reflect computation trees for the machine $M$. We have to make extensive reuse of variables in order to accumulate polynomially-many constraints using a fixed number of variables. One can picture a lopsided “tree of constraints” growing to one side with nested conjunctions that requantify old variables as needed in order to assert new constraints.

We use upper case words as names of nonterminals, sometimes abusing notation by indicating the free variables common to any formula in the language of the nonterminal, e.g., for a nonterminal $X$ the expression $X(y)$ simply indicates that nonterminal $X$ is being referred to, and any formula it generates has free variable $y$.

The grammar begins with start symbol $S$, which prepares some variables needed later:

$$
S ::= \exists c\ c'\ \text{turn}::= c = c' \land (\exists y.\text{Start}(y) \land \text{Init}_0(y, c, c', \text{turn}) \land \text{Game}(y, c, c', \text{turn}))
$$

The variables $c, c'$ actually stand for triples of variables $c_1, c_2, c_3$ and $c'_1, c'_2, c'_3$. These will be used to store three-cell “windows” of the machine’s tape. $\text{Start}(y)$ holds only for the unique special element, denoted $\star$ in §8, which represents the beginning of the cycle in each structure. Thus this causes $y$ to “point” to the beginning of the cycle in each structure.

$\text{Init}_0$ is a nonterminal that effects the initialization of the starting configuration by iteratively accumulating equalities between variables and tape symbols corresponding to the starting configuration. We skip this, since it is similar to later nonterminals for generating configurations.
Game($y, c, c', turn$) is the starting point for the game semantics of alternating Turing machines. Two players, Adam and Eve, take turns picking from one of two possible machine transitions. Eve must be able to pick transitions in such a way that for any strategy of Adam, the machine eventually halts in an accepting state. The Game nonterminal is as follows:

\[
\text{Game} ::= \text{Accepting}(c') \mid \text{MakeMove}(y, c, c', turn)
\]

This is the first choice available to a synthesizer. It either picks a formula from Accepting, in which case the machine should be in an accepting state, or it picks a formula from MakeMove, which accumulates constraints that simulate the game semantics.

The Accepting nonterminal is simple:

\[
\text{Accepting} ::= \text{EncodesState}(c') \rightarrow \text{AcceptState}(c')
\]

It asserts that if the window of tape symbols represented in $c'$ contains a symbol that represents a machine state, then that state is an accepting state (We don’t spell this out, but it is simple to do so.)

The other option is MakeMove:

\[
\begin{align*}
\text{MakeMove} & ::= \\
& \text{if } turn = \text{eve} \\
& \text{then EveMove}(y, c, c') \\
& \text{else AdamMove}(y, c, c')
\end{align*}
\]

The MakeMove nonterminal branches on whether the current move (which is kept in a variable) is for Adam or for Eve and simulates the appropriate player’s move. Note, here we use "if $\varphi$ then $\varphi_1$ else $\varphi_2$" as a shorthand for $\varphi \rightarrow \varphi_1 \land \neg \varphi \rightarrow \varphi_2$ (and similarly variants like "if $\varphi$ then $\varphi'$elif $\varphi_1$ then $\varphi_2$").

For an EveMove:

\[
\begin{align*}
\text{EveMove} & ::= \\
& \exists \text{move.} \text{Move}(\text{move}) \\
& \land \exists \text{turn.turn} = \text{adam} \\
& \land \text{GenerateConfig}(y, c, c', \text{move}, \text{turn})
\end{align*}
\]

where Move is another chance for the synthesizer to make a choice, namely, which of two transitions 0 or 1 Eve should take:

\[
\text{Move} ::= \text{move} = 0 \mid \text{move} = 1
\]

The next part of EveMove requantifies turn to give Adam the next turn and then enters a polynomial-size gadget called GenerateConfig whose purpose is to let the synthesizer choose tape symbols for the next machine configuration and to accumulate constraints that force each structure to equate their "window variables" with the symbols for their window.

\[
\text{GenerateConfig} ::= \exists c'.\text{Cell}_0
\]

We omitted Init_0 because it is similar to Cell_0, which begins the process of producing the tape symbols for the next machine configuration. Rather than bogging down in corner cases, we instead give the nonterminal for Cell_i, with $0 < i < s$:

\[
\begin{align*}
\text{Cell}_i & ::= \exists x. \text{Choose}(x) \\
& \land \text{if } \text{pred}(\star, y) \text{ then } c'_i = x \\
& \text{elif } y = \star \text{ then } c'_i = x \\
& \text{elif } \text{pred}(y, \star) \text{ then } c'_i = x \\
& \land \exists y'.\text{next}(y, y') \land \text{Cell}_{i+1}(y', c, c', \text{move}, \text{turn})
\end{align*}
\]
where \( \text{pred} \) is an immediate predecessor relation on the cycle of a structure and \( \text{next} \) is the immediate successor relation. We use \( \star \) here as a constant for the special element on the cycle (dark element in Figure 9) that marks the center of the window that a structure should track. The non-terminal \( \text{Choose} \) is another choice for the synthesizer:

\[
\text{Choose} ::= (x = a_1) \mid \ldots \mid (x = a_t)
\]

wherein it picks which of the \( a_1, \ldots, a_t \) tape symbols comes next at position \( i \) of the next configuration. This gadget is replicated again for the next cell, i.e., \( \text{Cell}_{i+1} \), and so on. The nonterminal for the final cell, i.e., \( \text{Cell}_5 \), finishes with a conjunct for a formula from a nonterminal \( \text{VerifyWindows} \):

\[
\text{VerifyWindow} ::= \text{if } \text{move} = 0 \text{ then } \delta(0, c_1, c_2, c_3, c'_2) \\
\text{elif } \text{move} = 1 \text{ then } \delta(1, c_1, c_2, c_3, c'_2) \\
\quad \land \exists c. c = c' \land \text{Game}(y, c, c', \text{turn})
\]

which asserts that the (move dependent) transition relation holds on the variables which hold the tape symbols for a given window. Finally, the previous configuration window variables are requantified and made equal to the current configuration window variables, and the game continues.

Finally, we return to the nonterminal for \( \text{AdamMove} \):

\[
\text{AdamMove} ::= \forall \text{move}. (\text{move} = 0 \lor \text{move} = 1) \\
\quad \rightarrow (\exists \text{turn}. \text{turn} = \text{eve} \land \\
\quad \text{if } \text{move} = 0 \\
\quad \text{then } \text{GenerateConfig}(y, c, c', \text{move}, \text{turn}) \\
\quad \text{else } \text{GenerateConfig}(y, c, c', \text{move}, \text{turn}))
\]

The fact that the grammar contains two synthesis obligations, one for when Adam picks transition 0 and another for transition 1, is crucial. It allows the synthesizer to build a sentence which witnesses a computation tree that depends on the history of Adam transitions.

Note that, for clarity, the description of the grammar above used more variables than strictly necessary. For instance, by encoding triples of tape symbols as single domain elements, and by tracking the current turn in the grammar rather than a variable, we can reduce the number of variables needed to 5.

**Correctness.** If \( M \) has an accepting computation tree on \( w \) the synthesizer can follow the strategy described by the computation tree to produce a sentence that is true in every structure. This involves synthesizing each tape symbol of each successive \( M \) configuration along each branch of the computation tree and picking the correct Eve transition for each Adam transition. If there is a sentence \( \varphi \in L(G) \) that is true in each structure, then an accepting computation tree can be reconstructed by traversing the sentence. Each window centered at cell \( i \) of each configuration constructed in this way must evolve in accordance with the transition relation since the sentence is true structure \( A_i \).

**F MUTUAL RECURSION IN FO-LFP**

In Figure 11 we give syntax for an extension to FO-LFP that allows for blocks of mutually recursive relation definitions. As discussed in §9.1, an automaton for evaluating such blocks of definitions needs state that increases with the number of distinct relations in a block. This number is bounded by the number of available definable symbols, given that reusing the same symbol for two distinct definitions in a single block is ambiguous (and thus we rule it out). The total amount of state an automaton needs to evaluate blocks of definitions thereby becomes independent of the size of formulas.
Datalog

We consider a

Theorem 12. Datalog

where

\( \text{Def} ::= \text{let Block in Def} \mid \varphi \)

Block ::= \( P(\bar{x}) \leftarrow \varphi \mid P(\bar{x}) \leftarrow \varphi \) and Block

\( \varphi ::= R(\bar{t}) \mid \varphi \lor \varphi \mid \varphi \land \varphi \mid \neg \varphi \mid \exists x.\varphi \mid \forall x.\varphi \mid P(\bar{t}) \)

\( t ::= x \mid c \mid f(\bar{i}) \mid \text{ite}(\varphi, t, t') \)

Fig. 11. Grammar for FO-LFP extended with blocks of (mutually) recursive definitions, where Def is the starting nonterminal, \( P \) ranges over a set of definable relation symbols, and \( R \) and \( f \) range over relations and functions from a signature \( \tau \).

G Datalog SYNTHESIS

We consider a Datalog synthesis problem as follows. Fix a set of variables \( V = \{ v_1, \ldots, v_k \} \), and let \( x, y \) range over \( V \). Let \( \tau = (R_1, \ldots, R_n) \) be a relational signature. Each \( R_i \) will be an extensional predicate, in Datalog parlance (encoding facts). Fix a set of \( \text{intensional} \) predicate symbols \( P = \{ P_1, \ldots, P_{k'} \} \) which the Datalog program should define. The Datalog programs we consider have the form \( (\Pi, Q) \), where \( \Pi \) is a set of rules and \( Q \in P \) is a distinguished intensional predicate. Logically, a \( \text{rule} \) is a formula of the form \( \rho := \exists \bar{x}.\bigwedge_i^\prime \psi_i(y, \bar{x}) \). Let each \( \psi_i \) be either of the form \( R_i(\cdot), \neg R_i(\cdot), \) or \( P_j(\cdot) \). Rules \( \rho \) of this kind are used to define the intensional predicates \( P_1, \ldots, P_{k'} \).

The reader may recognize a different syntax that makes this clear: \( P_j(y) \leftarrow \psi_1(y, x), \ldots, \psi_t(y, x) \). Note that rules may be recursive, that is, \( P_j \) may appear as one of the \( \psi_i \)'s.

The semantics of a Datalog program \( (\Pi, Q) \) over a \( \tau \)-structure \( A \) is defined in terms of a \textit{simultaneous fixed point}. Suppose an intensional predicate \( P_i \) has arity \( m_i \) and is defined by \( t \) clauses \( P_i(y) \leftarrow \psi_1^i(y, \bar{x}), \ldots, \psi_{t}^i(y, \bar{x}) \), for \( 1 \leq j \leq t \). Let \( X = 2^{A^{m_1}} \times \cdots \times 2^{A^{m_{k'}}} \). Then for a particular interpretation of the intensional predicates \( U = (U_1, \ldots, U_{k'}) \in X \) we can define an immediate consequence operator \( F_{P_i} : X \rightarrow 2^{A^{m_i}} \) for the predicate \( P_i \):

\[
F_{P_i}(U) = \left\{ \bar{a} \in A^{m_i} \mid (A, U) \models \exists \bar{x}. \bigwedge_{j=1}^t \psi_j^i(\bar{a}, \bar{x}) \right\}
\]

where \( (A, U) \models \varphi \) denotes entailment when the intensional predicates are interpreted as in \( U \). Each \( F_{P_i} \) is clearly monotonic since no \( P_i \) appears negatively in the rules of \( \Pi \). Thus we have a system of monotonic functions \( \{ F_{P_i} \} \), and the intensional predicates are interpreted according to the simultaneous least fixed point. For more on simultaneous fixed points we refer the reader to [Fritz 2002].

Similar to our handling of FO-LFP\( (k, k') \), we consider a variant of Datalog with only \( k \) first-order variables and \( k' \) intensional predicate symbols, denoted Datalog\( (k, k') \). It is easy to define a suitable ranked alphabet \( \Sigma_{\text{Datalog}(k, k')} \) (which is similar to earlier alphabets). Consider the problem in Problem 4:

**Theorem 12.** Datalog\( (k, k') \)-separator realizability and synthesis is decidable in EXPTIME for a fixed signature and fixed \( k, k' \in \mathbb{N} \).

**Sketch.** Follows very closely the construction from Section 6. The main novelty is to evaluate mutually recursive rules with a two-way automaton. Similar to the case of FO-LFP we can use a counter in the automaton state to ensure that we correctly check non-membership in least fixed point relations. The difference is that we need to keep multiple counters at once (one per intensional predicate), because evaluating any single definition may involve evaluating other definitions that can refer back to the original. Note that our solution for FO-LFP could get away with a
Problem 4: Datalog\((k, k')\)-separator realizability and synthesis

**Input:** \(((A, Pos, Neg, P_t, G)\) where
- \(A\) is a \(\tau\)-structure
- \(P_t\) is a target intensional predicate symbol
- \(Pos, Neg\) are sets of tuples over \(\text{dom}(A)^{\text{arity}(P_t)}\)
- \(G\) is an RTG over \(\Sigma_{\text{Datalog}(k,k')}\)

**Output:** \(\Pi \in L(G)\) s.t. \(Pos \subseteq \text{lfp}(F_{P_t})\) and \(Neg \cap \text{lfp}(F_{P_t}) = \emptyset\)
- Or "No" if no such program exists

single counter because each successive definition could only refer to previously defined relations. In contrast, there is no notion of scope within a single block, and all recursive definitions in the same block see each other.