Query Learning of Derived ω-Tree Languages in Polynomial Time

Dana Angluin[1], Timos Antonopoulos[2], and Dana Fisman[3]

1 **Yale University, New Haven, CT, USA**
2 **Yale University, New Haven, CT, USA**
3 **Ben-Gurion University, Be'er Sheva, Israel**

Abstract

We present the first polynomial time algorithm to learn nontrivial classes of languages of infinite trees. Specifically, our algorithm uses membership and equivalence queries to learn classes of ω-tree languages derived from weak regular ω-word languages in polynomial time. The method is a general polynomial time reduction of learning a class of derived ω-tree languages to learning the underlying class of ω-word languages, for any class of ω-word languages recognized by a deterministic Büchi acceptor. Our reduction, combined with the polynomial time learning algorithm of Maler and Pnueli [15] for the class of weak regular ω-word languages yields the main result. We also show that subset queries that return counterexamples can be implemented in polynomial time using subset queries that return no counterexamples for deterministic or non-deterministic finite word acceptors, and deterministic or non-deterministic Büchi ω-word acceptors.

A previous claim of an algorithm to learn regular ω-trees due to Jayasrirani, Begam and Thomas [11] is unfortunately incorrect, as shown in [4].

1998 ACM Subject Classification F.4.3 Formal Languages, F.1.1 Models of Computation, I.2.6 Learning

Keywords and phrases Learning, queries, infinite trees, derived tree languages, reactive systems

Digital Object Identifier 10.4230/LIPIcs.CSL.2017.10

1 Introduction

Query learning is a framework in which a learning algorithm attempts to identify a target concept using specified types of queries to an oracle (or teacher) about the target concept [2]. For example, if the target concept is a regular language L, a membership query asks whether a string x is a member of L, and is answered either "yes" or "no". An equivalence query asks whether a hypothesis language L' (represented, for example, by a deterministic finite acceptor) is equal to L. In the case of an equivalence query, the answer may be "yes", in which case the learning algorithm has succeeded in exact identification of the target concept, or it may be "no", accompanied by a counterexample, that is, a string x in L but not in L' (or vice versa). The counterexample is a witness that L' is not equal to L.

When L' is not equal to L, there is generally a choice (often an infinity) of possible counterexamples, and we require that the learning algorithm work well regardless of which counterexample is chosen by the teacher. To account for this in terms of quantifying the running time of the learning algorithm, we include a parameter that is the maximum length of any counterexample returned by the teacher at any point in the learning process. In this setting, the L^* algorithm of Angluin [1] learns any regular language L using membership and equivalence queries in time polynomial in the size of the smallest deterministic finite acceptor for L and the length of the longest counterexample chosen by the teacher. There

© Dana Angluin, Timos Antonopoulos and Dana Fisman;
licensed under Creative Commons License CC-BY

can be no such polynomial time algorithm using just membership queries or just equivalence queries [3].

The assumption that equivalence queries are available may seem unrealistic. How is a person or a program to judge the equivalence of the target concept to some large, complex, technical specification of a hypothesis? If the hypothesis and the target concept are both deterministic finite acceptors, there is a polynomial time algorithm to test equivalence and return a counterexample in case the answer is negative. Alternatively, if there is an efficient algorithm for exact learnability of a class $\mathbb{C}$ of concepts using membership and equivalence queries, then it is easily transformed to an efficient algorithm that approximately learns concepts from $\mathbb{C}$ using membership queries and a sufficiently large corpus of labeled examples [1]. In this transformation, an equivalence query is answered "yes" if the hypothesis L' agrees with the labels of all the examples in the corpus; otherwise, any exception supplies a counterexample x whose label disagrees with its classification by L'.

Since the publication of L^*, there have been a number of substantial improvements and extensions of the algorithm, as well as novel and unanticipated applications in the analysis, verification and synthesis of programs, protocols and hardware. In a recent CACM review article, Vaandrager [20] explains Model Learning, which takes a black box approach to learning a finite state model of a given hardware or software system using membership queries (implemented by giving the system a sequence of inputs and observing the sequence of outputs) and equivalence queries (implemented using a set of test sequences in which the outputs of the hypothesis are compared with the outputs of the given system.) The learned models may then be analyzed to find discrepancies between a specification and its implementation, or between different implementations. He cites applications in telecommunications, the automotive industry, online conference systems, as well as analyzing botnet protocols, smart card readers, bank card protocols, network protocols and legacy software.

Another application of finite state machine learning algorithms is in the assume-guarantee approach to verifying systems by dividing them into modules that can be verified individually. Cobleigh, Giannakopoulou and Păsareănu [8] first proposed using a learning algorithm to learn a correct and sufficient contextual assumption for the component being verified, and there has since been a great deal of research progress in this area.

If we consider *reactive systems*, that is, systems that maintain an ongoing interaction with their environment (e.g., operating systems, communication protocols, or robotic swarms), the restriction to models specified by finite automata processing finite sequences of inputs is too limiting. Instead, one trajectory of the behavior of a reactive system may be modeled using an infinite word (ω-word), each symbol of which specifies the current state of the system and the environment at a given time. The system itself may be modeled by an ω-automaton, that is, a finite state automaton that processes ω-words. The desired behavior of such a system may be specified by a linear temporal logic formula, that defines the set of ω-words that constitute "good" behaviors of the system.

Researchers have thus sought query learning algorithms for ω-automata that could be used in the settings of model learning or assume-guarantee verification for reactive systems. However, learning ω-automata seems to be a much more challenging problem than learning automata on finite words, in part because the Myhill-Nerode characterization for regular languages (that there is a unique minimum deterministic acceptor that can be constructed using the right congruence classes of the language) does not hold in general for regular ω-languages. The Myhill-Nerode characterization is the basis of the L^* algorithm and its successors.

There is no known polynomial time algorithm using membership and equivalence queries

to learn even the whole class $\mathbb{DBW}$ of languages recognized by deterministic Büchi acceptors, which is a strict subclass of the class of all regular ω-languages. Maler and Pnueli [15] have given a polynomial time algorithm using membership and equivalence queries to learn the *weak regular ω-languages*. This class, denoted $\mathbb{DwPW}$, is the set of languages accepted by deterministic weak parity automata, and is a non-trivial sublass of $\mathbb{DBW}$. The class $\mathbb{DwPW}$ does have a Myhill-Nerode property, and the Maler and Pnueli algorithm is a very interesting extension of the L^* approach.

In the context of assume-guarantee verification, Farzan et al. [9] proposed a direct application of L^* to learn the full class of regular ω-languages. Their approach is based on the result of Calbrix, Nivat and Podelski [7] showing that a regular ω-language L can be characterized by the regular language $L_\$$ of finite strings $u\$v$ representing the set of ultimately periodic words $u(v)^\omega$ in L. This establishes that a regular ω-language L is learnable using membership and equivalence queries in time polynomial in the size of the minimal deterministic finite acceptor for $L_\$$. However, the size of this representation may be exponentially larger than its ω-automaton representation. More recently, Angluin and Fisman [5] have given a learning algorithm using membership and equivalence queries for general regular ω-languages represented by families of deterministic finite acceptors, which improves on the $L_\$$ representation, however the running time is not bounded by a polynomial in the representation. Clearly, much more research is needed in the area of query learning of ω-automata.

Despite the difficulties in learning ω-automata (which are used in the analysis of linear temporal logic) in this paper we consider the theoretical question of learning ω-tree automata (which are used in the analysis of branching temporal logic). Though we have not identified any practical application for an algorithm for learning ω-tree languages, we note that our results shed new light on the phenomenon of learning, and we are confident that they will stimulate further work in this challenging area.

Because of the difficulty of the problem, we restrict our attention to ω-tree languages such that all of their paths satisfy a certain temporal logic formula, or equivalently, a property of ω-words. Given an ω-word language L, we use $\mathit{Trees}_d(L)$ to denote the set of all d-ary ω-trees t all of whose paths are in L. The ω-tree language $\mathit{Trees}_d(L)$ is often referred to as the *derived* language of L. In this context, it is natural to ask whether the question of learning a derived ω-tree language $\mathit{Trees}_d(L)$ can be reduced to learning the ω-word language L.

We answer this question affirmatively for the case that L can be recognized by a deterministic Büchi word automaton and learned using membership and equivalence queries. Applying this reduction to the result of Maler and Pnueli on polynomially learning languages in $\mathbb{DwPW}$ we obtain a polynomial learning algorithm for derived languages in $\mathit{Trees}_d(\mathbb{DwPW})$ using membership and equivalence queries. Moreover, any progress on polynomially learning an extended sub-class $\mathbb{C}$ of $\mathbb{DBW}$ using membership and equivalence queries can be automatically lifted to learning $\mathit{Trees}_d(\mathbb{C})$.

The framework of the reduction is depicted in Fig. 1. An algorithm $\mathbf{A}_{\mathit{Trees}}$ for learning $\mathit{Trees}_d(L)$ uses a learning algorithm $\mathbf{A}$ for L to complete its task. The membership and equivalence queries (MQ and EQ, henceforth) of algorithm $\mathbf{A}_{\mathit{Trees}}$ are answered by respective oracles MQ and EQ for $\mathit{Trees}_d(L)$. Since $\mathbf{A}$ asks membership and equivalence queries about L rather than $\mathit{Trees}_d(L)$, the learner $\mathbf{A}_{\mathit{Trees}}$ needs to find a way to answer these queries. If $\mathbf{A}$ asks a membership query about an ω-word, $\mathbf{A}_{\mathit{Trees}}$ can ask a membership query about an ω-tree all of whose paths are identical to the given ω-word. Since the tree is accepted by $\mathit{Trees}_d(L)$ iff the given word is accepted by L it can pass the answer as is to $\mathbf{A}$. If $\mathbf{A}$

asks an equivalence query, using an acceptor M for an ω-language, then $\mathbf{A}_{Trees}$ can ask an equivalence query using an acceptor M^T that accepts an ω-tree if all its paths are accepted by M. If this query is answered positively then $\mathbf{A}_{Trees}$ can output the tree acceptor M^T and halt. The challenge starts when this query is answered negatively.

When the EQ is answered negatively, a counterexample tree t is given. There are two cases to consider. Either this tree is in $Trees_d(L)$ but is rejected by the hypothesis acceptor M^T, in which case t is referred to as a *positive counterexample*; or this tree is not in $Trees_d(L)$ but is accepted by the hypothesis acceptor M^T, in which case t is referred to as a *negative counterexample*. If t is a positive counterexample, since M^T rejects t there must be a path in t which is rejected by M. It is not too dificult to extract that path. The real challenge is dealing with a negative counterexample. This part is grayed out in the figure. In this case the tree t is accepted by M^T yet it is not in $Trees_d(L)$. Thus, all the paths of the tree are accepted by M, yet at least one path is not accepted by L. Since L is not given, it is not clear how we can extract such a path. Since we know that not all paths of t are contained in L, a use of an unrestricted subset query could help us. Unrestricted subset queries (USQ) are queries on the inclusion of a current hypothesis in the unknown language that are answered by "yes" or "no" with an accompanying counterexample in the case the answer is negative.

Since we don't have access to USQs we investigate whether we can obtain such queries given the queries we have. We show that unrestricted subset queries can be simulated by restricted subset queries. Restricted subset queries (RSQ) on ω-words are subset queries that are *not* accompanied by counterexamples. This essentially means that there is a way to construct a desired counterexample without it being given. To discharge the use of restricted subset queries (as the learner is not provided such queries either) we investigate the relation between subsets of ω-words and ω-trees. Finally, we show that the desired subset queries on ω-words can be given to the ω-tree learning algorithm by means of ω-tree membership queries. From these we can construct a series of procedures to implement the grayed area.

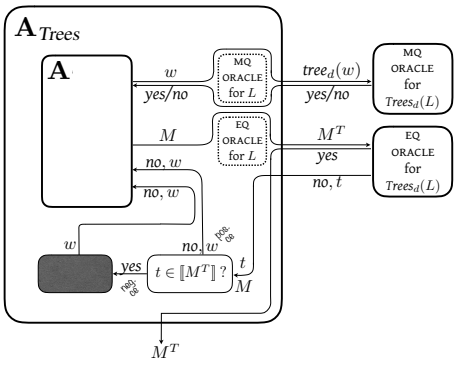

Figure 1 The reduction framework

The subsequent sections contain definitions of ω-words, ω-trees and automata processing them, derived ω-tree languages, the problem of learning classes of ω-word and ω-tree languages, preliminary results, the algorithm for the main reduction, and discussion. Many proof details are omitted because of space limitations.

2 Definitions

2.1 Words and trees

(For more details see Grädel, Thomas and Wilke [10], Perrin and Pin [18], and Löding [14].) Let Σ be a fixed finite alphabet of symbols. The set of all finite words over Σ is denoted Σ^*. The empty word is denoted ε, and the length of a finite word x is denoted $|x|$. Σ^+ is the set of all nonempty finite words over Σ, and for a nonnegative integer k, Σ^k is the set of all finite words over Σ of length equal to k. A finite word language is a subset of Σ^*.

An ω-word over Σ is an infinite sequence $w = \sigma_1\sigma_2\sigma_3\cdots$ where each $\sigma_i \in \Sigma$. The set of

all ω-words over Σ is denoted Σ^ω. An ω-word language is a subset of Σ^ω. The ω-regular expressions are analogous to finite regular expressions, with the added operation S^ω, where S is a set of finite words, and the restriction that concatenation may combine two sets of finite words, or a set of finite words and a set of ω-words. The set S^ω is the set of all ω-words $s_1 s_2 \cdots$ such that for each i, $s_i \in S$ and $s_i \neq \varepsilon$. For example, $(a+b)^*(a)^\omega$ is the set of all ω-words over $\{a, b\}$ that contain finitely many occurrences of b.

If $S \subseteq \Sigma^*$, n is a nonnegative integer and $u \in \Sigma^*$, we define the *length and prefix restricted* version of S by $S[n, u] = S \cap \Sigma^n \cap (u \cdot \Sigma^*)$. This is the set of all words in S of length n that begin with the prefix u. We also define the *length restricted* version of S by $S[n] = S[n, \varepsilon]$, that is, the set of all words in S of length n.

Let d be a positive integer. We consider T_d, the unlabeled complete d-ary ω-tree whose directions are specified by $D = \{1, \ldots, d\}$. The *nodes* of T_d are the elements of D^*. The *root* of T_d is the node ε, and the *children* of node v are $v \cdot i$ for $i \in D$. An *infinite path* π in T_d is a sequence $x_0, x_1, x_2, \ldots$ of nodes of T_d such that x_0 is the root and for all nonnegative integers n, x_{n+1} is a child of x_n. An infinite path in T_d corresponds to an ω-word over D giving the sequence of directions traversed by the path starting at the root.

A labeled d-ary ω-tree (or just *ω-tree*) is given by a mapping $t : D^+ \to \Sigma$ that assigns a symbol in Σ to each non-root node of T_d. We may think of t as assigning the symbol $t(v \cdot i)$ to the tree edge from node v to its child node $v \cdot i$. The set of all labeled d-ary ω-trees is denoted T_d^Σ. An ω-tree language is a subset of T_d^Σ. If $\pi = x_0, x_1, x_2, \ldots$ is an infinite path of T_d, then we define $t(\pi)$ to be the ω-word $t(x_1), t(x_2), \ldots$ consisting of the sequence of labels of the non-root nodes of π in t. (Recall that t does not label the root node.)

2.2 Automata on words

A *finite state word automaton* is given by a tuple $M = (Q, q_0, \delta)$, where Q is a finite set of states, $q_0 \in Q$ is the initial state, and $\delta : Q \times \Sigma \to 2^Q$ is the (nondeterministic) transition function. The automaton is *deterministic* if $\delta(q, \sigma)$ contains at most one state for every $(q, \sigma) \in Q \times \Sigma$, and *complete* if $\delta(q, \sigma)$ contains at least one state for every $(q, \sigma) \in Q \times \Sigma$. For a complete deterministic automaton we extend δ to map $Q \times \Sigma^*$ to Q in the usual way.

Let $x = \sigma_1 \sigma_2 \cdots \sigma_k$ be a finite word, where each $\sigma_n \in \Sigma$. A *run* of M on x is a sequence of $k+1$ states $r_0, r_1, \ldots, r_k$ such that $r_0 = q_0$ is the initial state and $r_n \in \delta(r_{n-1}, \sigma_n)$ for integers $1 \leq n \leq k$. Let $w = \sigma_1 \sigma_2 \cdots$ be an ω-word, where each $\sigma_n \in \Sigma$. A *run* of M on w is an infinite sequence of states $r_0, r_1, r_2, \ldots$ such that $r_0 = q_0$ is the initial state and $r_n \in \delta(r_{n-1}, \sigma_n)$ for all positive integers n.

A *nondeterministic finite acceptor* is given by $M = (Q, q_0, \delta, F)$, where (Q, q_0, δ) is a finite state word automaton, and the new component $F \subseteq Q$ is the set of accepting states. M is a *deterministic finite acceptor* if δ is deterministic. Let M be a nondeterministic finite acceptor and $x \in \Sigma^*$ a finite word of length n. M *accepts* x iff there is a run $r_0, r_1, \ldots, r_n$ of M on x such that $r_n \in F$. The language *recognized* by M is the set of all finite words accepted by M, denoted by $[\![M]\!]$. The class of all finite word languages recognized by deterministic finite acceptors is denoted by $\mathbb{DFW}$, and by nondeterministic finite acceptors, $\mathbb{NFW}$. These representations are equally expressive, that is, $\mathbb{NFW} = \mathbb{DFW}$.

Turning to finite state word automata processing ω-words, a variety of different acceptance criteria have been considered. Such an acceptor is given by a tuple $M = (Q, q_0, \delta, \alpha)$, where (Q, q_0, δ) is a finite state word automaton and α specifies a mapping from 2^Q to $\{0, 1\}$ which gives the criterion of acceptance for an ω-word w.

Given an ω-word w and a run $r = r_0, r_1, \ldots$ of M on w, we consider the set $Inf(r)$ of all states $q \in Q$ such that $r_n = q$ for infinitely many indices n. The acceptor M *accepts* the

ω-word w iff there exists a run r of M on w such that $\alpha(\mathit{Inf}(r)) = 1$. That is, M accepts w iff there exists a run of M on w such that the set of states visited infinitely often in the run satisfies the acceptance criterion α. The language *recognized* by M is the set of all ω-words accepted by M, denoted $[\![M]\!]$.

For a Büchi acceptor, the acceptance criterion α is specified by giving a set $F \subseteq Q$ of accepting states and defining $\alpha(S) = 1$ iff $S \cap F \neq \emptyset$. In words, a Büchi acceptor M accepts w if and only if there exists a run r of M on w such that at least one accepting state is visited infinitely often in r. For a co-Büchi acceptor, the acceptance criterion α is specified by giving a set $F \subseteq Q$ of rejecting states and defining $\alpha(S) = 1$ iff $S \cap F = \emptyset$. For a parity acceptor, α is specified by giving a function c mapping Q to an interval of integers $[i, j]$, (called *colors* or *priorities*) and defining $\alpha(S) = 1$ iff the minimum integer in $c(S)$ is even.

A *parity* automaton is said to be *weak* if no two strongly connected states have distinct colors, i.e., if looking at the partition of its states to maximal strongly connected components (MSCCs) all states of an MSCC have the same color. Clearly every weak parity automaton can be colored with only two colors, one even and one odd, in which case the colors are often referred to as *accepting* or *rejecting*. It follows that a weak parity automaton can be regarded as either a Büchi or a coBüchi automaton. If in addition no rejecting MSCC is reachable from an accepting MSCC, the acceptor is said to be *weak Büchi*. Likewise, a weak parity acceptor where no accepting MSCC is reachable from a rejecting MSCC, is said to be *weak coBüchi* acceptor.

The classes of languages of ω-words recognized by these kinds of acceptors will be denoted by three/four-letter acronyms, with $\mathbb{N}$ or $\mathbb{D}$ (for nondeterministic or deterministic), $\mathbb{B}$, $\mathbb{C}$, $\mathbb{P}$, $\mathbb{wB}$, $\mathbb{wC}$ or $\mathbb{wP}$ (for Büchi, co-Büchi, parity or their respective weak variants) and then $\mathbb{W}$ (for ω-words). Thus $\mathbb{DwBW}$ is the class of ω-word languages recognized by deterministic weak Büchi word acceptors. Concerning the expressive power of various types of acceptors, previous research has established the following results. The weak variants are strictly less expressive than the non-weak variants. Deterministic parity automata are more expressive than deterministic Büchi and coBüchi automata and the same is true for their weak variants. These results are summarized in Fig. 2. In addition, $\mathbb{NBW} = \mathbb{DPW} = \mathbb{NPW}$ and $\mathbb{DwPW} = \mathbb{DCW} \cap \mathbb{DBW}$. The class of *regular ω-languages* is the class $\mathbb{DPW}$, and the class of *weak regular ω-languages* is the class $\mathbb{DwPW}$.

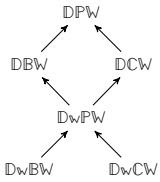

Figure 2 Expressiveness hierarchy of ω-acceptors [21, 16].

2.3 Automata on trees

Acceptors on d-ary ω-trees are equipped with analogous accepting conditions. Such an acceptor is given by a tuple $M = (Q, q_0, \delta, \alpha)$, where Q is a finite set of states, $q_0 \in Q$ is the initial state, the transition function δ is a map from Q and d-tuples of symbols to sets of d-tuples of states, that is, $\delta : Q \times \Sigma^d \to 2^{Q^d}$, and the acceptance criterion α specifies a function from 2^Q to $\{0, 1\}$. We may think of the acceptor as running top down from the root of the tree, at each node nondeterministically choosing a permissible d-tuple of states for the d children of the node depending on the state assigned to the node and the d-tuple of symbols on its outgoing edges.

We define a *run* of M on the ω-tree t as a mapping r from the nodes of T_d to Q such that $r(\varepsilon) = q_0$ and for every node x, we have $(r(x \cdot 1), \ldots, r(x \cdot d)) \in \delta(r(x), (t(x \cdot 1), \ldots, t(x \cdot d)))$. That is, the root is assigned the initial state and for every node, the ordered d-tuple of states assigned to its children is permitted by the transition function. The acceptor M *accepts* the ω-tree t iff there exists a run r of M on t such that for every infinite path π, we have

$\alpha(\mathit{Inf}(r(\pi))) = 1$. That is, there must be at least one run in which, for every infinite path, the set of states that occur infinitely often on the path satisfies the acceptance criterion α. The ω-tree language *recognized* by M is the set of all ω-trees accepted by M, denoted $[\![M]\!]$.

The specification of the acceptance criterion α is as for ω-word acceptors, yielding Büchi, co-Büchi and parity ω-tree acceptors. If the transition function specifies at most one permissible d-tuple of states for every element of $Q \times \Sigma^d$, then the acceptor is deterministic. The corresponding classes of ω-tree languages are also denoted by three-letter acronyms, where the last letter is $\mathbb{T}$ for ω-trees. For ω-trees, the class of all regular ω-tree languages is $\mathbb{NPT}$ and $\mathbb{NBT}$ is a proper subclass of $\mathbb{NPT}$. For any automaton or acceptor M, we denote the number of its states by $|M|$.

3 Derived ω-tree languages

Given an ω-tree t we define the ω-word language $\mathit{paths}(t)$ consisting of the ω-words labeling its infinite paths. That is, we define

$$\mathit{paths}(t) = \{t(\pi) \mid \pi \text{ is an infinite path in } T_d\}.$$

If L is an ω-word language and d is a positive integer, we define a corresponding language of d-ary ω-trees *derived from* L as follows:

$$\mathit{Trees}_d(L) = \{t \in T_d^\Sigma \mid \mathit{paths}(t) \subseteq L\}.$$

That is, $\mathit{Trees}_d(L)$ consists of all d-ary ω-trees such that every infinite path in the tree is labeled by an element of L. If $\mathbb{C}$ is any class of ω-word languages, $\mathit{Trees}_d(\mathbb{C})$ denotes the class of all ω-tree languages $\mathit{Trees}_d(L)$ such that $L \in \mathbb{C}$.

3.1 Derived tree languages

Not every regular d-ary ω-tree language can be derived in this way from an ω-word language. As an example, consider the language L_a of all binary ω-trees t over $\Sigma = \{a, b\}$ such that there is at least one node labeled with a. An NBT acceptor can recognize L_a by guessing and checking a path that leads to an a. However, if $L_a = \mathit{Trees}_2(L)$ for some ω-word language L, then because there are ω-trees in L_a that have infinite paths labeled exclusively with b, we must have $b^\omega \in L$, so the binary ω-tree labeled exclusively with b would also be in $\mathit{Trees}_2(L)$, a contradiction.

Given an ω-word acceptor $M = (Q, q_0, \delta, \alpha)$, we may construct a related ω-tree acceptor $M^{T,d} = (Q, q_0, \delta^{T,d}, \alpha)$ as follows. For all $q \in Q$ and all $(\sigma_1, \ldots, \sigma_d) \in \Sigma^d$, define

$$\delta^{T,d}(q, (\sigma_1, \ldots, \sigma_d)) = \{(q_1, \ldots, q_d) \mid \forall i \in D, q_i \in \delta(q, \sigma_i)\}.$$

That is, the acceptor $M^{T,d}$ may continue the computation at a child of a node with any state permitted by M, independently chosen. It is tempting to think that $[\![M^{T,d}]\!] = \mathit{Trees}_d([\![M]\!])$, but this may not be true when M is not deterministic.

▶ **Lemma 1.** *Given an ω-word acceptor M, we have that $[\![M^{T,d}]\!] \subseteq \mathit{Trees}_d([\![M]\!])$ with equality if M is deterministic.*

Boker et al. [6] give the following example to show that the containment asserted in Lemma 1 may be proper if M is not deterministic. The ω-language L specified by $(a+b)^*b^\omega$ can be recognized by the nondeterministic Büchi acceptor M with two states, q_0 and

q_1, transition function $\delta(q_0, a) = \{q_0\}$, $\delta(q_0, b) = \{q_0, q_1\}$, $\delta(q_1, b) = \{q_1\}$, and accepting state set $\{q_1\}$. Let $d = 2$, specifying binary trees with directions $\{1, 2\}$. Then $M^{T,2}$ is a nondeterministic ω-tree acceptor, but the following example shows $[\![M^{T,2}]\!] \subsetneq \mathit{Trees}_2(L)$. Consider the binary ω-tree t that labels every node in 1^*2 with a and every other non-root node with b. Clearly $t \in \mathit{Trees}_2(L)$ because every infinite path in t has at most one a, but no run of $M^{T,2}$ can satisfy the acceptance criterion on the path 1^ω. Suppose r were an accepting run of $M^{T,2}$ on t. Then for some $n \geq 0$, $r(1^n)$ would have to be equal to q_1. But then such a mapping r would not be a valid run because 1^n2 is labeled by a and $\delta^{T,2}(q_1, (b, a)) = \emptyset$ because $\delta(q_1, a) = \emptyset$.

3.2 Good for trees

This phenomenon motivates the following definition. An ω-word acceptor M is *good for trees* iff for any positive integer d, $[\![M^{T,d}]\!] = \mathit{Trees}_d([\![M]\!])$. Nondeterministic ω-word acceptors that are good for trees are equivalent in expressive power to deterministic ω-word acceptors, as stated by the following result of Boker et al.

▶ **Theorem 2.** *[6] Let L be a regular ω-word language and $d \geq 2$. If $\mathit{Trees}_d(L)$ is recognized by a nondeterministic ω-tree acceptor with acceptance criterion α, then L can be recognized by a deterministic ω-word acceptor with acceptance criterion α.*

This theorem generalizes prior results of Kupferman, Safra and Vardi for Büchi acceptors [13] and Niwiński and Walukiewicz for parity acceptors [17]. One consequence of Theorem 2 is that when $d \geq 2$, nondeterministic ω-word acceptors that are good for trees are not more expressive than the corresponding deterministic ω-word acceptors. Also, for $d \geq 2$, nondeterminism does not increase expressive power over determinism when recognizing ω-tree languages of the form $\mathit{Trees}_d(L)$. To see this, if N is a nondeterministic ω-tree acceptor with acceptance criterion α recognizing $\mathit{Trees}_d(L)$ then there is a deterministic ω-word acceptor M with acceptance criterion α such that $[\![M]\!] = L$, and $M^{T,d}$ is a deterministic ω-tree acceptor with acceptance criterion α that also recognizes $\mathit{Trees}_d(L)$.

However, it is possible that nondeterminism permits acceptors with smaller numbers of states. Kuperberg and Skrzypczak [12] have shown that for an NBT acceptor M recognizing $\mathit{Trees}_d(L)$, there is a DBW acceptor with at most $|M|^2$ states recognizing L, so nondeterminism gives at most a quadratic savings for Büchi tree acceptors that are good for trees. However, they have also shown that the blowup in the case of nondeterministic co-Büchi tree acceptors (and all higher parity conditions) is necessarily exponential in the worst case.

4 Learning tree languages

We address the problem of learning derived ω-tree languages by giving a polynomial time reduction of the problem of learning $\mathit{Trees}_d(\mathbb{C})$ to the problem of learning $\mathbb{C}$. The paradigm of learning we consider is exact learning with membership queries and equivalence queries. Maler and Pnueli [15] have given a polynomial time algorithm to learn the class of weak regular ω-languages using membership and equivalence queries. Their algorithm and the reduction we give in Theorem 10 prove the following theorem.

▶ **Theorem 3.** *For every positive integer d, there is a polynomial time algorithm to learn $\mathit{Trees}_d(\mathbb{D}\mathrm{w}\mathbb{P}\mathbb{W})$ using membership and equivalence queries.*

4.1 Representing examples

For a learning algorithm, the examples tested by membership queries and the counterexamples returned by equivalence queries need to be finitely represented. For learning regular ω-word languages, it suffices to consider *ultimately periodic ω-words*, that is, words of the form $u(v)^\omega$ for finite words $u \in \Sigma^*$ and $v \in \Sigma^+$. If two regular ω-word languages agree on all the ultimately periodic ω-words, then they are equal. The pair (u, v) of finite words represents the ultimately periodic word $u(v)^\omega$.

The corresponding class of examples in the case of regular ω-tree languages is the class of *regular ω-trees*. These are ω-trees that have a finite number of nonisomorphic complete infinite subtrees. We represent a regular ω-tree t by a *regular ω-tree automaton* $A_t = (Q, q_0, \delta, \tau)$, where (Q, q_0, δ) is a complete deterministic finite state word automaton over the input alphabet $D = \{1, \ldots, d\}$ and τ is an output function that labels each transition with an element of Σ. That is, $\tau : Q \times D \to \Sigma$. The regular ω-tree t represented by such an automaton A_t is defined as follows. For $x \in D^+$, let $i \in D$ be the last symbol of x and let x' be the rest of x, so that $x = x' \cdot i$. Then define $t(x) = \tau(\delta(q_0, x'), i)$, that is, $t(x)$ is the label assigned by τ to the last transition in the unique run of A_t on x.

Rabin [19] proved that if two regular ω-tree languages agree on all the regular ω-trees then they are equal. Thus, ultimately periodic ω-words and regular ω-trees are proper subsets of examples that are nonetheless sufficient to determine the behavior of regular ω-word and ω-tree acceptors on all ω-words and ω-trees, respectively.

4.2 Types of queries for learning

We consider the situation in which a learning algorithm $\mathbf{A}$ is attempting to learn an initially unknown target language L of ω-words from a known class $\mathbb{C} \subseteq \mathbb{DBW}$. The information that $\mathbf{A}$ gets about L is in the form of answers to queries of specific types [2]. The learning algorithm will use membership and equivalence queries; in addition, restricted and unrestricted subset queries will be considered in the proof.

In a *membership query about L*, abbreviated MQ, the algorithm $\mathbf{A}$ specifies an example as a pair of finite words (u, v) and receives the answer "yes" if $u(v)^\omega \in L$ and "no" otherwise. In an *equivalence query about L*, abbreviated EQ, the algorithm $\mathbf{A}$ specifies a hypothesis language $[\![M]\!]$ as a DBW acceptor M, and receives either the answer "yes" if $L = [\![M]\!]$, or "no" and a counterexample, that is, a pair of finite words (u, v) such that $u(v)^\omega \in (L \oplus [\![M]\!])$, where $B \oplus C$ denotes the symmetric difference of sets B and C.

In a *restricted subset query about L*, abbreviated RSQ, the algorithm $\mathbf{A}$ specifies a hypothesis language $[\![M]\!]$ as a DBW acceptor M, and receives the answer "yes" if $[\![M]\!] \subseteq L$ and "no" otherwise. An *unrestricted subset query about L*, abbreviated USQ, is like a restricted subset query, except that in addition to the answer of "no", a counterexample (u, v) is provided such that $u(v)^\omega \in ([\![M]\!] \setminus L)$.

A learning algorithm $\mathbf{A}$ using specific types of queries *exactly learns* a class $\mathbb{C}$ of ω-word languages iff for every $L \in \mathbb{C}$, the algorithm makes a finite number of queries of the specified types about L and eventually halts and outputs a DBW acceptor M such that $[\![M]\!] = L$. The algorithm runs in polynomial time iff there is a fixed polynomial p such that for every $L \in \mathbb{C}$, at every point the number of steps used by $\mathbf{A}$ is bounded by $p(n, m)$, where n is the size of the smallest DBW acceptor recognizing L, and m is the maximum length of any counterexample $\mathbf{A}$ has received up to that point.

The case of a learning algorithm for ω-tree languages is analogous, except that the examples and counterexamples are given by regular ω-tree automata, and the hypotheses

provided to equivalence or subset queries are represented by DBT acceptors. We also consider cases in which the inputs to equivalence or subset queries may be NBW or NBT acceptors.

5 Framework of a reduction

Suppose $\mathbf{A}$ is a learning algorithm that uses membership and equivalence queries and exactly learns a class $\mathbb{C} \subseteq \mathbb{DBW}$. We shall describe an algorithm $\mathbf{A}_{Trees}$ that uses membership and equivalence queries and exactly learns the derived class $Trees_d(\mathbb{C})$ of ω-tree languages. Note that $Trees_d(\mathbb{C}) \subseteq \mathbb{DBT}$.

The algorithm $\mathbf{A}_{Trees}$ with target concept $Trees_d(L)$ simulates algorithm $\mathbf{A}$ with target concept L. In order to do so, $\mathbf{A}_{Trees}$ must correctly answer membership and equivalence queries from $\mathbf{A}$ about L by making one or more membership and/or equivalence queries of its own about $Trees_d(L)$. Before describing $\mathbf{A}_{Trees}$ we establish some basic results about regular ω-trees.

5.1 Testing acceptance of a regular ω-tree

We describe a polynomial time algorithm $Accepted?(A_t, M)$ that takes as input a regular ω-tree t represented by a regular ω-tree automaton $A_t = (Q_1, q_{0,1}, \delta_1, \tau_1)$ and a DBW acceptor $M = (Q_2, q_{0,2}, \delta_2, F_2)$ and determines whether or not $M^{T,d}$ accepts t. If not, it also outputs a pair (u, v) of finite words such that $u(v)^\omega \in (paths(t) \setminus [\![M]\!])$.

Algorithm 1 : $Accepted?(A_t, M)$

Require: $A_t = (Q_1, q_{0,1}, \delta_1, \tau_1)$ representing t;
$M = (Q_2, q_{0,2}, \delta_2, F_2)$, a complete DBW acceptor
Ensure: Return "yes" if $M^{T,d}$ accepts t
else return "no" and (u, v) with $u(v)^\omega \in (paths(t) \setminus [\![M]\!])$.

```
let Q = Q_1 × Q_2
let q_0 = (q_{0,1}, q_{0,2})
for all (q_1, q_2) ∈ Q and i ∈ D do
    let δ((q_1, q_2), i) = (δ_1(q_1, i), δ_2(q_2, τ_1(q_1, i)))
let F = {(q_1, q_2) | q_2 ∈ F_2}
let M' = (Q, q_0, δ, F)
if [[M']] = D^ω then
    return "yes"
else
    find x(y)^ω ∈ (D^ω \ [[M']])
    let u(v)^ω = t(x(y)^ω)
    return "no" and (u, v)
```

We may assume M is complete by adding (if necessary) a new non-accepting sink state and directing all undefined transitions to the new state. We construct a DBW acceptor M' over the alphabet $D = \{1, \ldots, d\}$ by combining A_t and M as follows. The states are $Q = Q_1 \times Q_2$, the initial state is $q_0 = (q_{0,1}, q_{0,2})$, the set of accepting states is $F = \{(q_1, q_2) \mid q_2 \in F_2\}$, and the transition function δ is defined by $\delta((q_1, q_2), i) = (\delta_1(q_1, i), \delta_2(q_2, \tau_1(q_1, i)))$ for all $(q_1, q_2) \in Q$ and $i \in D$. For each transition, the output of the regular ω-tree automaton A_t is the input of the DBW acceptor M.

An infinite path π in t corresponds to an ω-word $z \in D^\omega$, giving the sequence of directions from the root. The unique run of M' on z traverses a sequence of states; if we project out the first component, we get the run of A_t on z, and if we project out the second component, we get the run of M on $t(\pi)$. Then $M^{T,d}$ accepts t iff M accepts $t(\pi)$ for every infinite path π, which is true iff $[\![M']\!] = D^\omega$. This in turn is true iff every nonempty accessible recurrent set of states in M' contains at least one element of F.

A set S of states is *recurrent* iff for all $q, q' \in S$, there is a nonempty finite word v such that $\delta(q, v) = q'$ and for every prefix u of v we have $\delta(q, u) \in S$. A set S of states is *accessible* iff for every $q \in S$ there exists a finite word u such that $\delta(q_0, u) = q$.

The algorithm to test whether $M^{T,d}$ accepts t first removes from the transition graph of M' all states that are not accessible. It then removes all states in F and tests whether there is any cycle in the remaining graph. If not, then $M^{T,d}$ accepts t. Otherwise, there is a state q in Q and finite words $x \in D^*$ and $y \in D^+$ such that $\delta(q_0, x) = q$ and $\delta(q, y) = q$ and none of the states traversed from q to q along the path y are in F. Thus, $x(y)^\omega$ is an ultimately periodic path π that does not visit F infinitely often, and letting $u(v)^\omega$ be $t(x(y)^\omega)$, we have $u(v)^\omega \in (\mathit{paths}(t) \setminus [\![M]\!])$, so the pair (u, v) is returned in this case. The required graph operations are standard and can be accomplished in time polynomial in $|M|$ and $|A_t|$.

5.2 Representing a language as paths of a tree

When the algorithm $\mathbf{A}_{\mathit{Trees}}$ makes a membership query about $\mathit{Trees}_d(L)$ with a regular ω-tree t, the answer is "yes" if $\mathit{paths}(t) \subseteq L$ and "no" otherwise. Thus, this query has the effect of a restricted subset query about L with $\mathit{paths}(t)$. But this does not give us restricted subset queries about L for arbitrary $\mathbb{DBW}$ languages. We show below that if $[\![M]\!]$ is a safety language, then it may be represented as $\mathit{paths}(t)$ for a regular ω-tree t.

An ω-word language L is a *safety language* iff L is a regular ω-word language and for every ω-word w not in L, there exists a finite prefix x of w such that no ω-word with prefix x is in L. A language is safety iff it is in the class $\mathbb{D}_\mathbb{W}\mathbb{CW}$. An alternative characterization is that there is an NBW acceptor $M = (Q, q_0, \delta, Q)$, all of whose states are accepting, such that $[\![M]\!] = L$. In this case, the acceptor is typically not complete (otherwise it recognizes Σ^ω). An example of a language in $\mathbb{D}_\mathbb{W}\mathbb{PW}$ that is not a safety language is $a^*b^*(a)^\omega$. Although b^ω is not in the language, every finite prefix b^k is a prefix of some ω-word in the language.

The following lemmas show a close connection between NBW acceptors of safety languages and $\mathit{paths}(t)$ for regular ω-trees t. The constructions in Lemmas 4 and 5 can be done in polynomial time. Lemma 5 is used in the proof of Lemma 9.

▶ **Lemma 4.** *If A_t is a regular ω-tree automaton representing an ω-tree t, then $\mathit{paths}(t)$ is a safety language recognizable by an NBW acceptor M with $|M| = |A_t|$.*

For the converse, representing a safety language as the paths of a regular ω-tree, we require a lower bound on d, the arity of the tree. If $M = (Q, q_0, \delta, F)$ is an NBW acceptor and $q \in Q$, we define the set of transitions out of q to be $\mathit{transitions}(q) = \{(\sigma, r) \mid \sigma \in \Sigma \wedge r \in \delta(q, \sigma)\}$. We define the *out-degree* of M to be the maximum over $q \in Q$ of the cardinality of $\mathit{transitions}(q)$.

▶ **Lemma 5.** *Let L be a safety language recognized by NBW acceptor $M = (Q, q_0, \delta, Q)$. Suppose the out-degree of M is at most d. Then there is a d-ary regular ω-tree t such that $\mathit{paths}(t) = L$, and t is representable by A_t with $|A_t| = |M|$.*

The NBW acceptor in the proof of Lemma 4 can be determinized via the subset construction to give a DBW acceptor of size at most $2^{|A_t|}$ recognizing the same language. In the worst case this exponential blow up in converting a regular ω-tree automaton to a DBW acceptor is necessary, as shown by the following.

▶ **Lemma 6.** *There exists a family of regular ω-trees $t_1, t_2, \ldots$ such that t_n can be represented by a regular ω-tree automaton of size $n + 2$, but the smallest DBW acceptor recognizing $\mathit{paths}(t_n)$ has size at least 2^n.*

If t is a d-ary regular ω-tree represented by the regular ω-tree automaton A_t, then $\mathit{acceptor}(A_t)$ denotes the NBW acceptor M recognizing $\mathit{paths}(t)$ constructed from A_t in the proof of Lemma 4. Note that the out-degree of $\mathit{acceptor}(A_t)$ is at most d.

If M is an NBW acceptor such that $[\![M]\!]$ is a safety language and the out-degree of M is at most d, then $\mathit{tree}_d(M)$ denotes the regular ω-tree automaton A_t constructed from M in the proof of Lemma 5. We also use the notation $\mathit{tree}_d(L)$ if L is a safety language and the implied acceptor for L is clear.

For example, given finite words $u \in \Sigma^*$ and $v \in \Sigma^+$, the singleton set containing $u(v)^\omega$ is a safety language recognized by a DBW of out-degree 1 and size linear in $|u| + |v|$. Then $\mathit{tree}_d(u(v)^\omega)$ represents the d-ary tree all of whose infinite paths are labeled with $u(v)^\omega$.

6 The algorithm $\mathbf{A}_{\mathit{Trees}}$

We now describe the algorithm $\mathbf{A}_{\mathit{Trees}}$, which learns $\mathit{Trees}_d(L)$ by simulating the algorithm $\mathbf{A}$ and answering the membership and equivalence queries of $\mathbf{A}$ about L. It is summarized in Algorithm 2 (at the end of the document).

Algorithm 2 : $\mathbf{A}_{\mathit{Trees}}$

Require: Learning algorithm $\mathbf{A}$ for $\mathbb{C}$;
MQ and EQ access to $\mathit{Trees}_d(L)$ for $L \in \mathbb{C}$
Ensure: Acceptor $M^{T,d}$ such that $[\![M^{T,d}]\!] = \mathit{Trees}_d(L)$

while $\mathbf{A}$ has not halted **do**
  if next step of $\mathbf{A}$ is not a query **then**
    simulate next step of $\mathbf{A}$
  else if $\mathbf{A}$ asks MQ(u, v) about L **then**
    answer $\mathbf{A}$ with MQ$(\mathit{tree}_d(u(v)^\omega))$ about $\mathit{Trees}_d(L)$
  else if $\mathbf{A}$ asks EQ(M) about L **then**
    ask EQ$(M^{T,d})$ about $\mathit{Trees}_d(L)$
    if EQ$(M^{T,d})$ answer is "yes" **then**
      return $M^{T,d}$ and halt
    else {EQ$(M^{T,d})$ answer is counterexample tree t given by A_t}
      if *Accepted?*(A_t, M) returns "no" with value (u, v) **then**
        answer $\mathbf{A}$ with (u, v)
      else {*Accepted?*(A_t, M) returns "yes"}
        let $M' = \mathit{acceptor}(A_t)$
        for all accepting states q of M' **do**
          simulate in parallel *Findctrex*(M', q)
          terminate all computations and answer $\mathbf{A}$ with the first (u, v) returned
[$\mathbf{A}$ halts with output M]
return $M^{T,d}$ and halt

If $\mathbf{A}$ asks a membership query with (u, v) then $\mathbf{A}_{\mathit{Trees}}$ constructs the regular ω-tree automaton $\mathit{tree}_d(u(v)^\omega)$ representing the d-ary regular ω-tree all of whose infinite paths are labeled $u(v)^\omega$, and makes a membership query with $\mathit{tree}_d(u(v)^\omega)$. Because $u(v)^\omega \in L$ iff the tree represented by $\mathit{tree}_d(u(v)^\omega)$ is in $\mathit{Trees}_d(L)$, the answer to the query about $\mathit{tree}_d(u(v)^\omega)$ is simply given to $\mathbf{A}$ as the answer to its membership query about (u, v).

For an equivalence query from $\mathbf{A}$ specified by a DBW acceptor M, the algorithm $\mathbf{A}_{\mathit{Trees}}$ constructs the corresponding DBT acceptor $M^{T,d}$, which recognizes $\mathit{Trees}_d([\![M]\!])$, and makes an equivalence query with $M^{T,d}$. If the answer is "yes", the algorithm $\mathbf{A}_{\mathit{Trees}}$ has succeeded in learning the target ω-tree language $\mathit{Trees}_d(L)$ and outputs $M^{T,d}$ and halts. Otherwise, the counterexample returned is a regular ω-tree t in $[\![M^{T,d}]\!] \oplus \mathit{Trees}_d(L)$, represented by a regular ω-tree automaton A_t. A call to the algorithm *Accepted?*(A_t, M) determines whether $M^{T,d}$ accepts t. If $M^{T,d}$ rejects t, then $t \in \mathit{Trees}_d(L)$ and t is a *positive counterexample*. If $M^{T,d}$ accepts t, then $t \notin \mathit{Trees}_d(L)$ and t is a *negative counterexample*. We consider these two cases.

If t is a positive counterexample then we know that $t \in \mathit{Trees}_d(L)$ and therefore $\mathit{paths}(t) \subseteq L$. Because $t \notin [\![M^{T,d}]\!]$, the acceptor M must reject at least one infinite path in t. In this case, the algorithm $\mathit{Accepted?}(A_t, M)$ returns a pair of finite words (u, v) such that $u(v)^\omega \in (\mathit{paths}(t) \setminus [\![M]\!])$, and therefore $u(v)^\omega \in (L \setminus [\![M]\!])$. The algorithm $\mathbf{A}_{\mathit{Trees}}$ returns the positive counterexample (u, v) to $\mathbf{A}$ in response to its equivalence query with M.

If t is a negative counterexample, that is, $t \in ([\![M^{T,d}]\!] \setminus \mathit{Trees}_d(L))$, then we know that $\mathit{paths}(t) \subseteq [\![M]\!]$, but at least one element of $\mathit{paths}(t)$ is not in L, so $([\![M]\!] \setminus L) \neq \emptyset$. Ideally, we would like to extract an ultimately periodic ω-word $u(v)^\omega \in (\mathit{paths}(t) \setminus L)$ and provide (u, v) to $\mathbf{A}$ as a negative counterexample in response to its equivalence query with M.

If we could make an *unrestricted subset query* with $\mathit{paths}(t)$ about L, then the counterexample returned would be precisely what we need.

As noted previously, if t is any regular ω-tree then we can simulate a restricted subset query with $\mathit{paths}(t)$ about L by making a membership query with t about $\mathit{Trees}_d(L)$, because $\mathit{paths}(t) \subseteq L$ iff $t \in \mathit{Trees}_d(L)$. In order to make use of this, we next show how to use restricted subset queries about L to implement an unrestricted subset query about L.

6.1 Restricted subset queries

To establish basic techniques, we show how to reduce unrestricted subset queries to restricted subset queries for nondeterministic or deterministic finite acceptors over finite words. Suppose $L \subseteq \Sigma^*$ and we may ask restricted subset queries about L. In such a query, the input is a nondeterministic (resp., deterministic) finite acceptor M, and the answer is "yes" if $[\![M]\!]$ is a subset of L, and "no" otherwise. If the answer is "no", we show how to find a shortest counterexample $u \in ([\![M]\!] \setminus L)$ in time polynomial in $|M|$ and $|u|$.

▶ **Theorem 7.** *There is an algorithm $\mathbf{R}^*$ with input M, an NFW (resp., DFW), and restricted subset query access to a language L with NFW (resp., DFW) acceptors as inputs, that correctly answers the unrestricted subset query with M about L. If L is recognized by a DFW T_L, then $\mathbf{R}^*(M)$ runs in time bounded by a polynomial in $|M|$ and $|T_L|$.*[1]

The idea of the proof is to first establish the minimal length ℓ of a counterexample, and then try to extend the prefix ϵ letter by letter until obtaining a full length minimal counterexample (see Appendix A). Note that trying to establish a prefix of a counterexample letter by letter, without obtaining a bound first, may not terminate. For instance, if $L = \Sigma^* \setminus a^*b$, one can establish the sequence of prefixes $\epsilon, a, aa, aaa, \ldots$ and never reach a counterexample.

We now turn to the ω-word case.

▶ **Theorem 8.** *There is an algorithm $\mathbf{R}^\omega$ with input M and restricted subset query access about L, (a language recognized by a DBW acceptor T_L) that correctly answers the unrestricted subset query with M about L. $\mathbf{R}^\omega(M)$ runs in time bounded by a polynomial in $|M|$ and $|T_L|$. If M is a DBW acceptor, then all the restricted subset queries will also be with DBW acceptors.*

For the sake of generality, the proof considers subset queries with NBW acceptors. The procedure $\mathbf{R}^\omega(M)$ takes as input an NBW acceptor M, and has restricted subset query access (with NBW acceptors as inputs) to L; it is summarized in Algorithm 3. It first asks a restricted subset query with M about L, returning the answer "yes" if its query is answered

[1] The cardinality of Σ is treated as a constant.

"yes". Otherwise, for each $q \in F$, it constructs the acceptor $M_q = (Q, q_0, \delta, \{q\})$ with the single accepting state q and asks a restricted subset query with M_q about L, until the first query answered "no". There will be at least one such query answered "no" because any element of $([\![M]\!] \setminus L)$ must visit at least one accepting state q of M infinitely many times, and will therefore be in $[\![M_q]\!]$. The procedure $\mathbf{R}^\omega(M)$ then calls the procedure *Findctrex*(M, q) to find a counterexample to return — i.e. a pair (u, v) such that $u(v)^\omega \in ([\![M]\!] \setminus L)$.

6.2 Producing a counterexample

The first challenge encountered in producing a counterexample, in comparison to the finite word case, is that one needs to work out both the period and the prefix of the counterexample to be found, and the two are correlated. Define $L_{q_0,q}$ to be the set of finite words that lead from the initial state q_0 to the state q in M, and define $L_{q,q}$ to be the set of nonempty finite words that lead from q back to q in M. Because the language $L_{q_0,q} \cdot (L_{q,q})^\omega$ is exactly the set of strings recognized by M_q, we know that $L_{q_0,q} \cdot (L_{q,q})^\omega \setminus L \neq \emptyset$.

The procedure *Findctrex*(M, q) first finds a suitable period, corresponding to a bounded size of a prefix yet to be found, and then finds a prefix of that size in a similar manner to the finite word case.

Since finding the period is more challenging than the prefix, we explain the procedure *Findprefix*(M, q, v) first. The procedure *Findprefix*(M, q, v), summarized in Algorithm 5, finds a prefix word u given a period word v which loops on state q and is guaranteed to be a period of a valid counterexample. It first finds a length k such that there exists $u \in L_{q_0,q}$ of length k such that $uv^\omega \notin L$. Then it finds such a word u symbol by symbol. Note that it uses length and prefix restricted versions of $L_{q_0,q}$.

Finding the periodic part is much more challenging. Indeed, even if one knows that there is a period of the form $(a\Sigma^\ell)^\omega$ for some ℓ then the size of the smallest period may be bigger than $\ell + 1$. For instance, if $L = \Sigma^\omega \setminus (abbaccadd)^\omega$ then there is a period of the form $(a\Sigma^2)^\omega$ but the shortest period of a counterexample is of size 9.

Procedure *Findperiod*(M, q), summarized in Algorithm 6, starts from the condition

$$L_{q_0,q} \cdot (L_{q,q})^\omega \setminus L \neq \emptyset$$

and finds a sequence of words $v_1, v_2, \ldots \in L_{q,q}$ such that for each $n \geq 1$,

$$L_{q_0,q} \cdot (v_1 v_2 \cdots v_n \cdot L_{q,q})^\omega \setminus L \neq \emptyset.$$

For a sufficiently long such sequence, there exists a subsequence $v = (v_i \cdots v_j)$ that is a suitable period word, as we illustrate in Appendix C.2.

The procedure *Nextword*(M, q, y), summarized in Algorithm 7, is called with $y = v_1 v_2 \cdots v_n$ and finds a suitable next word v_{n+1}. After determining a length ℓ, it repeatedly calls the procedure *Nextsymbol*(M, q, y, ℓ, v') to determine the next symbol of a suitable word of length ℓ.

The procedure *Nextsymbol*(M, q, y, ℓ, v'), summarized in Algorithm 8, is called to find a feasible next symbol with which to extend v' in the procedure *Nextword*.

7 Correctness

The main hurdle in proving correctness of algorithm $\mathbf{A}_{Trees}$ is to prove Theorem 8. The polynomial bound in the proof of Theorem 8 is obtained through a sequence of lemmas (Lemmas 12 to 14) bounding the size of the acceptors used in $\mathbf{A}_{Trees}$ subprocedures and the

length restrictions and running time in calls to RSQ made by these procedures. Appendix C.1 deals with bounding the acceptors, and Appendix C.2 deals with the more challenging part, providing the length restrictions. In the following we conclude with the theorem stating the correctness of algorithm $\mathbf{A}_{Trees}$.

The lemmas established in the previous subsection also show the correctness and running time of $Findctrex(M', q)$ when called by $\mathbf{A}_{Trees}$, provided that each RSQ about L is correctly answered and q satisfies the precondition of *Findctrex*.

To complete the consideration of representation issues, we must prove that $\mathbf{A}_{Trees}$ can successfully simulate *Findctrex* as stated in Lemma 9.

▶ **Lemma 9.** *When $\mathbf{A}_{Trees}$ simulates $Findctrex(M', q)$ in response to a negative counterexample t, every* RSQ *can be simulated with a* MQ *about $Trees_d(L)$.*

If q does not satisfy the precondition of *Findctrex*, then the procedure may run forever. However, at least one accepting state q satisfies the precondition, so at least one simulation will halt and return (u, v), at which point $\mathbf{A}_{Trees}$ terminates the other simulations. This concludes the proof of the reduction given by $\mathbf{A}_{Trees}$, whose general statement is given in Theorem 10.

▶ **Theorem 10.** *Suppose $\mathbb{C} \subseteq \mathbb{DBW}$ and $\mathbf{A}$ is a polynomial time algorithm that learns class $\mathbb{C}$ using membership and equivalence queries. Then for every positive integer d there is a polynomial time algorithm $\mathbf{A}_{Trees}$ that learns $Trees_d(\mathbb{C})$ using membership and equivalence queries.*

This theorem, together with Maler and Pnueli's [15] polynomial time algorithm to learn the class of weak regular ω-word languages using membership and equivalence queries proves our main result — Theorem 3.

8 Discussion

We have shown that if $\mathbb{C} \subseteq \mathbb{DBW}$ can be learned in polynomial time with membership and equivalence queries, then $Trees_d(\mathbb{C})$ can be learned in polynomial time with membership and equivalence queries for all $d \geq 1$. Consequently, there is a polynomial time algorithm to learn $Trees_d(\mathbb{DwPW})$ with membership and equivalence queries. We have also shown that there are polynomial time algorithms that implement unrestricted subset queries using restricted subset queries for $\mathbb{DFW}$, $\mathbb{NFW}$, $\mathbb{DBW}$ and $\mathbb{NBW}$.

One open question is whether there is an interesting subclass of $\mathbb{DBW}$ that is larger than $\mathbb{DwPW}$ but still learnable in polynomial time using membership and equivalence queries, to which Theorem 10 would also apply.

References

1. D. Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987.
2. D. Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.
3. D. Angluin. Negative results for equivalence queries. *Mach. Learn.*, 5(2):121–150, 1990.
4. D. Angluin. The class DBW ∩ DCW of ω-languages is identifiable in the limit from positive data and membership queries with polynomial time and data. Technical Report YALEU/DCS/TR-1528, Dept. of Computer Science, Yale University, August 2016.
5. D. Angluin and D. Fisman. Learning regular omega languages. *Theor. Comput. Sci.*, 650:57–72, 2016.

6 U. Boker, D. Kuperberg, O. Kupferman, and M. Skrzypczak. Nondeterminism in the presence of a diverse or unknown future. In *Automata, Languages, and Programming, 40th Inter. Colloq., ICALP*, pages 89–100, 2013.

7 H. Calbrix, M. Nivat, and A. Podelski. Ultimately periodic words of rational ω-languages. In *Proc. of the 9th Inter. Conf. on Mathematical Foundations of Programming Semantics*, pages 554–566, 1994.

8 J. M. Cobleigh, D. Giannakopoulou, and C. S. Păsăreanu. Learning assumptions for compositional verification. In *Proc. of the 9th Inter. Conf. on Tools and Algorithms for the Construction and Analysis of Systems, TACAS*, pages 331–346, 2003.

9 A. Farzan, Y. Chenand, E.M. Clarke, Y. Tsay, and B. Wang. Extending automated compositional verification to the full class of omega-regular languages. In *Tools and Algorithms for the Construction and Analysis of Systems, TACAS*, volume 4963 of *LNCS*, pages 2–17, 2008.

10 E. Grädel, W. Thomas, and T. Wilke, editors. *Automata, Logics, and Infinite Games: A Guide to Current Research*. Springer, 2002.

11 M. Jayasrirani, M. Humrosia Begam, and D. Gnanaraj Thomas. Learning of regular omega-tree languages. In *Grammatical Inference: Algorithms and Applications, 9th Inter. Colloq., ICGI*, pages 295–297, 2008.

12 D. Kuperberg and M. Skrzypczak. On determinisation of good-for-games automata. In *Automata, Languages, and Programming, 42nd Inter. Colloq., ICALP*, pages 299–310, 2015.

13 O. Kupferman, S. Safra, and M. Y. Vardi. Relating word and tree automata. *Annals of Pure and Applied Logic*, 138:126–146, 2006.

14 C. Löding. Automata on infinite trees. Available online from author, 2011.

15 O. Maler and A. Pnueli. On the learnability of infinitary regular sets. *Inform. Comput.*, 118(2):316–326, 1995.

16 Z. Manna and A. Pnueli. A hierarchy of temporal properties. In *Proc. of the Ninth Annual ACM Symposium on Principles of Distributed Computing, PODC*, pages 377–410, 1990.

17 D. Niwiński and I. Walukiewicz. Relating hierarchies of word and tree automata. In *In Symposium on Theoretical Aspects in Computer Science, STACS, LNCS 1373*, pages 320–331. Springer-Verlag, 1998.

18 D. Perrin and J.-É. Pin. *Infinite Words: Automata, Semigroups, Logic and Games*. Elsevier, 2004.

19 M. Rabin. *Automata on infinite objects and Church's problem*. Regional Conf. series in mathematics. Conf. Board of the Math. Sci. by the American Mathematical Society, 1972.

20 F. Vaandrager. Model learning. *Commun. ACM*, 60(2):86–95, 2017.

21 K. W. Wagner. A hierarchy of regular sequence sets. In *4th Symposium on Mathematical Foundations of Computer (MFCS)*, pages 445–449, 1975.

A Proof of Theorem 7

To prove Theorem 7 we first construct an acceptor $M_{\ell,v}$ for $[\![M]\!][\ell, v]$, the length and prefix restricted version of $[\![M]\!]$, given M, ℓ and v as inputs.

▶ **Lemma 11.** *There is a polynomial time algorithm to construct an acceptor $M_{\ell,v}$ for $[\![M]\!][\ell, v]$ given a NFW acceptor M, a nonnegative integer ℓ and a finite word v, such that*

1. *$M_{\ell,v}$ has at most one accepting state, which has no out-transitions,*
2. *the out-degree of $M_{\ell,v}$ is at most the out-degree of M,*
3. *$M_{\ell,v}$ is deterministic if M is deterministic.*

Proof of Theorem 7. For input M, define $M_{[\ell,v]}$ to be the finite acceptor constructed by the algorithm of Lemma 11 to recognize the length and prefix restricted language $[\![M]\!][\ell, v]$.

For $\ell = 0, 1, 2, \ldots$, ask a restricted subset query with $M_{[\ell,\varepsilon]}$, until the first query answered "no". At this point, ℓ is the shortest length of a counterexample in $([\![M]\!] \setminus L)$. Then a counterexample u of length ℓ is constructed symbol by symbol.

Assume we have found a prefix u' of a counterexample of length ℓ in $([\![M]\!] \setminus L)$, with $|u'| < \ell$. For each symbol $\sigma \in \Sigma$ we ask a restricted subset query with $M_{[\ell,u'\sigma]}$, until the first query answered "no". At this point, u' is extended to $u'\sigma$. If the length of $u'\sigma$ is now ℓ, then $u = u'\sigma$ is the desired counterexample; otherwise, we continue extending u'.

Note that if the input M is deterministic, then all of the restricted subset queries are made with deterministic finite acceptors. If L is recognized by a deterministic finite acceptor T_L, then the value of ℓ is bounded by $|M| \cdot |T_L|$, and the algorithm runs in time bounded by a polynomial in $|M|$ and $|T_L|$. ◀

B Algorithms

Algorithm 3 : $\mathbf{R}^{\omega}(M)$, implementing USQ(M)

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor
Ensure: "yes" if $[\![M]\!] \subseteq L$, else "no" and (u, v) s.t. $u(v)^{\omega} \in ([\![M]\!] \setminus L)$

if RSQ(M) = "yes" **then**
  return "yes"
else
  find $q \in F$ such that RSQ(M_q) = "no"
  return "no" and *Findctrex*(M, q)

Algorithm 4 : *Findctrex*(M, q)

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor;
$q \in F$;
$L_{q_0,q} \cdot (L_{q,q})^{\omega} \setminus L \neq \emptyset$
Ensure: (u, v) such that $u(v)^{\omega} \in (L_{q_0,q} \cdot (L_{q,q})^{\omega} \setminus L)$

let v = *Findperiod*(M, q)
let u = *Findprefix*(M, q, v)
return (u, v)

Algorithm 5 : *Findprefix*(M, q, v)

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor;
$q \in F$;
$v \in L_{q,q}$;
$L_{q_0,q} \cdot (v)^{\omega} \setminus L \neq \emptyset$
Ensure: $u \in L_{q_0,q}$ such that $u(v)^{\omega} \in (L_{q_0,q} \cdot (v)^{\omega} \setminus L)$

search for nonnegative integer k such that RSQ$(L_{q_0,q}[k] \cdot (v)^{\omega})$ = "no"
let $u = \varepsilon$
while $|u| < k$ **do**
  find $\sigma \in \Sigma$ such that RSQ$(L_{q_0,q}[k, u\sigma] \cdot (v)^{\omega})$ = "no"
  set $u = u \cdot \sigma$
return u

Algorithm 6 : *Findperiod*(M, q)

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor;
$q \in F$;
$L_{q_0,q} \cdot (L_{q,q})^\omega \setminus L \neq \emptyset$
Ensure: $v \in L_{q,q}$ such that $L_{q_0,q} \cdot (v)^\omega \setminus L \neq \emptyset$

let $y = \varepsilon$
for all integers $n = 1, 2, 3, \ldots$ **do**
  let $v_n = \mathit{Nextword}(M, q, y)$
  set $y = y \cdot v_n$
  for integers i, j with $1 \leq i \leq j \leq n$ **do**
    for $k = 0$ to $n|M|$ **do**
      if RSQ$(L_{q_0,q}[k] \cdot (v_i \cdots v_j)^\omega) =$ "no" **then**
        return $v = v_i \cdots v_j$

Algorithm 7 : *Nextword*(M, q, y)

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor;
$q \in F$;
$y \in L_{q,q}$ or $y = \varepsilon$;
$L_{q_0,q} \cdot (y \cdot L_{q,q})^\omega \setminus L \neq \emptyset$
Ensure: $v' \in L_{q,q}$ such that $L_{q_0,q} \cdot (y \cdot v' \cdot L_{q,q})^\omega \setminus L \neq \emptyset$

search for integers $k, \ell \geq 0$ s.t. RSQ$(L_{q_0,q}[k] \cdot (y \cdot L_{q,q}[\ell])^\omega) =$ "no"
let $v' = \varepsilon$
while $|v'| < \ell$ **do**
  let $\sigma = \mathit{Nextsymbol}(M, q, y, \ell, v')$
  set $v' = v' \cdot \sigma$
return v'

Algorithm 8 : *Nextsymbol*(M, q, y, ℓ, v')

Require: RSQ access to L;
$M = (Q, q_0, \delta, F)$, an NBW acceptor;
$q \in F$;
$y \in L_{q,q}$;
$v' \in \Sigma^*$, $|v'| < \ell$;
$L_{q_0,q} \cdot (y \cdot L_{q,q}[\ell, v'] \cdot L_{q,q})^\omega \setminus L \neq \emptyset$
Ensure: $\sigma \in \Sigma$ such that $L_{q_0,q} \cdot (y \cdot L_{q,q}[\ell, v'\sigma] \cdot L_{q,q})^\omega \setminus L \neq \emptyset$

find integers $k \geq 0$, $m \geq 1$, and $\sigma \in \Sigma$ such that
RSQ$(L_{q_0,q}[k] \cdot (y \cdot L_{q,q}[\ell, v'\sigma] \cdot L_{q,q}[m])^\omega) =$ "no"
return σ

C Correctness and bounds

C.1 Bounding the acceptors

We turn to the representation (as NBW or DBW acceptors) of the languages used in restricted subset queries by $\mathbf{R}^\omega(M)$ and its subprocedures. We consider the size, out-degree, and time to construct the acceptors.

In $\mathbf{R}^\omega(M)$, there is a restricted subset query with M itself, and if that query is answered "no", a sequence of restricted subset queries with M_q for accepting states q until an answer of "no". Clearly, if M is an NBW acceptor, each M_q is an NBW acceptor of the same size and out-degree and is easily constructed from M, and similarly if M is a DBW acceptor.

The restricted subset queries made in *Findctrex* and its subprocedures are of the form $P \cdot (S)^\omega$, where P is a length and prefix restricted version of $L_{q_0,q}$ and S is a concatenation of (at most) a finite word and two length and prefix restricted versions of $L_{q,q}$. Thus we consider the operations of concatenation and ω-repetition of regular languages of finite words.

These operations are particularly simple for DFW or NFW acceptors in *special form*, that is, containing at most one accepting state, which has no out-transitions defined. In general, any NFW acceptor can be converted to special form, possibly at the cost of increasing its out-degree. A regular language of finite words is recognized by a DFW acceptor in special form iff it is prefix-free.

However, if M is an NBW (resp., DBW) acceptor, then the finite word languages $L_{q_0,q}$ and $L_{q,q}$ are recognized by easily constructed NFW (resp., DFW) acceptors of size at most $|M|$ and out-degree at most the out-degree of M. Lemma 11 shows that the length and prefix restricted versions of $L_{q_0,q}$ and $L_{q,q}$ are recognized by NFW (resp., DFW) acceptors in special form which may be constructed in time polynomial in $|M|$, ℓ, and $|v|$ and have out-degree at most the out-degree of M.

▶ **Lemma 12.** *Suppose M_1 is an NFW acceptor in special form and M_2 is an NFW or NBW acceptor. Then an acceptor M for $[\![M_1]\!] \cdot [\![M_2]\!]$ can be constructed such that*

1. *$|M| \leq |M_1| + |M_2|$,*
2. *the out-degree of M is at most the maximum of out-degrees of M_1 and M_2,*
3. *M can be constructed in polynomial time,*
4. *M is deterministic if M_1 and M_2 are deterministic,*
5. *M is an NFW in special form if M_2 is an NFW in special form.*

▶ **Lemma 13.** *Suppose M_1 is an NFW acceptor in special form. Then an NBW acceptor M for $[\![M_1]\!]^\omega$ can be constructed such that*

1. *$|M| \leq |M_1|$,*
2. *the out-degree of M is at most the out-degree of M_1,*
3. *M can be constructed in polynomial time,*
4. *M is deterministic if M_1 is deterministic.*

The above give us the following corollary for the procedure $\mathbf{R}^\omega$.

▶ **Corollary 14.** *When the input to $\mathbf{R}^\omega(M)$ is an NBW (resp., DBW) acceptor M, each* RSQ *can be made with an NBW (resp., DBW) acceptor whose out-degree is at most the out-degree of M and can be constructed in time polynomial in $|M|$ and parameters giving the length restrictions and the lengths of any words that appear.*

C.2 Length restrictions and time bounds

We now turn to establish the correctness and running time of the subprocedures. The first two lemmas allow us to bound the parameters giving the length restrictions in inputs to RSQ.

▶ **Lemma 15.** *Let $S \subseteq L_{q,q}$ and suppose $L_{q_0,q} \cdot (S)^\omega \setminus L \neq \emptyset$. Then for some $k < |M| \cdot |T_L|$ we have $L_{q_0,q}[k] \cdot (S)^\omega \setminus L \neq \emptyset$.*

▶ **Lemma 16.** *Let $S \subseteq L_{q,q}$ and suppose $L_{q_0,q} \cdot (S \cdot L_{q,q})^\omega \setminus L \neq \emptyset$. Then for some $k, \ell < |M| \cdot |T_L|$, we have that $L_{q_0,q}[k] \cdot (S \cdot L_{q,q}[\ell])^\omega \setminus L \neq \emptyset$.*

We now prove the correctness and polynomial running time of *Findprefix* and *Findperiod*, which establishes the correctness and polynomial running time of *Findctrex*.

▶ **Lemma 17.** *Assume $v \in L_{q,q}$ is such that*

$L_{q_0,q} \cdot (v)^\omega \setminus L \neq \emptyset$.

Then in time polynomial in $|M|$, $|T_L|$ and $|v|$, the procedure Findprefix(M, q, v) returns a word $u \in L_{q_0,q}$ such that

$u(v)^\omega \in (L_{q_0,q} \cdot (v)^\omega \setminus L)$.

The procedure *Findperiod* depends on the procedures *Nextword* and *Nextsymbol*. The next lemma establishes the correctness and running time of the procedure *Nextsymbol*.

▶ **Lemma 18.** *Suppose ℓ is a positive integer, $y \in L_{q,q}$ or $y = \varepsilon$ and $v' \in \Sigma^*$ is such that $|v'| < \ell$ and we have*

$L_{q_0,q} \cdot (y \cdot L_{q,q}[\ell, v'] \cdot L_{q,q})^\omega \setminus L \neq \emptyset$.

Then in time polynomial in $|M|$, $|T_L|$, $|y|$ and ℓ, Nextsymbol(M, q, y, ℓ, v') finds a symbol $\sigma \in \Sigma$ such that

$L_{q_0,q} \cdot (y \cdot L_{q,q}[\ell, v'\sigma] \cdot L_{q,q})^\omega \setminus L \neq \emptyset$.

▶ **Lemma 19.** *Suppose $y \in L_{q,q}$ or $y = \varepsilon$ is such that*

$L_{q_0,q} \cdot (y \cdot L_{q,q})^\omega \setminus L \neq \emptyset$.

Then in time bounded by a polynomial in $|M|$, $|T_L|$ and $|y|$, Nextword(M, q, y) returns a word $v' \in L_{q,q}$ of length bounded by $|M| \cdot |T_L|$ such that

$L_{q_0,q} \cdot (yv' \cdot L_{q,q})^\omega \setminus L \neq \emptyset$.

The next lemma shows that *Findperiod* calls *Nextword* at most $|T_L|$ times.

▶ **Lemma 20.** *Suppose $v_1, v_2, \ldots, v_n \in L_{q,q}$ are such that*

$L_{q_0,q} \cdot (v_1 v_2 \cdots v_n \cdot L_{q,q})^\omega \setminus L \neq \emptyset$.

Also suppose that the number of states of T_L is less than n. Then there exist integers i and j with $1 \leq i \leq j \leq n$ such that

$L_{q_0,q} \cdot (v_i v_{i+1} \cdots v_j)^\omega \setminus L \neq \emptyset$.

The final lemma establishes the correctness and polynomial running time of the procedure *Findperiod*.

▶ **Lemma 21.** *Suppose $L_{q_0,q} \cdot (L_{q,q})^\omega \setminus L \neq \emptyset$. Then, in polynomial time in $|M|$ and $|T_L|$, the procedure Findperiod(M, q) with restricted query access to L returns a period word v satisfying the condition*

$L_{q_0,q} \cdot (v)^\omega \setminus L \neq \emptyset$.

These lemmas combine to prove Theorem 8, giving a polynomial time reduction of unrestricted subset queries to restricted subset queries for NBW acceptors (resp., DBW acceptors.)