Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 344 (2019) 189–208

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

On Distinguishing Sets of Structures by First-Order Sentences of Minimal Quantifier Rank

Thiago Alves Rocha[1]

Department of Computing
Federal Institute of Ceará
Maracanaú, Brazil

Ana Teresa Martins[2,4]

Department of Computing
Federal University of Ceará
Fortaleza, Brazil

Francicleber Martins Ferreira[3]

Department of Computing
Federal University of Ceará
Quixadá, Brazil

Abstract

We investigate the distinguishability of sets of relational structures concerning a class of structures in the following sense: for a fixed class of structures, given two sets of structures in this class, find a first-order formula of minimal quantifier rank that distinguishes one set from the other. We consider the following classes of structures: monadic structures, equivalence structures, and disjoint unions of linear orders. We use results of the Ehrenfeucht–Fraïssé game on these classes of structures in order to design an algorithm to find such a sentence. For these classes of structures, the problem of determining if the Duplicator has a winning strategy in an Ehrenfeucht–Fraïssé game is solved in polynomial time. We also introduce the distinguishability sentences which are sentences that distinguish between two given structures. We define the distinguishability sentences based on necessary and sufficient conditions for a winning strategy in an Ehrenfeucht–Fraïssé game. Our algorithm returns a boolean combination of such sentences. We also show that any first-order sentence is equivalent to a boolean combination of distinguishability sentences. Finally, we also show that our algorithm's running time is polynomial in the size of the input.

Keywords: Distinguishability, Ehrenfeucht–Fraïssé Games, Finite Model Theory.

[1] Email: thiago.alves@ifce.edu.br

[2] Email: ana@dc.ufc.br

[3] Email: francicleber@dc.ufc.br

[4] This author was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under the grant number 424188/2016-3.

https://doi.org/10.1016/j.entcs.2019.07.012

1 Introduction

Ehrenfeucht–Fraïssé games (EF games, for short) [6] is a fundamental technique of finite model theory [4,17,8] in proving the inexpressibility of certain properties in first-order logic (FO). The EF game is played on two structures by two players, the Spoiler and the Duplicator. If the Spoiler has a winning strategy for r rounds of such a game, it means that the structures can be distinguished by a first-order sentence φ whose quantifier rank is at most r, i.e., φ holds in exactly one of these structures. Besides providing a tool to measure the expressive power of a logic, EF games allow one to investigate the similarity between structures [19]. In a game played on structures $\mathcal{A}$ and $\mathcal{B}$, the EF-similarity between $\mathcal{A}$ and $\mathcal{B}$ is the minimum number of rounds such that the Spoiler has a winning strategy. Necessary and sufficient conditions characterizing the winning strategies for both players are needed in order to explore EF-similarity.

Explicit conditions characterizing winning strategies for the players on some standard classes of finite structures are provided in [15]. Examples of such classes are monadic structures (MS) and equivalence structures (ES). Besides, it is well known necessary and sufficient conditions characterizing the winning strategies on linear orders (LO) [17]. Using these results, the EF-similarity can be computed in polynomial time in the size of the structures.

Given a structure $\mathcal{A}$ and a natural number r, an r-Hintikka formula $\varphi^r_{\mathcal{A}}$ is a formula that describes the properties of $\mathcal{A}$ on EF games with r rounds [4]. An r-Hintikka formula $\varphi^r_{\mathcal{A}}$ holds exactly on all structures $\mathcal{B}$ such that the Duplicator has a winning strategy for the EF game with r rounds on $\mathcal{A}$ and $\mathcal{B}$. Also, $\varphi^r_{\mathcal{A}}$ has size exponential in the size of $\mathcal{A}$. Besides, any first-order formula is equivalent to a disjunction of Hintikka formulas.

An algorithm to deal with the problem of finding a formula of minimal quantifier rank that distinguishes two sets of structures over an arbitrary vocabulary is presented in [14]. An important part of this algorithm is the use of Hintikka formulas. As this algorithm works for arbitrary finite relational structures, it runs in exponential time. A general system for learning formulas defining board game rules uses this algorithm. These results are also used in reduction finding [13]. In [23], the authors define a variation of the problem in [14] which considers samples of classified strings.

In this work, we study a variation of the problem introduced in [14] when the class of structures is fixed. For a fixed class of structures $\mathcal{C}$, a sample $S = (P, N)$ consists of two finite sets $P, N \subseteq \mathcal{C}$ such that for each $\mathcal{A} \in P$, $\mathcal{B} \in N$, $\mathcal{A}$ and $\mathcal{B}$ are not isomorphic. For a fixed class of structures $\mathcal{C}$, given a sample S of structures in $\mathcal{C}$, the task is to find a first-order sentence φ_S of minimal quantifier rank that is consistent with S, i.e., it holds in all structures in P and does not hold in any structure in N. The size of the sample is the sum of the lengths of all structures in the sample. We call this problem the distinguishability problem.

We define an algorithm for the distinguishability problem on the following classes of structures: MS, ES, and disjoint unions of linear orders (DULO). We consider

MS and ES because necessary and sufficient conditions for a winning strategy of the players in an EF game on these classes are provided in the literature [15]. For DULO, we show our result on a characterization of the winning strategies for both players. We use these characterization results on EF games in order to design an algorithm to find a sentence of minimal quantifier rank which is consistent with the sample. Also, for MS, ES, and DULO, the EF-similarity can be computed in polynomial time in the size of the structures. Using these results, we show that our algorithm runs in polynomial time in the size of the sample. Therefore, this result improves the one in [14] for MS, ES, and DULO. We expect that our results can be adapted to other classes such as trees and strings with a built-in linear order relation.

In Artificial Intelligence, automated planning is the process of automatically constructing a sequence of actions that achieve a goal given some initial state [7]. In the elementary blocks world planning, states consist of a set of cubic blocks, with the same size and color, sitting on a table. A robot can pick up a block and moves it to another position, either onto the table or on the top of some other block [11]. Therefore, disjoint unions of linear orders are compelling because we may model a state of the elementary blocks world by using them [2]. We think that our results can be useful when it is required to compute sentences defining initial and goal states from positive and negative examples. For instance, one can use our results to obtain the formula $\neg\exists x_1 \exists x_2 (x_1 < x_2)$ which expresses that all blocks are on the table.

As the size of a Hintikka formula is exponential in the size of a given structure, our algorithm does not use Hintikka formulas. In our case, we define what we call the distinguishability sentences. They are defined based on conditions characterizing the winning strategies for the Spoiler. In this way, given two structures $\mathcal{A}, \mathcal{B}$ and a natural number r, we show that the distinguishability sentences hold in $\mathcal{A}$, do not hold in $\mathcal{B}$, and they have quantifier rank at most r. This result is essential for the definition of our algorithm and to guarantee its correctness. We define distinguishability sentences for each particular class of structures we are considering. Then, distinguishability sentences are easy to interpret because they express properties specific to the class. We also define distinguishability sentences in a way such that they have polynomial size. This result is essential in ensuring that our algorithm runs in polynomial time in the size of the sample. We also show that any first-order formula is equivalent to a boolean combination of distinguishability sentences.

A new logical framework to find a formula given a sample, also with a model-theoretic approach, can be found in [10,9]. In this framework, the input is only one structure, and its elements are classified as positive or negative. The problem is to find a hypothesis consistent with the classified elements where this hypothesis is a first-order formula in [10] and a monadic second-order formula in [9]. Also, [9] only considers strings as the input structure. The first main difference is that, in our approach, a sample consists of many structures classified as positive or negative. Second, the algorithm in [10] assumes that the quantifier rank is fixed while we

obtain a sentence of minimal quantifier rank.

Another logical framework for a similar problem is Inductive Logic Programming (ILP) [20,21,3]. ILP uses logic programming as a uniform representation for the sample and hypotheses. Therefore, due to the fundamental difference between the framework of ILP and our approach, there is no direct relationship between ILP and our work. Then, techniques used in the framework of ILP cannot easily be applied in our approach.

We organize this paper as follows: in Section 2, we give the basic definitions and results on Ehrenfeucht–Fraïssé games and Hintikka formulas. Moreover, in this section, we also give results on EF games and EF-similarity for monadic structures and equivalence structures. In Section 3, we show conditions characterizing the winning strategies for both players on disjoint union of linear orders. In Section 4, for each class of structures we are considering, we introduce the distinguishability sentences and provide some useful properties. In Section 5, we propose our algorithm for the distinguishability problem, we give an example of how it works, and we also show that the algorithm is correct. Finally, we conclude and show some future directions in Section 6.

2 Ehrenfeucht–Fraïssé Games

In this section, we present the basic notions about Ehrenfeucht–Fraïssé Games. We assume some background from first-order logic. For details, see [4,5]. In what follows, we are concerned with relational structures. The size of a first-order formula φ is the number of symbols occurring in φ. The quantifier rank $qr(\varphi)$ of a formula φ is the depth of nesting of quantifiers in φ as in the following:

Definition 2.1 [Quantifier Rank] Let φ be a first-order formula. The quantifier rank of φ, written $qr(\varphi)$, is defined as

$$qr(\varphi) := \begin{cases} 0, & \text{if } \varphi \text{ is atomic} \\ max(qr(\varphi_1), qr(\varphi_2)), & \text{if } \varphi = \varphi_1 \square \varphi_2 \text{ such that } \square \in \{\wedge, \vee, \leftarrow\} \\ qr(\psi), & \text{if } \varphi = \neg\psi \\ qr(\psi) + 1, & \text{if } \varphi = Qx\psi \text{ such that } Q \in \{\exists, \forall\} \end{cases}$$

Given a first-order sentence φ, the class of structures defined by φ is simply $MOD(\varphi) := \{\mathcal{A} \mid \mathcal{A} \models \varphi\}$. Now, we can formally define the distinguishability problem with respect to a class of structures. Let $\mathcal{C}$ be a class of structures. A sample $S = (P, N)$ consists of two finite sets $P, N \subseteq \mathcal{C}$ such that for each $\mathcal{A} \in P$, $\mathcal{B} \in N$, $\mathcal{A}$ and $\mathcal{B}$ are not isomorphic. Intuitively, P contains positively classified structures and N contains negatively classified structures. A sentence φ is consistent with a sample $S = (P, N)$ if $P \subseteq MOD(\varphi)$ and $N \cap MOD(\varphi) = \emptyset$. In what follows, we formally define the distinguishability problem.

Definition 2.2 [Distinguishability Problem] Let $\mathcal{C}$ be a fixed class of structures.

Given a sample S, the problem consists of finding a first-order sentence φ of minimal quantifier rank that is consistent with S.

It is well known that every finite relational structure can be characterized in first-order logic up to isomorphism, i.e., for every finite structure $\mathcal{A}$, there is a first-order sentence $\varphi_{\mathcal{A}}$ such that for all structures $\mathcal{B}$ we have $\mathcal{B} \models \varphi_{\mathcal{A}}$ if and only if $\mathcal{A}$ and $\mathcal{B}$ are isomorphic (Proposition 2.1.1 from [4]). Since we are considering a fixed vocabulary and samples are finite sets of finite relational structures, one can easily build in polynomial-time a first-order sentence consistent with a given sample. For example, let $\tau = \{P, R\}$ be a fixed vocabulary such that P has arity one, and R has arity two. Clearly, the size of $\varphi_{\mathcal{A}}$ is $O(n^2)$. Also, it takes polynomial time to check whether $\mathcal{A} \models \psi$ when ψ is atomic. Therefore, we build $\varphi_{\mathcal{A}}$ in polynomial time. Unfortunately, the quantifier rank of $\varphi_{\mathcal{A}}$ is the number of elements in the domain of $\mathcal{A}$ plus one. Therefore, we can not use these formulas in a solution to the distinguishability problem. Now, we focus on Ehrenfeucht–Fraïssé games and its importance in order to solve the problem we are considering in this work.

Definition 2.3 [EF Game] Let r be an integer such that $r \geq 0$, τ a vocabulary, $\mathcal{A}$ and $\mathcal{B}$ two τ-structures. The Ehrenfeucht–Fraïssé game (EF game, for short) $\mathcal{G}_r(\mathcal{A}, \mathcal{B})$ is played by two players called the Spoiler and the Duplicator. Each play of the game has r rounds and, in each round, the Spoiler plays first and picks an element from the domain A of $\mathcal{A}$, or from the domain B of $\mathcal{B}$. Then, the Duplicator responds by picking an element from the domain of the other structure. Let $a_i \in A$ and $b_i \in B$ be the two elements picked by the Spoiler and the Duplicator in the ith round. The Duplicator wins the play if the mapping $(a_1, b_1), ..., (a_r, b_r)$ is an isomorphism between the substructures induced by $a_1, ..., a_r$ and $b_1, ..., b_r$, respectively. Otherwise, Spoiler wins this play. We say that a player has a winning strategy in $\mathcal{G}_r(\mathcal{A}, \mathcal{B})$ if it is possible for him to win each play whatever choices are made by the opponent.

In this work, we always assume that $\mathcal{A}$ and $\mathcal{B}$ are not isomorphic. Observe that, for a structure $\mathcal{A}$ such that $|A| \leq r$, a structure $\mathcal{B}$ is isomorphic to $\mathcal{A}$ if and only if the Duplicator has a winning strategy in $\mathcal{G}_{r+1}(\mathcal{A}, \mathcal{B})$. Then, we can consider that the number of rounds r is bounded by the sizes of $\mathcal{A}$ and $\mathcal{B}$. Now, for a structure $\mathcal{A}$ and a natural number r, we define formulas describing the properties of $\mathcal{A}$ in any EF game with r rounds.

Definition 2.4 [r-Hintikka Formula] Let $\mathcal{A}$ be a structure, $\overline{a} = a_1...a_s \in A^s$, and $\overline{v} = v_1, ..., v_s$ a tuple of variables.

$$\varphi^0_{\mathcal{A},\bar{a}}(\bar{v}) := \bigwedge\{\varphi(\bar{v}) \mid \varphi \text{ is atomic or negated atomic and } \mathcal{A} \models \varphi[\bar{a}]\},$$

$$\text{and for } r > 0,\ \varphi^r_{\mathcal{A},\bar{a}}(\bar{v}) := \bigwedge_{a \in A} \exists v_{s+1} \varphi^{r-1}_{\mathcal{A},\bar{a}a}(\bar{v}, v_{s+1}) \wedge \forall v_{s+1} (\bigvee_{a \in A} \varphi^{r-1}_{\mathcal{A},\bar{a}a}(\bar{v}, v_{s+1})).$$

The Hintikka formula $\varphi^r_{\mathcal{A},\overline{a}}$ describes the isomorphism type of the substructure generated by $\overline{a}$ in $\mathcal{A}$. We write $\varphi^r_{\mathcal{A}}$ whenever $s = 0$. Given a structure $\mathcal{A}$ and a natural

number r, the size of $\varphi^r_{\mathcal{A}}$ is $O(2^r|A|^r)$. Therefore, since r is bounded by $|A|$, the size of $\varphi^r_{\mathcal{A}}$ is exponential in the size of $\mathcal{A}$. The following theorems are important to prove some of our results. These theorems are presented in [4].

Theorem 2.5 (Ehrenfeucht's Theorem) *Let $\mathcal{A}$ and $\mathcal{B}$ be structures, and r be a natural number. The Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$ iff $\mathcal{B} \models \varphi^r_{\mathcal{A}}$. The Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$ iff $\mathcal{A}$ and $\mathcal{B}$ satisfy the same sentences of quantifier rank at most r.*

Theorem 2.6 *Let φ be a sentence of quantifier rank at most r. Then, there exist structures $\mathcal{A}_1$, ..., $\mathcal{A}_k$ such that*

$$\models \varphi \leftrightarrow (\varphi^r_{\mathcal{A}_1} \vee ... \vee \varphi^r_{\mathcal{A}_k}).$$

EF games provide information about the similarity between structures. If two structures $\mathcal{A}$ and $\mathcal{B}$ are not isomorphic, then there is an r such that the Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$. The notion of EF-similarity below represents this information about similarity.

Definition 2.7 [EF-similarity] The EF-similarity between two structures $\mathcal{A}$ and $\mathcal{B}$, denoted by $EFsim(\mathcal{A},\mathcal{B})$, is the minimum number of rounds r such that Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$. If Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$ for all r, then $EFsim(\mathcal{A},\mathcal{B}) = \infty$.

EF games are important in our framework because if the Spoiler has a winning strategy in a game on $\mathcal{A}$ and $\mathcal{B}$ with r rounds, then there exists a first-order sentence φ of quantifier rank at most r that holds in $\mathcal{A}$ and does not hold in $\mathcal{B}$. Also, in this case, the sentence $\varphi^r_{\mathcal{A}}$ is an example of such a sentence. However, over arbitrary vocabularies, the problem of determining whether the Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{A},\mathcal{B})$ is $PSPACE$-complete [22]. Fortunately, it is possible to do better for EF games on MS, LO, and ES. In what follows, we show the case for EF games on MS and ES. For details of the results on these classes see [15].

A monadic structure is a structure $\mathcal{M} = \langle M, P_1^{\mathcal{M}}, ..., P_k^{\mathcal{M}}\rangle$ such that each P_i is monadic and $P_1^{\mathcal{M}}, ..., P_k^{\mathcal{M}}$ are pairwise disjoint. In what follows, we give a result on EF games on MS and we show how to compute the EF-similarity between two structures in MS.

Theorem 2.8 (EF Games on MS) *[15] For a structure $\mathcal{M}$ in* MS, *we set $P_{k+1}^{\mathcal{M}} = \overline{(P_1^{\mathcal{M}} \cup ... \cup P_k^{\mathcal{M}})}$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be structures in MS. The Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{M}_1,\mathcal{M}_2)$ iff there exists $i \in \{1, ..., k+1\}$ such that $(|P_i^{\mathcal{M}_1}| < r$ or $|P_i^{\mathcal{M}_2}| < r)$ and $|P_i^{\mathcal{M}_1}| \neq |P_i^{\mathcal{M}_2}|$.*

The EF-similarity between two structures in MS can be computed in polynomial time in the size of the structures in the following way.

$$EFsim(\mathcal{M}_1,\mathcal{M}_2) = min\{min(|P_i^{\mathcal{M}_1}|, |P_i^{\mathcal{M}_2}|) \mid |P_i^{\mathcal{M}_1}| \neq |P_i^{\mathcal{M}_2}|, 1 \leq i \leq k+1\} + 1.$$

An equivalence structure is a structure of the form $\mathcal{E} = \langle A, E^{\mathcal{E}}\rangle$ such that $E^{\mathcal{E}}$ is

an equivalence relation on A. Let $q_t^{\mathcal{E}}$ be the number of equivalence classes in $\mathcal{E}$ of size t. Let $q_{\geq t}^{\mathcal{E}}$ be the number of equivalence classes in $\mathcal{E}$ of size at least t.

Theorem 2.9 (EF Games on ES) *[15] Let r be a natural number, and $\mathcal{E}_1, \mathcal{E}_2$ be equivalence structures. The Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{E}_1, \mathcal{E}_2)$ iff (there exists a $t < r$ such that $q_t^{\mathcal{E}_1} \neq q_t^{\mathcal{E}_2}$ and $r \geq min\{q_t^{\mathcal{E}_1}, q_t^{\mathcal{E}_2}\} + t + 1$) or (there exists a $t \leq r$ such that $q_{\geq t}^{\mathcal{E}_1} \neq q_{\geq t}^{\mathcal{E}_2}$ and $r \geq min\{q_{\geq t}^{\mathcal{E}_1}, q_{\geq t}^{\mathcal{E}_2}\} + t$).*

By Theorem 2.9, the EF-similarity between equivalence structures $\mathcal{E}_1$ and $\mathcal{E}_2$ can be computed in the following way. Note that computing EF-similarity takes polynomial time in the size of the equivalence structures.

$$EFsim(\mathcal{E}_1, \mathcal{E}_2) = min(min\{min(q_t^{\mathcal{E}_1}, q_t^{\mathcal{E}_2}) + t \mid q_t^{\mathcal{E}_1} \neq q_t^{\mathcal{E}_2}\} + 1, min\{min(q_{\geq t}^{\mathcal{E}_1}, q_{\geq t}^{\mathcal{E}_2}) + t \mid q_{\geq t}^{\mathcal{E}_1} \neq q_{\geq t}^{\mathcal{E}_2}\}).$$

Besides the importance of EF games on a specific class of structures to our framework, we also use Theorem 2.8, Theorem 3.1, and Theorem 2.9 to define the distinguishability sentences. These sentences are defined based on the conditions characterizing the winning strategies for the Spoiler on MS, ES, and DULO.

Our algorithm's first step is to compute the quantifier rank necessary to distinguish between any two structures $\mathcal{A} \in P$ and $\mathcal{B} \in N$. Then, the fact that $EFsim(\mathcal{A}, \mathcal{B})$ can be computed in polynomial time is essential to show that our algorithm runs in polynomial time as well.

It is easy to build a first-order sentence of minimal quantifier rank that consists of a disjunction of Hintikka formulas, and that is consistent with a given sample. For example, let $P = \{\mathcal{M}_1\}$, $N = \{\mathcal{M}_2, \mathcal{M}_3\}$, $r = max\{EFsim(\mathcal{M}_1, \mathcal{M}_2), EFsim(\mathcal{M}_1, \mathcal{M}_3)\}$, and $S = (P, N)$. The sentence $\varphi^r_{\mathcal{M}_1}$ is a first-order sentence of minimal quantifier rank that is consistent with S. Unfortunately, the size of $\varphi^r_{\mathcal{M}_1}$ is exponential in the size of S. Therefore, $\varphi^r_{\mathcal{M}_1}$ can not be built in polynomial time in the size of the sample. This motivates the introduction of the distinguishability sentences in Section 4.

3 EF Games on Disjoint Unions of Linear Orders

In this section, we show conditions characterizing the winning strategies for both players on disjoint unions of linear orders. We also determine the EF-similarity in this context. First, we turn to linear orders. A linear order is a structure $\mathcal{L} = \langle L, <^{\mathcal{L}} \rangle$ such that $<^{\mathcal{L}}$ is a linear order on L. In what follows, for a linear order $\mathcal{L}$, let $q^{\mathcal{L}}$ be the number of elements in the domain of $\mathcal{L}$. For an element a, we define $\mathcal{L}^{>a}$ ($\mathcal{L}^{<a}$) as a substructure of $\mathcal{L}$ such that $\{b \in L \mid b > a\}$ ($\{b \in L \mid b < a\}$) is the domain of $\mathcal{L}^{>a}$ ($\mathcal{L}^{<a}$). The following result is well known in the literature [17]. In this result, the Spoiler's winning strategy consists of choosing, in her first round, an element from the linear order with more elements.

Theorem 3.1 (EF Games on LO) *Let r be a natural number, $\mathcal{L}_1$ and $\mathcal{L}_2$ be linear orders. The Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{L}_1, \mathcal{L}_2)$ if and only if $q^{\mathcal{L}_1} \neq q^{\mathcal{L}_2}$ and ($q^{\mathcal{L}_1} < 2^r - 1$ or $q^{\mathcal{L}_2} < 2^r - 1$).*

Now, we define disjoint unions of linear orders. Assume that $\mathcal{L}_1$ and $\mathcal{L}_2$ are linear orders such that $L_1 \cap L_2 = \emptyset$. Then, $\mathcal{L}_1 \uplus \mathcal{L}_2$, the disjoint union of $\mathcal{L}_1$ and $\mathcal{L}_2$, is the structure with domain $L_1 \cup L_2$ and $<^{\mathcal{L}_1 \uplus \mathcal{L}_2} = <^{\mathcal{L}_1} \cup <^{\mathcal{L}_2}$. We represent a disjoint unions of linear orders $\mathcal{W}$ by disjoint unions $(...(\mathcal{L}_1 \uplus \mathcal{L}_2) \uplus ... \uplus \mathcal{L}_l)$.

Equivalence structures also can be seen as disjoint unions of structures. Our results on DULO are inspired by the case of EF games on equivalence structures. In what follows, for a disjoint unions of linear orders $\mathcal{W}$, let $q_t^{\mathcal{W}}$ be the number of linear orders $\mathcal{L}$ in $\mathcal{W}$ such that $q^{\mathcal{L}} = t$. Also, let $q_{\geq t}^{\mathcal{W}}$ be the number of linear orders $\mathcal{L}$ in $\mathcal{W}$ such that $q^{\mathcal{L}} \geq t$. Given an element a in $\mathcal{W}_1$, $\mathcal{L}(a)$ denotes the linear order in $\mathcal{W}_1$ such that a is in the domain of $\mathcal{L}(a)$. Given a disjoint unions of linear orders $\mathcal{W}$, let $\mathcal{W}(a_1, ..., a_k)$ be the disjoint unions of linear orders obtained by removing $\mathcal{L}(a_1)$, ..., $\mathcal{L}(a_k)$ from $\mathcal{W}$.

Definition 3.2 [Disparity] Let r be a natural number, $\mathcal{W}_1$ and $\mathcal{W}_2$ be disjoint unions of linear orders. We say that

- $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ has a small disparity if there exists t such that $1 \leq t < 2^r - 2$, $q_t^{\mathcal{W}_1} \neq q_t^{\mathcal{W}_2}$, and $r \geq min\{q_t^{\mathcal{W}_1}, q_t^{\mathcal{W}_2}\} + \lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 2$.
- $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ has a large disparity if there exists t such that $1 \leq t \leq 2^r - 1$, $q_{\geq t}^{\mathcal{W}_1} \neq q_{\geq t}^{\mathcal{W}_2}$, and $r \geq min\{q_{\geq t}^{\mathcal{W}_1}, q_{\geq t}^{\mathcal{W}_2}\} + \lfloor log(t) \rfloor + 1$.

Lemma 3.3 *Let r be a natural number, $\mathcal{W}_1$ and $\mathcal{W}_2$ be two disjoint unions of linear orders. If $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ has a small or large disparity, then the Spoiler has a winning strategy.*

Proof. Suppose that $q_t^{\mathcal{W}_1} > q_t^{\mathcal{W}_2}$, and $r \geq q_t^{\mathcal{W}_2} + \lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 2$. The Spoiler has the following winning strategy: first, she chooses elements $a_1, a_2, ..., a_{q_t^{\mathcal{W}_2}}$ from distinct linear orders of size t in $\mathcal{W}_1$. The Duplicator must choose elements $b_1, ..., b_{q_t^{\mathcal{W}_2}}$ in $\mathcal{W}_2$ from distinct linear orders of size t. Next, the Spoiler chooses an element a from a distinct linear order $\mathcal{L}(a)$ of size t in $\mathcal{W}_1$ such that $q^{\mathcal{L}(a)^{>a}} < 2^{r-1} - 1$ and $q^{\mathcal{L}(a)^{<a}} < 2^{r-1} - 1$. Then, the Duplicator must select an element b from a linear order $\mathcal{L}(b)$ such that $q^{\mathcal{L}(b)} \neq t$. Then, the Spoiler has a winning strategy in $\mathcal{G}_{\lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 1}(\mathcal{L}(a)^{<a}, \mathcal{L}(b)^{<b})$ by Theorem 3.1. Furthermore, the Spoiler has a winning strategy in $\mathcal{G}_{\lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 2}(\mathcal{L}(a), \mathcal{L}(b))$. Therefore, the Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$.

Now, suppose that $q_{\geq t}^{\mathcal{W}_1} > q_{\geq t}^{\mathcal{W}_2}$, and $r \geq q_{\geq t}^{\mathcal{W}_2} + \lfloor log(t) \rfloor + 1$. The Spoiler has the following winning strategy: first, she chooses elements $a_1, a_2, ..., a_{q_{\geq t}^{\mathcal{W}_2}}$ from distinct linear orders of size at least t in $\mathcal{W}_1$. The Duplicator must choose elements $b_1, ..., b_{q_{\geq t}^{\mathcal{W}_2}}$ in $\mathcal{W}_2$ from distinct linear orders of size at least t. Next, by Theorem 3.1, the Spoiler uses her winning strategy in $\mathcal{G}_{\lfloor log(t) \rfloor + 1}(\mathcal{L}_1, \mathcal{L}_2)$ such that $\mathcal{L}$ is a linear order in $\mathcal{W}_1$, $q^{\mathcal{L}_1} \geq t$, $q^{\mathcal{L}_2} < t$. Then, the Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$. □

In the above result, different from the proof of Theorem 3.1, the Spoiler has a winning strategy which consists of choosing an element from a linear order with

fewer elements. Then, first, we need the following lemma in order to guarantee a winning strategy for the Duplicator.

Lemma 3.4 *Let r be a natural number, $\mathcal{L}_1$ and $\mathcal{L}_2$ be linear orders.*

- *If $q^{\mathcal{L}_1} \geq 2^r - 1$ and the Spoiler chooses an element from $\mathcal{L}_1$ in her first round, then the Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{L}_1, \mathcal{L}_2)$ if $q^{\mathcal{L}_2} \geq 2^r - 1$.*
- *If $q^{\mathcal{L}_1} = 2^r - 2$ and the Spoiler chooses an element from $\mathcal{L}_1$ in her first round, then the Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{L}_1, \mathcal{L}_2)$ if $q^{\mathcal{L}_2} \geq 2^r - 2$.*
- *If $q^{\mathcal{L}_1} < 2^r - 2$ and the Spoiler chooses an element from $\mathcal{L}_1$ in her first round, then the Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{L}_1, \mathcal{L}_2)$ if $q^{\mathcal{L}_2} = q^{\mathcal{L}_1}$.*

Proof. For the first and third part of the lemma, the Duplicator has a winning strategy as in Theorem 3.1. To prove the second part of the lemma, assume that $q^{\mathcal{L}_2} \geq 2^r - 2$. Let $a \in L_1$ be the element chosen by the Spoiler. Then, $q^{\mathcal{L}_1^{>a}} \geq 2^{r-1} - 1$ or $\mathcal{L}_1^{<a} \geq 2^{r-1} - 1$. Assume that $q^{\mathcal{L}_1^{>a}} \geq 2^{r-1} - 1$. The Duplicator chooses an element $b \in L_2$ such that $q^{\mathcal{L}_2^{>b}} \geq 2^{r-1} - 1$ and $q^{\mathcal{L}_2^{<b}} = q^{\mathcal{L}_2^{<a}}$. Then, the Duplicator has a winning strategy in $\mathcal{G}_{r-1}(\mathcal{L}_1^{>a}, \mathcal{L}_2^{>b})$. Therefore, the Duplicator has a winning strategy in r rounds. □

Theorem 3.5 (EF Games on DULO) *Let r be a natural number, $\mathcal{W}_1$ and $\mathcal{W}_2$ be disjoint unions of linear orders. The Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ iff $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ has a small or large disparity.*

Proof. One direction is Lemma 3.3. Conversely, assume that $\mathcal{G}_r(\mathcal{W}_1, \mathcal{W}_2)$ has neither small nor large disparity. Let $(a_1, b_1), ..., (a_k, b_k)$ be such that $a_i \in W_1$, $b_i \in W_2$ for $i \in \{1, ..., k\}$, and if the Spoiler has chosen a_i (b_i), then the Duplicator has selected b_i (a_i). We show that the Duplicator has a winning strategy by induction on k. Let $\mathcal{W}_1' = \mathcal{W}_1(a_1, ..., a_k)$ and $\mathcal{W}_2' = \mathcal{W}_2(b_1, ..., b_k)$. The following are our inductive hypotheses:

(i) For all $i, j \leq k$, $a_i < a_j$ iff $b_i < b_j$.
(ii) For all $i \leq k$, $q^{\mathcal{L}(a_i)} \geq 2^{r-i} - 1$, then $q^{\mathcal{L}(b_i)} \geq 2^{r-i} - 1$.
(iii) For all $i \leq k$, $q^{\mathcal{L}(a_i)} = 2^{r-i} - 2$, then $q^{\mathcal{L}(b_i)} \geq 2^{r-i} - 2$.
(iv) For all $i \leq k$, $q^{\mathcal{L}(a_i)} < 2^{r-i} - 2$, then $q^{\mathcal{L}(b_i)} = q^{\mathcal{L}(a_i)}$.
(v) $\mathcal{G}_{r-k}(\mathcal{W}_1', \mathcal{W}_2')$ has neither small nor large disparity.

If the Spoiler chooses an element $a_{k+1} \in W_1$ such that $a_i < a_{k+1}$, then the Duplicator responds by b_{k+1} such that $b_i < b_{k+1}$. The case where $a_{k+1} < a_i$ is analogous. Note that, by inductive hypotheses (ii), (iii), (iv), the Duplicator can select such an element. Now, assume $a_{k+1} \in W_1$ is in a linear order $\mathcal{L}(a_{k+1})$ different from $\mathcal{L}(a_i)$, for $i \leq k$. Suppose $q^{\mathcal{L}(a_{k+1})} \geq 2^{r-k} - 1$. By contradiction, assume that any linear order $\mathcal{L}$ in $\mathcal{W}_2'$ is such that $|L| < 2^{r-k} - 1$. Then, $\mathcal{G}_{r-k}(\mathcal{W}_1', \mathcal{W}_2')$ would have a large disparity as witnessed by $t = 2^{r-k} - 1$. Therefore, the Duplicator can choose b_{k+1} in $\mathcal{W}_2$ such that $\mathcal{L}(b_{k+1})$ is different from $\mathcal{L}(b_i)$, for $i \leq k$, and $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k} - 1$. Now, assume that $q^{\mathcal{L}(a_{k+1})} = 2^{r-k} - 2$. We assume, by absurd, that any linear order $\mathcal{L}$ in $\mathcal{W}_2'$ is such that $|L| < 2^{r-k} - 2$. Then, $\mathcal{G}_{r-k}(\mathcal{W}_1', \mathcal{W}_2')$ would have a

large disparity as witnessed by $t = 2^{r-k} - 2$. Therefore, there exists b_{k+1} in $\mathcal{W}_2$ such that $\mathcal{L}(b_{k+1})$ is different from $\mathcal{L}(b_i)$, for $i \leq k$, and $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k} - 2$. Duplicator chooses b_{k+1}. Finally, suppose that $q^{\mathcal{L}(a_{k+1})} < 2^{r-k} - 2$. Then, b_{k+1} must exist in $\mathcal{W}_2$ such that $\mathcal{L}(b_{k+1})$ is different from $\mathcal{L}(b_i)$, for $i \leq k$, and $q^{\mathcal{L}(b_{k+1})} = q^{\mathcal{L}(a_{k+1})}$. Otherwise $\mathcal{G}_{r-k}(\mathcal{W}_1', \mathcal{W}_2')$ would have a small disparity as witnessed by $t = q^{\mathcal{L}(a_{k+1})}$. Then, the Duplicator chooses b_{k+1}. The cases where the Spoiler chooses an element from $\mathcal{W}_2$ are analogous.

Now, we check that each of the inductive hypotheses hold on $(a_1, b_1), ..., (a_{k+1}, b_{k+1})$. Clearly, inductive hypothesis (i) holds. Now, assume that $q^{\mathcal{L}(a_{k+1})} \geq 2^{r-k-1} - 1$. If $q^{\mathcal{L}(a_{k+1})} \geq 2^{r-k} - 1$, then the strategy we describe ensures that $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k} - 1$. If $q^{\mathcal{L}(a_{k+1})} = 2^{r-k} - 2$, then the strategy ensures that $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k} - 2$. Finally, if $q^{\mathcal{L}(a_{k+1})} < 2^{r-k} - 2$, then $q^{\mathcal{L}(b_{k+1})} = q^{\mathcal{L}(a_{k+1})}$. Clearly, in all cases $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k-1} - 1$. Now, assume that $q^{\mathcal{L}(a_{k+1})} = 2^{r-k-1} - 2$. Obviously, $q^{\mathcal{L}(a_{k+1})} < 2^{r-k} - 2$. Then, $q^{\mathcal{L}(b_{k+1})} = q^{\mathcal{L}(a_{k+1})}$, and $q^{\mathcal{L}(b_{k+1})} \geq 2^{r-k-1} - 2$. At last, assume $q^{\mathcal{L}(a_{k+1})} < 2^{r-k-1} - 2$. Then, $q^{\mathcal{L}(a_{k+1})} < 2^{r-k} - 2$ and $q^{\mathcal{L}(b_{k+1})} = q^{\mathcal{L}(a_{k+1})}$. Therefore, $q^{\mathcal{L}(b_{k+1})} < 2^{r-k-1} - 2$.

Now, it remains to prove that (v) holds. Consider $\mathcal{W}_1'' = \mathcal{W}_1(a_1, ..., a_{k+1})$ and $\mathcal{W}_2'' = \mathcal{W}_2(b_1, ..., b_{k+1})$. By contradiction, assume that $\mathcal{G}_{r-k-1}(\mathcal{W}_1'', \mathcal{W}_2'')$ has a small disparity. Assume $t < 2^{r-k-1} - 2$ such that $q_t^{\mathcal{W}_1''} > q_t^{\mathcal{W}_2''}$ and $r - k - 1 \geq min\{q_t^{\mathcal{W}_1''}, q_t^{\mathcal{W}_2''}\} + \lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 2$. Then, $t < 2^{r-k} - 2$. If $q_t^{\mathcal{W}_1'} = q_t^{\mathcal{W}_2'}$, then $q^{\mathcal{L}}(b_k) = t$ and, as $t < 2^{r-k} - 2$, $q^{\mathcal{L}}(a_k) = t$. Therefore, $q_t^{\mathcal{W}_1'} \neq q_t^{\mathcal{W}_2'}$. Thus, there exists $t < 2^{r-k}$ such that $q_t^{\mathcal{W}_1'} \neq q_t^{\mathcal{W}_2'}$, and $r - k \geq min\{q_t^{\mathcal{W}_1'}, q_t^{\mathcal{W}_2'}\} \lfloor log(\lceil \frac{t+1}{2} \rceil) \rfloor + 2$ because $min\{q_t^{\mathcal{W}_1'}, q_t^{\mathcal{W}_2'}\} \leq min\{q_t^{\mathcal{W}_1''}, q_t^{\mathcal{W}_2''}\} + 1$. Then, $\mathcal{G}_{r-k-1}(\mathcal{W}_1'', \mathcal{W}_2'')$ do not have a small disparity. By an analogous way, $\mathcal{G}_{r-k-1}(\mathcal{W}_1'', \mathcal{W}_2'')$ do not have a large disparity. Finally, the Duplicator has a winning strategy because (i) holds at each round. □

Therefore, EF-similarity between disjoint unions of linear orders $\mathcal{W}_1$ and $\mathcal{W}_2$ can be computed as follows in polynomial time in the size of $\mathcal{W}_1$ and $\mathcal{W}_2$.

$$EFsim(\mathcal{W}_1, \mathcal{W}_2) = min(min\{min(q_t^{\mathcal{W}_1}, q_t^{\mathcal{W}_2}) + \lfloor log(\lceil \tfrac{t+1}{2} \rceil) \rfloor + 2 \mid q_t^{\mathcal{E}_1} \neq q_t^{\mathcal{E}_2}\},$$
$$min\{min(q_{\geq t}^{\mathcal{E}_1}, q_{\geq t}^{\mathcal{E}_2}) + \lfloor log(t) \rfloor + 1 \mid q_{\geq t}^{\mathcal{E}_1} \neq q_{\geq t}^{\mathcal{E}_2}\}).$$

4 Distinguishability Sentences

In this section, we define the distinguishability sentences for structures $\mathcal{A}$, $\mathcal{B}$ in a class of structures $\mathcal{C}$, and a natural number r. The distinguishability sentences hold in $\mathcal{A}$, do not hold in $\mathcal{B}$, and they have quantifier rank at most r. We define the set of distinguishability sentences $\Phi^r_{\mathcal{A},\mathcal{B}}$ in a way such that the Spoiler has a winning strategy in $\mathcal{G}_r(\mathcal{A}, \mathcal{B})$ if and only if there exists $\varphi \in \Phi^r_{\mathcal{A},\mathcal{B}}$. This result follows from Theorem 2.5. The first step is to show that the conditions characterizing winning strategies for the Spoiler can be expressed by first-order sentences of size polynomial in the size of the structures. This result is important in order to guarantee that our

algorithm runs in polynomial time in the size of the sample. These formulas are also important to help the explanation, and they improve readability of sentences returned by our algorithm.

4.1 Monadic Structures

First, we define $|P_i| \geq n$, for $i \in \{1, ..., k+1\}$, as a sentence describing that the number of elements in P_i is at least n:

$$|P_i| \geq n := \exists x_1 ... \exists x_n (\bigwedge_{l \neq j}^{n} x_l \neq x_j \wedge \bigwedge_{l=1}^{n} P_i(x_l)).$$

Clearly, $qr(|P_i| \geq n) = n$, the size of $|P_i| \geq n$ is $O(n^2)$, and $\mathcal{A} \models |P_i| \geq n$ iff $|P_i^{\mathcal{A}}| \geq n$. We also define abbreviations $|P_i| \leq n := \neg |P_i| \geq n+1$ and $|P_i| = n := |P_i| \geq n \wedge |P_i| \leq n$. Now, we can define the distinguishability sentences for monadic structures.

Definition 4.1 [Distinguishability Sentences for MS] Let $\mathcal{M}_1$, $\mathcal{M}_2$ be monadic structures. Let r be a natural number.

$$\Phi^r_{\mathcal{M}_1,\mathcal{M}_2} := \begin{array}{l} \{|P_i| < m \mid 1 \leq i \leq k+1, |P_i^{\mathcal{M}_1}| < |P_i^{\mathcal{M}_2}|, |P_i^{\mathcal{M}_1}| + 1 \leq m \leq min(r, |P_i^{\mathcal{M}_2}|)\} \cup \\ \{|P_i| \geq m \mid 1 \leq i \leq k+1, |P_i^{\mathcal{M}_1}| > |P_i^{\mathcal{M}_2}|, |P_i^{\mathcal{M}_2}| + 1 \leq m \leq min(r, |P_i^{\mathcal{M}_1}|)\}. \end{array}$$

Given $\mathcal{M}_1$, $\mathcal{M}_2$, and r, the size of a sentence $\varphi \in \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$ is $O((|M_1| + |M_2|)^2)$, and $|\Phi^r_{\mathcal{M}_1,\mathcal{M}_2}|$ is $O(k(|M_1| + |M_2|))$. Now, we give an example of the distinguishability sentences for MS. Then, we show results ensuring adequate properties of the distinguishability sentences.

Example 4.2 Let $\mathcal{M}_1 = \langle M_1, P_1^{\mathcal{M}_1} \rangle$ and $\mathcal{M}_2 = \langle M_2, P_1^{\mathcal{M}_2} \rangle$ such that $|P_1^{\mathcal{M}_1}| = 2$, $|P_2^{\mathcal{M}_1}| = 3$, $|P_1^{\mathcal{M}_2}| = 2$, $|P_2^{\mathcal{M}_1}| = 2$, and $r = 4$. Then, $|P_2| \geq 3 \in \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$ because $|P_2^{\mathcal{M}_1}| > |P_2^{\mathcal{M}_2}|$ and, for $m = 3$, $|P_2^{\mathcal{M}_2}| + 1 \leq m \leq min(r, |P_2^{\mathcal{M}_1}|)$. Also, observe that $|P_2| \geq 4 \notin \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$ since, for $m = 4$, $m > min(r, |P_2^{\mathcal{M}_1}|)$. Finally, $|P_1| \geq 2 \notin \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$ because $|P_1^{\mathcal{M}_1}| = |P_1^{\mathcal{M}_2}|$.

Lemma 4.3 *Let $\mathcal{M}_1$, $\mathcal{M}_2$ be structures in* MS, and r be a natural number. Let $\varphi \in \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$. Then, $\mathcal{M}_1 \models \varphi$ and $\mathcal{M}_2 \not\models \varphi$.

Proof. Suppose $\varphi = |P_i| < m$. Then, $|P_i^{\mathcal{M}_1}| < |P_i^{\mathcal{M}_2}|$ and $|P_i^{\mathcal{M}_1}| + 1 \leq m \leq min(r, |P_i^{\mathcal{M}_2}|)$. Therefore, $\mathcal{M}_1 \models \varphi$ because $|P_i^{\mathcal{M}_1}| < m$. Clearly, $m \leq |P_i^{\mathcal{M}_2}|$. Then, $\mathcal{M}_1 \not\models \varphi$. The case in which $\varphi = |P_i| \geq m$ is similar. □

Lemma 4.4 *Let $\mathcal{M}_1$, $\mathcal{M}_2$ be structures in* MS, and r be a natural number. Let $\varphi \in \Phi^r_{\mathcal{M}_1,\mathcal{M}_2}$. Then, $qr(\varphi) \leq r$.

Proof. Let $\varphi = |P_i| \trianglelefteq m$ where $\trianglelefteq \in \{<, \geq\}$. Hence, $qr(\varphi) = m$. As $m \leq r$, then $qr(\varphi) \leq r$. □

Now, we show that, over MS, any first-order sentence is equivalent to a boolean combination of distinguishability sentences. First, we define formulas equivalent to Hintikka formulas over MS.

Lemma 4.5 $\models \varphi^r_{\mathcal{M}} \leftrightarrow \bigwedge_{i=1}^{k+1} \varphi^{r,i}_{\mathcal{W}}$ *such that*

$$\varphi^{r,i}_{\mathcal{M}} := \begin{cases} |P_i| = |P_i^{\mathcal{M}}|, & \text{if } |P_i^{\mathcal{M}}| < r \\ |P_i| \geq r, & \text{otherwise.} \end{cases}$$

Proof. Let $\mathcal{M}' \models \varphi^r_{\mathcal{M}}$. Then, the Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{M}, \mathcal{M}')$. Then, for all i, $|P_i^{\mathcal{M}}| = |P_i^{\mathcal{M}'}|$ or $(|P_i^{\mathcal{M}}| \geq r$ and $|P_i^{\mathcal{M}'}| \geq r)$. Let i such that $|P_i^{\mathcal{M}}| < r$. Then, $|P_i^{\mathcal{M}}| = |P_i^{\mathcal{M}'}|$. Therefore, $\mathcal{M}' \models \varphi^{r,i}_{\mathcal{M}}$. Let i such that $|P_i^{\mathcal{M}}| \geq r$. Then, $|P_i^{\mathcal{M}'}| \geq r$. Therefore, $\mathcal{M}' \models \varphi^{r,i}_{\mathcal{M}}$. Finally, $\mathcal{M}' \models \bigwedge_{i=1}^{k+1} \varphi^{r,i}_{\mathcal{M}}$. Conversely, let $\mathcal{M}' \models \bigwedge_{i=1}^{k+1} \varphi^{r,i}_{\mathcal{M}}$. Then, $\mathcal{M}' \models \varphi^{r,i}_{\mathcal{M}}$, for all i. If $\varphi^{r,i}_{\mathcal{M}} = (|P_i| = |P_i^{\mathcal{M}}|)$, then $|P_i^{\mathcal{M}}| = |P_i^{\mathcal{M}'}|$. If $\varphi^{r,i}_{\mathcal{M}} = |P_i| \geq r$, then $(|P_i^{\mathcal{M}}| \geq r$ and $|P_i^{\mathcal{M}'}| \geq r)$. Then, for all i, $|P_i^{\mathcal{M}}| = |P_i^{\mathcal{M}'}|$ or $(|P_i^{\mathcal{M}}| \geq r$ and $|P_i^{\mathcal{M}'}| \geq r)$. Therefore, $\mathcal{M}' \models \varphi^r_{\mathcal{M}}$. □

Now, we need the following lemmas.

Lemma 4.6 *Let r be a natural number and $\mathcal{M}$ be a structure in* MS. There exists a set of monadic structures V_i such that $\varphi^{r,i}_{\mathcal{M}}$ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{M}' \in V_i} \Phi^r_{\mathcal{M},\mathcal{M}'}$.

Proof. If $\varphi^{r,i}_{\mathcal{M}} = |P_i| \geq r$, then $|P_i^{\mathcal{M}}| \geq r$. Let $V_i = \{\mathcal{M}'\}$ such that $|P_i^{\mathcal{M}'}| = r - 1$. Then, $|P_i| \geq r \in \Phi^r_{\mathcal{M},\mathcal{M}'}$ because $|P_i^{\mathcal{M}}| > |P_i^{\mathcal{M}'}|$ and $m = r$. If $\varphi^{r,i}_{\mathcal{M}} = (|P_i| = |P_i^{\mathcal{M}}|)$, then $|P_i^{\mathcal{M}}| < r$. Let $V_i = \{\mathcal{M}_1, \mathcal{M}_2\}$ such that $|P_i^{\mathcal{M}_1}| = |P_i^{\mathcal{M}}| - 1$ and $|P_i^{\mathcal{M}_2}| = |P_i^{\mathcal{M}}| + 1$. Then, $|P_i| \geq |P_i^{\mathcal{M}}| \in \Phi^r_{\mathcal{M},\mathcal{M}_1}$ because $|P_i^{\mathcal{M}}| > |P_i^{\mathcal{M}_1}|$ and $m = |P_i^{\mathcal{M}}|$. Also, $|P_i| < |P_i^{\mathcal{M}}| + 1 \in \Phi^r_{\mathcal{M},\mathcal{M}_2}$ because $|P_i^{\mathcal{M}}| < |P_i^{\mathcal{M}_2}|$ and $m = |P_i^{\mathcal{M}}| + 1$. Therefore, $\varphi^{r,i}_{\mathcal{M}}$ is equivalent to $|P_i| \geq |P_i^{\mathcal{M}}| \wedge |P_i| < |P_i^{\mathcal{M}}| + 1$. □

Lemma 4.7 *Let r be a natural number and $\mathcal{M}$ be a structure in* MS. There exists a set of monadic structures V such that $\varphi^r_{\mathcal{M}}$ is a boolean combination of sentences in $\bigcup_{\mathcal{M}' \in V} \Phi^r_{\mathcal{M},\mathcal{M}'}$.

Proof. By Lemma 4.5, $\models \varphi^r_{\mathcal{M}} \leftrightarrow \bigwedge_{i=1}^{k+1} \varphi^{r,i}_{\mathcal{W}}$. Let $V = \bigcup_{i=1}^{k+1} V_i$ such that V_i is as in Lemma 4.6. It follows that, $\varphi^r_{\mathcal{M}}$ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{M}' \in V} \Phi^r_{\mathcal{M},\mathcal{M}'}$. □

By Theorem 2.6, any first-order sentence is a disjunction of Hintikka formulas. Thus, the following result states that, over MS, any first-order sentence is equivalent to a boolean combination of distinguishability sentences.

Theorem 4.8 *Let φ be a first-order sentence over* MS. Then, there exists two sets U, V of monadic structures such that φ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{M} \in U, \mathcal{M}' \in V} \Phi^r_{\mathcal{M},\mathcal{M}'}$.

Proof. Let r such that $qr(\varphi) = r$. From Theorem 2.6 it follows that $\models \varphi \leftrightarrow \varphi^r_{\mathcal{M}_1} \vee ...\varphi^r_{\mathcal{M}_s}$. Let $U = \{\mathcal{M}_1, ..., \mathcal{M}_s\}$. In accord to Lemma 4.7, let V_j be such that $\varphi^r_{\mathcal{M}_j}$ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{M}' \in V_j} \Phi^r_{\mathcal{M}_j,\mathcal{M}'}$. Therefore, φ is equivalent to a boolean combination of sentences in $\bigcup_{j=1}^{s}(\bigcup_{\mathcal{M}' \in V_j} \Phi^r_{\mathcal{M}_j,\mathcal{M}'})$ and $\bigcup_{j=1}^{s}(\bigcup_{\mathcal{M}' \in V_j} \Phi^r_{\mathcal{M}_j,\mathcal{M}'}) \subseteq \bigcup_{\mathcal{M} \in U, \mathcal{M}' \in V} \Phi^r_{\mathcal{M},\mathcal{M}'}$. □

4.2 Equivalence Structures

In this subsection, we deal with the distinguishability sentences for equivalence structures. First, we define the following formulas:

$$\varphi_{q_{\geq t} \geq p} = \exists x_1 ... \exists x_p (\bigwedge_{l \neq j} \neg E(x_l, x_j) \wedge \bigwedge_{k=1}^{p} \exists y_2 ... \exists y_t (\bigwedge_{j=2}^{t} y_j \neq x_k \wedge \bigwedge_{l<j} y_l \neq y_j \wedge \bigwedge_{j=2}^{t} E(y_j, x_k))).$$

$$\varphi_{q_t \geq p} := \begin{array}{l} \exists x_1 ... \exists x_p (\bigwedge_{l<j} \neg E(x_l, x_j) \wedge \bigwedge_{k=1}^{p} \exists y_2 ... \exists y_t (\bigwedge_{j=2}^{t} y_j \neq x_k \wedge \bigwedge_{l \neq j} y_l \neq y_j \wedge \\ \bigwedge_{i=2}^{t} E(y_i, x_k) \wedge \forall z (E(z, x_k) \rightarrow (z = x_k \vee \bigvee_{j=2}^{t} z = y_j)))). \end{array}$$

Sentences of the form $\varphi_{q_{\geq t} \geq p}$ hold on equivalence structures such that the number of equivalence classes of size at least t is at least p. Each variable x_k represents an element in a distinct equivalence class. Variables y_i guarantee that a class has at least t elements. Formulas $\varphi_{q_t \geq p}$ are true when the number of equivalence classes of size t is at least p. Variable z forces that any element from an equivalence class represented by an element x_k is x_k or one of y_i for $i \in \{2, ..., t\}$. We also define $\varphi_{q_t < p} := \neg \varphi_{q_t \geq p}$ and $\varphi_{q_{\geq t} < p} := \neg \varphi_{q_{\geq t} \geq p}$. Clearly, $qr(\varphi_{q_{\geq t} \geq p}) = t + p - 1$ and $qr(\varphi_{q_t \geq p}) = t + p$. Also, the size of $\varphi_{q_{\geq t} \geq p}$ and $\varphi_{q_t \geq p}$ is $O((p+t)^3)$. Next, we define the distinguishability sentences for equivalence structures.

Definition 4.9 [Distinguishability Sentences for ES] Let $\mathcal{E}_1$, $\mathcal{E}_2$ be equivalence structures and r be a natural number.

$$\Phi^r_{\mathcal{E}_1,\mathcal{E}_2} := \begin{array}{l} \{\varphi_{q_t < m} \mid 1 \leq t \leq r, q_t^{\mathcal{E}_1} < q_t^{\mathcal{E}_2}, q_t^{\mathcal{E}_1} + 1 \leq m \leq min(r - t, q_t^{\mathcal{E}_2})\} \cup \\ \{\varphi_{q_t \geq m} \mid 1 \leq t \leq r, q_t^{\mathcal{E}_1} > q_t^{\mathcal{E}_2}, q_t^{\mathcal{E}_2} + 1 \leq m \leq min(r - t, q_t^{\mathcal{E}_1})\} \cup \\ \{\varphi_{q_{\geq t} < m} \mid 1 \leq t \leq r, q_{\geq t}^{\mathcal{E}_1} < q_{\geq t}^{\mathcal{E}_2}, q_{\geq t}^{\mathcal{E}_1} + 1 \leq m \leq min(r - t + 1, q_{\geq t}^{\mathcal{E}_2})\} \cup \\ \{\varphi_{\geq q_t \geq m} \mid 1 \leq t \leq r, q_{\geq t}^{\mathcal{E}_1} > q_{\geq t}^{\mathcal{E}_2}, q_{\geq t}^{\mathcal{E}_2} + 1 \leq m \leq min(r - t + 1, q_{\geq t}^{\mathcal{E}_1})\}. \end{array}$$

For equivalence structures $\mathcal{E}_1$, $\mathcal{E}_2$, and a natural number r, the size of a sentence $\varphi \in \Phi^r_{\mathcal{E}_1,\mathcal{E}_2}$ is $O((|E_1| + |E_2|)^3)$. Furthermore, $|\Phi^r_{\mathcal{E}_1,\mathcal{E}_2}|$ is $O((|E_1| + |E_2|)^2)$. In what follows, we give examples of the distinguishability sentences for equivalence structures.

Example 4.10 Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be equivalence structures such that $q_2^{\mathcal{E}_1} = 3$ and $q_2^{\mathcal{E}_2} = 2$, and $r = 5$. Then, $\varphi_{q_2 \geq 3} \in \Phi^r_{\mathcal{E}_1,\mathcal{E}_2}$ because $q_2^{\mathcal{E}_1} > q_2^{\mathcal{E}_2}$ and, for $m = 3$, $q_2^{\mathcal{E}_2} + 1 \leq m \leq min(r - 2, q_2^{\mathcal{E}_1})$.

Now, we show results ensuring that the distinguishability sentences for equivalence structures hold in the adequate equivalence structures and have quantifier rank at most r.

Lemma 4.11 *Let $\mathcal{E}_1, \mathcal{E}_2$ be equivalence structures, and r be a natural number. Let $\varphi \in \Phi^r_{\mathcal{E}_1,\mathcal{E}_1}$. Then, $\mathcal{E}_1 \models \varphi$ and $\mathcal{E}_2 \not\models \varphi$.*

Proof. Suppose $\varphi = \varphi_{q_t<m}$. Then, $q_t^{\mathcal{E}_1} < q_t^{\mathcal{E}_2}$ and $q_t^{\mathcal{E}_1} + 1 \leq m \leq min(r-t, q_t^{\mathcal{E}_2})$. Then, $\mathcal{E}_1 \models \varphi$ because $q_t^{\mathcal{E}_1} < m$. Also, as $m \leq q_t^{\mathcal{E}_2}$, $\mathcal{E}_2 \not\models \varphi$. The other cases are analogous. □

Lemma 4.12 *Let $\mathcal{E}_1, \mathcal{E}_2$ be equivalence structures, and r be a natural number. Let $\varphi \in \Phi^r_{\mathcal{E}_1,\mathcal{E}_2}$. Then, $qr(\varphi) \leq r$.*

Proof. If $\varphi = \varphi_{q_t \triangleright m}$ where $\triangleright \in \{<, \geq\}$, then $qr(\varphi) = t + m$. As $m \leq r - t$, then $qr(\varphi) \leq t + r - t = r$. If $\varphi = \varphi_{q_{\geq t} \triangleright m}$, then $qr(\varphi) = t + m - 1$. Therefore, as $m \leq r - t + 1$, $qr(\varphi) \leq t + r - t + 1 - 1 = r$. □

Now, we show that, over ES, any first-order sentence is equivalent to a boolean combination of distinguishability sentences. First, we need the following lemmas.

Lemma 4.13 $\models \varphi^r_{\mathcal{E}} \leftrightarrow (\bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_t} \wedge \bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_{\geq t}})$ *such that*

$$\varphi_{\mathcal{E}}^{r,q_t} := \begin{cases} \varphi_{q_t = q_t^{\mathcal{E}}}, & \text{if } q_t^{\mathcal{E}} + t + 1 \leq r \\ \varphi_{q_t > r-t-1}, & \text{otherwise.} \end{cases}$$

$$\varphi_{\mathcal{E}}^{r,q_{\geq t}} := \begin{cases} \varphi_{q_{\geq t} = q_{\geq t}^{\mathcal{E}}}, & \text{if } q_{\geq t}^{\mathcal{E}} + t \leq r \\ \varphi_{q_{\geq t} > r-t}, & \text{otherwise.} \end{cases}$$

Proof. Let $\mathcal{E}' \models \varphi^r_{\mathcal{E}}$. Then, the Duplicator has a winning strategy in $\mathcal{G}_r(\mathcal{E}, \mathcal{E}')$. Then, for all t, $q_t^{\mathcal{E}} = q_t^{\mathcal{E}'}$ or $r < min\{q_t^{\mathcal{E}}, q_t^{\mathcal{E}'}\} + t + 1$, and for all t, $q_{\geq t}^{\mathcal{E}} = q_{\geq t}^{\mathcal{E}'}$ or $r < min\{q_{\geq t}^{\mathcal{E}}, q_{\geq t}^{\mathcal{E}'}\} + t$. First, let t such that $r < q_t^{\mathcal{E}} + t + 1$. Then, $r < q_t^{\mathcal{E}'} + t + 1$ because $q_t^{\mathcal{E}} = q_t^{\mathcal{E}'}$ or $r < min\{q_t^{\mathcal{E}}, q_t^{\mathcal{E}'}\} + t + 1$. Besides, $\varphi_{\mathcal{E}}^{r,q_t} = \varphi_{q_t > r-t-1}$. Therefore, $\mathcal{E}' \models \varphi_{\mathcal{E}}^{r,q_t}$. If $r \geq q_t^{\mathcal{E}} + t + 1$, then $q_t^{\mathcal{E}} = q_t^{\mathcal{E}'}$. Therefore, $\mathcal{E}' \models \varphi_{q_t = q_t^{\mathcal{E}}}$. The case for $q_{\geq t}^{\mathcal{E}}$ is analogous. Then, $\mathcal{E}' \models (\bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_t} \wedge \bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_{\geq t}})$. Conversely, suppose that $\mathcal{E}' \models (\bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_t} \wedge \bigwedge_{t=1}^r \varphi_{\mathcal{E}}^{r,q_{\geq t}})$. Let t such that $\varphi_{\mathcal{E}}^{r,q_t} = \varphi_{q_t > r-t-1}$. Then, $r < q_t^{\mathcal{E}} + t + 1$ and $r < q_t^{\mathcal{E}'} + t + 1$. Therefore, $r < min\{q_t^{\mathcal{E}}, q_t^{\mathcal{E}'}\} + t + 1$. Let t such that $\varphi_{\mathcal{E}}^{r,q_t} = \varphi_{q_t = q_t^{\mathcal{E}}}$. Clearly, $q_t^{\mathcal{E}} = q_t^{\mathcal{E}'}$. The case for $\varphi_{\mathcal{E}}^{r,q_{\geq t}}$ is analogous. Then, for all t, $q_t^{\mathcal{E}} = q_t^{\mathcal{E}'}$ or $r < min\{q_t^{\mathcal{E}}, q_t^{\mathcal{E}'}\} + t + 1$, and for all t, $q_{\geq t}^{\mathcal{E}} = q_{\geq t}^{\mathcal{E}'}$ or $r < min\{q_{\geq t}^{\mathcal{E}}, q_{\geq t}^{\mathcal{E}'}\} + t$. Therefore, $\mathcal{E}' \models \varphi^r_{\mathcal{E}}$. □

Lemma 4.14 *Let r be a natural number, and $\mathcal{E}$ be an equivalence structure. There exists sets of equivalence structures V_t, $V_{\geq t}$ such that $\varphi_{\mathcal{E}}^{r,q_t}$, $\varphi_{\mathcal{E}}^{r,q_{\geq t}}$ are equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{E}' \in V_t} \Phi^r_{\mathcal{E},\mathcal{E}'}$ and $\bigcup_{\mathcal{E}' \in V_{\geq t}} \Phi^r_{\mathcal{E},\mathcal{E}'}$, respectively.*

Proof. If $\varphi_{\mathcal{E}}^{r,q_t} = \varphi_{q_t > r-t-1}$, then $q_t^{\mathcal{E}} \geq r-t$. Let $V_t = \{\mathcal{E}'\}$ such that $q_t^{\mathcal{E}'} = r-t-1$. Then, $\varphi_{q_t \geq r-t} \in \Phi^r_{\mathcal{E},\mathcal{E}'}$ because $r-t-1+1 \leq r-t \leq min(r-t, q_t^{\mathcal{E}})$. If $\varphi_{\mathcal{E}}^{r,q_t} = \varphi_{q_t = q_t^{\mathcal{E}}}$, then $q_t^{\mathcal{E}} < r-t$. $V_t = \{\mathcal{E}_1, \mathcal{E}_2\}$ such that $q_t^{\mathcal{E}_1} = q_t^{\mathcal{E}} - 1$ and $q_t^{\mathcal{E}_2} = q_t^{\mathcal{E}} + 1$. Then, $\varphi_{q_t \geq q_t^{\mathcal{E}}} \in \Phi^r_{\mathcal{E},\mathcal{E}_1}$ because, for $m = q_t^{\mathcal{E}}$, $q_t^{\mathcal{E}_1} \leq m \leq min(r-t, q_t^{\mathcal{E}})$ and $q_t^{\mathcal{E}} < r-t$. Also, $\varphi_{q_t < q_t^{\mathcal{E}}+1} \in \Phi^r_{\mathcal{E},\mathcal{E}_2}$ as long as $q_t^{\mathcal{E}} + 1 \leq min(r-t, q_t^{\mathcal{B}})$ and $q_t^{\mathcal{E}} + 1 \geq r-t$. For $\varphi_{\mathcal{E}}^{r,q_{\geq t}}$, we define $V_{\geq t}$ in an analogous way. □

Lemma 4.15 *Let r be a natural number, and $\mathcal{E}$ be an equivalence structure. There exists a set of equivalence structures V such that $\varphi^r_{\mathcal{E}}$ is a boolean combination of sentences in $\bigcup_{\mathcal{E}' \in V} \Phi^r_{\mathcal{E},\mathcal{E}'}$.*

Proof. This proof can be directly adapted from Lemma 4.7. □

The following result states that, over equivalence structures, any first-order sentence is equivalent to a boolean combination of distinguishability sentences.

Theorem 4.16 *Let φ be a first-order sentence over ES. Then, there exists a natural number r, and two sets U, V of equivalence structures such that φ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{E} \in U, \mathcal{E}' \in V} \Phi^r_{\mathcal{E},\mathcal{E}'}$.*

Proof. This proof can be directly adapted from Theorem 4.8. □

4.3 Disjoint Unions of Linear Orders

Now, we define the distinguishability sentences for disjoint unions of linear orders. First, we define the following formulas:

$$\varphi_{\geq n}^{<x,>y} = \begin{cases} \exists z(z = z), & \text{se } n = 0 \\ \exists z(z < x \wedge y < z), & \text{se } n = 1 \\ \exists z(z < x \wedge y < z \wedge \varphi_{\geq \lfloor \frac{n-1}{2} \rfloor}^{<z,>y} \wedge \varphi_{\geq \lfloor \frac{n}{2} \rfloor}^{<x,>z}), & \text{c.c.} \end{cases}$$

$$\varphi_{\geq n}^{<x} = \begin{cases} \exists y(y = y), & \text{se } n = 0 \\ \exists y(y < x), & \text{se } n = 1 \\ \exists y(y < x \wedge \varphi_{\geq \lfloor \frac{n-1}{2} \rfloor}^{<y} \wedge \varphi_{\geq \lfloor \frac{n}{2} \rfloor}^{<x,>y}), & \text{c.c.} \end{cases}$$

$$\varphi_{\geq n}^{>x} = \begin{cases} \exists y(y = y), & \text{se } n = 0 \\ \exists y(x < y), & \text{se } n = 1 \\ \exists y(y > x \wedge \varphi_{\geq \lfloor \frac{n-1}{2} \rfloor}^{>y} \wedge \varphi_{\geq \lfloor \frac{n}{2} \rfloor}^{<y,>x}), & \text{c.c.} \end{cases}$$

$$\varphi_{q_t \geq p} := \exists x_1 ... \exists x_p (\bigwedge_{l \neq j} (x_l \neq x_j \wedge \neg(x_l < x_j) \wedge \neg(x_l > x_j)) \wedge \bigwedge_{k=1}^{p} (\varphi_{=\lfloor \frac{t-1}{2} \rfloor}^{<x_k} \wedge \varphi_{=\lfloor \frac{t}{2} \rfloor}^{>x_k})).$$

$$\varphi_{q_{\geq t} \geq p} = \exists x_1 ... \exists x_p (\bigwedge_{l \neq j} (x_l \neq x_j \wedge \neg(x_l < x_j) \wedge \neg(x_l > x_j)) \wedge \bigwedge_{k=1}^{p} (\varphi_{\geq \lfloor \frac{t-1}{2} \rfloor}^{<x_k} \wedge \varphi_{\geq \lfloor \frac{t}{2} \rfloor}^{>x_k})).$$

Formulas of the form $\varphi^{\star x}_{\geq n}$ such that $\star \in \{<, >\}$ and $\varphi^{<x,>y}_{\geq n}$ are used in the definition of $\varphi_{q_t \geq p}$ and $\varphi_{q_{\geq t} \geq p}$. We also define abbreviations $\varphi^{*x}_{<n} := \neg\varphi^{*x}_{\geq n}$, $\varphi^{*x}_{\leq n} := \neg\varphi^{*x}_{\geq n+1}$, and $\varphi^{*x}_{=n} := \varphi_{\geq n} \wedge \varphi_{\leq n}$, for $* \in \{<, >\}$. These formulas are defined recursively in order to obtain the adequate quantifier rank. The recursive definitions can all be simplified to direct definitions with higher quantifier ranks but, in this case, we can not guarantee that the quantifier rank is minimum. Note that $qr(\varphi^{*x}_{\geq n}) = \lfloor log_2(n) \rfloor + 1$. Sentences of the form $\varphi_{q_t \geq p}$ express that the number q_t of linear orders of size t is at least p. Each variable x_k represents an element in a linear order. Analogously, $\varphi_{q_{\geq t} \geq p}$ holds in disjoint unions of linear orders such that the number $q_{\geq t}$ of linear orders of size at least t is at least p. We also define $\varphi_{q_t < p} := \neg\varphi_{q_t \geq p}$, $\varphi_{q_{\geq t} < p} := \neg\varphi_{q_{\geq t} \geq p}$. Finally, regarding the quantifier rank, $qr(\varphi_{q_t \geq p}) = p + \lfloor log_2(\lfloor \frac{t}{2} + 1 \rfloor) \rfloor + 1 = p + \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor + 1$, and $qr(\varphi_{q_{\geq t} \geq p}) = p + \lfloor log_2(t) \rfloor$. Furthermore, the size of $\varphi_{q_t \geq p}$ and $\varphi_{q_{\geq t} \geq p}$ is $O((p+n)^2)$. Now, we define the distinguishability sentences for disjoint unions of linear orders.

Definition 4.17 [Distinguishability Sentences for DULO] Let $\mathcal{W}_1$, $\mathcal{W}_2$ be disjoint unions of linear orders and r be a natural number.

$$\Phi^r_{\mathcal{W}_1,\mathcal{W}_2} := \begin{array}{l} \{\varphi_{q_t<p} \mid t < 2^r - 2, q_t^{\mathcal{W}_1} < q_t^{\mathcal{W}_2}, q_t^{\mathcal{W}_1} + 1 \leq p \leq min(q_t^{\mathcal{W}_2}, r - \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor - 1)\} \cup \\ \{\varphi_{q_t \geq p} \mid t < 2^r - 2, q_t^{\mathcal{W}_2} < q_t^{\mathcal{W}_1}, q_t^{\mathcal{W}_2} + 1 \leq p \leq min(q_t^{\mathcal{W}_1}, r - \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor - 1)\} \cup \\ \{\varphi_{q_{\geq t}<p} \mid t \leq 2^r - 1, q_{\geq t}^{\mathcal{W}_1} < q_{\geq t}^{\mathcal{W}_2}, q_t^{\mathcal{W}_1} + 1 \leq p \leq min(q_t^{\mathcal{W}_2}, r - \lfloor log_2(t) \rfloor)\} \cup \\ \{\varphi_{q_{\geq t} \geq p} \mid t \leq 2^r - 1, q_{\geq t}^{\mathcal{W}_2} < q_{\geq t}^{\mathcal{W}_1}, q_t^{\mathcal{W}_2} + 1 \leq p \leq min(q_t^{\mathcal{W}_1}, r - \lfloor log_2(t) \rfloor)\}. \end{array}$$

Given $\mathcal{W}_1$, $\mathcal{W}_2$, and r, the size of a sentence $\varphi \in \Phi^r_{\mathcal{W}_1,\mathcal{W}_2}$ is $O((|W_1| + |W_2|)^2)$. Since $t, p \leq |W_1| + |W_2|$, $|\Phi^r_{\mathcal{W}_1,\mathcal{W}_2}|$ is $O((|W_1| + |W_2|)^2)$. Now, we show results ensuring adequate properties of the distinguishability sentences for disjoint unions of linear orders. These results can be directly adapted from Lemma 4.11 and Lemma 4.12.

Lemma 4.18 *Let $\varphi \in \Phi^r_{\mathcal{W}_1,\mathcal{W}_2}$. Then, $\mathcal{W}_1 \models \varphi$ and $\mathcal{W}_2 \not\models \varphi$.*

Lemma 4.19 *Let $\varphi \in \Phi^r_{\mathcal{W}_1,\mathcal{W}_2}$. Then, $qr(\varphi) \leq r$.*

Now we show that, over DULO, any first-order sentence is equivalent to a boolean combination of distinguishability sentences. The following results can be directly adapted from Lemma 4.13, Lemma 4.14, Lemma 4.15, and Theorem 4.16.

Lemma 4.20 $\models \varphi^r_{\mathcal{W}} \leftrightarrow (\bigwedge_{t=1}^{2^r-3} \varphi^{r,q_t}_{\mathcal{W}} \wedge \bigwedge_{t=1}^{2^r-1} \varphi^{r,q_{\geq t}}_{\mathcal{W}})$ *such that*

$$\varphi^{r,q_t}_{\mathcal{W}} := \begin{cases} \varphi_{q_t = q_t^{\mathcal{W}}}, & \text{if } q_t^{\mathcal{W}} + \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor + 2 \leq r \\ \varphi_{q_t > r - \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor - 2}, & \text{otherwise.} \end{cases}$$

$$\varphi^{r,q_{\geq t}}_{\mathcal{W}} := \begin{cases} \varphi_{q_{\geq t} = q_{\geq t}^{\mathcal{W}}}, & \text{if } q_{\geq t}^{\mathcal{W}} + \lfloor log(t) \rfloor + 1 \leq r \\ \varphi_{q_{\geq t} > r - \lfloor log(t) \rfloor - 1}, & \text{otherwise.} \end{cases}$$

Theorem 4.21 *Let φ be a first-order sentence over disjoint unions of linear orders. Then, there exists sets V, U of disjoint unions of linear orders such that φ is equivalent to a boolean combination of sentences in $\bigcup_{\mathcal{W} \in V, \mathcal{W}' \in U} \Phi^r_{\mathcal{W},\mathcal{W}'}$.*

Algorithm 1.
Input: Sample $S = (P, N)$
$r \leftarrow max\{EFsim(\mathcal{A}, \mathcal{B}) \mid \mathcal{A} \in P, \mathcal{B} \in N\}$
$\varphi_S \leftarrow \bigvee_{\mathcal{A}\in P} \bigwedge_{\mathcal{B}\in N}$ **choose** $\varphi \in \Phi^r_{\mathcal{A},\mathcal{B}}$
return φ_S

5 The Algorithm for the Distinguishability Problem

In this section, we define an algorithm for finding a first-order sentence φ_S from a sample of structures S. Subformulas of φ_S are distinguishability sentences from sets of the form $\Phi^r_{\mathcal{A},\mathcal{B}}$ such that $\mathcal{A} \in P$ and $\mathcal{B} \in N$. We also give an example of how the algorithm works. We guarantee that our algorithm runs in polynomial time in the size of the input sample S. The size of the sample S is the sum of the sizes of all the structures it includes. We also show that φ_S returned by our algorithm is consistent with S. Furthermore, we also prove that φ_S is a sentence of minimal quantifier rank consistent with S. The pseudocode of our algorithm is in Algorithm 1.

First, the algorithm finds the minimum value r such that there exists a sentence of quantifier rank r that is consistent with the input sample S. After that, the algorithm builds φ_S. For each pair of structures $\mathcal{A} \in P$, $\mathcal{B} \in N$, it chooses $\varphi \in \Phi^r_{\mathcal{A},\mathcal{B}}$. Any choice of a sentence in $\Phi^r_{\mathcal{A},\mathcal{B}}$ leads to a formula consistent with S. In what follows, we show an example of how this algorithm works on a simple instance.

Example 5.1 Let $P = \{\mathcal{W}_1\}$, and $N = \{\mathcal{W}_2, \mathcal{W}_3\}$ such that $q_1^{\mathcal{W}_1} = 0$, $q_2^{\mathcal{W}_1} = 0$, $q_3^{\mathcal{W}_1} = 1$, $q_{\geq 4}^{\mathcal{W}_1} = 0$, $q_1^{\mathcal{W}_2} = 3$, $q_{\geq 2}^{\mathcal{W}_2} = 0$, $q_1^{\mathcal{W}_3} = 1$, $q_2^{\mathcal{W}_3} = 1$, $q_{\geq 3}^{\mathcal{W}_3} = 0$. Then, $EFsim(\mathcal{W}_1, \mathcal{W}_2) = 2$ as witnessed by $t = 2$, $q_{\geq t}^{\mathcal{W}_1} = 1$, $q_{\geq t}^{\mathcal{W}_2} = 0$, and $min(q_{\geq t}^{\mathcal{W}_1}, q_{\geq t}^{\mathcal{W}_2}) + \lfloor log_2(t) \rfloor + 1 = 2$. Furthermore, $EFsim(\mathcal{W}_1, \mathcal{W}_3) = 2$ as witnessed by $t = 1$, $q_t^{\mathcal{W}_1} = 0$, $q_t^{\mathcal{W}_3} = 1$, and $min(q_t^{\mathcal{W}_1}, q_t^{\mathcal{W}_3}) + \lfloor log_2(\lceil \frac{t+1}{2} \rceil) \rfloor + 2 = 2$. Therefore, $max\{EFsim(\mathcal{W}, \mathcal{W}') \mid \mathcal{W} \in P, \mathcal{W}' \in N\} = 2$. Then, $\varphi_{q_{\geq 2} \geq 1} \in \Phi^2_{\mathcal{W}_1,\mathcal{W}_2}$ because $q_{\geq 2}^{\mathcal{W}_1} > q_{\geq 2}^{\mathcal{W}_2}$ and, for $p = 1$, $1 \leq p \leq min(1, 1)$. Besides, $\varphi_{q_{\geq 1} < 2} \in \Phi^2_{\mathcal{W}_1,\mathcal{W}_3}$ since $q_{\geq 1}^{\mathcal{W}_1} < q_{\geq 1}^{\mathcal{W}_3}$, and $q_{\geq 1}^{\mathcal{W}_1} + 1 \leq 2 \leq min(q_{\geq 1}^{\mathcal{W}_3}, 2)$. Therefore, Algorithm 1 returns $\varphi_S = \varphi_{q_{\geq 2} \geq 1} \wedge \varphi_{q_{\geq 1} < 2}$.

In the following, we prove some properties of our algorithm. First, we show that it returns a sentence that is consistent with the sample. After that, we show that it returns a sentence of minimal quantifier rank. Then, we prove that the running time of our algorithm is polynomial in the size of the input sample.

Theorem 5.2 *Let S be a sample and φ_S returned by Algorithm 1. Then, φ_S is consistent with S.*

Proof. Let $\varphi_S = \bigvee_{\mathcal{A}\in P} \bigwedge_{\mathcal{B}\in N} \varphi^r_{\mathcal{A},\mathcal{B}}$ such that $r = max\{EFsim(\mathcal{A}, \mathcal{B}) \mid \mathcal{A} \in P, \mathcal{B} \in N\}$, and $\varphi^r_{\mathcal{A},\mathcal{B}}$ is the sentence chosen from $\Phi^r_{\mathcal{A},\mathcal{B}}$. Let $\mathcal{A}' \in P$. In this way, $\mathcal{A}' \models \bigwedge_{\mathcal{B}\in N} \varphi^r_{\mathcal{A}',\mathcal{B}}$, and then $\mathcal{A}' \models \varphi_S$. Now, let $\mathcal{B}' \in N$ and assume that $\mathcal{B}' \models \varphi_S$, i.e, $\mathcal{B}' \models \bigwedge_{\mathcal{B}\in N} \varphi^r_{\mathcal{A}',\mathcal{B}}$, for some $\mathcal{A}' \in P$. Therefore, $\mathcal{B}' \models \varphi^r_{\mathcal{A}',\mathcal{B}'}$. This is an absurd because, from Lemma 4.18, it follows that $\mathcal{B}' \not\models \varphi^r_{\mathcal{A}',\mathcal{B}'}$. □

Theorem 5.3 *The sentence φ_S returned by Algorithm 1 is a first-order sentence of minimal quantifier rank that is consistent with S.*

Proof. Suppose a first-order sentence ψ consistent with S such that $qr(\psi) < qr(\varphi_S) = max\{EFsim(\mathcal{A},\mathcal{B}) \mid \mathcal{A} \in P, \mathcal{B} \in N\}$. Let $\mathcal{A}' \in P$ and $\mathcal{B}' \in N$ such that $EFsim(\mathcal{A}',\mathcal{B}') = max\{EFsim(\mathcal{A},\mathcal{B}) \mid \mathcal{A} \in P, \mathcal{B} \in N\}$. Then, $\mathcal{A}'$ and $\mathcal{B}'$ are satisfied by the same first-order sentences of quantifier rank q such that $q < EFsim(\mathcal{A}',\mathcal{B}')$. Then, $\mathcal{A}' \models \psi$ iff $\mathcal{B}' \models \psi$. Therefore, ψ is not consistent with S. This is a contradiction. □

Theorem 5.4 *Given a sample S, Algorithm 1 returns φ_S in polynomial time in the size of S.*

Proof. First, the algorithm computes $max\{EFsim(\mathcal{A},\mathcal{B}) \mid \mathcal{A} \in P, \mathcal{B} \in N\}$ in order to use a suitable quantifier rank. This takes polynomial time because, for a given $\mathcal{A} \in P, \mathcal{B} \in N$, to compute $EFsim(\mathcal{A},\mathcal{B})$ takes polynomial time in the classes of structures we are considering in this work. Then, our algorithm loops over structures in the sample and, in each loop, it chooses a formula $\varphi \in \Phi^r_{\mathcal{A},\mathcal{B}}$. As the size of each $\varphi \in \Phi^r_{\mathcal{A},\mathcal{B}}$ is polynomial in the size of $\mathcal{A}$ and $\mathcal{B}$, and $|\Phi^r_{\mathcal{A},\mathcal{B}}|$ is polynomial in the size of S, the overall complexity of Algorithm 1 is polynomial time in the size of S. □

6 Conclusions and Future Work

We introduced an algorithm that returns, in polynomial time, a first-order sentence of minimal quantifier rank that is consistent with a given sample of structures. Our algorithm works with monadic structures, equivalence structures, and disjoint unions of linear orders. Our work is motivated by the algorithm defined in [14] that runs in exponential time for these classes of structures. Therefore, our approach is an improvement over the work in [14], for the particular problem on MS, ES, and DULO. Also, we think that our algorithm can be useful when it is desirable to compute sentences defining a class of structures from positive and negative structures. For example, as we have outlined in the introduction, disjoint unions of linear orders may represent states of the elementary blocks world planning. Therefore, for disjoint unions of linear orders, our algorithm can be used to define initial and final states of this problem. In the elementary blocks world planning, given initial and final states, the goal is to find a sequence of actions capable of transforming the initial state into a final state.

The algorithm defined in [14] uses Hintikka formulas which have size exponential in the size of a given structure. Then, we also introduced the distinguishability sentences which are defined based on the conditions characterizing winning strategies for the Spoiler on MS, ES, and DULO. Given two structures $\mathcal{A},\mathcal{B}$ and a natural number r, the distinguishability sentences hold in $\mathcal{A}$, do not hold in $\mathcal{B}$, and have quantifier rank at most r. Besides, the size of a distinguishability sentence is polynomial in the size of these structures. Furthermore, given $\mathcal{A},\mathcal{B}$ and r, the number of distinguishability formulas is also polynomial in the size of $\mathcal{A}$ and $\mathcal{B}$. Finally, we also show that any first-order sentence is equivalent to a boolean combination of

distinguishability sentences. This result suggests that our approach is likely to find any first-order sentence given a suitable sample of strings.

As future work, we intend to investigate algorithms for the distinguishability problem on other classes of structures. For example, equivalence structures with colors, embedded equivalence structures, and trees with level predicates [15]. For these classes of structures, the problem of determining if the Spoiler has a winning strategy is solved in exponential time. Besides, we intend to consider the class of strings with a built-in linear order [18]. These classes are interesting because first-order logic over strings with a built-in linear order captures star-free languages [24]. An algorithm for the distinguishability problem over strings can be used to derive a recognizer of a star-free language from a sample of strings. This result is significant because strings may be used to model text data, traces of program executions, DNA sequences, and sequences of symbolic data in general.

Finally, we plan to extend our approach to other logics such as monadic second-order logic. Regular languages are exactly the languages definable in monadic second-order logic over strings [1,16]. An algorithm which returns monadic second-order sentences consistent with a given sample of strings can be used in the problem of finding a model of a regular language consistent with a given sample of strings [12,25].

References

[1] Richard J. Büchi. Weak second-order arithmetic and finite automata. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 6(1-6):66–92, 1960.

[2] Stephen A. Cook and Yongmei Liu. A complete axiomatization for blocks world. *Journal of Logic and Computation*, 13(4):581–594, 2003.

[3] Luc De Raedt. *Logical and Relational Learning*. Springer, 2008.

[4] Heinz-Dieter Ebbinghaus and Jörg Flum. *Finite Model Theory*. Springer, 1995.

[5] Heinz-Dieter Ebbinghaus, Jörg Flum, and Wolfgang Thomas. *Mathematical Logic*. Springer, 1994.

[6] Andrzej Ehrenfeucht. An application of games to the completeness problem for formalized theories. *Fundamenta Mathematicae*, 49(2):129–141, 1961.

[7] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated Planning: Theory and Practice*. Elsevier, 2004.

[8] Erich Grädel, P. G. Kolaitis, L. Libkin, M. Marx, J. Spencer, Moshe Y. Vardi, Y. Venema, and Scott Weinstein. *Finite Model Theory and Its Applications*. Springer, 2005.

[9] Martin Grohe, Christof Löding, and Martin Ritzert. Learning MSO-definable hypotheses on strings. In *International Conference on Algorithmic Learning Theory*, pages 434–451. PMLR, 2017.

[10] Martin Grohe and Martin Ritzert. Learning first-order definable concepts over structures of small degree. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12. IEEE, 2017.

[11] Naresh Gupta and Dana S. Nau. On the complexity of blocks-world planning. *Artificial Intelligence*, 56(2-3):223–254, 1992.

[12] Marijn Heule and Sicco Verwer. Exact DFA identification using SAT solvers. In *International Colloquium on Grammatical Inference*, pages 66–79. Springer, 2010.

[13] Charles Jordan and Łukasz Kaiser. Experiments with reduction finding. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 192–207. Springer, 2013.

[14] Łukasz Kaiser. Learning games from videos guided by descriptive complexity. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 963–969. AAAI Press, 2012.

[15] Bakhadyr Khoussainov and Jiamou Liu. On complexity of Ehrenfeucht-Fraïssé games. *Annals of Pure and Applied Logic*, 161(3):404 – 415, 2009.

[16] Richard E. Ladner. Application of model theoretic games to discrete linear orders and finite automata. *Information and Control*, 33(4):281 – 303, 1977.

[17] Leonid Libkin. *Elements Of Finite Model Theory*. Springer, 2004.

[18] Elisabetta De Maria, Angelo Montanari, and Nicola Vitacolonna. Games on strings with a limited order relation. In *International Symposium on Logical Foundations of Computer Science*, pages 164–179. Springer, 2009.

[19] Angelo Montanari, Alberto Policriti, and Nicola Vitacolonna. An algorithmic account of Ehrenfeucht games on labeled successor structures. In *Logic for Programming, Artificial Intelligence, and Reasoning*, pages 139–153. Springer, 2005.

[20] Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, 1991.

[21] Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

[22] Elena Pezzoli. Computational complexity of Ehrenfeucht-Fraïssé games on finite structures. In *Computer Science Logic*, pages 159–170. Springer, 1999.

[23] Thiago Alves Rocha, Ana Teresa Martins, and Francicleber Martins Ferreira. On finding a first-order sentence consistent with a sample of strings. In *International Symposium on Games, Automata, Logics, and Formal Verification*, pages 220–234. Open Publishing Association, 2018.

[24] Wolfgang Thomas. Classifying regular events in symbolic logic. *Journal of Computer and System Sciences*, 25(3):360–376, 1982.

[25] Ilya Zakirzyanov, Anatoly Shalyto, and Vladimir Ulyantsev. Finding all minimum-size DFA consistent with given examples: SAT-based approach. In *International Conference on Software Engineering and Formal Methods*, pages 117–131. Springer, 2017.