On preservation theorems for two-variable logic

Erich Grädel and Eric Rosen
Mathematische Grundlagen der Informatik
RWTH Aachen
graedel,erosen@informatik.rwth-aachen.de

Abstract: We show that the existential preservation theorem fails for FO^2, two-variable first-order logic. It is known that for all $k \geq 3$, FO^k does not have an existential preservation theorem, so this settles the last open case, answering a question of Andréka, van Benthem, and Németi. In contrast, we prove that the homomorphism preservation theorem holds for FO^2.

Mathematics Subject Classification: 03C40, 03C13.

Keywords: Finite variable logic, Preservation theorems, Finite model theory.

1 Introduction

Classical preservation theorems establish connections between syntactic and semantic properties of first-order formulae. Typically they have the following form: a class $\mathcal{C}$ of structures, defined by a first-order sentence, is preserved under some specified algebraic operation if and only if it is definable by a first-order sentence of a certain syntactic form. A prominent example is the Łos-Tarski theorem, also known as the *existential preservation theorem*. It says that for every first-order sentence φ, the class of models of φ is preserved under extensions if and only if φ is equivalent to an existential sentence. Another example is the *homomorphism preservation theorem*: a first-order sentence φ defines a model class that is closed under homomorphisms if and only if φ is equivalent to a positive existential sentence. Further information on preservation theorems can be found in [4, 10].

More recently, a considerable amount of research has been devoted to preservation theorems for logics other than first-order logic and to preservation theorems on the class of all finite structures. It is known that most classical preservation theorems *fail* on the class of finite structures. In particular, Tait [17] proved that this is the case for the Łos-Tarski theorem. The corresponding question for the homomorphism preservation theorem remains open. For further information on preservation theorems on finite structures (and a collection of open problems), see [1].

Preservation theorems also play a role in the study of fragments of first-order logic, such as the finite variable fragments [2, 14, 16], the modal fragment [3, 15], and the guarded fragment [2]. If important preservation theorems, along with model theoretic principles like Craig interpolation and Beth definability, hold for a logical language, this indicates that the language possesses 'good' model theoretic properties and that there is a natural balance between its syntax and semantics.

In this paper, we determine the status of the existential preservation theorem and the homomorphism preservation theorem for FO^2, first-order logic with two variables. This logic has received considerable attention recently, for several reasons.

- The finite variable fragments FO^k of first-order logic play an important role in finite model theory (e.g., see [11]). They have a good balance between expressive power and amenability of model theoretic techniques, so that the study of the FO^k-theories of finite structures provides interesting insights into definability and complexity issues.
- FO^2 is the simplest non-trivial finite variable fragment of first-order logic, and has certain nice properties that are not shared by the fragments FO^k, for $k \geq 3$. In particular, FO^2 has the finite model property [12, 7], hence is decidable for satisfiability, and admits a positive solution of the canonization problem [13].
- Propositional modal logic can be embedded into FO^2 in a natural way (see [3]), and this has been offered as an explanation for the good properties of modal logics.

Andréka, van Benthem, and Németi [2] investigated the model theoretic properties of the finite variable fragments and proved, among other things, that the existential preservation theorem fails for all FO^k, with $k \geq 3$. The existential preservation is trivially true for FO^1, so FO^2 is the only remaining case, left as an open problem in [2]. We settle this case here. Our main results show that:

- The existential preservation theorem fails for FO^2.
- In contrast, the homomorphism preservation theorem is true for FO^2.

The failure of the existential preservation theorem for FO^2 further confirms the view that the embedding of propositional modal logic into FO^2 does not really explain its good algorithmic and model theoretic properties (e.g., see [2, 8]). Recall that the existential preservation theorem does hold for propositional modal logic (see [2, 15]).

A preservation theorem links a semantic property $\mathcal{P}$ of definable model classes to a syntactic property of formulae. Thus the failure of a preservation theorem raises the possibility of finding some alternative syntactic characterization. Alechina and Gurevich [1] formulate this question as follows. Suppose that a preservation theorem for a logic $\mathcal{L}$ and model classes with semantic property $\mathcal{P}$ fails. Is there a *recursive* class S of $\mathcal{L}$-sentences such that every sentence in S defines a model class with property $\mathcal{P}$ and, conversely, every $\mathcal{L}$-definable model class that has property $\mathcal{P}$ is in fact definable by a sentence from S? With respect to the failure of the existential preservation theorem for FO^2 we give a positive answer to the question, based on the fact that it has a decidable satisfiability problem.

Acknowledgments. We would like to thank Martin Otto for carefully reading this paper and an anonymous referee for helpful comments.

2 Definitions

FO denotes first-order logic. We restrict our attention to purely relational first-order formulae, that is, those without function symbols and constants, but including equality. Two-variable first-order logic, FO^2, is the class of first-order formulae that contain only the two variables x and y. The existential fragments of FO and FO^2, denoted $\mathrm{FO}(\exists)$ and $\mathrm{FO}^2(\exists)$, respectively, contain only those formulae that can be built from the atomic and the negated atomic formulae by means of $\wedge$, $\vee$, and $\exists$. The FO^2-type of an element a of a structure $\mathfrak{A}$ is the set of formulas $\psi(x) \in \mathrm{FO}^2$ such that $\mathfrak{A} \models \psi[a]$. The *quantifier rank* of a formula φ, $qr(\varphi)$, is the maximal nesting depth of quantifiers in φ. For a sentence φ, $\mathrm{Mod}(\varphi)$ is the class of models of φ, and $\mathrm{Mod}_f(\varphi)$ is the class of finite models of φ.

Given structures $\mathfrak{A}, \mathfrak{B}$, and $n \in \omega$, we write $\mathfrak{A} \equiv^{2,n} \mathfrak{B}$ iff for all sentences $\varphi \in \mathrm{FO}^2$ with $qr(\varphi) \leq n$, $\mathfrak{A} \models \varphi$ iff $\mathfrak{B} \models \varphi$. It is well-known that this relation can be characterized by a modified Ehrenfeucht-Fraïssé game (e.g., see [6, 13]). The *n-round two-pebble game* on $\mathfrak{A}$ and $\mathfrak{B}$ is a two player game, played with two pebbles, α_1, α_2, for structure $\mathfrak{A}$ and two pebbles, β_1, β_2, for structure $\mathfrak{B}$. In each round, Player I chooses one of the structures, and places one of its pebbles on an element of the structure. Player II responds by placing the corresponding pebble on an element of the other structure. Player II loses after a round iff the location of the pebbles does not determine a partial isomorphism between the structures. She wins the game just in case she does not lose after any of the n rounds.

Proposition 1. *For all $\mathfrak{A}, \mathfrak{B}$, and n, $\mathfrak{A} \equiv^{2,n} \mathfrak{B}$ iff Player II has a winning strategy in the n-round two-pebble game on $\mathfrak{A}$ and $\mathfrak{B}$.*

Likewise, we write $\mathfrak{A} \Rrightarrow^n_{\mathrm{FO}^2(\exists)} \mathfrak{B}$ iff for all $\varphi \in \mathrm{FO}^2(\exists)$ with $qr(\varphi) \leq n$, if $\mathfrak{A} \models \varphi$, then $\mathfrak{B} \models \varphi$; $\mathfrak{A} \Rrightarrow_{\mathrm{FO}^2(\exists)} \mathfrak{B}$ iff for all $n \in \omega$, $\mathfrak{A} \Rrightarrow^n_{\mathrm{FO}^2(\exists)} \mathfrak{B}$. The *existential n-round two-pebble game* from $\mathfrak{A}$ to $\mathfrak{B}$, is simply an n-round two-pebble game, in which Player I may only play on $\mathfrak{A}$.

Proposition 2. *For all $\mathfrak{A}, \mathfrak{B}$, and n, $\mathfrak{A} \Rrightarrow^n_{\mathrm{FO}^2(\exists)} \mathfrak{B}$ iff Player II has a winning strategy in the existential n-round two-pebble game from $\mathfrak{A}$ to $\mathfrak{B}$.*

3 Preservation under extensions

The main result in this section is the following.

Theorem 3. *There is a sentence $\varphi \in \mathrm{FO}^2$ such that $\mathrm{Mod}(\varphi)$ is closed under extensions and φ is not equivalent to any $\mathrm{FO}^2(\exists)$-sentence.*

Proof. The proof uses ideas from [16], where it is shown that for each $n \in \omega$, there is a sentence $\varphi_n \in \mathrm{FO}^3$ such that $\mathrm{Mod}(\varphi_n)$ is closed under extensions and φ_n is not equivalent to any existential FO^n-sentence. The structure of the argument is as follows.

1. We specify finite structures $\mathfrak{A}$, $\mathfrak{B}$, such that $\mathfrak{A} \Rrightarrow_{\mathrm{FO}^2(\exists)} \mathfrak{B}$, although $\mathfrak{A}$ cannot be embedded in $\mathfrak{B}$.

2. We define a class $\mathcal{D}$ of structures that is closed under substructures, $\mathfrak{B} \in \mathcal{D}, \mathfrak{A} \notin \mathcal{D}$, such that there is an FO^2-sentence θ with $\mathcal{D} = \mathrm{Mod}(\theta)$.

3. Then $\varphi := \neg\theta$ is the required sentence.

Let $\sigma = \{P, Q, S, T, E\}$, P, Q, S, T unary and E binary. We let $\mathfrak{A}, \mathfrak{B}$ be the σ-structures depicted in Figure 1, where the nodes labeled p, q, s, t are each the unique element in $\mathfrak{A}$, $\mathfrak{B}$,

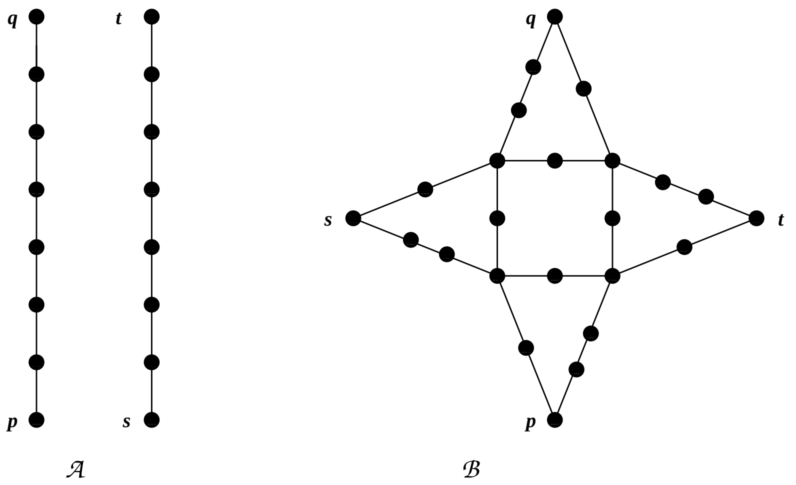

Figure 1: σ-structures

satisfying P, Q, S, T, respectively. Observe that $\mathfrak{B}$ does not contain disjoint paths from p to q and from s to t. In particular, $\mathfrak{A}$ cannot be embedded in $\mathfrak{B}$.

Since σ-structures are (colored) graphs, there is a natural notion of distance $d(a,b)$ between two points of a structure. If there is no path from a to b, define $d(a,b) = \omega$.

Lemma 4. $\mathfrak{A} \Rrightarrow_{\mathrm{FO}^2(\exists)} \mathfrak{B}$.

Proof. Let $\mathfrak{A}_1, \mathfrak{A}_2$ be the two connected components of $\mathfrak{A}$, each a chain of length 7. Observe that $\mathfrak{B}$ has two substructures isomorphic to $\mathfrak{A}_1$ and two substructures isomorphic to $\mathfrak{A}_2$.

We prove the lemma by showing that Player II has a winning strategy in the 'eternal' existential two-pebble game from $\mathfrak{A}$ to $\mathfrak{B}$. Without loss of generality, we can assume that Player I does not play the same pebble in two consecutive rounds. We argue by induction on the round number.

In round 1, Player I plays a pebble on some element a_1 of $\mathfrak{A}$, $a_1 \in \mathfrak{A}_i$. By the preceding remark, there is a substructure $\mathfrak{B}_1 \subseteq \mathfrak{B}$ and a partial isomorphism f_1 from $\mathfrak{A}$ to $\mathfrak{B}$ mapping $\mathfrak{A}_i$ onto $\mathfrak{B}_1$. Player II plays her pebble on the element $f_1(a_1)$. It is clear that a_1 and $f_1(a_1)$ realize the same atomic type, so that Player II does not lose at this stage.

Suppose now that Player II has not lost after n rounds, and that in round n, Player I placed a pebble on a_n in $\mathfrak{A}$ in component $\mathfrak{A}_j$, there is a partial isomorphism f_n from $\mathfrak{A}_j$ onto a substructure $\mathfrak{B}_n \subseteq \mathfrak{B}$, and Player II played on $f_n(a_n)$. We consider two cases.

One, in round $n+1$, Player I plays his other pebble on an element a_{n+1} also in $\mathfrak{A}_j$. Let $\mathfrak{B}_{n+1} := \mathfrak{B}_n$ and $f_{n+1} := f_n$. Player II responds by playing her pebble on $f_{n+1}(a_{n+1})$. It is clear that she thereby maintains a partial isomorphism between the pebbled elements.

Two, Player I plays his pebble on a_{n+1} in component $\mathfrak{A}_i, i \neq j$. Without loss, let $p, q \in \mathfrak{A}_i$. If Player I plays on p, resp. q, then it is clear that Player II maintains a partial isomorphism by playing on p, resp. q, in $\mathfrak{B}$. In particular, neither pair of pebbles is located on a pair of adjacent elements. (For the induction step, define $\mathfrak{B}_{n+1} \subseteq \mathfrak{B}$ to be one of the paths from p to q isomorphic to $\mathfrak{A}_{n+1}$.) Otherwise, suppose that $d(p, a_{n+1}) = l, 1 \leq l \leq 6$. Observe that there are two elements $b_1, b_2 \in \mathfrak{B}$ such that $d(p, b_i) = l, d(q, b_i) = 7 - l, i = 1, 2$. Furthermore, the pebbled element $f_n(a_n) \in \mathfrak{B}$ cannot be adjacent to both of them. Suppose it is not adjacent to b_1. Let $\mathfrak{B}_{n+1} \subseteq \mathfrak{B}$ be the unique substructure of $\mathfrak{B}$ containing b_1 and isomorphic to $\mathfrak{A}_i$, and let f_{n+1} be a partial isomorphism from $\mathfrak{A}$ to $\mathfrak{B}$ that maps $\mathfrak{A}_i$ onto $\mathfrak{B}_{n+1}$. Player II now plays her pebble on $f_{n+1}(a_{n+1})$. Neither $a_n, a_{n+1} \in \mathfrak{A}$ nor $f_{n+1}(a_n), f_{n+1}(a_{n+1}) \in \mathfrak{B}$ are adjacent, and Player II has again maintained a partial isomorphism. Therefore Player II does indeed have a winning strategy. ■

Lemma 5. *There is a class $\mathcal{D}$ of structures, closed under substructures, such that $\mathfrak{B} \in \mathcal{D}, \mathfrak{A} \notin \mathcal{D}$, and $\mathcal{D}$ is defined by an* FO^2*-sentence.*

Proof. First we describe the class $\mathcal{D}$; then we prove that it is defined by an FO^2-sentence.

Say that a σ-structure $\mathfrak{C}$ is *united* if

1. each unary relation P, Q, S, T is realized by exactly one element c_1, c_2, c_3, c_4, respectively;
2. for each $i, j \leq 4$, there is a path from c_i to c_j of length ≤ 9;
3. there is a path of length 7 from c_1 to c_2, resp. c_3 to c_4.

In this case, there is a single graph component containing each of the c_i, the *united component*. Observe that every substructure of $\mathfrak{B}$ with a single component containing p, q, s, and t is united. We are now ready to define the class $\mathcal{D}$.

Let $\psi = (\forall x \neg Px) \vee (\forall x \neg Qx) \vee (\forall x \neg Sx) \vee (\forall x \neg Tx)$. Let $\mathcal{E}_1$ be the set of all σ-structures $\mathfrak{C}$ such that each unary relation P, Q, S, T is realized by exactly one element c_1, c_2, c_3, c_4 in C, respectively, and $d(c_1, c_2) > 7$ or $d(c_3, c_4) > 7$. Let $\mathcal{E}_2$ be the set of all σ-structures $\mathfrak{C}$ such that $\mathfrak{C}$ is united and its united component is isomorphic to a substructure of $\mathfrak{B}$. Finally let $\mathcal{D} := \mathrm{Mod}(\psi) \cup \mathcal{E}_1 \cup \mathcal{E}_2$.

It is easy to see that $\mathfrak{A} \notin \mathcal{D}$: it is not united, each relation P, Q, S, T is realized by exactly one element a_1, a_2, a_3, a_4 in A, respectively, and $d(a_1, a_2) = d(a_3, a_4) = 7$. We claim that $\mathcal{D}$ is closed under substructures. If $\mathfrak{C} \in \mathcal{D}$ satisfies ψ, then so does every substructure of $\mathfrak{C}$, as ψ is universal. Suppose $\mathfrak{C} \in \mathcal{E}_1$, and $\mathfrak{C}' \subseteq \mathfrak{C}$. It is easy to verify that either $\mathfrak{C}' \models \psi$ or $\mathfrak{C}' \in \mathcal{E}_1$.

Finally, consider the case where $\mathfrak{C} \in \mathcal{E}_2$ and $\mathfrak{C}' \subseteq \mathfrak{C}$. If $\mathfrak{C}'$ is again united, then its united component is again isomorphic to a substructure of $\mathfrak{B}$, so $\mathfrak{C}'$ is also in $\mathcal{E}_2$. If one of the unary relations in σ is empty in $\mathfrak{C}'$, then $\mathfrak{C}' \models \psi$. So it only remains to consider the case where $\mathfrak{C}'$ is not united and contains exactly one realization of each of the relations P, Q, S, T, which we again call c_1, c_2, c_3, c_4, respectively. These cannot all be in the same component.

Suppose that in $\mathfrak{C}'$ there are paths connecting c_1 to c_2 and c_3 to c_4. These paths must be disjoint, since not all the c_i's are in the same component. Thus the united component of $\mathfrak{C}$ must contain a path from the (unique) P-element to the Q-element, and a path from the S-element to the T-element, both paths disjoint. But $\mathfrak{B}$ does not contain such disjoint paths, so $\mathfrak{C}$ cannot either. Therefore, in $\mathfrak{C}'$ either $d(c_1, c_2) = \omega$ or $d(c_3, c_4) = \omega$. Hence $\mathfrak{C}' \in \mathcal{E}_1$. This completes the claim that $\mathcal{D}$ is closed under substructures.

It only remains to prove that $\mathcal{D}$ is FO^2-definable. It is easy to see that there is an FO^2-sentence ψ_1 that defines $\mathcal{E}_1$ (e.g., using the δ-formulae defined below). Thus, we want to show that $\mathcal{E}_2$ is also defined by some sentence $\psi_2 \in \mathrm{FO}^2$. The desired definition of $\mathcal{D}$ is then $\psi \vee \psi_1 \vee \psi_2$.

Let $\{\mathfrak{C}_1, \dots, \mathfrak{C}_t\}$ be the set of substructures of $\mathfrak{B}$ that are connected and united. For each $s \leq t$, let $\mathcal{C}_s \subseteq \mathcal{E}_2$ be the class of united structures with united component isomorphic to $\mathfrak{C}_s$. Clearly $\mathcal{E}_2 = \bigcup_s \mathcal{C}_s$, so it suffices to show that each $\mathcal{C}_s$ is defined by a sentence $\gamma_s \in \mathrm{FO}^2$; then let $\psi_2 = \bigvee_s \gamma_s$.

Recall that a structure $\mathfrak{C}$ is *2-rigid* [5] if no two elements in $\mathfrak{C}$ realize the same FO^2-type. It is known that every finite 2-rigid structure is characterized up to isomorphism by a single FO^2-sentence. Below, we observe that each $\mathfrak{C}_s, s \leq t$, is in fact 2-rigid. From there, it is not hard to show that each $\mathcal{C}_s$ is defined by a single FO^2-sentence.

It will be helpful to introduce a number of defined formulae with one free variable. Each of these will say of a given element x that there is a path of some specified length, beginning at x, whose last element satisfies some quantifier free formula. For $n \in \omega$, we want $\delta_P^n(x)$ to say that there is a path of length at most n, whose first element is x and whose last element is in P. It is easy to show, by induction on n, that this property is expressible in FO^2. Let $\delta_P^0(x) := Px$ and

$$\delta_P^1(x) := Px \vee \exists y(Exy \wedge Py).$$

For $n \geq 1$, let

$$\delta_P^{n+1}(x) := \delta_P^n(x) \vee \exists y\big(Exy \wedge \exists x(x = y \wedge \delta_P^n(x))\big).$$

The corresponding formulae for the unary relations Q, S, T are defined in an analagous manner. Finally, for each $n \in \omega$, let

$$\delta^n(x) := \delta_P^n(x) \vee \delta_Q^n(x) \vee \delta_S^n(x) \vee \delta_T^n(x).$$

By a *δ-formula*, we mean any one of the formulae defined above.

We now observe that each structure $\mathfrak{C}_s, s \leq t$, is 2-rigid. Fix some $s \leq t$, and let $\{c_1, \dots, c_k\}$ be the universe of $\mathfrak{C}_s$. We claim that any two distinct elements c_i, c_j of $\mathfrak{C}_s$ are distinguished by some δ-formula. Suppose that c_i and c_j are equidistant both from the unique P-element and from the unique Q-element in $\mathfrak{C}_s$. (Otherwise, they are distinguished by a formula of the form δ_P^n or δ_Q^n.) Then it is easy to check they cannot be equidistant from the T-element. Thus, for each $c_i \in \mathfrak{C}_s$ there is a formula $\eta_i(x)$, a Boolean combination of δ-formulae, such that c_i is the only element in $\mathfrak{C}_s$ satisfying $\eta_i(x)$. Thus $\mathfrak{C}_s$ is 2-rigid.

We now show how to define the class $\mathcal{C}_s$. For each pair $i, j \leq k$, let $\zeta_{i,j}(x, y)$ be the complete atomic type of (c_i, c_j). Finally, let

$$\begin{aligned}\gamma_s := &\forall x(\delta^4(x) \to \delta^3(x)) \wedge \forall x(\delta^3(x) \to \bigvee_{i \leq k} \eta_i(x)) \wedge \\ &\bigwedge_i \forall x \forall y(\eta_i(x) \wedge \eta_i(y) \to x = y) \wedge \bigwedge_{i,j \leq k} \exists x \exists y(\eta_i(x) \wedge \eta_j(y) \wedge \zeta_{i,j}(x, y)).\end{aligned}$$

We now argue that γ_s does indeed define $\mathcal{C}_s$. Observe that for all $\mathfrak{C} \in \mathcal{C}_s, \mathfrak{C} \models \gamma_s$. In the other direction, the first conjunct implies, for united structures, that any element in the united component is of distance at most 3 from a P, Q, S or T-element. If some element were of distance $n > 4$, then there would also have to be some element of distance 4, but the first conjunct implies that such an element must be of distance ≤ 3. Together, the next two conjuncts then imply that the size of the united component is at most k, as every element in it must realize some formula $\eta_i(x), i \leq k$, and there is at most one such realization. Finally, the last conjunct says that each formula $\eta_i(x), i \leq k$, is satisfied by some (unique) element; for each pair $i, j \leq k$, it also determines the atomic 2-type of the pair of elements satisfying η_i and η_j. This completely specifies the united component. In particular, for every $\mathfrak{C}$, if $\mathfrak{C} \models \gamma_s$, then $\mathfrak{C} \in \mathcal{C}_s$. ∎

This also completes the proof of the theorem. ∎

3.1 Some positive results

We introduce some more notation. Given a τ-structure $\mathfrak{A}$ and a unary relation $P \in \tau$, $\mathfrak{A}|P$ denotes the substructure of $\mathfrak{A}$ with universe $P^{\mathfrak{A}}$. Let σ be a signature and P a unary relation, $P \notin \sigma$. For each σ-formula θ, there exists a $\sigma \cup \{P\}$-formula θ^P, the *relativization* of θ to P (with the same number of variables), such that for all structures $\mathfrak{A}$, $\mathfrak{A} \models \theta^P$ iff $(\mathfrak{A}|P) \models \theta$.

Proposition 6. *There is an effective procedure that determines whether, given an arbitrary sentence $\varphi \in \mathrm{FO}^2$, the class $\mathrm{Mod}(\varphi)$ is closed under extensions.*

Proof. Let φ be an FO^2-sentence, with signature σ. Choose some unary relation $P \notin \sigma$, and let $\varphi^* = \varphi^P \rightarrow \varphi$. Clearly, $\mathrm{Mod}(\varphi)$ is closed under extensions iff φ^* is valid. By the decidability of FO^2 [12], it is decidable whether φ^* is valid. ∎

The next proposition is a strengthening of the previous one. The proof involves a standard application of the completeness theorem.

Proposition 7. *There is a recursive function $f : \mathrm{FO}^2 \rightarrow \mathrm{FO}(\exists) \cup \{\star\}$, such that for all $\varphi \in \mathrm{FO}^2$, if $\mathrm{Mod}(\varphi)$ is not closed under extensions, then $f(\varphi) = \star$, and if $\mathrm{Mod}(\varphi)$ is closed under extensions, then $\varphi \equiv f(\varphi)$.*

Observe that we get no bound on the complexity of f. It would be interesting to find a more effective way to compute it.

Finally, we briefly consider FO^2-definability over the class of finite structures. The next proposition follows from the proof of Proposition 6 and the finite model property for FO^2.

Proposition 8. *For all $\varphi \in \mathrm{FO}^2$, $\mathrm{Mod}_f(\varphi)$ is closed under extensions iff $\mathrm{Mod}(\varphi)$ is closed under extensions. Therefore, by the existential preservation theorem for first-order logic, $\mathrm{Mod}_f(\varphi)$ is closed under extensions iff φ is equivalent to an $\mathrm{FO}(\exists)$-sentence.*

In contrast, it is known [9] that there is a sentence $\theta \in \mathrm{FO}^3$ such that $\mathrm{Mod}_f(\theta)$ is closed under extensions and θ is not equivalent to any $\mathrm{FO}(\exists)$-sentence.

4 Preservation under homomorphisms

Given σ-structures $\mathfrak{A}$ and $\mathfrak{B}$, a function $f : A \rightarrow B$ is a *homomorphism* from $\mathfrak{A}$ to $\mathfrak{B}$ iff for each k-ary relation $R \in \sigma$ and each k-tuple $(a_1, \ldots, a_k)$ of $\mathfrak{A}$, if $\mathfrak{A} \models Ra_1 \cdots a_k$, then $\mathfrak{B} \models Rfa_1 \cdots fa_k$. A class $\mathcal{C}$ of structures is *closed under homomorphisms* iff for all structures $\mathfrak{A}$ and $\mathfrak{B}$, if $\mathfrak{A} \in \mathcal{C}$ and there is a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$, then also

$\mathfrak{B} \in \mathcal{C}$. Recall that the homomorphism preservation theorem for first-order logic states that for each sentence $\varphi \in \mathrm{FO}$, $\mathrm{Mod}(\varphi)$ is closed under homomorphisms iff φ is equivalent to a positive existential FO-sentence. Below, we show that the analagous result holds for FO^2. (One direction is trivial.) It is an open question whether the homomorphism preservation theorem holds for any $\mathrm{FO}^k, k \geq 3$.

Let $\mathrm{FO}^2(\exists, +)$ denote the set of positive sentences in $\mathrm{FO}^2(\exists)$, that is, those that contain no negations. The *positive atomic type* of a pair of elements (a, a') of a structure $\mathfrak{A}$ is the conjunction of all atomic formulas $\psi(x, y)$ such that $\mathfrak{A} \models \psi[a, a']$. Clearly, (a, a') in $\mathfrak{A}$ and (b, b') in $\mathfrak{B}$ have the same positive atomic type iff they have the same atomic type.

As FO^2 contains only two variables, we assume, without loss of generality, that signatures contain no relations of arity ≥ 3. Distinct elements a and a' in a σ-structure $\mathfrak{A}$ are *adjacent* iff there is a binary relation $R \in \sigma$ such that $\mathfrak{A} \models Raa' \vee Ra'a$. Let $N(a)$, the *neighbours* of a, denote the set of elements adjacent to a.

Theorem 9. *The homomorphism preservation theorem holds for* FO^2*. In particular, for all* $\varphi \in \mathrm{FO}^2$*, if* $\mathrm{Mod}(\varphi)$ *defines a class that is closed under homomorphisms, then there is a* $\theta \in \mathrm{FO}^2(\exists, +)$*, with* $qr(\varphi) = qr(\theta)$*, such that* $\varphi \equiv \theta$*.*

Proof. Suppose that $\varphi \in \mathrm{FO}^2$ defines a class that is closed under homomorphisms, and σ is the set of relation symbols occuring in φ. Let $\mathcal{C} = \mathrm{Mod}(\varphi)$ and let $n = qr(\varphi)$. The argument proceeds as follows. For each model $\mathfrak{A} \in \mathcal{C}$, we define a sentence $\theta_{\mathfrak{A}} \in \mathrm{FO}^2(\exists, +)$, with $qr(\theta_{\mathfrak{A}}) \leq n$, such that $\mathfrak{A} \models \theta_{\mathfrak{A}}$ and $\mathrm{Mod}(\theta_{\mathfrak{A}}) \subseteq \mathcal{C}$. It is clear that, although $\mathcal{C}$ is infinite, there are, up to equivalence, only finitely many $\theta_{\mathfrak{A}}$ (because for any finite relational signature, up to equivalence there are only finitely many formulae of quantifer rank $\leq n$). Letting $\theta := \bigvee_{\mathfrak{A} \in \mathcal{C}} \theta_{\mathfrak{A}}$, it is immediate that $\mathrm{Mod}(\varphi) = \mathrm{Mod}(\theta)$, and hence $\varphi \equiv \theta$. The main part of the argument involves showing that, for each of the sentences $\theta_{\mathfrak{A}}, \mathfrak{A} \in \mathfrak{C}$, $\mathrm{Mod}(\theta_{\mathfrak{A}}) \subseteq \mathcal{C}$.

Defining the sentence $\theta_{\mathfrak{A}}$. For each structure $\mathfrak{A}$ and elements (a, b) of $\mathfrak{A}$, let $\rho_{a,b}(x, y)$ be the positive atomic type of (a, b) in $\mathfrak{A}$. For all elements a of $\mathfrak{A}$ and all $m \leq n$, we define a formula $\pi_a^m(x) \in \mathrm{FO}^2(\exists, +)$, with $qr(\pi_a^m(x)) \leq m$, such that $\mathfrak{A} \models \pi_a^m[a]$. $\pi_a^0(x)$ is just the positive atomic type of a in $\mathfrak{A}$. For all $m + 1$, we essentially want $\pi_a^{m+1}(x)$ to be

$$\pi_a^m(x) \wedge \bigwedge_{b \in N(a)} \exists y (\rho_{a,b}(x, y) \wedge \pi_a^m(y)),$$

except that we eliminate redundant, identical conjuncts. (Here, $\pi_a^m(y)$ denotes the formula obtained from $\pi_a^m(x)$ by exchanging all occurences of x and y.) This guarantees that, for fixed m, there are only finitely many formulae of the form $\pi_a^m(x)$. Finally, let $\theta_{\mathfrak{A}} := \bigwedge_{a \in A} \exists x \pi_a^{n-1}(x)$, again eliminating redundant conjuncts.

Proving $\mathbf{Mod}(\theta_{\mathfrak{A}}) \subseteq \mathcal{C}$. To show that $\mathrm{Mod}(\theta_{\mathfrak{A}}) \subseteq \mathcal{C}$ (or, equivalently, that $\theta_{\mathfrak{A}}$ implies φ), we define a structure $\mathfrak{C}$ such that:

1. $\mathfrak{C} \models \theta_{\mathfrak{A}}$;

2. $\mathfrak{C} \in \mathcal{C}$;

3. for all $\mathfrak{B}$ such that $\mathfrak{B} \models \theta_{\mathfrak{A}}$, there is a homomorphism from $\mathfrak{C}$ to $\mathfrak{B}$.

Since $\mathcal{C}$ is closed under homomorphisms, these conditions imply that every model of $\theta_{\mathfrak{A}}$ is in $\mathcal{C}$, as desired.

We now define $\mathfrak{C}$. Given $\theta_{\mathfrak{A}}$, let $Q = \{q_1, \ldots, q_t\}$ be the set of occurences of (existential) quantifiers in $\theta_{\mathfrak{A}}$. The universe of $\mathfrak{C}$ is Q. The interpretation of the relations $R \in \sigma$ on $\mathfrak{C}$ is determined straightforwardly from $\theta_{\mathfrak{A}}$, as follows. For each binary $R \in \sigma$, $\mathfrak{C} \models Rq_iq_j$ iff there is an occurence of an atomic formula, Rvw, such that at this place v and w are bound by q_i and q_j, respectively. Likewise for unary predicates. Every formula occcurs only positively, so $\mathfrak{C}$ is well-defined. It is easy to see that $\mathfrak{C}$ satisfies conditions 1 and 3. Indeed for all $\mathfrak{B}$, an 'assignment of variables', mapping each occurence of a variable following an existential quantifier to an element, so as to verify that $B \models \theta_{\mathfrak{A}}$, determines a homomorphism from $\mathfrak{C}$ to $\mathfrak{B}$.

All that remains is to prove that $\mathfrak{C} \in \mathcal{C}$. It suffices to show that there are $\mathfrak{A}'$ and $\mathfrak{D}$ such that (i) $\mathfrak{A} \subseteq \mathfrak{A}'$; (ii) $\mathfrak{A}' \equiv^{2,n} \mathfrak{D}$; and (iii) there is a homomorphism from $\mathfrak{D}$ to $\mathfrak{C}$. Since $\mathcal{C}$ is closed under homomorphisms and $\mathrm{FO}^{2,n}$-equivalence, $(i) - (iii)$ imply that $\mathfrak{C} \in \mathcal{C}$. Let $\mathfrak{D} := \mathfrak{C} \oplus \mathfrak{C}$ and $\mathfrak{A}' := \mathfrak{A} \oplus \mathfrak{D}$. ($\mathfrak{A} \oplus \mathfrak{B}$ denotes the structure which is the *disjoint* union of $\mathfrak{A}$ and $\mathfrak{B}$.) It is immediate that (i) and (iii) are satisfied.

We define the following supplementary relation S on $\mathfrak{C}$, and hence also on $\mathfrak{D}$. For all q_i, q_j in Q, Sq_iq_j iff q_j occurs in the scope of q_i and there is an occurence of a binary atomic formula in $\theta_{\mathfrak{A}}$ such that each of q_i and q_j bind a variable in this occurence of the formula. Observe that q_i and q_j are adjacent in $\mathfrak{C}$ iff they are adjacent in the directed graph (Q, S). It is easy to verify that (Q, S) is a directed forest, that is, the disjoint union of directed trees. Since (Q, S) is a forest, there is a well-defined function, $\nu(x)$ such that $\nu(q_i)$ is the height of q_i in (Q, S).

Proving $\mathfrak{A}' \equiv^{2,n} \mathfrak{D}$. To establish that $\mathfrak{A}' \equiv^{2,n} \mathfrak{D}$, we describe a winning strategy for Player II in the n-round 2-pebble game on $\mathfrak{A}'$ and $\mathfrak{D}$. The following claim simplifies matters.

Claim. *If Player I can win the game, then he can do so playing in accordance with the following 'normal form'.*

1. *In each round $m+1$, he plays the pebble pair that was not played in the previous round, and does not replay it on the same element it just occupied.*
2. *In round 1, he plays a pebble α_i on the $\mathfrak{A}$-component of $\mathfrak{A}'$.*
3. *In each round $m + 1$, he plays a pebble α_i, resp. β_i, on an element adjacent to α_{1-i}, resp. β_{1-i}.*

Condition 1 is obvious. To see that Condition 2 does not hinder Player I, suppose that he does not play his first move on the $\mathfrak{A}$-component of $\mathfrak{A}'$. Player II will then play all of her moves according to the bijection between $\mathfrak{D}$ and the $\mathfrak{D}$-component of $\mathfrak{A}'$, until Player I plays on the $\mathfrak{A}$-component of $\mathfrak{A}'$. Therefore, in order to win Player I must eventually play a pebble α_i on this component of $\mathfrak{A}'$. Player II can then play β_i on the vacant $\mathfrak{C}$-component of $\mathfrak{D}$. At the start of the next round, pebbles α_j and $\beta_j, i \neq j$, are removed from the board. Thus Player I could have reached the same position by playing on $\mathfrak{A} \subseteq \mathfrak{A}'$ in round 1.

Consider Condition 3. Suppose that in some round $m+1$, Player I plays β_i on an element of $\mathfrak{D}$ not adjacent to β_{1-i}. Player II can then play α_i on the corresponding element of a vacant $\mathfrak{C}$-component in $\mathfrak{A}'$. Since the pebbles $\alpha_j, \beta_j, i \neq j$ will be replayed in the next round, for the same reasons as above Player I has not made any progress. Likewise, if Player I plays on a non-adjacent element of $\mathfrak{A}'$. This completes the proof of the claim.

We now describe the winning strategy of Player II, assuming that Player I plays in accord with the above conditions. In round 1, Player I plays on some a_1 in $\mathfrak{A} \subseteq \mathfrak{A}'$. Player II then plays on one of the two copies of the element q_k (of either $\mathfrak{C}$-component) such that q_k occurs in the formula $\theta_{\mathfrak{A}}$ as the quantifier that binds the variable x in the formula $\pi_a^{n-1}(x)$. Let $b_1 := q_k$. By the definition of $\mathfrak{C}$, it is clear that a_1 and b_1 realize the same atomic type, so Player II maintains a partial isomorphism.

We now describe Player II's strategy in subsequent rounds. For each $m \leq n$, we let a_m, resp. b_m, denote the element in $\mathfrak{A}'$, resp. $\mathfrak{D}$, on which a pebble was placed in round m. Observe that b_1 is the root of an S-tree. In each later round, Player I, and hence Player II, always plays so that the pebbles on each structure are on adjacent elements.

We claim that Player II wins by maintaining the condition that, in each round m, the pebbles are played so that $\nu(b_m) = m - 1$ and b_m is the element q_l that binds the formula $\rho_{a_{m-1},a_m}(x,y) \wedge \pi_{a_m}^{n-m}(y)$. Again, it is easy to check from the definition of $\mathfrak{C}$ that by playing in accordance with this condition Player II always maintains a partial isomorphism. In particular, the occurence of the formula $\rho_{a_{m-1},a_m}(x,y)$ ensures that (a_{m-1}, a_m) and (b_{m-1}, b_m) realize the same atomic type. Observe that during the game the β pebbles will 'climb up the S-tree', so that the β pebble played in round $m+1$ will always be on the child of the β pebble played in the previous round. (Player II is prohibited from ever playing, in round $m+1$, a β pebble on the parent of b_m, since to do so would be to replay a pebble on the same element, in violation of Condition 1 of the normal form.)

All that remains is to prove this last claim. We argue by induction on the round number. Round 1 has already been taken care of. There are two cases, depending on which structure Player I selects. One, in round $m+1$, Player I chooses $\mathfrak{A}'$, and plays on $a_{m+1} \in N(a_m)$ in $\mathfrak{A}'$. Player II can then play her pebble on the quantifier occurence that binds the formula $\rho_{a_m,a_{m+1}}(x,y) \wedge \pi_{a_{m+1}}^{n-(m+1)}(y)$, which is a conjunct of $\pi_{a_m}^{n-m}(y)$.

Two, Player I chooses $\mathfrak{D}$, and plays β_j on some element b_{m+1}, a child of b_m. By the induction hypothesis, b_m is the element q_i in $\mathfrak{D}$ such that q_i is the quantifier that binds the formula $\rho_{a_{m-1},a_m}(x,y) \wedge \pi_{a_m}^{n-m}(y)$. As Player I has played on a child of b_m, he must have played on some element q_j which binds the formula $\rho_{a_m,a}(x,y) \wedge \pi_a^{n-(m+1)}(y)$, for some $a \in \mathfrak{A}$, adjacent to a_m in $\mathfrak{A}'$. Thus, Player II can play her pebble on this element a.

This now establishes that $\mathfrak{A}' \equiv^{2,n} \mathfrak{D}$.

Thus $\theta_{\mathfrak{A}}$ implies φ, and $\theta := \bigvee_{\mathfrak{A} \in \mathcal{C}} \theta_{\mathfrak{A}}$ is equivalent to φ. This completes the proof of the theorem. ■

Observe that the same argument also works over the class of finite stuctures.

5 Some open problems

We mention a number of questions that are suggested by our results. These fall into two natural categories. In the introduction, we mentioned that many problems connected with FO^2 are recursive and even computationally easy. In this spirit, one can ask about the computational complexity of some problems arising in our work; it seems reasonable to hope for positive solutions. For example:

- Is there a good complexity bound on the function in Proposition 7?
- Suppose that φ defines a class closed under extensions. Can one compute the minimal n such that φ is equivalent to an $\mathrm{FO}^n(\exists)$-sentence? (Proposition 7 implies that one can compute an upper bound.)

The latter problem is a variant on a question of Andréka, van Benthem, and Németi [2].

Secondly, there are still a number of open questions about the status of various preservation theorems for FO^2 and also for the logics FO^k, $k \geq 3$. We mention just two.

- Does the analog of Lyndon's lemma hold for FO^2? That is, is each sentence $\varphi \in \mathrm{FO}^2$ such that $\mathrm{Mod}(\varphi)$ is monotone in every relation equivalent to a positive FO^2-sentence?
- Does the homomorphism preservation theorem hold for FO^k, $k \geq 3$?

References

[1] N. Alechina and Y. Gurevich. Syntax vs. semantics on finite structures. In J. Mycielski, G. Rozenberg, and A. Salomaa, editors, *Structures in Logic and Computer Science. A Selection of Essays in Honor of A. Ehrenfeucht*, pages 14–33. Springer LNCS 1261, 1997.

[2] H. Andréka, J. van Benthem, and I. Németi. Modal languages and bounded fragments of predicate logic. ILLC Research Report ML-96-03, 1996.

[3] J. van Benthem. *Modal Logic and Classical Logic.* Bibliopolis, 1983.

[4] C. C. Chang and H. J. Keisler. *Model Theory.* North-Holland, 1990.

[5] A. Dawar, S. Lindell, and S. Weinstein. Infinitary logic and inductive definability over finite structures. *Information and Computation*, 119:160–175, 1994.

[6] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory.* Springer, 1995.

[7] E. Grädel, Ph. Kolaitis, and M. Vardi. On the decision problem for two-variable first-order logic. *The Bulletin of Symbolic Logic*, 3:53–69, 1997.

[8] E. Grädel and M. Otto. On logics with two variables. *Theoretical Computer Science*, 1998. To appear.

[9] Y. Gurevich. Toward logic tailored for computational complexity. In *Computation and Proof Theory*, pages 175–216. Springer Lecture Notes in Mathematics 1104, 1984.

[10] W. Hodges. *Model Theory.* Cambridge University Press, 1993.

[11] I. Hodkinson. Finite variable logics. *Bulletin of the European Association for Theoretical Computer Science*, 51:111–140, October 1993.

[12] M. Mortimer. On languages with two variables. *Z. Math. Logik und Grundlagen Math.*, 21:135–140, 1975.

[13] M. Otto. *Bounded Variable Logics and Counting.* Springer, 1997.

[14] E. Rosen. *Finite model theory and finite variable logic.* PhD thesis, University of Pennsylvania, 1995.

[15] E. Rosen. Modal logic over finite structures. *Journal of Logic, Language and Information*, 6:427–439, 1997.

[16] E. Rosen and S. Weinstein. Preservation theorems in finite model theory. In D. Leivant, editor, *Logic and Computational Complexity*, pages 480–502. Springer LNCS 960, 1995.

[17] W. W. Tait. A counterexample to a conjecture of Scott and Suppes. *The Journal of Symbolic Logic*, 24:15–16, 1959.