

Description Complexity of Unary Structures in First-Order Logic with Links to Entropy

Reijo Jaakkola   

Mathematics Research Centre, Tampere University, Finland

Antti Kuusisto   

Mathematics Research Centre, Tampere University, Finland

Miikka Vilander  

Mathematics Research Centre, Tampere University, Finland

Abstract

The description complexity of a model is the length of the shortest formula that defines the model. We study the description complexity of unary structures in first-order logic FO, also drawing links to semantic complexity in the form of entropy. The class of unary structures provides, e.g., a simple way to represent tabular Boolean data sets as relational structures. We define structures with FO-formulas that are strictly linear in the size of the model as opposed to using the naive quadratic ones, and we use arguments based on formula size games to obtain related lower bounds for description complexity. For a typical structure the upper and lower bounds in fact match up to a sublinear term, leading to a precise asymptotic result on the expected description complexity of a randomly selected structure. We then give bounds on the relationship between Shannon entropy and description complexity. We extend this relationship also to Boltzmann entropy by establishing an asymptotic match between the two entropies. Despite the simplicity of unary structures, our arguments require the use of formula size games, Stirling's approximation and Chernoff bounds.

2012 ACM Subject Classification Theory of computation → Finite Model Theory; Mathematics of computing → Information theory

Keywords and phrases formula size, finite model theory, formula size games, entropy, randomness

Digital Object Identifier 10.4230/LIPIcs.CSL.2025.17

Related Version *Full Version*: <https://arxiv.org/abs/2406.02108>

Funding Antti Kuusisto and Miikka Vilander were supported by the Academy of Finland projects *Explaining AI via Logic* (XAILOG), grant number 345612 and *Theory of computational logics*, grant numbers 352419, 352420, 353027, 324435 and 328987.

1 Introduction

This paper investigates the resources needed to define finite models with a unary relational vocabulary. While unary models are very simple, it turns out that proving limits on the formula sizes for defining them is non-trivial. Furthermore, unary models are important as they give a direct relational representation of Boolean data sets, consisting simply of data points and their properties – thereby providing one of the simplest data representation schemes available. In practice all tabular data can be discretized and modeled via a Boolean data set. This relates to applications in, e.g., explainability and compression.

Given a logic \mathcal{L} and a class \mathcal{M} of models, the *description complexity* $C(\mathfrak{M})$ of a model \mathfrak{M} is the minimum length of a formula $\varphi \in \mathcal{L}$ that defines \mathfrak{M} with respect to \mathcal{M} . In the main scenario of this paper, \mathcal{M} is the class of models with the same domain of a finite size n and with the same unary vocabulary τ . We mostly study the setting via first-order logic FO. However, as description complexity links to the themes of *compressibility* and *compression*,



© Reijo Jaakkola, Antti Kuusisto, and Miikka Vilander;
licensed under Creative Commons License CC-BY 4.0

33rd EACSL Annual Conference on Computer Science Logic (CSL 2025).

Editors: Jörg Endrullis and Sylvain Schmitz; Article No. 17; pp. 17:1–17:20

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

we also investigate the restricted languages FO_d where the quantifier rank of every formula is limited to a positive integer d . This will lead to dramatically shorter description lengths (cf. Section 3) via a natural lossy compression phenomenon.

We also investigate how the *Shannon entropies* of unary structures are linked to their description complexities, the general trend being that higher entropy relates to higher description complexity. Shannon entropy is a well-known measure of intrinsic complexity, or randomness, from information theory. The Shannon entropy of a probability distribution $\mathbb{P} : X \rightarrow [0, 1]$ over a finite set X is given by $-\sum_{x \in X} \mathbb{P}(x) \log_2 \mathbb{P}(x)$. A relational structure \mathfrak{M} of size n over a unary vocabulary τ naturally defines a probability distribution over its domain. Indeed, let T be the set of unary quantifier-free *types* over τ , i.e., subsets of τ . A point a of a model \mathfrak{M} *realizes* a type $\pi \subseteq \tau$ if π is the set of relation symbols corresponding to exactly those unary relations that contain the point a . Now a τ -model \mathfrak{M} of size n naturally defines the probability distribution $\mathbb{P} : T \rightarrow [0, 1]$ such that $\mathbb{P}(\pi) = \frac{|\pi|}{n}$, where $|\pi|$ is the number of points of \mathfrak{M} realizing the type π . The Shannon entropy of \mathfrak{M} is then naturally defined to be equal to the Shannon entropy of the distribution $\mathbb{P} : T \rightarrow [0, 1]$.

While the Shannon entropy of \mathfrak{M} gives an intrinsic measure of complexity (or randomness) of \mathfrak{M} , another entropy measure may perhaps be easier to grasp intuitively. *Boltzmann entropy* has its origins in statistical mechanics, and it was originally defined as $k \ln \Omega$, where k is the Boltzmann constant and Ω the number of *microstates* of a system. In our setting, we follow [14] and define Boltzmann entropy of a *model class* \mathcal{A} as $\log_2 |\mathcal{A}|$, thus dropping the Boltzmann constant k , using binary logarithms and associating models with microstates. Now, it is natural to then define the Boltzmann entropy of a model \mathfrak{M} as $\log_2 |\mathcal{M}|$, where \mathcal{M} is the isomorphism class of \mathfrak{M} (recall here that in our setting, all models have the same domain of size n , so \mathcal{M} is finite). The reason why the Boltzmann entropy of \mathfrak{M} is a reasonable measure of intrinsic complexity of \mathfrak{M} is now easy to motivate. Firstly, consider a τ -model \mathfrak{M}_0 of size n where each $P \in \tau$ is interpreted as the empty relation. This is a very simple model whose isomorphism class has size 1 and the Boltzmann entropy of \mathfrak{M}_0 is thus very low: $\log_2 1 = 0$. On the other hand, models with the predicates in τ distributed in more disordered ways have larger isomorphism classes and thus greater Boltzmann entropies.

1.1 Contributions

Concerning upper bounds on description complexity, we show how to define unary structures via FO-formulas that are linear in model size. This contrasts the standard quadratic formulas that use equalities for counting cardinalities in a naive way. We also give analogous formulas for FO_d with quantifier rank at most d . Concerning lower bounds, we use formula size games to provide bounds with a worst case gap of a constant factor of 2 in relation to the upper bounds. This is done both for full FO and FO_d .

For a random structure the upper and lower bounds in fact match up to a sublinear additive term. Using this, we show that – asymptotically – the expected description complexity of a random unary structure of size n and over the vocabulary τ is exactly $3n/2^{|\tau|}$.

We then turn our attention to entropy. We show a close relationship between the Shannon entropy and Boltzmann entropy of a unary structure. We obtain related upper and lower bounds and thereby also establish the following asymptotic equivalence for *every* sequence \mathfrak{M}_n of models of increasing size n : $H_S(\mathfrak{M}_n) \sim \frac{1}{n} H_B(\mathfrak{M}_n)$. We note that a result bearing a resemblance to this one has been obtained in a slightly different framework in [15].

Finally, we relate the description complexity of a model to its entropy. We investigate the general picture of the relationship by giving upper and lower bounds on the description complexity of a model in terms of its entropy. See Figure 1a for the case of FO and Figure 1b

for FO_d . The bounds allow us to *exclude* a large portion of the (a priori) possible combinations of description complexity and entropy. In particular, we see that models with very high entropy have higher description complexity than models with very low entropy. Moreover, *models with a very low entropy are guaranteed to have a reasonably low description complexity, while models with very high entropies must have a notable description complexity.*

1.2 Related work, techniques and applications

Description complexity is conceptually related to Kolmogorov complexity, and it is also well known that entropy and Kolmogorov complexity are linked. Indeed, for computable distributions, Shannon entropy links to Kolmogorov complexity to within a constant. This is discussed, e.g., in [17, 9, 16]. However, [23] shows that the general link fails for Rényi and Tsallis entropies. See, e.g., [9, 16, 23] for discussions on Rényi and Tsallis entropies.

Concerning work in the intersection of logic and entropy, the recent article [14] by Jaakkola et al. provides related results for a graded modal logic GMLU over Kripke-models with the universal accessibility relation.

They show that the expected Boltzmann entropy of the equivalence classes of GMLU is asymptotically equivalent to the expected description complexity times the vocabulary size. While [14] concerns GMLU, the current paper studies (monadic) FO. Because of the multi-variable nature of FO, this leads to some major differences in the techniques required. The upper bound formulas of the current paper use some clever tricks that are not possible in the modal logic GMLU. Indeed, together with the results of [14], our upper bound formulas show that FO is more succinct than GMLU. Furthermore, the techniques used for the lower bounds for GMLU do not suffice for FO, necessitating new arguments.

Surprisingly, the relationship to entropy also turns out to be different. Indeed, in the case of GMLU, models with maximal entropy have maximal description complexity, while in the case of FO this is no longer the case.

For proving bounds on formula sizes, we use *formula size games* for FO. Indeed, variants of standard Ehrenfeucht-Fraïssé games would not suffice, as we need to deal with formula length, and thereby with all logical operators, including connectives. The formula size game that we use for FO is a slight modification of the game of Hella and Väänänen [10]. The first formula size game, developed by Razborov in [20], dealt with propositional logic. A later variant of the game was defined by Adler and Immerman for CTL in [1]. In [11] the formula size game for modal logic ML was used by Hella and Vilander to establish that bisimulation invariant FO is non-elementarily more succinct than ML. For a further example, we also mention the frame validity games of Balbiani et al. [2]. Recently, Fagin et al. in [7, 8] and Carmosino et al. in [4, 5, 6] have developed and used *multi-structural games* to prove lower bounds on the *number of quantifiers* that are needed for separating two structures in a given logic. In [8] they have also pointed out that strong lower bounds on the number of quantifiers would imply new lower bounds in circuit complexity.

Description complexity is relevant in many applications, one interesting link being data compression. It is natural to consider unary models \mathfrak{M} as data sets to be compressed into corresponding FO-sentences. To give a *simplified* example, let \mathcal{M} be the class of models over the unary alphabet $\tau = \{P, Q\}$ and with domain $M = \{1, \dots, 10\}$. Let \mathfrak{M}_1 be the model where $P^{\mathfrak{M}_1} = Q^{\mathfrak{M}_1} = M$ and \mathfrak{M}_2 be the model where $P^{\mathfrak{M}_2} = \{1, 2, 3\}$ and $Q^{\mathfrak{M}_2}$ is, say, $\{3, 4, 5, 6, 7\}$. Now, the simple formula $\forall x(P(x) \wedge Q(x))$ fully defines the model \mathfrak{M}_1 with respect to \mathcal{M} , while the model \mathfrak{M}_2 clearly requires a more complex formula. Suppose then that our models are represented as tabular Boolean data, meaning that each model corresponds to a 0-1-matrix with ten rows (one row for each domain element $m \in M$) and

two columns, one column for P and another one for Q . In this framework, when using FO as a compression language, the Boolean matrix for \mathfrak{M}_1 then compresses nicely into the formula $\forall x(P(x) \wedge Q(x))$, while the matrix for \mathfrak{M}_2 compresses to a notably more complex formula.

Many of the technical goals in *explainable artificial intelligence* (XAI) relate to compression [22], often revolving around issues of compressing information given by probability distributions. It is natural to expect representations of distributions with very high values of Shannon entropy to be more difficult to compress than ones with very low values. Concerning formula length, recent articles on XAI using minimum length formulas of logics as explanations of longer specifications include, e.g., [3, 18, 12, 13], and numerous others. For work on using short Boolean formulas as general explanations of real-life data given in the form of unary relational structures (i.e., tabular Boolean data sets), see [13]. In that paper, surprisingly short Boolean formulas are shown to give similar error rates to ones obtained by more sophisticated classifiers, e.g., neural networks and naive Bayesian classifiers.

Concerning further directions in explainability, minimum size descriptions ψ of unary relational models \mathfrak{M} can be useful for finding explanations in the context of the *special explainability problem* [12]. The positive case of this problem amounts to finding formulas χ with a given bound k on length such that $\mathfrak{M} \models \chi \models \varphi$, where φ acts as a classifier. In this context, it often suffices to find a short interpolant χ such that $\psi \models \chi \models \varphi$, where ψ is a minimum description of \mathfrak{M} . In applications, this latter task can often be more efficient than the first one, especially when ψ is significantly smaller than \mathfrak{M} . One way to ensure ψ is short enough is to describe \mathfrak{M} in a sufficiently incomplete way, such as with FO_d with small d .

Finally, in applications, it is typically easy to compute the Shannon entropy of structures, while description complexity and thereby issues relating to compressibility and explainability are *much more difficult to determine*. Therefore, even a rough picture of the links between entropy and description complexity can be useful.

The plan of the paper is as follows. After the preliminaries in Section 2, we provide upper bounds for the description complexity of unary structures in Section 3. In Section 4 we establish related lower bounds using games. In Section 5 we determine asymptotically the expected description complexity of a random unary structure. In Section 6 we give bounds on the relationship between entropy and description complexity. In Section 7 we conclude.

2 Preliminaries

Let $\tau = \{P_1, \dots, P_k\}$ be a monadic vocabulary and let $Var = \{x_1, x_2, \dots\}$ be a countably infinite set of variables. The syntax of first-order logic $\text{FO}[\tau]$ is generated by the grammar: $\varphi ::= x = y \mid P(x) \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \exists x\varphi \mid \forall x\varphi$, where $x, y \in Var$ and $P \in \tau$. The **quantifier rank** of a formula $\varphi \in \text{FO}[\tau]$ is the maximum number of nested quantifiers in the formula. We denote by $\text{FO}_d[\tau]$ the fragment of $\text{FO}[\tau]$ that only includes the formulas with quantifier rank at most d . A formula $\varphi \in \text{FO}[\tau]$ is in **negation normal form** if negations are only applied to **atomic formulas** $x = y$ or $P(x)$. We assume all formulas are in negation normal form and treat the notation $\neg\varphi$ as shorthand for the negation normal form formula obtained from φ by pushing the negation to the level of atomic formulas.

The **size** of a formula $\varphi \in \text{FO}[\tau]$ is defined as the number of atomic formulas, conjunctions, disjunctions and quantifiers in φ . Note that negations do not contribute to the size of φ . This choice together with using negation normal form means that positive and negative atomic information is treated as equal in terms of formula size. In line with this thinking, we will refer also to $x \neq y$ and $\neg P(x)$ as atomic formulas in the sequel.

A formula $\varphi \in \text{FO}[\tau]$ is in **prenex normal form** if it is of the form $Q_1x_1 \dots Q_mx_m\psi$, where $Q_i \in \{\exists, \forall\}$ for $i \in \{1, \dots, m\}$ and $\psi \in \text{FO}[\tau]$ has no quantifiers. It is well-known that every FO-formula can be transformed into an equivalent formula in prenex normal form which has the same size as the original formula.

A τ -**model** is a tuple $\mathfrak{M} = (M, P_1^{\mathfrak{M}}, \dots, P_k^{\mathfrak{M}})$, where $M = \{1, \dots, n\}$ and $P_i^{\mathfrak{M}} \subseteq M$ for $i \in \{1, \dots, k\}$. A model \mathfrak{M} is a **model of size n** if $|M| = n$. A partial function $s : \text{Var} \rightarrow M$ is called an **interpretation**. We also call pairs (\mathfrak{M}, s) models and identify the pair $(\mathfrak{M}, \emptyset)$ with the model \mathfrak{M} . The truth relation $(\mathfrak{M}, s) \models \varphi$ is defined in the usual way for $\text{FO}[\tau]$.

Let $\mathfrak{M} = (M, P_1^{\mathfrak{M}}, \dots, P_k^{\mathfrak{M}})$ be a τ -model of size n . We say that a formula $\varphi \in \text{FO}[\tau]$ **defines** \mathfrak{M} if for all τ -models \mathfrak{M}' of size n we have $(\mathfrak{M}', \emptyset) \models \varphi$ iff \mathfrak{M}' is isomorphic to \mathfrak{M} . As first-order logic cannot distinguish between isomorphic structures, we can in some sense identify the model \mathfrak{M} with the class of models isomorphic to \mathfrak{M} . The **description complexity** $C(\mathfrak{M})$ of \mathfrak{M} is the size of the smallest formula in $\text{FO}[\tau]$ that defines \mathfrak{M} .

Note that our definition of description complexity concerns separating \mathfrak{M} only from other models of the same size n . Requiring separation from all other models would unduly emphasize the size of the model, making even very simple models have a high description complexity. For example, the model $\mathfrak{M} = (M, P^{\mathfrak{M}})$ of size n , where $P^{\mathfrak{M}} = M$, would already require a formula with size in the order of n . In our setting, $C(\mathfrak{M}) = 2$, because \mathfrak{M} is defined by the formula $\forall xP(x)$.

A τ -**type** π is a subset of τ . A point $a \in M$ **realizes** a τ -type π if for all $P \in \tau$ we have $a \in P^{\mathfrak{M}}$ iff $P \in \pi$. We let $|\pi|_{\mathfrak{M}}$ denote the number of points in \mathfrak{M} realizing π . We often omit the subscript when the model is clear from the context. Note that two τ -models \mathfrak{M} and \mathfrak{M}' are isomorphic iff each type is realized in the same number of points in both models.

We also consider more coarse ways to divide models into classes than isomorphism. For each positive integer d we can define an equivalence relation \equiv_d over τ -models of size n as follows. Given two τ -models \mathfrak{M} and \mathfrak{M}' of size n , we define that $\mathfrak{M} \equiv_d \mathfrak{M}'$ iff for each τ -type π with $|\pi|_{\mathfrak{M}} < d$, we have that $|\pi|_{\mathfrak{M}} = |\pi|_{\mathfrak{M}'}$. In other words, $\mathfrak{M} \equiv_d \mathfrak{M}'$ iff each type that is realized in less than d points in \mathfrak{M} is realized in the same number of points in both models. It is easy to show that $\mathfrak{M} \equiv_d \mathfrak{M}'$ iff they satisfy the same sentences of $\text{FO}_d[\tau]$. The **d -description complexity** $C_d(\mathfrak{M})$ of a τ -model \mathfrak{M} is the size of the smallest $\text{FO}_d[\tau]$ -formula that defines the equivalence class of \mathfrak{M} in \equiv_d .

To characterize model classes, we use tuples with $t = 2^{|\tau|}$ numbers. For an isomorphism class, the tuple is simply $(|\pi_1|, \dots, |\pi_t|)$. For an equivalence class \mathcal{M} of \equiv_d , we only use numbers up to d . For a tuple $\bar{m} = (m_1, \dots, m_t)$, if $m_i = d$, then there are at least d realizing points of type π_i in models of the class \mathcal{M} . If $m_i < d$, then each model has exactly m_i points realizing the type π_i . The notation $\mathcal{M}_{\bar{m}}$ refers to classes of \equiv_d via these tuples. The tuples that correspond to some class of \equiv_d are characterized by the conditions $m_i \leq d$ for $i \in \{1, \dots, t\}$, $\sum_{i=1}^t m_i \leq n$ and if $\sum_{i=1}^t m_i < n$, then $m_j = d$ for some $j \in \{1, \dots, t\}$. If $\sum_{i=1}^t m_i = n$, then $\mathcal{M}_{\bar{m}}$ is an isomorphism class.

Since τ -types partition the points of a τ -model \mathfrak{M} , we may consider a natural probability distribution over the types in \mathfrak{M} . The probability p_π of a type π is simply $|\pi|/n$, that is, the probability of hitting a point of type π when selecting a point from \mathfrak{M} randomly. The **Shannon entropy** of \mathfrak{M} is the quantity $H_S(\mathfrak{M}) := \sum_{i=1}^t -p_{\pi_i} \log(p_{\pi_i}) = \sum_{i=1}^t -\frac{|\pi_i|}{n} \log\left(\frac{|\pi_i|}{n}\right)$. Here we follow the convention $0 \log(0) = 0$. Shannon entropy is an information theoretic way of measuring randomness of probability distributions. Uniform distributions have maximal Shannon entropy, as the uncertainty of the outcome of choosing a random point is maximized. Conversely, for a distribution that places all of the probability mass on a single event, Shannon

entropy is zero. Hence, a model realizing each type the same number of times (or as close as possible) has maximal Shannon entropy, while for a model that realizes only a single type Shannon entropy is zero.

Another way to define entropy of a model \mathfrak{M} uses the model class \mathfrak{M} belongs to. Given an equivalence relation \equiv over models of size n (and thus domain $\{1, \dots, n\}$), the **Boltzmann entropy** of \mathfrak{M} with respect to \equiv is $H_B(\mathfrak{M}) := \log(|\mathcal{M}|)$, where \mathcal{M} is the equivalence class of \mathfrak{M} . In this paper the equivalence relation \equiv is either isomorphism in the case of full FO or \equiv_d for FO_d. For isomorphism, we write $H_B(\mathfrak{M})$ and for \equiv_d we write $H_B^d(\mathfrak{M})$.

Boltzmann entropy originates from statistical mechanics, where it measures the randomness of a macrostate (= a model class) via the number of microstates (= models) that correspond to it. The idea is that a larger macrostate is “more random” (or “less specific”) since it is more likely to be hit by a random selection. We show in Section 6 that $H_S(\mathfrak{M}) \sim \frac{1}{n} H_B(\mathfrak{M})$, where n is the size of the domain of \mathfrak{M} . Thus the two notions of entropy are asymptotically equivalent up to normalization. This shows that both entropies indeed measure the randomness of a model from different points of view.

3 Upper bound formulas

In this section we define arbitrary τ -models via formulas of size linear in the size of the model. Recall that defining a model means separating it from all non-isomorphic models with the same domain size. To see why linear size formulas are quite succinct, note that the following naive formula $\bigwedge_{\ell=1}^{2^{|\tau|}} \exists x_1 \dots \exists x_{|\pi_\ell|} \left(\bigwedge_{i=1}^{|\pi_\ell|} \pi_\ell(x_i) \wedge \bigwedge_{j=i+1}^{|\pi_\ell|} x_i \neq x_j \right)$, which expresses that for each $1 \leq \ell \leq 2^{|\tau|}$ the type π_ℓ is realized by at least $|\pi_\ell|$ distinct points, is of quadratic size in the size n of the model.

For clean results on formula size, we define a constant $c_\tau := 15|\tau|2^{|\tau|}$. Note that we consider c_τ to be constant as it only depends on the size of the alphabet τ , which in our context is constant.

► **Theorem 1.** *Let \mathfrak{M} be a model of size n . Let $T = \{\pi_1, \dots, \pi_\ell\}$ be the types realized in \mathfrak{M} , enumerated in ascending order of numbers of realizing points. Now we have the bound $C(\mathfrak{M}) \leq \min(3|\pi_\ell| + c_\tau, 6|\pi_{\ell-1}| + c_\tau)$.*

Proof. We obtain two different upper bound formulas. Due to lack of space, we only give one of them in full here; see A.1 for details on the second formula.

We begin with an easy formula we use extensively below. For a type π and $x \in \text{Var}$, let

$$\pi(x) := \bigwedge_{P \in \pi} P(x) \wedge \bigwedge_{P \notin \pi} \neg P(x).$$

The formula $\pi(x)$ states that the point x realizes the type π .

Let $T = \{\pi_1, \dots, \pi_\ell\}$ be a set of τ -types and let \bar{m} be a sequence of $r \leq \ell$ positive integers with $0 < m_1 \leq \dots \leq m_r$. Let \mathfrak{M} be a model of size n , where exactly the types in T are realized. We will make sure of this with a separate formula later. The formula $\varphi(T, \bar{m})$ below is satisfied by such a model \mathfrak{M} if and only if for every $i \in \{1, \dots, r\}$, the model \mathfrak{M} has *at least* m_i points that realize the type π_i . Note that we do not assert anything about the types $\pi_{r+1}, \dots, \pi_\ell$, but we still need to mention them in the formula. We define

$$\psi_{m_r} := y \neq x_{m_r-1} \wedge \bigvee_{\substack{j \in \{1, \dots, r\} \\ m_j = m_r}} (\pi_j(x_1) \wedge \pi_j(y))$$

$$\begin{aligned}
\psi_i &:= y \neq x_{i-1} \wedge \psi_{i+1}, \text{ if } m_j \neq i \text{ for all } j \in \{1, \dots, r\}, \text{ and} \\
\psi_i &:= y \neq x_{i-1} \wedge \left(\bigvee_{\substack{j \in \{1, \dots, r\} \\ m_j = i}} (\pi_j(x_1) \wedge \pi_j(y)) \vee \psi_{i+1} \right), \text{ otherwise.} \\
\psi_1 &:= \psi_2, \text{ if } m_j \neq 1 \text{ for all } j \in \{1, \dots, r\}, \text{ and} \\
\psi_1 &:= \bigvee_{\substack{j \in \{1, \dots, r\} \\ m_j = 1}} \pi_j(x_1) \vee \psi_2, \text{ otherwise.} \\
\varphi(T, \bar{m}) &:= \forall x_1 \dots \forall x_{m_r-1} \exists y \left(\bigvee_{j \in \{r+1, \dots, \ell\}} \pi_j(x_1) \vee \psi_1 \right)
\end{aligned}$$

We proceed with an explanation of how the formula $\varphi(T, \bar{m})$ works. We assume that precisely the types in T are realized in the model \mathfrak{M} to be evaluated, so we know that the first universal variable x_1 is always attached to a point that realizes one of the types in T . The formula first checks if x_1 realizes one of the types $\pi_{r+1}, \dots, \pi_\ell$ that we wish to ignore. The recursion then handles the rest of the types, starting with the smallest ones. If the type π_j of x_1 has $m_j = 1$, nothing further is stated as we already know the type is realized in \mathfrak{M} by our assumption.

Now, consider a type π_j with, say, $m_j = 5$. Up to the subformula ψ_5 , the recursion of our formula has insisted that $y \neq x_i$ for $i \in \{1, 2, 3, 4\}$. Note that the formula does not contain any atomic formulas $x_{i_1} \neq x_{i_2}$. The crucial point is that since the variables x_1, \dots, x_4 are universally quantified, the existence of y must hold also in the case, where x_1, \dots, x_4 happen to all be different points of the same type π_j . If the evaluated model \mathfrak{M} has at least 5 points that realize π_j , then the formula holds as another point y that realizes π_j can be found. If, however, \mathfrak{M} has only 4 points that realize π_j , then one of the universally quantified tuples includes precisely those 4 points and another y of the same type cannot be found.

We adopt the notation $k = |\tau|$ and compute the size of $\varphi(T, \bar{m})$. The formula has m_r quantifiers. For each type $\pi \in T$, there are at most two occurrences of the subformula $\pi(x)$ (with different variables x). Each subformula $\pi(x)$ contains k atomic formulas. Thus there are at most $2k|T|$ atomic formulas of the form $P(x)$ or $\neg P(x)$. Each inequality $y \neq x_i$ for $1 \leq i \leq m_r - 1$ occurs exactly once, so there are $m_r - 1$ atomic formulas that are equalities or inequalities. Finally we multiply the number of atomic formulas by two and subtract one to also account for the binary connectives. The size of $\varphi(T, \bar{m})$ is thus at most

$$m_r + 2(m_r - 1 + 2k|T|) - 1 = 3m_r + 4k|T| - 3.$$

We proceed to define our first complete upper bound formula that defines an isomorphism class of models. Let \mathfrak{M} be a τ -model with domain $M = \{1, \dots, n\}$. Let $T = \{\pi_1, \dots, \pi_\ell\}$ be the set of τ -types realized in \mathfrak{M} and let $\bar{m} = (|\pi_1|, \dots, |\pi_\ell|)$. Assume further that \bar{m} is increasing. The full formula $\varphi(\mathfrak{M})$ is based on bounding the size of every type in T from below, thus separating it from all non-isomorphic models with the same domain size.

$$\varphi(\mathfrak{M}) := \bigwedge_{i=1}^{\ell} \exists x \pi_i(x) \wedge \forall x \bigvee_{i=1}^{\ell} \pi_i(x) \wedge \varphi(T, \bar{m})$$

In addition to the size of $\varphi(T, \bar{m})$ computed above, $\varphi(\mathfrak{M})$ includes $|T| + 1$ quantifiers and two occurrences of $\pi(x)$ for each type $\pi \in T$, resulting in $2k|T|$ atomic formulas. Accounting for the added binary connectives, the size of $\varphi(\mathfrak{M})$ is thus at most

$$|T| + 1 + 2 \cdot 2k|T| + 3|\pi_\ell| + 4k|T| - 3 = 3|\pi_\ell| + 8k|T| + |T| - 2 \leq 3|\pi_\ell| + c_\tau.$$

17:8 Description Complexity in FO with Links to Entropy

The second formula $\psi(\mathfrak{M})$ of size at most $6|\pi_{\ell-1}| + c_\tau$ states that each type π_i with $i \neq \ell$ has exactly $|\pi_i|$ points. See A.1 for details. Both formulas define any model \mathfrak{M} so we can always use whichever is smaller, thus proving the claim. \blacktriangleleft

► **Corollary 2.** *Let \mathfrak{M} be a model of size n . Now $C(\mathfrak{M}) \leq 2n + c_\tau$.*

Proof. A model \mathfrak{M} corresponding to the tuple $(0, \dots, 0, n/3, 2n/3)$ maximises the value of the expression $\min(3|\pi_\ell| + c_\tau, 6|\pi_{\ell-1}| + c_\tau)$, getting the value $2n + c_\tau$. \blacktriangleleft

We now consider defining equivalence classes of \equiv_d . Recall that an equivalence class of \equiv_d corresponds to a tuple $\bar{m} = (m_1, \dots, m_t)$, where $t = 2^{|\tau|}$, $m_i \leq d$ for all $i \in \{1, \dots, t\}$, $\sum_{i=1}^t m_i \leq n$ and if $\sum_{i=1}^t m_i < n$, then $m_j = d$ for some $j \in \{1, \dots, t\}$.

► **Theorem 3.** *Let \mathfrak{M} be a τ -model of size n . Let $\mathcal{M}_{\bar{m}}$ be the equivalence class of \mathfrak{M} in \equiv_d , where $\bar{m} = (m_1, \dots, m_t)$ is the corresponding tuple with the numbers in ascending order. Let m_r be the highest number in \bar{m} below d . Now $C_d(\mathfrak{M}) \leq 3d + 3m_r + c_\tau$. Additionally, if $m_{t-1} < d$, then $C_d(\mathfrak{M}) \leq 6m_{t-1} + c_\tau$.*

Proof. We use the same subformulas from Theorem 1 to obtain two linear size formulas. See A.2 for details. The first formula of size $3d + 3m_r + c_\tau$ works for any tuple \bar{m} and states that each type π_i has exactly m_i points if $m_i < d$ and at least d points if $m_i = d$. The second formula of size $6m_{t-1} + c_\tau$ states that each type π_i with $i \neq t$ has exactly m_i points and works only if all types except possibly π_t have less than d points. \blacktriangleleft

Note that since $m_r < d$, we have $6m_r < 3d + 3m_r$ so the bound for the special case is tighter than the general one. While we must use the more general bound for any \bar{m} with at least two instances of d , the tighter bound is significantly better for small classes with only one instance of d in their tuple. For example, the class with the tuple $(0, \dots, 0, 1, d)$ gets an upper bound of $6 + c_\tau$ regardless of the number d . At the other extreme, the class with the tuple $(0, \dots, 0, d-1, d, d)$ gets an upper bound of $3d + 3(d-1) + c_\tau = 6d - 3 + c_\tau$.

We again directly obtain a global upper bound on description complexity.

► **Corollary 4.** *Let \mathfrak{M} be a τ -model of size n . Now $C_d(\mathfrak{M}) \leq 6d - 3 + c_\tau$.*

4 Lower bounds via formula size games

In this section, we show lower bounds that match the upper bounds of Section 3 up to a factor of 2. We use the formula size game for first-order logic defined in [10]. We modify the game slightly to correspond to formulas in prenex normal form as this form does not affect the size of the formula. In addition, we introduce a second resource parameter q that corresponds to the number of quantifiers in the separating formula. The game consists of two phases: a quantifier phase, where only \exists -moves and \forall -moves can be made by S, and an atomic phase, where only \vee -moves, \wedge -moves and atomic moves can be made. Before the definition of the game, we define some notation.

Let \mathcal{A} be a set of τ -models and let $\varphi \in \text{FO}[\tau]$. We denote $\mathcal{A} \models \varphi$ to mean $(\mathfrak{M}, s) \models \varphi$ for all $(\mathfrak{M}, s) \in \mathcal{A}$. Similarly, we denote $\mathcal{A} \models \neg\varphi$ to mean $(\mathfrak{M}, s) \not\models \varphi$ for all $(\mathfrak{M}, s) \in \mathcal{A}$.

For an interpretation s , a point $a \in M$ and a variable $x \in \text{Var}$, we denote by $s[a/x]$ the interpretation s' such that $s'(x) = a$ and $s'(y) = s(y)$ for all $y \in \text{dom}(s)$, $y \neq x$. Let \mathcal{A} be a set of τ -models with the same domain M and let $f : \mathcal{A} \rightarrow M$ be a function. We denote by $\mathcal{A}[f/x]$ the set $\{(\mathfrak{M}, s[f(\mathfrak{M}, s)/x]) \mid (\mathfrak{M}, s) \in \mathcal{A}\}$. Intuitively, the function f gives the

new interpretation of the variable x for each model $(\mathfrak{M}, s) \in \mathcal{A}$. Additionally, we denote $\mathcal{A}[M/x] := \{(\mathfrak{M}, s[a/x]) \mid (\mathfrak{M}, s) \in \mathcal{A}, a \in M\}$. Here the variable x is given all possible interpretations, usually leading to a larger set of models. We next define the game.

Let \mathcal{A}_0 and \mathcal{B}_0 be sets of τ -models and let $r_0, q_0 \in \mathbb{N}$ with $r_0 > q_0$. The FO **prenex formula size game** $\text{FS}^\tau(r_0, q_0, \mathcal{A}_0, \mathcal{B}_0)$ has two players: Samson (S) and Delilah (D). Positions of the game are of the form $(r, q, \mathcal{A}, \mathcal{B})$, where $r, q \in \mathbb{N}$ and \mathcal{A} and \mathcal{B} are sets of τ -models. The starting position is $(r_0, q_0, \mathcal{A}_0, \mathcal{B}_0)$. In a position $(r, q, \mathcal{A}, \mathcal{B})$, if $r = 0$, then the game ends and D wins. Otherwise, if $q > 0$, the game is said to be in the **quantifier phase** and S can choose from the following three moves:

- **\exists -move:** S chooses $f : \mathcal{A} \rightarrow M$ and $x_i \in \text{Var}$. The new position is $(r - 1, q - 1, \mathcal{A}[f/x_i], \mathcal{B}[M/x_i])$.

- **\forall -move:** The same as the \exists -move with the roles of \mathcal{A} and \mathcal{B} switched.

- **Phase change:** S moves on to the atomic phase and the new position is $(r, 0, \mathcal{A}, \mathcal{B})$.

In a position $(r, q, \mathcal{A}, \mathcal{B})$, if $q = 0$, the game is said to be in the **atomic phase** and S can choose from the following three moves:

- **\wedge -move:** S chooses $r_1, r_2 \in \mathbb{N}$ and $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{B}$ such that $r_1 + r_2 + 1 = r$ and $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B}$. Then D chooses the next position from the options $(r_1, 0, \mathcal{A}, \mathcal{B}_1)$ and $(r_2, 0, \mathcal{A}, \mathcal{B}_2)$.

- **\vee -move:** The same as the \wedge -move with the roles of \mathcal{A} and \mathcal{B} switched.

- **Atomic move:** S chooses an atomic formula α . The game ends. If $\mathcal{A} \models \alpha$ and $\mathcal{B} \models \neg\alpha$, then S wins. Otherwise, D wins.

The prenex formula size game characterizes separation of model classes with formulas of limited size in the following way.

► **Theorem 5.** *Let \mathcal{A}_0 and \mathcal{B}_0 be sets of τ -models and let $r_0, q_0 \in \mathbb{N}$ with $r_0 > q_0$. The following are equivalent*

1. *S has a winning strategy in the game $\text{FS}^\tau(r_0, q_0, \mathcal{A}_0, \mathcal{B}_0)$,*
2. *there is an FO $[\tau]$ -formula φ in prenex normal form with size at most r_0 and at most q_0 quantifiers such that $\mathcal{A}_0 \models \varphi$ and $\mathcal{B}_0 \models \neg\varphi$,*
3. *there is an FO $[\tau]$ -formula φ with size at most r_0 and at most q_0 quantifiers such that $\mathcal{A}_0 \models \varphi$ and $\mathcal{B}_0 \models \neg\varphi$.*

Proof. For the simple inductive proof on how the game works, see [10]. The slight modifications of the separate parameter q for quantifiers and prenex normal form do not change the proof in any meaningful way so we omit it. For the equivalence between the second and third item, note that transforming a formula into prenex form and renaming variables as needed, does not increase its size in full FO with no restrictions on, say, the number of variables. ◀

We take a moment to build some intuition on the formula size game. The role of player S is to show that the model sets \mathcal{A}_0 and \mathcal{B}_0 can be separated by some FO formula with restrictions on size and number of quantifiers. To achieve this, S starts building the supposedly separating formula, starting from the quantifiers.

Each move of the game corresponds to an operator or atomic formula. When making a move, S makes choices for each model that reflect how that particular model is going to satisfy the formula, in the case of models in \mathcal{A} , or not satisfy it, in the case of models in \mathcal{B} . For example, for an \exists -move, S must choose for each model in \mathcal{A} the point to quantify. This is done via the function f . For a \wedge -move, S chooses for each model in \mathcal{B} one of the conjuncts, asserting that the model will not satisfy that conjunct.

The resources r_0 and q_0 restrict the moves of S. He can only make at most q_0 quantifier moves in the quantifier phase of the game. The resource r_0 limits the size of the entire separating formula, including the quantifiers. In the atomic phase, for \wedge -moves, S must

17:10 Description Complexity in FO with Links to Entropy

divide the remaining resource r between the two conjuncts. It is then the role of D to choose the conjunct she thinks cannot be completed in such a way that the models present are separated. Once D has chosen a conjunct, the other conjunct not chosen is discarded for the rest of the game. Thus, the entire separating formula need not be constructed.

We move on to our lower bounds. Let \mathfrak{M} be a τ -model with domain $M = \{1, \dots, n\}$ and let $T = \{\pi_1, \dots, \pi_\ell\}$ be the types realized in \mathfrak{M} , enumerated in ascending order of numbers of realizing points, like in the previous section. We assume that $\ell \geq 2$ as a model, where all points are of the same type, is easily defined by a constant-sized formula. We use the formula size game to show a lower bound of the order $3|\pi_{\ell-1}|$ for the description complexity of \mathfrak{M} .

Let \mathfrak{M}' be the model obtained from \mathfrak{M} by changing the type of one point from $\pi_{\ell-1}$ to π_ℓ . We define $\mathcal{A}_0 = \{(\mathfrak{M}, \emptyset)\}$ and $\mathcal{B}_0 = \{(\mathfrak{M}', \emptyset)\}$. We will show that separating the sets \mathcal{A}_0 and \mathcal{B}_0 requires a formula of size at least $3|\pi_{\ell-1}| - 3$. We begin with an easy lemma on the number of quantifiers required to separate \mathcal{A}_0 from \mathcal{B}_0 .

► **Lemma 6.** *If φ separates \mathcal{A}_0 from \mathcal{B}_0 , then φ has at least $|\pi_{\ell-1}|$ quantifiers.*

Proof. Let $r_0 > |\pi_{\ell-1}| - 1$. We show that D has a winning strategy for the formula size game $\text{FS}^\tau(r_0, |\pi_{\ell-1}| - 1, \mathcal{A}_0, \mathcal{B}_0)$. By Theorem 5, this proves the claim.

We show that in any position of such a game, there is a pair $(\mathfrak{M}, s) \in \mathcal{A}$ and $(\mathfrak{M}', s') \in \mathcal{B}$ of models that cannot be separated by any atomic formula. At the starting position, the single models in \mathcal{A}_0 and \mathcal{B}_0 are such a pair as no variables have been quantified. We proceed to show that D can maintain this pair of models through any move of S. We only treat one of each pair of dual moves as the other is handled the same way.

∃-**move:** S chooses a function $f : \mathcal{A} \rightarrow M$. We focus on the point $a = f(\mathfrak{M}, s)$ chosen for the model $(\mathfrak{M}, s) \in \mathcal{A}$. On the other side, copies of $(\mathfrak{M}', s') \in \mathcal{B}$ are generated for each point $b \in M$, but we restrict attention to only one as follows. If there is a previously quantified variable x with $s(x) = a$, then we choose $b = s'(x)$. Otherwise we choose a new point b of the same type as a . If the type of a is π_i with $i < \ell - 1$, then \mathfrak{M} and \mathfrak{M}' have the same points of type π_i so we may choose $b = a$. If $i \in \{\ell - 1, \ell\}$, then both \mathfrak{M} and \mathfrak{M}' have at least $|\pi_{\ell-1}| - 1$ points of the type π_i so we may choose a fresh b of the same type. The new pair of models found in this manner is clearly atomic-equivalent.

Phase change: With no changes to the sets of models \mathcal{A} and \mathcal{B} , the important pair of models is still clearly present in the next position.

∧-**move:** S chooses splits $r_1 + r_2 + 1 = r$ and $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B}$. Now the model $(\mathfrak{M}', s') \in \mathcal{B}$ is in \mathcal{B}_1 or \mathcal{B}_2 and \mathcal{A} remains unchanged. Thus our model pair is present in one of the positions $(r_1, 0, \mathcal{A}, \mathcal{B}_1)$ and $(r_2, 0, \mathcal{A}, \mathcal{B}_2)$. By choosing such a position, D maintains the pair of models.

Atomic move: The model pair is atomic-equivalent, so D wins after any atomic move. ◀

The next lemma concerns the atomic phase. We show that if the number of different atomic formulas required to separate the model sets \mathcal{A} and \mathcal{B} is too large, D wins the game.

► **Lemma 7.** *In a game $\text{FS}^\tau(r_0, q_0, \mathcal{A}_0, \mathcal{B}_0)$, let $(r, 0, \mathcal{A}, \mathcal{B})$ be the first position of the atomic phase and let Γ be a minimum size set of atomic formulas such that for every $(\mathfrak{M}, s) \in \mathcal{A}$ and $(\mathfrak{M}', s') \in \mathcal{B}$, there is $\alpha \in \Gamma$ with $(\mathfrak{M}, s) \models \alpha$ and $(\mathfrak{M}', s') \not\models \alpha$. If $r < 2|\Gamma| - 1$, then D has a winning strategy from the position $(r, 0, \mathcal{A}, \mathcal{B})$.*

Proof. We show that every move of S either ends the game in a win for D, or maintains the condition $r < 2|\Gamma| - 1$. Assume this condition holds in position $(r, 0, \mathcal{A}, \mathcal{B})$.

Atomic move: S chooses an atomic formula α . Since $1 \leq r < 2|\Gamma| - 1$, we have $|\Gamma| \geq 2$ so the single atomic formula α does not separate \mathcal{A} from \mathcal{B} and D wins.

\wedge -move: S chooses splits $r_1 + r_2 + 1 = r$ and $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B}$. Assume for contradiction that there are sets Γ_1 and Γ_2 of atomic formulas such that Γ_i separates \mathcal{A} from \mathcal{B}_i and $r_i \geq 2|\Gamma_i| - 1$. Now for every pair of models $(\mathfrak{M}, s) \in \mathcal{A}$ and $(\mathfrak{M}', s') \in \mathcal{B}$ we have $(\mathfrak{M}', s') \in \mathcal{B}_1$ or $(\mathfrak{M}', s') \in \mathcal{B}_2$ so the set $\Gamma_1 \cup \Gamma_2$ separates \mathcal{A} from \mathcal{B} . Recalling that Γ is a separating set of minimum size and $r < 2|\Gamma| - 1$, we also have $r < 2|\Gamma_1 \cup \Gamma_2| - 1 \leq 2(|\Gamma_1| + |\Gamma_2|) - 1 \leq r_1 + r_2 + 1 = r$, which is a contradiction. Thus we have $r_1 < 2|\Gamma_1| - 1$ or $r_2 < 2|\Gamma_2| - 1$. By choosing the correct position D can maintain the required condition.

\vee -move: Identical to the \wedge -move with the roles of \mathcal{A} and \mathcal{B} switched. \blacktriangleleft

We are now ready for the main theorem of this section.

► Theorem 8. *Let \mathfrak{M} be a model of size n . Let $T = \{\pi_1, \dots, \pi_\ell\}$ be the types realized in \mathfrak{M} , enumerated in ascending order of numbers of realizing points, where $\ell \geq 2$. Now $C(\mathfrak{M}) \geq 3|\pi_{\ell-1}| - 3$.*

Proof. We begin with a definition. Let Γ be a set of atomic FO-formulas. We denote the set of variables occurring in formulas of Γ by $V(\Gamma)$. We define the **variable graph of Γ** as $G(\Gamma) = (V(\Gamma), E(\Gamma))$, where $(x, y) \in E(\Gamma)$ iff $x = y \in \Gamma$ or $x \neq y \in \Gamma$. We say that $\Delta \subseteq \Gamma$ is a **connected component of Γ** if $G(\Delta)$ is a maximal connected subgraph of $G(\Gamma)$.

For convenience, we denote here $m := |\pi_{\ell-1}|$. Consider a formula size game $\text{FS}^\tau(3m - 4, q_0, \mathcal{A}_0, \mathcal{B}_0)$. We show that D has a winning strategy for this game, thus proving the claim by Theorem 5. By Lemma 6 we see that to have a chance of winning, S must begin the game with at least m quantifiers. We then move on to the first position $(r, 0, \mathcal{A}, \mathcal{B})$ of the atomic phase, where $r \leq 2m - 4$. Let Γ be a set of atomic formulas such that for every $(\mathfrak{M}, s) \in \mathcal{A}$ and $(\mathfrak{M}', s') \in \mathcal{B}$, there is $\alpha \in \Gamma$ such that $(\mathfrak{M}, s) \models \alpha$ and $(\mathfrak{M}', s') \not\models \alpha$. If $|\Gamma| \geq m - 1$ for every such Γ , then $r \leq 2m - 4 = 2(m - 1) - 2 < 2|\Gamma| - 1$ so D has a winning strategy by Lemma 7. We now assume for contradiction that there exists such a Γ with $|\Gamma| \leq m - 2$.

Consider the connected components Δ of Γ . Since a connected graph with k edges has at most $k + 1$ vertices, for every Δ at most $m - 1$ variables occur in the formulas of Δ .

We now explain why there is a *single* pair of models $(\mathfrak{M}, s) \in \mathcal{A}$ and $(\mathfrak{M}', s') \in \mathcal{B}$ such that they are atomic equivalent with respect to the variables in $V(\Delta)$ for *every* connected component Δ of Γ . We consider the quantifier moves S made in the quantifier phase in the order the moves were made. For every variable x used in a \exists -move, we consider Δ such that $x \in V(\Delta)$. We proceed as in the proof of Lemma 6, with respect to only the variables in $V(\Delta)$. That is, if there is a previously quantified variable $y \in V(\Delta)$ such that $s(y) = s(x)$, we choose the opposing model where $s'(x) = s'(y)$. Otherwise, we choose a point with no variables of $V(\Delta)$ attached. Each Δ uses at most $m - 1$ variables so we do not run out of fresh points of any type. The same protocol works for \forall -moves as well.

Note that the choices of models are made based on the connected component Δ of x , completely independently of other components. Since every variable x is in exactly one component Δ , this means that the resulting pair of models is simultaneously atomic equivalent with regards to each component separately. Thus this model pair cannot be separated by any atomic formula in Γ . This contradiction with the definition of Γ proves the claim. \blacktriangleleft

We now consider lower bounds in the setting of FO_d . Recall that an equivalence class of \equiv_d is characterized by a tuple (m_1, \dots, m_t) , where $t = 2^{|\tau|}$, $m_i \leq d$, $\sum_{i=1}^t m_i \leq n$ and if $\sum_{i=1}^t m_i < n$, then $m_j = d$ for some j . Let $\bar{m} = (m_1, \dots, m_t)$ be such a tuple in ascending order of the numbers m_i . If $\sum_{i=1}^t m_i = n$, then \bar{m} corresponds to an isomorphism class and the lower bounds above work as is. Thus we assume that $\sum_{i=1}^t m_i < n$ and consequently $m_t = d$. By taking a model \mathfrak{M} in the equivalence class $\mathcal{M}_{\bar{m}}$ with a maximal number of

17:12 Description Complexity in FO with Links to Entropy

points of the type π_t , we can directly obtain the model \mathfrak{M}' as above and get a lower bound on defining the class $\mathcal{M}_{\bar{m}}$ in full FO. This bound directly extends also to FO_d , as limiting quantifier rank gives no advantage in terms of formula size.

► **Corollary 9.** *Let $\mathcal{M}_{\bar{m}}$ be an equivalence class of \equiv_d , where $\bar{m} = (m_1, \dots, m_t)$ is the corresponding tuple with the numbers in ascending order. Now $C(\mathcal{M}_{\bar{m}}) \geq 3m_{t-1} - 3$.*

5 Expected description complexity

Using Theorems 1 and 8, we can determine asymptotically the expected description complexity of a *random* τ -model. Here by random we mean that the model is sampled uniformly at random from the set of all τ -models of size n . That is, we determine the asymptotic behavior of the quantity $\mathbb{E}_n[C] := \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M}} C(\mathfrak{M})$ as $n \rightarrow \infty$, where the sum is taken over all the τ -models \mathfrak{M} of size n .

We say that a τ -model \mathfrak{M} is **balanced**, if for every τ -type π , we have $|\pi|_{\mathfrak{M}} - \frac{n}{2^{|\tau|}} = o(n)$. In other words, a model is balanced if every type is realized roughly the same number of times, allowing for a sublinear discrepancy. We use the well-known Chernoff bounds to establish that a random model is very likely balanced.

► **Proposition 10** (Multiplicative Chernoff bound). *Let $X := \sum_{i=1}^n X_i$ be a sum of independent 0-1-valued random variables, where $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$. Let $\mu := \mathbb{E}[X]$. Now, for every $0 \leq \delta < 1$ we have that $\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\delta^2\mu/3}$*

Proof. See for example Corollary 4.6 in [19]. ◀

► **Lemma 11.** *The probability that a random τ -model of size n is balanced is at least $1 - 2^{|\tau|+1}/n$.*

Proof. A routine calculation using Proposition 10. See A.3 for details. ◀

The previous lemma gives a rough characterization of random τ -models. Using this characterization together with Theorem 8 we can determine asymptotically the expected description complexity of a random τ -model.

► **Theorem 12.** $\mathbb{E}_n[C] \sim \frac{3n}{2^{|\tau|}}$

Proof. To give an upper bound on $\mathbb{E}_n[C]$ we first rewrite it as follows:

$$\mathbb{E}_n[C] = \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ balanced}} C(\mathfrak{M}) + \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ not balanced}} C(\mathfrak{M}) \quad (1)$$

Using Corollary 2 and Lemma 11 we see that

$$\begin{aligned} \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ not balanced}} C(\mathfrak{M}) &\leq \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ not balanced}} 2n + c_\tau = \Pr[\mathfrak{M} \text{ is not balanced}] \cdot (2n + c_\tau) \\ &\leq \frac{2^{|\tau|+1}}{n} \cdot (2n + c_\tau) = 2^{|\tau|+2} + \frac{c_\tau 2^{|\tau|+1}}{n} = \mathcal{O}(1). \end{aligned}$$

Since we are interested in the asymptotic behavior of $\mathbb{E}_n[C]$, the above shows that we can safely concentrate on the first sum in Equation (1). Using Theorems 1 and 8 we see that if \mathfrak{M} is balanced, then $\frac{3n}{2^{|\tau|}} - o(n) \leq C(\mathfrak{M}) \leq \frac{3n}{2^{|\tau|}} + o(n)$. Hence

$$\Pr[\mathfrak{M} \text{ is balanced}] \cdot \left(\frac{3n}{2^{|\tau|}} - o(n) \right) \leq \frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ balanced}} C(\mathfrak{M}) \leq \Pr[\mathfrak{M} \text{ is balanced}] \cdot \left(\frac{3n}{2^{|\tau|}} + o(n) \right).$$

Since $\Pr[\mathfrak{M} \text{ is balanced}]$ goes to one as $n \rightarrow \infty$, we see that $\frac{1}{2^{|\tau|n}} \sum_{\mathfrak{M} \text{ balanced}} C(\mathfrak{M}) \sim \frac{3n}{2^{|\tau|}}$, which is what we wanted to show. \blacktriangleleft

6 Entropy and description complexity

In this section we establish results that illustrate how entropy and description complexity relate to each other. As one can already imagine after seeing our results on description complexity, there can be models with very close entropies and quite different description complexities. We can nevertheless use our results to *exclude* many a priori possible combinations of description complexity and entropy. For notational simplicity, we adopt the notation $t := 2^{|\tau|}$.

We begin by showing that the Boltzmann and Shannon entropies of a single model are essentially the same up to normalization. This underlines the fact that both entropies measure the same thing: the randomness of a model.

► **Theorem 13.** *Let \mathfrak{M} be a τ -model of size n . Now*

$$H_S(\mathfrak{M}) - \frac{1}{n} H_B(\mathfrak{M}) < \frac{(t-1) \log(\sqrt{2\pi n})}{n} - \frac{\log(e)}{12n^2} + \frac{t \log(e)}{12n^2 + n}.$$

Proof. Using the quantitative version of Stirling's approximation given in [21], we obtain

$$\begin{aligned} H_B(\mathfrak{M}) &= \log \binom{n}{n_1 \dots n_t} = \log \frac{n!}{n_1! \dots n_t!} = \log(n!) - \sum_{i=1}^t \log(n_i!) \\ &< \log \left(\sqrt{2\pi n} \left(\frac{n}{e} \right)^n e^{\frac{1}{12n}} \right) - \sum_{i=1}^t \log \left(\sqrt{2\pi n_i} \left(\frac{n_i}{e} \right)^{n_i} e^{\frac{1}{12n_i+1}} \right) \\ &= \log(\sqrt{2\pi n}) + n \log(n) - n \log(e) + \frac{\log(e)}{12n} \\ &\quad - \sum_{i=1}^t \left(\log(\sqrt{2\pi n_i}) + n_i \log(n_i) - n_i \log(e) + \frac{\log(e)}{12n_i+1} \right) \\ &\leq n \log(n) - \sum_{i=1}^t n_i \log(n_i) - (t-1) \log(\sqrt{2\pi n}) + \frac{\log(e)}{12n} - \frac{t \log(e)}{12n+1}. \end{aligned}$$

Note that the term $n \log(e)$ is cancelled out above because $n_1 + \dots + n_t = n$. Using this same fact we also easily see that

$$H_S(\mathfrak{M}) = \sum_{i=1}^t -\frac{n_i}{n} \log \frac{n_i}{n} = \sum_{i=1}^t \frac{n_i}{n} \log(n) - \sum_{i=1}^t \frac{n_i}{n} \log(n_i) = \log(n) - \sum_{i=1}^t \frac{n_i}{n} \log(n_i).$$

Finally, by dividing $H_B(\mathfrak{M})$ with n we obtain

$$H_S(\mathfrak{M}) - \frac{1}{n} H_B(\mathfrak{M}) < \frac{(t-1) \log(\sqrt{2\pi n})}{n} - \frac{\log(e)}{12n^2} + \frac{t \log(e)}{12n^2 + n}. \quad \blacktriangleleft$$

The above quantitative result readily implies that the Boltzmann and Shannon entropies of a single model are asymptotically the same up to normalization. A connection that bears a similarity to the one pointed out here has also been noted briefly in [15].

► **Corollary 14.** *Let $(\mathfrak{M}_n)_{n \in \mathbb{Z}_+}$ be a sequence of τ -models where each \mathfrak{M}_n has size n . Now $H_S(\mathfrak{M}_n) \sim \frac{1}{n} H_B(\mathfrak{M}_n)$ as $n \rightarrow \infty$.*

17:14 Description Complexity in FO with Links to Entropy

The above results show that for the connections to description complexity, we could use either of the two notions of entropy. We opt for Shannon entropy here.

We will next use results from Sections 3 and 4 to prove two theorems that give bounds on description complexity in terms of Shannon entropy. Recall from Section 3 the constant $c_\tau := 15|\tau|2^{|\tau|}$. The first of our two theorems gives global upper and lower bounds on description complexity based on the same edge case distributions.

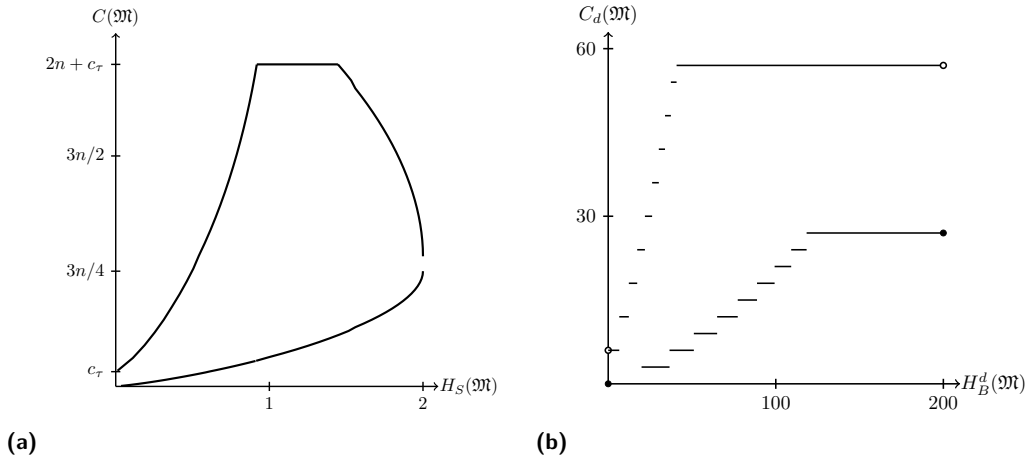
► **Theorem 15.** *Let $p \in [0, \frac{1}{t}]$. If $H_S(\mathfrak{M}) > ((t-1)p - 1) \log(1 - (t-1)p) - (t-1)p \log(p)$, then $3np - 3 < C(\mathfrak{M}) < 3n(1 - (t-1)p) + c_\tau$.*

Proof. Let $f(p) := ((t-1)p - 1) \log(1 - (t-1)p) - (t-1)p \log(p)$. The function $f(p)$ gives the entropy of a τ -model \mathfrak{M}' corresponding to the tuple $(np, \dots, np, n(1 - (t-1)p))$, where $n(1 - (t-1)p) > np$ for the given values of p . Since all types but the largest are evenly distributed, any model, where the largest type has at least $n(1 - (t-1)p)$ realizing points has entropy at most $H_S(\mathfrak{M}') = f(p)$. Therefore if $H_S(\mathfrak{M}) > f(p)$, then the largest type of \mathfrak{M} has less than $n(1 - (t-1)p)$ realizing points. By Theorem 1, we obtain $C(\mathfrak{M}) < 3n(1 - (t-1)p) + c_\tau$. On the other hand, since the largest type of \mathfrak{M} has less realizing points than in \mathfrak{M}' , those points realize some other type. Therefore the second largest type of \mathfrak{M} has more than np realizing points. By Theorem 8, we obtain $C(\mathfrak{M}) > 3np - 3$. ◀

The next theorem uses low entropy models with only two realized types to show a better upper bound on description complexity for low entropy models than the above global one.

► **Theorem 16.** *Let $p \in [0, \frac{1}{2}]$. If $H_S(\mathfrak{M}) < (p-1) \log(1-p) - p \log(p)$, then $C(\mathfrak{M}) < 6np + c_\tau$.*

Proof. Let $h(p) := (p-1) \log(1-p) - p \log(p)$. The function $h(p)$ gives the entropy of a τ -model \mathfrak{M} corresponding to the tuple $(0, \dots, 0, np, n(1-p))$. If $H_S(\mathfrak{M}) < h(p)$, then the second largest type of \mathfrak{M} must be smaller than np . Thus, by Theorem 1, $C(\mathfrak{M}) < 6np + c_\tau$. ◀



■ **Figure 1** Figure 1a on the left shows an area that encapsulates all combinations of Shannon entropy and FO-description complexity for the values $|\tau| = 2$ and $n = 1000$. Figure 1b on the right concerns the case of FO_d and shows bounds on description complexity in terms of Boltzmann entropy for values $|\tau| = 2$, $n = 100$ and $d = 10$ with the constants -3 and c_τ omitted.

Figure 1a incorporates both of the above theorems as well as Corollary 2 to show an area, where all possible combinations of Shannon entropy and description complexity must fall. First, comparing the left side of the plot to the right, we can see that models with very high entropy have significantly higher description complexity than models with very low entropy.

We can also see from Figure 1a that the gap between our upper bounds and lower bounds is only constant at both extremes of entropy. For models with middling entropy, the gap is at its largest. This is because middling values of entropy can be realized by models with very different distributions of types, leading to different description complexity.

We conjecture that the upper bound given by Theorem 1 is in reality tight up to the constant c_τ . Now, recall that for any single model, our upper and lower bounds have a worst case gap of a factor of 2. Therefore, assuming that our conjecture is true, the lower bound would only rise to at most double its current height. In other words, the general picture illustrated by Figure 1a would not be significantly different under our conjecture.

We proceed to show that similar relationships between description complexity and entropy hold also in the case of limited quantifier rank. As the classes of Ξ_d contain multiple different isomorphism types of models, it is not clear how to define Shannon entropy. Boltzmann entropy, however, is still straightforward so we use Boltzmann entropy here. We formulate similar theorems to those above for full FO.

► **Theorem 17.** *Let $h \in \{1, \dots, d-1\}$. If $H_B^d(\mathfrak{M}) > \log \binom{n}{h \dots h \ n-(t-1)h}$, then $C_d(\mathfrak{M}) > 3h-3$.*

Proof. Let $f(n, h) = \log \binom{n}{h \dots h \ n-(t-1)h}$. The function $f(n, h)$ gives the Boltzmann entropy of the class of models $\mathcal{M}_{\bar{m}}$, where $\bar{m} = (h, \dots, h, d)$. Any class of models obtained from this one by lowering any of the numbers in the tuple is clearly smaller than $\mathcal{M}_{\bar{m}}$ and thus has lower Boltzmann entropy. Thus, for any larger class of models the second largest number in its tuple must be greater than h . By Corollary 9, we obtain $C_d(\mathfrak{M}) > 3h-3$. ◀

► **Theorem 18.** *Let $h \in \{1, \dots, d-1\}$. If $H_B^d(\mathfrak{M}) < \log \binom{n}{h}$, then $C_d(\mathfrak{M}) < 6h + c_\tau$.*

Proof. The function $g(n, h) = \log \binom{n}{h}$ gives the Boltzmann entropy of a class $\mathcal{M}_{\bar{m}}$ of models, where $\bar{m} = (0, \dots, 0, h, d)$. Now every class of models, where the second largest number in the tuple is at least h , is larger than or equal to $\mathcal{M}_{\bar{m}}$. Thus if $H_B^d(\mathfrak{M}) < g(n, h)$, then the class of \mathfrak{M} is smaller and the second largest number in its tuple is smaller than h . By Theorem 3 we obtain $C_d(\mathfrak{M}) < 6h + c_\tau$. ◀

We again have a plot in Figure 1b, where the possible combinations of entropy and description complexity lie between the two chopped lines. This time, we plotted from the above theorems $3h$ for the lower bound and $6h$ for the upper bound, omitting the constants -3 and c_τ . For these low values of n and d , the constants would have warped the picture in a significant way. With high enough n and d , the constants are clearly negligible, but for such values, the Boltzmann entropy quickly becomes impractical to calculate as the model class sizes explode. We provide a plot of the leading terms for the values $n = 100$ and $d = 10$ without the constants to illustrate the trends one would see for higher values of n and d .

We see that the first observation we made for full FO still holds. The models with very high entropy have significantly higher description complexity than those with very low entropy. Concerning the gap between the upper and lower bounds, it is again constant at the extremes. The largest gap can now be found significantly before the halfway point of entropy, unlike for full FO. This is because the limit d of quantifier rank quite quickly cuts short the growth of the upper bound while the lower bound grows slower.

7 Conclusion

We have studied the description complexity of unary models, obtaining bounds for FO and FO_d . We have found the asymptotic description complexity of a random unary structure and studied the relation between Shannon entropy and description complexity – also observing a connection between Boltzmann and Shannon entropy. Links to entropy can be useful as computing entropy is *significantly easier than determining description complexity*.

An obvious future goal would be to close the gaps between the upper and lower bounds. Generalizing to full relational vocabularies is also interesting, although this seems to require highly involved arguments. The part on entropy would there relate to Boltzmann entropy, as there is no obvious unique definition for Shannon entropy in the k -ary scenario.

References

- 1 Micah Adler and Neil Immerman. An $n!$ lower bound on formula size. *ACM Trans. Comput. Log.*, 4(3):296–314, 2003. doi:10.1145/772062.772064.
- 2 Philippe Balbiani, David Fernández-Duque, Andreas Herzig, and Petar Iliev. Frame-validity games and lower bounds on the complexity of modal axioms. *Log. J. IGPL*, 30(1):155–185, 2022. doi:10.1093/jigpal/jzaa068.
- 3 Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/b1adda14824f50ef24ff1c05bb66faf3-Abstract.html>.
- 4 Marco Carmosino, Ronald Fagin, Neil Immerman, Phokion Kolaitis, Jonathan Lenchner, and Rik Sengupta. On the number of quantifiers needed to define boolean functions, 2024.
- 5 Marco Carmosino, Ronald Fagin, Neil Immerman, Phokion G. Kolaitis, Jonathan Lenchner, and Rik Sengupta. A finer analysis of multi-structural games and beyond. *CoRR*, abs/2301.13329, 2023. doi:10.48550/arXiv.2301.13329.
- 6 Marco Leandro Carmosino, Ronald Fagin, Neil Immerman, Ph. G. Kolaitis, Jonathan Lenchner, Rik Sengupta, and Ryan Williams. Parallel play saves quantifiers. *ArXiv*, 2024.
- 7 Ronald Fagin, Jonathan Lenchner, Kenneth W. Regan, and Nikhil Vyas. Multi-structural games and number of quantifiers. In *2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–13, 2021. doi:10.1109/LICS52264.2021.9470756.
- 8 Ronald Fagin, Jonathan Lenchner, Nikhil Vyas, and Ryan Williams. On the Number of Quantifiers as a Complexity Measure. In Stefan Szeider, Robert Ganian, and Alexandra Silva, editors, *47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022)*, volume 241 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 48:1–48:14, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.MFCS.2022.48.
- 9 Peter Grünwald and Paul M. B. Vitányi. Shannon information and Kolmogorov complexity. *CoRR*, cs.IT/0410002, 2004. URL: <http://arxiv.org/abs/cs.IT/0410002>, doi:10.48550/arXiv.cs/0410002.
- 10 Lauri Hella and Jouko Väänänen. The size of a formula as a measure of complexity. In Åsa Hirvonen, Juha Kontinen, Roman Kossak, and Andrés Villaveces, editors, *Logic Without Borders - Essays on Set Theory, Model Theory, Philosophical Logic and Philosophy of Mathematics*, volume 5 of *Ontos Mathematical Logic*, pages 193–214. De Gruyter, 2015. doi:10.1515/9781614516873.193.
- 11 Lauri Hella and Miikka Vilander. Formula size games for modal logic and μ -calculus. *J. Log. Comput.*, 29(8):1311–1344, 2019. doi:10.1093/logcom/exz025.

- 12 Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Feyzbakhsh Rankooh, and Miikka Vilander. Explainability via short formulas: the case of propositional logic with implementation. In *Joint Proceedings of (HYDRA 2022) and the RCRA Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion*, volume 3281 of *CEUR Workshop Proceedings*, pages 64–77, 2022. URL: <https://ceur-ws.org/Vol-3281/paper6.pdf>.
- 13 Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Feyzbakhsh Rankooh, and Miikka Vilander. Short boolean formulas as explanations in practice. In Sarah Alice Gaggl, Maria Vanina Martinez, and Magdalena Ortiz, editors, *Logics in Artificial Intelligence - 18th European Conference, JELIA 2023, Dresden, Germany, September 20-22, 2023, Proceedings*, volume 14281 of *Lecture Notes in Computer Science*, pages 90–105. Springer, 2023. doi:10.1007/978-3-031-43619-2_7.
- 14 Reijo Jaakkola, Antti Kuusisto, and Miikka Vilander. Relating description complexity to entropy. In Petra Berenbrink, Patricia Bouyer, Anuj Dawar, and Mamadou Moustapha Kanté, editors, *40th International Symposium on Theoretical Aspects of Computer Science, STACS 2023, March 7-9, 2023, Hamburg, Germany*, volume 254 of *LIPICs*, pages 38:1–38:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPICs.STACS.2023.38.
- 15 Andrey Kolmogorov. The theory of transmission of information. In *Selected Works of A. N. Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*, pages 6–32. Springer Netherlands, 1993. doi:10.1007/978-94-017-2973-4_3.
- 16 Sik K. Leung-Yan-Cheong and Thomas M. Cover. Some equivalences between Shannon entropy and Kolmogorov complexity. *IEEE Trans. Inf. Theory*, 24(3):331–338, 1978. doi:10.1109/TIT.1978.1055891.
- 17 Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 4th Edition*. Texts in Computer Science. Springer, 2019. doi:10.1007/978-3-030-11298-1.
- 18 João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7469–7479. PMLR, 2021. URL: <http://proceedings.mlr.press/v139/marques-silva21a.html>.
- 19 Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005. doi:10.1017/CB09780511813603.
- 20 Alexander A. Razborov. Applications of matrix methods to the theory of lower bounds in computational complexity. *Comb.*, 10(1):81–93, 1990. doi:10.1007/BF02122698.
- 21 Herbert Robbins. A remark on stirling’s formula. *The American Mathematical Monthly*, 62(1):26–29, 1955. doi:10.2307/2308012.
- 22 Advait Sarkar. Is explainable AI a race against model complexity? In *Workshop on Transparency and Explanations in Smart Systems (TeXSS), in conjunction with ACM Intelligent User Interfaces (IUI 2022)*, volume 3124 of *CEUR Workshop Proceedings*, pages 192–199, 2022.
- 23 Andreia Teixeira, Armando Matos, Andre Souto, and Luis Filipe Coelho Antunes. Entropy measures vs. Kolmogorov complexity. *Entropy*, 13(3):595–611, 2011. doi:10.3390/e13030595.

A Appendix

A.1 Proof of Theorem 1 continued

We define here the second upper bound formula $\psi(\mathfrak{M})$ of size at most $6|\pi_{\ell-1}| + c_\tau$, along with required subformulas.

Let T , \bar{m} and \mathfrak{M} be as in the proof so far. We define another formula $\chi(T, \bar{m})$ below. Now the model \mathfrak{M} satisfies $\chi(T, \bar{m})$ if and only if for every $i \in \{1, \dots, r\}$, the model \mathfrak{M} has at most m_i points that realize the type π_i . We again do not assert anything about the types π_j with no corresponding m_j .

17:18 Description Complexity in FO with Links to Entropy

$$\theta_{m_r} := y = x_{m_r} \vee \bigvee_{\substack{j \in \{1, \dots, r\} \\ m_j = m_r}} (\pi_j(x_1) \wedge \neg \pi_j(y))$$

$\theta_i := y = x_i \vee \theta_{i+1}$, if $m_j \neq i$ for all $j \in \{1, \dots, r\}$, and

$$\theta_i := y = x_i \vee \left(\bigvee_{\substack{j \in \{1, \dots, r\} \\ m_j = i}} (\pi_j(x_1) \wedge \neg \pi_j(y)) \vee \left(\bigwedge_{\substack{j \in \{1, \dots, r\} \\ m_j = i}} \neg \pi_j(x_1) \wedge \theta_{i+1} \right) \right), \text{ otherwise.}$$

$$\chi(T, \bar{m}) := \forall x_1 \exists x_2 \dots \exists x_{m_r} \forall y \left(\bigvee_{j \in \{r+1, \dots, \ell\}} \pi_j(x_1) \vee \theta_1 \right)$$

We again explain how the above formula works. Note that directly taking the negation of the formula $\varphi(T, \bar{m})$ would not work as we are dealing with all types at once. We instead again start with a universally quantified variable x_1 that is attached to a point realizing a type $\pi_j \in T$. We first check if π_j is one of the types we can safely ignore. Assume then that $m_j = 5$. The existentially quantified variables x_2, \dots, x_5 are then chosen to be of the same type π_j as x_1 in such a way that every point of the type π_j has at least one x_i attached to it. Since $m_j = 5$, the first step of the recursion insists that either y is the same as x_1 or the recursion continues. When the recursion arrives at θ_5 , we cannot go any further, as to continue, we would need $m_j \neq 5$. We are instead left with the two options of either $y = x_5$ or y realizes a different type than x_1 . This amounts to saying that there are no more than 5 points that realize the type π_j .

The crucial point of the formula $\chi(T, \bar{m})$ is that the first universally quantified variable x_1 allows us to use the same existential quantifiers to count all types at once. To ensure that we do not require all of the types to be the same size, we restrict the type realized by x_1 before continuing with the recursion.

We compute the size of $\chi(T, \bar{m})$. The formula has $m_r + 1$ quantifiers. For each type π , the subformula $\pi(x)$ occurs at most three times and for at least one type with $|\pi| = m_r$, only two times. This results in $3k|T| - k$ atomic formulas of the form $P(x)$ or $\neg P(x)$. For the equalities and inequalities, each equality $y = x_i$ for $1 \leq i \leq m_r$ occurs exactly once, for a total of m_r such atomic formulas. Accounting for the binary connectives, the size of $\chi(T, \bar{m})$ is thus at most

$$m_r + 1 + 2(m_r + 3k|T| - k) - 1 = 3m_r + 6k|T| - 2k.$$

Our second complete upper bound formula $\psi(\mathfrak{M})$ avoids counting the type π_ℓ with the most realizing points by bounding the size of all other types from above and from below. For this formula we denote by $\bar{m} \setminus |\pi_\ell|$ the sequence $(|\pi_1|, \dots, |\pi_{\ell-1}|)$. We define

$$\psi(\mathfrak{M}) := \bigwedge_{i=1}^{\ell} \exists x \pi_i(x) \wedge \forall x \bigvee_{i=1}^{\ell} \pi_i(x) \wedge \varphi(T, \bar{m} \setminus |\pi_\ell|) \wedge \chi(T, \bar{m} \setminus |\pi_\ell|).$$

The numbers of new quantifiers and atomic formulas are the same as for $\varphi(\mathfrak{M})$. Accounting for the binary connectives, including the one connecting $\varphi(T, \bar{m} \setminus |\pi_\ell|)$ and $\chi(T, \bar{m} \setminus |\pi_\ell|)$, the size of $\psi(\mathfrak{M})$ is now at most

$$\begin{aligned} & |T| + 1 + 2(k|T| + k|T|) + 3|\pi_{\ell-1}| + 4k|T| - 3 + 3|\pi_{\ell-1}| + 6k|T| - 2k + 1 \\ &= 6|\pi_{\ell-1}| + 14k|T| + |T| - 2k - 1 \leq 6|\pi_{\ell-1}| + c_T. \end{aligned}$$

A.2 Proof of Theorem 3

Let $\bar{m} = (m_1, \dots, m_\ell)$ be a tuple corresponding to a class of \equiv_d , ordered in the following way. The first numbers m_1, \dots, m_r are the ones greater than 0 and smaller than d in ascending order. The numbers m_{r+1}, \dots, m_ℓ are all equal to d , and finally the numbers $m_{\ell+1}, \dots, m_t$ are all equal to 0.

Using this order for the types, the set $T = \{\pi_1, \dots, \pi_\ell\}$ is now the set of types realized in models of the class and the first r types are each realized exactly $m_i < d$ times. This is in line with the notation of the formulas for full FO above.

Our first formula works for any \bar{m} . The formula states that each type π_j is realized at least m_j times and furthermore, the ones with $m_j < d$ are realized at most m_j times.

$$\varphi_d(\bar{m}) := \bigwedge_{i=1}^{\ell} \exists x \pi_i(x) \wedge \forall x \bigvee_{i=1}^{\ell} \pi_i(x) \wedge \varphi(T, (m_1, \dots, m_\ell)) \wedge \chi(T, (m_1, \dots, m_r))$$

In the same way as for $\psi(\mathfrak{M})$ in the proof of Theorem 1, the size of $\varphi_d(\bar{m})$ is at most

$$\begin{aligned} & |T| + 1 + 2(k|T| + k|T|) + 3d + 4k|T| - 3 + 3m_r + 6k|T| - 2k + 1 \\ &= 3d + 3m_r + 14k|T| + |T| - 2k - 1 \leq 3d + 3m_r + c_\tau. \end{aligned}$$

Our second formula is only for the special case, where there is exactly one m_j equal to d . In this case, as with full FO, we can avoid counting the type with the most realizing points. The rest of the types π_j have $m_j < d$ and the formula states that each π_j is realized at least and at most m_j times.

$$\psi_d(\bar{m}) := \bigwedge_{i=1}^{\ell} \exists x \pi_i(x) \wedge \forall x \bigvee_{i=1}^{\ell} \pi_i(x) \wedge \varphi(T, (m_1, \dots, m_r)) \wedge \chi(T, (m_1, \dots, m_r))$$

Again in the same way as for $\psi(\mathfrak{M})$ in the proof of Theorem 1, the size of $\psi_d(\bar{m})$ is at most

$$\begin{aligned} & |T| + 1 + 2(k|T| + k|T|) + 3m_r + 4k|T| - 3 + 3m_r + 6k|T| - 2k + 1 \\ &= 6m_r + 14k|T| + |T| - 2k - 1 \leq 6m_r + c_\tau. \end{aligned}$$

The upper bounds of the claim follow.

A.3 Proof of Lemma 11

We will use Proposition 10. For every type π and $1 \leq i \leq n$ we associate a 0-1-valued random variable $X_{\pi,i}$ such that $X_{\pi,i} = 1$ with probability $2^{-|\tau|}$ and $X_{\pi,i} = 0$ with probability $1 - 2^{-|\tau|}$. Intuitively this is an indicator random variable for the event “the i th element received the type π ”. Now $X_\pi = \sum_{i=1}^n X_{\pi,i}$ is a random variable that counts the number of times π is realized. Clearly $\mathbb{E}[X_\pi] = n/2^{|\tau|}$, which also holds for every type π . Set $\mu := n/2^{|\tau|}$ and $\delta(n) := \sqrt{\frac{3}{2^{|\tau|}} \frac{\ln(n)}{n}}$. Now

$$2e^{-\delta(n)^2 \mu / 3} = 2n^{-1}$$

and

$$\delta(n)\mu = \frac{\sqrt{3}}{2^{|\tau|} \sqrt{2^{|\tau|}}} \sqrt{\ln(n)n}.$$

17:20 Description Complexity in FO with Links to Entropy

Thus, by Proposition 10, we know that

$$\Pr \left[|X_\pi - \mu| \geq \frac{\sqrt{3}}{2^{|\tau|}\sqrt{2^{|\tau|}}} \sqrt{\ln(n)n} \right] \leq 2n^{-1}$$

Applying the union bound, we also see that

$$\begin{aligned} & \Pr \left(\exists \pi : |X_\pi - \mu| \geq \frac{\sqrt{3}}{2^{|\tau|}\sqrt{2^{|\tau|}}} \sqrt{\ln(n)n} \right) \\ & \leq \sum_{\pi} \Pr \left[|X_\pi - \mu| \geq \frac{\sqrt{3}}{2^{|\tau|}\sqrt{2^{|\tau|}}} \sqrt{\ln(n)n} \right] \\ & \leq 2^{|\tau|+1} n^{-1} \end{aligned}$$

Thus, with probability at least $1 - 2^{|\tau|+1}/n$ in a random model \mathfrak{M} of size n we have for every type π that

$$\left| |\pi|_{\mathfrak{M}} - \frac{n}{2^{|\tau|}} \right| \leq \frac{\sqrt{3}}{2^{|\tau|}\sqrt{2^{|\tau|}}} \sqrt{\ln(n)n}.$$

Hence, with probability at least $1 - 2^{|\tau|+1}/n$ a random model of size n is balanced.