Probabilistic Characterisation of Models of First-Order Theories

Soroush Rafiee Rad*

Abstract

We study probabilistic characterisation of a random model of a finite set of first order axioms. Given a set of first order axioms $\mathcal{T}$ and a structure $\mathcal{M}$ which we only know is a model of $\mathcal{T}$, we are interested in the probability that $\mathcal{M}$ would satisfy a sentence ψ. Answering this question for all sentences in the language will give a probability distribution over the set of sentences which can be regarded as the probabilistic characterisation of the model $\mathcal{M}$. We investigate defining these probabilistic characterisations as the limit of probability functions imposed on the set of finite models of $\mathcal{T}$. We show how a symmetry axiom can uniquely specify the probability function over finite models and will study the existence of the limit in terms of the quantifier complexity of $\mathcal{T}$.

keywords: probabilistic models, first order theories, renaming principle, probabilistic logic

1 Introduction

Let L be a first order language, SL the set of sentences of L and $\mathcal{T} \subset SL$ a finite consistent set of sentences. We are interested in the extent to which such a set of sentences $\mathcal{T}$ can characterise a model over L. To be more precise the question is

Given a finite consistent set of sentences $\mathcal{T}$ of first order axioms, what should we take as the default or most normal model of $\mathcal{T}$?

The first thing to clarify before one can answer this question is how to interpret the *normality* requirement. There are indeed different ways that one can understand this. In a model theoretic view, for example, one can expect the most normal model

*Department of Philosophy, Bayreuth University – Soroush.R.Rad@gmail.com

to be the smallest and canonical model, thus interpreting the most normal model as a prime model (see for example [9], [17]). Or one might require some closure conditions from such models and require them to be existentially closed. Another approach is to understand the most normal as the 'average' model and investigate this question by looking at the distribution of models (see for example [1], [2], [12], [14] and [15]). Here we take a different approach and look at the question

Given a finite (consistent) set $\mathcal{T}$ of first order axioms, from a language L and a structure $\mathcal{M}$ with domain $\{a_1, a_2, \ldots\}$ over L which we only know to be a model of $\mathcal{T}$, what probability should we assign to a sentence $\theta(a_1, \ldots, a_n)$ being true in $\mathcal{M}$?

In other words we are interested in how a set of first order axioms can probabilistically characterise a random model in the *most normal* or natural way. In this sense, $\mathcal{T}$ imposes a probability assignment on the set of sentences, SL, assigning probability 1 to each $\phi \in \mathcal{T}$. We will call such probability assignments *Probabilistic Models* of $\mathcal{T}$. The probability assigned to each sentence ψ is understood as the probability that a random model of $\mathcal{T}$ will satisfy ψ. Notice that the only constraint imposed here is that $\mathcal{M}$ is a model for $\mathcal{T}$, which ensures that the probability assignment should give probability 1 to all sentences in $\mathcal{T}$. This leaves a lot of freedom for choosing the assignment of probabilities to other sentences, and different ways of making this choice will capture different structural properties that one imposes on the way that $\mathcal{T}$ should characterise $\mathcal{M}$ or, in other words, how one interprets the *normality* requirement for $\mathcal{M}$. For example one such property that has attracted a lot of attention in the literature is to require $\mathcal{T}$ to impose the least informative of such probability assignments, called the Maximum Entropy model of $\mathcal{T}$. This probability assignment is understood as a probabilistic description of $\mathcal{M}$ to the extent that it is characterised by $\mathcal{T}$ while remaining maximally unconstrained beyond that. In this case, the condition of normality is interpreted as being minimally constrained. Other approaches to make this assignment of probabilities will capture different notions of normality such as *averageness* or *typicality*[1], etc. The mapping that assigns to each finite set of axioms (or more generally, to each set of consistent) one such probability assignment over sentences of the language is called an *Inference Process*. Inference processes are of interest in many areas and a wide range of them have been proposed and studied in the literature. The most extensively studied amongst which is arguably the Maximum Entropy inference process mentioned above, which is of interest in several disciplines; from statistics [18, 19], physics [22], statistical mechanics and thermo-dynamics [24] to economics and finance [20, 42], and more recently from computer science [10, 11, 7] to formal epistemology, Bayesian

[1]Centre of Mass or Minimal Distance models for example which we shall refer to later on.

inference [21, 16, 28] and belief formation [31, 32, 39, 41], see [23] for an extensive list of examples of Maximum Entropy application in science and engineering. This is, however, by no means the only one that is of interest in the literature. Many other examples such as Centre of Mass, Minimum Distance [28] as well as a spectrum of other such assignments given by generalised Renyi Entropies [36] have been extensively studied and employed in different contexts, along with other approaches for deriving probability distribution over the set of models, especially in computer science and data base theory [8].

The main result in this paper is a unified analysis of these inference processes in terms of a structural property, referred to as the Renaming Principle (RP). Thus, we will not be dealing with any specific inference process. Instead, we will investigate the generalisation of a class of inference processes, characterised by this property to first order languages. Indeed, we show that RP, that is satisfied by a wide range of inference processes including the ones just mentioned, uniquely characterises the probability distribution in the context that we are interested in, i.e., when investigating models of a set of axioms that hold categorically. So, our analysis here will show that the only property of these widely different inference processes, investigated and employed in rather different contexts, that is really relevant is the Renaming Principle and the differences between them, that are numerous, are by and large irrelevant in the context of characterising categorical constraints. This is important because indeed the Renaming Principle seems a very natural condition to impose on inference processes. Although we will not deal with justification of this principle in this paper, it will become clear immediately that, at least in the context of the question we asked above, violation of RP is much more in need of justification than its satisfaction.

In what follows we shall investigate two things: first we will show that any two inference processes that satisfy Renaming Principle will be equivalent in characterising a set of first order axioms. That is, if one such inference process is well-defined on a first order language, so are all, and moreover, when they are well-defined, they agree. Next we will investigate when such an inference process is guaranteed to be well-defined for a first order language in terms of the quantifier complexity of the set of axioms we wish to consider. There have been relatively recent studies in generalising specific inference processes, which have been studied rather extensively for propositional languages, to first order case; in particular the Maximum Entropy, see for example [6], [25], [26], [27], [30], [34], [40], [41]. For the second part of our analysis, we give a survey of these results, some of which we have previously only hinted at without proofs or full analysis. In this sense our survey here will put previous results in a more general light and will show the existence, or lack thereof, of a well-defined inference process satisfying the Renaming Principle, for sets of axioms of specific complexities. We will also point out how in certain cases, namely for unary first order languages, the symmetry requirement imposed by the RP will capture some model theoretic interpretation of normality such as existentially closeness. The results which we will review, and give full detail of, in this paper settle all but one case for the complexity of the set of first order axioms $\mathcal{T}$. One of the main goals of the second part of our analysis is to put these results and the detail of their analysis together in the hope that it will facilitate the search for an answer to this only case that still remains open.

2 Preliminaries and Notation

Throughout this paper, we will work with a first order language L with finitely many relation symbols, no function symbols and countably many constant symbols $a_1, a_2, a_3, ...$ which we assume to exhaust the universe. Let RL and SL denote the sets of relation symbols and the set of sentences of L respectively, and let a *term model* for L be a structure M for the language L with domain $M = \{ a_i \,|\, i = 1, 2, ...\}$ where every constant symbol is interpreted as itself.

Definition 1. A probability function[2] on SL is a function $w : SL \to [0\,,\,1]$ such that for every $\theta, \phi, \exists x\psi(x) \in SL$,

- P1. If $\models \theta$ then $w(\theta) = 1$.
- P2. $w(\theta \vee \phi) = w(\theta) + w(\phi) - w(\theta \wedge \phi)$.
- P3. $w(\exists x\psi(x)) = \lim_{n\to\infty} w(\bigvee_{i=1}^{n} \psi(a_i))$.

Let us first briefly consider a propositional language L_{Prop} in order to lay the grounds for the first order case.

Definition 2. Let L_{Prop} be a propositional language with propositional variables $p_1, p_2, ..., p_n$. By *atoms* of L_{Prop} we mean the set of sentences $\{ \alpha_i \,|\, i = 1, ...J\}$, $J = 2^n$ of the form

$$\pm p_1 \wedge \pm p_2 \wedge ... \wedge \pm p_n.$$

For every sentence $\phi \in SL_{Prop}$ there is unique set $\Gamma_\phi \subseteq \{ \alpha_i | \, i = 1, ..., J \}$ such that $\models \phi \leftrightarrow \bigvee_{\alpha_i \in \Gamma_\phi} \alpha_i$. Since the α_i's are mutually inconsistent, for every probability

[2]To see how this connects to the standard definition of probability measures in measure theory, notice that we can identify sentences of L with complex events in the probability space $(\Omega, \mathcal{P}(\Omega), w)$ where Ω is the set of all term models for L. The axioms P1-P2 will then correspond to standards probability axioms for w.

function w

$$w(\phi) = w(\bigvee_{\alpha_i \models \phi} \alpha_i) = \sum_{\alpha_i \models \phi} w(\alpha_i).$$

On the other hand since $\models \bigvee_{i=1}^{J} \alpha_i$, we have $\sum_{i=1}^{J} w(\alpha_i) = 1$. So the probability function w will be uniquely determined by its values on the atoms α_i's, that is by the vector $\vec{w} = (w(\alpha_1), ..., w(\alpha_J)) \in \mathbb{D}^{L_{Prop}}$ where $\mathbb{D}^{\mathbb{L}_{\mathbb{P}\searrow\rtimes\imath}} = \{\, \vec{x} \in \mathbb{R}^J \,|\, \vec{x} \geq 0,\ \sum_{i=1}^{J} x_i = 1\}$. Conversely if $\vec{a} \in \mathbb{D}^{L_{Prop}}$, we can define a probability function $w : SL_{Prop} \to [0\,,\,1]$ such that $(w(\alpha_1), ..., w(\alpha_J)) = \vec{a}$ by setting $w(\phi) = \sum_{\alpha_i \models \phi} a_i$. This gives a one to one correspondence between the probability functions on SL_{Prop} and the points in $\mathbb{D}^{L_{Prop}f}$.

Let $\mathcal{T} = \{\phi_1, \ldots, \phi_n\} \subseteq SL_{Prop}$ be a consistent set of sentences. We are interested in the probabilistic assignments on the set of sentences of the language induced by $\mathcal{T}$, i.e. the probability functions on SL_{Prop} that assign probability 1 to each sentence $\phi \in \mathcal{T}$. In this sense, each such $\mathcal{T}$ imposes a constraint set $C_{\mathcal{T}} = \{w(\phi_1) = 1, \ldots w(\phi_n) = 1\}$ and we are interested in probability functions w on SL_{Prop} that satisfy $C_{\mathcal{T}}$. We shall call these probability assignments *probabilistic models of* $\mathcal{T}$. We are, in particular, interested in investigating systematic ways of picking one such assignment in a way that captures a (possibly context dependent) notion of *normality*.

Replacing each $w(\phi_j)$ in $C_{\mathcal{T}}$ with $\sum_{\alpha_i \models \phi_j} w(\alpha_i)$ and adding the equation $\sum_{i=1}^{J} w(\alpha_i) = 1$ we will get a system of linear equations $(w(\alpha_1), ..., w(\alpha_J))A_{\mathcal{T}} = \vec{1}$. If the probability function w satisfies $C_{\mathcal{T}}$, the vector $(w(\alpha_1), ..., w(\alpha_J))$ will be a solution for the equation $\vec{x}A_{\mathcal{T}} = \vec{1}$. We will denote the set of non-negative solutions to this equation by

$$V^{L_{Prop}}(C_{\mathcal{T}}) = \{\, \vec{x} \in \mathbb{R}^J \,|\vec{x} \geq 0,\ \ \vec{x}A_{\mathcal{T}} = \vec{1}\,\} \subseteq \mathbb{D}^{L_{Prop}}.$$

Thus the set of probabilistic models of $\mathcal{T}$ will be in a one to one correspondence with the set $V^{L_{Prop}}(C_{\mathcal{T}})$.

We talked of a systematic way to pick for each $\mathcal{T}$ a probability function that satisfies $C_{\mathcal{T}}$. This is made precise in the notion of an *Inference Process*. Let L_{Prop} be a propositional language.

Definition 3. Let $\mathbb{P}$ be the set of probability functions on SL_{Prop} and CL_{Prop} be the set of sets of linear constraints of the form $\{\sum_{j=1}^{m} a_{ij} w(\phi_j) = b_i\}$. An inference process is a function $N : CL_{Prop} \to \mathbb{P}$ that for every set of linear constraints $C_i \in CL_{Prop}$ picks a probability function $w_i \in \mathbb{P}$ that satisfies C_i, or equivalently a point in $V^{L_{Prop}}(C_i)$.

Put simply, an inference process is a mapping, N, that assigns to each set of constraints C a model $N(C)$, that is a probability function that satisfies the constraints given in C. Notice that constraints set imposed by a set of sentences $\mathcal{T}$ is a set of linear constraints.

Example 1. Let w be a probability function on SL_{Prop}. The Shannon Entropy and the Centre of Mass Infinity of w are defined as, see [28],

$$E(w) = -\sum_{i=1}^{J} w(\alpha_i)\log(w(\alpha_i)) = -\sum_{i=1}^{J} w_i \log(w_i), \textit{and}$$

$$CM_\infty(w) = \sum_{i=1}^{J} \log(w(\alpha_i)) = \sum_{i=1}^{J} \log(w_i).$$

Let $\mathcal{T} \subset SL_{Prop}$ be a finite set of sentences.

- The Maximum Entropy inference process, assigns to each set of sentences $\mathcal{T}$ the Maximum Entropy model of $\mathcal{T}$, $ME(\mathcal{T})$ that is the probability function that satisfies $C_{\mathcal{T}}$ and for which $E(w)$ is maximal.

- The Centre of Mass Infinity inference process, assigns to each $\mathcal{T} \subset SL_{Prop}$ (or more precisely to each $C_{\mathcal{T}}$) the Centre of Mass model of $\mathcal{T}$, $CM_\infty(\mathcal{T})$ that is the probability function that satisfies $C_{\mathcal{T}}$ and for which $CM_\infty(w)$ is maximal.

Shannon's entropy is the most commonly accepted measure for the informational content of a probability function. To be precise the informational content of a probability function is inversely proportional to its Shannon entropy. That is, the higher the entropy of a probability function, the lower its informational content. The probability function that assigns the full probability mass (of 1) to a single atom, say α_1 for example, is maximally informative and has the lowest entropy, while the probability function that gives equal probability to all atoms is minimally informative and has the highest entropy. So the Maximum Entropy model of $\mathcal{T}$ is the most uninformative probability function that assigns probability 1 to sentences in $\mathcal{T}$.

Notice that a probability function over the set of sentences SL_{Prop} imposes a unique probability function over the models of L_{Prop}. These are exactly the atoms of L_{Prop}. In this sense the Maximum Entropy model of $\mathcal{T}$ can be regarded as the most equivocal characterisation of a random model by $\mathcal{T}$: it assigns probability 1 to the set of models satisfying $\mathcal{T}$ and beyond that remains completely equivocal amongst them. In a similar way the Centre of Mass Infinity captures the notion of *typicality* thus the Centre of Mass model of $\mathcal{T}$ is the probabilistic characterisation of the most typical (or average) model by $\mathcal{T}$. There are many other inference processes that are proposed and studied in the literature for different purposes, each capturing a different notion of normality.

Having set this up, we can now return our focus to first order languages. Although one does not have the notion of atoms for a first order language L (as they would require infinite conjunctions), the *state descriptions* for finite sub-languages will play a similar role to that of atoms in the propositional case.

Definition 4. Let L be a first order language with the set of relation symbols RL and let L^k be the sub-language of L with the domain restricted to constants $a_1, ..., a_k$. The state descriptions of L^k are defined as the sentences $\Theta^k_1, ..., \Theta^k_{n_k}$ of the form

$$\bigwedge_{\substack{i_1,...,i_j\leq k\\ R\ j\text{-ary}\\ R\in RL, j\in N^+}} \pm R(a_{i_1}, ..., a_{i_j}).$$

For a quantifier free sentence $\theta \in SL$ let k be an upper bound on the i such that a_i appears in θ. Then θ can be thought of as being from the propositional language L^k_{Prop} with propositional variables $R(a_{i_1}, ..., a_{i_j})$ for $i_1, ..., i_j \leq k$, $R \in RL$ and R j-ary[3]. To be more precise for $r \geq k$ define $(-)^r : SL^r \to SL^r_{Prop}$ as

$$\begin{aligned}
(R_j(a_{i_1}, \ldots, a_{i_n}))^r &= R_j(a_{i_1}, \ldots, a_{i_n})\\
(\neg\phi)^r &= \neg(\phi)^r\\
(\phi \vee \psi)^r &= (\phi)^r \vee (\psi)^r\\
(\phi \wedge \psi)^r &= (\phi)^r \wedge (\psi)^r\\
(\exists x\phi(x))^r &= \bigvee_{i=1}^{r} (\phi(a_i))^r.
\end{aligned}$$

For a set of sentences $\mathcal{T} \subset SL$, let $\mathcal{T}^r = \{\phi^r \,|\, \phi \in \mathcal{T}\}$ and $C^r_{\mathcal{T}}$ the set of constraints imposed by $\mathcal{T}^r$. Then $\mathcal{T}^r$ and $C^r_{\mathcal{T}}$ give the restriction of $\mathcal{T}$ to the finite fragment L^r of L, and the constraints imposed by $\mathcal{T}$ on this finite sublanguage (and the corresponding propositional language L^r_{Prop}), respectively.

The sentences Θ^k_i will be the atoms of L^k_{Prop} and for the quantifier free sentence θ, $\models \theta \leftrightarrow \bigvee_{\Theta^k_i \models \theta} \Theta^k_i$. Then since Θ^k_i's are mutually inconsistent, for every probability function w on SL, $w(\theta) = w(\bigvee_{\Theta^k_i \models \theta} \Theta^k_i) = \sum_{\Theta^k_i \models \theta} w(\Theta^k_i)$. Thus to determine w on quantifier free sentences we only need to determine the values $w(\Theta^k_i)$ (for all k), and to require

[3]Take for example the quantifier free sentence $P(a_1) \wedge \neg R(a_1, a_2)$. Here $k = 2$ and this sentence can be thought of as a sentence from the propositional language L^2_{Prop} with propositional variables $P(a_1), P(a_2), R(a_1, a_1), R(a_1, a_2), R(a_2, a_1)$ and $R(a_2, a_2)$.

$$w(\Theta_i^k) \geq 0 \text{ and } \sum_{i=1}^{n_k} w(\Theta_i^k) = 1 \tag{1}$$

$$w(\Theta_i^k) = \sum_{\Theta_j^{k+1} \vDash \Theta_i^k} w(\Theta_j^{k+1}) \tag{2}$$

to ensure that w satisfies P1 and P2. To see this assume w is given on all Θ_i^k for all k, satisfying Equations (1) and (2) and define w on all quantifier free sentences as $w(\phi) = \sum_{\Theta_i^k \vDash \phi} w(\Theta_i^k)$ where k is the largest that a_k appears in ϕ.

Notice that a sentence from the propositional language L^k_{Prop} is also a sentence in L^n_{Prop} for all $n \geq k$. Equation (2) ensures that w is consistent over increasing values of k. This gaurantees that for a entence ϕ the probability $w(\phi)$ does not depend on the choice of the language, L^k (i.e. on considering ϕ from language L^k or L^n for some $n \geq k$). To see this remember that k is the largest such that a_k appears in ϕ and notice that since ϕ is quantifier free, a state description $\Theta_j^{k+1} \vDash \phi$ if and only if it extends a state description Θ_i^k that satisfies ϕ. Then

$$\sum_{\Theta_j^{k+1} \vDash \phi} w(\Theta_j^{k+1}) = \sum_{\Theta_i^k \vDash \phi} \sum_{\Theta_j^{k+1} \vDash \Theta_i^k} w(\Theta_j^{k+1}) = \sum_{\Theta_i^k \vDash \phi} w(\Theta_i^k).$$

Now, let θ and ψ be quantifier free sentences and k be the largest such that a_k appears in either of θ or ψ. Since $\Theta_i^k \vDash \theta \vee \psi$ if and only if $\Theta_i^k \vDash \theta$ or $\Theta_i^k \vDash \psi$, we have $w(\theta \vee \psi) = \sum_{\Theta_i^k \vDash \theta \vee \psi} w(\Theta_i^k) = \sum_{\Theta_i^k \vDash \theta} w(\Theta_i^k) + \sum_{\Theta_i^k \vDash \psi} w(\Theta_i^k) - \sum_{\Theta_i^k \vDash \theta \wedge \psi} w(\Theta_i^k) = w(\theta) + w(\psi) - w(\theta \wedge \psi)$, which gives P2. And, if $\vDash \theta$ then by Equation (1), $w(\theta) = \sum_{\Theta_i^k \vDash \theta} w(\Theta_i^k) = \sum_{i=1}^{n_k} w(\Theta_i^k) = 1$ which gives P1.

Thus specifying the probabilities of all state descriptions and ensuring Equations (1) and (2) will determine a probability function (i.e. satisfying P1 and P2) on all qunstifier free sentences. The following theorem due to Gaifman [13] ensures that this is indeed enough to determine w on all sentences. Let $QFSL$ be the set of quantifier free sentences of L.

Theorem 1. Let $v : QFSL \to [0\,,\,1]$ satisfy P1 and P2 for $\theta, \phi \in QFSL$. Then v has a unique extension $w : SL \to [0\,,\,1]$ that satisfies P1, P2 and P3. In particular if $w : SL \to [0\,,\,1]$ satisfies P1, P2 and P3, then w is uniquely determined by its restriction to $QFSL$.

Just as a probability function on the set of sentences of a propositional language is determined by its values on the atoms, a probability function on the set of sentences of a first order language is determined by its values on the state descriptions. Although dealing with state descriptions is more complicated than working with atoms (one has to consider state descriptions of L^k for all k), they play a crucial and indispensable role in the analysis that will follow. Note that the set of state descriptions of L^k is the same as the set of term models over L^k.

Definition 5. Let $\{b_1, \ldots, b_n\} \subset \{a_1, a_2, \ldots\}$. By state descriptions of L over $\{b_1, \ldots, b_n\} \subset \{a_1, a_2, \ldots\}$ we mean sentences $\Psi(b_1, \ldots, b_n)$ of the form

$$\bigwedge_{\substack{a_{i_1},\ldots,a_{i_j} \subset \{b_1,\ldots,b_n\} \\ R \in RL,\ R\ j\text{-ary}}} \pm R(a_{i_1}, \ldots, a_{i_j}).$$

If Θ^r is a state description of L^r with $r > n$ such that $\{b_1, \ldots, b_n\} \subseteq \{a_1, \ldots, a_r\}$, we say $\Psi(b_1, \ldots, b_n)$ *is determined* by Θ^r if and only if for all $R \in RL$ and all $t_1, \ldots, t_j \in \{b_1, \ldots, b_n\}$, $\Psi(b_1, \ldots, b_n) \models R(t_1, \ldots, t_j) \iff \Theta^r \models R(t_1, \ldots, t_j)$. That is when Θ^r agrees with Ψ when restricted to $\{b_1, \ldots, b_n\}$.

Notice however that our definition of an inference process as well as the examples of inference processes given above are defined for propositional languages. Indeed to calculate $E(w)$ or $CM_\infty(w)$ we make reference to atoms of the language, which we cannot do if we move to first order languages because this would require the use of infinite conjunctions.

2.1 Inference Processes on First Order Languages

There has been extensive work on extending inference processes to first order languages, especially for the Maximum Entropy inference process, which is of interest in many different areas. This literature, however, is mostly concerned with some particular inference process and the generalisations to first order language is usually specific to that inference process. See for example [40] for Williamson's generalisation of Maximum Entropy to first order languages and [35] for a detailed analysis. Another proposal is given in [6] for an alternative generalisation of Maximum Entropy to unary first order languages, and employed in [29] to generalise Centre of Mass Infinity and Minimum Distance inference process also to unary first order languages and, a modified version of it, in [34] and [30] to investigate the extension of Maximum Entropy inference process to polyadic languages. This second approach has a more general flavour that can be adopted for any inference process defined on propositional languages. This approach (in the modified version of [30]), which we shall discuss in detail, will be the focus of this paper.

Let N be an inference process defined for propositional languages. The idea here follows from the observation that the finite sub-languages L^r can essentially be treated as propositional languages for which N is assumed to be well-defined. The proposal is then to define N for a set of first order sentences $\mathcal{T}$, $N(C_{\mathcal{T}})$, on some state descriptions Θ^n of L^n by looking at the sublanguages L^r and $N(C^r_{\mathcal{T}})$ and taking the limit as r grows. To be more precise let L be a first order language and $\mathcal{T} \subset SL$, the proposal is to define for each state description Θ^n of L^n

$$N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})(\Theta^n)$$

if the limit exists and undefined otherwise. When the limit exists for all n and Θ^n, this will define $N(C_{\mathcal{T}})$ on all state descriptions and thus on all quantifier free sentences which then extends uniquely to all SL by Gaifman theorem. The main question when working with this proposal is thus to investigate the conditions under which the limit exists for all n and state descriptions Θ^n, and when it does, that $N(C_{\mathcal{T}})$ does indeed satisfy $C_{\mathcal{T}}$. We will investigate the existence of this limit in terms of the quantifier complexity of the sentences in T.
For the Maximum Entropy inference process, in particular, the existence of this limit for any set of sentences T from a *unary* first order languages as well as for sets of Σ_1 or Π_1 sentences on arbitrary polyadic languages has been investigated by Barnett and Paris [6] and Rafiee Rad [34] and Paris and Rafiee Rad [30]. In what follows we will survey these results and put them in a more general light. Indeed, the main property used in the above mentioned analysis for the Maximum Entropy inference process is the symmetry condition of the Renaming Principle.

Definition 6. (Renaming Principle) An inference process N, defined on propositional language L_{Prop}, satisfies *Renaming Principle* if for two sets of linear constraints C_1 and C_2 of the from

$$C_1 = \{\sum_{j=1}^{J} a_{ji} w(\gamma_j) = b_i \mid i = 1, \dots m\}$$

$$C_2 = \{\sum_{j=1}^{J} a_{ji} w(\delta_j) = b_i \mid i = 1, \dots m\}$$

where $\gamma_1, \dots \gamma_J$ and $\delta_1, \dots, \delta_J$ are permutations of atoms of L_{Prop}, $\alpha_1, \dots \alpha_J$,

$$N(C_1)(\gamma_j) = N(C_2)(\delta_j),$$

for all $j = 1, \dots, J$.

Fact 1. Let L_{Prop} be a propositional language with atoms $\alpha_1, \ldots, \alpha_J$ and let N be an inference process defined on L_{Prop}. Then N satisfies Renaming Principle if for every set of linear constraint $C \in CL_{Prop}$ and permutation σ of $1, \ldots, J$ we have $\sigma(N(V^{L_{Prop}}(C))) = N(\sigma V^{L_{Prop}}(C))$.[4]

See [28, pages 97-98].
In Sections 4 and 5, we will look at unary first order languages and sets of Σ_1 sentences from any polyadic language. In Section 6, we will focus on sets of Π_1 sentences and will show the existence of the limit for any set $\mathcal{T}$ of Π_1 sentences from a unary language with equality as well as for any set $\mathcal{T}$ containing only what we shall call *slow* Π_1 sentences from arbitrary polyadic languages. Finally, in Section 7 we will show that the limit does not necessarily exist in general.

3 Renaming Principle and the Equivalence of Inference Processes on sets of First Order Axioms

The renaming Principle is satisfied, not only by the Maximum Entropy inference process but also by a wide range of other inference processes proposed and studied in the literature. The results in this paper thus cover not only the Maximum Entropy inference process but also other well studied instances such as the Centre of Mass Infinity, Minimum Distance, and more generally the spectrum of inference processes based on generalised Renyi Entropies [36], as they all satisfy RP.

Proposition 1. i. Maximum Entropy and Centre of Mass Infinity and Minimum Distance inference processes satisfy the Renaming Principle.

ii. More generally an inference process based on generalised Renyi entropies satisfies the Renaming Principle.

Proof. For (i) see [28, page 98]. For (ii) For a probability function w generalised Renyi entropeis are difined as

$$H_\delta(w) = \frac{1}{1-\delta} \sum_{i=1}^{J} w(\alpha_i)^\delta$$

for $\delta \geq 0$, $\delta \neq 1$. Let N_δ be the inference process defined in terms of maximising H_δ then it is easy to see that for any set of linear constraints $C \in CL_{Prop}$ and permutation σ of $1, \ldots, J$ we have $\sigma N_\delta(V^{L_{Prop}}(C)) = N_\delta(\sigma V^{L_{Prop}}(C))$. The result now follows using Fact 1.

□

[4] Here the inference process N is considered as a choice function on the set $V^{L_{Prop}}(C)$ directly.

Although there are inference processes that violate the Renaming Principle, it does seem as a minimally demanding condition that is very natural to impose on an inference process, at least in the context that is relevant to us.[5] Remember the goal here is to define the most normal models for a set of first order axioms, or to be more precise, to give a probabilistic characterisation of models to the extent they are specified by a set of axioms. As such, it is natural to expect that renaming the structures that satisfy the axioms should have no bearing on this probabilistic characterisation.

Proposition 2. Let N be an inference process defined on propositional languages that satisfies Renaming Principle and let $\mathcal{T} = \{\phi_1, \ldots, \phi_n\}$. If $N(C_{\mathcal{T}})$ is defined, then for a state description Θ^n of L^n

$$N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} \frac{|\{\,\Theta^r \mid {}^{\Theta^r \text{extends } \Theta^n}_{\Theta^r \text{consistent with } \mathcal{T}}\}|}{|\{\,\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} \tag{3}$$

Proof. Let k be the largest number such that a_k appears in $\mathcal{T}$ and let $r > k$. As was discussed above the finite sub-languages L^r can be treated as propositional languages and so N is defined and satisfies Renaming Principle on these languages. Notice that the constraint imposed by $C^r_{\mathcal{T}}$ is of the form

$$w(\Theta^r_{i_1}) + \ldots + w(\Theta^r_{i_n}) = 1$$

where $S = \{\Theta^r_{i_1}, \ldots, \Theta^r_{i_n}\}$ is the set of all state descriptions of L^r that are consistent with $\mathcal{T}$. Let σ be *any* permutation of state descriptions of L^r such that $\sigma(\Theta^r) \in S$ for all $\Theta^r \in S$. Let $C^r_{\mathcal{T}} = \{w(\Theta^r_{i_1}) + \ldots + w(\Theta^r_{i_n}) = 1\}$ and $C'^r_{\mathcal{T}} = \{w(\sigma(\Theta^r_{i_1})) + \ldots + w(\sigma(\Theta^r_{i_n})) = 1\}$. By Renaming Principle we have

$$N(C^r_{\mathcal{T}})(\Theta^r) = N(C'^r_{\mathcal{T}})(\sigma(\Theta^r)).$$

But the sum in $C'^r_{\mathcal{T}}$ is just a rearranging of the sum in $C^r_{\mathcal{T}}$ (by the way σ was defined). Hence we have $C'^r_{\mathcal{T}} = C^r_{\mathcal{T}}$ and so $N(C^r_{\mathcal{T}})(\phi) = N(C'^r_{\mathcal{T}})(\phi)$ for all $\phi \in SL^r$ and so for all $\Theta^r \in S$

$$N(C^r_{\mathcal{T}})(\Theta^r) = N(C'^r_{\mathcal{T}})(\sigma(\Theta^r)) = N(C^r_{\mathcal{T}})(\sigma(\Theta^r)).$$

Since σ is any permutation that respects the consistency with $\mathcal{T}$, we have that $N(C^r_{\mathcal{T}})$ assigns equal probability to all state descriptions of L^r that are consistent

[5] Notice that we defined RP for *propositional languages* and we take it as a natural constraints for inference processes on propositional languages only. As it shall become clear, in extending an inference process N that is defined on propositional languages to first order languages, we only assume RP for the application of N on propositional languages. The corresponding notion to RP for first order languages is Constant Exchangeability wich we will not deal with in this paper, please see [28] for more detailed analysis of Constant Exchangeability.

with $\mathcal{T}$.[6] By equation (2) and noticing that $N(C^r_{\mathcal{T}})$ assigns probability zero to those state descriptions of L^r that are inconsistent with $\mathcal{T}$,

$$N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} \sum_{\Theta^r \models \Theta^n} N(C^r_{\mathcal{T}})(\Theta^r)$$

$$= \lim_{r\to\infty} \sum_{\Theta^r \models \Theta^n} \frac{1}{|\{\,\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} = \lim_{r\to\infty} \frac{|\{\,\Theta^r \mid {}^{\Theta^r \text{extends } \Theta^n}_{\Theta^r \text{consistent with } \mathcal{T}}\}|}{|\{\,\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|}.$$

■

The result above guarantees that for a set of axioms $\mathcal{T}$, either $N(C_{\mathcal{T}})$ exists for all inference processes N that satisfy Renaming Principle, or does not for any. Next result which follows from this shows that when they exist, they all agree.

Corollary 1. Let N be an inference process defined on propositional languages that satisfies Renaming Principle and $\mathcal{T}$ a set of first order axioms, and assume that $N(C_{\mathcal{T}}) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})$ exists. Then for every inference process M that also satisfies the Renaming Principle,

$$N(C_{\mathcal{T}}) = M(C_{\mathcal{T}}).$$

4 Probabilistic Models of First Order Theories: unary languages

We start by looking at the simplest case where the language contains only finitely many unary predicates. We will show that in this simple case the proposal for generalising an inference process N, that satisfies RP, as the limit of its application on constraints restricted to finite sublanguages is well-defined.

Let L be a first order language with finitely many unary predicates $P_1, \ldots, P_n$ and domain $\{a_1, a_2, \ldots\}$. For the rest of this section let $\mathcal{T} \subseteq SL$ be a finite satisfiable set of sentences from L. Let $Q_1, \ldots, Q_J$ enumerate all the formulas of the form

$$\pm P_1(x) \wedge \pm P_2(x) \wedge \ldots \wedge \pm P_n(x).$$

We will call these *types* for L. Remember that L^k is the language L with domain restricted to $\{a_1, \ldots, a_k\}$ and for $k < r$, let $()^r : SL^k \to SL^r_{Prop}$ be the translation

[6]Notice that our assumption that $\mathcal{T}$ consists of sentences that hold categorically (hence probability 1) is essential. This ensures that fort each r, for any permutation σ that permutes state descriptions of L^r with non-zero probability amongst themselves, the original set of constraints C^r and the set of constraints generated by applying the permutation, C'^r, are the same and thus $N(C^r_{\mathcal{T}})(\phi) = N(C'^r_{\mathcal{T}})(\phi)$ for all $\phi \in SL^r$. For an example of a non-categorical set of constraints where the ME, CM_∞ and MD give different answers see [33, page 43-46].

from SL^k to the propositional language L^r_{Prop} from the previous section. Let Θ_i, $i = 1, \ldots, J^k$ enumerate the state descriptions of L^k, that is, the exhaustive and exclusive set of sentences of the form

$$\bigwedge_{i=1}^{k} Q_{m_i}(a_i).$$

Lemma 2. [6] Any sentence $\theta \in SL$, with k largest such that a_k appears in θ, is equivalent to a disjunction of consistent sentences $\phi_{i,\vec{\epsilon}}$ of the form

$$\Theta_i \wedge \bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon_j}$$

where $\epsilon_j \in \{0, 1\}$ and $\theta^0 = \neg\theta$, $\theta^1 = \theta$, $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_J)$ is a sequence of 0's and 1's and $\vDash \neg(\phi_{i,\vec{\epsilon}} \wedge \phi_{j,\vec{\delta}})$ when $(i, \vec{\epsilon}) \neq (j, \vec{\delta})$.

Proof. The proof is a straightforward adoption of the proof of Theorem 3.5 in [12]. ■

Let w^r be probability functions defined on SL^r, and $\theta(a_1, \ldots, a_k) \in SL$. Then for all $r > k$, $w^r(\theta^r(a_1, \ldots, a_k)) = \sum_{\Theta^r \vDash \theta(\vec{a})} w^r(\Theta^r)$ and is thus determined by the probabilities, $w^r(\Theta^r)$, of those state descriptions of L^r that are consistent with $\theta(a_1, \ldots, a_k)$. From Lemma 2 the same holds for the sentences $\phi_{i,\vec{\epsilon}}$. That is to specify $w^r(\theta^r(a_1, \ldots, a_k))$ one only needs to specify the probabilities $w^r(\phi^r_{i,\vec{\epsilon}})$ for those $\phi_{i,\vec{\epsilon}}$ that entail $\theta(a_1, \ldots, a_k)$. Notice, however, that the number of state descriptions of L^r depends on r while, for large enough r, the number of sentences $\phi_{i,\vec{\epsilon}}$ is independent of r. As we move from L^r to L^{r+1} what changes is the number of state descriptions that satisfy each $\phi^{r+1}_{i,\vec{\epsilon}}$, but the number of these sentences remain the same for all r eventually. Thus to specify $w^r(\theta(a_1, \ldots, a_k))$ we need to specify w^r on the same set of sentences (namely those $\phi_{i,\vec{\epsilon}}$ that satisfy $\theta(a_1, \ldots, a_k)$) for all r eventually. This is an advantage of working with unary languages which will be lost when moving to more expressive languages.

We now define some short-hand notations for the ease of writing which we shall use for the rest of this section. For each $\phi_{i,\vec{\epsilon}} = \Theta_i \wedge \bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon_j}$, let

$$A_i = \{\, m_j \mid j = 1, ..., k \,\}, \; P_{\vec{\epsilon}} = \{\, j \mid \epsilon_j = 1 \,\}, \; P_{i,\vec{\epsilon}} = \{\, j \mid j \in P_{\vec{\epsilon}} \textit{ and } j \notin A_i \,\}$$

then, A_i enumerates the types that are satisfied by some a_j, $j = 1, \ldots, k$, in Θ_i, $P_{\vec{\epsilon}}$ enumerates the types that $\phi_{i,\vec{\epsilon}}$ requires to be satisfied and $P_{i,\vec{\epsilon}}$ enumerates those that are required to be satisfied but are not satisfied in Θ_i.

$$\phi^r_{i,\vec{\epsilon}} = \Theta_i \wedge \bigwedge_{j=1}^{J} (\bigvee_{i=1}^{r} Q_j(a_i))^{\epsilon_j} \equiv \bigvee_{\substack{m_j \in P_{\vec{\epsilon}} \text{for} j=k+1,\dots,r \\ P_{i,\vec{\epsilon}} \subseteq \{m_j \,|\, k+1 \le j \le r\}}} (\Theta_i \wedge \bigwedge_{j=k+1}^{r} Q_{m_j}(a_j)) \tag{4}$$

Set $p_{\vec{\epsilon}} = |P_{\vec{\epsilon}}|$, and $p_{i,\vec{\epsilon}} = |P_{i,\vec{\epsilon}}|$, then the number of disjuncts in 4, i.e. the number of state descriptions of L^r that logically imply $\phi_{i,\vec{\epsilon}}$ will be

$$n^r_{i,\vec{\epsilon}} = \sum_{j=0}^{p_{i,\vec{\epsilon}}} (-1)^j \binom{p_{i,\vec{\epsilon}}}{j} (p_{\vec{\epsilon}} - j)^{r-k}.$$

To see this notice that we need to choose $r-k$ elements from $P_{\vec{\epsilon}}$ while covering $P_{i,\vec{\epsilon}}$. This is the total number of ways we can choose $r-k$ elements from $P_{\vec{\epsilon}}$ that is $(p_{\vec{\epsilon}})^{r-k}$ minus the number of ways we can make this choice and miss at least one element in $P_{i,\vec{\epsilon}}$, that is $\sum_{j=1}^{p_{i,\vec{\epsilon}}} (-1)^{j-1} \binom{p_{i,\vec{\epsilon}}}{j} (p_{\vec{\epsilon}} - j)^{r-k}$, which gives (4).

Theorem 3. Let L be a unary first order language, $\mathcal{T} \subset SL$ a finite set of satisfiable sentences and N an inference process defined on propositional languages that satisfies the Renaming Principle. Let $\vec{\epsilon_1}, \dots, \vec{\epsilon_s}$ be all those vectors $\vec{\epsilon}$ for which $\bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon_j}$ is consistent with $\mathcal{T}$ and for which $p_{\vec{\epsilon}}$ takes its largest possible value, and let $\theta \in SL$ such that k is the largest that a_k appears in θ or $\mathcal{T}$. Then

$$N(C_{\mathcal{T}})(\theta) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})(\theta^r) = |H|/|K|, \tag{5}$$

where $K = \{\phi_{i,\vec{\epsilon^t}} \,|\, \phi_{i,\vec{\epsilon^t}}$ is consistent with $\bigwedge \mathcal{T}, 1 \le i \le J^k, 1 \le t \le s\}$ and

$$H = \{\phi_{i,\vec{\epsilon^t}} \,|\, \phi_{i,\vec{\epsilon^t}} \text{ is consistent with } \theta \wedge \bigwedge \mathcal{T}, 1 \le i \le J^k, 1 \le t \le s\}.$$

and, as above, $\phi_{i,\vec{\epsilon}} = \Theta_i \wedge \bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon_j}$ and Θ_i, $i = 1, \dots, J^k$ enumerate the state descriptions of L^k, i.e., sentences of the form $\bigwedge_{i=1}^{k} Q_{m_i}(a_i)$. Furthermore, $N(C_{\mathcal{T}})$ is a probability function on SL and satisfies $C_{\mathcal{T}}$.

Proof.
It is clear that if $N(C_{\mathcal{T}})$ exists, it is a probability function and that for all $\psi \in \mathcal{T}$, $N(C_{\mathcal{T}})(\psi) = 1$. To show (5), let

$$K' = \{\phi_{i,\vec{\epsilon}} \,|\, \phi_{i,\vec{\epsilon}} \text{ is consistent with } \bigwedge \mathcal{T}\}, \textit{and}$$

$$H' = \{\phi_{i,\vec{\epsilon}} \,|\, \phi_{i,\vec{\epsilon}} \text{ is consistent with } \theta \wedge \bigwedge \mathcal{T}\}.$$

Let Θ_i^r run through the state descriptions of L^r and let $\Gamma_{\mathcal{T}}^r$ be the set of state descriptions of L^r that are consistent with $\bigwedge \mathcal{T}$. By Proposition 2, all the state descriptions of in $\Gamma_{\mathcal{T}}^r$ will get the same probability, namely $\frac{1}{|\Gamma_{\mathcal{T}}^r|}$, by $N(C_{\mathcal{T}}^r)$. Let $n_{i,\vec{\epsilon}}^r = \sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}(p_{\vec{\epsilon}} - j)^{r-k}$ be the number of state descriptions of L^r that logically imply $\phi_{i,\vec{\epsilon}}$, as above, and notice that $(1 - \frac{j}{p_\epsilon})^{r-k} \to 0$ as $r \to \infty$ for $0 < j < p_{\vec{\epsilon}}$, so $\lim_{r\to\infty}\sum_{j=0}^{p_{i,\vec{\epsilon}}}(1 - \frac{j}{p_{\vec{\epsilon}}})^{r-k} \to 1$, and thus we have

$$\lim_{r\to\infty}\frac{n_{i,\vec{\epsilon}}^r}{(p_{\vec{\epsilon}})^{r-k}} = 1. \tag{6}$$

So

$$N(C_{\mathcal{T}}^r)(\theta^r) = \sum_{\Theta_j^r \models \theta^r} N(C_{\mathcal{T}}^r)(\Theta_j^r) = \sum_{\phi_{i,\vec{\epsilon}}\in H'}\sum_{\Theta_j^r\models\phi_{i,\vec{\epsilon}}} N(C_{\mathcal{T}}^r)(\Theta_j^r)$$
$$= \sum_{\phi_{i,\vec{\epsilon}}\in H'}\frac{n_{\phi_{i,\vec{\epsilon}}}^r}{|\Gamma_{\mathcal{T}}^r|} = \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K'} n_{\phi_{i,\vec{\epsilon}}}^r}$$

where the last equality uses the fact that $|\Gamma_{\mathcal{T}}^r| = \sum_{\phi_{i,\vec{\epsilon}}\in K'} n_{\phi_{i,\vec{\epsilon}}}^r$ which holds as both sides count the number of state descriptions of L^r that are consistent with $\bigwedge \mathcal{T}$.
Let $c_1 > c_2 > \ldots > c_t$ be the distinct values for $p_{\vec{\epsilon}}$ for the sentences in K', so we have $p_{\vec{\epsilon}} = c_1$ for all $\phi_{i,\epsilon} \in K$ (and thus for $\phi_{i,\epsilon} \in H$) and for all $\phi_{i,\epsilon} \notin K$, $p_{\vec{\epsilon}} \le c_2$. Then we will have

$$N(C_{\mathcal{T}})(\theta) = \lim_{r\to\infty} N(C_{\mathcal{T}}^r)(\theta^r) =$$
$$\lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K'} n_{\phi_{i,\vec{\epsilon}}}^r} = \lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in H'\setminus H} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n_{\phi_{i,\vec{\epsilon}}}^r} =$$
$$\lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n_{\phi_{i,\vec{\epsilon}}}^r} + \lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'\setminus H} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n_{\phi_{i,\vec{\epsilon}}}^r}.$$

Next notice that

$$\lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n_{\phi_{i,\vec{\epsilon}}}^r} \ge \lim_{r\to\infty}\frac{c_1^{r-k}|K|}{c_1^{r-k}|K| + c_2^{r-k}|K'\setminus K|} = 1$$

To see this notice that by (6) $\lim_{r\to\infty}\frac{n_{\phi_{i,\vec{\epsilon}}}^r}{(p_{\vec{\epsilon}})^{r-k}} = 1$ and that $p_{\vec{\epsilon}} = c_1$ for all $\phi_{i,\vec{\epsilon}} \in K$, $p_{\vec{\epsilon}} < c_2$ for all $\phi_{i,\vec{\epsilon}} \in K' \setminus K$ and we have $c_2 < c_1$. Thus

$$\lim_{r\to\infty}\frac{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n_{\phi_{i,\vec{\epsilon}}}^r + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n_{\phi_{i,\vec{\epsilon}}}^r} = 1. \tag{7}$$

From (7) we have

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'\setminus H} n^r_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n^r_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n^r_{\phi_{i,\vec{\epsilon}}}} = \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'\setminus H} n^r_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n^r_{\phi_{i,\vec{\epsilon}}}} \leq \lim_{r\to\infty} \frac{c_2^{r-k}|H'\setminus H|}{c_1^{r-k}|K|} = 0,$$

since $c_2 < c_1$. In consequence we get

$$N(C_{\mathcal{T}})(\theta) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})(\theta^r) = \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} n^r_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n^r_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}}\in K'\setminus K} n^r_{\phi_{i,\vec{\epsilon}}}}$$

$$= \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} n^r_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in K} n^r_{\phi_{i,\vec{\epsilon}}}} = \frac{c_1^{r-k}|H|}{c_1^{r-k}|K|} = \frac{|H|}{|K|}.$$

It is immediate to check that $N(C_{\mathcal{T}})$ defined as above satisfies $C_{\mathcal{T}}$.

■

In particular then $N(C_{\mathcal{T}})$ assigns probability 1 to $\bigvee_{i=1}^{s} \bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon^i_j}$ (and $1/s$ to each disjunct), thus exclusively favouring those structures $\mathcal{M}$ that model $\mathcal{T}$ in which as many of the Q_j as possible are satisfied, that is existentially closed models of $\mathcal{T}$. To see this remember that $\vec{\epsilon_1}, \ldots, \vec{\epsilon_s}$ were taken to be those for which $\bigwedge_{j=1}^{J} \left(\exists x Q_j(x)\right)^{\epsilon_j}$ is consistent with $\mathcal{T}$ and p_ϵ is maximal, that is those $\vec{\epsilon}$ which has maximal number of j with $\epsilon_j = 1$.
This shows that an inference process N satisfying the RP can be correctly generalised to a unary first order language as the limit of its application on finite sublanguages. What is more, for this simple case of unary languages, the Renaming Principle is enough to ensure N implies *existential closedness* . That is, $N(C_{\mathcal{T}})$ only assigns positive probability to those structures $\mathcal{M}$ that are a model of $\mathcal{T}$ and are existentially closed.

5 Probabilistic Models of Σ_1 sentences

The next case concerns a polyadic language L, and sets of axioms $\mathcal{T}$ that include only Σ_1 sentences. We will show that for all inference processes N and all such set $\mathcal{T}$, $N(C^r_{\mathcal{T}})$ converge and the limit in r satisfies $C_{\mathcal{T}}$. Indeed we show something stronger: we will show that for such N and $\mathcal{T}$, $N(C_{\mathcal{T}})$ is always obtained by an appropriate conditionalisation of the equivocator $P_=$.

Definition 7. Define the equivocator, $P_=$, as the probability function that for each k, assigns equal probability to Θ^k_i's (the state descriptions of $L^{(k)}$), i.e. the most non-committal probability function.

Notice that a set $\mathcal{T}$ of Σ_1 sentences can be written as a single Σ_1 sentence by taking the conjunction of $\mathcal{T}$. Thus without loose of generality we will focus on singleton sets of Σ_1 sentences of the form $\mathcal{T} = \{\exists x_1, \ldots, x_t \theta(\vec{a}, x_1, \ldots, x_t)\}$.

Theorem 4. Let L be a first order language, N an inference process defined for propositional languages that satisfies Renaming Principle. Let $\phi \in SL$ be the satisfiable Σ_1 sentence $\exists x_1, \ldots, x_t \theta(\vec{a}, x_1, \ldots, x_t)$, $\mathcal{T} = \{\phi\}$ and $C = \{w(\phi) = 1\}$[7] the corresponding set of constraints. For each n and state description Θ^n of L^n define.

$$N(C)(\Theta^n) = \lim_{r \to \infty} N(C^r)(\Theta^n)$$

Then $N(C)$ is well-defined and is a probability function on SL that satisfies C. Furthermore, or all $\psi \in SL$,

$$N(C)(\psi) = P_{=}(\psi \mid \Lambda)$$

where k be the largest that a_k appears in ϕ, Γ^k_ϕ is the set of state descriptions of L^k that are consistent with ϕ and $\Lambda = \bigvee \Gamma^k_\phi$.

The proof follows exactly as for the proof for Maximum Entropy inference process given in [34] as the only property of the Maximum Entropy used in that proof was the Renaming Principle. We will summarize the proof here for the sake of completeness. Let ϕ be as above and k be largest such that a_k appears in ϕ. There can be state descriptions of L^k that entail $\neg\phi$. These state descriptions obviously have no extensions to $r > k$ that are consistent with ϕ. The main idea of the proof is that for those state descriptions of L^k that *do not* entail $\neg\phi$, almost all their extensions are consistent with ϕ. The result will then follow by noticing that first, each state description of L^k will have the same number of extensions to L^r, $r > k$ and second, by Renaming Principle all these extensions have the same probability. Lemma 5 makes the main observation precise, but first we introduce some short-hand notations following [34]:
Let Γ^r be the set of state descriptions of L^r and Γ^r_C be the subset of Γ^r that are consistent with ϕ^r. For $\Theta^l_i \in \Gamma^l$ and $r > l$ let $\Gamma^r_{l,i} = \{\Psi^r_j \in \Gamma^r \mid \Psi^r_j \vDash \Theta^l_i\}$ be the set of state description of L^r that extend the state description Θ^l_i of L^l and ${}^C\Gamma^r_{l,i} \subseteq \Gamma^r_{l,i}$ those of which that satisfy C^r, that is ${}^C\Gamma^r_{l,i} = \Gamma^r_C \cap \Gamma^r_{l,i}$. State descriptions of L^l will all have the same number of extensions to state descriptions of L^{l+1} thus $|\Gamma^r_{l,i}| = |\Gamma^r_{l,j}|$ for $\Theta^l_i, \Theta^l_j \in \Gamma^l$. Take Γ^k_ϕ as the set of state descriptions of L^k that are consistent with ϕ, and let $\Gamma^k_{\neg\phi} = \Gamma^k - \Gamma^k_\phi$.

[7] we eliminated the subscript $\mathcal{T}$ from $C_{\mathcal{T}}$ for ease of notation.

Lemma 5. If Θ_i^l is a state description of L^l that extends some state description in Γ_ϕ^k then

$$\lim_{r\to\infty} \frac{|^C\Gamma^r_{l,i}|}{|\Gamma^r_{l,i}|} = 1.$$

In other words, if Θ_i^l extends a state description of L^k that is consistent with ϕ then almost all its extensions to a state description of L^r will also be consistent with ϕ.

Proof.

Notice that $|\frac{^C\Gamma^r_{l,i}}{\Gamma^r_{l,i}}|$ is the probability that a random extension of the state description $\Theta_i^l \in \Gamma^l$ to L^r will satisfy C^r.[8] Remember that Θ_i^l extends a state description in Γ_ϕ^k, say Ψ^k. We can now calculate this probability. Take $\Theta_i^l \in \Gamma^l$ and let's consider its extensions to state descriptions of L^{l+t}, remembering that t is the number of variables in ϕ. Let $L^{a_{i_1},\dots a_{i_n}}$ be language L with only constant symbols $a_{i_1},...,a_{i_n}$ and let Δ_i, $i = 1,...,M$ enumerate the state descriptions of $L^{\{a_1,...,a_k\}\cup\{a_{l+1},...,a_{l+t}\}}$ that extend Ψ^k (thus they agree with Θ_i^l when restricted to $a_1,\dots,a_k$) . Then state descriptions of L^{l+t} that are extension of Θ_i^l can be written in the form $\Theta^{l+t}_{i,m} \equiv \Theta_i^l \wedge \Delta_j \wedge V_h(a_1,...,a_{l+t})$,[9] with $m = 1,...,|\Gamma^{l+t}_{l,i}|$, $j = 1,...,M$, and $h = 1,...,\frac{|\Gamma^{l+t}_{l,i}|}{M}$. At least one of the Δ_j's satisfies $\theta(\vec{a},a_{l+1},...,a_{l+t})$ and will hence satisfies C^{l+t}. The probability that an arbitrary $\Theta^{l+t}_{i,m}$ satisfies C^{l+t} will be the number of $\Theta^{l+t}_{i,m}$'s that satisfies C^{l+t} divided by the total number of $\Theta^{l+t}_{i,m}$'s that is *at least*, $\frac{|\Gamma^{l+t}_{l,i}|}{M}.\frac{1}{|\Gamma^{l+t}_{l,i}|} = \frac{1}{M}$, and so the probability that a random $\Theta^{l+t}_{i,m}$ does not satisfy C^{l+t} will be *at most* $1-\frac{1}{M}$. Now consider the extension of Θ_i^l to a state description of L^{l+pt},

$$\Theta^{l+pt}_{i,m} \equiv \Theta_i^l \wedge \Delta^1_{j_1} \wedge \Delta^2_{j_2} \wedge ... \wedge \Delta^p_{j_p} \wedge V'_h(a_1,...,a_{l+pt})$$

with $m = 1,...,|\Gamma^{l+pt}_{k,i}|$, $j_1,...,j_p \in \{1,...,M\}$, $h = 1,...,\frac{|\Gamma^{l+pt}_{l,i}|}{M^p}$ and where Δ^s_j enumerate the state description of $L^{\{a_1,...,a_k\}\cup\{a_{l+(s-1)t+1},...,a_{l+st}\}}$ that extend Ψ^k. The probability that $\Theta^{l+pt}_{i,m}$ does not satisfy C^{l+pt} is *at most* as high as the probability that $\Delta^1_j \nvDash \theta(\vec{a},a_{l+1},...,a_{l+t}),...,\Delta^p_j \nvDash \theta(\vec{a},a_{l+(p-1)t+1},...,a_{l+pt})$ so $0 \le 1 - \frac{|^C\Gamma^{l+pt}_{l,i}|}{|\Gamma^{l+pt}_{l,i}|} \le (1-\frac{1}{M})^p$. Let $p \to \infty$, then $0 \le \lim_{r\to\infty} 1 - \frac{|^C\Gamma^r_{l,i}|}{|\Gamma^r_{l,i}|} \le \lim_{p\to\infty}(1-\frac{1}{M})^p = 0$. Hence, we have $\lim_{r\to\infty} 1 - \frac{|^C\Gamma^r_{l,i}|}{|\Gamma^r_{l,i}|} = 0$ and $\lim_{r\to\infty} \frac{|^C\Gamma^r_{l,i}|}{|\Gamma^r_{l,i}|} = 1$ as required. ■

[8]The denominator is the total number of extensions of $\Theta_i^l \in \Gamma^l$ to L^r and the numerator is the number of those extensions of $\Theta_i^l \in \Gamma^l$ to L^r that satisfy C^r.

[9]$V_h(a_1,...,a_{l+t})$ enumerate sentence of the form $\bigwedge_{\substack{i_1,...,i_j \le l+t \\ R\in RL\, j\text{-ary}}} R_i(a_{i_1},\dots,a_{i_j})^{\epsilon_{i_1,...,i_j}}$ where $\{a_{i_1},\dots,a_{i_j}\}$ intersects both $\{a_{k+1},\dots a_l\}$ and $\{a_{l+1},\dots,a_{l+t}\}$.

The proof of Theorem 4 now follows using Lemma 5. To see this notice that since $N(C^r)$ satisfies ϕ^r it should assign all probability mass to those state descriptions that extend state descriptions in Λ. Then by Lemma 5 if Θ^n is consistent with ϕ then almost all its extensions to L^r, $r > n$ will be consistent with ϕ. By Renaming Principle all these extensions will have the same probability. So for each n, $N(C)(\Theta^n) = P_=(\Theta^n|\Lambda)$. Then $N(C)$ is clearly a probability function and the fact that it satisfies C follows from the fact that $P_=(\exists\vec{x}\theta(\vec{a},\vec{x}) \mid \Lambda) = 1$, see [34, Lemma 2] for this and more detail on the proof.
In particular then, for a set of Σ_1 sentences $\mathcal{T}$, if k is the largest that a_k appears in $\mathcal{T}$ and all state descriptions of L^k are consistent with $\bigwedge \mathcal{T}$, then $N(C_{\mathcal{T}}) = P_=$ and this is so for *any* such $\mathcal{T}$. We will now move to the more problematic case of Π_1 theories.

6 Probabilistic Models of Π_1 sentences

The next case is for sets of sentences $\mathcal{T}$ consisting of Π_1 sentences. Following as in the previous sections we ask whether $N(C^r_{\mathcal{T}})$ converges as $r \to \infty$ for a set $\mathcal{T}$ consisting of Π_1 sentences. We have already seen that this holds if L is a unary language as our result for that case does not depend on the quantifier complexity of sentences in $\mathcal{T}$. We conjecture that this is the case for any finite predicate language L, without function symbols and whose only constant symbols are $a_1, a_2, \ldots$, though our results to date fall short of proving that. This is the only case of this analysis that still remains open. Nevertheless we will show this for two special cases below: first we will show this for a unary languages *with equality*. It should be noted that the approach of Section 4 can be directly adopted to this case but we will give an alternative proof here which we shall also use for the second special case we will consider. That is for a polyadic language L when $\mathcal{T}$ consists only of what we shall call *slow* Π_1 sentences. For this second case we will give the full details of the analysis given briefly in [30] for the maximum entropy inference process, which can be immediately adopted for any inference process satisfying the Renaming Principle.
To make clear what 'equality' means in this context we require that our probability functions give probability 1 to the axioms of equality and probability 0 to $a_i = a_j$ for $i \neq j$.

6.1 Π_1 sentences from Unary Languages with Equality

Let $\mathcal{T}$ be a set of Π_1 sentences and L from a unary first order language L with equality and with predicate symbols $P_1, \ldots, P_m$. As before we notice that a set of

Π_1 sentences $\mathcal{T}$ can be written as a single Π_1 sentence by taking the conjunction of $\mathcal{T}$. Thus without loose of generality we will focus on singleton sets of sentences of the form $\mathcal{T} = \{\forall x_1, \ldots, x_t \theta(x_1, \ldots, x_t)\}$.
Let $Q_1, \ldots, Q_J$ enumerate formulas of the form

$$\pm P_1(x) \wedge \pm P_2(x) \wedge \ldots \wedge \pm P_m(x)$$

which as before we shall call the types for L *with equality removed*. Let $n \gg k \geq t$. Given a state description Θ^n of L^n, let M_Θ be the unique structure for L with universe $\{a_1, \ldots, a_n\}$ specified by Θ^n. Say that Θ^n is of *sort* κ, where $\kappa : \{1, \ldots, J\} \to \{0, 1, \ldots, t\}$, if for $1 \leq i \leq J$,

$$\kappa(i) = \min\{|\{ j \,|\, \Theta^n \models Q_i(a_j) \}|, t\}.$$

The types Q_i describe different ways that constants can behave, in other words different kinds of constants that we can have. The sort κ for the state description Θ then gives information about how many constants of each kind there are in Θ. To be more precise, the sort of each state description Θ determines, for each kind Q_i, if there are at least t constants of type Q_i in Θ and if not how many constants of this type are in Θ exactly.

Lemma 6. Suppose that $\phi(x_1, \ldots, x_k)$ is quantifier free and Θ_1^n, Θ_2^n are state descriptions with the same sort. Then

$$M_{\Theta_1} \models \forall x_1, \ldots, x_k \, \phi(x_1, \ldots, x_k) \iff M_{\Theta_2} \models \forall x_1, \ldots, x_k \, \phi(x_1, \ldots, x_k).$$

Proof. Suppose $M_{\Theta_1} \models \forall x_1, \ldots, x_k \, \phi(x_1, \ldots, x_k)$ but $M_{\Theta_2} \not\models \forall x_1, \ldots, x_k \, \phi(x_1, \ldots, x_k)$. This means that there are $a_{i_1}, \ldots, a_{i_k}$ such that $M_{\Theta_2} \models \neg\phi(a_{i_1}, \ldots, a_{i_k})$ and suppose that

$$M_{\Theta_2} \models Q_{i_j}(a_{i_j}).$$

Since $\Theta_1(a_1, \ldots, a_n)$ and $\Theta_2(a_1, \ldots, a_n)$ are state descriptions with the same sort we should have $a_{t_1}, \ldots, a_{t_k}$ such that

$$M_{\Theta_1} \models Q_{i_j}(a_{t_j})$$

and such that whenever $a_{i_j} \neq a_{i_k}$, $a_{t_j} \neq a_{t_k}$.
Thus $M_{\Theta_1} \models \neg\phi(a_{t_1}, \ldots, a_{t_k})$ and so $M_{\Theta_1} \not\models \forall x_1, \ldots, x_k \, \phi(x_1, \ldots, x_k)$ that is a contradiction. The other direction of the proof will be similar. ∎

Theorem 7. Let L be a unary first order language with equality, ϕ a Π_1 sentences in L of the form $\forall x_1,\dots,x_t\theta(x_1,\dots,x_t)$, $C=\{w(\phi)=1\}$ and N an inference process defined on propositional languages that satisfies the Renaming Principle. For every n, and state description Θ^n of L^n let

$$N(C)(\Theta^n)=\lim_{r\to\infty}N(C)(\Theta^n)$$

then $N(C)$ is well-defined. Furthermore $N(C)$ is a probability function on SL and satisfies C.

Proof.
If $N(C)$ is well-defined then it satisfies (1) and (2), as it is a limit of probability functions and thus by Theorem 1 uniquely extends to a probability function on SL. To see that it satisfies C notice that

$$N(C)(\phi)=\lim_{n\to\infty}N(C)(\bigwedge_{i_1,\dots,i_t\in\{1,\dots,n\}}\theta(a_{i_1},\dots,a_{i_t}))$$

$$=\lim_{n\to\infty}\lim_{r\to\infty}N(C^r)(\bigwedge_{i_1,\dots,i_t\in\{1,\dots,n\}}\theta(a_{i_1},\dots,a_{i_t}))=\lim_{n\to\infty}1=1$$

where the equality before last is given by the fact that $N(C^r)$ satisfies $\bigwedge_{i=1,\dots,t\in\{1,\dots,r\}}\theta(a_{i_1},\dots,a_{i_t})$ and hence for $r>n$ $N(C^r)$ satisfies $\bigwedge_{i=1,\dots,t\in\{1,\dots,n\}}\theta(a_{i_1},\dots,a_{i_t})$.
To see that $N(C)$ is well-defined, let $\Theta^n_1,\dots,\Theta^n_M$ be all the state descriptions of L^n consistent with $\forall\vec{x}\theta(x_1,\dots,x_t)$. Let $\kappa_1,\dots,\kappa_R$ be the *distinct* sorts appearing where the ordering has been chosen so that if $\kappa_i(m)\le\kappa_j(m)$ for all $1\le m\le J$, then $j\le i$.[10]
Given a state description Θ^n consistent with $\forall x_1,\dots,x_t\,\theta(x_1,\dots,x_t)$ and of sort κ_g let b_{gh} be the number of state descriptions of L^{n+1} of sort κ_h extending Θ^n and consistent with $\forall x_1,\dots,x_t\,\Theta(x_1,\dots,x_t)$.[11] These $\langle b_{gh}\rangle$ form a lower triangular matrix B and if we start from a state description Θ^n of sort κ_i the number of state descriptions of L^{n+r} of sort $\kappa_1,\kappa_2,...,\kappa_R$ is given by $(B^r)^T\vec{e_i}$ where $\vec{e_i}$ is the column vector with 1 in i-th place and zero elsewhere and $(B^r)^T$ is the transpose of the matrix $B^r=\Pi^r_{j=1}B$.
For Θ^n a state description of sort κ_i consistent with $\forall x_1,\dots,x_t\,\theta(x_1,\dots,x_t)$ the number of state descriptions of L^{n+r} extending it and still consistent with $\forall\vec{x}\,\theta(x_1,\dots,x_t)$ is

$$\langle 1,1,\dots,1\rangle(B^r)^T\vec{e}_i.$$

[10] Notice the reversing of the inequalities here.
[11] Notice that provided n is large this number does not depend on n.

Similarly the total number of state descriptions of L^{n+r} consistent with $\forall x_1, \ldots, x_t$ $\theta(x_1, \ldots, x_t)$ is

$$\sum_{j=1}^{R} N_j \langle 1, 1, \ldots, 1 \rangle (B^r)^T \vec{e}_j$$

where N_j is the number of state description of L^n of sort κ_j.
By Renaming Principle $N(C)$ will give each of these the same probability, namely

$$\left(\sum_{j=1}^{R} N_j \langle 1, 1, \ldots, 1 \rangle (B^r)^T \vec{e}_j \right)^{-1}$$

thus

$$N(C^{n+r})(\Theta^n) = \frac{\langle 1, 1, \ldots, 1 \rangle (B^r)^T \vec{e}_i}{\sum_{j=1}^{R} N_j \langle 1, 1, \ldots, 1 \rangle (B)^{r\,T} \vec{e}_j}$$

and

$$N(C)(\Theta^n) = \lim_{r \to \infty} N(C^r)(\Theta^n) = \lim_{r \to \infty} \frac{\langle 1, 1, \ldots, 1 \rangle (B^r)^T \vec{e}_i}{\sum_{j=1}^{R} N_j \langle 1, 1, \ldots, 1 \rangle (B^r)^T \vec{e}_j} \tag{8}$$

Thus to complete the proof it is enough to show that the limit in the RHS of (8) exists. Notice that to show that the limit in (8) exists it would be enough to show that for each $h \neq i$, $\lim_{r \to \infty} \frac{\langle 1,1,\ldots,1 \rangle (B^r)^T \vec{e}_i}{\langle 1,1,\ldots,1 \rangle (B^r)^T (\vec{e}_i + \vec{e}_h)}$ exists.

Claim 1.

$$\lim_{r \to \infty} \frac{\langle 1, 1, ..., 1 \rangle (B^r)^T \vec{e}_i}{\langle 1, 1, ..., 1 \rangle (B^r)^T (\vec{e}_i + \vec{e}_h)}, \qquad exists$$

Proof. See Appendix ■

This shows that $N(C)$ is well-defined on all state descriptions Θ^n for all n and completes the proof.

■

6.2 Probabilistic models of slow Π_1 sentences

We will now look at a general polyadic language L and will show the existence of the limit $\lim_{r \to \infty} N(C^r_{\mathcal{T}})$ for $\mathcal{T} \subseteq SL$ consisting of only those Π_1 sentences that for each k, have exponentially bounded number of models of size k. As before let $\mathcal{T} = \{\forall x_1, ..., x_t\ \theta(x_1, ..., x_t)\}$, $C_{\mathcal{T}} = \{w(\forall x_1, ..., x_t\ \theta(x_1, ..., x_t)) = 1\}$ the corresponding constraints set and k be an upper bound on i such that a_i appears in $\mathcal{T}$.

Definition 8. For $\Theta(b_1, \dots, b_r)$, a state description in L over $b_1, \dots, b_r$[12], we say b_i, b_j are *indistinguishable* mode $\Theta(\vec{b})$, denoted $b_i \sim_{\Theta(\vec{b})} b_j$, if

$$\Theta(b_1, \dots, b_r) \wedge b_i = b_j$$

is consistent with the axioms of equality for the language L plus =. The relation $\sim_{\Theta(\vec{b})}$ is an equivalence relation. The *spectrum* of $\Theta(\vec{b})$ is the multi-set of sizes of the equivalence classes of $\sim_{\Theta(\vec{b})}$ and the *length* of its spectrum, denoted $||\Theta(\vec{b})||$, is the number of equivalence classes.

Definition 9. [30] We say that a quantifier free formula $\theta(x_1, x_2, \dots, x_n)$ is *slow* if there are some constants c, d such that for all r the number of term models with domain $\{a_1, \dots, a_r\}$ that satisfy $\forall\vec{x}\theta(\vec{x})$ is at most dc^r.

Theorem 8. Let p be the largest arrity of any relation symbol in L. If $\theta(x_1, x_2, \dots, x_n)$ is slow with bound dk^r, then there is a finite set S of state descriptions Θ^{k+p} (of L^{k+p}), of spectrum length at most k such that

$$\bigwedge_{i_1,\dots,i_n=1}^{k+p} \theta(a_{i_i}, \dots, a_{i_n}) \equiv \bigvee_{\Theta^{k+p}\in S} \Theta^{k+p}. \tag{9}$$

Proof. Since the LHS is a sentence of L^{k+p}, there is a finite set of state descriptions of L^{k+p} that gives the above equivalence. We prove that any state description consistent with $\bigwedge_{i_1,\dots,i_n=1}^{k+p} \theta(a_{i_i}, \dots, a_{i_n})$ has spectrum length at most k. Suppose that there is a state description Θ^{k+p} of L^{k+p} consistent with $\forall\vec{x}\theta(x_1, \dots, x_n)$ with

$$||\Theta^{k+p}|| > k.$$

We can extend this state description to an state description on, $a_1, a_2, \dots, a_q$, for $q > k + p$ by making the new elements clones of existing elements. In other words, we just add the new elements to the equivalence classes of existing elements. Furthermore, we can do this in $||\Theta^{k+p}||^{q-k-p}$ ways. Thus, we will have at least $||\Theta^{k+p}||^{q-k-p}$ many models of $\forall\vec{x}\theta(x_1, \dots, x_n)$ of size q. But this clearly exceeds dk^q for sufficiently large q, and this is a contradiction. Thus if $\theta(x_1, \dots, x_n)$ is slow with bound dk^r, then for large r each state descriptions L^r that is consistent with $\forall x_1, \dots x_n\theta(x_1, \dots, x_n)$ has at most k distinguishable elements. ∎

Theorem 9. Let L be a first order language and let $\mathcal{T} = \{\forall\vec{x}\theta(\vec{x})\}$ where $\theta(\vec{x})$ is slow. For all n and state description Θ^n of L^n, let

$$N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} N(C^r_{\mathcal{T}})(\Theta^n).$$

[12] As given in Definition 5.

Then $N(C_{\mathcal{T}})$ is well-defined. Furthermore $N(C_{\mathcal{T}})$ is a probability function on SL that satisfies $C_{\mathcal{T}}$.

The idea of the proof is as follows. Remember from Proposition 2 that since N satisfies Renaming,

$$N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} \frac{|\{\Theta^r \mid {}^{\Theta^r \text{extends } \Theta^n}_{\Theta^r \text{consistent with} \mathcal{T}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} \tag{10}$$

We will show that for a slow formula $\theta(\vec{x})$ and large r, almost all models of $\forall\vec{x}\theta(x)$ of size r will have as many mutually distinguishable constants as possible. By Theorem 8 the maximum number of mutually distinguishable constants is bounded by k where dk^r is the bound for $\forall\vec{x}\theta(\vec{x})$. So the asymptotic number of models of size r is the same as models of size r with k equivalence classes of constants. This will give an expression for the denominator of (10). Next, we shall use the same intuitions to find an expression for the numerator of (10). To do this, we will find the number of models of $\forall\vec{x}\theta(x)$ that extend some given state description by looking at the number of possible extensions for each spectrum length, of which there are at most k.

Proof. If $N(C_{\mathcal{T}})$ is well-defined it will satisfy (1) and (2) by definition and thus, by Theorem 1 uniquely extends to a probability function on SL and clearly satisfies $N(C_{\mathcal{T}})$. To see that it is well-defined, let $\forall\vec{x}\theta(\vec{x})$ be slow with bound dk^r. Using Proposition 2, to show Theorem 9, it is enough to show that for any state description Θ^n, the limit

$$\lim_{r\to\infty} N(C_{\mathcal{T}})(\Theta^n) = \lim_{r\to\infty} \frac{|\{\Theta^r \mid {}^{\Theta^r \text{extends } \Theta^n}_{\Theta^r \text{consistent with} \mathcal{T}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} \tag{11}$$

exists.

Let Θ^r be a state description of L^r consistent with $\mathcal{T}$ and with equivalence classes $S_1, S_2, ..., S_q$ ordered such that if i_u is minimal with $a_{i_u} \in S_u$, then $i_1 < i_2 < ... < i_q$. This means that the equivalence classes are ordered by the minimum index of their constants. Notice that by Theorem 8, $q \leq k$. Take the constants $a_{i_1} \ldots, a_{i_q}$ from $S_1, \ldots, S_q$ respectively and let us consider the state description $\Psi(a_{i_1}, ..., a_{i_q})$ on $a_{i_1}, ..., a_{i_q}$ (see Definition 5) logically implied by Θ^r. So, $\Psi(a_{i_1}, ..., a_{i_q})$ is a sentence of the form

$$\Psi(a_{i_1}, ..., a_{i_q}) = \bigwedge_{b_1,...,b_{k_R} \in \{a_{i_1},...a_{i_q}\}} \pm R(b_1, \ldots, b_{k_R})$$

and $\Theta^r \vDash \Psi(a_{i_1}, ..., a_{i_q})$. Then $\Psi(a_{i_1}, ..., a_{i_q})$ has spectrum $\{1, ..., 1\}$ with length $q \leq k$. This means $\Psi(a_{i_1}, ..., a_{i_q})$ divides $a_{i_1}, \ldots, a_{i_q}$ into q equivalence classes, i.e., $a_{i_1}, ..., a_{i_q}$ are mutually distinguishable mod $\Psi(a_{i_1}, ..., a_{i_q})$. To see this, notice that any two a_{i_s}, a_{i_t} among these are distinguishable mod Θ^r because they are from different equivalence classes of $\sim_{\Theta^r}$. This means that there is some $\vec{a}$, and R such that $\Theta^r \vDash R(a_{i_s}, \vec{a}) \wedge \neg R(a_{i_t}, \vec{a})$ or $\Theta^r \vDash \neg R(a_{i_s}, \vec{a}) \wedge R(a_{i_t}, \vec{a})$, etc. But since Θ^r divides $\{a_1, \ldots, a_r\}$ into equivalence classes $S_1, \ldots, S_q$, for each a_u appearing in $\vec{a}$, we should have $a_u \sim_{\Theta^r} b_u$ for some $b_u \in \{a_{i_1}, \ldots, a_{i_q}\}$. Let $\vec{b} = (b_u)_{a_u \in \vec{a}}$. Then a_{i_s} and a_{i_t} can be distinguished by $\vec{b}$ and so they will be distinguishable by $\Psi(a_{i_1}, ..., a_{i_q})$. Then since every two $a_{i_s}, a_{i_t} \in \{a_{i_1}, \ldots, a_{i_q}\}$ are distinguishable for $\Psi(a_{i_1}, ..., a_{i_q})$, it has spectrum $\{1, 1, \ldots, 1\}$ of size q. Next, we should note that we can recover Θ^r from $\Psi(a_1, ..., a_q)$ and $S_1, ..., S_q$. So, the number of state descriptions Θ^r is the number of choices of $\Psi(a_{i_1}, ..., a_{i_q})$ and the choices of $S_1, ..., S_q$. Let d_q be the number of state descriptions Ψ^q (state descriptions on q constants) consistent with $\mathcal{T}$ that have spectrum length q.

The only condition on the equivalence classes is that they should be non-empty and form a partition of $\{a_1, \ldots, a_r\}$, so the number of choices of $S_1, ..., S_q$ will be the Stirling number of second kind, $S_r^q = \left\{ {r \atop q} \right\}$. Hence, the number of choices for the Θ^r above will be

$$d_q S_r^q = \frac{d_q}{q!} \sum_{j=0}^{q} (-1)^{q-j} \binom{q}{j} j^r.$$

This is the number of state descriptions Θ^r that are consistent with $\mathcal{T}$ and have spectrum length q.

It now follows that the number of state descriptions of L^r consistent with $\mathcal{T}$ is

$$\sum_{i=1}^{s} \frac{d_{q_i}}{q_i!} \sum_{j=0}^{q_i} (-1)^{q_i-j} \binom{q_i}{j} j^r$$

where $q_s < q_{s-1} < ... < q_1 \leq k$ are the distinct possible spectrum lengths of the state descriptions on L^r consistent with $\mathcal{T}$. The proportion of state descriptions with spectrum length q_1, as $r \to \infty$, will be

$$\lim_{r\to\infty} \left(\sum_{i=1}^{s} \frac{d_{q_i}}{q_i!} \sum_{j=0}^{q_i} (-1)^{q_i-j} \binom{q_i}{j} j^r \right) \left(\frac{d_{q_1}}{q_1!} \left(q_1^r - \ldots + (-1)^{q_1} \right) \right)^{-1} =$$

$$\lim_{r\to\infty} \left(\frac{d_{q_1}}{q_1!} \left(q_1^r - \ldots + (-1)^{q_1} \right) + \ldots + \frac{d_{q_s}}{q_s!} \left(q_s^r - \ldots + (-1)^{q_s} \right) \right) \left(\frac{d_{q_1}}{q_1!} q_1^r \right)^{-1} = 1. \quad (12)$$

What equation (12) says is that for large enough r, the number of models of $\mathcal{T}$ with domain $\{a_1, \ldots, a_r\}$ and maximum spectrum length (q_1) is the same as the number of all models with domain $\{a_1, \ldots, a_r\}$. This means that for large r, almost all models of $\forall\vec{x}\theta(\vec{x})$ with domain $\{a_1, \ldots, a_r\}$ have as many mutually distinct constants as possible. Thus, as $r \to \infty$ the number of state descriptions of L^r consistent with $\mathcal{T}$ will be asymptotically

$$\frac{d_{q_1}}{q_1!} q_1^r. \tag{13}$$

So (13) gives an expression for the denominator of (11). We will next try to find an expression for the numerator.

Fix a state description Θ^n of L^n. We are interested in the number of models of $\forall\vec{x}\theta(\vec{x})$ that extend this state description. Let Θ^r be as such, that is, a state description on L^r that extends Θ^n and is consistent with $\mathcal{T}$. By Theorem 8, Θ^r will have spectrum of length at most k, say with equivalence classes $S_1, ..., S_{q'}$, $q' \leq k$, again ordered as before, by the lowest indices appearing in them so that if i_t is minimal such that $a_{i_t} \in S_t$, then $i_1 < i_2 < \ldots < i_q$. Let h be maximal such that $i_h \leq n$. So, for $l \leq h$, every S_l includes some of $\{a_1, \ldots, a_n\}$ and for $h < k$, $S_k \cap \{a_1, \ldots, a_n\} = \emptyset$. We now take the constant with minimum index from $S_{h+1}, \ldots, S_{q'}$, that is $a_{i_{h+1}}, a_{i_{h+2}}, \ldots, a_{i_{q'}}$ respectively such that for all $a_j \in S_t$, $i_t \leq j$ for $t = h+1, \ldots, q'$.

Let $\Psi(a_1, a_2, \ldots, a_n, a_{i_{h+1}}, a_{i_{h+2}}, \ldots, a_{i_{q'}})$ be the state description on $a_1, a_2, \ldots, a_n$, $a_{i_{h+1}}, \ldots, a_{i_{q'}}$ determined by Θ^r (Definition 5). By the discussion above, and same as before, Θ^r can be recovered from Ψ and the equivalence classes $S_1, S_2, \ldots, S_{q'}$. To see this, notice that the constants appearing in Ψ cover all equivalence classes of $\sim_{\Theta^r}$, $S_1, \ldots, S_{q'}$, because it explicitly includes elements from $S_{h+1}, \ldots, S_{q'}$ and all $S_1, \ldots, S_h$ include some of $\{a_1, \ldots, a_r\}$ by definition. So, every other constant in $a_1, \ldots, a_r$ not appearing in Ψ is indistinguishable for Θ^r from one of $a_1, \ldots, a_n, a_{i_{h+1}}, \ldots, a_{q'}$. The difference now from our analysis for the denominator is that there we looked at all state descriptions as opposed to those extending Θ^n. So, we no longer have a free choice of partition $S_1, S_2, \ldots, S_{q'}$ because the non-empty members of

$$S_1 \cap \{1, 2, \ldots, n\},\ S_2 \cap \{1, 2, \ldots, n\}, \ldots, S_{q'} \cap \{1, 2, \ldots, n\} \tag{14}$$

should form a refinement of the partition of the equivalence classes $T_1, T_2, \ldots, T_q$ of Θ^n. These non-empty intersections will be a refinement of $T_1, T_2, \ldots, T_q$ because those constants from $\{a_1, \ldots, a_n\}$ that were distinguishable by Θ^n will remain so by Θ^r and also by Ψ as they extend Θ^n, but some of the constants that were indistinguishable by Θ^n (and therefore were in the same equivalence class T_i) might now be distinguishable for Θ^r and Ψ by means of new constants $a_{n+1}, \dots, a_r$.

Notice that there are finitely many of such possible Ψ's for each possible spectrum lengths. Let $\Psi_1, \dots, \Psi_s$ enumerate them, then all the state descriptions in the numerator of (11) will be recovered from one of these Ψ's. Hence to show that the limit in (11) exists, it will be enough to show that

$$\lim_{r\to\infty} \frac{|\{\Theta^r \mid {\scriptstyle \Theta^r \text{extends } \Theta^n \atop {\Theta^r \text{consistent with} \mathcal{T} \atop \Theta^r \text{recovered from} \Psi_j}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} \tag{15}$$

exists for $j = 1, \dots, s$ because

$$\lim_{r\to\infty} \frac{|\{\Theta^r \mid {\Theta^r \text{extends } \Theta^n \atop \Theta^r \text{consistent with} \mathcal{T}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} = \lim_{r\to\infty} \sum_{j=1}^{s} \frac{|\{\Theta^r \mid {\scriptstyle \Theta^r \text{extends } \Theta^n \atop {\Theta^r \text{consistent with} \mathcal{T} \atop \Theta^r \text{recovered from} \Psi_j}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|}.$$

For a fixed Ψ let q' be the spectrum length and $R_1, R_2, \dots, R_p$ denote the refinement as in (14). For this particular refinement the number of choices of $S_1, S_2, \dots, S_{q'}$ for which the non-empty members of (14) are $R_1, R_2, \dots, R_p$ is

$$\sum_{\substack{U \subseteq \{n+1,\dots,r\} \\ |U| \geq q'-p}} p^{r-n-|U|} \left\{ {|U| \atop q'-p} \right\}. \tag{16}$$

To see this, notice that the number of possible $S_1, \dots, S_{q'}$ is the number of ways one can distribute $a_{n+1}, \dots, a_r$ into q' equivalence classes, p of them given by $R_1, \dots, R_p$ (which already include $a_1, \dots, a_n$). That is the number of ways one can take a subset U of $a_{n+1}, \dots, a_r$ and distribute it between $q' - p$ equivalence classes with at least one for each class (to make sure we end up with right number of equivalence classes) that is $\left\{ {|U| \atop q'-p} \right\}$ times the number of ways to distribute the remaining $r - n - |U|$ between $R_1, \dots, R_p$ which are already non-empty and that is $p^{r-n-|U|}$.

Thus, the number of state descriptions corresponding to this Ψ that extend Θ^n, are consistent with $\mathcal{T}$, and have spectrum length q', will be given by (16). If we expand this, we get

$$\sum_{z=q'-p}^{r-n} \frac{p^{r-n-z}}{(q'-p)!} \left(\sum_{j=0}^{q'-p} (-1)^{q'-p-j} \binom{q'-p}{j} j^z \right) \binom{r-n}{z}$$

and inserting this in (15) we get

$$\lim_{r\to\infty} \frac{|\{\Theta^r \mid {}^{\Theta^r \text{extends } \Theta^n}_{\substack{\Theta^r \text{consistent with } \mathcal{T} \\ \Theta^r \text{recovered from} \Psi_j}}\}|}{|\{\Theta^r \mid \Theta^r \text{consistent with } \mathcal{T}\}|} =$$

$$\lim_{r\to\infty} \frac{\frac{1}{(q'-p)!}\sum_{z=q'-p}^{r-n} p^{r-n-z}\left(\sum_{j=0}^{q'-p}(-1)^{q'-p-j}\binom{q'-p}{j}j^z\right)\binom{r-n}{z}}{\frac{d_{q_1}}{q_1!}q_1^r} =$$

$$\lim_{r\to\infty} \frac{\frac{1}{(q'-p)!}\sum_{j=0}^{q'-p}(-1)^{q'-p-j}\binom{q'-p}{j}\sum_{z=q'-p}^{r-n} p^{r-n-z}j^z\binom{r-n}{z}}{\frac{d_{q_1}}{q_1!}q_1^r}. \tag{17}$$

Again, notice that there are finitely many j in the numerator of (17). Thus to show that the limit in (17) exists, it will be enough to show that it exists for each particular j. Since $\sum_{z=q'-p}^{r-n} p^{r-n-z}j^z\binom{r-n}{z}$ is asymptotic with $\sum_{z=0}^{r-n} p^{r-n-z}j^z\binom{r-n}{z} = (p+j)^{r-n}$, it is enough to show that

$$\lim_{r\to\infty} \frac{(-1)^{q'-p-j}\binom{q'-p}{j}(p+j)^{r-n}}{\frac{d_{q_1}}{q_1!}q_1^r}$$

exists for $j = 0, \ldots, q'-p$. But since $p+j \leq q' \leq q_1$, this is clearly zero unless $p+j = q' = q_1$, in which case it exists. Hence the limit in (15) exists for each j and as a result, the the limit in (11) exists. This completes the proof.

■

7 Probabilistic Models of arbitrary sets of sentences

We will end with a negative result. Giving the full detail of the result mentioned in [30] for Maximum Entropy inference process, we will show, by means of an example, that extending an inference process N, defined over propositional languages, to a first order language by taking the limit of its application on finite sublanguages as defined above, is not always well-defined.

Example Assume a predicate language L with a ternary relation symbol G and a binary relation symbol R and a unary predicate P and let $\mathcal{E}$ be the conjunction of:

$$\forall x, y, z(x =_G y \to (R(x,z) \to R(y,z)))$$

$$\forall x, y(R(x,y) \leftrightarrow R(y,x))$$

$$\forall x, y, z((R(x,y) \wedge R(x,z)) \to (x =_G y \vee x =_G z \vee y =_G z))$$

$$\forall x \exists y(x \neq_G y \wedge R(x,y))$$

$$\forall x \neg R(x, x)$$

and $\mathcal{O}$ be the conjunction of:

$$\forall x, y, z(x =_G y \to (R(x, z) \to R(y, z)))$$

$$\forall x, y(R(x, y) \leftrightarrow R(y, x))$$

$$\forall x, y, z((R(x, y) \wedge R(x, z)) \to (y =_G z))$$

$$\forall x, y, z, t((R(x, y) \wedge R(z, t) \wedge (x =_G y) \wedge (z =_G t)) \to (x =_G z))$$

$$\forall x \exists y R(x, y)$$

$$\exists x R(x, x)$$

where

$$x =_G y \leftrightarrow \forall u, t(G(x, u, t) \leftrightarrow G(y, u, t)).$$

Let $\mathcal{M}_{\mathcal{E}}^n$ and $\mathcal{M}_{\mathcal{O}}^n$ denote the models of $\mathcal{E}$ and $\mathcal{O}$ of size n respectively [13]. The result then follows from the following two claims which we prove in the appendix.

Claim 2. Let $\#\mathcal{M}_{\mathcal{E}}^n$ be the number of models of $\mathcal{E}$ of size n.

- If n is an even number, $\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} \leq \#\mathcal{M}_{\mathcal{E}}^n \leq \frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} + n^n\binom{2^{n^2}}{n-2}$.
- If n is an odd number, $\#\mathcal{M}_{\mathcal{E}}^n \leq n^n\binom{2^{n^2}}{n-1}$.

Comparing the upper bound calculated for models of size n for large n when n is odd with the lower bound of models of size n when n is even, we can see that $\mathcal{E}$ has significantly more models of even size than models of odd size. We will now follow the same way to find an estimation of the number of models of $\mathcal{O}$.

Claim 3. Let $\#\mathcal{M}_{\mathcal{O}}^n$ be the number of models of $\mathcal{O}$ of size n.

- If n is an even number, $\#\mathcal{M}_{\mathcal{O}}^n \leq n^n\binom{2^{n^2}}{n-1}$.
- If n is an odd number, $\frac{n!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\binom{2^{n^2}}{n} \leq \#\mathcal{M}_{\mathcal{O}}^n$.

[13] Notice that $\mathcal{E}$ and $\mathcal{O}$ are Π_2

Thus for even n,

$$\frac{\#\mathcal{M}_{O}^{n}}{\#\mathcal{M}_{\mathcal{E}}^{n}} \leq \frac{n^{n}\binom{2^{n^2}}{n-1}}{\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n}} \leq \frac{n^{n+1}2^{\frac{n}{2}}(\frac{n}{2})!}{n!(2^{n^2}-n+1)} \leq \frac{n^{n+1}2^{\frac{n}{2}}}{2^{n^2}-n+1}$$

but we have $n^{n+1}2^{\frac{n}{2}} = 2^{(n+1)\log n+\frac{n}{2}}$ and $2^{(n+1)\log n+\frac{n}{2}} << 2^{n^2}$ since for large enough n, $\log n + \frac{1}{2} << n$. Thus $\lim_{\substack{n\to\infty \\ n\,even}} \frac{\#\mathcal{M}_{O}^{n}}{\#\mathcal{M}_{\mathcal{E}}^{n}} = 0$. Using the same pattern, for odd n,

$$\frac{\#\mathcal{M}_{\mathcal{E}}^{n}}{\#\mathcal{M}_{O}^{n}} \leq \frac{n^{n}\binom{2^{n^2}}{n-2}}{\frac{n!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\binom{2^{n^2}}{n}} \leq \frac{n^{n}2^{\frac{n-1}{2}}}{(2^{n^2}-n+2)(2^{n^2}-n+1)} \leq \frac{2^{n\log n+\frac{n-1}{2}}}{(2^{n^2}-n+2)(2^{n^2}-n+1)}$$

and so $\lim_{\substack{n\to\infty \\ n\,odd}} \frac{\#\mathcal{M}_{\mathcal{E}}^{n}}{\#\mathcal{M}_{O}^{n}} = 0$. Now let $\mathcal{T} = \{(\mathcal{E} \wedge \forall x P(x)) \vee (O \wedge \forall x \neg P(x))\}$. Then for a sentence like $P(a_1)$

$$\lim_{n\to\infty} N(C_{\mathcal{T}}^{n})(P(a_1))$$

does not exist. To see this notice that for very large n, if n is even almost all models of $\mathcal{T}$ of size n should satisfy $\mathcal{E} \wedge \forall x P(x)$ and thus $N(C_{\mathcal{T}}^{n})(P(a_1)) = 1$ while for odd n, almost all models of $\mathcal{T}$ will satisfy $(O \wedge \forall x \neg P(x))$ and thus $N(C_{\mathcal{T}}^{n})(P(a_1)) = 0$. This example shows that even in the case of a Π_2 sets of sentences, we cannot in general define $N(C_{\mathcal{T}})$, the extension of the probability function defined by taking the limit of $N(C_{\mathcal{T}}^{r})$ of quantifier free sentences (or state descriptions) over finite sub languages L^r, simply because the relevant asymptotic limit does not necessarily exist even when we drop the equality from language.[14]

8 Conclusion

We studied most normal probabilistic characterisation of under-determined models specified by some finite consistent set of axioms. Specific instances of this problem is of interest in many areas and there is extensive literature studying them for propositional languages and different approaches have been proposed and studied, each promoting some notion of normality for the way such probabilistic characterisation is carried out. The situation for first order languages, however, seem significantly different. One approach to answer this for the first order case is to attempt to define this probabilistic characterisation directly on the first order

[14]The idea of using a relation symbol, here G, to 'approximate' the equality via $=_G$ is due to Grove, Halpern and Koller [15] to my knowledge.

language. This, however, seems to depend strongly on the specific conditions that one assumes for the way that the characterisation has to be carried out. In our terminology, on how exactly the notion of normality for the probabilistic characterisation is formalised. However, even for specific cases, and indeed even for the most extensively studied notion of normality, i.e. the Maximum Entropy models, there is no proposal on how to define these probabilistic models directly on the first order language in general, see [35]. A second approach is to try to define the probabilistic models on first order languages as the limit of such models on finite sublanguages. These sublanguages can be essentially treated as propositional languages where the situation is much better understood. There are, however, at least two issues with this approach. The first comes from dealing with sets of first order axioms that have no finite models where this approach fails immediately. But even assuming that the set of axioms will have finite models of sufficiently large size, still this limit does not necessarily exist in general as we showed in the previous section. Nevertheless as we showed for *simple* sets of axioms, i.e those with quantifier complexity of at most Π_1 the approach looks promising. We showed this for any set of axioms from a unary first order language and for sets of axioms with quantifier complexity of Σ_1 as well as for special cases of Π_1 sets of axioms. We conjecture that this is also the case for all Π_1 sets of axioms.
Finally it is worth emphasising that we did not deal with the computational complexity of the problem of calculating probability distributions given by our extension of inference processes to first order languages. This analysis falls outside the scope of this paper but it is important to mention that for connecting these results to applications a proper understanding of the computational complexities involved would be important and necessary.

References

[1] Bacchus, F, Grove, A.J., Halpern, J.Y. and Koller, D., Generating new beliefs from old, Proceedings of the Tenth Annual Conference on Uncertainty in Arti - cial Intelligence, (UAI-94), 1994, 37-45.

[2] Bacchus, F, Grove, A.J., Halpern, J.Y. & Koller, D., "From Statistical Knowledge to Degrees of Belief" in *Artificial Intelligence,* 1996, 87:75–143.

[3] Balogh, Bollobas & Weinreich, "The Speed of Hereditary Properties of Graphs" in *Journal of Combinatorial Theory*, 2000, series B 79:131–156.

[4] Balogh, Bollobas & Weinreich, "The Penultimate Range of Growth for Graph Properties" in *European Journal of Combinatorics*, 2001, 22(3):277–289.

[5] Ballobas & Thomason, "Projections of Bodies and Hereditary properties of hyper graphs" in *Bulletin of the London Mathematical Society*, 1995, 27(5):417–424.

[6] Barnett, O.W. & Paris, J.B., "Maximum Entropy Inference with Quantified Knowledge" in *Logic Journal of the IGPL,* 2008, 16(1):85–98.

[7] Berger, A., Della Pietra, S. & Della Pietra, V., "A maximum Entropy Approach to Natural Language Processing" in *Computational Linguistics*, 1996, 22(1):39–71.

[8] Broecheler, M., Simari, G. I., & Subrahmanian V. S., "Using Histograms to Better Answer Queries to Probabilistic Logic Programs",*Proceedings of 25th International Conference on Logic Programming, ICLP*, 40–45, 2009

[9] Chang, C.C and Keisler, H.J., Model Theory, Studies in Logic and the Founda- tions of Mathematics, Vol. 73., North Holland Publishing Co., 1973.

[10] Chen, C. H., "Maximum Entropy Analysis for Pattern Recognition", in *Maximum Entropy and Bayesian Methods*, P. F. Fougere (eds), Kluwer Academic Publisher, 1990.

[11] Grotenhuis, M. G., "An Overview of the Maximum Entropy Method of Image Deconvolution", A University of Minnesota – Twin Cities "Plan B" Master's paper.

[12] Grove, A.J., Halpern, J.Y. & Koller, D., "Random Worlds and Maximum Entropy" in *Journal of Artificial Intelligence Research*, 1994, 2:33–88.

[13] Gaifman, H., "Concerning measures in first order calculi" in *Israel Journal of Mathematics*, 1964, 2(1):1–18.

[14] A.J. Grove, J.Y. Halpern, D. Koller, Asymptotic conditional probabilities: the unary case. SIAM J. of Computing, 1996, 25(1): 1–51.

[15] A.J. Grove, J.Y. Halpern, D. Koller, Asymptotic conditional probabilities: the non-unary case. J. Symbolic Logic, 1996, 61(1): 250–276.

[16] Guiasu, S. & Shenitzer, A., "The Principle of Maximum Entropy", in*The Mathematical Intelligencer*, 1985, 7(1).

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

[18] Jaynes, E. T., "Information Theory and Statistical Mechanics" in *Physical Reviews*, 1957, 106: 620–630, 108:171–190.

[19] Jaynes, E. T., "Notes on Present Status and Future Prospects" in *Maximum Entropy and Bayesian Methods*, W.T. Grandy & L.H. Schick, (edt), Kluwer, 1990, pp.1–13.

[20] Jaynes, E. T., "How Should We Use Entropy in Economics?", 1991,manuscript available at: http://www.leibniz.imag.fr/LAPLACE/Jaynes/prob.html.

[21] Jaynes, E. T., "Clearing up mysteries – The original goal", in *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.), Kluwer, Dordrecht, 1989.

[22] Kapur, J. N., "Twenty Five Years of Maximum Entropy" in *Journal of Math. Phy Sci*, 1983, 17(2):103–156

[23] Kapur, J. N., *Maximum Entropy Models in Science and Engineering*, John Wiley & Sons, 1990.

[24] Kapur, J. N., "Non-Additive Measures of Entropy and Distributions of Statistical Mechanics" in *Ind Jour Pure App Math*, 1983, 14(11):1372—1384.

[25] Landes, J. & Masterton, G., "Invariant Equivocation" *Erkenntnis*, 82: 141–167, 2017.

[26] Landes, J. & Williamson, J. "Justifying Objective Bayesianism on Predicate Languages", *Entropy* 17(4):2459–2543, 2015.

[27] Landes, J. & Williamson, J., "Objective Bayesianism and the maximum entropy principle", *Entropy*, 15 (9): 3528–3591, 2013.

[28] Paris, J.B., *The Uncertain Reasoner's Companion*, Cambridge University Press, 1994.

[29] Paris, J.B., & Rafiee Rad, S., "Inference Processes for Quantified Predicate Knowledge", in *Logic, Language, Information and Computation*, WoLLIC, Edinburgh, 2008, Eds. W.Hodges and R. de Queiroz, Springer LNAI, 5110, pp249-259.

[30] Paris, J. B. & Rafiee Rad, S., "A note on the least informative model of a theory", in *Programs, Proofs, Processes*, CiE 2010, Eds. F. Ferreira, B. Löwe, E. Mayordomo, & L. Mendes Gomes, Sprnger LNCS 6158, pp342-351, 2010.

[31] Paris, J.B. & Vencovská, "On the Applicability of Maximum Entropy to Inexact Reasoning" in *International Journal of Approximate Reasoning*, 1989, 3(1): 1–34.

[32] Paris, J.B. & Vencovská, "A Note on the Inevitability of Maximum Entropy" in *International Journal of Approximate Reasoning*, 1990, 4(3):183–224.

[33] Rafiee Rad, Soroush, *Inference Processes for First Order Probabilistic Languages*, PhD Thesis, University of Manchester 2009.

[34] Rafiee Rad, S., "Maximum Entropy Models for Σ_1 Sentences", in *Journal of Applied Logic*, 2018, 5(1), 287–300.

[35] Rafiee Rad, S., "Equivocation Principle on First Order Languages", in *Studia Logica*, 2017, 105(1), 121–152.

[36] Rényi, A., 'On measures of information and entropy", in *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1960, 547–561.

[37] Shannon,C. E. & Weaver, W. *The Mathematical Theory of Communication*, University of Illinois Press, 1949.

[38] Scheinerman E. R. & Zito, J., "On the Size of Hereditary Classes of Graphs" in *Journal Combinatorial Theory*, 1994, Series B 61: 16–39.

[39] Williamson, J., *In Defence of Objective Bayesianism* Oxford University Press, Oxford, 2010.

[40] Williamson, J., "Objective Bayesian probabilistic logic" in *Journal of Algorithms in Cognition, Informatics and Logic*, 2008, 63: 167–183.

[41] Williamson, J. "Motivating Objective Bayesianism: From Empirical Constraints to Objective Probabilities", in *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr*, Harper, W. L. and Wheeler, G. R. (edts) London, UK: College Publications, 2007:155–183.

[42] Zellner, A. "Bayesian Methods and Entropy in Economics and Econometrics" in *Maximum /entropy and Bayesian Methods*, Grandy W. T. & Schick, L. H. (eds), Kluwer Academic Publishing, 1991.

9 Appendix

Proof of Claim 1 We will show that for every two element, $b_{ij}^{(n)}$, $b_{st}^{(n)}$ in the matrix B^n either the limit of the ratio of these elements, $\lim_{n\to\infty} \frac{b_{ij}^{(n)}}{b_{st}^{(n)}}$, is finite or one of them grow much faster than the other, i.e., the limit is infinite. Thus if we consider

$$\lim_{k\to\infty} \frac{< 1, 1, ..., 1 > B^{k\,T} \vec{e_i}}{< 1, 1, ..., 1 > B^{k\,T} (\vec{e_i} + \vec{e_h})}$$

the limit will be finite if the ratio of every two elements has a finite limit and if not all of them have a finite limit then the one that grows fastest will appear in the denominator and this makes the overall limit zero or 1 and this completes the proof of Claim 1.

Proof. Let $B = (b_{ij})$ be an $R \times R$ lower triangular matrix with positive entries. Then the ij entry of B^n, for $i \geq j$ is given by

$$\sum_{i=t_0>t_1>...>t_m=j} \sum_{r_0+...+r_m=n-m} \prod_{s=0}^{m-1} b_{t_s t_{s+1}} \prod_{s=0}^{m} b_{t_s t_s}^{r_s}.$$

There are only a finite fixed number of possible $t_0, ..., t_m$ so it would be enough to show that for two particular choices (possibly at different i, j) the limit

$$\lim_{n\to\infty} \frac{\sum_{r_0+...+r_m=n} \prod_{s=0}^{m} b_{t_s t_s}^{r_s}}{\sum_{u_0+...+u_q=n} \prod_{s=0}^{q} b_{g_s g_s}^{u_s}} \tag{18}$$

either exists or is ∞. To show this we will first find a better expression for, say, the numerator. We will consider this in two cases.

Before proceeding with the proof we will try to find simpler expression for the terms in nominator and denominator of (18).

Claim 4.

$$\sum_{r_0+...+r_m=n} \prod_{s=0}^{m} b_{t_s t_s}^{r_s} = \sum_{s=0}^{m} b_{t_s t_s}^{n+m} \prod_{y\neq s} (b_{t_s t_s} - b_{t_y t_y})^{-1}.$$

We first show the following two technical lemmas that will be usefult for proving Claim 4.

Lemma 10.

$$\sum_{i=0}^{m} \frac{1}{(b_{m+1} - b_i) \prod_{\substack{j=0 \\ j\neq i}}^{m} (b_i - b_j)} = \frac{1}{\prod_{k=0}^{m} (b_{m+1} - b_k)}$$

Proof. we will show that :

$$\sum_{i=0}^{m} \frac{1}{(b_{m+1}-b_i)\prod_{\substack{j=0\\ j\neq i}}^{m}(b_i-b_j)} - \frac{1}{\prod_{k=0}^{m}(b_{m+1}-b_k)} = 0$$

To see this multiply both sides by $\prod_{k=0}^{m}(b_{m+1}-b_k)$ and we will have :

$$\sum_{i=0}^{m} \frac{\prod_{\substack{k=0\\ k\neq i}}^{m}(b_{m+1}-b_k)}{\prod_{\substack{j=0\\ j\neq i}}^{m}(b_i-b_j)} - 1 = 0$$

The left hand side is polynomial in b_{m+1} with degree m and $m+1$ distinct zeros, namely $\{b_0, b_1, ..., b_m\}$, so it should be identical with zero. ■

Lemma 11.

$$\sum_{j=1}^{m}\left(\sum_{i=0}^{m} \frac{b_{m+1}^{n+j} b_i^{m-j}}{\prod_{\substack{k=0\\ k\neq i}}^{m}(b_i-b_k)}\right) = 0$$

Proof.

$$\sum_{j=1}^{m}\left(\sum_{i=0}^{m} \frac{b_{m+1}^{n+j} b_i^{m-j}}{\prod_{\substack{k=0\\ k\neq i}}^{m}(b_i-b_k)}\right) = b_{m+1}^{n+1}\left[\sum_{i=0}^{m} \frac{\sum_{j=0}^{m-1} b_{m+1}^{j} b_i^{m-1-j}}{\prod_{\substack{k=0\\ k\neq i}}^{m}(b_i-b_k)}\right] = b_{m+1}^{n+1}\left[\sum_{i=0}^{m} \frac{b_{m+1}^{m}-b_i^{m}}{(b_{m+1}-b_i)\prod_{\substack{k=0\\ k\neq i}}^{m}(b_i-b_k)}\right]$$

$$= b_{m+1}^{n+1}\left[\sum_{i=0}^{m} \frac{b_{m+1}^{m}}{(b_{m+1}-b_i)\prod_{\substack{k=0\\ k\neq i}}^{m}(b_i-b_k)} + \sum_{i=0}^{m} \frac{b_i^{m}}{\prod_{\substack{k=0\\ k\neq i}}^{m+1}(b_i-b_k)}\right]$$

$$= b_{m+1}^{n+1}\left[\frac{b_{m+1}^{m}}{\prod_{k=0}^{m}(b_{m+1}-b_k)} + \sum_{i=0}^{m} \frac{b_i^{m}}{\prod_{\substack{k=0\\ k\neq i}}^{m+1}(b_i-b_k)}\right]$$

where the last equality is given by Lemma 10. Thus it would be enough to show that

$$\left[\frac{b_{m+1}^{m}}{\prod_{k=0}^{m}(b_{m+1}-b_k)} + \sum_{i=0}^{m} \frac{b_i^{m}}{\prod_{\substack{k=0\\ k\neq i}}^{m+1}(b_i-b_k)}\right] = 0$$

To see this multiply both sides by $\prod_{k=0}^{m}(b_{m+1}-b_k)$ and we will have

$$b_{m+1}^{m} - \left(\frac{b_0^{m}(b_{m+1}-b_1)...(b_{m+1}-b_m)}{(b_0-b_1)...(b_0-b_m)} + ... + \frac{b_m^{m}(b_{m+1}-b_0)...(b_{m+1}-b_{m-1})}{(b_m-b_0)...(b_m-b_{m-1})}\right)$$

the above expression is a polynomial of degree m with respect to b_{m+1} which has $m+1$ roots, namely $\{b_0, ...b_m\}$ so it should be identical with zero. ■

Proof. of Claim 4

Proof by induction on m; For the base case, where $m = 0$ we have

$$b^n_{t_0t_0} = b^n_{t_0t_0}$$

which is clearly true. Suppose the result is true for m and we will prove it for $m+1$. To simplify the notation we will write b_s for $b_{t_st_s}$, etc.

$$\sum_{r_0+...+r_m+r_{m+1}=n} \prod_{s=0}^{m+1} b_s^{r_s} = \sum_{r_{m+1}=0}^{n} \left[\sum_{r_0+...+r_m=n-r_{m+1}} b_{m+1}^{r_{m+1}} \prod_{s=0}^{m} b_s^{r_s} \right]$$

$$= \sum_{r_{m+1}=0}^{n} \left[b_{m+1}^{r_{m+1}} \sum_{s=0}^{m} b_s^{n+m-r_{m+1}} \prod_{y\neq s} (b_s - b_y)^{-1} \right] = \sum_{j=0}^{n} \left(\sum_{s=0}^{m} \frac{b_{m+1}^{j} b_s^{n+m-j}}{\prod_{\substack{k=0\\k\neq s}}^{m} (b_s - b_k)} \right)$$

The equality before last is given by the induction hypothesis. We add to this the expression in Lemma 11(which is equal to zero),
So we will have

$$\sum_{r_0+...+r_m+r_{m+1}=n} \prod_{s=0}^{m+1} b_s^{r_s} = \sum_{j=0}^{n} \left(\sum_{s=0}^{m} \frac{b_{m+1}^{j} b_s^{n+m-j}}{\prod_{\substack{k=0\\k\neq s}}^{m} (b_s - b_k)} \right) + \sum_{j=1}^{m} \left(\sum_{s=0}^{m} \frac{b_{m+1}^{n+j} b_s^{m-j}}{\prod_{\substack{k=0\\k\neq s}}^{m} (b_s - b_k)} \right) =$$

$$\sum_{s=0}^{m} \left(\sum_{j=0}^{n+m} \frac{b_{m+1}^{j} b_s^{n+m-j}}{\prod_{\substack{k=0\\k\neq s}}^{m} (b_s - b_k)} \right) = \sum_{s=0}^{m} \left(\frac{b_s^{n+m+1} - b_{m+1}^{n+m+1}}{\prod_{\substack{k=0\\k\neq s}}^{m+1} (b_s - b_k)} \right)$$

$$= \sum_{s=0}^{m} \left(\frac{b_s^{n+m+1}}{\prod_{\substack{k=0\\k\neq s}}^{m+1} (b_s - b_k)} \right) - \sum_{s=0}^{m} \left(\frac{b_{m+1}^{n+m+1}}{\prod_{\substack{k=0\\k\neq s}}^{m+1} (b_s - b_k)} \right)$$

and using Lemma 10, we have

$$\sum_{r_0+...+r_m+r_{m+1}=n} \prod_{s=0}^{m+1} b_s^{r_s} = \sum_{s=0}^{m} \left(\frac{b_s^{n+m+1}}{\prod_{\substack{k=0\\k\neq s}}^{m+1} (b_s - b_k)} \right) + \frac{b_{m+1}^{n+m+1}}{\prod_{k=0}^{m} (b_{m+1} - b_k)} = \sum_{s=0}^{m+1} b_s^{n+m+1} \prod_{y\neq s} (b_s - b_y)^{-1}$$

and this completes the proof of Claim 4. ■

We now return to the proof of Claim 1.

Case1 Assume that all the $b_{t_s t_s}$ are different. Using Claim 4 for the case when all $b_{t_s t_s}$ are distinct, we can see that the limit in (18) clearly exists, if $\max\{b_{t_s t_s}\} \leq \max\{b_{g_s g_s}\}$ and is ∞ otherwise.

Case 2 If not all the $b_{t_s t_s}$ are distinct, suppose that the distinct values are $a_0, ..., a_p$ and let $A_j = \{t \,|\, b_{tt} = a_j\}$, $d_j = |A_j|$ and $r'_j = \sum_{b_{t_i t_i} = a_j} r_i$ then

$$\sum_{r_0+...+r_m=n} \prod_{s=0}^{m} b_{t_s t_s}^{r_s} = \lim_{A_p \to a_p} ... \lim_{A_0 \to a_0} \sum_{r_0+...+r_m=n} \prod_{s=0}^{m} z_{t_s t_s}^{r_s}$$

$$= \lim_{A_p \to a_p} ... \lim_{A_0 \to a_0} \sum_{s=0}^{m} z_{t_s t_s}^{n+m} \prod_{y \neq s} (z_{t_s t_s} - z_{t_y t_y})^{-1} =$$

$$\lim_{A_p \to a_p} ... \lim_{A_0 \to a_0} \sum_{t_s \in A_0} z_{t_s t_s}^{n+m} \prod_{y \neq s} (z_{t_s t_s} - z_{t_y t_y})^{-1} + ... + \lim_{A_p \to a_p} ... \lim_{A_0 \to a_0} \sum_{t_s \in A_p} z_{t_s t_s}^{n+m} \prod_{y \neq s} (z_{t_s t_s} - z_{t_y t_y})^{-1}$$

$$= \lim_{A_0 \to a_0} \sum_{t_s \in A_0} \frac{z_{t_s t_s}^{n+m}}{\prod_{j=1}^{p} (z_{t_s t_s} - a_j)^{d_j}} \prod_{\substack{y \neq s \\ t_y \in A_0}} (z_{t_s t_s} - z_{t_y t_y})^{-1} + ... + \lim_{A_p \to a_p} \sum_{t_s \in A_p} \frac{z_{t_s t_s}^{n+m}}{\prod_{j=0}^{p-1} (z_{t_s t_s} - a_j)^{d_j}} \prod_{\substack{y \neq s \\ t_y \in A_p}} (z_{t_s t_s} - z_{t_y t_y})^{-1}$$

where for $A_i = \{t_1, ..., t_k\}$, $A_i \to a_i$ is intended as short for $\lim_{z_{t_k t_k} \to a_i} ... \lim_{z_{t_1 t_1} \to a_i}$ and the second equality is given by Claim 4. The following two lemmas allow us to simplify this even further.

Lemma 12. For an infinitely differentiable function f,

$$\lim_{z \to x} (k!)^{-1} \frac{\partial^k}{\partial x^k} \left(\frac{f(x)}{x-z} - \frac{f(z)}{x-z} \right) = \frac{1}{(k+1)!} \frac{\partial^{k+1}}{\partial x^{k+1}} f(x)$$

Proof.
Using the infinite Taylor expansion

$$f(z) = f(x) + (z-x) \frac{\partial}{\partial x} f(x) + \frac{(z-x)^2}{2!} \frac{\partial^2}{\partial x^2} f(x) + ...$$

since $f(x)$ is infinitely differentiable we have

$$\frac{f(x)}{x-z} - \frac{f(z)}{x-z} = \sum_{n=1}^{\infty} \frac{(z-x)^{n-1}}{n!} \frac{\partial^n}{\partial x^n} f(x)$$

and thus

$$\frac{1}{k!} \lim_{z \to x} \left(\frac{\partial^k}{\partial x^k} \left(\frac{f(x)}{x-z} - \frac{f(z)}{x-z} \right) \right) = \frac{1}{k!} \lim_{z \to x} \left(\frac{\partial^k}{\partial x^k} \left(\sum_{n=1}^{\infty} \frac{(z-x)^{n-1}}{n!} \frac{\partial^n}{\partial x^n} f(x) \right) \right) \tag{19}$$

any term in the right hand side with $n > k+1$ will include a positive power of $(z-x)$ after k derivative and so will approach zero as $z \to x$. So from (19)

$$\frac{1}{k!}\lim_{z\to x}\left(\frac{\partial^k}{\partial x^k}\left(\frac{f(x)}{x-z}-\frac{f(z)}{x-z}\right)\right)=\frac{1}{k!}\lim_{z\to x}\left(\frac{\partial^k}{\partial x^k}\left(\sum_{n=1}^{k+1}\frac{(z-x)^{n-1}}{n!}\frac{\partial^n}{\partial x^n}f(x)\right)\right)$$

$$=\frac{1}{k!}\lim_{z\to x}\left(\sum_{n=1}^{k+1}\left(\sum_{i=0}^{k}\binom{k}{i}\frac{\partial^i}{\partial x^i}\left(\frac{(z-x)^{n-1}}{n!}\right)\frac{\partial^{n+k-i}}{\partial x^{n+k-i}}f(x)\right)\right). \tag{20}$$

Any terms in the inner sum of the rightmost expression with $i \geq n$ is zero because $\frac{\partial^i}{\partial x^i}\left(\frac{(z-x)^{n-1}}{n!}\right) = 0$ for $i \geq n$. Moreover, for $i < n-1$, $\frac{\partial^i}{\partial x^i}\left(\frac{(z-x)^{n-1}}{n!}\right)$ will include a positive power of $(z-x)$. So for every term , say T, in the above expression with $i \neq n-1$ we have

$$\lim_{z\to x} T = 0$$

so from 20 we have

$$\frac{1}{k!}\lim_{z\to x}\left(\frac{\partial^k}{\partial x^k}\left(\frac{f(x)}{x-z}-\frac{f(z)}{x-z}\right)\right)=\frac{1}{k!}\lim_{z\to x}\left(\sum_{n=1}^{k+1}\binom{k}{n-1}\frac{\partial^{n-1}}{\partial x^{n-1}}\left(\frac{(z-x)^{n-1}}{n!}\right)\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)\right)$$

$$=\frac{1}{k!}\sum_{n=1}^{k+1}\frac{(-1)^{n-1}}{n}\binom{k}{n-1}\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)=\frac{1}{k!}\sum_{n=1}^{k+1}\frac{(-1)^{n-1}}{n}\frac{k!}{(n-1)!(k+1-n)!}\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)$$

$$=\frac{1}{(k+1)!}\sum_{n=1}^{k+1}(-1)^{n-1}\frac{(k+1)!}{n!(k+1-n)!}\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)=\frac{1}{(k+1)!}\sum_{n=1}^{k+1}(-1)^{n-1}\binom{k+1}{n}\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)$$

$$=\frac{1}{(k+1)!}\frac{\partial^{k+1}}{\partial x^{k+1}}f(x)$$

■

Lemma 13. For an infinitely differentiable function $g(x)$:

$$\lim_{x_k\to x_1}\lim_{x_{k-1}\to x_1}\ldots\lim_{x_2\to x_1}\sum_{i=1}^{k}g(x_i)\prod_{i\neq j}(x_i-x_j)^{-1}=\left((k-1)!^{-1}\frac{\partial^{k-1}}{\partial x^{k-1}}g(x)\right)_{x_1}$$

Proof. By induction on k. The result is obvious for $k = 2$. Suppose the lemma is true for k and we will show it for $k+1$.

$$\lim_{x_{k+1}\to x_1}\lim_{x_k\to x_1}\ldots\lim_{x_2\to x_1}\sum_{i=1}^{k+1}g(x_i)\prod_{i\neq j}(x_i-x_j)^{-1}=\lim_{x_{k+1}\to x_1}\left(\lim_{x_k\to x_1}\ldots\lim_{x_2\to x_1}\sum_{i=1}^{k}g(x_i)\prod_{j=1,j\neq i}^{k+1}(x_i-x_j)^{-1}+\right.$$

$$\lim_{x_k \to x_1} \dots \lim_{x_2 \to x_1} g(x_{k+1}) \prod_{i=1}^{k} (x_{k+1} - x_i)^{-1} \Bigg).$$

Notice that for an infinitely differentiable g we have

$$\left((k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} g(x) \right)_{x_1} = \lim_{x \to x_1} (k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} g(x).$$

Now using the induction hypothesis for $\frac{g(x)}{x - x_{k+1}}$ we will have

$$\lim_{x_{k+1} \to x_1} \lim_{x_k \to x_1} \dots \lim_{x_2 \to x_1} \sum_{i=1}^{k+1} g(x_i) \prod_{j \neq i} (x_i - x_j)^{-1}$$

$$= \lim_{x_{k+1} \to x_1} \left(\left((k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \frac{g(x)}{(x - x_{k+1})} \right)_{x_1} - (-1)^k \frac{g(x_{k+1})}{(x_1 - x_{k+1})^k} \right)$$

$$= \lim_{x_{k+1} \to x_1} \left(\lim_{x \to x_1} \left((k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \frac{g(x)}{(x - x_{k+1})} \right) - \lim_{x \to x_1} \left((-1)^k \frac{g(x_{k+1})}{(x - x_{k+1})^k} \right) \right)$$

$$= \lim_{x_{k+1} \to x_1} \lim_{x \to x_1} \left((k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \frac{g(x)}{(x - x_{k+1})} - (k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \frac{g(x_{k+1})}{(x - x_{k+1})} \right)$$

$$= \lim_{x \to x_1} \lim_{x_{k+1} \to x} (k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \left(\frac{g(x)}{x - x_{k+1}} - \frac{g(x_{k+1})}{x - x_{k+1}} \right) \tag{21}$$

(since all the functions involved are infinitely differentiable). Using Lemma 12 from (21) we have:

$$\lim_{x_{k+1} \to x_1} \lim_{x_k \to x_1} \dots \lim_{x_2 \to x_1} \sum_{i=1}^{k+1} g(x_i) \prod_{i \neq j} (x_i - x_j)^{-1} = \lim_{x \to x_1} \lim_{x_{k+1} \to x} (k!)^{-1} \frac{\partial^{k-1}}{\partial x^{k-1}} \left(\frac{g(x)}{x - x_{k+1}} - \frac{g(x_{k+1})}{x - x_{k+1}} \right)$$

$$= \lim_{x \to x_1} \frac{1}{k!} \left(\frac{\partial^k}{\partial x^k} g(x) \right) = \frac{1}{k!} \left(\frac{\partial^k}{\partial x^k} g(x) \right)_{x_1}$$

as required and this completes the proof of Lemma 13. ■

Using Lemma 13[15] the expressions in the numerator and denominator of (18) will be in the form

$$\sum_{r_0 + \dots + r_m = n} \prod_{s=0}^{m} b_{t_s t_s}^{r_s} = \sum_{i=0}^{p} \lim_{A_i \to a_i} \sum_{z_{t_s t_s} \in A_i} \frac{z_{t_s t_s}^{n+m}}{\prod_{j \neq i} (z_{t_s t_s} - a_j)^{d_j}} \prod_{\substack{y \neq s \\ t_y \in A_i}} (z_{t_s t_s} - z_{t_y t_y})^{-1} =$$

[15]To see how Lemma 13 applies notice that here the limit $\lim_{A_i \to a_i} = \lim_{z_{t_k t_k} \to a_i} \dots \lim_{z_{t_1 t_1} \to a_i}$ can be equivalently written as $\lim_{z_{t_1 t_1} \to a_i} \lim_{z_{t_k t_k} \to z_{t_1 t_1}} \dots \lim_{z_{t_2 t_2} \to z_{t_1 t_1}}$.

$$\frac{1}{(d_0-1)!}\left(\frac{\partial^{(d_0-1)}}{\partial z_{t_st_s}^{(d_0-1)}}\left(\frac{z_{t_st_s}^{n+m}}{\prod_{j=1}^{p}(z_{t_st_s}-a_j)^{d_j}}\right)\right)_{a_0}$$

$$+\ldots+\frac{1}{(d_p-1)!}\left(\frac{\partial^{(d_p-1)}}{\partial z_{t_st_s}^{(d_p-1)}}\left(\frac{z_{t_st_s}^{n+m}}{\prod_{j=0}^{p-1}(z_{t_st_s}-a_j)^{d_j}}\right)\right)_{a_p}$$

Again in this case we can see that limit in 18 exists if $\max\{b_{t_st_s}\} \leq \max\{b_{g_sg_s}\}$ and is ∞ otherwise.

■

Proof.[**of Claim 2**] Let n be an even number. Then there will be $\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n}$ many models for which we have $\mathcal{M} \models a_i \neq_G a_j \;\; 1 \leq i, j \leq n$. That is the number of models of $\mathcal{E}$ where the $a_1, ..., a_n$ are different with respect to $=_G$ and there will be at most $n^n\binom{2^{n^2}}{n-2}$ many models where not all of $a_1, ..., a_n$ are different. To see this notice that $\begin{pmatrix} 2^{n^2} \\ n \end{pmatrix}$ is the number of ways we can interpret G so that $a_1, ..., a_n$ are all different according to $=_G$. Let $P_1(x), ..., P_{2^{n^2}}(x)$ denote the sentences of the form $\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{n} \pm G(x, a_i, a_j)$. When G is interpreted on L^n each $a_i \;\; 1 \leq i \leq n$ will satisfy one of the $P_k(x) \;\; 1 \leq k \leq 2^{n^2}$. The fact that $a_1, ..., a_n$ are different according to G means that each $P_k(x)$ is satisfied by at most one a_i. So the number of ways we can interpret G such that $a_1, ..., a_n$ are all different with respect to $=_G$, will be the number of ways we can choose $P_{i_1}(x), ..., P_{i_n}(x)$ all different among $P_1(x), ..., P_{2^{n^2}}(x)$ (each being intended for a different a_i) and that will be $\binom{2^{n^2}}{n}$.
After G is interpreted and $a_1, ..., a_n$ are all chosen to be different according to $=_G$, R will put $a_1, ..., a_n$ into groups of 2. To see this notice that in $\mathcal{E}$, we have $\forall x \exists y(x \neq_G y \wedge R(x, y)$. So each element is paired with at least one element and it cannot be paired with more than one because if we have $R(x, y) \wedge R(x, z)$ then we should have $x =_G y$ or $x =_G z$ or $y =_G z$ but $a_1, ..., a_n$ are chosen to be different according to $=_G$. So the number of different possibilities for R will be the number of ways we can put $a_1, ..., a_n$, into groups of 2, that is $\frac{n!}{2^{\frac{n}{2}}}$ and this should be divided by $(\frac{n}{2})!$ because the order in which these groups of 2 are chosen is not important and so the number of possibilities for R will be $\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}$. Thus the number of models of size n for even n (where the a_i's are mutually different with respect to $=_G$) will be $\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n}$. For models where not all of a_i's are different according to $=_G$, assume that $n-2k$ of them are different and then we will take the sum over $k = 1, ..., \frac{n}{2}$. Notice that it is not possible to have an odd number of a_i's mutually different with respect to $=_G$ because R is dividing those elements of the model that are mutually different with respect to $=_G$ into groups of 2 and this will not be possible if the number of

these elements is odd. To see this notice that R will be grouping the elements of the model such that each group contains at least 2 different elements with respect to $=_G$ (because of the conjunct $\forall x \exists y(x \neq_G y \wedge R(x,y))$). On the other hand if a group contains more than 2 elements, say 3, $\mathcal{E}$ will force the third element to be equal (according to $=_G$) with one of the other two. So when all the elements of the model are different according to $=_G$, there cannot be any group with more than 2 elements hence R will be dividing the elements of the model into disjoint pairs. Then, the number of ways we can define G such that $n-2k$ of a_i's are different will be $\binom{2^{n^2}}{n-2k}$ and the number of ways we can put these $n-2k$ many a_i's into groups of 2 will be $\frac{(n-2k)!}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}$, whilst each of the remaining $2k$ elements, say $a_{n-2k+1}, ..., a_n$, can be equal (with respect to $=_G$) to any of the $n-2k$ elements, $a_1, ..., a_{n-2k}$, and so will belong to corresponding group of 2. Hence each of these $2k$ elements can belong to any of $\frac{n-2k}{2}$ groups and there will be $(\frac{n-2k}{2})^{2k}$ possibilities. Thus, for an even number n, the number of models of size n where $n-2k$ elements are different according to $=_G$ will be

$$(\frac{n-2k}{2})^{2k}\frac{(n-2k)!}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}\binom{2^{n^2}}{n-2k} \leq n^{n-1}\binom{2^{n^2}}{n-2} \tag{22}$$

and using (22) the total number of models of $\mathcal{E}$ of size n where not all the a_i's are different with respect to $=_G$ will be $\sum_{k=1}^{\frac{n}{2}}(\frac{n-2k}{2})^{2k}\frac{(n-2k)!}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}\binom{2^{n^2}}{n-2k} \leq \frac{n}{2}n^{n-1}\binom{2^{n^2}}{n-2} \leq n^n\binom{2^{n^2}}{n-2}$. And this gives us an upper bound on the number of models in this case. Hence for an even number n, we have $\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} \leq \#\mathcal{M}^n_{\mathcal{E}} \leq \frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} + n^n\binom{2^{n^2}}{n-2}$.
If $n = 2k+1$ is an odd number, notice that there will be no model of $\mathcal{E}$ where $a_1, ..., a_n$ are mutually different with respect to $=_G$. To see this remember that R will be grouping the elements of the model into disjoint pairs and this is not possible when the number of elements is odd. Thus the only models of $\mathcal{E}$ of size odd will be those in which some of the a_i's are equal according to $=_G$. In exactly the same way as above we can show that the number of models of size n for odd n, will be at most

$$n^n\binom{2^{n^2}}{n-1}$$

and this will be an upper bound on the number of models of $\mathcal{E}$ of size n where n is odd.

∎

Proof.[of Claim 3]

Let n be even. According to O there should exists at least one element a_i with $R(a_i, a_i)$. This means that O cannot have models of size even where all the elements are different with respect to $=_G$. To see this assume that all the elements are different according to $=_G$ and let a_i be such that $R(a_i, a_i)$ holds. Then $\neg R(a_i, a_j)$ for $i \neq j$ because otherwise we will have $R(a_i, a_i) \wedge R(a_i, a_j)$ and so $a_i =_G a_j$ which is a contradiction. On the other hand a_i will be the only element with $R(a_i, a_i)$ because if for $k \neq i$, $R(a_k, a_k)$ then $a_i =_G a_k$ which again is a contradiction. So R will connect a_i only to itself and then will divide the rest of the elements into groups of two which is impossible as there will be an odd number of elements left. So for an even number n, there will are no model of size n where the elements are all different with respect to $=_G$.
For the number of models of size n where not all the elements are different with respect to $=_G$, suppose first that there are $n - 2k$ distinguishable elements. There will be an element connected to itself through R which should be one of these $n-2k$ elements but as above this cannot be the case because there can be at most one of them with this property and if there exists one such element in the domain there will be an odd number left and it will not be possible to interpret R in a way to put them into groups of two. Hence there will be no model where $n - 2k$ elements are different with respect to $=_G$. It remain the case where the models are of size n for an even number n and $n - 2k + 1$ elements are different with respect to $=_G$. In this case exactly one of these $n - 2k + 1$ elements will be connected to itself and to no other of the remaining $n - 2k$ and the other $n - 2k$ will again be divided into groups of two. The rest of the domain ($2k - 1$ elements) each can be connected through R to one element in one of these groups of two or can be connected to the one element that is connected to itself. Hence the number of possibilities will be

$$\sum_{k=1}^{\frac{n}{2}-1} \frac{(n-2k)!}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}\binom{n-2k+1}{1}\binom{2^{n^2}}{n-2k+1}(\frac{n-2k+2}{2})^{2k-1} \leq$$

$$\sum_{k=1}^{\frac{n}{2}-1} \frac{n^{n-2k-1}}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}(\frac{n^{2k-1}}{2^{2k-1}})(n-2k+1)\binom{2^{n^2}}{n-2k+1} \leq$$

$$\sum_{k=1}^{\frac{n}{2}-1} n^{n-1}\binom{2^{n^2}}{n-1} \leq n^n\binom{2^{n^2}}{n-1}.$$

Thus for an even number n, the number of models of size n is at most $n^n\binom{2^{n^2}}{n-1}$. This gives an upper bound on the number of models of O of even size. For an odd number n, O has $\frac{n!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\binom{2^{n^2}}{n}$ many models where all the elements are different with respect to $=_G$. This is because we can choose n different elements with respect to $=_G$ in $\binom{2^{n^2}}{n}$ many ways and among them exactly one should be connected only to itself for which there are n possibilities and then the remaining $n-1$ should be divided into groups of two for which there are $\frac{(n-1)!}{2^{\frac{n-1}{2}}}$ possibilities. And there are at most $n^n\binom{2^{n^2}}{n-1}$ many models where not all the elements are different according to $=_G$ in the same way that it is calculated above. Hence $\frac{n!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\binom{2^{n^2}}{n}$ gives a lower bound on the number of models of O of odd size.

■