Soroush Rafiee Rad

Equivocation Axiom on First Order Languages

Abstract. In this paper we investigate some mathematical consequences of the Equivocation Principle, and the Maximum Entropy models arising from that, for first order languages. We study the existence of Maximum Entropy models for these theories in terms of the quantifier complexity of the theory and will investigate some invariance and structural properties of such models.

Keywords: Maximum Entropy models, Probabilistic reasoning, Equivocation principle.

1. Introduction

In this paper we study the most "uninformative" model for a probabilistic theory K over a first order language L. By a probabilistic theory we mean a set of assertions regarding the probabilities of some sentences in the language L. The theory K for our purpose is identified with a *satisfiable* set of linear constraints on these probabilities, of the form

$$\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i, \quad i = 1, 2, ..., m, \quad a_{ij}, b_j \in \mathbb{R},$$

where θ_j's are sentences from the language L. A categorical (non-probabilistic) theory $K = \{\theta_1, \ldots, \theta_n\}$ will be a special case by setting $K = \{w(\theta_1) = 1, \ldots, w(\theta_n) = 1\}$. A model for such a theory will then be a probability function over the sentences of L, which will be defined shortly, that satisfies the constraints given in K. The problem we are interested in is to investigate the most non-committal model of K. That is to investigate the probability function, amongst all that satisfy K, that admits the *Equivocation Principle:*

> Equivocation Principle: The assignment of probabilities should otherwise [beyond what is enforced by the constraints] be as equivocal as possible.

Such a probability function can be regarded as the most representative model of K, in the sense of best approximating a model that characterises K. In

Presented by **Daniele Mundici**; *Received* December 11, 2015

Studia Logica (2017) 105: 121–152
DOI: 10.1007/s11225-016-9684-x

other words, the Equivocation Principle ensures that the model satisfies K and remains as *free* as possible beyond that, in a sense analogous to that of *free* algebraic structures (free vector spaces, groups, etc) that are defined by a set of equations. With this intuition, the most representative model of K will be identified with its most uninformative model, i.e., the probability function that remains maximally *uninformative* beyond what is given in K.

This problem has attracted a lot of attention from different areas and the literature investigating it is extensive and spreads across several disciplines; from statistics [8,9] and physics [11,12], to computer science, pattern recognition [4], image processing [6], computational linguistics [3] as well as economics and finance [10,25]. It also plays a central role in Formal Epistemology and in particular the Objective Bayesian account [1,15,20,22,23]. This literature almost exclusively promotes some formalisation of the Equivocation Principle and the notion of un-informativeness that involves Shannon entropy. The major part of this literature, however, is concerned with propositional languages which have been extensively studied. Although the case of first order languages has been addressed, for example by Paris [15], Paris and Rafiee Rad [18] and Williamson [24], there are still gaps in the literature regarding a detailed analysis of the Equivocation Principle for first order languages. It is to this aspect of the literature that we hope to contribute in this paper.

In particular, we will not be concerned with the justification and defence of Maximum Entropy, as there is already a large literature addressing this issue from different perspectives, for example in the works of Paris and Vencovská, [16,17], where they argue on behalf of the Maximum Entropy from an axiomatic point of view and by adhering to a set of rationality principles (see also [15]) or in the recent works by Landes and Williamson [13, 14] where they argue for it from a decision theoretic perspective. What we will be concerned with here, is to investigate the extent to which Maximum Entropy models, however well justified, are well defined for constraints given on a first order language.

We will focus on Williamson's formalisation of the Equivocation Principle and investigate the most equivocal models in terms of the quantifier complexity of the theory under consideration. In this sense our work here is more in line with [18].

1.1. Preliminaries and Notation

Throughout this paper, we work with a first order language L with finitely many relation symbols, countably many constant symbols $a_1, a_2, a_3, ...$ and

no function symbols. We also assume that these individuals exhaust the universe. Let RL, FL and SL denote the sets of relation symbols, formulae and sentences for L respectively.

DEFINITION 1. $w : SL \to [0\,,1]$ is a probability function if for $\theta, \phi, \exists x\psi(x) \in SL$,
P1. If $\models \theta$ then $w(\theta) = 1$.
P2. $w(\theta \vee \phi) = w(\theta) + w(\phi) - w(\theta \wedge \phi)$.
P3. $w(\exists x\psi(x)) = \lim_{n\to\infty} w(\bigvee_{i=1}^{n} \psi(a_i))$.

DEFINITION 2. A probabilistic theory K is defined to be a *satisfiable* set of linear constraints of the form $\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i, \quad i = 1, 2, ..., m$, where $\theta_j \in SL$, $a_{ij}, b_j \in \mathbb{R}$ and w is a probability function.

DEFINITION 3. Let $\mathcal{L}$ be a finite propositional language with propositional variables $p_1, ..., p_n$. By *atoms* of $\mathcal{L}$ we mean the set of sentences $\{\, \alpha_i \,|\, i = 1, ...J\}$, of the form

$$\bigwedge_{i=1}^{n} p_i^{\epsilon_i}$$

where $\epsilon_i \in \{0, 1\}$, $p^1 = p$ and $p^0 = \neg p$.

For every $\phi \in S\mathcal{L}$ we can find a unique set of atoms $\Gamma_\phi \subseteq \{\, \alpha_i |\, i = 1, ..., J \,\}$ with $\models \phi \leftrightarrow \bigvee_{\alpha_i \in \Gamma_\phi} \alpha_i$. It is easy to check that $\Gamma_\phi = \{\, \alpha_j \,|\, \alpha_j \vDash \phi \,\}$. Since the α_i's are mutually inconsistent, for every probability function w, $w(\phi) = w(\bigvee_{\alpha_i \vDash \phi} \alpha_i) = \sum_{\alpha_i \vDash \phi} w(\alpha_i)$ and 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 determined uniquely by its values on the atoms, that is, by the vector $< w(\alpha_1), ..., w(\alpha_J) >$ in $\mathbb{D}^{\mathcal{L}} = \{\, \vec{x} \in \mathbb{R}^J \,|\, \vec{x} \geq 0, \sum_{i=1}^{J} x_i = 1\}$. On the other hand if $\vec{a} \in \mathbb{D}^{\mathcal{L}}$ we can define a probability function $w : S\mathcal{L} \to [0\,,\,1]$ by $w(\phi) = \sum_{\alpha_i \vDash \phi} a_i$ so that $< w(\alpha_1), ..., w(\alpha_J) >= \vec{a}$.

This gives a one to one correspondence between the probability functions on $S\mathcal{L}$ and the points in $\mathbb{D}^{\mathcal{L}}$. Given $K = \{\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i, \mid i = 1, 2, ..., m\}$, replacing each $w(\theta_j)$ with $\sum_{\alpha_i \vDash \theta_j} w(\alpha_i)$ and adding the equation $\sum_{i=1}^{J} w(\alpha_i) = 1$ we will get a system of linear equations $< w(\alpha_1), ..., w(\alpha_J) > A_K = \vec{b}_K$. Thus if the probability function w is a model for K (i.e. w satisfies constraints in K) the vector $< w(\alpha_1), ..., w(\alpha_J) >$ will be a solution for the equation $\vec{x} A_K = \vec{b}_K$. We will denote the set of non-negative solutions to this equation by $V^{\mathcal{L}}(K) = \{\, \vec{x} \in \mathbb{R}^J \,|\, \vec{x} \geq 0,\ \vec{x} A_K = \vec{b}_K \,\} \subseteq \mathbb{D}^{\mathcal{L}}$. In this setting, the question of choosing a probability function satisfying K will be equivalent to the question of choosing a point in $V^{\mathcal{L}}(K)$.

For a first order language, however, the atoms are not expressible as sentences since they will involve infinite conjunctions. Instead, in the first order case one works with the set of *state descriptions* for finite sub-languages that can play a similar role.

Definition 4. Let L be a first order language and let L^k be a sub-language of L with only constant symbols $a_1, ..., a_k$. The state descriptions of L^k are defined as the sentences $\Theta_1^{(k)}, ..., \Theta_{J_k}^{(k)}$ of the form

$$\bigwedge_{\substack{i_1,\dots,i_j\leq k \\ R_i\ j-ary \\ R_i\in RL}} R_i(a_{i_1}, ..., a_{i_j})^{\epsilon_{i_1,\dots,i_j}}$$

where $\epsilon_{i_1,...,i_j} \in \{0, 1\}$ and $R_i^0 = \neg R_i$ and $R_i^1 = R_i$.

The set of state descriptions of L^k is the set of term models of L with domain $\{a_1, \dots, a_k\}$.

Given a quantifier free sentence θ if k is the maximum such that a_k appears in θ, then θ can be regarded as a sentence of the propositional language $\mathcal{L}^{(k)}$ with propositional variables $R_i(a_{i_1}, ..., a_{i_j})$, $i_1, ..., i_j \leq k$, $R_i \in RL$. Notice that the state descriptions $\Theta_i^{(k)}$ are the atoms of $\mathcal{L}^{(K)}$ and so $\vDash \theta \leftrightarrow \bigvee_{\Theta_i^{(k)} \vDash \theta} \Theta_i^{(k)}$. Thus for every probability function w,

$$w(\theta) = w\left(\bigvee_{\Theta_i^{(k)} \vDash \theta} \Theta_i^{(k)}\right) = \sum_{\Theta_i^{(k)} \vDash \theta} w\left(\Theta_i^{(k)}\right),$$

and to determine $w(\theta)$ one needs to determine the values $w(\Theta_i^{(k)})$ in such a way that

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

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

to guarantee that w satisfies P1 and P2. By the following theorem of Gaifman [5], this will be enough to determine w on all SL.

Theorem 1. *Let $QFSL$ be the set of quantifier free sentences of L and 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 $QFSL$.

Definition 5. By *state descriptions of* L on $\{b_1, \ldots, b_r\}$ we mean sentences $\Psi(b_1, \ldots, b_r)$ of the form

$$\bigwedge_{\substack{a_{i_1},\ldots,a_{i_j} \subset \{b_1,\ldots,b_r\} \\ R_i \in RL,\ R_i\ j-ary}} R_i(a_{i_1}, ..., a_{i_j})^{\epsilon_{i_1,\ldots,i_j}}$$

where $\epsilon_{i_1,\ldots,i_j} \in \{0, 1\}$, $R_i^1 = R_i$, $R_i^0 = \neg R_i$ and $\{b_1, \ldots, b_r\} \subset \{a_1, a_2, \ldots\}$. If $\Theta^{(m)}$ is a state description of L^m with $m > r$ such that $\{b_1, \ldots, b_r\} \subset \{a_1, \ldots, a_m\}$, we say $\Psi(b_1, \ldots, b_r)$ is determined by $\Theta^{(m)}$ if and only if for all $R \in RL$ and all $t_1, \ldots, t_j \in \{b_1, \ldots, b_r\}$

$$\Psi(b_1, \ldots, b_r) \vDash R(t_1, \ldots, t_j) \iff \Theta^{(m)} \vDash R(t_1, \ldots, t_j).$$

Notice that a state description of L, $\Psi(b_1, \ldots, b_r)$, is a term models for L with domain $\{b_1, \ldots, b_r\}$.

Definition 6. Define the equivocator, $P_=$, as the probability function that for each k, assigns equal probabilities to the $\Theta_i^{(k)}$'s (the state descriptions of L^k). Notice that this determines $P_=$ on all quantifier free sentences, and by Theorem 1, on all SL.

1.2. Maximum Entropy

Shannon Entropy is a widely accepted measure for the information of a probability function [21]. For a probability function W defined on a set $A = \{a_1, \ldots, a_n\}$ i.e., $0 \leq W(a_i) \leq 1$ and $\sum_i W(a_i) = 1$, the Shannon entropy of W is defined as

$$E(W) = -\sum_{i=1}^{n} W(a_i) \log(W(a_i)).$$

Definition 7. An inference process, N, on L is a function that on each set of linear constraints K, returns a probability function on SL, say $N(K)$, that satisfies K.

We shall denote the inference process that on each set of constraints K, returns the Maximum Entropy solution of K by ME, that is, $ME(K)$ is the probability function satisfying K with maximum Shannon entropy. There are two approaches for defining ME on a set of constraints.
First consider a set of linear constraints K from a propositional language $\mathcal{L}$ with atoms $\alpha_1, \ldots, \alpha_J$. The first approach for defining the Maximum

Entropy solution for K is to take the unique probability function w, that satisfies K, or equivalently the unique point $\vec{w} \in V^L(K)$, with maximum Shannon Entropy

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

We notice that, when K involves only linear constraints, the set $V^L(K)$ is convex and so is the function $f(x) = -\sum_{i=1}^{J} x_i \log(x_i)$ and these guarantee the uniqueness.
The second approach for defining Maximum Entropy models followed for example by Williamson, [23], which we shall write as ME_W, uses relative Shannon Entropy. In this approach equivocation is achieved by minimising the information theoretic divergence from the probability function $P_=$ (defi 6). The idea here is that $P_=$ is the most uninformative probability function over $S\mathcal{L}$. In this sense, $P_=$ is taken as a point of reference and the information theoretic divergence of a probability function W from $P_=$ is taken as a measure for its informativeness. The information theoretic divergence of a probability function W from the probability function V is defined by:

$$RE(W, V) = \sum_{i=1}^{J} W(\alpha_i) \log\left(\frac{W(\alpha_i)}{V(\alpha_i)}\right).$$

Williamson defines the Maximum Entropy solution for a set of constraints K, $ME_W(K)$, as the unique probability function w, that minimises the relative entropy

$$\sum_{i=1}^{J} w(\alpha_i) \log\left(\frac{w(\alpha_i)}{P_=(\alpha_i)}\right)$$

or equivalently the unique point $\vec{w} \in V^L(K)$ with minimum $\sum_{i=1}^{J} w_i \log(\frac{w_i}{1/J})$.

Proposition 1. *Let $\mathcal{L}$ be a propositional language, K a set of linear constraints and $\psi \in S\mathcal{L}$. Then*

$$ME(K)(\psi) = ME_W(K)(\psi)$$

There have been proposals in the literature to generalise both definitions to first order languages. The obvious problem is that for first order languages one cannot express the Shannon Entropy or the relative Entropy using atomic sentences since in the first order case only the atoms of finite sub-languages are expressible in the language.

In [2], Barnett and Paris propose to define the Maximum Entropy solutions for a set of linear constraints, K, from a first order language L, as the limit of the Maximum Entropy solutions of K on finite sub-languages, L^k, as k increases. These finite sub-languages can be regarded as propositional languages for which the Maximum Entropy solutions are defined uniquely. More precisely, take a first order language L with relation symbols $R_1, \ldots, R_t$, domain $a_1, a_2, \ldots$, and a set of linear constraints K. Let $\mathcal{L}^{(r)}$ be the propositional language with $R_j(a_{i_1}, \ldots, a_{i_n})$, $1 \leq j \leq t$, $1 \leq i_1, \ldots, i_n \leq r$ as its propositional variables. Let k be the maximum such that a_k appears in K and define $(-)^{(r)} : SL^k \to S\mathcal{L}^{(r)}$ for $r > k$ by

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

Let $K^{(r)}$ be the result of replacing every θ appearing in K by $\theta^{(r)}$, then $K^{(r)}$ is a set of constraints over the propositional language $\mathcal{L}^{(r)}$.

Definition 8. (ME) Let L be a first order language and K as above. For a state description $\Theta_i^{(k)}$ of L^k,

$$ME(K)(\Theta_i^{(k)}) = \lim_{r \to \infty} ME(K^{(r)})(\Theta_i^{(k)}).$$

This defines ME on the state descriptions and thus the quantifier free sentences which is uniquely extended to all SL by Theorem 1.

To extend ME_W to first order languages, Williamson defines the r-divergence of a probability function W from a probability function V as

$$d_r(W, V) = \sum_{i=1}^{J_r} W(\Theta_i^{(r)}) \log \left(\frac{W(\Theta_i^{(r)})}{V(\Theta_i^{(r)})} \right)$$

where $\Theta_i^{(r)}$'s are the state descriptions of L^r. Then for probability functions W, V and U defined on SL, W is closer to U than V if $d_r(W, U) < d_r(V, U)$ for all r eventually. Williamson defines the Maximum Entropy solutions for K, $ME_W(K)$ as follows:

Definition 9. (ME_W) Let K be a set of linear constraints. A Maximum Entropy solution for K, is a probability function, w satisfying K such that

there is no other probability function v that satisfies K and no N such that for all $r > N$, $d_r(v, P_=) < d_r(w, P_=)$.

In this paper we will focus on Williamson's formulation of the Maximum Entropy models. Notice that defining maximum entropy models as the limit of such models for finite sub-languages suffers from the finite model problem, that is, it will fail when dealing with a constraint set K with no finite models. This is so because for such a set of constraints, K, the corresponding $K^{(r)}$ is not satisfiable. Williamson's definition, however, does not suffer from this problem as one does not need to consider probability functions defined on any finite sub-language.

2. The ME_W On First Order Languages

2.1. The ME_W On Unary Languages

We will start our investigation from first order languages with only unary predicates. Our goal in this section is to show that the ME_W model is unique for first order theories coming from a unary language. Let L be the first order language with only unary predicates $P_1, \ldots, P_n$ and domain $\{a_1, a_2, \ldots\}$, in [2], Barnett and Paris showed the following result.

Proposition 2. *For a set of satisfiable linear constraint K on a unary first order language L,*

$$ME(K)(\psi) = \lim_{r\to\infty} ME(K^{(r)})(\psi^{(r)})$$

is well defined for all $\psi \in SL$ and $ME(K)$ defined in this way is a probability function on SL that satisfies K.

To show that ME_W is unique we will show that the probability function defined by this limit, $ME(K)$, which is well defined and satisfies K by Proposition 2, is closer to $P_=$ than any other probability function that satisfies K. Hence $ME_W(K)$ will be uniquely defined on SL and in this sense our result in this section extends Proposition 1 to unary first order languages.

For a unary first order language L with predicate symbols $P_1, \ldots, P_n$, let Q_i, $i = 1, \ldots, J$ enumerate all the *formulae* of the form $\bigwedge_{j=1}^{n} P_i^{\epsilon_i}(x)$ where $\epsilon_j \in \{0, 1\}$ and $P_i^0 = \neg P_i$ and $P_i^1 = P_i$ and let α_i for $i = 1, ..., J^k$ enumerate the exhaustive and exclusive set of sentences of the form

$$\bigwedge_{j=1}^{k} Q_{m_j}(a_j).$$

LEMMA 1. (Barnett & Paris) *Any sentence* $\theta(a_1, \ldots, a_k) \in SL$ *is equivalent to a disjunction of the consistent sentences* $\phi_{i,\vec{\epsilon}}$ *of the form*

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

where $\vec{\epsilon} = (\epsilon_1, ..., \epsilon_J)$ *is a sequence of* 0*s and* 1*s and* $\models \neg(\phi_{i,\vec{\epsilon}} \wedge \phi_{j,\vec{\delta}})$ *when* $(i, \vec{\epsilon}) \neq (j, \vec{\delta})$.

Notice that since $\phi_{i,\vec{\epsilon}}$'s are mutually inconsistent and exhaustive, each state description of L^k satisfies exactly one of these sentences and so these $\phi_{i,\vec{\epsilon}}$'s give a partition of the state descriptions of L^k. The same is true for every L^r with $r > k$: Since $\phi_{i,\vec{\epsilon}}$'s are also sentences of L^r, every state description of L^r also satisfies exactly one of these $\phi_{i,\vec{\epsilon}}$'s.

Let w be a probability function on SL and w^k its restriction to SL^k. As pointed out above, the probability function w^k on SL^k can be identified with the vector $\vec{w}^k = (w(\zeta_1), \ldots, w(\zeta_{J^k})) \in \mathbb{D}^{(r)}$ where the ζ_j are the state descriptions of L^k. This is so, because any sentence of L^k can be written as a disjunction of a subset of these mutually inconsistent sentences. By Lemma 1 and the discussion above, the same holds for the sentences $\phi_{i,\vec{\epsilon}}$ and the same argument allows us to identify the probability function w^k on SL^k by its value on $\phi_{i,\vec{\epsilon}}$'s or equivalently by the vector $\vec{w}^k = (w(\phi_{i,\vec{\epsilon}}))_{i,\vec{\epsilon}}$. Take

$$\phi_{i,\vec{\epsilon}} = \alpha_i \wedge \bigwedge_{j=1}^{J} (\exists x Q_j(x))^{\epsilon_j}$$

with $\alpha_i = \bigwedge_{j=1}^{k} Q_{m_j}(a_j)$, and let $A_i = \{ m_j \mid j = 1, ..., k \}$, $P_{\vec{\epsilon}} = \{ j \mid \epsilon_j = 1 \}$ and $P_{i,\vec{\epsilon}} = \{ j \mid j \in P_{\vec{\epsilon}} \text{ and } j \notin A_i \}$. So

$$\phi_{i,\vec{\epsilon}}^{(r)} = \alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\epsilon_j} \equiv \bigvee_{\substack{m_j \in P_{\vec{\epsilon}} \text{for} j=k+1,\ldots,r \\ P_{i,\vec{\epsilon}} \subseteq \{ m_j \mid k+1 \leq j \leq r \}}} \left(\alpha_i \wedge \bigwedge_{j=k+1}^{r} Q_{m_j}(a_j) \right) \tag{III}$$

Setting $p_{\vec{\epsilon}} = |P_{\vec{\epsilon}}|$, *and* $p_{i,\vec{\epsilon}} = |P_{i,\vec{\epsilon}}|$ the number of disjuncts in (III) will be equal to

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

The disjunction in (III), is the disjunction of those state descriptions of L^r that logically imply $\phi^{(r)}_{i,\vec{\epsilon}}$. Notice that each state description of L^r implies precisely one of the sentences $\phi^{(r)}_{i,\vec{\epsilon}}$ (since every two of them are mutually inconsistent) and for each $\phi^{(r)}_{i,\vec{\epsilon}}$ there are precisely $\sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}(p_{\vec{\epsilon}}-j)^{r-k}$ many state description implying it (the number of disjuncts in (III)). Next, consider a set of constraints K and let k be the maximum such that a_k appears in K, then each sentence θ appearing in K is by Lemma 1 logically equivalent to a disjunction of sentences $\phi_{i,\vec{\epsilon}}$. Thus each $\theta^{(r)}$ appearing in $K^{(r)}$ is similarly equivalent to the corresponding disjunction of sentences $\phi^{(r)}_{i,\vec{\epsilon}}$.

Proposition 3. *If $W = ME(K) = \lim_{r\to\infty} ME(K^{(r)})$, for every state description $\Theta^{(r)}_j$ of L^r if $\Theta^{(r)}_j \vDash \phi^{(r)}_{i,\vec{\epsilon}}$, then*

$$W(\Theta^{(r)}_j) = \frac{W(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}(p_{\vec{\epsilon}}-j)^{r-k}}. \tag{IV}$$

Proof. Remember that any sentence in K (resp. $K^{(r)}$) is equivalent to a disjunction of sentences $\phi_{i,\vec{\epsilon}}$ (resp. $\phi^{(r)}_{i,\vec{\epsilon}}$) and that $\phi^{(r)}_{i,\vec{\epsilon}}$'s partition the state descriptions of L^r. What (IV) asserts is that all state descriptions in a partition cell (all those satisfying the same $\phi^{(r)}_{i,\vec{\epsilon}}$) receive equal probability by W. To see this notice that if $\Theta^{(k)}_i$ and $\Theta^{(k)}_j$ are state descriptions that satisfy $\phi^{(r)}_{i,\vec{\epsilon}}$ then $ME(K^{(r)})(\Theta^{(k)}_i) = ME(K^{(r)})(\Theta^{(k)}_j)$ otherwise take probability function v on SL^r with

$$v(\Theta^{(k)}_l) = ME(K^{(r)})(\Theta^{(k)}_l) \quad l \neq i,j$$

$$v(\Theta^{(k)}_i) = v(\Theta^{(k)}_j) = \frac{ME(K^{(r)})(\Theta^{(k)}_j) + ME(K^{(r)})(\Theta^{(k)}_j)}{2}$$

then v satisfies $K^{(r)}$ because it assigns the same probabilities to $\phi^{(r)}_{i,\vec{\epsilon}}$'s as $ME(K^{(r)})$ while $E(ME(K^{(r)})) < E(v)$ which is a contradiction with $ME(K^{(r)})$ being the maximum entropy solution to $K^{(r)}$. Thus for every r, $ME(K^{(r)})(\Theta^{(k)}_i) = ME(K^{(r)})(\Theta^{(k)}_j)$ and since $W = \lim_{r\to\infty} ME(K^{(r)})$ we have $W(\Theta^{(k)}_i) = W(\Theta^{(k)}_j)$. ∎

For a given K, let k be the upper bound on i such that a_i appears in K as before and let $\phi_{i,\vec{\epsilon}}$ be as defined in Lemma 1. Then by Lemma 1 and Proposition 3 the probability function $ME(K^r)$ is identified with its values on $\phi^{(r)}_{i,\vec{\epsilon}}$. What is important to notice here, and is also the main reason for working with these sentences $\phi^{(r)}_{i,\vec{\epsilon}}$ is that the number of these sentences is

independent of r. Notice that as we move from L^r to L^{r+1} the number of state descriptions that satisfy each $\phi^{(r)}_{i,\vec{\epsilon}}$ changes but not the number of these sentences. This means we can represent the ME solution for $K^{(r)}$ and $K^{(r+1)}$ (on languages L^r and L^{r+1}) with vectors of the same length.

Next, for $W = \lim_{r\to\infty} ME(K^{(r)})$, let W^r, be the restriction of W to SL^r and let $\alpha_1, \ldots, \alpha_{J^r}$ be the state descriptions of L^r, then

$$\begin{aligned} E(W^r) &= -\sum_{i=1}^{J^r} W^r(\alpha_i)\log(w^r(\alpha_i)) \\ &= -\sum_{i,\vec{\epsilon}} W^r(\phi^{(r)}_{i,\vec{\epsilon}})\log\left(\frac{W^r(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}(p_{\vec{\epsilon}}-j)^{r-k}}\right) \\ &= -\sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log(W^r(\phi_{i,\vec{\epsilon}})) + (r-k)\sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log(p_{\vec{\epsilon}}) \\ &\quad + \sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log\left(\sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}\left(1-\frac{j}{p_{\vec{\epsilon}}}\right)^{r-k}\right). \end{aligned}$$

Let $\delta(W,r) = \sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log\left(\sum_{j=0}^{p_{i,\vec{\epsilon}}}(-1)^j\binom{p_{i,\vec{\epsilon}}}{j}(1-\frac{j}{p_{\vec{\epsilon}}})^{r-k}\right)$. So

$$E(W^r) = -\sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log(W^r(\phi_{i,\vec{\epsilon}})) + (r-k)\sum_{i,\vec{\epsilon}} W^r(\phi_{i,\vec{\epsilon}})\log(p_{\vec{\epsilon}}) + \delta(W,r) \tag{V}$$

The summation is over a finite number of terms and as $r \to \infty$, $\left(1-\frac{j}{p_{\vec{\epsilon}}}\right)^{r-k}$ $\to 0$ and consequently, as $r \to \infty$, $\delta(W,r) \to 0$. Also notice that in the same way we can represent $E(U^r)$ as (V) for any probability function U that satisfies (IV). We are now in the position to state the main result of this section.

THEOREM 3. *Let L be a language with only finitely many unary predicates and constant symbols $a_1, a_2, \ldots$. Let K be a finite set of linear constraints as before. Then $ME_W(K)$ is unique and agrees with $ME(K)$.*

PROOF. We will show that for $W = \lim_{n\to\infty} ME(K^{(n)})$

$$(\forall w \in V^L(K)\ \ ((w \neq W) \Rightarrow \exists N \forall n \geq N\ d_n(W, P_=) \leq d_n(w, P_=))$$

where $V^L(K)$ is the set of probability functions that satisfy K. Notice that this proves something stronger than what is required by Definition 9. Definition 9 requires that no probability function is closer to $P_=$ than W on L^n for all n eventually. We shall prove that the W given here is closer than any

probability function to $P_=$ on L^n for all n eventually thus establishing both the existence and the uniqueness of the Maximum Entropy solution for K. Suppose not and let $w \neq W$ be a probability function satisfying K such that for infinitely many n,

$$d_n(w, P_=) < d_n(W, P_=). \tag{VI}$$

Notice that

$$d_n(w, P_=) < d_n(W, P_=) \iff E(W^n) < E(w^n). \tag{VII}$$

where W^n and w^n are restrictions of W and w to L^n. We will make use of the following claim: ■

Claim 1.

$$w(\phi_{i,\vec{\epsilon}}) = \lim_{n\to\infty} w(\phi^{(n)}_{i,\vec{\epsilon}}) \tag{VIII}$$

$$W(\phi_{i,\vec{\epsilon}}) = \lim_{n\to\infty} W(\phi^{(n)}_{i,\vec{\epsilon}}) \tag{IX}$$

By Lemma 1, each sentence in K is logically equivalent to a disjunction of sentences $\phi_{i,\vec{\epsilon}}$ similarly each $\theta^{(n)}$ in $K^{(n)}$ is equivalent to the corresponding disjunction of sentences $\phi^{(n)}_{i,\vec{\epsilon}}$. If we fix an order for these $\phi_{i,\vec{\epsilon}}$'s, and let $\vec{x} =< w(\phi_{i,\vec{\epsilon}}) >_{i,\vec{\epsilon}}$, then, as before, the knowledge base $K^{(n)}$ will be equivalent to a system of linear equations $\vec{x}A_K = \vec{b}$. As explained above the number of the sentences $\phi_{i,\vec{\epsilon}}$ does not depend on n, and so the matrix A_K will not depend on n. Let

$$X = \left\{ \vec{x} \,|\, \vec{x}A_K = \vec{b}\} \quad Y = \{\vec{x} \in X \,|\, \sum_{i,\vec{\epsilon}} x_{i,\vec{\epsilon}} \log p_{\vec{\epsilon}} \text{ is maximal.} \right\}$$

It can be easily checked that X and Y are convex. Let $\vec{v} \in Y$ be the point for which $-\sum_{i,\vec{\epsilon}} v(\phi_{i,\vec{\epsilon}}) \log(v(\phi_{i,\vec{\epsilon}}))$ is maximal. This $\vec{v}$ is unique by convexity of Y and let $W^n = ME(K^{(n)})$ so $W^n \in X$ and $E(W^n)$ is maximal, so in particular, $E(\vec{v}) \leq E(W^n)$ and thus by (V)

$$\begin{aligned} &-\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(v_{i,\vec{\epsilon}}) + (n-k) \sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(p_{\vec{\epsilon}}) + \delta(v, n) \\ &\leq -\sum_{i,\vec{\epsilon}} W^n(\phi^{(n)}_{i,\vec{\epsilon}})) \log(W^n(\phi_{i,\vec{\epsilon}})^{(n)})) \\ &\quad +(n-k) \sum_{i,\vec{\epsilon}} W^n(\phi^{(n)}_{i,\vec{\epsilon}})) \log(p_{\vec{\epsilon}}) + \delta(W^n, n) \end{aligned}$$

hence

$$\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(p_{\vec{\epsilon}}) - \sum_{i,\vec{\epsilon}} W^n(\phi_{i,\vec{\epsilon}}^{(n)})) \log(p_{\vec{\epsilon}})$$

$$\leq \frac{-\sum_{i,\vec{\epsilon}} W^n(\phi_{i,\vec{\epsilon}}^{(n)})) \log(W^n(\phi_{i,\vec{\epsilon}}^{(n)}))) + \sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(v_{i,\vec{\epsilon}}) + \delta(W^n, n) - \delta(\vec{v}, n)}{n-k} \tag{X}$$

and by the choice of $\vec{v} \in X$

$$0 \leq \sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(p_{\vec{\epsilon}}) - \sum_{i,\vec{\epsilon}} W^n(\phi_{i,\vec{\epsilon}}^{(n)})) \log(p_{\vec{\epsilon}}). \tag{XI}$$

Since as $n \to \infty$ the right hand side of (X) approaches 0, we have, as $n \to \infty$

$$\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}} \to \sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log p_{\vec{\epsilon}} \tag{XII}$$

since $W = \lim_{n\to\infty} W^n$. By the choice of $\vec{v}$ we should have $\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}$ is maximal.

Assuming Claim 1, let $\Theta_1^{(n)}, \ldots, \Theta_{J^n}^{(n)}$ range over the state descriptions of L^n and take n large and satisfying (VI) and we have

$$\sum_{i=1}^{J^n} w(\Theta_i^{(n)}) \log w(\Theta_i^{(n)}) < \sum_{i=1}^{J^n} W(\Theta_i^{(n)}) \log W(\Theta_i^{(n)}).$$

Hence

$$\sum_{i,\vec{\epsilon}} \sum_{\Theta^{(n)} \vDash \phi_{i,\vec{\epsilon}}^{(n)}} w(\Theta^{(n)}) \log w(\Theta^{(n)}) < \sum_{i,\vec{\epsilon}} \sum_{\Theta^{(n)} \vDash \phi_{i,\vec{\epsilon}}^{(n)}} W(\Theta^{(n)}) \log W(\Theta^{(n)}). \tag{XIII}$$

In this inequality the left hand side is, by convexity, at least

$$\sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}^{(n)}) \log \left(\frac{w(\phi_{i,\vec{\epsilon}}^{(n)})}{p_{\vec{\epsilon}}^{n-k} \sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} (1 - \frac{j}{p_{\vec{\epsilon}}})^{n-k}} \right).$$

Remember that if $\Theta_j^{(n)}, \Theta_k^{(n)}$ are state descriptions of L^n that logically imply the same $\phi_{i,\vec{\epsilon}}^{(n)}$ then $W(\Theta_j^{(n)}) = W(\Theta_k^{(n)})$. So the right hand side of (XIII) is actually equal to

$$\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}^{(n)}) \log \left(\frac{W(\phi_{i,\vec{\epsilon}}^{(n)})}{p_{\vec{\epsilon}}^{n-k} \sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} (1 - \frac{j}{p_{\vec{\epsilon}}})^{n-k}} \right).$$

Simplifying this gives

$$\sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}^{(n)}) \log\left(w(\phi_{i\vec{\epsilon}}^{(n)})\right) - (n-k)\sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}^{(n)}) \log p_{\vec{\epsilon}} + \delta(w,n)$$
$$< \sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}^{(n)}) \log\left(W(\phi_{i,\vec{\epsilon}}^{(n)})\right) - (n-k)\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}^{(n)}) \log p_{\vec{\epsilon}} + \delta(W,n) \tag{XIV}$$

where $\delta(W,n), \delta(w,n) \to 0$ as $n \to \infty$. Hence, using (VIII), (IX), we must have

$$\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}} \ \leq \ \sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}. \tag{XV}$$

By (XII) and the explanation immediately after that, $\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}$ is maximal so from (XV) we should have $\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}} = \sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}$. Using (VIII), (IX) and (XIV) it must be the case that

$$\sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log\left(w(\phi_{i,\vec{\epsilon}})\right) \ \leq \ \sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log\left(W(\phi_{i,\vec{\epsilon}})\right) \tag{XVI}$$

Next notice that $\vec{W^n} =< W^n(\phi_{i,\vec{\epsilon}}) >_{i,\vec{\epsilon}}$ is a bounded sequence and so has a convergent subsequence. The limit of this subsequence will also be in Y by (XII). However, by (XI) and (XII) we should have

$$-\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(v_{i,\vec{\epsilon}}) - \delta(W^n,n) + \delta(\vec{v},n) \leq -\sum_{i,\vec{\epsilon}} W^n(\phi_{i,\vec{\epsilon}}) \log(W^n(\phi_{i,\vec{\epsilon}}))$$

So if $\vec{t}$ is the limit of any convergent subsequence of $\vec{W^n}$ then $\vec{t} \in Y$ and

$$-\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(v_{i,\vec{\epsilon}}) \leq -\sum_{i,\vec{\epsilon}} t_{i,\vec{\epsilon}} \log(t_{i,\vec{\epsilon}}).$$

But by the choice of $\vec{v} \in Y$ we should have $\vec{t} = \vec{v}$ and thus $W = \lim_{n\to\infty} ME(K^{(n)}) = lim_{n\to\infty} W^n = v$. From $v = W$ and by the choice of $v \in Y$ that maximises both $\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log p_{\vec{\epsilon}}$ (in X) and $\sum_{i,\vec{\epsilon}} v_{i,\vec{\epsilon}} \log(v_{\vec{\epsilon}})$ (in Y), for every, $\vec{U} \in X$ we should have $E(\vec{U}) \leq E(W)$ and moreover, by uniqueness of $W \in Y$ if $\vec{U} \in Y$ and $U \neq W$ then $E(\vec{U}) < E(W)$. In particular for w in (VI), we have $E(w) \leq E(W)$

$$\sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log W(\phi_{i\vec{\epsilon}}) - (n-k)\sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}} + \delta(W,n)$$
$$\leq \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}}) - (n-k)\sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}} + \delta(w,n). \tag{XVII}$$

with $\delta(W,n), \delta(w,n) \to 0$ as $n \to \infty$. In consequence, for large n, we have $\sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log W(\phi_{i\vec{\epsilon}}) - (n-k) \sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}}$

$$\leq \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}}) - (n-k) \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}}. \tag{XVIII}$$

But from (XVIII) and $\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}} = \sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}$ and we have

$$\sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log W(\phi_{i\vec{\epsilon}}) \leq \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}})$$

and because $\sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}$ is maximal, $w \in Y$ and since $w \neq W$ from uniqueness of W we should have the strict inequality

$$\sum_{i,\vec{\epsilon}} W(\phi_{i\vec{\epsilon}}) \log W(\phi_{i\vec{\epsilon}}) < \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}})$$

and these give a contradiction with (XV). To complete the proof, it remains to prove Claim 1.

Proof of Claim. **1.** For a probability function v and distinct Q_i, Q_j

$$v(\exists x Q_i(x)) = \lim_{n\to\infty} v\left(\bigvee_{k=1}^{n} Q_i(a_k)\right) = \lim_{n\to\infty} v\left(\bigvee \Gamma^{(n)}_{Q_i}\right),$$

where $\Gamma^{(n)}_{Q_i}$ are those state descriptions of L^n containing as a conjunct $Q_i(a_j)$ for some $1 \leq j \leq n$. Similarly (see Chapter 11 in [15])

$$v(\exists x Q_i(x) \wedge \exists x Q_j(x)) = v(\exists x, y Q_i(x) \wedge Q_j(y)) = \lim_{n\to\infty} v\left(\bigvee \Gamma^{(n)}_{Q_i,Q_j}\right). \tag{XIX}$$

$$v(\exists x Q_i(x) \wedge \neg\exists x Q_j(x)) = v(\exists x Q_i(x)) - v(\exists x Q_i(x) \wedge \exists x Q_j(x)) \tag{XX}$$

$$= \lim_{n\to\infty} v\left(\bigvee \Gamma^{(n)}_{Q_i}\right) - \lim_{n\to\infty} v\left(\bigvee \Gamma^{(n)}_{Q_i,Q_j}\right) \tag{XXI}$$

$$= \lim_{n\to\infty} v\left(\bigvee \left(\Gamma^{(n)}_{Q_i} - \Gamma^{(n)}_{Q_i,Q_j}\right)\right) \tag{XXII}$$

$$= \lim_{n\to\infty} v\left(\Gamma^{(n)}_{Q_i,\neg Q_j}\right) \tag{XXIII}$$

where $\Gamma^{(n)}_{Q_i,\neg Q_j}$ are those state descriptions of L^n which contain $Q_i(a_k)$ as a conjunct for some $1 \leq k \leq n$ but do not contain as a conjunct $Q_j(a_k)$ for any k. We will now show that

$$v\left(\bigwedge_{k=1}^{m}\exists xQ_k(x)\wedge\bigwedge_{l=m+1}^{J}\neg\exists xQ_l(x)\right)=\lim_{n\to\infty}v\left(\bigvee\Gamma^{(n)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J}\right)\tag{XXIV}$$

by induction on $J-m$. The result for $J-m=0$ is given by the following theorem proved in [15].

Proposition 4. *For $v:SL\to[0\,,\,1]$ satisfying (P1-3) introduced in the 1.1 and $\psi(x)\in FL$,*

$$v(\exists x\psi(x))=sup_r v\left(\bigvee_{i=1}^{r}\psi(a_i)\right).$$

So we will have

$$v\left(\bigwedge_{k=1}^{m}\exists xQ_k(x)\right)=\lim_{n\to\infty}v\left(\bigwedge_{k=1}^{m}\bigvee_{i=1}^{n}Q_k(a_i)\right)=\lim_{n\to\infty}v\left(\bigvee\Gamma^{(n)}_{Q_1,\ldots,Q_m}\right).$$

Assume that (XXIV) is true for $J-m$. Then

$$\begin{aligned}v\left(\bigwedge_{k=1}^{m}\exists xQ_k(x)\wedge\bigwedge_{k=m+1}^{J+1}\neg\exists xQ_k(x)\right)&=v\left(\bigwedge_{k=1}^{m}\exists xQ_k(x)\wedge\bigwedge_{k=m+1}^{J}\neg\exists xQ_k(x)\right)\\&\quad-v\left(\bigwedge_{k=1}^{m}\exists xQ_k(x)\wedge\bigwedge_{k=m+1}^{J}\neg\exists xQ_k(x)\wedge\exists xQ_{J+1}(x)\right)\\&=\lim_{n\to\infty}v\left(\bigvee\Gamma^{(n)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J}\right)\\&\quad-\lim_{n\to\infty}v\left(\bigvee\Gamma^{(n)}_{Q_1,\ldots,Q_m,Q_{J+1},\neg Q_{m+1},\ldots,\neg Q_J}\right)\\&=\lim_{n\to\infty}v\left(\bigvee\Gamma^{(n)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J,\neg Q_{J+1}}\right)\end{aligned}$$

as required. Now we have

$$\begin{aligned}w(\phi_{i,\vec{\epsilon}})&=w\left(\alpha_i\wedge\bigwedge_{j=1}^{J}(\exists xQ_j(x))^{\epsilon_j}\right)\\&=\lim_{n\to\infty}w\left(\bigvee\Gamma^{(n)}_{\alpha_i,Q_{j_1},\ldots,Q_{j_m},\neg Q_{j_{m+1}},\ldots,\neg Q_{j_J}}\right)\\&=\lim_{n\to\infty}w\left(\alpha_i\wedge\bigwedge_{j=1}^{J}\left(\bigvee_{l=1}^{n}Q_j(a_l)\right)^{\epsilon_j}\right)=\lim_{n\to\infty}w(\phi^{(n)}_{i,\vec{\epsilon}})\end{aligned}$$

where $\epsilon_{j_1},\ldots,\epsilon_{j_m}=1$ and $\epsilon_{j_{m+1}},\ldots,\epsilon_{j_J}=0$ and similarly for W. ∎

So the Maximum Entropy model for a set of constraints K as characterised by ME_W, is unique for unary languages. We will now move to general polyadic languages.

2.2. ME_W and the General Polyadic Case

In this section we investigate the existence of Maximum Entropy solutions for sets of constraints from a general polyadic language. We will show by an example that there exists a set of constraints K with quantifier complexity of Σ_2, such that the closest solution of K to $P_=$, in the sense of Definition 9, does not exist uniquely. In particular, we will show that for any probability function w satisfying K one can find a probability function W closer to $P_=$ than w that also satisfies K. Williamson anticipates cases where the closest probability function to $P_=$ does not exist and addresses this by considering "sufficient closeness" to $P_=$ where the sufficiency is assumed to be determined on contextual and pragmatic grounds. Nevertheless, our example below not only establishes a case like that, but also makes the underlying reasons precise.

Example Let L be a first order language with only a binary relation symbol R and $K = \{\, w(\exists x \forall y\ R(x,y)) = 1 \,\}$. Suppose $ME_W(K)$ is uniquely defined and let $w = ME_W(K)$. So w is a probability function on SL and $w(\exists x \forall y\ R(x,y)) = 1$. We will show that there is some probability function W on SL, also satisfying K, such that for each N there will be some $r > N$ with $d_r(W, P_=) < d_r(w, P_=)$. This will give a contradiction to $w = ME_W(K)$. To see this let $e_i = w(\forall y\, R(a_i, y))$, pick k such that $e_k > 0$ and let r be large so in particular $e_k > 2^{-r}$.

Claim 2. *Let w be defined on SL and define W^r on the state description $\Theta_{\vec{\epsilon}} = \bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i, a_j)$ of L^r as*

$$W^r(\Theta_{\vec{\epsilon}}) = 2^{-r} w \left(\bigwedge_{\substack{1 \le i,j \le r \\ i \ne k}} R^{\epsilon_{ij}}(a_i, a_j) \right),$$

then $d_r(W, P_=) < d_r(w, P_=)$.

We will now proceed to define the required probability function W on SL. We consider two cases:

Case 1 There are arbitrarily large k such that $e_k > 0$.

In this case pick an infinite sequence $k_0 < k_1 < k_2 < \ldots$ of such k and define W on L^{r_s} where, $k_{s-1} \le r_s \le k_s - 1$, $s \ge 2$

$$W\left(\bigwedge_{i,j=1}^{r_s} R^{\epsilon_{ij}}(a_i,a_j)\right)$$
$$= 2^{-r_s} w\left(\bigwedge_{\substack{i,j=1\\ i\neq k_m, 0\leq m<s}}^{r_s} R^{\epsilon_{ij}}(a_i,a_j) \wedge \bigwedge_{m=1}^{s-1}\bigwedge_{j=1}^{r_s} R^{\epsilon_{k_{m-1}j}}(a_{k_m},a_j)\right).$$

An explanation here is that in forming W we use w but replace a_{k_0} by a 'random element', replace a_{k_1} by a_{k_0}, a_{k_2} by a_{k_1} and so on. The net effect of these constructions is that for W

$$W\left(\bigvee_{i=1}^{r_s} \forall y\, R(a_i,y)\right) \geq w\left(\bigvee_{i=1}^{r_{s-1}} \forall y\, R(a_i,y)\right).$$

To see this notice that

$$W\left(\bigvee_{i=1}^{r_s}\bigwedge_{j=1}^{n} R(a_i,a_j)\right)$$
$$\geq W\left(\bigvee_{\substack{i=1\\ i\neq k_0,\ldots,k_{s-2}}}^{r_{s-1}} \bigwedge_{j=1}^{n} R(a_i,a_j) \vee \bigvee_{m=1}^{s-1}\bigwedge_{j=1}^{n} R(a_{k_m},a_j)\right)$$
$$= w\left(\bigvee_{i=1}^{r_{s-1}}\bigwedge_{j=1}^{n} R(a_i,a_j)\right)$$

Taking the limit as $n\to\infty$ here gives

$$W(\bigvee_{i=1}^{r_s} \forall y R(a_i,y)) \geq w(\bigvee_{i=1}^{r_{s-1}} \forall y R(a_i,y))$$

and hence by taking the limit as $s\to\infty$,

$$W(\exists x \forall y R(x,y)) \geq w(\exists x \forall y R(x,y)).$$

Hence we have $W(\exists x\, \forall y R(x,y)) = 1$. This establishes that W also satisfies K.

Let, $k_{s-1} \leq r \leq k_s - 1$ and define w' on SL^r as follows:

$$w'\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right) = 2^{-r} w\left(\bigwedge_{\substack{i,j=1\\ i\neq k_{s-1}}}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right).$$

So

$$\sum_{\vec{\epsilon}} W\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)\log W\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)$$
$$=\sum_{\vec{\epsilon}} w'\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)\log w'\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)$$

and so $d_r(W,P_=)=d_r(w',P_=)$.
Using Claim 2, with r sufficiently large, $d_r(w',P_=)<d_r(w,P_=)$ and so

$$d_r(W,P_=)<d_r(w,P_=).$$

as required.

Case 2 There is some g such that $e_k=0$ for $k\geq g$. In this case pick a $0<j$ such that $e_j>0$ and the permutation σ of $\mathbb{N}^+$ such that for $i\neq j,g+1$, $\sigma(i)=i$ and $\sigma(j)=g+1$ and $\sigma(g+1)=j$. For $r\in\mathbb{N}^+$ let

$$W\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)$$
$$=2^{-1}\left(w\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i,a_j)\right)+w\left(\bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_{\sigma(i)},a_{\sigma(j)})\right)\right).$$

Then for $n>g$,

$$W\left(\bigvee_{i=1}^{g+1}\bigwedge_{k=1}^{n} R(a_i,a_k)\right)$$
$$=2^{-1}\left(w(\bigvee_{i=1}^{g+1}\bigwedge_{k=1}^{n} R(a_i,a_k))+w(\bigvee_{i=1}^{g+1}\bigwedge_{k=1}^{n} R(a_{\sigma(i)},a_{\sigma(k)}))\right)$$

Since $\{1,2,...,g+1\}=\{\sigma(1),\sigma(2),...,\sigma(g+1)\}$ we will have

$$W\left(\bigvee_{i=1}^{g+1}\bigwedge_{k=1}^{n} R(a_i,a_k)\right)=w\left(\bigvee_{i=1}^{g+1}\bigwedge_{k=1}^{n} R(a_i,a_k)\right).$$

Taking the limit as $n\to\infty$ and noticing that $w(\forall yR(a_i,y))=0$ for $i>g+1$,

$$W\left(\bigvee_{i=1}^{g+1}\forall yR(a_i,y)\right)=w\left(\bigvee_{i=1}^{g+1}\forall yR(a_i,y)\right)=w(\exists x\forall yR(x,y))=1.$$

We show that for large r, $d_r(W, P_=) < d_r(w, P_=)$. To show this it is enough to show that

$$-\sum_{\vec{\epsilon}} W(\Theta_{\vec{\epsilon}}) \log W(\Theta_{\vec{\epsilon}}) > -\sum_{\vec{\epsilon}} w(\Theta_{\vec{\epsilon}}) \log w(\Theta_{\vec{\epsilon}}). \tag{XXV}$$

Notice that the permutation σ can be also considered as a permutation of state descriptions and let $\sigma(\Theta_{\vec{\epsilon}})$ have the obvious meaning. If $\sigma(\Theta_{\vec{\epsilon_1}}) = \Theta_{\vec{\epsilon_2}}$ then $\sigma(\Theta_{\vec{\epsilon_2}}) = \Theta_{\vec{\epsilon_1}}$. So to show (XXV) it is enough to show that for each $\vec{\epsilon}$,

$$\begin{aligned} &W(\Theta_{\vec{\epsilon}}) \log W(\Theta_{\vec{\epsilon}}) + W(\Theta_{\vec{\epsilon'}}) \log W(\Theta_{\vec{\epsilon'}}) \\ &\quad \leq w(\Theta_{\vec{\epsilon}}) \log w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}}) \log w(\Theta_{\vec{\epsilon'}}) \end{aligned} \tag{XXVI}$$

where $\Theta_{\vec{\epsilon'}} = \sigma(\Theta_{\vec{\epsilon}})$ and that this inequality is strict for some $\Theta_{\vec{\epsilon}}$ eventually. But (XXVI) is by definition of W:

$$\begin{aligned} &\frac{w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}})}{2} \log \left(\frac{w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}})}{2} \right) \\ &\quad + \frac{w(\Theta_{\vec{\epsilon'}}) + w(\Theta_{\vec{\epsilon}})}{2} \log \left(\frac{w(\Theta_{\vec{\epsilon'}}) + w(\Theta_{\vec{\epsilon}})}{2} \right) \\ &\leq w(\Theta_{\vec{\epsilon}}) \log w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}}) \log w(\Theta_{\vec{\epsilon'}}) \end{aligned}$$

that is

$$\begin{aligned} &\big(w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}})\big) \log\left(\frac{w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}})}{2}\right) \\ &\quad \leq w(\Theta_{\vec{\epsilon}}) \log w(\Theta_{\vec{\epsilon}}) + w(\Theta_{\vec{\epsilon'}}) \log w(\Theta_{\vec{\epsilon'}}) \end{aligned} \tag{XXVII}$$

which holds by the convexity of the function $x \log x$. Furthermore this inequality will eventually (for large r) be strict for some $\Theta_{\vec{\epsilon}}$ because otherwise we will have $W \restriction_{L^r} = w \restriction_{L^r}$ but

$$W(\forall y R(a_{g+1}, y)) = 2^{-1} w(\forall y R(a_j, y)) = 2^{-1} e_j > 0$$

while by the choice of g, $w(\forall y R(a_{g+1}, y)) = 0$ so $W \neq w$ and thus there exists some M such that for $r > M$, $W \restriction_{L^r} \neq w \restriction_{L^r}$.

So with any probability function w satisfying this K, one can always use w to construct a probability function W, also satisfying K, that is closer to the $P_=$ on L^n for all n eventually and thus increase the entropy in the sense of ME_W. Hence the closest solution of K to $P_=$ does not exist and ME_W fails to provide a generalisation of the Maximum Entropy solution applicable in general. To complete this example we will now prove Claim 2.

Proof of Claim. **2.** By (VII) it is enough to show that

$$-\sum_{\vec{\epsilon}} W(\Theta_{\vec{\epsilon}}) \log W(\Theta_{\vec{\epsilon}}) > -\sum_{\vec{\epsilon}} w(\Theta_{\vec{\epsilon}}) \log w(\Theta_{\vec{\epsilon}}). \tag{XXVIII}$$

Let δ and τ respectively range over the maps from

$$\{\, \langle i,j\rangle \,|\, 1 \le i,j \le r,\, i \ne k \,\} \to \{0,1\} \text{ and } \{\, \langle k,j\rangle \,|\, 1 \le j \le r,\, \} \to \{0,1\}$$

Then (XXVIII) will be

$$-\sum_{\vec{\delta}\cup\vec{\tau}} W(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log W(\Theta_{\vec{\delta}\cup\vec{\tau}}) > -\sum_{\vec{\delta}\cup\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w(\Theta_{\vec{\delta}\cup\vec{\tau}}).$$

To show this we will show that for each $\vec{\delta}$,

$$-\sum_{\vec{\tau}} W(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log W(\Theta_{\vec{\delta}\cup\vec{\tau}}) \ge -\sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \tag{XXIX}$$

and that the inequality should be strict for some $\vec{\delta}$. For two state descriptions $\Theta_{\vec{\delta}\cup\vec{\tau_1}}$ and $\Theta_{\vec{\delta}\cup\vec{\tau_2}}$ we have, by definition, $W(\Theta_{\vec{\delta}\cup\vec{\tau_1}}) = W(\Theta_{\vec{\delta}\cup\vec{\tau_2}}) = 2^{-r} w(\bigvee_{\vec{\tau}} \Theta_{\vec{\delta}\cup\vec{\tau}})$ and notice that the number of possible $\vec{\tau}$ is 2^r. Using this (XXIX) will be

$$-w(\bigvee_{\vec{\tau}} \Theta_{\vec{\delta}\cup\vec{\tau}}) \log(2^{-r} w(\bigvee_{\vec{\tau}} \Theta_{\vec{\delta}\cup\vec{\tau}})) \ge -\sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w(\Theta_{\vec{\delta}\cup\vec{\tau}}). \tag{XXX}$$

The state descriptions are pairwise disjoint and so (XXX) will be

$$-\sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log(2^{-r} \sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}})) \ge -\sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w(\Theta_{\vec{\delta}\cup\vec{\tau}}). \tag{XXXI}$$

But $x \log x$ is a convex function and the number of possible $\vec{\tau}$'s in (XXXI) is 2^r. Hence by convexity we should have

$$\left(\sum_{\vec{\tau}} 2^{-r} w(\Theta_{\vec{\delta}\cup\vec{\tau}})\right) \left(\log(\sum_{\vec{\tau}} 2^{-r} w(\Theta_{\vec{\delta}\cup\vec{\tau}}))\right)$$
$$\le 2^{-r} \left(\sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w(\Theta_{\vec{\delta}\cup\vec{\tau}})\right)$$

that is (XXXI). Furthermore the inequality in (XXVIII) is strict because if we had equality for all such $\vec{\delta}$ then we would have $W \restriction_{L^r} = w \restriction_{L^r}$. To see that this leads to a contradiction, let ν be the map from $\{\, \langle k,j\rangle \,|\, 1 \le j \le r,\, \} \to \{0,1\}$ taking everything to 1. Then we will have

$$W(\bigwedge_{j=1}^{r} R(a_k, a_j)) = W(\bigvee_{\vec{\delta}} \Theta_{\vec{\delta}\cup\vec{\nu}}) = \sum_{\vec{\delta}} W(\Theta_{\vec{\delta}\cup\vec{\nu}})$$
$$= 2^{-r} \sum_{\vec{\delta}} w(\bigvee_{\vec{\tau}} \Theta_{\vec{\delta}\cup\vec{\tau}}) = 2^{-r} \sum_{\vec{\delta}} \sum_{\vec{\tau}} w(\Theta_{\vec{\delta}\cup\vec{\tau}})$$
$$= 2^{-r} \sum_{\vec{\epsilon}} w(\Theta_{\vec{\epsilon}}) = 2^{-r}$$

If $W^r = W \restriction_{L^r} = w \restriction_{L^r} = w^r$ then $e_k = w(\forall x R(a_k, x)) \leq w^r(\bigwedge_{j=1}^{r} R(a_k, a_j))$ $= W^r(\bigwedge_{j=1}^{r} R(a_k, a_j)) = 2^{-r}$, and this is a contradiction as r has been chosen large, so $2^{-r} < e_k$, This finishes the proof of Claim 2. ■

The quantifier complexity of K above is Σ_2. Thus our result here shows that for sentences with quantifier complexity of Σ_2 or above and the constraint sets induced by them, the Maximum Entropy models are not always uniquely defined. In the next section we will consider Σ_1 sentences and the constraints sets induced by them.

2.3. Constraints from Σ_1 Sentences

Let $K = \{w(\exists \vec{x}\theta(\vec{x})) = 1\}$ be the constraint induced by a Σ_1 sentence from a first order language L. In this section we will show that $ME_W(K)$ is unique. To show this we will show that there exists a probability function w defined on SL that satisfies K and is closer than any other probability function that satisfies K to $P_=$, on L^n for all n eventually. Notice again that this is stronger than what is required by Definition 9 as it establishes both the existence of a Maximum Entropy solution as well as its uniqueness. To this end, we will show that $P_=$ itself satisfies K and will thus be the $ME_W(K)$.

THEOREM 4. *Let $K = \{w(\exists \vec{x}\theta(\vec{x})) = 1\}$ where $\exists \vec{x}\theta(\vec{x})$ is a consistent Σ_1 sentence. Then $P_=$ is the Maximum Entropy solution for K, i.e., $ME_W(K)$ $= P_=$.*

PROOF. Let $\phi \in SL$ be of the form $\exists x_1, ..., x_t \psi(\vec{x})$ where ψ is quantifier free. We will show that if ϕ is satisfiable then $P_=(\phi) = 1$. Equivalently we will show that for a universal sentence ϕ' of the form $\forall x_1, ..., x_t \psi'(\vec{x})$ that is not a tautology we have $P_=(\phi') = 0$. Let $Q_i(x_1, ..., x_t)$, $i \in I$ enumerate formulae of the form

$$\bigwedge_{\substack{i_1,\ldots,i_j \leq t \\ R_i j-ary \\ R_i \in RL}} R_i^{\epsilon_{x_{i_1},\ldots,x_{i_j}}}(x_{i_1}, ..., x_{i_j}).$$

where, as before, $\epsilon_{x_{i_1},...,x_{i_j}} \in \{0,1\}$, $R_i^1 = R_i$ and $R_i^0 = \neg R_i$. Since $\forall x_1, ...,$ $x_t\psi'(\vec{x})$ is not a tautology then there is some proper subset J of I such that

$$\vDash \psi'(\vec{x}) \leftrightarrow \bigvee_{j\in J} Q_j(\vec{x}).$$

For $i_1 < i_2 < ... < i_t < q$ the number of extensions of $Q_i(a_{i_1}, ..., a_{i_t})$ is the same for each i so $P_=(Q_i(a_{i_1}, ..., a_{i_t})) = \frac{1}{|I|}$ and for disjoint $\vec{a}^1, ..., \vec{a}^r$, $P_=(Q_{n_1}(\vec{a}^1) \wedge ... \wedge Q_{n_r}(\vec{a}^r)) = \frac{1}{|I|^r}$. So

$$P_=(\forall x_1, ..., x_t\psi'(\vec{x})) \leq P_=(\psi'(\vec{a}^1) \wedge ... \wedge \psi'(\vec{a}^r))$$

$$= \sum_{n_1,...,n_r\in J} P_=(Q_{n_1}(\vec{a}^1) \wedge ... \wedge Q_{n_r}(\vec{a}^r)) = \left(\frac{|J|}{|I|}\right)^r \to 0 \; as \; r \to \infty.$$

So for every non tautology universal sentence ϕ', $P_=(\phi') = 0$ and so every satisfiable existential sentence will get value 1. This completes the proof. ■

Thus, although the ME_W fails to provide a unique extension to first order languages it is uniquely defined on such languages for constraint sets involving sentences of quantifier complexity Σ_1.

3. The ME_W, Permutation of Constants and Cloned State Descriptions

We will now turn to the investigation of some structural properties of the Maximum Entropy models. In particular we show first that these models are invariant under permutation of those constants that do not appear in the set of constraints and second that Maximum Entropy models will in the limit put all probability on those structures (state descriptions) that admit as many mutually distinguishable constants as possible.

Let σ be the permutation of $a_1, a_2, ...$ that transposes a_i and a_j, that is, $\sigma(a_i) = a_j$, $\sigma(a_j) = a_i$ and $\sigma(a_k) = a_k$ for $k \neq i, j$.

THEOREM 5. *Let $K = \{\sum_{j=1}^n a_{ji}v(\phi_j) = b_i \,|\, i = 1, ..., m\}$ be a set of linear constraints such that the constants a_i and a_j do not appear in K and let $w = ME_W(K)$. Then $w(\sigma(\psi)) = w(\psi)$ for $\psi \in SL$ where $\sigma(\psi)$ is the result of transposing a_i and a_j throughout ψ.*

PROOF. Assume $w(\sigma(\psi)) \neq w(\psi)$ for some ψ and define the probability function W as follows,

$$W^n(\Theta^{(n)}) = 2^{-1}(w^n(\Theta^{(n)}) + w^n(\sigma(\Theta^{(n)})))$$

First notice that for all ϕ_i appearing in K, $\sigma(\phi_i) = \phi_i$ and so $W(\phi_i) = w(\phi_i)$. Hence W satisfies K as w was a solution for K.

Claim 3. *$d_n(W, P_=) < d_n(w, P_=)$ for large n eventually.*

Claim 3 gives the required contradiction as we assumed w to be the closest probability function to $P_=$ that satisfies K. Thus we should have $w(\psi) = w(\sigma(\psi))$ and $ME_W(K)$ remains invariant under the permutations that permute those individuals not appearing explicitly in K. To prove Claim 3 it is enough to show that for large n

$$\sum_{\Theta_i^{(n)}} W^n(\Theta_i^{(n)}) \log(W^n(\Theta_i^{(n)})) < \sum_{\Theta_i^{(n)}} w^n(\Theta_i^{(n)}) \log(w^n(\Theta_i^{(n)})). \quad \text{(XXXII)}$$

By definition,

$$\begin{aligned} &\sum_{\Theta_i^{(n)}} W^n(\Theta_i^{(n)}) \log(W^n(\Theta_i^{(n)})) \\ &\quad = \sum_{\Theta_i^{(n)}} 2^{-1} \left(w^n(\Theta_i^{(n)}) + w^n(\sigma(\Theta_i^{(n)})) \right) \log(2^{-1}(w^n(\Theta_i^{(n)}) \\ &\qquad + w^n(\sigma(\Theta_i^{(n)})))) \end{aligned}$$

and for each $\Theta_i^{(n)}$ there is exactly one $\Theta_j^{(n)}$ such that $\Theta_j^{(n)} = \sigma(\Theta_i^{(n)})$ and $\Theta_i^{(n)} = \sigma(\Theta_j^{(n)})$. Notice that $x \log(x)$ is convex so

$$\begin{aligned} &2 \left(\frac{w^n(\Theta_i^{(n)}) + w^n(\Theta_j^{(n)})}{2} \log \left(\frac{w^n(\Theta_i^{(n)}) + w^n(\Theta_j^{(n)})}{2} \right) \right) \\ &\quad \leq w^n(\Theta_i^{(n)}) \log(w^n(\Theta_i^{(n)})) \\ &\qquad + w^n(\Theta_j^{(n)}) \log(w^n(\Theta_j^{(n)})) \end{aligned}$$

and thus

$$\sum_{\Theta_i^{(n)}} W^n(\Theta_i^{(n)}) \log(W^n(\Theta_i^{(n)})) \leq \sum_{\Theta_i^{(n)}} w^n(\Theta_i^{(n)}) \log(w^n(\Theta_i^{(n)})).$$

This inequality should be strict eventually otherwise $W = w$ which is a contradiction as $w(\psi) \neq w(\sigma(\psi))$ while $W(\psi) = W(\sigma(\psi))$ and this completes the proof of Claim 3. ∎

3.1. ME_W and the Cloned State Descriptions

The state description $\Phi^{(n)}$ is a clone of $\Psi^{(m)}$ with $n > m$ if the behaviour of each constants in $\Phi^{(n)}$ is the same as the behaviour of some constant in

$\Psi^{(m)}$. For a permutation σ of the constants that permutes a_i and a_j and keeps other constants fixed, and a state description $\Psi^{(m)}$, let $\sigma(\Psi^{(m)})$ be the result of swapping a_i and a_j in $\Psi^{(m)}$. We say that constants a_i and a_j are indistinguishable for $\Psi^{(m)}$, $a_i \sim_{\Psi^{(m)}} a_j$, if $\sigma(\Psi^{(m)}) = \Psi^{(m)}$. Notice that each state description $\Psi^{(m)}$ gives a partition of the $\{a_1, \ldots, a_m\}$ into equivalence classes of the indistinguishability relation $\sim_{\Psi^{(m)}}$. If $\Phi^{(n)}$ is a clone of $\Psi^{(m)}$, then $\sim_{\Phi^{(n)}}$ has the same set of equivalence classes and each constant $a_{m+1}, \ldots, a_n$ is added to one of these existing equivalence classes and will thus have the same behaviour as some constant $\{a_1, \ldots, a_m\}$ in $\Psi^{(m)}$. We will show that the Maximum Entropy models do not favour the cloned state descriptions. Indeed we will show that in the limit, maximum entropy models will assign all the probability to those state descriptions that are not a clone of some other state description on a smaller number of constants. In this way, the maximum entropy models will favour those state descriptions that admit as many distinguishable constants as possible.

Definition 10. For $m \leq p$, we say that the state description $\Phi(a_1, ..., a_p)$ is a *clone* of the state description $\Psi(a_1, ..., a_m)$ if there is a function τ from p to m such that

$$\Phi(a_{\tau(1)}, ..., a_{\tau(p)}) \equiv \Psi(a_1, ..., a_m).$$

Theorem 5. *When it is consistent with K, $ME_W(K)$ will, in the limit, put all the probability on the structures in which there are as many explicitly distinct individuals as possible. In other words, if there is a state description on $a_1, a_2, ..., a_{m+1}$ that is consistent with K which is not the clone of any state description on $a_1, ..., a_m$ then if $\bigvee \beta_p$ is the disjunction of those state descriptions $\Theta^{(p)}(a_1, ..., a_p)$ which are clones of some state description on $a_1, ..., a_m$,*

$$\lim_{p \to \infty} ME_W(K)(\bigvee \beta_p) = 0.$$

In other words, if it is consistent with K to have $m + 1$ distinguishable constants, then $ME_W(K)$ will, for large enough r, assign zero probability to those state descriptions on $a_1, \ldots, a_r$ which have at most m distinguishable constants.

Proof. Suppose not. Set $w = ME_W(K)$ and let $a > 0$ be the largest such that $w(\bigvee \beta_p) \geq a$, for all p eventually. We shall show that for $n > m$ any state description $\Delta^{(n)}(a_1, ..., a_n)$ that is consistent with K must be a clone of *some* state description on $a_1, ..., a_m$. This will give a contradiction with

the assumption that there is an state description on $a_1, a_2, ..., a_{m+1}$ that is consistent with K and is not the clone of any state description on $a_1, ..., a_m$. Suppose on the contrary that a state description $\Delta^{(n)}(a_1, ..., a_n)$ (where $n > m$) did exist and was consistent with K but was not a clone of any state description on $a_1, ..., a_m$. We may assume that $a_1, .., a_n$ are all distinguishable in $\Delta^{(n)}$, in other words replacing any a_i in $\Delta^{(n)}(a_1, ..., a_n)$ by a_j, $1 \leq j \leq n$, $i \neq j$ gives a contradiction. Define for the state description $\Phi^{(p)}$ with $p \geq m$,

$$w^c(\Phi(a_1, ..., a_p)) = \lim_{r\to\infty} w\left(\bigvee \beta_r\right)$$

where $\bigvee \beta_r$ is the disjunction of those state descriptions on $\{a_1, \ldots, a_r\}$ which extend $\Phi^{(p)}$ and are clones of some state description on $a_1, ..., a_m$. Notice that this limit exists and for $p > m$, $\sum_{\beta_p} w^c(\beta_p) = a > 0$.

We define the probability function W as follows. For a state description $\Lambda^{(r)}$ where $r \geq n$:

$$W(\Lambda^{(r)}) := \begin{cases} w(\Lambda^{(r)}) + Q_r^{-1}a & \text{If } \Lambda^{(r)} \text{ extends } \Delta^{(n)} \text{ and is a clone of } \Delta^(n) \\ w(\Lambda^{(r)}) - w^c(\Lambda^{(r)}) & \text{If } \Lambda^(r) \text{ is a clone of some } \Psi^{(m)}(a_1, ..., a_m) \\ w(\Lambda^{(r)}) & \textit{Otherwise} \end{cases}$$

where Q_r is the number of clones of $\Delta^{(n)}(a_1, ..., a_n)$ on $a_1, ..., a_r$.

Claim 4. *W extends to a probability function on SL and is closer to $P_=$ than w, that is $d_n(W, P_=) < d_n(w, P_=)$ for all n eventually.*

This provides the required contradiction by the choice of w. So, if there is an m and $a > 0$ such that $\lim_{r\to\infty} w(\bigvee \beta_r) = a$, where $\bigvee \beta_r$ is the disjunction of state descriptions on $a_1, ..., a_r$ that are clones of some state description on $a_1, ..., a_m$, then eventually every state description consistent with K should be clone of some state description $a_1, ..., a_m$. As pointed out earlier, this contradicts the assumption of the existence of a state description on $a_1, ..., a_{m+1}$ consistent with K that is not a clone of any state description on $a_1, ..., a_m$. ■

Before we proceed to prove Claim 4 it might be helpful to mention that the idea here is that because $\Delta(a_1, ..., a_n)$ is not a clone of any state description $\Psi(a_1, ..., a_m)$, for large $r > p$, $\Delta(a_1, ..., a_n)$ has far more clones extending it than there are clones of state descriptions on $a_1, ..., a_m$. Then, in the long run, it will be more advantageous in terms of entropy to spread the probability measure uniformly onto these clones of $\Delta(a_1, ..., a_n)$ than (possibly non-uniformly) on the clones of state descriptions on $a_1, ..., a_m$. This

is exactly what is happening in the definition of W above. We take some probability off from the clones of the state descriptions on $\{a_1, \ldots, a_m\}$ and divide it equally among the state descriptions that are clones of $\Delta^{(n)}$.

Proof of Claim. **4.** Let Γ_r denote the set of state descriptions of L^r and remember that $w = ME_w(K)$ and W as defined above. We will first show that W extends to a probability function on SL. To show this, by (II), it is enough to show that $\sum_{\Lambda_i^{(r)}} W(\Lambda_i^{(r)}) = 1$ where $\Lambda_i^{(r)}$ ranges over Γ_r and that $W(\Lambda^{(r)}) = \sum_{\Lambda_j^{(r+1)} \models \Lambda^{(r)}} W(\Lambda_j^{(r+1)})$ where $\Lambda_j^{(r+1)}$ ranges over state descriptions of L^{r+1} , i.e. Γ_{r+1}. To see that $\sum_{\Lambda_i^{(r)}} W(\Lambda_i^{(r)}) = 1$, let Γ_r^1 be those state descriptions in Γ_r that extend $\Delta^{(n)}$ and are clones of $\Delta^{(n)}$ and Γ_r^2 be those that are clones of some state descriptions on $a_1, ..., a_m$. Set $\Gamma_r^3 = \Gamma_r - (\Gamma_r^1 \cup \Gamma_r^2)$. Thus

$$\begin{aligned}
\sum_{\Lambda_i^{(r)}} W(\Lambda_i^{(r)}) &= \sum_{\Lambda_i^{(r)} \in \Gamma_r^1} W(\Lambda_i^{(r)}) + \sum_{\Lambda_i^{(r)} \in \Gamma_r^2} W(\Lambda_i^{(r)}) + \sum_{\Lambda_i^{(r)} \in \Gamma_r^3} W(\Lambda_i^{(r)}) \\
&= \sum_{\Lambda_i^{(r)} \in \Gamma_r^1} \left(w(\Lambda_i^{(r)}) + Q_r^{-1} a \right) + \sum_{\Lambda_i^{(r)} \in \Gamma_r^2} \left(w(\Lambda_i^{(r)}) - w^c(\Lambda_i^{(r)}) \right) \\
&\quad + \sum_{\Lambda_i^{(r)} \in \Gamma_r^3} w(\Lambda_i^{(r)}) \\
&= \sum_{\Lambda^{(r)} \in \Gamma_r} w(\Lambda^{(r)}) - \sum_{\Lambda^{(r)} \in \Gamma_r^2} w^c(\Lambda^{(r)}) + a = 1 + a \\
&\quad - \sum_{\Lambda^{(r)} \in \Gamma_r^2} w^c(\Lambda^{(r)}) = 1.
\end{aligned}$$

To see that W extends correctly to a probability function on SL, as in (II), we will consider each case separately. For the first case, that is when $\Lambda^{(r)}$ extends $\Delta^{(n)}$ and is a clone of $\Delta^{(n)}$,

$$W(\Lambda^{(r)}) = w(\Lambda^{(r)}) + Q_r^{-1} a = \left(\sum_{\substack{\Lambda_j^{(r+1)} \in \Gamma_{r+1} \\ \Lambda_j^{(r+1)} \models \Lambda^{(r)}}} w(\Lambda_j^{(r+1)}) \right) + Q_r^{-1} a. \quad \text{(XXXIII)}$$

Let $\Gamma_{r+1} = \Gamma_{r+1}^{\Delta} \cup \overline{\Gamma_{r+1}^{\Delta}}$ where Γ_{r+1}^{Δ} is the set of those state descriptions in Γ_{r+1} that are clones of $\Delta^{(n)}$. Notice that state descriptions in $\overline{\Gamma_{r+1}^{\Delta}}$ that extend $\Lambda^{(r)}$ are *not* clones of any state description on $a_1, ..., a_m$ since $\Delta^{(n)}$

is not a clone of any state description on $a_1, \ldots, a_m$ and $\Lambda^{(r)}$ extends $\Delta^{(n)}$. Thus for $\Lambda(a_1, ..., a_{r+1}) \in \overline{\Gamma^{\Delta}_{r+1}}$, $W(\Lambda(a_1, ..., a_{r+1})) = w(\Lambda(a_1, ..., a_{r+1}))$ also $Q_{r+1} = |\Gamma^{\Delta}_{r+1}| = n|\Gamma^{\Delta}_{r}| = nQ_r$ as every state description in Γ^{Δ}_{k} has exactly n extensions to state descriptions Γ^{Δ}_{k+1}[1]. So for (XXXIII) we have,

$$
\begin{aligned}
W(\Lambda(a_1, ..., a_r)) &= Q_r^{-1}a + \sum_{\substack{\Lambda_j \in \Gamma_{r+1} \\ \Lambda_j \vDash \Lambda}} w(\Lambda_j(a_1, ..., a_{r+1})) \\
&= \sum_{\substack{\Lambda_j \in \Gamma^{\Delta}_{r+1} \\ \Lambda_j \vDash \Lambda}} \left(w(\Lambda_j(a_1, ..., a_{r+1})) + Q_{r+1}^{-1}a\right) \\
&\quad + \sum_{\substack{\Lambda_j \in \overline{\Gamma^{\Delta}_{r+1}} \\ \Lambda_j \vDash \Lambda}} W(\Lambda_j(a_1, ..., a_{r+1})) \\
&= \sum_{\substack{\Lambda_j \in \Gamma^{\Delta}_{r+1} \\ \Lambda_j \vDash \Lambda}} W(\Lambda_j(a_1, ..., a_{r+1})) + \sum_{\substack{\Lambda_j \in \overline{\Gamma^{\Delta}_{r+1}} \\ \Lambda_j \vDash \Lambda}} W(\Lambda_j(a_1, ..., a_{r+1})) \\
&= \sum_{\substack{\Lambda_j \in \Gamma_{r+1} \\ \Lambda_j \vDash \Lambda}} W(\Lambda_j(a_a, ..., a_{r+1}))
\end{aligned}
$$

We shall now show that W is closer to $P_=$ than w. To this end, it is enough to show that for r large enough $\sum_{\Lambda^{(r)} \in \Gamma_r} W(\Lambda^{(r)}) \log(W(\Lambda^{(r)})) < \sum_{\Lambda^{(r)} \in \Gamma_r} w(\Lambda^{(r)}) \log(w(\Lambda^{(r)}))$ or

$$
\begin{aligned}
&\sum_{\Lambda^{(r)} \in \Gamma^1_r} W(\Lambda^{(r)}) \log(W(\Lambda^{(r)})) + \sum_{\Lambda^{(r)} \in \Gamma^2_r} W(\Lambda^{(r)}) \log(W(\Lambda^{(r)})) \\
&\quad + \sum_{\Lambda \in \Gamma^3_r} W(\Lambda^{(r)}) \log(W(\Lambda^{(r)})) < \sum_{\Lambda^{(r)} \in \Gamma_r} w(\Lambda^{(r)}) \log(w(\Lambda^{(r)})).
\end{aligned}
$$

Expanding the left hand side we have

$$
\begin{aligned}
&\sum_{\Lambda^{(r)} \in \Gamma^1_r} \left(w(\Lambda^{(r)}) + Q_r^{-1}a\right) \log \left(w(\Lambda^{(r)}) + Q_r^{-1}a\right) \\
&\quad + \sum_{\Lambda^{(r)} \in \Gamma^2_r} \left(w(\Lambda^{(r)}) - w^c(\Lambda^{(r)})\right) \log \left(w(\Lambda^{(r)}) - w^c(\Lambda^{(r)})\right) \\
&\quad + \sum_{\Lambda^{(r)} \in \Gamma^3_r} w(\Lambda^{(r)}) \log \left(w(\Lambda^{(r)})\right) < \sum_{\Lambda^{(r)} \in \Gamma_r} w(\Lambda^{(r)}) \log \left(w(\Lambda^{(r)})\right).
\end{aligned}
$$

[1]Notice that here we are using the fact that $a_1, ..., a_n$ are all distinguishable in $\Delta(a_1, ..., a_n)$.

Notice that $0 < w(\Lambda^{(r)}) - w^c(\Lambda^{(r)}) \leq 1$ and so $\log(w(\Lambda^{(r)}) - w^c(\Lambda^{(r)})) \leq 0$ and $\sum_{\Lambda^{(r)} \in \Gamma^2_r} (w(\Lambda^{(r)}) - w^c(\Lambda^{(r)})) \log(w(\Lambda^{(r)}) - w^c(\Lambda^{(r)})) \leq 0$ so expanding the left hand side and rearranging the equation, it is enough to show that

$$\sum_{\Lambda^{(r)} \in \Gamma^1_r} w(\Lambda^{(r)}) \log\left(1 + \frac{a}{Q_r w(\Lambda^{(r)})}\right) + Q_r^{-1} a \sum_{\Lambda^{(r)} \in \Gamma^1_r} \log\left(w(\Lambda^{(r)}) + Q_r^{-1} a\right)$$
$$< \sum_{\Lambda^{(r)} \in \Gamma^2_r} w(\Lambda^{(r)}) \log(w(\Lambda^{(r)})) \qquad \text{(XXXIV)}$$

The first thing to notice here is that if $\Lambda_1^{(r)}, \Lambda_2^{(r)} \in \Gamma^1_r$ then we can assume that w gives them the same probability, otherwise we can define a bijection σ_s from $\Delta^{(s)} \in \Gamma^1_s$ extending $\Lambda_1^{(r)}$ to $\Delta^{(s)} \in \Gamma^1_s$ that extend $\Lambda_2^{(r)}$ for $s \geq r$ such that if $\Delta'^{(s+1)} \in \Gamma^1_{s+1}$ extends $\Delta^{(s)} \in \Gamma^1_s$ then $\sigma_{s+1}(\Delta'^{(s+1)}) \in \Gamma^1_{s+1}$ extends $\sigma_s(\Delta^{(s)}) \in \Gamma^1_s$. Now defining for $s \geq r$

$$w'(\Delta^{(s)}) = 2^{-1}(w(\Delta^{(s)}) + w(\sigma_s(\Delta^{(s)})))$$

for $\Delta^{(s)} \in \Gamma^1_s$ extending $\Lambda_1^{(r)}$ and

$$w'(\Delta^{(s)}) = 2^{-1}(w(\Delta^{(s)}) + w(\sigma_s^{-1}(\Delta^{(s)})))$$

for $\Delta^{(s)} \in \Gamma^1_s$ extending $\Lambda_2^{(r)}$ and $w'(\Delta^{(s)}) = w(\Delta^{(s)})$ on other state descriptions gives a probability function satisfying K that is closer to $P_=$ than w. Thus for $\Lambda_1^{(r)}, \Lambda_2^{(r)} \in \Gamma^1_r$, $w(\Lambda_1^{(r)}) = w(\Lambda_2^{(r)})$.

Let $w(\Lambda^{(r)}) = \frac{b}{Q_r}$ for $\Lambda^{(r)} \in \Gamma^1_r$. Then (XXXIV) will become

$$(a+b)\log\left(\frac{a+b}{Q_r}\right) + b\log\left(\frac{b}{Q_r}\right) < \sum_{\Lambda^{(r)} \in \Gamma^2_r} w(\Lambda^{(r)}) \log(w(\Lambda^{(r)})).$$

For the right hand side, let $\sum_{\Lambda^{(r)} \in \Gamma^2_r} w(\Lambda^{(r)}) = d$ and notice that since any $\Lambda^{(r)} \in \Gamma^2_r$ is a clone of some state description on L^m we should have $|\Gamma^2_r| \leq D\,(m)^{r-m} \leq D\,(n-1)^r$ where D is the number of state descriptions of L^m consistent with K. On the other hand, by convexity,

$$\sum_{\Lambda^{(r)} \in \Gamma^2_r} w(\Lambda^{(r)}) \log(w(\Lambda^{(r)})) \geq |\Gamma^2_r| \frac{d}{|\Gamma^2_r|} \log\left(\frac{d}{|\Gamma^2_r|}\right) \geq d \log\left(\frac{d}{D\,(n-1)^r}\right)$$

whilst the left hand side of (XXXIV) is at most $c - a\log(Q_r) = c' - a\log(n^r)$ for some constants c and c' and to show (XXXIV) it will be enough to show that $c' - a\log(n^r) < d\log(\frac{d}{D\,(n-1)^r})$ that is $(1/r)(c' + d(\log(D) - \log(d))) < a\log(n) - d\log(n-1)$ which holds for r large enough. This completes the proof of Claim 4. ∎

So where K allows at least $m+1$ distinct constants, for any m the limit as $p \to \infty$ of the probability of the state descriptions on $a_1, ..., a_p$ that are a clone of some state descriptions on $a_1, ..., a_m$ will tend to zero. In other words ME_W will in the limit put all the probability on the structures in which there are as many explicitly distinct individuals as possible.
This result, concerning the treatment of cloned state descriptions, is not meant as a shortcoming nor an advantage of the Maximum Entropy models. The relevance (or lack thereof) of the cloned state descriptions is highly contextual. It does, however, provide a structural analysis that, in our opinion, links the behaviour of the Maximum Entropy models (that we have, so far, studied in terms of the treatment of the state descriptions) to the treatment of the constants, which can hopefully provide better intuition regarding the behaviour of these models. One way of reading this result is thus that, as one would expect, the most entropic models admit as many different types of constants as possible.

4. Conclusions

We studied the problem of determining the least committal model of a probabilistic theory K. The problem has attracted a lot of attention from different disciplines and is relevant to many scientific areas. There are two approaches for defining such models on propositional languages; either as the probability function with maximum Shannon entropy, $ME(K)$, or as the one that minimises the informational distance to the most non-committal probability function over all $(P_=)$, ME_W. It is known that these approaches agree for theories from propositional languages.

We focused on the second characterisation, ME_W, and studied the generalisation of $ME_W(K)$ to a set of constraints K over a first order language in terms of the quantifier complexity of K. We showed that ME_W is unique for purely unary languages as well as for Σ_1 constraints and in these cases it agrees with $\lim_{r\to\infty} ME(K^{(r)})$. The case of Π_1 constraint sets remains open. It is also not known whether $\lim_{r\to\infty} ME(K^{(r)})$ for these constraint sets exists in general or not. In [18] Paris and Rafiee Rad showed the existence of this limit for theories consisting of only *slow* formulae, that are those whose models of any size are bounded exponentially.

However, for a set of constraints K with quantifier complexity of Σ_2 or higher, $ME_W(K)$ is not always unique. In particular we showed that for such constraint sets one can always increase the entropy of the model by making the witness of the existential quantifier scarcer. Although we established

this by means of an example, our analysis gives a general account of why the maximum entropy models for these theories do not exist. Finally, we proved that the ME_W solution exclusively favours those models with as many explicitly distinct individuals as possible.

Acknowledgements. This work is greatly indebted to Jeff Paris for his invaluable guidance and advice in the development of these results. I would also like to thank the referee for useful comments that helped greatly in improving this paper. The Author's research is funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 283963. The research leading to these results has been partially supported by MATHLOGAP studentship.

References

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

[2] Barnett, O. W., and J. B. Paris, Maximum Entropy inference with qualified knowledge, *Logic Journal of the IGPL* 16(1):85–98, 2008.

[3] Berger, A., S. Della Pietra, and V. Della Pietra, A maximum entropy approach to natural language processing, *Computational Linguistics* 22(1):39–71, 1996.

[4] Chen, C. H., Maximum entropy analysis for pattern recognition, in P. F. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publisher, London, 1990.

[5] Gaifman, H., Concerning measures in first order calculi, *Israel Journal of Mathematics* 24:1–18, 1964.

[6] Grotenhuis, M. G., *An Overview of the Maximum Entropy Method of Image Deconvolution*, A University of Minnesota Twin Cities Plan B Masters paper.

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

[8] Jaynes, E. T., Information theory and statistical mechanics, *Physical Reviews* 106:620–630, 108:171–190, 1957.

[9] Jaynes, E. T., Notes on present status and future prospects, in W. T. Grandy and L. H. Schick (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, London, 1990, pp. 1–13.
[10] Jaynes, E. T., How Should We Use Entropy in Economics? 1991, manuscript available at: http://www.leibniz.imag.fr/LAPLACE/Jaynes/prob.html.
[11] Kapur, J. N., Twenty five years of maximum entropy, *Journal of Mathematical and Physical Sciences* 17(2):103–156, 1983.
[12] Kapur, J. N., Non-additive measures of entropy and distributions of statistical mechanics, *Indian Journal of Pure and Applied Mathematics* 14(11):1372–1384, 1983.
[13] Landes, J., and J. Williamson, Objective Bayesianism and the maximum entropy principle, *Entropy* 15(9):3528–3591, 2013.
[14] Landes, J., and J. Williamson, Justifying objective bayesianism on predicate languages, *Entropy* 17:2459–2543, 2015.
[15] Paris, J. B., *The Uncertain Reasoner's Companion*, Cambridge University Press, Cambridge, 1994.
[16] Paris, J. B., and A. Vencovská, A note on the inevitability of maximum entropy, *International Journal of Approximate Reasoning* 4(3):183–224, 1990.
[17] Paris, J. B., and A. Vencovská, In defense of the maximum entropy inference process, *International Journal of Approximate Reasoning* 17(1):77–103, 1997.
[18] Paris, J. B., and S. Rafiee Rad, A note on the least informative model of a theory, in F. Ferreira, B. Löwe, E. Mayordomo, and L. Mendes Gomes (eds.), *Programs Proofs Processes, CiE 2010*, Springer LNCS 6158, 2010, pp. 342–351.
[19] Rafiee Rad, S., *Inference Processes For Probabilistic First Order Languages*. PhD Thesis, University of Manchester, 2009. http://www.maths.manchester.ac.uk/~jeff/
[20] Rosenkrantz, R. D., *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Reidel, Dordrecht, 1977.
[21] Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Champaign, 1949.
[22] Williamson, J., *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford, 2005.
[23] Williamson, J., Objective Bayesian probabilistic logic, *Journal of Algorithms in Cognition, Informatics and Logic* 63:167–183, 2008.
[24] Williamson, J., *In Defence of Objective Bayesianism*, Oxford University Press, Oxford, 2010, pp. 167–183.
[25] Zellner, A., Bayesian methods and entropy in economics and econometrics, in W. T. Grandy and L. H. Schick (eds.), *Maximum Entropy and Bayesian Methods*, 1991.

S. Rafiee Rad
Institute for Logic, Language and Computation, UvA
Amsterdam, Netherlands
soroush.r.rad@gmail.com