Maximum Entropy Models for Σ_1 Sentences

Soroush Rafiee Rad*

Abstract

In this paper we investigate the most uninformative models of Σ_1 sentences. We will show that the two main approaches for defining the Maximum Entropy models on first order languages are well defined for Σ_1 sentences and that they agree on sets of sentences consisting of only Σ_1 sentences.

Keywords: Maximum Entropy, probabilistic models, existential sentence, Objective Bayesian Epistemology

1 Introduction

The Maximum Entropy model for a sentence ϕ represents the most uninformative model of ϕ. To be more precise, given a consistent sentence ϕ and a formula $\psi(x_1, \ldots, x_n)$ from a first order language L, let M be an structure for L with domain $\{a_1, a_2, \ldots\}$ which we only know to be a model of ϕ. A natural question about this M is to ask how likely it is for M to be also a model of ψ, in other words, what probability should one assign to M being also a model of ψ. When ϕ identifies a unique model N (i.e. $M = N$), this question may be answered by checking the validity of $\psi(a_{i_1}, \ldots, a_{i_n})$ in N. If ϕ admits more than one model, however, knowing that M is a model of ϕ under-determines M and the validity of ψ in M may be uncertain. In this sense ϕ induces an assignment of probabilities to the sentences of the language, where the probability assigned to ψ is intended as the probability that a random model of ϕ is also a model of ψ. This will in turn induce a probability distribution on the set of structures for L with domain $\{a_1, a_2, \ldots, \}$.

We are interested in the least informative of such assignments with respect to M which we shall call the Maximum Entropy model of ϕ, i.e., the Maximum Entropy model of ϕ is identified with the assignment of probabilities that leaves M as unconstrained as possible beyond being a model of ϕ. In this sense it gives a probabilistic description that specifies M to the extent that it is characterised by ϕ while remaining as free as

*Institute for Logic, Language and Computation, UvA. soroush.r.rad@gmail.com

possible beyond that. Note, however, that a Maximum Entropy model is not a model in the sense of a structure for the language, but rather a probability function on the set of sentences of the language that characterises an uncertain (i.e. under-determined) structure. It is important to emphasise at this point that in what follows, we shall say "model" to refer to these probability functions. We shall instead say "term models" to refer to the structures. More generally, given a set of linear constraints K the Maximum Entropy model of K is the probability function over the sentences of L which satisfies the constraints given in K while remaining maximally uninformative beyond that. When considering a set of linear constraints K, we use "models of K" and "solutions for K", interchangeably.

These probability functions have been extensively investigated and applied in various disciplines from statistics [5] and physics, [7] to computer science, pattern recognition [3], computational linguistics [2] as well as economics and finance [6]. Another prototypical example where Maximum Entropy models are of great relevance is formal epistemology and the study of rational belief formation [8, 16, 17]. In this setting the problem of interest is how should an agent in possession of some evidence form rational belief? To be slightly more precise, the question is; given sentences $\phi_1, \ldots, \phi_n$ as the agent's evidence, what would be the credence x she has to assign to some arbitrary sentence ψ such that x represents a rational belief of the agent in the context of her evidence. Equivalently, one can ask which probability function over the sentences of the language best represents the degrees of belief of the agent.

The most popular proposal for formalising the concept of *least informative* is to take Shannon's entropy as the measure for the informational content of a probability function. Given a set of constraints, one approach, see for example [11], is to choose the probability function satisfying the constraints with maximum Shannon entropy as the least informative one. A second approach, followed for example by Williamson [16, 17], uses the relative Shannon entropy instead. To make the idea clear, consider the problem we started with and a case in which there is no information (and thus no restrictions) concerning the structure M. In this case the satisfaction of a sentence ψ in M is maximally uncertain and thus the assignment of probabilities should be maximally equivocal. We shall call this probability function (which we shall shortly define precisely) $P_=$. The second approach for defining the "least informative", requires the assignment of probabilities to satisfy the given constraints and remain informationally as close as possible to $P_=$, where the informational difference between two probability functions is measured by their relative entropy. It is not hard to check that on propositional languages both approaches are well defined and result in the same unique answer [13].

The literature on justification of Maximum Entropy or its underlying principles is extensive and it remains the strongest candidate for the formalisation of the least informative probability function [11, 15, 17]. The major part of this literature is concerned with propositional languages, however, there have been attempts to generalise both these approaches to the first order case. To generalise the first approach, Barnett and

Paris, [1], propose to define the Maximum Entropy models on a first order language as the limit of the Maximum Entropy models on finite sub-languages. They showed that for constraint sets from languages with only unary predicates, this limit exists and the resulting probability function does satisfy the constraints. To generalise the second approach one has to move to a more sophisticated notion of informational distance.
This paper further investigates the Maximum Entropy models and the extent to which they can be defined for first order languages; in particular we shall investigate the Maximum Entropy models for existential sentences. The paper unfolds as follows: Section 2 reviews preliminaries and notation, as well as the definition of the Maximum Entropy probability functions over propositional and first order languages; and Section 3 proves the main theorems. We will then conclude with a discussion in Section 4.

2 Preliminaries and Notation

Throughout this paper, we will work with a first order language L with finitely many relation symbols, no function symbols, no equality and countably many constant symbols $a_1, a_2, a_3, ...$ Furthermore we assume that these constants exhaust the universe. Let RL, SL and TL denote the sets of relation symbols, sentences and the term models for L respectively, where a *term model* is a structure M for the language L with domain $M = \{ a_i \mid i = 1, 2, ...\}$ where every constant symbol is interpreted as itself. For more details on the preliminary definitions and results please see [9, 12].

Definition 1 *$w : SL \to [0\,, 1]$ is a probability function if for every $\theta, \phi, \exists x\psi(x) \in SL$,*

P1. If $\models \theta$ then $w(\theta) = 1$.

P2. If $\models \neg(\theta \wedge \phi)$ then $w(\theta \vee \phi) = w(\theta) + w(\phi)$.

P3. $w(\exists x\psi(x)) = \lim_{n\to\infty} w(\bigvee_{i=1}^{n} \psi(a_i))$.

Definition 2 *Let $\mathcal{L}$ be a finite propositional language with propositional variables $p_1, ..., p_n$. Atoms of $\mathcal{L}$ are the sentences $\{ \alpha_i \mid i = 1, ...J\}$, of the form $\bigwedge_{i=1}^{n} p_i^{\epsilon_i}$ where $\epsilon_i \in \{0, 1\}$, $p^1 = p$ and $p^0 = \neg p$.*

Take a propositional language $\mathcal{L}$. For every sentence $\phi \in S\mathcal{L}$, there is unique set $\Gamma_\phi \subseteq \{ \alpha_i \mid i = 1, ..., J \}$ such that $\models \phi \leftrightarrow \bigvee_{\alpha_i \in \Gamma_\phi} \alpha_i$. It can be easily checked that $\Gamma_\phi = \{ \alpha_j \mid \alpha_j \vdash \phi \}$. Thus if w is a probability function $w(\phi) = w(\bigvee_{\alpha_i \models \phi} \alpha_i) = \sum_{\alpha_i \models \phi} w(\alpha_i)$ as the α_i's are mutually inconsistent. 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 α_i's, i.e., by the vector $< w(\alpha_1), ..., w(\alpha_J) > \in \mathbb{D}^{\mathcal{L}} = \{ \vec{x} \in \mathbb{R}^J \mid \vec{x} \geq 0, \sum_{i=1}^{J} x_i = 1 \}$. Conversely if $\vec{a} \in \mathbb{D}^{\mathcal{L}}$ we can define a probability function $w' : S\mathcal{L} \to [0\,, 1]$ such that $< w'(\alpha_1), ..., w'(\alpha_J) >= \vec{a}$ by setting $w'(\phi) = \sum_{\alpha_i \models \phi} a_i$.
Now consider a first order language L. Although the atoms of L are not expressible in

the language (as they will require infinite conjunctions), the *state descriptions* for the finite sub-languages will play a similar role to that of atoms in the propositional case.

Definition 3 *Let L be a first order language with the finite set of relation symbols RL and let L^k be the sub-language of L with only constant symbols $a_1, ..., a_k$. The state descriptions of L^k are the sentences $\Theta^k_1, ..., \Theta^k_{n_k}$ of the form*

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

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

Throughout this paper we will denote the set of state descriptions for L and L^r by Γ and Γ^r respectively. Furthermore, we will write Γ_ϕ (res. Γ^r_ϕ) for the set of state descriptions of L (res. L^r) that are consistent with the sentence ϕ.

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 $\mathcal{L}^k$ with propositional variables $R_i(a_{i_1}, ..., a_{i_j})$ for $i_1, ..., i_j \leq k$, $R_i \in RL$. The sentences Θ^k_i will be the atoms of $\mathcal{L}^k$ and as before $\models \theta \leftrightarrow \bigvee_{\Theta^k_i \models \theta} \Theta^k_i$ and for every probability function w, $w(\theta) = w(\bigvee_{\Theta^k_i \models \theta} \Theta^k_i) = \sum_{\Theta^k_i \models \theta} w(\Theta^k_i)$. Thus to determine $w(\theta)$ we only need to determine the values $w(\Theta^k_i)$ and to require

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

$$w(\Theta^k_i) = \sum_{\Theta^{k+1}_j \models \Theta^k_i} w(\Theta^{k+1}_j) \tag{2}$$

to ensure that w satisfies P1 and P2. The following theorem due to Gaifman [4], 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. We note that the set of state descriptions of L^k is the same as the set of term models for L^k with domain $\{a_1, \ldots, a_k\}$.

Definition 4 *Define the equivocator, $P_=$, as the probability function that for each k, assigns equal probabilities to the Θ_i^k's (the state descriptions of L^k), i.e., the most non-committal probability function.*

Notice that this determines $P_=$ on all of SL by Theorem 1 and the preceding argument.

Definition 5 *A sentence ϕ from a first order language L is called a Σ_1 sentence iff ϕ is logically equivalent to a sentence of the form $\exists \vec{x} \theta(\vec{x})$ where $\theta(\vec{x})$ is quantifier free.*

Definition 6 *A constraint set K is a finite satisfiable set of linear constraints of the from $\{\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i \mid i = 1, \ldots, m\}$, where $\theta_j \in SL, a_{ij}, b_i \in \mathbb{R}$ and w is a probability function. Every finite satisfiable set of sentences $K = \{\phi_1, \ldots, \phi_n\}$ is identified with the constraint set $\{w(\phi_1) = 1, \ldots, w(\phi_n) = 1\}$ induced by it and in particular we shall identify every sentence ϕ with the constraint $w(\phi) = 1$.*

We shall next give the definition of Maximum Entropy solutions for a set of linear constraints K as above. Our results in Section 3, however, are concerned only with the constraints that are induced by a sentence. In particular, by the Maximum Entropy model of the sentences ϕ we mean the Maximum Entropy probability function that satisfies the corresponding constraint $w(\phi) = 1$.

Definition 7 *The Shannon entropy of the probability function, W, defined on a set $X = \{x_1, \ldots, x_n\}$ (so $0 \leq W(x_i) \leq 1$ and $\sum_i W(x_i) = 1$), is given by*

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

The Shannon entropy is the most commonly used measures for the informational content of a probability function, [14].

Definition 8 *An inference process, N, on L, is a function that on each set of linear constraints K, returns a probability function on SL, N(K), that satisfies K.*

We will write ME for the inference process that on each set of constraints K, returns the maximum entropy probability function that satisfies K, denoted as $ME(K)$. There are two approaches for defining Maximum Entropy probability functions that satisfy a set of constraints. We shall start from a propositional case first and then move to the first order languages. Let $\mathcal{L}$ be a propositional language with atoms $\alpha_1, \ldots, \alpha_J$ and K a set of linear constraints. The first approach is to define $ME(K)$ as the unique probability function over the sentences of the language that satisfies K and for which the Shannon entropy $-\sum_{i=1}^{J} w(\alpha_i) \log(w(\alpha_i))$ is maximised. Since K consists of only linear

constraints, the set of probability functions that satisfy K is convex and so is the function $f(x) = -\sum_{i=1}^{J} x_i \log(x_i)$, hence the uniqueness.
An alternative approach is studied by Williamson [16], which we will denote by ME_W. In this approach Maximum Entropy probability functions that satisfy a set of constraints K are defined by minimising the divergence from the probability function $P_=$, which has the maximum Shannon entropy. The information theoretic divergence of a probability function W from the probability function V is given by their relative entropy and defined as:

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

Williamson defines the Maximum Entropy probability function for a set of constraints K, $ME_W(K)$, as the probability function w, that satisfies K and has the minimum relative entropy to $P_=$, i.e. $\sum_{i=1}^{J} w(\alpha_i) \log\left(\frac{w(\alpha_i)}{P_=(\alpha_i)}\right)$, amongst all those probability functions that satisfy K.

Proposition 1 *Let $\mathcal{L}$ be a propositional language and K a set of linear constraints. Then $ME(K)(\phi) = ME_W(K)(\phi)$ for all $\phi \in S\mathcal{L}$.*

Proof. Let $\alpha_1, \ldots, \alpha_J$ be the atoms of $\mathcal{L}$. Notice that

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

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

Let w be a probability function that satisfies K then w minimises $RE(w, P_=)$ if and only if w maximises $E(w)$. Hence $ME_W(K)$ and $ME(K)$ specify the same probability function. ■

Thus the two approaches agree for constraint sets from a propositional language. The main difficulty for extending these definitions to first order languages is that in the case of a first order language one does not have access to the atomic sentences in order to express the entropy or the relative entropy. In the first order case one has only access to state descriptions over finite sub-languages.
To extend the first approach to a first order language L, Barnett and Paris [1], propose to define the Maximum Entropy probability function that satisfies K as the limit of the Maximum Entropy models of K restricted to finite sub-languages, L^k. These finite sub-languages can essentially be treated as propositional languages where the Maximum Entropy models are well defined for every set of linear constraints. To be more precise let L be a first order language with relation symbols $RL = \{R_1, \ldots, R_t\}$ and constant

[1] Notice that RE is not a distance measure since it is not symmetric, so it is not the distance between W and V but rather the divergence of W from V.

symbols $\{a_1, a_2, \ldots\}$, and let K be a set of linear constraints as above. Define $\mathcal{L}^r$ to be the propositional language with propositional variables $R_i(a_{i_1}, \ldots, a_{i_j})$ for $R_i \in RL$ and $a_{i_1}, \ldots, a_{i_j} \in \{a_1, \ldots a_r\}$. If k is the maximum such that a_k appears in K, for $r \geq k$ define $(-)^{(r)} : SL^k \to S\mathcal{L}^r$ as

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

For a set of linear constraints K, let $K^{(r)}$ be the result of replacing every θ appearing in K with $\theta^{(r)}$ and notice that for a state description Θ^k of L^k and $r \geq k$, $(\Theta^k)^{(r)} = \Theta^k$. Barnett and Paris [1], propose to define the Maximum Entropy probability function on first order languages as follows:

Definition 9 *(ME) Let L be a first order language and K a set of linear constraints. For a state description Θ_i^k of L^k, let $ME(K)(\Theta_i^k) = \lim_{r\to\infty} ME(K^{(r)})(\Theta_i^k)$.*

This determines $ME(K)$ on all state descriptions and thus on all quantifier free sentences, which is uniquely extended to all $\psi \in SL$ by Theorem 1.

For the second approach, ME_W, Williamson first defines the r-divergence of a probability function W from a probability function V by

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

where Θ_i^r's are state descriptions of L^r. Thus the r-divergence of W from V is the divergence of W from V when they are restricted to L^r. Then for probability functions U, V and W, U is closer to V than W if there exists N such that for all $r > N$, $d_r(U, V) < d_r(U, W)$. Williamson [16] defines the Maximum Entropy probability functions on first order languages as:

Definition 10 **(ME_W)** *Let K be a set of linear constraints as before. The Maximum Entropy model of K, $ME_W(K)$, is the probability function, w, satisfying K such that there is no probability function v that satisfies K and $d_r(v, P_=) < d_r(w, P_=)$ for all r eventually.*

The main questions here are whether or not the Maximum Entropy probability functions, given by Definitions 9 and 10, are well defined for every constraint set K from a first order language, i.e, whether or not the limit in Definition 9, or the closest probability function to $P_=$ as in Definition 10, exist for every K, and when they are well

defined, whether or not the resulting probability functions satisfy K. In [1] Barnett and Paris showed that for any set of linear constraints over monadic first order languages, the Maximum Entropy probability function is indeed well defined and that it satisfies the constraints. On the other hand in the general case for constraint sets containing sentences with quantifier complexity of Σ_2, Π_2 or higher the Maximum Entropy probability functions that satisfy the constraints are not always well defined (see [13]). The case of Π_1 sentences has been studied and partially answered by Paris and Rafiee Rad in [10] and in this paper we will focus on knowledge bases consisting of a Σ_1 sentence, i.e., constraint sets of the form $\{w(\exists \vec{x}\, \phi(\vec{x})) = 1\}$ where $\phi(\vec{x})$ is quantifier free.

3 The Maximum Entropy Models for Σ_1 Sentences

We will now turn to our main result concerning the Maximum Entropy models of sentences with quantifier complexity of Σ_1. We will show that both approaches for defining Maximum Entropy models are well defined for these sentences and agree with each other. As was pointed out before, for our purpose, every Σ_1 sentence $\exists \vec{x}\theta(\vec{x})$ is identified with the constraint set $\{w(\exists \vec{x}\theta(\vec{x})) = 1\}$.

Lemma 2 *Let $\phi \in SL$ be a satisfiable Σ_1 sentence of the form $\exists x_1, ..., x_t \theta(a_1, \ldots, a_l, \vec{x})$ and let Γ^l_ϕ be the set of state descriptions of L^l that are consistent with ϕ. Then $P_=(\phi \mid \bigvee \Gamma^l_\phi) = 1$.*

Proof. Let $\gamma = \neg\phi = \forall x_1, ..., x_t \neg\theta(\vec{a}, \vec{x})$. Let $\vec{a}$ be all the constants appearing in θ with l the largest such that a_l appears in $\vec{a}$ and let Γ^l be the set of state descriptions of L^l. First notice that for $\Theta^{(l)}_j \in \Gamma^l$ if $\Theta^{(l)}_j \vDash \gamma$ then $\Theta^{(l)}_j \vDash \neg\phi$ and thus $\Theta^{(l)}_j \notin \Gamma^l_\phi$. We show that for every $\Theta^{(l)}_j \in \Gamma^l_\phi$, $P_=(\Theta^{(l)}_j \wedge \gamma) = 0$. If $\Theta^{(l)}_j$ is inconsistent with $\gamma^{(l)}$ then[2] $P_=(\Theta^{(l)}_j \wedge \gamma) = 0$. This is so because if $\Theta^{(l)}_j$ is inconsistent with $\gamma^{(l)}$ then $\Theta^{(l)}_j \vDash \bigvee_{i_1,\ldots,i_t \leq l} \theta(\vec{a}, a_{i_1}, \ldots, a_{i_t})$ so $\Theta^{(l)}_j \vDash \exists \vec{x}\theta(\vec{a}, \vec{x}) \equiv \neg\gamma$. So $P_=(\Theta^{(l)}_j \wedge \gamma) \leq P_=(\neg\gamma \wedge \gamma) = 0$.
Let $\Gamma^l_{\phi,\gamma^{(l)}}$ be the set of state descriptions in Γ^l_ϕ that are consistent with $\gamma^{(l)}$. For $\Theta^{(l)}_j \in \Gamma^l_{\phi,\gamma^{(l)}}$ let $Q_i(\vec{a}, x_1, ..., x_t)$, $i \in I$ enumerate formulae of the form

$$\Theta^{(l)}_j \wedge \bigwedge_{\substack{y_{i_1},\ldots,y_{i_j} \in \{a_1,\ldots,a_l\} \cup \{x_1,\ldots,x_t\} \\ \{y_{i_1},\ldots,y_{i_j}\} \cap \{x_1,\ldots,x_t\} \neq \emptyset \\ R \in RL,\ j\text{-}ary}} \pm R(y_{i_1}, ..., y_{i_j}).$$

Since $\neg\theta(\vec{a}, \vec{x})$ is not a tautology, and since $\Theta^{(l)}_j \nvDash \gamma$ there is some strict subset J of I such that $\vDash \Theta^{(l)}_j \wedge \neg\theta(\vec{a}, \vec{x}) \leftrightarrow \bigvee_{j \in J} Q_j(\vec{a}, \vec{x})$. To see this notice that the sentences $Q_i(\vec{a}, x_1, ..., x_t)$ are state descriptions of a language L but with constants $a_1, \ldots, a_l, x_1, \ldots x_t$, which extend the state description $\Theta^{(l)}_j \in \Gamma^l_{\phi,\gamma^{(l)}}$. Then since $\Theta^{(l)}_j \wedge \neg\theta(\vec{a}, \vec{x})$ is a sentence

[2] Remember that $\gamma^{(l)} = \bigwedge_{i_1,\ldots,i_t \leq l} \neg\theta(\vec{a}, a_{i_1}, \ldots, a_{i_t})$

in the language $L^{a_1,...,a_n,x_1,...x_t}$ that implies $\Theta^{(l)}_j$, it will be equivalent to a disjunction of some of these state descriptions. Now, for $i_1 < i_2 < ... < i_t < r$ the number of extensions of $Q_i(\vec{a}, a_{i_1}, ..., a_{i_t})$ to a state description of L^r is the same for each i so $P_=(Q_i(\vec{a}, a_{i_1}, ..., a_{i_t})) = \frac{1}{|I|}$ and for disjoint $\vec{a}^1, ..., \vec{a}^r$, $P_=(Q_{n_1}(\vec{a}, \vec{a}^1) \wedge ... \wedge Q_{n_r}(\vec{a}, \vec{a}^r)) = \frac{1}{|I|^r}$. So

$$P_=(\Theta^{(l)}_j \wedge \forall x_1, ..., x_t \neg\theta(\vec{a}, \vec{x})) \leq P_=(\Theta^{(l)}_j \wedge \bigwedge_{i=1}^{r} \neg\theta(\vec{a}, \vec{a}^i)) = \sum_{n_1,...,n_r \in J} P_=(\bigwedge_{i=1}^{r} Q_{n_i}(\vec{a}, \vec{a}^i)) = \left(\frac{|J|}{|I|}\right)^r.$$

And $\left(\frac{|J|}{|I|}\right)^r \to 0$ *as* $r \to \infty$. Thus for all $\Theta^{(l)}_j \in \Gamma^l_\phi$, $P_=(\Theta^{(l)}_j \wedge \gamma) = 0$ and thus $P_=(\gamma \,|\, \Theta^{(l)}_j) = 0$. So for every $\Theta^{(l)}_j \in \Gamma^l_\phi$, $P_=(\phi \,|\, \Theta^{(l)}_j) = 1$ and thus $P_=(\phi \,|\, \bigvee \Gamma^l_\phi) = 1$ as required. ■

Theorem 3 *Let ϕ be a satisfiable Σ_1 sentence of the form $\exists x_1, ..., x_t \theta(a_1, \ldots, a_l, \vec{x})$ and let Γ^l_ϕ be the set of state descriptions of L^l that are consistent with ϕ. For $K = \{w(\phi) = 1\}$ and $\psi \in SL$, $ME_W(K)(\psi) = P_=(\psi \,|\, \bigvee \Gamma^l_\phi)$.*

Proof. First by Lemma 2, $P_=(- \,|\, \bigvee \Gamma^l_\phi)$ satisfies K. It is also the closest probability function to $P_=$ that satisfies K. To see this notice that if w is a probability function that satisfies K then $w(\phi) = 1$. Thus for all $k \geq l$, both w and $P_=(- \,|\, \bigvee \Gamma^l_\phi)$ assign probability zero to the state descriptions of L^k that are inconsistent with $\phi^{(k)}$. For those state descriptions that are consistent with $\phi^{(k)}$, $P_=(- \,|\, \bigvee \Gamma^l_\phi)$ assigns equal probability while w assigns different probability to at least some of them. Thus for $k \geq l$ on each L^k, $P_=(- \,|\, \bigvee \Gamma^l_\phi)$ has a higher entropy that w and thus has a smaller k-divergence from $P_=$. Hence by definition $P_=(- \,|\, \bigvee \Gamma^l_\phi)$ is closer than w to $P_=$. ■

Theorem 3 specifies the Maximum Entropy models for Σ_1 sentences as characterised by ME_W and Definition 10. We shall now turn to the Maximum Entropy models as characterised by ME and the limit in the Definition 9.

Theorem 4 *Let ϕ be the satisfiable Σ_1 sentence $\exists \vec{x} \theta(a_1, \ldots, a_l, \vec{x})$, Γ^l_ϕ be the set of state descriptions of L^l that are consistent with ϕ and $K = \{w(\phi) = 1\}$. Then for $\psi \in SL$, $ME(K)(\psi) = P_=(\psi \,|\, \bigvee \Gamma^l_\phi)$.*

Proof.
Let $\Lambda = \bigvee \Gamma^l_\phi$. We will show that for quantifier free ψ, $ME(K)(\psi) = P_=(\psi \,|\, \Lambda)$. This establishes that $ME(K)$ agrees with $P_=(- \,|\, \Lambda)$ on quantifier free sentences and thus, by Theorem 1, they will agree on all SL, that is, for all $\psi \in SL$, $ME(K)(\psi) = P_=(\psi \,|\, \Lambda)$.
Let Γ^r be the set of state descriptions of L^r and Γ^r_K be the subset of Γ^r that satisfy $\phi^{(r)}$. For $\Theta^k_i \in \Gamma^k$ define for $r \geq k$, $\Gamma^r_{k,i} = \{\Psi^r_j \in \Gamma^r \,|\, \Psi^r_j \vDash \Theta^k_i\}$. In other words, $\Gamma^r_{k,i}$ is the set of state description of L^r that extend the state description Θ^k_i of L^k. Notice that $|\Gamma^r_{k,i}| = |\Gamma^r_{k,j}|$ for $\Theta^k_i, \Theta^k_j \in \Gamma^k$ because state descriptions of L^k will all have the same number of extensions to state descriptions of L^{k+1}. Let ${}^K\Gamma^r_{k,i} = \Gamma^r_K \cap \Gamma^r_{k,i}$ be the set of

extensions of Θ^k_i to a state description of L^r that satisfies $\phi^{(r)}$. Take Γ^l_ϕ as the set of state descriptions of L^l that are consistent with ϕ, and let $\Gamma^l_{\neg\phi} = \Gamma^l - \Gamma^l_\phi$.

Notice that $ME(K^{(r)})$ assigns probability zero to those state descriptions of L^r that are inconsistent with $\phi^{(r)}$ (so those not in Γ^r_K) since it should assign probability 1 to $\phi^{(r)}$,

$$\Psi^r \in \Gamma^r \setminus \Gamma^r_K, \quad ME(K^{(r)})(\Psi^r) = 0. \tag{3}$$

Next notice also that $ME(K^{(r)})$ assigns equal probability to those state descriptions that are consistent with $\phi^{(r)}$ (i.e to those in Γ^r_K). To see this, suppose not and define the probability function w on SL^r that agrees with $ME(K^{(r)})$ (i.e. assigns zero probability) on those state descriptions that are inconsistent with $\phi^{(r)}$ but divides the full probability measure equally among those in Γ^r_K. Then w satisfies $K^{(r)}$ but it is easy to check that w has strictly higher entropy than $ME(K^{(r)})$, on L^r, which is a contradiction with the choice of $ME(K^{(r)})$ as the Maximum Entropy probability function on L^r that satisfies $K^{(r)}$, so

$$\Psi^r \in \Gamma^r_K, \quad ME(K^{(r)})(\Psi^r) = \frac{1}{|\Gamma^r_K|}. \tag{4}$$

Thus by (3) and (4), for the state description Θ^k_i, $k \geq l$,

$$ME(K^{(r)})(\Theta^k_i) = \sum_{\substack{\Psi^r \in \Gamma^r \\ \Psi^r \vDash \Theta^k_i}} ME(K^{(r)})(\Psi^r) = \sum_{\substack{\Psi^r \in \Gamma^r_K \\ \Psi^r \vDash \Theta^k_i}} ME(K^{(r)})(\Psi^r) = \frac{|{}^K\Gamma^r_{k,i}|}{|\Gamma^r_K|}.$$

The state descriptions in $\Gamma^l_{\neg\phi}$ are inconsistent with ϕ and thus have no extension to a state description of L^r that satisfies $\phi^{(r)}$. Hence Γ^r_K includes only extensions of state descriptions in Γ^l_ϕ and we have $\Gamma^r_K = \bigcup_{\Theta^l_j \in \Gamma^l_\phi} {}^K\Gamma^r_{l,j}$ and since ${}^K\Gamma^r_{l,j}$'s include extensions of different state description of L^l and are thus disjoint,

$$|\Gamma^r_K| = \sum_{\Theta^l_j \in \Gamma^l_\phi} |{}^K\Gamma^r_{l,j}|. \tag{5}$$

On the other hand, for $k \geq l$, $P_=(- \mid \Lambda)$ assigns equal probabilities to all state descriptions of L^k that are consistent with $\Lambda = \bigvee \Gamma^l_\phi$ and zero to those that are not. Thus those with non-zero probability are exactly those state descriptions of L^k that are extensions of some state description in Γ^l_ϕ and the number of these state descriptions is $\sum_{\Theta^l_j \in \Gamma^l_\phi} |\Gamma^k_{l,j}|$. Thus $P_=(\Theta^k_i|\Lambda) = 0$ if Θ^k_i extends a state description in $\Gamma^l_{\neg\phi}$ and $P_=(\Theta^k_i|\Lambda) = \frac{1}{\sum_{\Theta^l_j \in \Gamma^l_\phi} |\Gamma^k_{l,j}|}$ if Θ^k_i extends a state description in Γ^l_ϕ.

To show $ME(K)(\psi) = P_=(\psi \mid \Lambda)$ for quantifier free ψ, it is enough to show that for each k and each state description $\Theta^k_i \in \Gamma^k$, $ME(K)(\Theta^k_i) = P_=(\Theta^k_i \mid \Lambda)$. By definition, this is

$$\lim_{r \to \infty} ME(K^{(r)})(\Theta^k_i) = P_=(\Theta^k_i \mid \Lambda). \tag{6}$$

For $k \geq l$, the state descriptions of L^k are extensions of either a state description in Γ^l_ϕ or a state description in $\Gamma^l_{\neg\phi}$. The state description in $\Gamma^l_{\neg\phi}$ are inconsistent with ϕ and thus have no extension to L^r that satisfies $\phi^{(r)}$, that is

$$^K\Gamma^r_{k,s} = \emptyset \quad \text{for} \quad \Theta^k_s \in \Gamma^l_{\neg\phi},$$

and so $ME(K^{(r)})(\Theta^k_s) = 0$. Hence for those Θ^k_i that extend a state description in $\Gamma^l_{\neg\phi}$,

$$\lim_{r\to\infty} ME(K^{(r)})(\Theta^k_i) = 0 = P_=(\Theta^k_i|\Lambda).$$

For those Θ^k_i that extend a state description in Γ^l_ϕ, we have to show that

$$\lim_{r\to\infty} \frac{|^K\Gamma^r_{k,i}|}{|\Gamma^r_K|} = \frac{1}{\sum_{\Theta^l_j \in \Gamma^l_\phi} |\Gamma^k_{l,j}|}. \tag{7}$$

Using, (5) and the fact that $|\Gamma^k_{l,j}|$ is the same for all $\Theta^l_j \in \Gamma^l_\phi$, to show 7 we will show that[3]

$$\lim_{r\to\infty} \frac{|^K\Gamma^r_{k,i}| \sum_{\Theta^l_j \in \Gamma^l_\phi} |\Gamma^k_{l,j}|}{\sum_{\Theta^l_j \in \Gamma^l_\phi} |^K\Gamma^r_{l,j}|} = \lim_{r\to\infty} \frac{|^K\Gamma^r_{k,i}|\,|\Gamma^l_\phi|\,|\Gamma^k_{l,j}|}{\sum_{\Theta^l_j \in \Gamma^l_\phi} |^K\Gamma^r_{l,j}|} = 1. \tag{8}$$

Lemma 5 *Let K, $^K\Gamma^r_{k,i}$ and $\Gamma^r_{k,i}$ be as defined above then* $\lim_{r\to\infty} \frac{|^K\Gamma^r_{k,i}|}{|\Gamma^r_{k,i}|} = 1$.

Proof.
Notice that $|\frac{^K\Gamma^r_{k,i}}{\Gamma^r_{k,i}}|$ is the probability that a random extension of the state description $\Theta^k_i \in \Gamma^k$ to a state description of L^r will satisfy the $K^{(r)}$.[4] Remember that K consists of a Σ_1 sentence $\exists x_1, ..., x_t \theta(\vec{a}, x_1, ..., x_t)$, l is the largest that a_l appears in $\theta(\vec{a}, \vec{x})$, and that Θ^k_i extends description in Γ^l_ϕ, say Ψ^l, and let's calculate this probability.
Take $\Theta^k_i \in \Gamma^k$ and let's consider its extensions to state descriptions of L^{k+t}. Let $L^{a_{i_1},...a_{i_n}}$ 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 $L^{\{a_1,...,a_l\}\cup\{a_{k+1},...,a_{k+t}\}}$ that extend Ψ^l (thus they agree with Θ^k_i when restricted to $a_1, \ldots, a_l$) . Then state descriptions of L^{k+t} that are extension of Θ^k_i can be written in the form $\Theta^{k+t}_{i,m} \equiv \Theta^k_i \wedge \Delta_j \wedge V_h(a_1, ..., a_{k+t})$[5] with $m = 1, ..., |\Gamma^{k+t}_{k,i}|$, $j = 1, ..., M$, and $h = 1, ..., \frac{|\Gamma^{k+t}_{k,i}|}{M}$. At least one of the Δ_j's satisfies $\theta(\vec{a}, a_{k+1}, ..., a_{k+t})$ and will hence satisfies $K^{(k+t)}$. The probability that an arbitrary $\Theta^{k+t}_{i,m}$ satisfies $K^{(k+t)}$ will be the number of $\Theta^{k+t}_{i,m}$'s that satisfies $K^{(k+t)}$ divided by the total number of $\Theta^{k+t}_{i,m}$'s that is *at least*, $\frac{|\Gamma^{k+t}_{k,i}|}{M} \cdot \frac{1}{|\Gamma^{k+t}_{k,i}|} = \frac{1}{M}$, and so the probability that a random $\Theta^{k+t}_{i,m}$ does not satisfy $K^{(k+t)}$ will be

[3] Notice that $\sum_{\Theta^l_j \in \Gamma^l_\phi} |\Gamma^k_{l,j}| \neq 0$ and does not depend on r.

[4] The denominator is the total number of extensions of $\Theta^k_i \in \Gamma^k$ to a state description of L^r and the nominator is the number of those extensions of $\Theta^k_i \in \Gamma^k$ to a state description of L^r that satisfy $K^{(r)}$.

[5] $V_h(a_1, ..., a_{k+t})$ enumerate sentence of the form $\bigwedge_{\substack{i_1,...,i_j \leq k+t \\ R \in RLj-arey}} R_i(a_{i_1}, \ldots, a_{i_j})^{\epsilon_{i_1,...,i_j}}$ where $\{a_{i_1}, \ldots, a_{i_j}\}$ intersects both $\{a_{l+1}, \ldots a_k\}$ and $\{a_{k+1}, \ldots, a_{k+t}\}$.

at most as much as the maximum probability that Δ_j does not satisfy $\theta(\vec{d}, a_{k+1}, ..., a_{k+t})$ that is $1 - \frac{1}{M}$. Now consider the extension of Θ_i^k to a state description of L^{k+pt},

$$\Theta_{i,m}^{k+pt} \equiv \Theta_i^k \wedge \Delta_{j_1}^1 \wedge \Delta_{j_2}^2 \wedge ... \wedge \Delta_{j_p}^p \wedge V_h'(a_1, ..., a_{k+pt})$$

with $m = 1, ..., |\Gamma_{k,i}^{k+pt}|$, $j_1, ..., j_p = 1, ..., M$, $h = 1, ..., \frac{|\Gamma_{k,i}^{k+pt}|}{M^p}$ and where Δ_j^s enumerate the state description of $L^{\{a_1,...,a_l\}\cup\{a_{k+(s-1)t+1},...,a_{k+st}\}}$ that extend Ψ^l. The probability that $\Theta_{i,m}^{k+pt}$ does not satisfy $K^{(k+pt)}$ is *at most* as high as the probability that $\Delta_j^1 \nvdash \theta(\vec{d}, a_{k+1}, ..., a_{k+t}), ..., \Delta_j^p \nvdash \theta(\vec{d}, a_{k+(p-1)t+1}, ..., a_{k+pt})$ so $0 \leq 1 - \frac{|^K\Gamma_{k,i}^{k+pt}|}{|\Gamma_{k,i}^{k+pt}|} \leq (1 - \frac{1}{M})^p$. Let $p \to \infty$, then $0 \leq \lim_{r\to\infty} 1 - \frac{|^K\Gamma_{k,i}^r|}{|\Gamma_{k,i}^r|} \leq \lim_{p\to\infty}(1 - \frac{1}{M})^p = 0$. Hence, we have $\lim_{r\to\infty} 1 - \frac{|^K\Gamma_{k,i}^r|}{|\Gamma_{k,i}^r|} = 0$ and $\lim_{r\to\infty} \frac{|^K\Gamma_{k,i}^r|}{|\Gamma_{k,i}^r|} = 1$ as required. ■

All state descriptions of L^k have the same number of extensions to a state description of L^r for $k < r$ thus $|\Gamma_{k,i}^r| = |\Gamma_{k,j}^r|$ for $\Theta_i^k, \Theta_j^k \in \Gamma^k$ and also $|\Gamma_{l,j}^r|$ is the same for all $\Theta_j^l \in \Gamma_\phi^l$. Hence, $|\Gamma_{l,j}^k|\,|\Gamma_{k,i}^r| = |\Gamma_{l,j}^r|$[6] and so,

$$\lim_{r\to\infty} \frac{\sum_{\Theta_j^l\in\Gamma_\phi^l} |^K\Gamma_{l,j}^r|}{|\Gamma_{l,j}^k|\,|\Gamma_{k,i}^r|} = \lim_{r\to\infty} \sum_{\Theta_j^l\in\Gamma_\phi^l} \frac{|^K\Gamma_{l,j}^r|}{|\Gamma_{l,j}^r|} = \sum_{\Theta_j^l\in\Gamma_\phi^l} \lim_{r\to\infty} \frac{|^K\Gamma_{l,j}^r|}{|\Gamma_{l,j}^r|} = |\Gamma_\phi^l|$$

where the last equality follows from Lemma 5. Then

$$\lim_{r\to\infty} \frac{|^K\Gamma_{k,i}^r|\,|\Gamma_\phi^l|\,|\Gamma_{l,j}^k|}{\sum_{\Theta_j^l\in\Gamma_\phi^l} |^K\Gamma_{l,j}^r|} = |\Gamma_\phi^l| \lim_{r\to\infty} \frac{|^K\Gamma_{k,i}^r|}{|\Gamma_{k,i}^r|} \lim_{r\to\infty} \frac{|\Gamma_{l,j}^k|\,|\Gamma_{k,i}^r|}{\sum_{\Theta_j^l\in\Gamma_\phi^l} |^K\Gamma_{l,j}^r|} = 1$$

and this establishes 8 as required and completes the proof. ■

Corollary 1 *For a knowledge base K consisting of a Σ_1 sentence, and a sentence $\psi \in SL$, $ME(K)(\psi) = ME_w(K)(\psi)$.*

4 Discussion

We studied the Maximum Entropy probability functions as the canonical characterisation of some under-determined structure about which we have some partial information. The strongest candidate for this characterisation is the "least informative" probability function that satisfies the given partial information which is in turn formalised in terms of (relative) Shannon Entropy.

For propositional languages, the Maximum Entropy probability function that satisfies

[6] What this says is that the number of extensions of Θ_j^l to a state description of L^k times the number of extensions of a state description of L^k to an state description of L^r (which is the same for all $\Theta_i^k \in \Gamma^k$), is equal to the number of extensions of Θ_j^l to an state description of L^r.

a given set of linear constraints is well defined and has been extensively studied. Our goal in this paper was to contribute to the investigation of these probability functions for first order languages. Barnett and Paris had shown in [1] that such probability functions are well defined for constraint sets from a monadic first order language. The case of Π_1 sentences has been investigated and partially answered by Paris and Rafiee Rad in [10] while for the sentences with the quantifier complexity of Σ_2, Π_2 or above these models are not necessarily well defined.
In this paper we have proved that the Maximum Entropy models are well defined for Σ_1 sentences and showed how these models are closely related to $P_=$, the most non-committal probability function. Furthermore, we showed that the two main approaches to defining Maximum Entropy models on first order languages, agree on the Σ_1 sentences.

References

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

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

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

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

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

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

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

[8] Landes, J. and Williamson, J. "Objective Bayesianism and Maximum Entropy Principle", in *Entropy*,15(9):3528–3591, 2013.

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

[10] Paris, J.B. and Rad, S.R., "A note on the least informative model of a theory", in *Programs, Proofs, Processes, CiE 2010*, Eds. F. Ferreira, B. Lwe, E. Mayordomo, and L. Mendes Gomes, Springer LNCS 6158, 342–351, 2010.

[11] Paris, J.B. and Vencovská, "In defence of the maximum entropy inference process", in *International Journal of Approximate Reasoning*, 17(1):77–103, 1997.

[12] Paris, J.B. and Vencovská, *Pure Inductive Logic*, Cambridge University Press, 2015.

[13] Rafiee Rad, S., *Inference Processes for First Order Probabilistic Languages*, PhD Thesis, University of Manchester 2009. available at `http://www.maths.manchester.ac.uk/~jeff/`

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

[15] Williamson, J., "From Bayesian epistemology to inductive logic", in *Journal of Applied Logic*, 2, 2013.

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

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