INFERENCE PROCESSES FOR PROBABILISTIC FIRST ORDER LANGUAGES

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2009

By
Soroush Rafiee Rad
Mathematics

Contents

ABSTRACT.

In this thesis we will investigate inference processes for predicate languages. The main question we are concerned with in this thesis is how to choose a probability function amongst those that satisfy a certain knowledge base. This question has been extensively studied for propositional logic and we shall investigate it for first order languages. We will first study the generalisation of Minimum Distance, MD, and Centre of Mass, CM_∞, inference processes to unary predicate languages and then we will investigate the generalisations of the Maximum Entropy inference process to general polyadic languages.

For the case of the Maximum Entropy inference process we will study and compare two generalisations, the BP-method and the W-method. We will show that the two methods agree for the unary and Σ_1 knowledge bases and we conjecture that the result holds for the Π_1 knowledge bases too.

We shall show that neither of these generalisations for the Maximum Entropy inference process is universally well defined for a first order language and we shall study some of the problems associated with generalising this inference process to polyadic languages.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

I would like to thank my father and my mother for their constant support and help and for providing me with a lifetime of opportunities that have led to this point.

I am enormously grateful and heavily indebted to Jeff Paris for his supervision, guidance, inspiration and patience in the last three years. It has been a great privilege to work under his supervision.

I am also very grateful to George Wilmers for his support and help through out the last four years without which this course of study would have not been possible for me.

A considerable part of this research has been funded by MATHLOGAP fellowship provided by the Marie Curie Fellowship Association to whom I am also very grateful.

Chapter 1

Introduction

In this thesis we are concerned with the general question of deriving information from uncertain and probabilistic knowledge bases. The problem is that given a knowledge base K consisting of linear information about the degree of belief of an agent in certain sentences in the language what assertions are 'rational' to make about the agent's degree of belief in other sentences in the language.

Such knowledge bases appear constantly in everyday situations. The uncertainty of the information in our knowledge base can be the result of many different factors such as insufficient information, margins of error in measurements or reliability of sources, etc; but it is hardly the case that we can come up with a knowledge base consisting of only certain truth values (0 or 1) or we will be left with very restricted knowledge bases and will lose the advantages we can have in using the partial information available.

A fundamental underlying assumption for all the material covered in this thesis is the assumption that a 'rational' agent's degree of belief (in sentences of a language) shall be represented with a *probability function*. An extensive amount of work has been done on this subject and many justifications have been presented for this assumption, for example the Dutch Book argument, to name one, which we believe to have provided a final answer to this debate and we shall not pursue this matter any further here [see [24]-Chapter 3 for more discussion].

Being convinced by these justifications we will work in the framework of inductive logic with the agent's knowledge being given by a set of linear constraints on the probabilities given to certain sentences of the language.

In this setting, the question of what degree of belief should the agent assign to a sentence ϕ on the basis of a knowledge base K, will be answered by taking a probability function representing the agent's knowledge base, w_K, and then calculating the probability of the sentence ϕ, $w_K(\phi)$, accordingly. However this assumption alone will leave us with a (usually infinite) set of probability functions consistent with the knowledge base and thus the justification for the agent's answer for $w_K(\phi)$ will depend on the justifiability of the choice of probability function w_K from this set.[1]

We will thus need to impose further conditions in order to narrow down this choice and ideally be left with a unique probability function where possible. These conditions will be the main tool for comparing different choices of probability functions and one such choice shall be preferred to another when a larger number of these conditions are satisfied. Therefore the justifiability of our choice shall be assessed via these conditions. This means that the further restrictions we wish to impose should enforce only the properties that are a priori accepted as justified or desirable by common sense and hence comes the name *Common Sense Principles*.

On this basis we shall investigate different methods of choosing a probability function satisfying a certain knowledge base, to represent an agent's belief function based on common sense principles.

A lot of effort has been put into this problem for the case of propositional languages and widely accepted and justified answers have been given to the question of how to choose this probability function, see for example [1], [2], [6], [13], [24], [26], [29]. The same question for first order languages, however, is far from being settled.

In this thesis we will study some of the proposals given for the propositional case for first order languages to find out the extent to which they can be generalized both to unary first order languages where the generalization seems to go through without any difficulties and the more problematic polyadic cases.

We shall first introduce some notations and definitions that we will be using throughout

[1] In what follows we will drop this subscript K since the knowledge base is always fixed in the context.

this thesis and introduce the conditions and criteria we consider for the choice of the probability function. The important role of these conditions is to ensure rationality or more precisely justifiability of the choice of a certain probability function.

1.1 Framework and Notation

Throughout this thesis we will work with a first order language L with finitely many relation symbols, no function symbols and countably many constant symbols $a_1, a_2, a_3, \ldots$. Furthermore we assume that these individuals exhaust the universe. Equality shall only be assumed to exist in the language when it is explicitly stated.

Let RL, FL, SL and TL denote the set of relation symbols, the set of formulae, the set of sentences and the set of term models for L respectively, where a term model is a structure M for the language L with domain $|M| = \{a_i \mid i = 1, 2, \ldots\}$ where every constant symbol is interpreted as itself.

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

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

A knowledge base K is taken to be a *satisfiable* set of linear constraints of the form

$$\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i, \quad i = 1, 2, \ldots, m$$

where $\theta_j \in SL$, $a_{ij}, b_j \in \mathbb{R}$ and w is a probability function. In this setting then, the main question we are interested in is: Given a set K as above, what value should be given to $w(\phi)$ for an arbitrary $\phi \in SL$ on the basis of K.

In the case when the knowledge base K determines the probability function w uniquely the answer to the above question will of course be clear. However this is hardly ever the case and when it is not, deciding the values $w(\phi)$ for all $\phi \in SL$ on the basis of K will be equivalent to choosing a probability function amongst all probability functions

that satisfy K. We shall expand our notation further before returning to this question.

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

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

For every sentence $\phi \in S\mathcal{L}$ there is unique set $\Gamma_\phi \subseteq \{\alpha_i \mid i = 1, \ldots, J\}$ such that

$$\models \phi \leftrightarrow \bigvee_{\alpha_i \in \Gamma_\phi} \alpha_i.$$

It can be easily checked that $\Gamma_\phi = \{\alpha_j \mid \alpha_j \models \phi\}$.

Thus if $w : S\mathcal{L} \to [0\,,\ 1]$ is a probability function then

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

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, that is by the vector

$$< w(\alpha_1), \ldots, w(\alpha_J) > \in \mathbb{D}^{\mathcal{L}} \quad \text{where} \quad \mathbb{D}^{\mathcal{L}} = \{\vec{x} \in \mathbb{R}^J \mid \vec{x} \geq 0, \sum_{i=1}^{J} x_i = 1\}.$$

Conversely if $\vec{a} \in \mathbb{D}^{\mathcal{L}}$ we can define a probability function $w' : S\mathcal{L} \to [0\,,\ 1]$ such that $< w'(\alpha_1), \ldots, w'(\alpha_J) >= \vec{a}$ by setting

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

This gives a one to one correspondence between the probability functions on $\mathcal{L}$ and the points in $\mathbb{D}^{\mathcal{L}}$.

Let K be a knowledge base as above, this time for the propositional language $\mathcal{L}$. So K

will be a consistent set of linear constraints

$$\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i \quad i = 1, \ldots, m,$$

where each θ_j is a sentence of $\mathcal{L}$. Replacing each $w(\theta_j)$ in K with $\sum_{\alpha_i \models \theta_j} w(\alpha_i)$ and adding the equation $\sum_{i=1}^{J} w(\alpha_i) = 1$ we will get a system of linear equations

$$< w(\alpha_1), \ldots, w(\alpha_J) > A_K = \vec{b}_K.$$

Thus if the probability function w satisfies K the vector $< w(\alpha_1), \ldots, w(\alpha_J)) >$ 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 $V^{\mathcal{L}}(K)$, that is

$$V^{\mathcal{L}}(K) = \{ \vec{x} \in \mathbb{R}^J \,|\vec{x} \geq 0,\ \vec{x} A_K = \vec{b}_K \} \subseteq \mathbb{D}^{\mathcal{L}}.$$

Thus the question of choosing a probability function satisfying K will be equivalent to the question of choosing a point in $V^L(K)$. We shall use this equivalence frequently in what follows.

The following theorem, due to Gaifman, provides a similar result for the case of a first order language L. Let $QFSL$ be the set of quantifier free sentences of the 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 $QFSL$.*

See [24] for a proof.

Let L be a first order language with relation symbols $R_1, R_2, \ldots, R_t$, let $L^{(k)}$ be a sub-language of L with only constant symbols $a_1, \ldots, a_k$ and let $\Theta_1^{(k)}, \ldots, \Theta_{n_k}^{(k)}$ enumerate all the sentences of the form

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

We shall call these $\Theta_i^{(k)}$'s the *state descriptions* of $L^{(k)}$. For $\theta \in QFSL$ 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 $R(a_{i_1}, \ldots, a_{i_j})$ for $i_1, \ldots, i_j \leq k$, $R \in RL$ and R $j-ary$. Then the sentences $\Theta_i^{(k)}$ will be the atoms of $\mathcal{L}^{(k)}$ and

$$\theta \leftrightarrow \bigvee_{\Theta_i^{(k)} \models \theta} \Theta_i^{(k)}$$

so

$$w(\theta) = \sum_{\Theta_i^{(k)} \models \theta} w(\Theta_i^{(k)}).$$

Thus to determine the value $w(\theta)$ we only need to determine the values $w(\Theta_i^{(k)})$ and to require

- $w(\Theta_i^{(k)}) \geq 0$ and $\sum_{i=1}^{n_k} w(\Theta_i^{(k)}) = 1$.
- $w(\Theta_i^{(k)}) = \sum_{\Theta_j^{(k+1)} \models \Theta_i^{(k)}} w(\Theta_j^{(k+1)})$.

to ensure that w satisfies P1 and P2.

Having set our framework we will return to the main question, that is, how to choose a probability function satisfying a given set K of linear constraints. To be able to answer this question we will first need to define and formalize criteria that allow us to compare these choice processes or inference processes as we shall call them. These criteria are intended to reflect the 'rational' and 'common-sensical' behavior we shall expect the chosen probability function and the choice process to demonstrate.

1.2 Principles Of Uncertain Reasoning

Principles of Uncertain Reasoning [see [24]], are conditions and restrictions on the way a probability function is chosen from a set of probability functions. These principles we shall now state are considered desirable or 'rational' for inference processes to satisfy and they shall remain our main criteria to favor one such process to another. These principles are all well formulated and studied for propositional languages and while generalizing an inference precess to first order languages we shall seek a generalization that preserves the principles that were satisfied by the inference process in the propositional case where possible.

Let $N(K)$ be a choice function that chooses a probability function satisfying K. The following is a list of common sense conditions desirable for N.

Equivalence Principle

If K_1 and K_2 are equivalent in the sense that $V^L(K_1) = V^L(K_2)$ then $N(K_1) = N(K_2)$.

Principle of Irrelevant Information

Let K_1, K_2 as above and $\theta \in SL$ but no propositional variable appearing in θ or any sentence in K_1 also appears in K_2. Then

$$N(K_1 + K_2)(\theta) = N(K_1)(\theta).$$

Continuity

For $\theta \in SL$ a microscopic change in the knowledge base K should not result in macroscopic changes in the value $N(K)(\theta)$.

Open-Mindedness Principle

For K as above and $\theta \in SL$, if $K + w(\theta) \neq 0$ is consistent then $N(K)(\theta) \neq 0$.

Renaming Principle

Suppose

$$K_1 = \{ \sum_{j=1}^{J} a_{ji} w(\gamma_j) = b_i \mid i = 1, 2, \ldots, m \},$$

$$K_2 = \{ \sum_{j=1}^{J} a_{ji} w(\delta_j) = b_i \mid i = 1, 2, \ldots, m \},$$

where $\gamma_1, \ldots, \gamma_J, \delta_1, \ldots, \delta_J$ are permutations of $\alpha_1, \ldots, \alpha_J$. Then

$$N(K_1)(\gamma_j) = N(K_2)(\delta_j).$$

Obstinacy Principle

Let K_1 and K_2 as above be such that $N(K_1)$ satisfies K_2. Then $N(K_1 + K_2) = N(K_1)$.

Language Invariance

Suppose we have a family of inference processes N^L, one for each finite language L. Then this family is said to be language invariant if whenever $L_1 \subseteq L_2$ (so $SL_1 \subseteq SL_2$ and $CL_1 \subseteq CL_2$) and $K \in CL_1$ then $N^{L_2}(K)$ agrees with $N^{L_1}(K)$ on SL_1.

Relativisation Principle

Suppose $K_1, K_2 \in CL$, $0 < c < 1$ and

$$K_1 = \{ w(\phi) = c \} + \left\{ \sum_{j=1}^{r} a_{ji} w(\theta_j \mid \phi) = b_i \mid i = 1, \ldots, m \right\},$$

$$K_2 = K_1 + \left\{ \sum_{j=1}^{q} e_{ji} w(\psi_j \mid \neg\phi) = f_i \mid i = 1, \ldots, s \right\}.$$

Then for $\theta \in SL$, $N(K_1)(\theta \mid \phi) = N(K_2)(\theta \mid \phi)$.

See [24] for a more comprehensive list of principles and more discussions and justifications.

The main reason that these principles are important to our purpose is the crucial role they play in justifying the Maximum Entropy inference process (to which most of this thesis is dedicated) as the most 'rational' choice for the probability function.

In the next chapter we will study two well known inference processes, Limiting Centre of Mass CM_∞ and Minimum Distance MD (both defined for propositional logic) for first order languages. In chapters 3 and 4 we will study two methods for generalizing the Maximum Entropy inference process to the first order languages and we shall provide a comparison between the two.

Chapter 2

Inference Processes for quantified knowledge

In the view of the main question introduced in the previous chapter we are interested to find out what probability is 'rational' to assign to an arbitrary sentence of the language on the basis of a knowledge base consisting of a set of probabilistic constraints on the intelligent agent's belief function.

Answering this question will amount to choosing a probability function amongst all probability functions that satisfy the required constraints and take the value for our arbitrary sentence accordingly. Hence the above question will be equivalent to the question that given a set of linear probabilistic constraints what will be the most 'rational' probability function satisfying this set. A number of possible answers to this question have been proposed both for propositional and predicate languages, for example [1], [2], [6], [13], [14], [15], [24], [26], [27], [28], [29], based on various underlying assumptions about the form and origin of the knowledge and the probability function w, see [6] for a discussion.

An inference process is a process of choosing one such probability function in the set of all probability functions satisfying the constraints in the knowledge base.

Definition 1 *An inference process on L is a function that on each $K \in CL$ gives a probability function on SL that satisfies K.*

The result of applying an inference process to a knowledge base will then be a probability function on SL and this probability function is taken to represents the agent's degree of belief in any sentence of the language. The inference process itself can be regarded as a model for the agent's deduction system.

The possibility of being used as a tool to model the deduction system of intelligent agents is one of the main reasons that make inference processes an interesting topic of investigation. In this context intelligent agents can be regarded as inference processes being applied to a knowledge base, assuming of course that, enough time being spent, we can find such a set K containing all knowledge accessible to the agent. This approach provides us with a powerful mathematical machinery to formulate and study the consistency and common sense criteria we expect from intelligent agents in terms of mathematical constraints on these inference processes [see [24]-chapter 6].

The main criterion to prefer one such inference process over another will thus be the extent to which that inference process, or more precisely the probability function chosen by that inference process, satisfies consistency and common sense principles. It is important to note that 'rational' in the above context is indeed an attempt to emphasis the importance of the justifiability of our answer for our intended purpose rather than an invitation to follow a certain theory of rationality. Thus the notion of rationality for our answer will not necessarily have so much of a universal interpretation. Rather one might believe that it does actually depend on the purpose or the context. However there is an inference process, namely Maximum Entropy, that is widely accepted to be the most rational or more precisely the most satisfactory according to the number of common sense principles it satisfies. Investigation of this inference process will form the major part of this thesis [see chapter 3 and 4].

The justifications for different inference processes can generally be classified into two categories; in the first the inference process is justified by typicality of its output. That is, the inference process will choose the probability function that is as representative or as average as possible among all the probability functions that satisfy the knowledge base. Another group of inference processes are justified by the amount of information contained in the probability function chosen by that inference process where it is intended to choose the probability function that contains the least amount of information beyond a knowledge base among all the probability functions satisfying it. In what

follows we shall study three examples of inference processes; The Centre of Mass inference process of the former group and Minimum Distance and Maximum Entropy inference processes of the latter.

As mentioned above these inference processes differ in the set of common sense principles they satisfy and thus will be found suitable for different purposes and contexts. However the Maximum Entropy is most widely agreed upon and accepted as the most commonsensical inference process not only because it satisfies the largest number of principles but also because it is the only inference process satisfying them all [see [24]].

In the rest of his chapter we will present a generalisation of Minimum Distance, henceforth referred to as MD, and limiting Centre of Mass, henceforth referred to as CM_∞ (both defined for propositional logic) to unary predicate languages by using the method introduced by Paris and Barnett [6] for generalising the Maximum Entropy inference process, which we will refer to as the BP-method.

The BP-method

What we refer to here as the BP-method is the general approach of defining the application of an inference process N for a first order language L to a knowledge base K as the limiting case of its application on the propositional finite sublanguages of L, should this limit exist. We will study the application of this method to different inference processes and we will refer to all cases as the BP-method. The reason for this convention is that, except for when it is explicitly pointed out that we are working with general inference processes, the specific inference process we are working with is clear from context and this will help avoiding unnecessary complexity in the notation.

In the next two chapters we will introduce an alternative method for generalising Maximum Entropy to unary languages now augmented with equality and present a more detailed study of this inference process for special cases of polyadic languages. We will also study an alternative generalization of Maximum entropy introduced by Jon Williamson which we shall refer to as the W-method.

2.1 The Minimum Distance Inference Process

The Minimum Distance inference process is based on minimising the amount of information beyond the knowledge base which is included in the probability function. Thus it falls into the second category mentioned above. Here we try to minimise this information by minimising the Euclidean distance between the candidate probability function and the probability function representing total ignorance, that is the probability function with minimum information.

To be more precise if a finite propositional language L has propositional variables $p_1, p_2, \ldots, p_n$, a probability function *Bel* on SL is determined by its values on 2^n sentences $\alpha_1, \alpha_2, \ldots, \alpha_{2^n}$ of the form

$$\pm p_1 \wedge \pm p_2 \wedge \ldots \wedge \pm p_n$$

We refer to these as the atoms of the language.

This, as mentioned in Chapter 1, gives a one to one correspondence between the probability functions and the points in $V^L(K)$. Then the minimum distance inference process applied to K, $MD(K)$, is defined to be the probability function satisfying K, for which

$$D(\vec{x}) = \sum_{i=1}^{n} x_i^2$$

is minimal. Equivalently, such that the distance between the associated $\vec{x}$ in $V^L(K)$ and the point $< \frac{1}{2^n}, \ldots, \frac{1}{2^n} >$ representing the least informative probability function is minimal.

As a justification for MD one can argue that since the point $< \frac{1}{2^n}, \ldots, \frac{1}{2^n} >$ represents the probability function with least information to pick the least informative solution for K we should choose the point that is as close as possible to $< \frac{1}{2^n}, \ldots, \frac{1}{2^n} >$. However the usual Euclidean metric will prove not to be the best measure for information at least from the point of view of common sense principles since replacing this with Shannon's measure of uncertainty will introduce the Maximum Entropy solution which provides a more satisfactory answer in the sense that the resulting probability function satisfies more of the principles of uncertain reasoning. The question might still be valid whether all these principles are desirable requirements for every purpose or whether in different

situations some might carry a stronger weight or even lose their validity.

Theorem 2 *The Minimum Distance Inference Process satisfies the principles of equivalence, continuity and renaming and obstinacy but not the principles of irrelevant information, open-mindedness or relativisation.*

see [6].

Here we generalise this inference process to unary predicate languages. The main idea is to define the $MD(K)$ over a language with countable universe M as the limiting case of the MD solutions over finite sublanguages of M. These finite sublanguages can then be treated as propositional languages where the inference process is already defined.

Let L be a language with just constants $a_1, a_2, \ldots$ and finitely many unary predicates $P_1, P_2, \ldots, P_n$ and with no function symbols, nor identity, and define a linear knowledge base on a first order language as a consistent set of linear constraints of the form

$$K = \left\{ \sum_{j=1}^{r} a_{ji} w(\theta_j) = b_i \,|\, i = 1, \ldots, m \right\}$$

where $\theta_j \in SL$ for $j = 1, \ldots, r$[1].

Now let $L^{(k)}$ be the language L with only constant symbols $a_1, \ldots, a_k$ and let $Q_1, \ldots, Q_J$, $J = 2^n$ enumerate all formulas of the form

$$\pm P_1(x) \wedge \ldots \wedge \pm P_n(x),$$

referred to as the atoms of L.

Let $\mathfrak{L}^r$ be the propositional language with the propositional variables $P_j(a_i)$, $i = 1, \ldots, r$ $j = 1, \ldots, n$. For $k < r$ define $()^{(r)} : SL^k \to S\mathfrak{L}^r$ inductively as follows:

$$(P_j(a_i))^{(r)} = P_j(a_i)$$

[1]This is the same as the definition of a knowledge base for a propositional language except that the sentences θ_j here, are from a first order language.

$$(\neg\phi)^{(r)} = \neg\phi^{(r)}$$

$$(\phi \vee \theta)^{(r)} = \phi^{(r)} \vee \theta^{(r)}$$

$$(\phi \wedge \theta)^{(r)} = \phi^{(r)} \wedge \theta^{(r)}$$

$$(\exists x\psi(x))^{(r)} = \bigvee_{i=1}^{r} \psi(a_i)^{(r)}$$

Barnett and Paris proved the following Lemmas 3, 4 and Theorem 5 in [6] which we shall use in the later sections:

Lemma 3 *If $\theta, \phi \in SL^{(k)}$ and $k \leq r$ and $\theta \equiv \phi$ then $\theta^{(r)} \equiv \phi^{(r)}$.*

Let α_i for $i = 1, \ldots, J^k$ enumerate the exhaustive and exclusive set of sentences of the form

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

Lemma 4 *Any sentence $\theta \in SL^k$ is equivalent to a disjunction of consistent sentences $\phi_{i,\vec{\epsilon}}$ of the form*

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

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

Theorem 5 *If K is finite, satisfiable set of linear constraints over L then $K^{(r)}$ is also satisfiable over Ω^r for large enough r, where $K^{(r)}$ is the set K in which every sentence $\theta \in SL$ is replaced with $\theta^{(r)} \in SL^{(r)}$, i.e. where if*

$$K = \left\{ \sum_{j=1}^{r} a_{ji} w(\theta_j) = b_i \,|\, i = 1, \ldots, m \right\}$$

then

$$K^{(r)} = \left\{ \sum_{j=1}^{r} a_{ji} w(\theta_j^{(r)}) = b_i \,|\, i = 1, \ldots, m \right\}.$$

Theorem 6 *For $\theta \in SL$:*

$$Bel(\theta) = \lim_{r \to \infty} MD(K^{(r)})(\theta^{(r)})$$

exists and is a probability function on L that satisfies K.

Proof. By Lemma 4 every consistent sentence $\theta(a_1, \ldots, a_k) \in SL$ is equivalent to a disjunction of consistent sentences of the form

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

If $\alpha_i = \bigwedge_{j=1}^{k} Q_{m_j}(a_j)$ then let

$$A_i = \{ m_j \mid j = 1, \ldots, k \},\ P_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = \{ j \mid \epsilon_j = 1 \},\ P_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = \{ j \mid j \in P_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \text{ and } j \notin A_i \}$$

so

$$\phi_{i,\vec{\epsilon}}^{(r)} = \alpha_i \wedge \bigwedge_{j=1}^{J} (\bigvee_{t=1}^{r} Q_j(a_t))^{\epsilon_j}$$

will be equivalent to

$$\bigvee_{\substack{m_j \in P_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \text{ for } j=k+1,\ldots,r \\ P_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \subseteq \{ m_j \mid k+1 \le j \le r \}}} (\alpha_i \wedge \bigwedge_{j=k+1}^{r} Q_{m_j}(a_j)) \tag{2.1}$$

If we set

$$p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = |P_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}|, \text{ and } p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = |P_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}|$$

the number of disjuncts in 2.1 will be

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

so if we set

$$x_{i,\vec{\epsilon}} = Bel(\phi_{i,\vec{\epsilon}}^{(r)})$$

where $Bel(\theta) = MD(K^{(r)})(\theta^{(r)})$, then since MD satisfies renaming, for every atom ζ [2] of $\mathfrak{L}^r$ such that $\zeta \models \phi^{(r)}_{i,\vec{\epsilon}}$

$$Bel(\zeta) = MD(K^{(r)})(\zeta) = \frac{Bel(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}}$$

Notice that each atom ζ implies precisely one of the sentences $\phi^{(r)}_{i,\vec{\epsilon}}$, and so for each $(i, \vec{\epsilon})$ there are precisely $\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}$ many atoms ζ.

For $\vec{x} = \langle Bel(\zeta_1), Bel(\zeta_2), \ldots, Bel(\zeta_{J^r}) \rangle$

$$D_r(\vec{x}) = \sum_{i=1}^{J^r} x_i^2 = \sum_{i=1}^{J^r} (Bel(\zeta_i))^2 =$$

$$\sum_{i=1}^{J^r} \left(\frac{Bel(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}} \right)^2$$

$$= \sum_{i,\vec{\epsilon}} \left(\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k} \right) \left(\frac{Bel(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}} \right)^2$$

$$= \sum_{i,\vec{\epsilon}} \frac{x^2_{i,\vec{\epsilon}}}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}}.$$

[2] It is important to notice the difference between the notion of *atom* in propositional and first order languages. For a propositional language $L = \{ p_1, \ldots, p_n \}$ atoms are the set of sentences α_i of the form

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

while for a first order language with only unary predicates $P_1, \ldots, P_n$, atoms are referred to *formulas* $Q_i(x)$, of the form

$$\bigwedge_{i=1}^{n} P(x)^{\epsilon_i}.$$

For a first order language $L^{(k)}$ that is the same as L with only constant symbols $a_1, \ldots, a_k$ the set of sentences ζ_i of the form $\bigwedge_{i=1}^{k} Q_{m_i}(a_i)$ are called *state descriptions* of $L^{(k)}$. As discussed, language $L^{(k)}$ can be considered as a propositional language for which these ζ_i's will be atoms. Thus it is important to notice that when $L^{(k)}$ is considered as a first order language these will be called state descriptions and when it is considered as a propositional language they will be called atoms.

As pointed out in Chapter 1, page 11, a probability function Bel on $L^{(r)}$ can be identified with the vector

$$\vec{x} = \langle Bel(\zeta_1), Bel(\zeta_2), \ldots, Bel(\zeta_{J^r}) \rangle$$

where ζ_i's are the state descriptions of $L^{(r)}$.[3] The justification for this identification is that any sentence of the language can be written as a disjunction of a subset of these state descriptions and the state descriptions are mutually inconsistent. By Lemma 4 and the discussion above the same holds for the sentences $\phi_{i,\vec{\epsilon}}$. Thus the same argument allows us to identify a probability function Bel on $L^{(r)}$ with the vector

$$\vec{x} = \langle Bel(\phi_{i,\vec{\epsilon}}) \rangle.$$

The advantage of identifying Bel with the vector $\vec{x} = \langle Bel(\phi_{i,\vec{\epsilon}}) \rangle$ rather than the vector $\vec{x} = \langle Bel(\zeta_1), Bel(\zeta_2), \ldots, Bel(\zeta_{J^r}) \rangle$ is that the number of state descriptions of $L^{(r)}$ depends on r and as we move from $L^{(r)}$ to $L^{(r+1)}$ the number of state descriptions will change from J^r to J^{r+1} and thus the vectors identifying Bel on $L^{(r)}$ and $L^{(r+1)}$ will be of different dimensions. However the number of sentences $\phi^{(r)}_{i,\vec{\epsilon}}$ is independent of r. Notice that as we move from $L^{(r)}$ to $L^{(r+1)}$, what changes is the number of state descriptions that satisfy each $\phi^{(r)}_{i,\vec{\epsilon}}$ but the number of these sentences remains the same for all r eventually, and thus the probability function Bel on $L^{(r)}$ and $L^{(r+1)}$ can be identified with vectors of the same size. This is the main motivation for defining and using these $\phi_{i,\vec{\epsilon}}$'s rather than the actual state descriptions. Thus for the rest of this chapter the vectors $\vec{x}$ identifying probability functions are constructed on the basis of the sentences $\phi_{i,\vec{\epsilon}}$ rather than the actual state descriptions.

For this $\theta(a_1, \ldots, a_k)$ let $c_1 < c_2 < \ldots < c_s$ be the values for $p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}}$ which occur. Without loss of generality we can assume that $\phi^{(r)}_{i,\vec{\epsilon}}$ are such that the first n_1 coordinates of $\vec{x^{(r)}}$ have $p^{\phi_{i,\vec{\epsilon}}}_{\epsilon} = c_1$ and the next n_2 coordinates have $p^{\phi_{i,\vec{\epsilon}}}_{\epsilon} = c_2$ and so on for $\vec{x^{(r)}} =< Bel(\phi_{i,\vec{\epsilon}}) >$. This is true because the order in which we consider the sentences $\phi_{i,\vec{\epsilon}}$ is not important. Using the same pattern as on page 11 let K be of the form

$$\sum_{j=1}^{n} a_{ij} w(\theta_j) = b_i \quad i = 1, \ldots, m,$$

[3] or atoms of $L^{(r)}$ if we consider $L^{(r)}$ as a propositional language.

where each θ_j is a sentence of $L^{(k)}$. Replacing each $w(\theta_j)$ in K with $\sum_{\phi_{i,\vec{\epsilon}} \models \theta_j} w(\phi_{i,\vec{\epsilon}})$ (possible by Lemma 4) and adding the equation $\sum_{\phi_{i,\vec{\epsilon}}} w(\phi_{i,\vec{\epsilon}}) = 1$ we will get a system of linear equations

$$< w(\phi_{i,\vec{\epsilon}}) > A_K = \vec{b}_K.$$

Thus if the probability function w satisfies K the vector $< w(\phi_{i,\vec{\epsilon}}) >$ will be a solution for the equation

$$\vec{x} A_K = \vec{b}_K.$$

Let

$$S = \{ \vec{x} \mid \vec{x} A_K = \vec{b_K} \}$$

be the set of solutions for this equation i.e. S is the set of probability functions that satisfy K, and define inductively,

$$T_1 = \{ \vec{x} \in S \mid \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = c_1} x_{i,\vec{\epsilon}}^2 \ \ is \ \ minimal \}$$

and

$$T_j = \{ \vec{x} \in T_{j-1} \mid \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = c_j} x_{i,\vec{\epsilon}}^2 \ \ is \ \ minimal \}.$$

If f^i are projection functions onto the first $n_1 + \ldots + n_i$ coordinates, then the sets $S^i = \{ f^i(\vec{x}) \mid \vec{x} \in S \}$ are convex and closed so for every two points $\vec{u}, \vec{v} \in T_i$, $f^i(\vec{u}) = f^i(\vec{v})$. This means that all the points in T_i have the same first $n_1 + n_2 + \ldots + n_i$ coordinates and thus the T_s will contain a single point, $\vec{X}$. Now let

$$x_{i,\vec{\epsilon}}^{(r)} = MD(K^{(r)})(\phi_{i,\vec{\epsilon}}^{(r)})$$

So $\vec{x}^{(r)}$ is a point in S such that $D_r(\vec{x})$ is minimal and so;

$$D_r(\vec{x}^{(r)}) \leq D_r(\vec{X}). \tag{2.2}$$

We will show that $\lim_{r\to\infty} \vec{x}^{(r)} = \vec{X}$.

First we will show that

$$\lim_{r\to\infty} f^1(\vec{x}^{(r)}) = f^1(\vec{X}).$$

Then we will show the same thing for the $f^2(\vec{x^{(r)}})$ and so on.

$\vec{x}^{(r)}$ is a bounded sequence and so has a convergent subsequence (for the simplicity of notation we will assume that it is actually the whole sequence), say convergent to $\vec{Y}$. We should show that

$$f^1(\vec{Y}) = f^1(\vec{X})$$

To see this notice that by (2.2) we have

$$c_1^{r-k} D_r(\vec{x^{(r)}}) \leq c_1^{r-k} D_r(X)$$

so

$$\sum_{i,\vec{\epsilon}} \frac{x_{i,\vec{\epsilon}}^{(r)2} c_1^{r-k}}{\sum_{j=0}^{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}} (-1)^j \binom{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}}{j} (p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - j)^{r-k}} \leq \sum_{i,\vec{\epsilon}} \frac{X_{i,\vec{\epsilon}}^2 c_1^{r-k}}{\sum_{j=0}^{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}} (-1)^j \binom{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}}{j} (p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - j)^{r-k}}$$

If we set $\gamma_i^{(r)} = \sum_{j=0}^{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}} (-1)^j \binom{p_{i,\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}}{j} (p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - j)^{r-k}$ we can rewrite this as:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} \frac{x_{i,\vec{\epsilon}}^{(r)2} c_1^{r-k}}{\gamma_i^{(r)}} + \sum_{i,\vec{\epsilon},c_1<p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}} \frac{x_{i,\vec{\epsilon}}^{(r)2} c_1^{r-k}}{\gamma_i^{(r)}} \leq \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} \frac{X_{i,\vec{\epsilon}}^2 c_1^{(r-k)}}{\gamma_i^{(r)}} + \sum_{i,\vec{\epsilon},c_1<p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}} \frac{X_{i,\vec{\epsilon}}^2 c_1^{r-k}}{\gamma_i^{(r)}}. \quad (2.3)$$

Since

$$(1 - \frac{j}{p_{\epsilon}^{\phi_{i,\vec{\epsilon}}}})^{(r-k)} \to 0 \ \ as \ \ r \to \infty$$

for $0 < j < p_{\epsilon}^{\phi_{i,\vec{\epsilon}}}$, so

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

and we have,

$$\frac{\gamma_i^{(r)}}{p_{\epsilon}^{\phi_{i,\vec{\epsilon}}\, r-k}} \to 1$$

using this for $c_1 < p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}$ we have:

$$\sum_{i,\vec{\epsilon},c_1<p_{\epsilon}^{\phi_{i,\vec{\epsilon}}}} \frac{x_{i,\vec{\epsilon}}^{(r)2} c_1^{r-k}}{\gamma_i^{(r)}} \to 0$$

and

$$\sum_{i,\vec{\epsilon},c_1<p_\epsilon^{\phi_{i,\vec{\epsilon}}}} \frac{X_{i,\vec{\epsilon}}^2 c_1^{r-k}}{\gamma_i^{(r)}} \to 0$$

From (2.3) we have:

$$\sum_{p_\epsilon^{\phi_{i,\vec{\epsilon}}}=c_1} Y_{i,\vec{\epsilon}}^2 \leq \sum_{p_\epsilon^{\phi_{i,\vec{\epsilon}}}=c_1} X_{i,\vec{\epsilon}}^2 \tag{2.4}$$

Notice that $\vec{Y} = \lim_{r\to\infty} \vec{x}^{(r)}$ and so $\vec{Y} \in S$ and as $\vec{X} \in T_1$ by 2.4 we should have $\vec{Y} \in T_1$, then by what was explained above, $f^1(\vec{X}) = f^1(\vec{Y})$ as required.

We will need the following lemma for the rest of the proof:

Lemma 7 *Let $B \subseteq R^m$ be a convex polyhedron with corners $\vec{a_1}, \vec{a_2}, \ldots, \vec{a_q}$. Let $\vec{c} \in B$ and let $f : R^m \to R^n$ be the projection function given by*

$$f\langle x_1, x_2, \ldots, x_m\rangle = \langle x_1, x_2, \ldots, x_n\rangle$$

Now suppose that $\vec{y_j} \in R^n$ for $j \in \mathbb{N}$ are such that $f^{-1}(\vec{y_j}) \cap B \neq \emptyset$ for all j and

$$lim_{j\to\infty}\vec{y_j} = f(\vec{c}).$$

Then there is a sequence $\vec{z_j} \in B$ converging to $\vec{c}$ such that $f(\vec{z_j})$ form a subsequence of the $\vec{y_j}$.

Proof. Any point in B can be written as a linear combination

$$\vec{c} + \sum_{i=1}^{q} \lambda_i \vec{e_i}$$

where $\vec{e_i} = \vec{a_i} - \vec{c}$ and the $\lambda_i \geq 0$ with sum ≤ 1 and so any $\vec{x} \in f(B)$ can be written as

$$f(\vec{c}) + \sum_{i=1}^{q} \lambda_i f(\vec{e_i})$$

with $\lambda_i > 0$ with sum at most 1 where we drop any $f(\vec{e_i})$ which are $\vec{0}$.

Now for each $\vec{y_j}$ pick one such presentation:

$$\vec{y}_j = f(\vec{c}) + \sum_{i=1}^{q} \lambda_{ij} f(\vec{e}_i)$$

with:

$$\vec{z}_j = \vec{c} + \sum_{i=1}^{q} \lambda_{ij} \vec{e}_i \in B$$

Notice that this is possible since $f^{-1}(\vec{y}_j) \cap B \neq \emptyset$.
It is obvious that $f(\vec{z_j})$ is a subsequence of $\vec{y_j}$. To show

$$\lim_{j\to\infty} \vec{z_j} = \vec{c}$$

it is enough to show that

$$\lim_{j\to\infty} \sum_{i=1}^{q} \lambda_{ij} \vec{e}_i = \vec{0}$$

To show this we will show that $\lim_{j\to\infty} \lambda_{ij} = 0$.
We know that $lim_{j\to\infty} \vec{y_j} = f(c)$ and so we have $lim_{j\to\infty} \sum_{i=1}^{q} \lambda_{ij} f(\vec{e}_i) = \vec{0}$

Let

$$\vec{t_j} = \sum_{i=1}^{q} \lambda_{ij} f(\vec{e}_i)$$

We have

$$\lim_{j\to\infty} \vec{t_j} = \vec{0}$$

and $\vec{t_j}$ is in the convex polyhedron with corners $f(\vec{e}_i)$. For each $\vec{t_j}$ pick a smallest set $f(\vec{e_{i_1}}), \ldots, f(\vec{e_{i_h}})$ such that:

$$\vec{t_j} = \sum_{k=1}^{h} \lambda_{i_k j} f(\vec{e_{i_k}}) \tag{2.5}$$

with $\lambda_{i_k j} > 0$ and $\sum_{i_k} \lambda_{i_k j} \leq 1$. By taking a subsequence if necessary we can assume that $\vec{z_j}$ all have the same smallest set and that $\lambda_{i_k j} \to \lambda_{i_k}$ as $j \to \infty$. For simplicity of notation we can assume that these smallest sets are all the $\vec{e_i}$, so (2.5) will become:

$$\vec{t_j} = \sum_{i=1}^{q} \lambda_{ij} f(\vec{e}_i) \tag{2.6}$$

and

$$\vec{0} = \sum_{i=1}^{q} \lambda_i f(\vec{e_i}). \tag{2.7}$$

Now if all the $\lambda_i = 0$ we have the required result, otherwise suppose some of the $\lambda_i > 0$. Then from (2.6) and (2.7) we will have:

$$\vec{t_j} = \sum_{i=1}^{q} (\lambda_{ij} - \nu\lambda_i) f(\vec{e_i}) \tag{2.8}$$

Now if we increase ν from 0, one of the coefficients will become zero while others are still positive and this contradicts the choice of smallest set. Hence we should have all $\lambda_i = 0$ as required. ■

To continue the proof, consider f^2, the projection function onto the first $n_1 + n_2$ coordinates. We will show that

$$f^2(\vec{Y}) = f^2(\vec{X})$$

Remember that $\vec{x}^{(r)}$ is a bounded sequence and so has a convergent subsequence, converging to $\vec{Y}$. Suppose that $f^2(\vec{Y}) \neq f^2(\vec{X})$. By above discussion we have that

$$\lim_{r\to\infty} f^1(\vec{x}^{(r)}) \to f^1(\vec{Y}) = f^1(\vec{X})$$

So by Lemma 7 there is a sequence, say $\vec{z}^{(r)} \in S$ such that

$$\lim_{r\to\infty} \vec{z}^{(r)} = \vec{X}$$

and $f^1(\vec{z}^{(r)})$ is a subsequence of $f^1(\vec{x}^{(r)})$, for simplicity of notation we will assume that this subsequence is the whole sequence $f^1(\vec{x}^{(r)})$.

We will show that for large enough r, we have $D_r(\vec{z}^{(r)}) < D_r(\vec{x}^{(r)})$ which is a contradiction because as $\vec{x}^{(r)}$ is defined to be $MD(\phi^{(r)}_{i,\vec{e}})$, we should have $D_r(\vec{x}^{(r)}) \leq D_r(\vec{z}^{(r)})$.

To show that $D_r(\vec{z}^{(r)}) < D_r(\vec{x}^{(r)})$ we will show that $c_2^{r-k} D_r(\vec{z}^{(r)}) < c_2^{r-k} D_r(\vec{x}^{(r)})$

We can rewrite this as

$$\sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_1}} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_2}} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}>c_2}} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} <$$

$$\sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_1}} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_2}} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}>c_2}} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}}$$

By what is said above we have that:

$$c_2^{r-k} \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_1}} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} = c_2^{r-k} \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_1}} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} \tag{2.9}$$

and as $r \to \infty$

$$\sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}>c_2}} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{2}}{\gamma_i^{(r)}} \to 0 \tag{2.10}$$

$$\sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}>c_2}} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{2}}{\gamma_i^{(r)}} \to 0 \tag{2.11}$$

and so it is enough to show that r can be large enough that,

$$\sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_2}} z_{i,\vec{\epsilon}}^{(r)2} < \sum_{\substack{\phi_{i,\vec{\epsilon}} \\ p_{\vec{\epsilon}}=c_2}} x_{i,\vec{\epsilon}}^{(r)2} - \delta' \tag{2.12}$$

for some fixed $\delta' > 0$.

As $r \to \infty$ we have:

$$\vec{z}^{(r)} \to \vec{X}$$

so

$$\vec{z}_{i,\vec{\epsilon}}^{(r)} \to \vec{X}_{i,\vec{\epsilon}}$$

so

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{z}_{i,\vec{\epsilon}}^{(r)2} \to \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{X}_{i,\vec{\epsilon}}^{2} \tag{2.13}$$

the same way we have

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{x}_{i,\vec{\epsilon}}^{(r)2} \to \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{Y}_{i,\vec{\epsilon}}^{2} \tag{2.14}$$

by definition of $\vec{X}$ and our assumption that $f^2(\vec{X}) \neq f^2(\vec{Y})$,

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{X}_{i,\vec{\epsilon}}^{2} < \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{Y}_{i,\vec{\epsilon}}^{2}$$

Let

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{Y}_{i,\vec{\epsilon}}^{2} - \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{X}_{i,\vec{\epsilon}}^{2} = \delta$$

Using (2.13) and (2.14) we can choose r large enough so that

$$\left| \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{X}_{i,\vec{\epsilon}}^{2} - \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{z}_{i,\vec{\epsilon}}^{(r)2} \right| \leq \delta/3$$

and

$$\left| \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{Y}_{i,\vec{\epsilon}}^{2} - \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{x}_{i,\vec{\epsilon}}^{(r)2} \right| \leq \delta/3$$

so we have

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{x}_{i,\vec{\epsilon}}^{(r)2} - \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{z}_{i,\vec{\epsilon}}^{(r)2} \geq \delta/3$$

and so

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{z}_{i,\vec{\epsilon}}^{(r)2} < \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \vec{x}_{i,\vec{\epsilon}}^{(r)2} - \delta/6$$

So by (2.9) we have:

$$c_2^{r-k} \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + c_2^{r-k} \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} < c_2^{r-k} \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + c_2^{r-k} \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} - \delta/6$$

by (2.10) and (2.11) we can choose r large enough such that,

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}>c_2} c_2^{r-k} \frac{z_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} <$$

$$\sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_1} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=c_2} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}} + \sum_{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}>c_2} c_2^{r-k} \frac{x_{i,\vec{\epsilon}}^{(r)2}}{\gamma_i^{(r)}}$$

This gives us the required contradiction and so we must have

$$f^2(\vec{Y}) = f^2(\vec{X}).$$

For the next step we repeat the same process with f^3, the projection function onto the first $n_1 + n_2 + n_3$ coordinates, repeating this process $s - 1$ times (s is the number of different $p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}$s) we will have

$$\lim_{r\to\infty} \vec{x}^{(r)} = \vec{Y} = \vec{X}.$$

Notice that $\vec{x^{(r)}}$ are all in S and S is closed so we have

$$\vec{X} = \lim_{r\to\infty} \vec{x^{(r)}} \in S$$

and thus $\vec{X}$ is a probability function and satisfies K as required.

■

2.2 The Centre of Mass Inference Process

The Centre of Mass inference process on a propositional language L, CM^L, is defined so as to choose the most typical (for a certain notion of typicality) probability function. Here for a knowledge base K the probability function is chosen to be as average and representative as possible among all the probability functions satisfying K. More precisely the centre of mass inference process is defined to choose the centre of mass of $V^L(K)$, the set of solutions for K. Notice that since $V^L(K)$ is a convex set its centre of mass will be in $V^L(K)$, which will not necessarily be the case if we generalise K to contain non linear constraints.

The justifications for CM^L fall into the first category mentioned before. This inference process can to some extent be justified by the principle of indifference (or Laplace's principle). The main idea is that when dealing with a set of facts all the possible words that are consistent with these facts should be regarded as equally likely. In our terminology then, all the probability functions satisfying K or, equivalently, all the points in $V^L(K)$ should be regarded as equally likely. Thus choosing the centre of mass of $V^L(K)$ will correspond to choosing the average or the most representative point in $V^L(K)$ and there appear to be situations where choosing the probability function through this approach provides a more plausible and better justified answer when compared to inference processes based on minimising information [see [23]].

Although CM^L seems very intuitive, it suffers from some important shortcomings one of which is the failure of language invariance. Extending the language by adding new propositional variables may change the probability given to a sentence θ on the basis of K, even if the new propositional variables do not explicitly appear in θ or K. To correct this shortcoming we will define the *limiting centre of mass inference process,* CM_∞.

The following three theorems are stated and proved in [24].

Theorem 8 $\lim_{\substack{L\subset L' \\ |L'|\to\infty}} CM^{L'}(K)(\theta)$ *exists and is equal to* $CM_\infty(K)(\theta)$ *where* $CM_\infty(K)(\theta)$ *is an inference process that choose the probability function or equivalently the point in* $V^L(K)$ *where* $\sum_{i\notin I^L(K)} \log x_i$ *is maximal, where*

$$I^L(K) = \{\, i \mid \forall \vec{y} \in V^L(K), y_i = 0 \,\}.$$

Theorem 9 *CM_∞ is language invariant.*

Theorem 10 *The limiting Centre of Mass inference process, CM_∞ satisfies the principles of equivalence, continuity, open-mindedness, renaming and obstinacy but not the principles of irrelevant information nor the principle of relativisation.*

By the above discussion, the limiting centre of mass inference process CM_∞ will be defined to be the unique point in $V^L(K)$ where $\sum_{i \notin I^L(K)} \log x_i$ is maximal. Here we will use the same pattern as the previous section to generalize this inference process to the case of unary predicate languages.

Theorem 11 *For $\theta \in SL$:*

$$Bel(\theta) = \lim_{r \to \infty} CM_\infty(K^{(r)})(\theta^{(r)})$$

exists and is a probability function on L that satisfies K.

Proof. Using the same construction as for Theorem 6, if we set

$$x_{i,\vec{\epsilon}} = Bel(\phi^{(r)}_{i,\vec{\epsilon}})$$

the function $CM_\infty(\vec{x})$ for $\vec{x} =< Bel(\zeta_1), \ldots, Bel(\zeta_{J^r}) >$ will be given by[4]:

$$CM_\infty(\vec{x}) = \sum_{i=1}^{J^r} \log x_i = \sum_{i=1}^{J^r} \log Bel(\zeta_i)$$

$$= \sum_{i=1}^{J^r} \log \left(\frac{Bel(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}} \right)$$

as CM_∞ satisfies renaming,

$$= \sum_{i,\vec{\epsilon}} \sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k} \log \left(\frac{x_{i,\vec{\epsilon}}}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k}} \right) =$$

$$\sum_{i,\vec{\epsilon}} \sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} - j)^{r-k} \log x_{i,\vec{\epsilon}} -$$

[4]For simplicity of notation we will assume that $I^L(K) = \emptyset$.

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

Now let

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

and

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

So we will have :

$$CM_\infty(\vec{x}) = \sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log x_{i,\vec{\epsilon}} - (r-k) \sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - \delta_r$$

We know that as $r \to \infty$, $\frac{\gamma_{i,\vec{\epsilon}}^{(r)}}{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}\, r-k}} \to 1$ and $\delta_r \to 0$. Set

$$S = \{ \vec{x} \mid \vec{x} A_K = \vec{b_K} \}$$

$$T_1 = \{ \vec{x} \in S \mid \sum_{i,\vec{\epsilon}, p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = b_1} \log x_{i,\vec{\epsilon}} \quad is \ \ maximal \}$$

$$T_j = \{ \vec{x} \in T_{j-1} \mid \sum_{i,\vec{\epsilon}, p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = b_j} \log x_{i,\vec{\epsilon}} \quad is \ \ maximal \}$$

where $b_1 > b_2 > \ldots > b_s$ is the set of $p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}$ arranged in a decreasing order. With the same argument as for the Theorem 6, for two points $\vec{u}, \vec{v} \in T_i$, $f^i(\vec{u}) = f^i(\vec{v})$ and T_s will be a single point, $\vec{X}$. Let $x_{i,\vec{\epsilon}}^{(r)} = CM_\infty(\phi_{i,\vec{\epsilon}}^{(r)})$. We want to prove that

$$\lim_{r \to \infty} \vec{x}^{(r)} = \vec{X}.$$

Equivalently we want to show that

$$lim_{r \to \infty} f^i(\vec{x}^{(r)}) = f^i(\vec{X})$$

for $i = 1, \ldots, s$.

We can assume that the first n_1 coordinates of $\vec{x}^{(r)}$ have $p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = b_1$, the next n_2 coordinates have $p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = b_2$ and so on. By definition of $\vec{x}^{(r)}$ we will have :

$$CM_\infty(\vec{x}^{(r)}) \geq CM_\infty(\vec{X})$$

$$\sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log x_{i,\vec{\epsilon}}^{(r)} - (r-k) \sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - \delta_r \geq \sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log X_{i,\vec{\epsilon}} - (r-k) \sum_{i,\vec{\epsilon}} \gamma_{i,\vec{\epsilon}}^{(r)} \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} - \delta_r. \tag{2.15}$$

First we will show that

$$\lim_{r\to\infty} f^1(\vec{x}^{(r)}) = f^1(\vec{X})$$

To see this, suppose this is not the case. $\vec{x}^{(r)}$ is a bounded sequence and so has a convergent subsequence, say converging to $\vec{Y}$, where $f^1(\vec{Y}) \neq f^1(\vec{X})$.

Using (2.15) we have:

$$\sum_{i,\vec{\epsilon}} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log x_{i,\vec{\epsilon}}^{(r)}}{b_1^{r-k}} \geq \sum_{i,\vec{\epsilon}} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{X}_{i,\vec{\epsilon}}}{b_1^{r-k}}$$

which we can rewrite as:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_1^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_1^{r-k}} \geq \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{X}_{i,\vec{\epsilon}}}{b_1^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{X}_{i,\vec{\epsilon}}}{b_1^{r-k}} \tag{2.16}$$

As $r \to \infty$ we will have

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{Y}_{i,\vec{\epsilon}}}{b_1^{r-k}} \to 0$$

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_1} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{X}_{i,\vec{\epsilon}}}{b_1^{r-k}} \to 0$$

and using this and (2.16) we have:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_1} \log \vec{Y}_{i,\vec{\epsilon}} \geq \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_1} \log \vec{X}_{i,\vec{\epsilon}} \tag{2.17}$$

which is a contradiction as $\vec{X} \in T_1$. Notice that by (2.17) and the fact that $\vec{Y} = \lim_{r\to\infty} \vec{x}^{(r)}$ and so $\vec{Y} \in S$, we should have that $\vec{Y} \in T_1$, then as noted above we should have

$$f^1(\vec{X}) = f^1(\vec{Y}).$$

And this gives the required contradiction so

$$\lim_{r\to\infty} f^1(\vec{x}^{(r)}) = f^1(\vec{X}). \tag{2.18}$$

Next step is to show that

$$\lim_{r\to\infty} f^2(\vec{x}^{(r)}) = f^2(\vec{X})$$

Again suppose that this is not the case, that is $f^2(\vec{Y}) \neq f^2(\vec{X})$. By Lemma 7 we know that there is a sequence $z^{(r)} \in S$ converging to $\vec{X}$ and $f^1(\vec{z}^{(r)})$ is a subsequence of $f^1(\vec{x}^{(r)})$ (for simplicity of notation we will assume that this subsequence is the whole sequence $f^1(\vec{x}^{(r)})$). To get the required contradiction we will show that for large enough r

$$CM_\infty(\vec{x}^{(r)}) < CM_\infty(\vec{z}^{(r)})$$

which is a contradiction with the choice of $\vec{x}^{(r)}$.

To show this we will show that for large r

$$\frac{1}{b_2^{r-k}} CM_\infty(\vec{x}^{(r)}) < \frac{1}{b_2^{r-k}} CM_\infty(\vec{z}^{(r)}).$$

We can rewrite this as:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} < \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \geq b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} \tag{2.19}$$

As $r \to \infty$ we have:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} \to 0$$

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} < b_2} \frac{\gamma_{i,\vec{\epsilon}}^{(r)} \log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} \to 0$$

so it is enough to show that:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}\geq b_2} \frac{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}\,r-k}\log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} < \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}\geq b_2} \frac{p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}\,r-k}\log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} - \delta' \tag{2.20}$$

for some fixed $\delta' > 0$.

Which is:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_1} \frac{b_1^{r-k}\log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)} < \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_1} \frac{b_1^{r-k}\log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} + \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} - \delta'. \tag{2.21}$$

We know that $f^1(\vec{z}^{(r)}) = f^1(\vec{x}^{(r)})$, so we have:

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_1} \frac{b_1^{r-k}\log \vec{x}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}} = \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_1} \frac{b_1^{r-k}\log \vec{z}_{i,\vec{\epsilon}}^{(r)}}{b_2^{r-k}}$$

and so it is enough to show that

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)} < \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} - \delta' \tag{2.22}$$

but we have

$$\vec{x}^{(r)} \to \vec{Y}$$

and

$$\vec{z}^{(r)} \to \vec{X}$$

so

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)} \to \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{Y}_{i,\vec{\epsilon}}$$

and

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} \to \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{X}_{i,\vec{\epsilon}}$$

and by definition of $\vec{X}$ we have

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{X}_{i,\vec{\epsilon}} - \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{Y}_{i,\vec{\epsilon}} > 0$$

notice that inequality is strict because $f^2(\vec{Y}) \neq f^2(\vec{X})$.

Now if we set

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{X}_{i,\vec{\epsilon}} - \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{Y}_{i,\vec{\epsilon}} = \delta \tag{2.23}$$

and we take r large enough so that:

$$\left| \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)} - \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{Y}_{i,\vec{\epsilon}} \right| < \frac{\delta}{3} \tag{2.24}$$

and

$$\left| \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} - \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{X}_{i,\vec{\epsilon}} \right| < \frac{\delta}{3} \tag{2.25}$$

we will have

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} - \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)} > \frac{\delta}{3}$$

and so

$$\sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{z}_{i,\vec{\epsilon}}^{(r)} - \delta/6 > \sum_{i,\vec{\epsilon},p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}=b_2} \log \vec{x}_{i,\vec{\epsilon}}^{(r)}$$

which is (2.22), as required. Repeating the same process we can show that $f^3(\lim_{r\to\infty} \vec{x}^{(r)}) = f^3(\vec{X})$ and so on to have the required result.
Notice that $\vec{x^{(r)}}$ are all in S and S is closed so we have

$$\vec{X} = \lim_{r\to\infty} \vec{x^{(r)}} \in S$$

and thus $\vec{X}$ is a probability function that satisfies K as required. ■

Here we have proved the required result for the generalization of two specific inference processes, Minimum Distance and Centre of Mass, but in fact analogous proofs would also give the result for the Maximum Entropy Inference Process (already proved in [6]) and the spectrum of other inference processes based on generalized Renyi Entropies.

In our original question we imagined an agent wishing to assign probabilities to all sentences on the basis of quantified knowledge K. A special case of this is when K simply amounts to the assertion that some consistent, finite, set of axioms $\mathcal{T}$ hold categorically, i.e.

$$K = \{ w(\phi) = 1 \mid \phi \in \mathcal{T} \}.$$

In this case our question might be reformulated as

> *Given a finite (consistent) set $\mathcal{T}$ of first order axioms what should we take as the default or most normal model of $\mathcal{T}$? More precisely, if we know only that the structure M with universe $\{ a_i \mid i \in \mathbb{N} \}$ is a model of $\mathcal{T}$ what probability should we give to a sentence $\theta(a_1, a_2, \ldots, a_n)$ being true in M?*

There are various approaches one might take to this question depending on the interpretation of 'most normal'. For example within a model theory context one might consider a *prime model*, where such exists, to be the 'most normal' in the sense of being the smallest and the canonical example (see for example [8, p96], [12, p336]). On the other hand one might feel that if possible the default model should be existentially closed in the sense that any quantifier free formula which could be satisfied in a superstructure model of $\mathcal{T}$ was already satisfied in the default model. Alternatively we might consider arguing via the distribution of models, see for example [1], [2], [13], [14], [15], in order to make the default the 'average' model.

Furthermore, at first sight it would appear that there was already a rather well studied approach to this problem via Inductive Logic. In that subject, see for example [7], [11], [19], [22], this same problem with $\mathcal{T} = \emptyset$ is quite central. So it might seem that a solution to our problem here could be had by simply taking a rationally justified probability function w championed within Inductive Logic for the case of a completely empty knowledge base and then conditioning w on $\bigwedge \mathcal{T}$. The first problem with that approach however is that there is currently no clearly favored rational solution to the Inductive Logic problem. But more seriously, those solutions w which have been proposed generally give non-tautologous universal sentences probability 0, see

for example [16], [20], [21], [22, p22-23], [24, p196-197], and once $w(\bigwedge \mathcal{T}) = 0$ such conditioning will not be possible.[5] [6]

However if we assume that the sentences of $\mathcal{T}$ come from the *purely unary language* of the preceding sections then the method described, based on any of the above inference processes (in fact in this simple case on *any* inference process satisfying the Renaming Principle) with $K = \{ w(\phi) = 1 \mid \phi \in \mathcal{T} \}$, can be applied, and in fact always yield the same answer.

Theorem 12 *Let $K = \{ w(\phi) = 1 \mid \phi \in \mathcal{T} \}$, and let N be an inference process satisfying the Renaming Principle. If $\vec{\epsilon}^1, \ldots, \vec{\epsilon}^s$ are all those vectors $\vec{\epsilon}$ for which $\bigwedge_{j=1}^{J} (\exists x Q_j(x))^{\epsilon_j}$ is consistent with $\mathcal{T}$ and for which $p_{\vec{\epsilon}}^{\phi_{i,\vec{e}}}$ takes its largest possible value, then*

$$N(K)(\theta(a_1, \ldots, a_k)) = |H|/|L|$$

where

$$L = \{ \phi_{i,\vec{e}^r} \mid \phi_{i,\vec{e}^r} \text{ is consistent with } \bigwedge \mathcal{T},\ 1 \leq i \leq J^k,\ 1 \leq r \leq s \},$$

$$H = \{ \phi_{i,\vec{e}^r} \mid \phi_{i,\vec{e}^r} \text{ is consistent with } \theta(a_1, \ldots, a_k) \wedge \bigwedge \mathcal{T},\ 1 \leq i \leq J^k,\ 1 \leq r \leq s \}.$$

Proof. Let

$$L' = \{ \phi_{i,\vec{e}} \mid \phi_{i,\vec{e}} \text{ is consistent with } \bigwedge \mathcal{T} \}$$

$$H' = \{ \phi_{i,\vec{e}} \mid \phi_{i,\vec{e}} \text{ is consistent with } \theta(a_1, \ldots, a_n) \wedge \bigwedge \mathcal{T} \},$$

and let s_i's run through the state descriptions of $L^{(r)}$. Let $S^{(r)}$ be the set of those state descriptions of $L^{(r)}$ that are consistent with K. By the Renaming Principle all the state descriptions in $S^{(r)}$ will get the same probability, namely $\frac{1}{|S^{(r)}|}$ and let $\gamma_{\phi_{i,\vec{e}}}^{(r)}$ be the number of state descriptions of $L^{(r)}$ that logically imply $\phi_{i,\vec{e}}$. By previous discussions

[5] It is true that proposals have been made for solutions to the Inductive Logic problem which give some non-tautologous universal sentences non-zero probability, see for example [9], [16], [20], [21], [25]. However they seem (to us) too ad hoc to be seriously considered 'logical'.

[6] This apparent discontinuity between the cases when $\mathcal{T} =/\neq \emptyset$ is intriguing – the method we shall apply here still works when $\mathcal{T} = \emptyset$ but gives an unsatisfactory solution to the inductive logic problem, unsatisfactory in that it corresponds to the so called completely independent solution which entertains no induction i.e. learning by example, see for example [24, p172].

in the proof of Theorem 6, we know that

$$\lim_{r\to\infty} \frac{\gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{(p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}})^{r-k}} = 1.$$

So

$$N(K^{(r)})(\theta^{((r))}(a_1,\ldots,a_k)) = \sum_{s_i \vDash \theta} N(K^{(r)})(s_i) =$$

$$\sum_{\phi_{i,\vec{\epsilon}} \in H'} \sum_{s_i \vDash \phi_{i,\vec{\epsilon}}} N(K)(s_i) =$$

$$\sum_{\phi_{i,\vec{\epsilon}} \in H'} \frac{\gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{|S^{(r)}|} =$$

$$\frac{\sum_{\phi_{i,\vec{\epsilon}} \in H'} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L'} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}.$$

To see this notice that

$$|S^{(r)}| = \sum_{\phi_{i,\vec{\epsilon}} \in L'} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}$$

as both are the number of state descriptions of $L^{(r)}$ consistent with K.

Let $c_1 > c_2 > \ldots > c_t$ be the distinct values for $p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}}$ for the sentences in L' so we have $p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} = c_1$ for $\phi_{i,\vec{\epsilon}} \in L$ (and thus for $\phi_{i,\vec{\epsilon}} \in H$) and for every $\phi_{i,\vec{\delta}} \notin L$ we have $p^{\phi_{i,\vec{\delta}}}_{\vec{\delta}} \leq c_2$. Thus we will have

$$N(K)(\theta) = \lim_{r\to\infty} N(K^{(r)})(\theta^{(r)}) =$$

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}} \in H'} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L'} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} = \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}} \in H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}} \in H'-H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}} \in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} =$$

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}} \in H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}} \in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} + \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}} \in H'-H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}} \in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}$$

First notice that

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}} \in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}} \in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}} \in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} \geq \lim_{r\to\infty} \frac{c_1^{r-k}|L|}{c_1^{r-k}|L| + c_2^{r-k}|L'-L|} = 1$$

so

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}}\in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} = 1.$$

Thus

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'-H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}}\in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} = \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H'-H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} \leq$$

$$\lim_{r\to\infty} \frac{c_2^{r-k}|H'-H|}{c_1^{r-k}|L|} = 0,$$

since $c_2 < c_1$. In consequence we will have

$$N(K)(\theta) = \lim_{r\to\infty} N(K^{(r)})(\theta^{(r)}) =$$

$$\lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}} + \sum_{\phi_{i,\vec{\epsilon}}\in L'-L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} = \lim_{r\to\infty} \frac{\sum_{\phi_{i,\vec{\epsilon}}\in H} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}}{\sum_{\phi_{i,\vec{\epsilon}}\in L} \gamma^{(r)}_{\phi_{i,\vec{\epsilon}}}} =$$

$$\frac{c_1^{r-k}|H|}{c_1^{r-k}|L|} = \frac{|H|}{|L|}.$$

■

In particular then w gives probability 1 to

$$\bigvee_{i=1}^{s} \bigwedge_{j=1}^{J} (\exists x Q_j(x))^{\epsilon^i_j},$$

(and probability $1/s$ to each of the disjuncts), thus exclusively favoring those models M of $\mathcal{T}$ in which as that as many of the Q_j are satisfied as possible, that is the existentially closed models of $\mathcal{T}$.

We will conclude this chapter with an example where the inference processes CM_∞, MD and ME will provide different answers.

Example. Let L be a language with only unary predicates P and Q. Let $K = \{w(\forall x P(x)) = 1/3, w(\forall x Q(x)) = 1/3\}$ and $\theta = \forall x(P(x) \wedge Q(x))$. We will be interested in the value $w(\theta)$.

Atoms of this language will be formulas,

$$Q_1 : P(x) \wedge Q(x)$$

$$Q_2 : P(x) \wedge \neg Q(x)$$

$$Q_3 : \neg P(x) \wedge Q(x)$$

$$Q_4 : \neg P(x) \wedge \neg Q(x)$$

and the state descriptions of $L^{(r)}$ are sentences, $S = \{\alpha_1, \ldots, \alpha_{4^r}\}$, of the form

$$\alpha_m = \bigwedge_{j=1}^{r} Q_{m_j}(a_j)$$

.

Define

$$A = \{\alpha_m \,|\, \alpha_m \vDash \bigwedge_{j=1}^{r} P(a_j)\},$$

and

$$B = \{\alpha_m \,|\, \alpha_m \vDash \bigwedge_{j=1}^{r} Q(a_j)\},$$

so for each $\alpha_m \in A$, we have

$$\alpha_m \vDash (Q_1(a_i) \vee Q_2(a_i)) \qquad i = 1, \ldots, r.$$

similarly for each $\alpha_m \in B$, we have

$$\alpha_m \vDash (Q_1(a_i) \vee Q_3(a_i)) \qquad i = 1, \ldots, r.$$

Notice the $|A| = |B| = 2^r$. Let $\alpha_1 \in A$, be the state description that logically implies $\bigwedge_{i=1}^{r} Q_1(a_i)$. Notice that $A \cap B = \{\alpha_1\}$ and set

$$A' = A - \{\alpha_1\}. \quad \text{and} \quad B' = B - \{\alpha_1\},$$

so $A' \cap B' = \emptyset$. From constraints in K we have

$$\sum_{\alpha_i \in A} w(\alpha_i) = \sum_{\alpha_j \in B} w(\alpha_j) = \frac{1}{3}.$$

By renaming all the state descriptions in A' should get the same probability, say u, and similarly all the state descriptions in B' will get value v and those in $S - (A \cup B)$, probability t. In this setting and using the vector notation for probability functions (based on state descriptions rather than $\phi_{i,\vec{e}}$), the knowledge base K will be equivalent to the system of equations,

$$x_1 + \sum_{\alpha_i \in A'} x_i = x_1 + (2^r - 1)u = \frac{1}{3},$$

$$x_1 + \sum_{\alpha_i \in B'} x_i = x_1 + (2^r - 1)v = \frac{1}{3},$$

$$x_1 + \sum_{\alpha_i \in A'} x_i + \sum_{\alpha_i \in B'} x_i + \sum_{\alpha_i \in S-(A \cup B)} x_i = x_1 + (2^r - 1)u + (2^r - 1)v + (2^{2r} - 2^{r+1} + 1)t = 1.$$

We can simplify this by setting $y_1 = x_1, y_2 = (2^r - 1)u, y_3 = (2^r - 1)v, y_4 = (2^{2r} - 2^{r+1} + 1)t$ as

$$y_1 + y_2 = \frac{1}{3}, y_1 + y_3 = \frac{1}{3},$$

$$y_1 + y_2 + y_3 + y_4 = 1.$$

Thus the set of solutions for K is in a one to one correspondence withe set

$$\{ < x_1, \ldots, x_4 > \in \mathbb{R}^4 \mid 0 \leq x_i \leq 1, \sum_{j=1}^{4} x_j = 1, x_1 + x_2 = x_1 + x_3 = 1/3 \}.$$

The inference process MD will choose the point $y = < y_1, \ldots, y_4 >$ in this set for which $\sum_{i=1}^{4} y_i^2$ is minimal. Putting this equations into Lagrange multiplier method we will get,

$$y_1 = \frac{1}{12}, \quad y_2 = \frac{3}{12}, \quad y_3 = \frac{3}{12}, \quad y_4 = \frac{5}{12}.$$

and we have

$$MD(K^{(r)})(\alpha_1) = y_1 = \frac{1}{12}, \quad MD(K^{(r)})(\alpha_m) = \frac{y_2}{2^r - 1} = \frac{3}{12(2^r - 1)} \quad for\ \alpha_m \in A' \cup B',$$

$$MD(K^{(r)})(\alpha_m) = \frac{y_4}{(2^{2r} - 2^{r+1} + 1)} = \frac{5}{12(2^{2r} - 2^{r+1} + 1)} \quad for\ \alpha_m \in S - (A \cup B).$$

Thus we have

$$MD(K)(\theta) = \lim_{r \to \infty} MD(K^{(r)})(\theta^{(r)}) =$$

$$MD(K^{(r)})(\alpha_1) = \frac{1}{12}.$$

The inference process CM_∞ will choose the point $< t_1, \ldots, t_4 >$ for which $\sum_{j=1}^{4} \log(t_j)$ is maximal. Again using Lagrange multipliers method we will get

$$t_1 = 0.1301294011, \quad t_2 = t_3 = 0.2032039322, \quad t_4 = 0.463462735.$$

So

$$CM_\infty(K^{(r)})(\alpha_1) = t_1 = 0.1301294011,$$

$$CM_\infty(K^{(r)})(\alpha_m) = \frac{t_2}{2^r - 1} = \frac{0.2032039322}{2^r - 1} \quad for\ \alpha_m \in A' \cup B',$$

$$MD(K^{(r)})(\alpha_m) = \frac{t_4}{(2^{2r} - 2^{r+1} + 1)} = \frac{0.463462735}{2^{2r} - 2^{r+1} + 1} \quad for\ \alpha_m \in S - (A \cup B).$$

Thus we have

$$CM_\infty(K)(\theta) = \lim_{r\to\infty} CM_\infty(K^{(r)})(\theta^{(r)}) =$$

$$CM_\infty(K^{(r)})(\alpha_1) = 0.1301294011.$$

The maximum entropy inference process will choose the point $< z_1, \ldots, z_4 >$ for which $\sum_{j=1}^{4} z_j \log(z_j)$ is minimal. Using Lagrange multiplier methods as before we will get

$$z_1 = \frac{1}{9}, \quad z_2 = z_3 = \frac{2}{9}, \quad z_4 = \frac{4}{9}.$$

$$ME(K^{(r)})(\alpha_1) = z_1 = \frac{1}{9}, \quad ME(K^{(r)})(\alpha_m) = \frac{z_2}{2^r - 1} = \frac{2}{9(2^r - 1)} \quad for\ \alpha_m \in A' \cup B',$$

$$MD(K^{(r)})(\alpha_m) = \frac{y_4}{(2^{2r} - 2^{r+1} + 1)} = \frac{4}{9(2^{2r} - 2^{r+1} + 1)} \quad for\ \alpha_m \in S - (A \cup B).$$

Thus we have

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

$$ME(K^{(r)})(\alpha_1) = \frac{1}{9}$$

and as we can see the three inference processes provide different answers.

Chapter 3

Maximum Entropy Inference Process On Quantified Knowledge

Maximum Entropy (henceforth referred to as ME) is the most well studied and commonly accepted inference process to work with probability logic. Here again, like the case of MD, the choice of probability function emphasises on increasing the uncertainty to its maximum possible level by minimizing the amount of information included in the candidate probability function beyond the knowledge base. However the choice of measure used here for comparing two probability functions will prove to be much better justified than the choice of Euclidean measure in the case of MD. Here the uncertainty or entropy is maximised through maximising the Shannon measure of uncertainty

$$-\sum x_i \log(x_i).$$

This choice, unlike the choice of Euclidean measure for MD, benefits from a very strong mathematical justification. To see this let $H(\vec{x})$ be a measure for uncertainty on the set $\bigcup_{k\geq 1} D_k$ where

$$D_k = \{ < x_1, \ldots, x_k > \in \mathbb{R}^k \mid \vec{x} \geq 0 \, and \, \sum x_i = 1 \},$$

then the following properties will be expected from H as a measure for uncertainty:

1) For each $k > 0$, $H \upharpoonright D_k$ should be continuous.
2) For $0 < n < m$, $H(\frac{1}{n}, \ldots, \frac{1}{n}) < H(\frac{1}{m}, \ldots, \frac{1}{m})$.

3) If $\sum_{j=1}^{m_i} y_{ij} = 1$ and $y_{ij} \geq 0$ for $i = 1, \ldots, k$ and $\vec{x} \in D_k$ then

$$H(x_1 y_{11}, x_1 y_{12}, \ldots, x_1 y_{1m_1}, x_2 y_{21}, \ldots, x_i y_{ij}, \ldots)$$

$$= H(\vec{x}) + \sum_{i=1}^{k} x_i H(y_{i1}, \ldots, y_{im_i}).$$

If we take x_i $i = 1 \ldots k$ to be probabilities assigned to some disjoint events $E_1 \ldots E_k$ then the above requirements can be justified as follow: The first requirement is justified by saying that small changes in $\vec{x} \in D_k$ are expected to result only in small changes in $H(\vec{x})$.

Considering uncertainty as the amount of information we will gain by observing which of the exclusive and exhaustive events $E_1, \ldots E_k$ actually holds, the second requirement will be justified by saying that where $n < m$, the amount of information we gain by learning which of n equally probable events holds is less than that gained by learning which of m equally probable events holds. The third requirement is justified by saying that the information gained by learning the event corresponding to $x_i y_{ij}$ is the same as the information gained by first learning the event corresponding to x_i and then the event corresponding to y_{ij}.

Expecting these requirements will then (see for example [24] for the proof), force the function $H(\vec{x})$ to be of the form

$$H(x_1, \ldots, x_m) = -\sum_{i=1}^{m} c x_i \log(x_i) \quad \text{for some } c > 0.$$

Thus using the Shannon's measure seems to be the best mathematically justified choice for a measure of uncertainty. This remains true from the point of view of principles of uncertain reasoning.

Theorem 13 *ME satisfies continuity and the principles of equivalence, irrelevant information, open-mindedness, renaming, obstinacy and relativisation. Furthermore ME is the only inference process that satisfies all these principles at the same time.*

Proof. See for example [24]. ■

ME has become a well accepted inference process partly because it ensures that the process through which the probability function is being chosen affects as little as possible the final result through extra restrictions and assumptions beyond those already enforced by the knowledge base, and thus can be argued to provide a probability distribution which is based as much as possible on the evidence and the evidence alone rather than the reasoner or the process of analysing the evidence.

A lot of investigations and studies have been made regarding the justifications and properties of ME, especially for the case of propositional logic. This has given this inference process a great strength and has made it into a topic of great interest both for mathematicians, see for example [15], [24], [26], [27], [28], and philosophers some of whom have identified it as the fundamental thesis for Objective Baysianism, see Jaynes, [17], [18], Rosenkrantz, [31] and Williamson, [33].

In this and the next chapter we will present a study of Maximum Entropy for predicate languages. In the rest of this chapter we investigate a generalisation of ME to a predicate language L as the limiting case of ME defined on finite sublanguages of L. Here we introduce an alternative machinery to the one presented for the BP-method in [6] as used in previous chapter. This machinery is developed with the hope of facilitating the generalisation from unary to more general polyadic languages. As we shall see however the BP-method will unfortunately prove not to be applicable to the most general case. In the rest of the chapter we investigate this method for two special cases and in Chapter 4 we will study an alternative generalisation suggested by Williamson in [34], W-method, and will provide a comparison between the two methods.

3.1 The BP-Method For Π_1 Knowledge Bases From A Unary Language With Identity

In this section we will investigate whether the BP-method on inference processes satisfying Renaming Principle will converge for knowledge bases of the form

$$K = \{ w(\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)) = 1 \}$$

where $\Theta(x_1, \ldots, x_q)$ is quantifier free and is coming from a finite unary first order language without function symbols and augmented with equality. We conjecture that this is the case for any finite predicate language L, without function symbols and whose only constant symbols are $a_1, a_2, \ldots$, though our results to date fall short of proving that. Indeed we would add to this the conjecture that the same is also true of the W-method and that they both give the same answer.

In the case when the language L is purely unary and without equality we have seen that convergence does hold for Maximum entropy, MD and CM_∞, and indeed this holds, if we take general linear knowledge bases without any restriction on the quantifier complexity of the sentences involved. We now consider the next simplest case, where L is a finite unary language augmented with equality, =, and show that the BP-method does converge for any inference process that satisfies the Renaming Principle[1]. To make clear what 'equality' means in this context we require that our probability functions give probability 1 to the axioms of equality and probability 0 to $a_i = a_j$ for $i \neq j$.

Fix L to be this language (finite unary first order language with equality and without function symbols) until otherwise indicated and let $\alpha_1(x), \ldots, \alpha_J(x)$ be the atoms of L with equality removed. Let $n \gg q$. Given a state description $\eta(a_1, \ldots, a_n)$ of $L^{(n)}$ (with equality), that is consistent with the axioms of equality, let M_η be the unique structure for L with universe $\{a_1, \ldots, a_n\}$ in which $\eta(a_1, \ldots, a_n)$ is true. Say that $\eta(a_1, \ldots, a_n)$ is of *type* κ, where $\kappa : \{1, \ldots, J\} \to \{0, 1, \ldots, q\}$, if for $1 \leq i \leq J$,

$$\kappa(i) = \min\{ |\{ j \,|\, \eta(a_1, \ldots, a_n) \models \alpha_i(a_j) \}|, q \}.$$

[1]Actually the machinery of the previous chapter can be directly adapted to this case but we will give an alternative proof here in the hope that this approach may eventually allow of wider generalizations.

Lemma 14 *Suppose that $\Theta(x_1, \ldots, x_q)$ is quantifier free and $\eta_1(a_1, \ldots, a_n), \eta_2(a_1, \ldots, a_n)$ are state descriptions with the same type. Then*

$$M_{\eta_1} \models \forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q) \iff M_{\eta_2} \models \forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q).$$

Proof. Suppose $M_{\eta_1} \models \forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q)$ but $M_{\eta_2} \not\models \forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q)$. This means that there are $a_{i_1}, \ldots, a_{i_q}$ such that $M_{\eta_2} \models \neg\Theta(a_{i_1}, \ldots, a_{i_q})$ and suppose that

$$M_{\eta_2} \models \alpha_{i_j}(a_{i_j}).$$

Since $\eta_1(a_1, \ldots, a_n)$ and $\eta_2(a_1, \ldots, a_n)$ are state descriptions with the same type we should have $a_{t_1}, \ldots, a_{t_q}$ such that

$$M_{\eta_1} \models \alpha_{i_j}(a_{t_j}).$$

Thus $M_{\eta_1} \models \neg\Theta(a_{t_1}, \ldots, a_{t_q})$ and so $M_{\eta_1} \not\models \forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q)$ that is a contradiction. The other direction of the proof will be similar. ■

Theorem 15 *Let L be a unary first order language with finitely many predicate symbols and equality whose only constant symbols are $a_1, a_2, \ldots$ and has no function symbols and let K be a knowledge base as above. If N is an inference process defined on finite propositional languages that satisfies the Renaming Principle then*

$$N(K)(\theta) = \lim_{r\to\infty} N(K^{(r)})(\theta^{(r)})$$

exists and is a probability function on L that satisfies K.

Proof. To show Theorem 15 it is enough to show that for a state description $\Delta^{(n)}$ of $L^{(n)}$

$$\lim_{r\to\infty} N(K^{(r)})(\Delta^{(n)})$$

exists and that it is a probability function that satisfies K. Notice that if the limit does exist it is obviously a probability function that satisfies K just as in Theorem 6, because the set of solutions for K is closed and thus the limit of elements of this set will still be in the set.

Notice that $N(K^{(r)})$ is a probability function on $L^{(r)}$ so if $\eta^{(r)}$ run through the state

descriptions of $L^{(r)}$,

$$N(K^{(r)})(\Delta^{(n)}) = \sum_{\eta^{(r)} \models \Delta^{(n)}} N(K^{(r)})(\eta^{(r)}).$$

Since N satisfies Renaming all the state descriptions of $L^{(r)}$ that are consistent with K (in other words are models of $K^{(r)}$) will get the same probability, namely $\frac{1}{|\{\eta^{(r)} \mid \eta^{(r)} \text{is consistent with } K^{(r)}\}|}$.

Thus

$$N(K^{(r)})(\Delta^{(n)}) = \sum_{\eta^{(r)} \models \Delta^{(n)}} N(K^{(r)})(\eta^{(r)}) =$$

$$\frac{|\{\eta^{(r)} \mid {}^{\eta^{(r)} \text{extends} \Delta^{(n)}}_{\eta^{(r)} \text{is consistent with } K^{(r)}}\}|}{|\{\eta^{(r)} \mid \eta^{(r)} \text{is consistent with } K^{(r)}\}|}$$

and it will be enough to show that

$$\lim_{r \to \infty} \frac{|\{\eta^{(r)} \mid {}^{\eta^{(r)} \text{extends} \Delta^{(n)}}_{\eta^{(r)} \text{is consistent with } K^{(r)}}\}|}{|\{\eta^{(r)} \mid \eta^{(r)} \text{is consistent with } K^{(r)}\}|}$$

exists. Let $\eta_1(a_1, \ldots, a_n), \ldots, \eta'_R(a_1, \ldots, a_n)$ be those state descriptions consistent with

$\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)$. Let $\kappa_1, \ldots, \kappa_R$ be the *distinct* types appearing where the ordering has been chosen so that if $\kappa_i(m) \leq \kappa_j(m)$ for all $1 \leq m \leq J$ then $j \leq i$.[2]

Given a state description $\eta(a_1, \ldots, a_n)$ consistent with $\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)$ and of type κ_g let b_{gh} be the number of state descriptions $\eta'(a_1, \ldots, a_n, a_{n+1})$ of type κ_h extending $\eta(a_1, \ldots, a_n)$ and consistent with $\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)$. Notice that provided n is large this number does not depend on n.

These $< b_{gh} >$ form a lower triangular matrix B and if we start from a state description $\eta(a_1, \ldots, a_n)$ of type κ_i the number of state descriptions $\eta'(a_1, \ldots, a_{n+k})$ of type $\kappa_1, \kappa_2, \ldots, \kappa_R$ that extend $\eta(a_1, \ldots, a_n)$ is given as a vector by $(B^k)^T \vec{e}_i$ where $\vec{e}_i$ is the column vector with 1 in i-th place and zero elsewhere.

Now for $\Delta(a_1, \ldots, a_n)$ a state description of type κ_i consistent $\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)$ the number of state descriptions $\eta'(a_1, \ldots, a_{n+k})$ extending it and still consistent with $\forall x_1, \ldots, x_q \, \Theta(x_1, \ldots, x_q)$ is

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

[2]Notice the reverse of the inequalities here.

Similarly the total number of state descriptions $\eta'(a_1, \ldots, a_{n+k})$ consistent with $\forall x_1, \ldots, x_q$ $\Theta(x_1, \ldots, x_q)$ is

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

where N_j is the number of state description of type κ_j.

By renaming $N(K^{(n+k)})$ will give each of these the same probability, namely

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

so that $\Delta(a_1, \ldots, a_n)$ gets value

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

and it will be enough to show that

$$\lim_{k \to \infty} \frac{\langle 1, 1, \ldots, 1 \rangle (B^k)^T \vec{e}_i}{\sum_{j=1}^{R} N_j \langle 1, 1, \ldots, 1 \rangle (B^k)^T \vec{e}_j}$$

exists. since there are finitely many j it will be enough to show that

$$\lim_{k \to \infty} \frac{< 1, 1, \ldots, 1 > (B^k)^T \vec{e}_i}{< 1, 1, \ldots, 1 > (B^k)^T (\vec{e}_i + \vec{e}_h)}$$

exists.

Claim 1

$$\lim_{k \to \infty} \frac{< 1, 1, \ldots, 1 > (B^k)^T \vec{e}_i}{< 1, 1, \ldots, 1 > (B^k)^T (\vec{e}_i + \vec{e}_h)}, \qquad exists$$

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

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

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

$$\lim_{n\to\infty} \frac{\sum_{r_0+\ldots+r_m=n} \prod_{s=0}^{m} b_{t_s t_s}^{r_s}}{\sum_{u_0+\ldots+u_q=n} \prod_{s=0}^{q} b_{g_s g_s}^{u_s}} \tag{3.1}$$

either exists or is ∞.

To show this we will first find a better expression for, say, the numerator. We will consider this in two cases. First assume for simplicity that all the $b_{t_s t_s}$ are different.

Lemma 16

$$\frac{1}{(b_{m+1}-b_0)(b_0-b_1)\ldots(b_0-b_m)} + \frac{1}{(b_{m+1}-b_1)(b_1-b_0)\ldots(b_1-b_m)} + \ldots$$
$$+\frac{1}{(b_{m+1}-b_m)(b_m-b_0)\ldots(b_m-b_{m-1})}$$
$$= \frac{1}{(b_{m+1}-b_0)(b_{m+1}-b_1)\ldots(b_{m+1}-b_m)}$$

Proof. we will show that :

$$\frac{1}{(b_{m+1}-b_0)(b_0-b_1)\ldots(b_0-b_m)} + \frac{1}{(b_{m+1}-b_1)(b_1-b_0)\ldots(b_1-b_m)} + \ldots$$
$$+\frac{1}{(b_{m+1}-b_m)(b_m-b_0)\ldots(b_m-b_{m-1})}$$
$$-\frac{1}{(b_{m+1}-b_0)(b_{m+1}-b_1)\ldots(b_{m+1}-b_m)} = 0$$

First multiply both sides by $(b_{m+1}-b_0)\ldots(b_{m+1}-b_m)$ and we will have :

$$\frac{(b_{m+1}-b_1)\ldots(b_{m+1}-b_m)}{(b_0-b_1)\ldots(b_0-b_m)} + \frac{(b_{m+1}-b_0)(b_{m+1}-b_2)\ldots(b_{m+1}-b_m)}{(b_1-b_0)\ldots(b_1-b_m)} + \ldots$$
$$+\frac{(b_{m+1}-b_0)\ldots(b_{m+1}-b_{m-1})}{(b_m-b_0)\ldots(b_m-b_{m-1})} - 1 = 0$$

We consider this as a polynomial in b_{m+1}. It has degree m but $m+1$ distinct zeros, namely $\{b_0, b_1, \ldots, b_m\}$, so it should be identical with zero. ■

Lemma 17 *Let A be the following expression*

$$\frac{b_{m+1}^{n+1} b_0^{m-1}}{(b_0-b_1)\ldots(b_0-b_m)} + \frac{b_{m+1}^{n+1} b_1^{m-1}}{(b_1-b_0)\ldots(b_1-b_m)} + \ldots + \frac{b_{m+1}^{n+1} b_m^{m-1}}{(b_m-b_0)\ldots(b_m-b_{m-1})} +$$

$$\frac{b_{m+1}^{n+2} b_0^{m-2}}{(b_0-b_1)\ldots(b_0-b_m)} + \frac{b_{m+1}^{n+2} b_1^{m-2}}{(b_1-b_0)\ldots(b_1-b_m)} + \ldots + \frac{b_{m+1}^{n+2} b_m^{m-2}}{(b_m-b_0)\ldots(b_m-b_{m-1})} +$$

$$\vdots$$

$$\frac{b_{m+1}^{n+m} b_0^{0}}{(b_0-b_1)\ldots(b_0-b_m)} + \frac{b_{m+1}^{n+m} b_1^{0}}{(b_1-b_0)\ldots(b_1-b_m)} + \ldots + \frac{b_{m+1}^{n+m} b_m^{0}}{(b_m-b_0)\ldots(b_m-b_{m-1})}.$$

Then $A = 0$.

Proof.

$$A = b_{m+1}^{n+1}\left[\frac{b_{m+1}^{0} b_0^{m-1} + \ldots + b_{m+1}^{m-1} b_0^{0}}{(b_0-b_1)\ldots(b_0-b_m)} + \ldots + \frac{b_{m+1}^{0} b_m^{m-1} + \ldots + b_{m+1}^{m-1} b_m^{0}}{(b_m-b_0)\ldots(b_m-b_{m-1})}\right] =$$

$$b_{m+1}^{n+1}\left[\frac{b_{m+1}^{m} - b_0^{m}}{(b_0-b_1)\ldots(b_0-b_m)(b_{m+1}-b_0)} + \ldots + \frac{b_{m+1}^{m} - b_m^{m}}{(b_m-b_0)\ldots(b_m-b_{m-1})(b_{m+1}-b_m)}\right] =$$

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

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

It will be enough to show that

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

$$\frac{b_0^{m}}{(b_0-b_1)\ldots(b_0-b_m)(b_0-b_{m+1})} + \ldots + \frac{b_m^{m}}{(b_m-b_0)\ldots(b_m-b_{m-1})(b_m-b_{m+1})} = 0$$

By Lemma 16 it will be enough to show that:

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

$$\dots+\frac{b_m^m}{(b_m-b_0)\dots(b_m-b_{m-1})(b_m-b_{m+1})}=0$$

To see this multiply both sides by $(b_{m+1}-b_0)\dots(b_{m+1}-b_m)$ and we will have

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

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

Claim 2

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

Proof.

Proof by induction on m;

For the base case, where $m=0$ we have

$$b_{t_0t_0}^n=b_{t_0t_o}^n$$

which is clearly true.

Suppose the result is true for m and we will prove it for $m+1$ to simplify the notation we will show $b_{t_st_s}$ by b_s etc.:

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

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

by induction hypothesis. Expanding the rightmost expression we will have:

$$\frac{b_{m+1}^0 b_0^{n+m}}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^0 b_1^{n+m}}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^0 b_m^{n+m}}{(b_m-b_0)\dots(b_m-b_{m-1})}+$$

$$\frac{b_{m+1}^1 b_0^{n+m-1}}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^1 b_1^{n+m-1}}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^1 b_m^{n+m-1}}{(b_m-b_0)\dots(b_m-b_{m-1})}+$$

$$\vdots$$

$$\frac{b_{m+1}^n b_0^m}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^n b_1^m}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^n b_m^m}{(b_m-b_0)\dots(b_m-b_{m-1})}$$

to the above expression we will add the following expression (the expression A that is equal to zero by Lemma 17):

$$\frac{b_{m+1}^{n+1} b_0^{m-1}}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^{n+1} b_1^{m-1}}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^{n+1} b_m^{m-1}}{(b_m-b_0)\dots(b_m-b_{m-1})}+$$

$$\frac{b_{m+1}^{n+2} b_0^{m-2}}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^{n+2} b_1^{m-2}}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^{n+2} b_m^{m-2}}{(b_m-b_0)\dots(b_m-b_{m-1})}+$$

$$\vdots$$

$$\frac{b_{m+1}^{n+m} b_0^0}{(b_0-b_1)\dots(b_0-b_m)}+\frac{b_{m+1}^{n+m} b_1^0}{(b_1-b_0)\dots(b_1-b_m)}+\dots+\frac{b_{m+1}^{n+m} b_m^0}{(b_m-b_0)\dots(b_m-b_{m-1})}$$

So we will have

$$\sum_{r_0+\dots+r_m+r_{m+1}=n}\prod_{s=0}^{m+1} b_s^{r_s}=\frac{b_0^{n+m+1}-b_{m+1}^{n+m+1}}{(b_0-b_1)\dots(b_0-b_m)(b_0-b_{m+1})}+\frac{b_1^{n+m+1}-b_{m+1}^{n+m+1}}{(b_1-b_0)\dots(b_1-b_m)(b_1-b_{m+1})}$$

$$+\dots+\frac{b_m^{n+m+1}-b_{m+1}^{n+m+1}}{(b_m-b_0)\dots(b_m-b_{m-1})(b_m-b_{m+1})}$$

$$=\frac{b_0^{n+m+1}}{(b_0-b_1)\dots(b_0-b_m)(b_0-b_{m+1})}+\dots+\frac{b_m^{n+m+1}}{(b_m-b_0)\dots(b_m-b_{m-1})(b_m-b_{m+1})}$$

$$-\frac{b_{m+1}^{n+m+1}}{(b_0-b_1)\dots(b_0-b_m)(b_0-b_{m+1})}-\dots-\frac{b_{m+1}^{n+m+1}}{(b_m-b_0)\dots(b_m-b_{m-1})(b_m-b_{m+1})} \quad (3.2)$$

Using Lemma 16 and (3.2) we will have:

$$\sum_{r_0+\dots+r_m+r_{m+1}=n}\prod_{s=0}^{m+1}b_s^{r_s}=\frac{b_0^{n+m+1}}{(b_0-b_1)\dots(b_0-b_m)(b_0-b_{m+1})}$$

$$+\dots+\frac{b_m^{n+m+1}}{(b_m-b_0)\dots(b_m-b_{m-1})(b_m-b_{m+1})}$$

$$+\frac{b_{m+1}^{n+m+1}}{(b_{m+1}-b_0)(b_{m+1}-b_1)\dots(b_{m+1}-b_m)}=\sum_{s=0}^{m+1}b_s^{n+m+1}\prod_{y\neq s}(b_s-b_y)^{-1}$$

and this completes the proof of Claim 2. ■

Now using Claim 2 for the case when all $b_{t_st_s}$ are distinct, we can see that the limit in (3.1) clearly exists, if $\max\{b_{t_st_s}\}\leq\max\{b_{g_sg_s}\}$ and is ∞ otherwise.

For the second case where not all the $b_{t_st_s}$ are distinct, suppose that the distinct values are $a_0,\dots,a_p$ and let $A_j=\{t\,|\,b_{tt}=a_j\}$, $d_j=|A_j|$ and $r'_j=\sum_{b_{t_it_i}=a_j}r_i$ then

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

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

where $A_i\to a_i$ is intended as short for $\lim_{z_{t_kt_k}\to a_i}\dots\lim_{z_{t_1t_1}\to a_i}$ where $A_i=\{t_1,\dots,t_k\}$.

Now we can rewrite this as

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

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

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

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

Lemma 18 *For an infinitely differentiable function f,*

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

Proof.

Using the infinite Taylor expansion

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

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

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

and thus

$$\frac{1}{k!} \lim_{z \to x} \left(\frac{\partial^k}{\partial x^k} \left(\frac{f(x)}{x - z} - \frac{f(z)}{x - z} \right) \right) = \frac{1}{k!} \lim_{z \to x} \left(\frac{\partial^k}{\partial x^k} \left(\sum_{n=1}^{\infty} \frac{(z - x)^{n-1}}{n!} \frac{\partial^n}{\partial x^n} f(x) \right) \right) \tag{3.3}$$

any term in the right hand side with $n > k + 1$ will include a positive power of $(z - x)$ after k derivative and so will approach zero as $z \to x$. So from (3.3)

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

$$= \frac{1}{k!} \lim_{z \to x} \left(\sum_{n=1}^{k+1} \left(\sum_{i=0}^{k} \binom{k}{i} \frac{\partial^i}{\partial x^i} \left(\frac{(z - x)^{n-1}}{n!} \right) \frac{\partial^{n+k-i}}{\partial x^{n+k-i}} f(x) \right) \right). \tag{3.4}$$

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

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

so from 3.4 we have

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

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

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

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

■

Lemma 19 *For an infinitely differentiable $g(x)$:*

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

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

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

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

Notice that since g is infinitely differentiable we have

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

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

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

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

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

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

$$= \lim_{x\to x_1}\lim_{x_{k+1}\to x}(k-1)!^{-1}\frac{\partial^{k-1}}{\partial x^{k-1}}\left(\frac{g(x)}{x-x_{k+1}}-\frac{g(x_{k+1})}{x-x_{k+1}}\right). \tag{3.5}$$

Now using Lemma 18 from (3.5) we have:

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

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

as required and this completes the proof of Lemma 19. ■

Now we return to the second case of Claim 1 where not all $b_{t_st_s}$ are different. Using Lemma 19 the expressions in the numerator and denominator of (3.1) will be in the form

$$\sum_{r_0+\dots+r_m=n}\prod_{s=0}^{m}b_{t_st_s}^{r_s}=\lim_{A_0\to a_0}\sum_{t_s\in A_0}\frac{z_{t_st_s}^{n+m}}{\prod_{j=1}^{p}(z_{t_st_s}-a_j)^{d_j}}\prod_{\substack{y\neq s\\ t_y\in A_0}}(z_{t_st_s}-z_{t_yt_y})^{-1}+\dots+$$

$$\lim_{A_p\to a_p}\sum_{t_s\in A_p}\frac{z_{t_st_s}^{n+m}}{\prod_{j=0}^{p-1}(z_{t_st_s}-a_j)^{d_j}}\prod_{\substack{y\neq s\\ t_y\in A_p}}(z_{t_st_s}-z_{t_yt_y})^{-1}=\frac{1}{(d_0-1)!}\left(\frac{\partial^{d_0-1}}{\partial z_{t_st_s}^{d_0-1}}\left(\frac{z_{t_st_s}^{n+m}}{\prod_{j=1}^{p}(z_{t_st_s}-a_j)^{d_j}}\right)\right)_{a_0}$$

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

Again in this case we can see that limit in 3.1 exists if $\max\{b_{t_st_s}\}\leq\max\{b_{g_sg_s}\}$ and is ∞ otherwise. Now that we have established 3.1 we can see that for every two element in the matrix B^n either the limit of the ratio of these elements is finite or one of them grow much faster than the other. In other words there are some $b_{i_1j_1}^{(n)},\dots,b_{i_qj_q}^{(n)}$ such that the ratio of any two of these tends to a finite non-zero limit whilst for any other $b_{ks}^{(n)}$ in

B^n

$$\lim_{n\to\infty} \frac{b_{ks}^{(n)}}{b_{ij}^{(n)}} = 0.$$

Thus if we consider

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

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

Using Claim 1 we have that the required limit in Theorem 15 exists and this completes the proof of Theorem 15. ■

Remarks As noted earlier this result could have been derived directly by the techniques in [6]. One reason for giving the alternative derivation however, is that it seems that this approach may have other applications, and may possibly allow further improvements. The point is that, as given, the proof relies on two key facts:

- That there is a notion of the *type* of a state description such that there are only finitely many types and for these Lemma 14 holds;
- The number and types of state descriptions $\eta'(a_1, \ldots, a_{n+1})$ (consistent with $\forall x_1, \ldots, x_q\, \Theta(x_1, \ldots, x_q)$) extending a state description $\eta(a_1, \ldots, a_n)$ depends only on the type of this latter state description and not on n for large n;

and there are other situations in which these conditions hold, for example when the knowledge base forces there to be just finitely many distinguishable individuals.

3.2 The BP-method And The General Polyadic Case

The BP-method as discussed above defines the application of an inference process on a predicate language as the limiting case of its application on the finite sublanguages. Here a finite sublanguage, as we discussed in Chapter 2, can be treated as a propositional language for which the inference process is well defined. The essence of defining ME on a finite sublanguage lies in the random structure method and the principle of indifference[3]. Here, given a sentence θ as our knowledge base we take the sample space to be those structures that satisfy θ and by the principle of indifference all these structures (possible situations) are considered equally likely. Then to assign probability to a sentence ϕ we take the proportion of structures that satisfy ϕ from the structures that satisfy θ and hence define the conditional probabilities. To assign a probability to ϕ on the original language L this method will then take the limit of these proportions as the size of these sub languages increase.

This method was proved to converge for the unary languages in [6] and we proved in the previous section that the same holds for unary languages this time augmented with equality. A natural question to ask here would be whether or not this method can be used for predicate languages in general. Unfortunately however, the limit of these conditional probabilities does not necessarily exist in the general case. In this section we will present an example to demonstrate a situation where the limit of these conditional probabilities does not exist and the BP-method does not converge, and this will show that the attempt to use the BP-method to define the Maximum Entropy solution on a predicate language will fail in general. However an important point to notice is that the example presented here is based on a Π_2 knowledge base i.e. a knowledge base equivalent to a Π_2 sentence.

Example Let L be a first order language with a ternary relation symbol G, a binary relation symbol R, and a unary predicate P and define

$$x =_G y \leftrightarrow \forall u, t(G(x,u,t) \leftrightarrow G(y,u,t)).^4$$

[3]It is important to note that the BP-method is not in general the same as the random structure idea however here we are dealing with knowledge bases of the form $\{w(\theta) = 1\}$ when they are the same.

[4]Notice that the underlying language is not assumed to include equality and with this definition we intend to approximate the equality via the relation G.

Let $\mathcal{E}$ be the conjunction of:

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

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

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

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

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

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

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

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

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

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

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

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

Let $\mathcal{M}^n_{\mathcal{E}}$ and $\mathcal{M}^n_{\mathcal{O}}$ denote the models of $\mathcal{E}$ and $\mathcal{O}$ of size n respectively [5].

To have an estimation of the number of models of $\mathcal{E}$, $\#\mathcal{M}^n_{\mathcal{E}}$, first let n be an even number. Then, as we will shortly explain in details, there will be

$$\frac{n!^2}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n}$$

many models for which we have

$$\mathcal{M} \models a_i \neq_G a_j \qquad 1 \leq i < j \leq n.$$

That is the number of models of $\mathcal{E}$ where the $a_1, \ldots, a_n$ are different according to $=_G$

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

and there will be at most

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

many models where not all of $a_1, \ldots, a_n$ are different according to $=_G$.

To see this notice that $n!\binom{2^{n^2}}{n}$ is the number of ways we can interpret G so that $a_1, \ldots, a_n$ are all different according to $=_G$. Let $P_1(x), \ldots, P_{2^{n^2}}(x)$ denote the sentences of the form

$$\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{n} \pm G(x, a_i, a_j)$$

When G in interpreted on $L^{(n)}$ each a_i $1 \leq i \leq n$ will satisfy one of the $P_k(x)$ $1 \leq k \leq 2^{n^2}$. The fact that $a_1, \ldots, a_n$ are different according to G means that each $P_k(x)$ is satisfied by at most one a_i or in other words for $i \neq j$, a_i and a_j will not satisfy the same $P_k(x)$. So the number of ways we can interpret G such that $a_1, \ldots, a_n$ are all different in respect to $=_G$ will be the number of ways we can choose $P_{i_1}(x), \ldots, P_{i_n}(x)$ all different among $P_1(x), \ldots, P_{2^{n^2}}(x)$ each being intended for a different a_i that will be

$$n!\binom{2^{n^2}}{n}.$$

After G is interpreted and $a_1, \ldots, a_n$ are all chosen to be different in respect to $=_G$, R will put $a_1, \ldots, a_n$ into groups of 2. To see this notice that in $\mathcal{E}$, we have $\forall x \exists y (x \neq_G y \wedge R(x, y))$. So each element is paired with at least one element and it cannot be paired with more than one because if we have $R(x, y) \wedge R(x, z)$ then we should have $x =_G y$ or $x =_G z$ or $y =_G z$ but $a_1, \ldots, a_n$ are chosen to be different according to $=_G$. So the number of different possibilities for R will be the number of ways we can put $a_1, \ldots, a_n$, into groups of 2, that is $\frac{n!}{2^{\frac{n}{2}}}$ and this should be divided by $(\frac{n}{2})!$ because the order in which these groups of 2 are chosen is not important and so the number of possibilities for R will be

$$\frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}.$$

Thus the number of models of size n for even n where the a_i's are mutually different in respect to $=_G$ will be

$$\frac{n!^2}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n}$$

For models where not all of a_i's are different according to $=_G$ assume that $n - 2k$ of

them are different and then we will take the sum over $k = 1, \ldots, \frac{n}{2}$. Notice that it is not possible to have an odd number of a_i's mutually different with respect to $=_G$, as we will shortly explain in detail, because R is dividing those elements of the model that are mutually different with respect to $=_G$ into groups of 2 and this will not be possible if the number of these elements is odd.

The number of ways we can define G such that $n - 2k$ of a_i's are different will be

$$(n-2k)!\binom{2^{n^2}}{n-2k}$$

and the number of ways we can put these $n - 2k$ many a_i's into groups of 2 will be

$$\frac{(n-2k)!}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}$$

the same way as above whilst each of the remaining $2k$ elements, say $a_{n-2k+1}, \ldots, a_n$, can be equal with respect to $=_G$ to any of the $n - 2k$ elements, $a_1, \ldots, a_{n-2k}$, and so will belong to corresponding group of 2. So for each of these $2k$ elements there will be $(n - 2k)^{2k}$ possibilities.

Thus the number of models of size even n where $n - 2k$ elements are different according to $=_G$ will be

$$(n-2k)^{2k}\frac{(n-2k)!^2}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}\binom{2^{n^2}}{n-2k} \tag{3.6}$$

but we have

$$(n-2k)^{2k}\frac{(n-2k)!^2}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!} \leq n^{2k}.\frac{n^{2n-4k-2}}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!} \leq \frac{n^{2n-1}}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!} \leq n^{2n-1}$$

and also

$$\binom{2^{n^2}}{n-2k} \leq \binom{2^{n^2}}{n-2}$$

Hence for (3.6) we have

$$(n-2k)^{2k}\frac{(n-2k)!^2}{2^{\frac{n-2k}{2}}(\frac{n-2k}{2})!}\binom{2^{n^2}}{n-2k} \leq n^{2n-1}\binom{2^{n^2}}{n-2} \tag{3.7}$$

and using (3.7) for the total number of models of $\mathcal{E}$ of size n where not all the a_i's are different with respect to $=_G$ we will have

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

And this gives us an upper bound on the number of models in this case.

Hence for n even, we have

$$\frac{n!^2}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} \leq \#\mathcal{M}^n_{\mathcal{E}} \leq \frac{n!^2}{2^{\frac{n}{2}}(\frac{n}{2})!}\binom{2^{n^2}}{n} + n^{2n}\binom{2^{n^2}}{n-2}.$$

We will now continue as before to find an estimation of $\#\mathcal{M}^n_{\mathcal{E}}$ where n is an odd number.

Notice that in this case there will be no model of $\mathcal{E}$ where $a_1, \ldots, a_n$ are mutually different with respect to $=_G$. To see this we should remember that an interpretation of R will be grouping the elements of the model such that each group contains at least 2 different elements with respect to $=_G$ (because of the conjunct $\forall x \exists y(x \neq_G y \wedge R(x,y))$). On the other hand if a group contains more than 2 elements, say 3, $\mathcal{E}$ will force the third element to be equal according to $=_G$ with one of the other two. So when all the elements of the model are different with each other according to $=_G$, there cannot be any group with more than 2 elements hence R will be dividing the elements of the model into disjoint pairs and this is not possible when the number of elements is odd.

Thus the only models of $\mathcal{E}$ of size odd will be those in which some of the a_i's are equal according to $=_G$. In exactly the same way as above we can show that the number of models of size n for odd n, will be less than

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

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

Comparing the upper bound calculated for models of size n for odd n with the lower bound of models of size n for even n, we can see that $\mathcal{E}$ has significantly more models of even size than models of odd size.

We will now try the same method to find an estimation of the number of models of O. First consider n to be even.

According to O there should exists at least one element a_i for which we have $R(a_i, a_i)$. This means that O cannot have models of size even where all the elements are different with respect to $=_G$. To see this assume that all the elements are different according to $=_G$ and let a_i be such that we have $R(a_i, a_i)$. Then first of all we cannot have $R(a_i, a_j)$ for $i \neq j$ because otherwise we will have $R(a_i, a_i) \wedge R(a_i, a_j)$ and so we should have $a_i =_G a_j$ which is a contradiction. So if we have $R(a_i, a_i)$ then $\neg R(a_i, a_j)$ for all $j \neq i$. On the other hand a_i will be the only element which is in the relation R with itself because if there is another element a_k for which we have $R(a_k, a_k)$ then we should have $a_i =_G a_k$ which is again a contradiction. So R will connect a_i only to itself and then will divide the rest of the elements into groups of two which is impossible as there will be an odd number of elements left. Thus there will be no model of size n where the elements are all different with respect to G.

For the number of models of size n where not all the elements are different with respect to $=_G$ suppose first that there are $n - 2k$ distinguishable elements. There will be an element connected to itself through R which should be one of these $n - 2k$ elements but as above this cannot be the case because there can be at most one of them with this property and if there exists one such element among them there will be an odd number of them left and it will not be possible to interpret R in a way to put them into groups of two. Hence there will be no model where $n - 2k$ elements are different with respect to $=_G$.
Next will be the case where the models are of size even n and $n - 2k + 1$ elements are different with respect to $=_G$. In this case exactly one of these $n - 2k + 1$ elements will be connected to itself and not any other of the remaining $n - 2k$ elements because there should be an odd number of them that are connected to themselves so the remaining will be of an even number and so can be divided into groups of two by R. So there should be at least one and there cannot be more than one because they are different with respect to $=_G$ and the other $n - 2k$ will again be divided into groups of two. For the remaining $2k - 1$ elements each can be equal to one of the $n - 2k + 1$ elements.

Hence the number of possibilities will be

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

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

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

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

Thus the number of models of size even n will be **at most**

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

And this gives us an upper bound on the number of models of $\mathcal{O}$ of even size.

We should find an estimation for the number of models of $\mathcal{O}$ of odd size. For an odd number n, $\mathcal{O}$ has

$$\frac{n!(n-1)!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}.n.\binom{2^{n^2}}{n}$$

many models where all the elements are different with respect to $=_G$. This is because we can choose n different elements with respect to $=_G$ in $n!\binom{2^{n^2}}{n}$ many ways and among them exactly one should be connected only to itself for which there are n possibilities and then the remaining $n-1$ should be divided into groups of two for which there are $\frac{(n-1)!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}$ possibilities. And there are at most

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

many models where not all the elements are different according to $=_G$ in the same way that it is calculated above. Hence

$$\frac{n!(n-1)!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}.n.\binom{2^{n^2}}{n}$$

gives a lower bound on the number of models of O of odd size.

By the discussion above, for even n, we have

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

$$\leq \frac{n^{2n}\frac{2^{n^2}!}{(n-1)!(2^{n^2}-n+1)!}}{\frac{n!^2}{2^{\frac{n}{2}}(\frac{n}{2})!}\frac{2^{n^2}!}{n!(2^{n^2}-n)!}} \leq \frac{n^{2n+1}2^{\frac{n}{2}}(\frac{n}{2})!}{n!^2(2^{n^2}-n+1)}$$

$$\leq \frac{n^{2n+1}2^{\frac{n}{2}}}{2^{n^2}-n+1}$$

but we have

$$n^{2n+1}2^{\frac{n}{2}} = 2^{(2n+1)\log n+\frac{n}{2}}$$

and

$$2^{(2n+1)\log n+\frac{n}{2}} << 2^{n^2}$$

because for large enough n we have $3\log n + \frac{1}{2} << n$. Thus

$$\lim_{\substack{n\to\infty \\ n\ even}} \frac{\#\mathcal{M}_O^n}{\#\mathcal{M}_\mathcal{E}^n} = 0.$$

Using the same pattern, for odd n, we have

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

$$\leq \frac{n^{2n}\frac{2^{n^2}!}{(2^{n^2}-n+2)!(n-2)!}}{\frac{n!(n-1)!}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}n\frac{2^{n^2}!}{(2^{n^2}-n)!n!}} \leq \frac{n^{2n}(n-1)2^{\frac{n-1}{2}}(\frac{n-1}{2})!}{n!(n-2)!(2^{n^2}-n+2)(2^{n^2}-n+1)}$$

$$\leq \frac{n^{2n+1}2^{\frac{n-1}{2}}}{(2^{n^2}-n+2)(2^{n^2}-n+1)}$$

$$\leq \frac{2^{(2n+1)\log n+\frac{n-1}{2}}}{(2^{n^2}-n+2)(2^{n^2}-n+1)}$$

and so again we have

$$\lim_{\substack{n\to\infty \\ n\ odd}} \frac{\#\mathcal{M}_{\mathcal{E}}^n}{\#\mathcal{M}_{\mathcal{O}}^n} = 0.$$

Now let's consider the sentence ϕ as follow

$$(\mathcal{E} \wedge \forall x P(x)) \vee (\mathcal{O} \wedge \forall x \neg P(x))$$

and let the knowledge base be $K = \{ w(\phi) = 1 \}$. Then for a sentence like $P(a_1)$ and a probability function w satisfying this knowledge base the limit

$$\lim_{n\to\infty} w^{(n)}(P(a_1))$$

does not exist.

This example shows that even in the case of a Π_2 knowledge base over a language L, we cannot in general define the probability of a sentence ϕ, $w(\phi)$, as the limiting case of probabilities assigned to it over the finite sub languages $L^{(n)}$, $w^n(\phi^{(n)})$, simply because the relevant asymptotic limit does not necessarily exist even when we drop the equality from language (the idea of using a relation symbol, here G, to 'approximate' the equality via $=_G$ is due to Grove, Halpern and Koller [15] to my knowledge).

In the rest of this chapter we will thus turn our attention to knowledge bases of lower complexity level. We will show that this method converges for Σ_1 and for special cases of Π_1 knowledge bases. However as mentioned before we conjecture that this is true for any Π_1 knowledge base from a finite first order language without function symbols or constant symbols other that $a_i's$.

3.3 The BP-Method for Σ_1 Knowledge Bases

Following our investigation of BP-method we will show in this section that this method converges for knowledge bases of the form

$$K = \{ w(\exists \vec{x} \theta(\vec{x})) = 1 \}$$

where $\theta(\vec{x})$ is a quantifier free consistent sentence. Here we consider $\theta(\vec{x})$ to be from a general predicate language without any restriction on the arity of relation symbols but with no function symbols nor equality. As we shall shortly see, in this case our method will provide the same answer independent of the sentence $\exists\vec{x}\theta(\vec{x})$ in the knowledge base!

In the case of an existential sentence as above one might expect that as the number of constants increases, the Maximum Entropy solution gets closer to $P_=$, that is defined to be the completely independent solution over an empty knowledge base [see [24]]. This is by definition the probability function that for $\Delta(a_1, \ldots, a_r)$, a state description of $L^{(r)}$, we have

$$P_=(\Delta(a_1, \ldots, a_r)) = \frac{1}{q_r},$$

where q_r is the number of state descriptions of $L^{(r)}$.

To see this notice that the $ME(K^{(r)})$ divides the probability equally between the state descriptions of $L^{(r)}$ that are models of $K^{(r)}$. But as r increases the proportion of state descriptions that satisfies $K^{(r)}$ will also increase and the ME solution will look more and more like the $P_=$. We will show that this proportion will converge to 1.

Thus consider the knowledge base K as

$$K = \{w(\exists\vec{x}\theta(\vec{x})) = 1\}$$

where θ is a quantifier free consistent sentence.

Theorem 20 $\lim_{r\to\infty} ME(K^{(r)})(\phi) = P_=(\phi)$ *for any sentence* ϕ *in* SL *and* K *as above.*

Proof. To see this let $S^{(r)}$ be the set of state descriptions over $L^{(r)}$ and $S'^{(r)}$ be the subset of $S^{(r)}$ that satisfies $K^{(r)}$ and for $s_i^{(k)} \in S^{(k)}$ define $S_{k,i}^{(r)} = \{ s_j^{(r)} \in S^{(r)} \mid s_j^{(r)} \vDash s_i^{(k)}\}$. In other words $S_{k,i}^{(r)}$ is the set of state description on $L^{(r)}$ that extend the state description $s_i^{(k)}$ in $L^{(k)}$. Notice that $|S_{k,i}^{(r)}| = |S_{k,j}^{(r)}|$ where $s_i^{(k)}, s_j^{(k)}$ are both state descriptions of $L^{(k)}$ because all state descriptions of $L^{(k)}$ will have the same number of extensions to state descriptions of $L^{(k+1)}$.

Now it is enough to show that for each $s_i \in S^{(k)}$,

$$\lim_{r\to\infty} ME(K^{(r)})(s_i) = P_=(s_i).$$

that is

$$\forall \epsilon > 0\, \exists N\, \forall r > N\, |ME(K^{(r)})(s_i) - P_=(s_i)| < \epsilon$$

We will discuss this in two cases,

Case 1 If $s_i \vDash K^{(r)}$ then $ME(K^{(r)})(s_i) = \frac{|S^{(r)}_{k,i}|}{|S'^{(r)}|}$ and $P_=(s_i) = \frac{1}{|S^{(k)}|}$ so it is enough to show that we can take N large enough such that for all $r > N$,

$$|\frac{|S^{(r)}_{k,i}|}{|S'^{(r)}|} - \frac{1}{|S^{(k)}|}| < \epsilon.$$

To see this notice that

$$\frac{|S^{(r)}_{k,i}|}{|S'^{(r)}|} - \frac{1}{|S^{(k)}|} \leq 1 - \frac{|S'^{(r)}|}{|S^{(r)}|}$$

because $1 - \frac{|S'^{(r)}|}{|S^{(r)}|} = \frac{|S'^{(r)}|}{|S^{(r)}_{k,i}|}\left(\frac{|S^{(r)}_{k,i}|}{|S'^{(r)}|} - \frac{1}{|S^{(k)}|}\right)$ and $|S^{(r)}_{k,i}| \leq |S'^{(r)}|$ as we have $s_i \vDash K^{(r)}$ and so every extension of s_i will also be a model of $K^{(r)}$, and notice that $|S^{(k)}|.|S^{(r)}_{k,i}| = |S^{(r)}|$.

So it would be enough to show that we can take N large enough so that

$$1 - \frac{|S'^{(r)}|}{|S^{(r)}|} < \epsilon. \tag{3.8}$$

Now if we define the probability function Bel_n on $SL^{(n)}$ as

$$Bel_n(\psi) = \frac{|\{\, M \in TL^{(n)} \mid M \vDash \psi \,\}|}{|TL^{(n)}|}$$

where TL is the set of term models on $L^{(n)}$ and let $Bel_\infty(\psi) = \lim_{n\to\infty} Bel_n(\psi)$ then by a theorem due to Fagin [10], we know that Bel_∞ agrees with $P_=$.

Lemma 21 *Let $\phi \in SL$ be of the form $\exists x_1, \ldots, x_t \psi(\vec{x})$ where ψ is quantifier free. If ϕ is satisfiable then $P_=(\phi) = 1$.*

Proof. To show this we will show that for a universal sentence ϕ' of the form $\forall x_1, \ldots, x_t \psi'(\vec{x})$ that is not a tautology we have $P_=(\phi') = 0$.

Let $Q_i(x_1, \ldots, x_t)$, $i \in I$ enumerate formulae of the form

$$\bigwedge_{\substack{i_1,\ldots,i_j \le t \\ Rj-ary \\ R \in RL, j \in N^+}} \pm R(x_{i_1}, \ldots, x_{i_j}).$$

Since $\forall x_1, \ldots, x_t \psi'(\vec{x})$ is not a tautology then there is some strict subset J of I such that

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

For $i_1 < i_2 < \ldots < i_t < q$ the number of extensions of $Q_i(a_{i_1}, \ldots, a_{i_t})$ is the same for each i so

$$P_=(Q_i(a_{i_1}, \ldots, a_{i_t})) = \frac{1}{|I|}$$

and for disjoint $\vec{a}^1, \ldots, \vec{a}^r$,

$$P_=(Q_{n_1}(\vec{a}^1) \wedge \ldots \wedge Q_{n_r}(\vec{a}^r)) = \frac{1}{|I|^r}.$$

So

$$\begin{aligned} P_=(\forall x_1, \ldots, x_t \psi'(\vec{x})) &\le P_=(\psi'(\vec{a}^1) \wedge \ldots \wedge \psi'(\vec{a}^r)) \\ &= \sum_{n_1,\ldots,n_r \in J} P_=(Q_{n_1}(\vec{a}^1) \wedge \ldots \wedge Q_{n_r}(\vec{a}^r)) \\ &= \left(\frac{|J|}{|I|}\right)^r \to 0 \ as \ r \to \infty. \end{aligned}$$

So for every non tautology universal sentence ϕ' we will have $P_=(\phi') = 0$ and so every satisfiable existential sentence will get value 1 and this completes the proof of Lemma 21. ■

Now we have $1 = P_=(\exists \vec{x} \theta(\vec{x})) = Bel_\infty(\exists \vec{x} \theta(\vec{x})) = \lim_{r\to\infty} Bel_r(\exists \vec{x} \theta(\vec{x})) = \lim_{r\to\infty} \frac{|S'^{(r)}|}{|S^{(r)}|}$ and this will give (3.8).

Case 2 If $s_i \not\models K^{(r)}$ then $ME(K^{(r)})(s_i) = \frac{|S'^{(r)}_{k,i}|}{|S'^{(r)}|}$ and $P_=(s_i) = \frac{1}{|S^{(k)}|}$ where $S'^{(r)}_{k,i}$ is the set of state descriptions of $L^{(r)}$ satisfying $K^{(r)}$ that extends the state description $s_i \in S^{(k)}$

so it is enough to show that we can take N_2 large enough such that for all $r > N_2$,

$$|\frac{|S'^{(r)}_{k,i}|}{|S'^{(r)}|} - \frac{1}{|S^{(k)}|}| < \epsilon \tag{3.9}$$

Now we have:

$$|\frac{|S'^{(r)}_{k,i}|}{|S'^{(r)}|} - \frac{1}{|S^{(k)}|}| = \frac{1}{|S^{(k)}|} - \frac{|S'^{(r)}_{k,i}|}{|S'^{(r)}|} \leq 1 - \frac{|S'^{(r)}_{k,i}||S^{(k)}|}{|S'^{(r)}|}$$

So to show (3.9) it will be enough to show that

$$\lim_{r\to\infty} \frac{|S'^{(r)}_{k,i}||S^{(k)}|}{|S'^{(r)}|} = 1 \tag{3.10}$$

Lemma 22 *Let $S'^{(r)}_{k,i}$ and $S^{(r)}_{k,i}$ be as defined above then*

$$\lim_{r\to\infty} \frac{|S'^{(r)}_{k,i}|}{|S^{(r)}_{k,i}|} = 1.$$

Proof. Notice that $|\frac{S'^{(r)}_{k,i}}{S^{(r)}_{k,i}}|$ is the probability that a random extension of $s^{(k)}_i$ to a state description in $L^{(r)}$ will satisfy the $K^{(r)}$. Remember that K consists of a single existential sentence $\exists x_1, \ldots, x_t \theta(x_1, \ldots, x_t)$ and let's calculate this probability.

Take the state description $s^{(k)}_i \in S^{(k)}$ and let's consider its extensions to state descriptions of $L^{(k+t)}$. Let $L^{a_{k+1},\ldots,a_{k+t}}$ be language L with only individuals $a_{k+1}, \ldots, a_{k+t}$ and let u_i $i = 1, \ldots, M$ enumerate the state descriptions of $L^{a_{k+1},\ldots,a_{k+t}}$. Then state descriptions of $L^{(k+t)}$ that are extension of $s^{(k)}_i$ will be of the form

$$s^{(k+t)}_{i,l} = s^{(k)}_i \wedge u_j \wedge V_h(a_1, \ldots, a_{k+t})$$

$$l = 1, \ldots, |S^{(k+t)}_{k,i}|, j = 1, \ldots, M, h = 1, \ldots, \frac{|S^{(k+t)}_{k,i}|}{M}.$$

At least one of the u_j's satisfies $\theta(a_{k+1}, \ldots, a_{k+t})$ and will hence satisfies $K^{(k+t)}$.

The probability that an arbitrary $s^{(k+t)}_{i,l}$ satisfies $K^{(k+t)}$ will be the number of $s^{(k+t)}_{i,l}$'s that

satisfies $K^{(k+t)}$ divided by the total number of $s_{i,l}^{(k+t)}$'s that is **at least**

$$\frac{\frac{|S_{k,i}^{(k+t)}|}{M}}{|S_{k,i}^{(k+t)}|} = \frac{1}{M}$$

and so the probability that a random $s_{i,l}^{(k+t)}$ does not satisfy $K^{(k+t)}$ will be **at most** as much as the maximum probability that u_j does not satisfy $\theta(a_{k+1}, \ldots, a_{k+t})$ that is $1 - \frac{1}{M}$.

Now consider the extension of $s_i^{(k)}$ to a state description on L^{k+pt}. We have

$$s_{i,l}^{(k+pt)} = s_i^{(k)} \wedge u_{j_1}^1 \wedge u_{j_2}^2 \wedge \ldots \wedge u_{j_p}^p \wedge V_h'(a_1, \ldots, a_{k+pt})$$

$$l = 1, \ldots, |S_{k,i}^{(k+pt)}|, \;\; j_1, \ldots, j_p = 1, \ldots, M, \;\; h = 1, \ldots, \frac{|S_{k,i}^{(k+pt)}|}{M^p}$$

where u_j^s enumerate the state description of $L^{a_{k+(s-1)t+1}, \ldots, a_{k+st}}$.

The probability that a random $s_{i,l}^{(k+pt)}$ does not satisfy $K^{(k+pt)}$ is **at most** as high as the probability that

$$u_j^1 \nvdash \theta(a_{k+1}, \ldots, a_{k+t}), \ldots, u_j^p \nvdash \theta(a_{k+(p-1)t+1}, \ldots, a_{k+pt})$$

so

$$0 \leq 1 - \frac{|S_{k,i}'^{(k+pt)}|}{|S_{k,i}^{(k+pt)}|} \leq (1 - \frac{1}{M})^p$$

Since we can have $p \to \infty$ as $r \to \infty$

$$0 \leq \lim_{r\to\infty} 1 - \frac{|S_{k,i}'^{(r)}|}{|S_{k,i}^{(r)}|} \leq \lim_{r\to\infty} (1 - \frac{1}{M})^p = 0$$

Hence,

$$\lim_{r\to\infty} 1 - \frac{|S_{k,i}'^{(r)}|}{|S_{k,i}^{(r)}|} = 0$$

and

$$\lim_{r\to\infty} \frac{|S_{k,i}'^{(r)}|}{|S_{k,i}^{(r)}|} = 1$$

as required and this completes the proof of Lemma 22. ■

To continue to show (3.10) we will show that

$$\lim_{r\to\infty} \frac{|S'^{(r)}|}{|S'^{(r)}_{k,i}||S^{(k)}|} = 1.$$

But

$$\lim_{r\to\infty} \frac{|S'^{(r)}|}{|S'^{(r)}_{k,i}||S^{(k)}|} = 1.\lim_{r\to\infty} \frac{|S'^{(r)}|}{|S'^{(r)}_{k,i}||S^{(k)}|}$$

$$= \lim_{r\to\infty} \frac{|S'^{(r)}_{k,i}|}{|S^{(r)}_{k,i}|}.\lim_{r\to\infty} \frac{|S'^{(r)}|}{|S'^{(r)}_{k,i}||S^{(k)}|}$$

$$= \lim_{r\to\infty} \frac{|S'^{(r)}_{k,i}|}{|S^{(r)}_{k,i}|}.\frac{|S'^{(r)}|}{|S'^{(r)}_{k,i}||S^{(k)}|}$$

$$= \lim_{r\to\infty} \frac{|S'^{(r)}|}{|S^{(r)}_{k,i}|.|S^{(k)}|}$$

$$= \lim_{r\to\infty} \frac{|S'^{(r)}|}{|S^{(r)}|}$$

and so to show (3.10) it will be enough to show that $\lim_{r\to\infty} \frac{|S'^{(r)}|}{|S^{(r)}|} = 1$, which we have already proved in case 1. This completes case 2.

■

Thus the BP-method is well defined and converges on the general predicate language where the knowledge base is equivalent to a Σ_1 sentence. We will study another special polyadic case in the next section.

3.4 BP-Method And Slow Formulae

So far we have investigated the BP-method for Π_1 knowledge bases from unary languages with equality as well as the Σ_1 knowledge bases on polyadic languages and we have seen that it converges and, as we shall see in the next chapter, the answer agrees with the one obtained through the W-method, an alternative generalization of Maximum Entropy to first order languages. However as we have shown in section 3.2 this method is not well defined in the general case.

In this section we will first investigate this method on the next simplest case, for a language with a single binary relation symbol and a Π_1 knowledge base. We will then introduce the notion of slow formulae and consider the application of BP-method to knowledge bases consisting of slow formulae.

Let $K = \{w(\forall x_1, \ldots, x_q \theta(x_1, \ldots, x_q)) = 1\}$ be a knowledge base on a language L with a single binary relation R and let $\Theta_j^{(l)}$ run through the state descriptions of $L^{(l)}$. The maximum entropy solution $ME(K)$, by the BP-method, will be defined as the limiting case of $ME(K^{(r)})$ as r increases. As the maximum entropy satisfies the renaming principle, on the language $L^{(r)}$ all the state descriptions consistent with $K^{(r)}$ will get the same probability. Thus for a state description $\Theta_i^{(r)}$ of $L^{(r)}$ we have

$$ME(K)(\Theta_i^{(r)}) = \lim_{m\to\infty} \frac{|\{\, \Theta^{(m)} \mid \substack{\Theta^{(m)} \text{extends} \Theta_i^{(r)} \\ \Theta^{(m)} \text{consistent with } K^{(m)}} \,\}|}{|\{\, \Theta^{(m)} \mid \Theta^{(m)} \text{consistent with } K^{(m)} \}|} \tag{3.11}$$

To study this we will first introduce some definitions and results in graph theory which would seem to throw some light on this problem.

Definition 2 *A graph property is an infinite class of graphs closed under isomorphism. A property is called hereditary if it is closed under taking induced subgraphs. For a property $\mathscr{P}$ we will denote by $|\mathscr{P}^m|$ the number of graphs of size m in $\mathscr{P}$.*

The hereditary properties and the speed of growth of $|\mathscr{P}^m|$ where $\mathscr{P}$ is a hereditary property has been studied in graph theory and we shall consider those results to study the limit in (3.11).

Scheinerman and Zito proved the following result in [32],

Theorem 23 *Let $\mathscr{P}$ be a hereditary property of graphs. Then one of the following holds.*

1. *For all m sufficiently large* $|\mathscr{P}^m|$ *is identically zero, one or two.*
2. $|\mathscr{P}^m| = \Theta(1)m^k$ *for some positive integer k (i.e.* $c_1 m^k < |\mathscr{P}^m| < c_2 m^k$ *for some* $c_1, c_2 > 0$*).*
3. *For some positive* $c_1, c_2, c_1^m < |\mathscr{P}^m| < c_2^m$.
4. *For some* $c > 0$, $m^{cm} \leq |\mathscr{P}^m|$.

They also proved in [32], that for the first case, for sufficiently large m, $\mathscr{P}^m = \emptyset$ or $\{K_m\}$ or $\{\overline{K_m}\}$ or $\{K_m, \overline{K_m}\}$, where K_m is the complete graph on m vertices and $\overline{K_m}$ is the graph on m vertices with no edges.

We will now consider the other three cases.

Definition 3 *Let G be a graph and* $x \in V(G)$*, the set of vertices of G, and let* $N(x)$ *denote the neighborhood of x that is the set of y such that* $\{x, y\} \in E(G)$*, the edge set of G. For* $x, y \in V(G)$ *we will write* $x \sim y$ *if* $N(x) - \{y\} = N(y) - \{x\}$*. This is an equivalence relation and we will call the equivalence classes of* $\sim$*, homogeneous sets. If* $x \sim y$ *in G we will say that x and y are G-equivalent.*

Balogh, Bollobas and Weinreich proved in [3] that if $|\mathscr{P}^m|$ has polynomial growth then every $G \in \mathscr{P}$ has a bounded number of homogeneous sets (more precisely, at most $k + 1$ if $|\mathscr{P}^m| = O(m^k)$) and only one of them can have unbounded order.

For a hereditary property $\mathscr{Q}$, let $l_{\mathscr{Q}}$ be the maximal number of homogeneous sets of any graph $G \in \mathscr{Q}$. It was shown in [3] that $l_{\mathscr{Q}} < \infty$ if and only if $|\mathscr{Q}^m| = O(k^m)$ for some k. Let $k_{\mathscr{Q}}$ be the maximal number of unbounded homogeneous sets for any graph $G \in \mathscr{Q}$ and $t_{\mathscr{Q}}$ be a bound on the size of bounded homogeneous sets. As was mentioned above when $k_{\mathscr{Q}} = 1$, $|\mathscr{Q}^m|$ is polynomial (of order at most $t(l_{\mathscr{Q}} - 1)$).

In [3] Balogh et al. also proved a finer result than the one in Theorem 23 for the case when $|\mathscr{Q}^m|$ has an exponential growth as well as the following results for the cases when the order of growth is above exponential.

Theorem 24 *Let* $\mathscr{Q}$ *be a hereditary property for which* $l_{\mathscr{Q}} < \infty$*. Then there exists* $k, t \in \mathbb{N}$ *such that* $|\mathscr{Q}^m| = O(m^t k^m)$*. In particular, there exists polynomials* $\{p_i\}_{i=1}^k$ *such that, for sufficiently large m, we have:* $|\mathscr{Q}^m| = \sum_{i=1}^k p_i(m) i^m$*, where k is the maximal number of unbounded homogeneous sets in* $\mathscr{Q}$*.*

Lemma 25 *If there are graphs in $\mathscr{P}$ with an arbitrary large number of homogeneous classes, then $|\mathscr{P}^m| \geq m^{(\frac{1}{2}-o(1))m}$.*

Theorem 26 *Assume $|\mathscr{Q}^m| = \Omega(k^m)$ (that is, there is $c > 0$ such that $|\mathscr{Q}^m| > ck^m$ for all m). Then either $|\mathscr{Q}^m| \geq m^{(1+o(1))m}$ or $|\mathscr{Q}^m| = m^{(1-\frac{1}{k}+o(1))m}$ for some k.*

Thus if $|\mathscr{Q}^m|$ grows faster than exponential then either $|\mathscr{Q}^m| \geq m^{(1+o(1))m}$ or $|\mathscr{Q}^m| = m^{(1-\frac{1}{k}+o(1))m}$.

For the cases where $|\mathscr{Q}^m| \geq m^{(1+o(1))m}$ they divide this into two cases. The first case is when $m^m \leq |\mathscr{Q}^m| \leq 2^{o(m^2)}$. This will be the problematic case. Scheinerman and Zito, in [32] asked the question wether for a property $\mathscr{Q}$, with $|\mathscr{Q}^m|$ in this range, the limits

$$\lim_{m\to\infty} \frac{\log|\mathscr{Q}^m|}{m\log m} \quad \textit{and} \quad \lim_{m\to\infty} \frac{\log\log|\mathscr{Q}^m|}{\log m}$$

always exists. In [4] Balogh et al. show that these limits do not necessarily exist and properties in this range can oscillate infinitely often which means that there can be large difference between $\lim inf_m(\frac{|\mathscr{Q}^m|}{m\log m})$ and $\lim sup_m(\frac{|\mathscr{Q}^m|}{m\log m})$.

However above this range, that is where $|\mathscr{Q}^m| > 2^{o(n^2)}$, $|\mathscr{Q}^m|$ will again become well behaved. In [5], Ballobas and Thomason proved the following result.

Theorem 27 *If $\lim sup_{m\to\infty} \frac{\log|\mathscr{Q}^m|}{m^2} > 0$ and $\mathscr{Q}$ is not the trivial class of all graphs, then there is an integer $k \geq 2$ such that $|\mathscr{Q}^m| = 2^{(1-\frac{1}{k}+o(1))\frac{m^2}{2}}$.*

The results on the speed of hereditary properties do not immediately apply to our general problem because they assume equality and $a_i \neq a_j$ for $i \neq j$ and more ever they are concerned with undirected graphs.[6] Nevertheless the results obtained in that study suggest some lines of inquiry in our more general case. It is in that direction that we now turn.

We will now see how we can relate our problem to this setting.

[6]The presence of equality itself is not a problem for us provided we allow the possibility that $a_i = a_j$ even though $i \neq j$.

Notice that for the language L, with only one binary relation symbol, the state descriptions of $L^{(m)}$ are of the form

$$\bigwedge_{j=1}^{m}\bigwedge_{k=1}^{m} R^{\epsilon_{jk}}(a_j, a_k).$$

These can be identified with all the possible directed graphs on the set of vertices $\{a_1, \ldots, a_m\}$. We shall use this identification in what follows.

A state description $\Theta_k^{(m)}$ is consistent with $K^{(m)}$ if and only if for every $a_{i_1}, \ldots, a_{i_q} \in \{a_1, \ldots, a_m\}$ we have $\theta(a_{i_1}, \ldots, a_{i_q})$. Let $\Theta_j^{(q)}$, $j = 1, \ldots, J$ be the state descriptions of $L^{(q)}$ and let

$$\Gamma = \{\, \Theta_j^{(q)} \mid \Theta_j^{(q)} \models \neg \bigwedge_{l_1,\ldots,l_n=1}^{q} \theta(a_{l_1}, \ldots, a_{l_q}) \,\}.$$

So Γ will be the set of state description on $\{a_1, \ldots, a_q\}$ or equivalently the set of directed graphs with $\{a_1, \ldots, a_q\}$ as the vertex set that are inconsistent with $K^{(q)}$. Thus considering the state descriptions as graphs, a state description $\Theta_k^{(m)}$ is consistent with $K^{(m)}$ if and only if the graph with vertices $\{a_1, \ldots, a_m\}$ associated with it does not have any induced subgraph of size q isomorphic to any graph in Γ.

By the discussion above the denominator in (3.11), $|\{\, \Theta^{(m)} \mid \Theta^{(m)}\text{consistent with } K^{(m)}\}|$, will be the number of graphs of size m with no induced subgraph of size n isomorphic to any graph in Γ.

Let Δ be a set of finite graphs and define $\mathscr{P}_\Delta$ to be the property of having no induced subgraph isomorphic to a graph in Δ. The property $\mathscr{P}_\Delta$ defined in this way will then be a hereditary property.

Thus for the denominator in (3.11) and Γ defined above, we will have

$$|\{\, \Theta^{(m)} \mid \Theta^{(m)}\text{consistent with } K^{(m)}\}| = |\mathscr{P}_\Gamma^m|.$$

As usual let L be a language with constants $a_1, a_2, a_3, \ldots$ and finitely many relation symbols but no functions nor equality. In what follows $b_1, \ldots, b_n$ will denote some distinct constants a_i.

Definition 4 *For $\Theta(b_1, \ldots, b_n)$, a state description (in L), we say b_i, b_j are indistinguishable mode $\Theta(\vec{b})$, denoted $b_i \sim_{\Theta(\vec{b})} b_j$, if*

$$\Theta(b_1, \ldots, b_n) \wedge b_i = b_j$$

is consistent with the axioms of equality for the language L plus =[7].

The relation $\sim_{\Theta(\vec{b})}$ is an equivalence relation. The *spectrum* of $\Theta(\vec{b})$ [see [30]] is the multiset of sizes of its equivalence classes. Let *length* of its spectrum, denoted $\|\Theta(\vec{b})\|$, be the number of non-empty equivalence classes.

Definition 5 *We say that a quantifier free formula $\theta(x_1, x_2, \ldots, x_n)$ is slow if there are some constants c, d such that for all r the number of models of $\forall \vec{x} \theta(\vec{x})$ with universe $\{a_1, \ldots, a_r\}$ is at most dc^r.*

Theorem 28 *If $\theta(x_1, x_2, \ldots, x_n)$ is slow, with bound $d(k-1)^r$ then there is a finite set A of state descriptions $\Theta(a_1, a_2, \ldots, a_k)$ of spectrum length at most $k-1$ such that*

$$\bigwedge_{i_1, \ldots, i_n = 1}^{k} \theta(a_{i_i}, \ldots, a_{i_n}) \equiv \bigvee_{\Theta(a_1, \ldots, a_k) \in A} \Theta(a_1, \ldots, a_k). \tag{3.12}$$

Proof. Suppose that $\theta(x_1, \ldots, x_n)$ is slow, with bound $d(k-1)^r$ but there is a state description $\Theta(a_1, \ldots, a_p)$ which determines a model of $\forall \vec{x} \theta(x_1, \ldots, x_n)$ (equivalently consistent with this sentence) with

$$\|\Theta(a_1, \ldots, a_p)\| > k - 1.$$

Then we can extend this state description to one with q individuals, $a_1, a_2, \ldots, a_q, q > p$ by making the new elements clones of existing elements. In other words we just add the new elements to the equivalence classes of existing elements. Furthermore we can do this in $\|\Theta(a_1, \ldots, a_p)\|^{q-p}$ ways. Thus we will have at least $\|\Theta(a_1, \ldots, a_p)\|^{q-p}$ many models of $\forall \vec{x} \theta(x_1, \ldots, x_n)$ of size q. But this clearly exceeds $d(k-1)^q$ for sufficiently large q, and this is a contradiction.

Thus if we represent

$$\bigwedge_{i_1, \ldots, i_n = 1}^{k} \theta(a_{i_1}, \ldots, a_{i_n})$$

[7] This is the analogy to the equivalence in Definition 3.

as a logically equivalent disjunction of state descriptions $\Theta_j(a_1, \ldots, a_k)$, $j = 1, \ldots, t$ because we have

$$\|\Theta_j(a_1, \ldots, a_k)\| \leq k - 1$$

for $j = 1, \ldots, t$ we will have the required representation.

In other words if $\theta(x_1, \ldots, x_n)$ is slow then there is a number k such that for each state descriptions over $a_1, \ldots, a_k$ that logically imply $\bigwedge_{i_1,\ldots,i_n=1}^{k} \theta(a_{i_1}, \ldots, a_{i_n})$ has at most $k-1$ distinguishable elements (or equivalently has spectrum of size at most $k - 1$).

■

Let the knowledge base K be given by

$$\{w(\forall \vec{x}\theta(\vec{x})) = 1\}$$

where θ is slow with bound $d(k-1)^m$. So there is a finite set A of state descriptions of $L^{(k)}$ with spectrum length at most $k - 1$ such that,

$$\bigwedge_{i_1,\ldots,i_n=1}^{k} \theta(a_{i_i}, \ldots, a_{i_n}) \equiv \bigvee_{\Theta(a_1,\ldots,a_k)\in A} \Theta(a_1, \ldots, a_k). \tag{3.13}$$

We will next show that the BP-method converges for the knowledge bases consisting of slow formula. To show this we have to show that for a state description $\Xi^{(r)}(a_1, \ldots, a_r)$ the limit

$$\lim_{m\to\infty} ME(K^{(m)})(\Xi^{(r)}) = \lim_{m\to\infty} \frac{|\{\, \Phi^{(m)} \mid \substack{\Phi^{(m)}\text{extends}\Xi_i^{(r)} \\ \Phi^{(m)}\text{consistent with } K^{(m)}} \,\}|}{|\{\, \Phi^{(m)} \mid \Phi^{(m)}\text{consistent with } K^{(m)}\}|} \tag{3.14}$$

exists and then by BP-method we can define

$$ME(K)(\Xi^{(r)}) = \lim_{m\to\infty} ME(K^{(m)})(\Xi^{(r)}).$$

To show this we will first find the number of state descriptions of $L^{(m)}$ consistent with $K^{(m)}$.

Let $\Phi(a_1, \ldots, a_m)$ be a state description consistent with K, say with equivalence classes $S_1, S_2, \ldots, S_q$ such that if i_t is minimal with $a_{i_t} \in S_t$ then $i_1 < i_2 < \ldots < i_q$.
Notice that by the discussion in the proof of Theorem 28 $q < k$.

Let $\Psi(a_{i_1},\ldots,a_{i_q})$ be the state description of $a_{i_1},\ldots,a_{i_q}$ logically implied by $\Phi(a_1,\ldots,a_m)$. Then $\Psi(a_1,\ldots,a_q)$ has spectrum $\{1,\ldots,1\}$ with length $q < k$. Notice that we can recover $\Phi(a_1,\ldots,a_m)$ from $\Psi(a_1,\ldots,a_q)$ and $S_1,\ldots,S_q$. So the number of state descriptions $\Phi(a_1,\ldots,a_m)$ is the number of choices of $\Psi(a_1,\ldots,a_q)$ and the choices of $S_1,\ldots,S_q$.

The number of state descriptions $\Psi(a_1,\ldots,a_q)$ above is the number of state descriptions on $a_1,\ldots,a_q$ consistent with $K^{(q)}$ with spectrum length q, say d_q. The only condition on the equivalence classes S's is that they should be non-empty and form a partition of $\{1,2,\ldots,m\}$, (their subscripts being determined by their least elements) so the number of choices of $S_1,\ldots,S_q$ will be the Stirling number of second kind, $S_m^{(q)} = \left\{ {m \atop q} \right\}$. So the number of choices for the $\Phi(a_1,\ldots,a_m)$ above will be

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

$$\frac{d_q}{q!}\left(q^m - q(q-1)^m + \frac{q(q-1)}{q}(q-2)^m - \ldots + (-1)^q\right).$$

Hence the number of state descriptions of $L^{(m)}$ consistent with $K^{(m)}$ is

$$\frac{d_{q_1}}{q_1!}\left(q_1^m - q_1(q_1-1)^m + \frac{q_1(q_1-1)}{q_1}(q_1-2)^m - \ldots + (-1)^{q_1}\right) +$$

$$\frac{d_{q_2}}{q_2!}\left(q_2^m - q_2(q_2-1)^m + \frac{q_2(q_2-1)}{q_2}(q_2-2)^m - \ldots + (-1)^{q_2}\right) + \ldots +$$

$$\frac{d_{q_s}}{q_s!}\left(q_s^m - q_s(q_s-1)^m + \frac{q_s(q_s-1)}{q_s}(q_s-2)^m - \ldots + (-1)^{q_s}\right)$$

where $q_s < q_{s-1} < \ldots < q_1 < k$ are the distinct spectrum lengths of the state descriptions on $L^{(k)}$ consistent with $\bigwedge_{i_1,\ldots,i_n=1}^{k}\theta(a_{i_1},\ldots,a_{i_n})$. Hence the proportion of these state descriptions as $m \to \infty$ with spectrum length q_1 will be

$$\lim_{m\to\infty}\left(\frac{d_{q_1}}{q_1!}(q_1^m - \ldots + (-1)^{q_1}) + \ldots +\right.$$

$$\left.\frac{d_{q_s}}{q_s!}(q_s^m - \ldots + (-1)^{q_s})\right)\left(\frac{d_{q_1}}{q_1!}(q_1^m - \ldots + (-1)^{q_1})\right)^{-1} =$$

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

Thus as $m\to\infty$ the number of these state descriptions will be asymptotically

$$\frac{d_{q_1}}{q_1!}q_1^m.$$

Next we fix a state description $\Xi(a_1,\ldots,a_r)$ and look at the number of models of $\forall\vec{x}\theta(\vec{x})$ extending this state description. Let $\Phi(a_1,\ldots,a_m)$ be a state description on $L^{(m)}$ extending $\Xi(a_1,\ldots,a_r)$ that is a model of $\forall\vec{x}\theta(\vec{x})$. Again by the discussion in the proof of Theorem 28, $\Phi(a_1,\ldots,a_m)$ will have spectrum of length at most $k-1$, say with equivalence classes spectrum $S_1,\ldots,S_{q'}$, $q'<k$, again ordered so that if i_t is minimal such that $a_{i_t}\in S_t$ then $i_1<i_2<\ldots<i_q$. Let h be maximal such that $i_h\leq r$.

Let $\Psi(a_1,a_2,\ldots,a_r,a_{i_{h+1}},a_{i_{h+2}},\ldots,a_{i_{q'}})$ be the state description on $a_1,a_2,\ldots,a_r,a_{i_{h+1}}$ $,\ldots,a_{i_{q'}}$ determined by $\Phi(a_1,\ldots,a_m)$. Again $\Phi(a_1,\ldots,a_m)$ can be recovered from $\Psi(a_1,a_2,\ldots,a_r,a_{r+1},\ldots,a_{r+q'-h})$ and the classes $S_1,S_2,\ldots,S_{q'}$, their order being determined as before by $i_1<i_2<\ldots<i_{q'}$ where i_t is minimal such that $a_{i_t}\in S_t$. The only difference now from the previous case (when effectively $\Xi(a_1,\ldots,a_r)=\top$) is that we no longer have an essentially free choice of partition $S_1,S_2,\ldots,S_{q'}$ because the non-empty members of

$$S_1\cap\{1,2,\ldots,r\},\ S_2\cap\{1,2,\ldots,r\},\ldots,S_{q'}\cap\{1,2,\ldots,r\} \tag{3.15}$$

form a refinement of the partition of the equivalence classes $T_1,T_2,\ldots,T_g$ of $\Xi(a_1,\ldots,a_r)$ which is determined by Ψ. Notice that there are finitely many of such possible Ψ's for all the possible spectrum lengths, say $\Psi_1,\ldots,\Psi_s$ and all the state descriptions in the nominator of (3.14) will be recovered from one of these Ψ's. Hence to show that the limit in (3.14) exists it will be enough to show that

$$\lim_{m\to\infty}\frac{\left|\{\,\Phi^{(m)}\mid {}^{\Phi^{(m)}\text{extends}\,\Xi_i^{(r)}}_{\Phi^{(m)}\text{recovered from}\,\Psi_j}{}^{\Phi^{(m)}\text{consistent with}\,K^{(m)}}\}\right|}{\left|\{\,\Phi^{(m)}\mid\Phi^{(m)}\text{consistent with }K^{(m)}\}\right|} \tag{3.16}$$

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

$$\lim_{m\to\infty} \frac{|\{\, \Phi^{(m)} \mid \substack{\Phi^{(m)}\text{extends}\Xi_i^{(r)} \\ \Phi^{(m)}\text{consistent with } K^{(m)}}\}|}{|\{\, \Phi^{(m)} \mid \Phi^{(m)}\text{consistent with } K^{(m)}\}|} = \lim_{m\to\infty} \sum_{j=1}^{s} \frac{|\{\, \Phi^{(m)} \mid \substack{\Phi^{(m)}\text{extends}\Xi_i^{(r)} \\ \Phi^{(m)}\text{consistent with } K^{(m)} \\ \Phi^{(m)}\text{recovered from}\Psi_j}\}|}{|\{\, \Phi^{(m)} \mid \Phi^{(m)}\text{consistent with } K^{(m)}\}|}.$$

For a fixed Ψ let $R_1, R_2, \ldots, R_p$ denote this refinement as in (3.15) and let q' be the spectrum length. Then for this particular refinement the number of choices of $S_1, S_2, \ldots, S_{q'}$ for which the non-empty members of (3.15) are $R_1, R_2, \ldots, R_p$ is

$$\sum_{\substack{U\subseteq\{r+1,\ldots,m\} \\ |U|\geq q'-p}} p^{m-r-|U|} \left\{ \begin{matrix} |U| \\ q'-p \end{matrix} \right\}$$

Thus the number of state descriptions corresponding to this Ψ with spectrum length q' that extend $\Xi(a_1, \ldots, a_r)$ and are consistent with $K^{(m)}$ will be

$$\sum_{\substack{U\subseteq\{r+1,\ldots,m\} \\ |U|\geq q'-p}} p^{m-r-|U|} \left\{ \begin{matrix} |U| \\ q'-p \end{matrix} \right\}.$$

If we expand this we will have

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

Thus for (3.16) we will have

$$\lim_{m\to\infty} \frac{|\{\, \Phi^{(m)} \mid \substack{\Phi^{(m)}\text{extends}\Xi_i^{(r)} \\ \Phi^{(m)}\text{consistent with } K^{(m)} \\ \Phi^{(m)}\text{recovered from}\Psi_i}\}|}{|\{\, \Phi^{(m)} \mid \Phi^{(m)}\text{consistent with } K^{(m)}\}|} =$$

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

$$\lim_{m\to\infty} \frac{\frac{1}{(q'-p)!} \sum_{j=0}^{q'-p} (-1)^{q'-p-j} \binom{q'-p}{j} \sum_{n=q'-p}^{m-r} p^{m-r-n} j^n \binom{m-r}{n}}{\frac{d_{q_1}}{q_1!} q_1^m}. \tag{3.17}$$

Notice that as there are finitely many j in the nominator of (3.17), to show that the limit in (3.17) exists it will be enough to show that it exists for each particular j.

Since $\sum_{n=q'-p}^{m-r} p^{m-r-n} j^n \binom{m-r}{n}$ is asymptotic with $\sum_{n=0}^{m-r} p^{m-r-n} j^n \binom{m-r}{n} = (p+j)^{m-r}$ it is enough to show that

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

exists for $j = 0, \ldots, q' - p$.

But since $p + j \leq q' \leq q_1$ this is clearly zero unless $p + j = q' = q_1$, in which case it exists. Hence the limit in (3.16) exists and so the BP-method will be well defined and converges on the knowledge bases consisting of slow formula.

Although we started our investigation in this section from a language with a single binary relation symbol and hence used the ideas from graph theory, which seems rather related in the obvious way, the definition of slow formulae and the proof of theorem 28 and the discussion for the existence of the limit following it does not rely in any way on the underlying language. Thus the results of the discussion above holds for any polyadic language in general as long as we are dealing with Π_1 knowledge bases consisting of slow formulae.

Thus the BP-method will converge for the Π_1 knowledge bases of the form $\{w(\forall \vec{x}\theta(\vec{x})) = 1\}$ where θ is slow without any restrictions on the language. The case where θ is not slow remains a problem of course, though perhaps the graph theory analysis will eventually provide some clues here.

Chapter 4

An Alternative Generalization Of Maximum Entropy

In this chapter we will investigate an alternative definition introduced by Jon Williamson for extending the Maximum Entropy inference process to a first order language, and try to study some of its properties to provide a comparison between this and the BP-method.

As part of studying Objective Baysianism for countably infinite domains, in [34] Jon Williamson investigates the generalization of the equivocation principle to first order languages. In [34] the equivocation principle is presented as follows:

> Equivocation: The agent's degrees of belief should otherwise [beyond what is enforced by the knowledge base] be as equivocal as possible.

Here again when we are interested in choosing a probability function to represent an agent's degree of belief based on a knowledge base K, the equivocation principle will require our inference process to be justified by the amount of information included in the resulting probability function. In other words the equivocation principle is forcing the inference process to choose the probability function with the least amount of information beyond our knowledge base K where the amount of information is measured by how close the chosen probability function is to the one called the *equivocator*.

In the case of a first order predicate language L, let $L^{(r)}$ be the restriction of L to the first r constants as before and let $\Delta_1, \ldots, \Delta_{q_r}$ be the state descriptions for $L^{(r)}$. Then

the equivocator probability function, $P_=$, is defined to be the probability function for which

$$P_=(\Delta_i) = \frac{1}{q_r} \quad i = 1, \ldots, q_r$$

where q_r is the number of state descriptions of $L^{(r)}$.

The equivocation principle will then require the candidate probability function to be as close as possible to $P_=$ where the distance is measured by the cross entropy. This forces the inference process to be information theoretically as close as possible to $P_=$ i.e. the point of total ignorance.

To be more precise, Williamson defines the r-distance between two probability function to be

$$d_r(P, Q) = \sum_{i=1}^{q_r} P(\Delta_i^{(r)}) \log(\frac{P(\Delta_i^{(r)})}{Q(\Delta_i^{(r)})})$$

where the q_r is the number of state descriptions on the first r individuals and $\Delta_i^{(r)}$ runs through the state descriptions of $L^{(r)}$.

Then for probability functions P, Q and R, P is said to be closer to R than Q if there is some N such that for all $r \geq N$, $d_r(P, R) < d_r(Q, R)$. Similarly the r-entropy of a probability function P is defined to be

$$H_r(P) = -\sum_{i=1}^{q_r} P(\Delta_i^{(r)}) \log(P(\Delta_i^{(r)})).$$

Then P is said to have greater entropy than Q, written $P \gg Q$, if there is some N such that for all $r \geq N$, $H_r(P) > H_r(Q)$. Using these definitions the equivocation principle can be rephrased as follow:

An agent's degrees of belief should be representable by a probability function that satisfies the knowledge base and is maximal with respect to $\gg$.

Thus Williamson defines the Maximum Entropy solution to a knowledge base K, $ME(K)$ as follow:

Definition 6 *Let K be a set of linear constraint as before. The Maximum Entropy solution for K, ME(K), is the probability function satisfying K such that for any other*

probability function w that satisfies K there will be an N such that for all $r > N$ we have

$$d_r(ME(K), P_=) < d_r(w, P_=).$$

This is the definition referred to as the W-method.

Notice that this definition will satisfy the equivocation principle above since,

$$d_r(w, P_=) = -H_r(w) - \log(q_r).$$

Corollary 1 *Let K be a knowledge base equivalent to a consistent Σ_1 sentence. Then $P_=$ is the Maximum Entropy solution for K chosen by W-method.*

Proof. By Lemma 21, $P_=$ is itself a solution for a knowledge base K of this form. ■

4.1 The W-Method And The Finite Model Problem

An advantage of this definition that seems rather immediate is that it does not suffer from the finite model problem. Notice that in the BP-method to define $ME(K)$ we take the limit as $r \to \infty$ of $ME(K^{(r)})$ that is the probability function defined on $L^{(r)}$ that satisfies $K^{(r)}$ and has the maximum entropy, where $L^{(r)}$ is the finite sub language of L described above. Thus the BP-method will fail when dealing with knowledge bases without any finite models while for the W-method the probability functions are all defined on L itself and the method will not suffer from the non existence of finite models. To see this consider the following example.

Example Consider the knowledge base K defined on the language L with a single binary relation symbol R and equality. [1]

$$K = \{ \forall x \neg R(x,x),\ \forall x,y(\neg(x=y) \to (R(x,y) \vee R(y,x)),$$

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

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

[1] We could use the same idea as in the previous chapter to use a ternary relation symbol G to approximate equality via $=_G$ to achieve the same result for a language without equality.

As one can easily check K defines a strict linear order without an endpoint and thus K does not have a finite model of any cardinality. Hence the attempt to define the $ME(K)$ on L as the $\lim_{r\to\infty} ME(K^{(r)})$ will fail. However as we will now show, one can still find a probability function that is closest to the equivocator in the sense of the W-method.

To see this let W be the probability function defined on L as follow, Let

$$\Upsilon_r = \{\, \vec{\epsilon} =< \epsilon_{i,j} > \mid 1 \leq i, j, k \leq r,\ \epsilon_{i,j} \in \{0,1\},\ \epsilon_{i,i} = 0,$$

$$((\epsilon_{i,j} = \epsilon_{j,k} = 1) \rightarrow (\epsilon_{i,k} = 1)),$$

$$((\epsilon_{i,j} = 1) \vee (\epsilon_{j,i} = 1)\, for\ i \neq j)\ \}.$$

Then for

$$\Theta_{\vec{\epsilon}} = \bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i, a_j)$$

let,

$$W(\Theta_{\vec{\epsilon}}) = \begin{cases} \frac{1}{r!} & if \quad \vec{\epsilon} \in \Upsilon_r \\ 0 & otherwise \end{cases}$$

Notice that this is well defined because if $\vec{\epsilon} \in \Upsilon_{r+1}$ then $\vec{\epsilon}\ \restriction_{1\leq i,j\leq r} \in \Upsilon_r$. An explanation here is that W on $L^{(r)}$ divides the whole probability equally between those state description that correspond to a strict linear ordering of $\{a_1, \ldots, a_r\}$.

We will now show that this is the closest probability function to $P_=$ that satisfies K.

W obviously satisfies K. Let w be another probability function satisfying K and let $w^{(r)}$ be the restriction of w to $L^{(r)}$ as usual. First notice that for

$$\Theta_{\vec{\epsilon}} = \bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i, a_j)$$

where $\vec{\epsilon} \notin \Upsilon_r$ we should have,

$$w(\Theta_{\vec{\epsilon}}) = 0.$$

To see this notice that, if $\vec{\epsilon} \notin \Upsilon_r$, $\Theta_{\vec{\epsilon}}$ does not correspond to a strict linear ordering of

$\{a_1, \ldots, a_r\}$ and this means that we should have,

$$\Theta_{\vec{\epsilon}} \models \left(\bigvee_{i=1}^{r} R(a_i, a_i) \vee \right.$$

$$\bigvee_{i,j=1}^{r} \left(\neg(a_i = a_j) \wedge \neg R(a_i, a_j) \wedge \neg R(a_j, a_i) \right) \vee$$

$$\left. \bigvee_{i,j,k=1}^{r} \left(R(a_i, a_j) \wedge R(a_j, a_k) \wedge \neg R(a_i, a_k) \right) \right)$$

Now if we have $w(\Theta_{\vec{\epsilon}}) > 0$ then we will have

$$w \left(\bigvee_{i=1}^{r} R(a_i, a_i) \vee \right.$$

$$\bigvee_{i,j=1}^{r} \left(\neg(a_i = a_j) \wedge \neg R(a_i, a_j) \wedge \neg R(a_j, a_i) \right) \vee$$

$$\left. \bigvee_{i,j,k=1}^{r} \left(R(a_i, a_j) \wedge R(a_j, a_k) \wedge \neg R(a_i, a_k) \right) \right) > 0.$$

which means,

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

$$\forall x, y((\neg(x = y)) \rightarrow (R(x, y) \vee R(y, x))) \wedge$$

$$\forall x, y, z(R(x, y) \wedge R(y, z)) \rightarrow R(x, z))) < 1.$$

Hence w will not satisfy K which is a contradiction.

So if w satisfies K then $w^{(r)}$ can only assign non zero probability to those state descriptions $\Theta_{\vec{\epsilon}}$ in $L^{(r)}$ for which $\vec{\epsilon} \in \Upsilon_r$ which is at most $r!$ many state descriptions.

We can now show that for each r,

$$d_r(W, P_=) \leq d_r(w, P_=),$$

that is

$$\sum_{\vec{\epsilon}} W(\Theta_{\vec{\epsilon}}) \log(\frac{W(\Theta_{\vec{\epsilon}})}{P_=(\Theta_{\vec{\epsilon}})}) \leq \sum_{\vec{\epsilon}} w(\Theta_{\vec{\epsilon}}) \log(\frac{w(\Theta_{\vec{\epsilon}})}{P_=(\Theta_{\vec{\epsilon}})}).$$

For this it will be enough to show that

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

or equivalently

$$\sum_{\vec{\epsilon} \in \Upsilon_r} \frac{1}{r!} \log(\frac{1}{r!}) \leq \sum_{\substack{\vec{\epsilon} \in \Upsilon_r \\ w(\Theta_{\vec{\epsilon}})>0}} w(\Theta_{\vec{\epsilon}}) \log(w(\Theta_{\vec{\epsilon}})),$$

that is

$$\log(\frac{1}{r!}) \leq \sum_{\substack{\vec{\epsilon} \in \Upsilon_r \\ w(\Theta_{\vec{\epsilon}})>0}} w(\Theta_{\vec{\epsilon}}) \log(w(\Theta_{\vec{\epsilon}})).$$

To see that this inequality holds notice that $\sum_{i=1}^{r!} w(\Theta_i) = 1$ and that $x \log(x)$ is a convex function and for a convex function f we have,

$$m.f\left(\frac{\sum_{i=1}^{m} x_i}{m}\right) \leq \sum_{i=1}^{m} f(x_i).$$

Also the inequality should be strict for at least one r otherwise we will have $W = w$. Thus we will have

$$\log(\frac{1}{r!}) < \sum_{\substack{\vec{\epsilon} \in \Upsilon_r \\ w(\Theta_{\vec{\epsilon}})>0}} w(\Theta_{\vec{\epsilon}}) \log(w(\Theta_{\vec{\epsilon}})).$$

and so

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

for r large enough eventually, as required.

As shown by this example there are situations where the non-existence of finite models will make the application of BP-method impossible whilst the W-method can still be applied without a problem to find the least informative probability function satisfying the knowledge base.

Although the W-method does not suffer from the finite model problem it still fails to provide a universally well defined method to make the choice of probability function as we shall shortly see in an example of these situations for a Σ_2 knowledge base [see 4.3].

In the next section we will show that the W-method is well defined for unary languages and furthermore it will give the same answer as the BP-method independent of

the quantifier complexity of the knowledge base. The same is true for polyadic languages and Σ_1 knowledge bases as seen in Theorem 20 and Corollary 1. As mentioned in the previous chapter we conjecture that the same holds for Π_1 knowledge bases.

4.2 The W-method On A Unary Language

In this section we will investigate the W-method for a knowledge base K from a unary language. We will show that, as for the BP-method, the W-method is well defined for this case and furthermore, the two methods give the same answer in this case.

It has been shown in [6] that the BP-method is well defined and converges when applied to a unary language for any knowledge base K. Here we will show that the same probability function chosen by BP-method will serve as the right answer for the W-method too. In other words, we will show that the probability function chosen by the BP-method will be the closest (in the sense defined by the W-method) probability function to $P_=$ that satisfies K.

As mentioned in Chapter 1 there is a one to one correspondence between the probability functions on $L^{(r)}$ satisfying $K^{(r)}$ and the points in the convex set $VL(K^{(r)})$. In other words the probability functions on $L^{(r)}$ can be identified with the vectors $< x_1, \ldots, x_m >$ where $x_i \geq 0$ and $\sum_{i=1}^{m} x_i = 1$ and m is the number of state descriptions on the first r individuals. We shall use this identification in what follows.

As in Chapter 2 let L be a language with only the unary predicate $P_1, \ldots, P_t$ and the constants $a_1, a_2, \ldots$. Let $Q_1, \ldots, Q_J$, $J = 2^t$ be an enumeration of the formulae $\pm P_1 \wedge \pm P_2 \wedge \ldots \wedge \pm P_t$ in some fixed order and as before let $\phi_{i,\vec{\epsilon}}$ run through the sentences of the form

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

where α_i for $i = 1, \ldots, J^k$ enumerate the exhaustive and exclusive set of sentences of the form

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

and $\vec{\epsilon} =< \epsilon_1, \ldots, \epsilon_J >$ is a sequence of 0's and 1's. Notice that $\models \neg(\phi_{i,\vec{\epsilon}} \wedge \phi_{j,\vec{\delta}})$ when $< i, \vec{\epsilon} > \neq < j, \vec{\delta} >$.

By Lemma 4, any sentence $\theta \in L^{(k)}$ is equivalent to a disjunction of some sentences $\phi_{i,\vec{\epsilon}}$ defined as above. Remember that

$$\phi^{(r)}_{i,\vec{\epsilon}} = \alpha_i \wedge \bigwedge_{j=1}^{J} (\bigvee_{k=1}^{r} Q_j(a_k))^{\epsilon_j}$$

that is equivalent to

$$\bigvee_{\substack{m_j \in P^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} \text{ for } j=k+1,\ldots,r \\ P^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}} \subset \{m_j \mid k+1 \le j \le r\}}} \left(\alpha_i \wedge \bigwedge_{j=k+1}^{r} Q_{m_j}(a_j) \right) \tag{4.1}$$

where $P^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} = \{ j \mid \epsilon_j = 1 \}$ and $P^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}} = \{ j \mid j \in P^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} \text{ and } j \notin A_i \}$ and $A_i = \{ m_j \mid j = 1, \ldots, k \}$.

This is the disjunction of those atoms of $L^{(r)}$ that logically imply $\phi^{(r)}_{i,\vec{\epsilon}}$ and each atom implies precisely one of the sentences $\phi^{(r)}_{i,\vec{\epsilon}}$ and the number of disjuncts in (4.1) is

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

where $p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}} = |P^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}|$ and $p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}} = |P^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}}|$. If W is the probability chosen through BP-method then for the state description s_j on $a_1, \ldots, a_r$ we have

$$W(s_j) = \frac{W(\phi^{(r)}_{i,\vec{\epsilon}})}{\sum_{j=0}^{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}} (-1)^j \binom{p^{\phi_{i,\vec{\epsilon}}}_{i,\vec{\epsilon}}}{j} (p^{\phi_{i,\vec{\epsilon}}}_{\vec{\epsilon}}{}^{r-k} - j)}$$

where $s_j \vDash \phi^{(r)}_{i,\vec{\epsilon}}$ as we will shortly discuss [see [6] for further discussions on these].

We will now return to our main question in this section.

Theorem 29 *Let L be a language with only finitely many unary predicates and the universe $a_1, a_2, \ldots$ and let K be a finite set of linear constraints. Then the W-method is well defined for K and gives the same answer as the BP-method*[2].

[2]It should be noted that Jon Williamson has mentioned in a private conversation that he has proved

Proof. We will show that the solution from the BP-method, that is $W = \lim_{r\to\infty} ME(K^{(r)})$ satisfies the condition in the W-method. That is

$$(\forall w \models K)\,((w \neq W) \Rightarrow \exists N \forall r \geq N\ d_r(W, P_=) < d_r(w, P_=))\,.$$

Suppose not and let $w \neq W$ be a probability function satisfying K such that for infinitely many r,

$$d_r(w^{(r)}, P_=) \leq d_r(W^{(r)}, P_=) \tag{4.2}$$

where as usual $w^{(r)}$ is the restriction of w to $L^{(r)}$.

Since $w \neq W$ we can pick a large i and $\vec{\tau}$ such that $w(\phi_{i,\vec{\tau}}) \neq W(\phi_{i,\vec{\tau}})$. From the characterization of W given in [6] we have that for large r,

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

$$\sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}}) - (r-k) \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} + \delta(w, r)$$

where

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

and we have

$$\delta(W, r), \delta(w, r) \to 0$$

as $r \to \infty$, in consequence we have

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

$$< \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log w(\phi_{i\vec{\epsilon}}) - (r-k) \sum_{i,\vec{\epsilon}} w(\phi_{i\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} \tag{4.3}$$

Notice that

$$w(\phi_{i,\vec{\epsilon}}) = \lim_{r\to\infty} w(\phi_{i,\vec{\epsilon}}^{(r)}) \tag{4.4}$$

$$W(\phi_{i,\vec{\epsilon}}) = \lim_{r\to\infty} W(\phi_{i,\vec{\epsilon}}^{(r)}) \tag{4.5}$$

---

the equivalence of the two methods in the unary languages too.

To see this notice that for a probability function v and distinct Q_i, Q_j

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

where $\Gamma^{(r)}_{Q_i}$ are those state descriptions of $L^{(r)}$ containing as a conjunct $Q_i(a_i)$ for some $1 \leq i \leq r$. Similarly [see [24]-Chapter 11]

$$\begin{aligned} v(\exists x Q_i(x) \wedge \exists x Q_j(x)) &= v(\exists x, y Q_i(x) \wedge Q_j(y)) && (4.6)\\ &= \lim_{r\to\infty} v\left(\bigvee \Gamma^{(r)}_{Q_i,Q_j}\right). && (4.7)\end{aligned}$$

So,

$$\begin{aligned} v(\exists x Q_i(x) \wedge \neg\exists x Q_j(x)) &= v(\exists x Q_i(x)) - v(\exists x Q_i(x) \wedge \exists x Q_j(x)) && (4.8)\\ &= \lim_{r\to\infty} v(\bigvee \Gamma^{(r)}_{Q_i}) - \lim_{r\to\infty} v\left(\bigvee \Gamma^{(r)}_{Q_i,Q_j}\right) && (4.9)\\ &= \lim_{r\to\infty} v\left(\bigvee (\Gamma^{(r)}_{Q_i} - \Gamma^{(r)}_{Q_i,Q_j})\right) && (4.10)\\ &= \lim_{r\to\infty} v\left(\Gamma^{(r)}_{Q_i,\neg Q_j}\right) && (4.11)\end{aligned}$$

where $\Gamma^{(r)}_{Q_i,\neg Q_j}$ are those atoms of $L^{(r)}$ which contain $Q_i(a_k)$ as a conjunct for some $1 \leq k \leq r$ but do not contain as a conjunct $Q_j(a_k)$ for any k.

We will now show that

$$v\left(\bigwedge_{k=1}^{m} \exists x Q_k(x) \wedge \bigwedge_{l=m+1}^{J} \neg\exists x Q_l(x)\right) = \lim_{r\to\infty} v\left(\bigvee \Gamma^{(r)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J}\right) \quad (4.12)$$

by induction on $J - m$.

The result for $J - m = 0$ is given by the following theorem proved in [24].

Theorem 30 *For $v : SL \to [0, 1]$ satisfying (P1-3) from Chapter 1 and $\psi(x) \in SL$,*

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

So we will have

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

Assume that (4.12) is true for $J-m$. Then

$$v\left(\bigwedge_{k=1}^{m}\exists x Q_k(x)\wedge\bigwedge_{k=m+1}^{J+1}\neg\exists x Q_k(x)\right)$$

$$=v\left(\bigwedge_{k=1}^{m}\exists x Q_k(x)\wedge\bigwedge_{k=m+1}^{J}\neg\exists x Q_k(x)\right)-v\left(\bigwedge_{k=1}^{m}\exists x Q_k(x)\wedge\bigwedge_{k=m+1}^{J}\neg\exists x Q_k(x)\wedge\exists x Q_{J+1}(x)\right)$$

$$=\lim_{r\to\infty}v\left(\bigvee\Gamma^{(r)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J}\right)-\lim_{r\to\infty}v\left(\bigvee\Gamma^{(r)}_{Q_1,\ldots,Q_m,Q_{J+1},\neg Q_{m+1},\ldots,\neg Q_J}\right)$$

$$=\lim_{r\to\infty}v\left(\bigvee\Gamma^{(r)}_{Q_1,\ldots,Q_m,\neg Q_{m+1},\ldots,\neg Q_J,\neg Q_{J+1}}\right)$$

as required.

Now we have

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

$$=\lim_{r\to\infty}w\left(\bigvee\Gamma^{(r)}_{\alpha_i,Q_{j_1},\ldots,Q_{j_m},\neg Q_{j_{m+1}},\ldots,\neg Q_J}\right)$$

$$=\lim_{r\to\infty}w\left(\alpha_i\wedge\bigwedge_{j=1}^{J}\left(\bigvee_{l=1}^{r}Q_j(a_l)\right)^{\epsilon_j}\right)$$

$$=\lim_{r\to\infty}w(\phi^{(r)}_{i,\vec{\epsilon}})$$

where $\epsilon_{j_1},\ldots,\epsilon_{j_m}=1$ and $\epsilon_{j_{m+1}},\ldots,\epsilon_{j_J}=0$ and similarly for W.
Now take r large and satisfying (4.2). Then we have that

$$\sum_s w(s)\log w(s)\leq\sum_s W(s)\log W(s)$$

where the s range over the atoms of $L^{(r)}$. Hence

$$\sum_{i,\vec{\epsilon}}\sum_{s\models\phi^{(r)}_{i,\vec{\epsilon}}}w(s)\log w(s)\leq\sum_{i,\vec{\epsilon}}\sum_{s\models\phi^{(r)}_{i,\vec{\epsilon}}}W(s)\log W(s).\tag{4.13}$$

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

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

Now notice that if s_u, s_v are atoms of $L^{(r)}$ that logically imply the same $\phi^{(r)}_{i,\vec{\epsilon}}$ then for any $m \geq r$ the number of extensions of s_u to an atom of $L^{(m)}$ logically implying $\phi^{(m)}_{i,\vec{\delta}}$ is the same as the number for s_v and hence $W(s_v) = W(s_u)$ since by renaming we should have $ME(K^{(r)})(s_v) = ME(K^{(r)})(s_u)$ and we have $W = \lim_{r\to\infty} ME(K^{(r)})$. So the right hand side of (4.13) is equal to

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

Simplifying now gives that

$$\sum_{i,\vec{\epsilon}} w(\phi^{(r)}_{i,\vec{\epsilon}}) \log \left(w(\phi^{(r)}_{i,\vec{\epsilon}}) \right) - (r-k) \sum_{i,\vec{\epsilon}} w(\phi^{(r)}_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} + \delta(\vec{w}, r)$$

$$\leq \sum_{i,\vec{\epsilon}} W(\phi^{(r)}_{i,\vec{\epsilon}}) \log \left(W(\phi^{(r)}_{i,\vec{\epsilon}}) \right) - (r-k) \sum_{i,\vec{\epsilon}} W(\phi^{(r)}_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} + \delta(\vec{W}, r) \qquad (4.14)$$

where $\delta(\vec{w}, r), \delta(\vec{W}, r) \to 0$ as $r \to \infty$. Hence, using (4.4), (4.5), we must have

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

It is proved in [6] that W is such that $\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}$ is maximal among those probability functions that satisfy K. Hence we should have

$$\sum_{i,\vec{\epsilon}} W(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}} = \sum_{i,\vec{\epsilon}} w(\phi_{i,\vec{\epsilon}}) \log p_{\vec{\epsilon}}^{\phi_{i,\vec{\epsilon}}}.$$

Hence, again using (4.4), (4.5) and (4.14) it must be the case that

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

which contradicts (4.3) and this completes the proof of Theorem 29. ■

So for a unary language the W-method is well defined and gives the same answer as BP-method.

When working with a Σ_1 knowledge base (possibly from a language with relation symbols of higher arities) $P_=$ itself will be a solution for the knowledge base as we have proved before. Thus in this case the $P_=$ itself will be the chosen probability function by the W-method which agrees with the answer from the BP-method.

In the following sections we will investigate some situations where the W-method will not be well defined and study some of its properties.

4.3 The W-method And The General Polyadic Case

As discussed before by Williamson's definition a probability function w satisfying a set of constraints K is closest to $P_=$, with respect to d, if

$$(\forall W \vDash K)[W \neq w \Rightarrow \exists N\ \forall n \geq N\ d_n(w, P_=) < d_n(W, P_=)].$$

An important issue here will be the existence of such a probability function satisfying this condition in general. Here we will show, by an example, that it is not always possible to find a probability function closest to the equivocator in the above sense and thus the W-method will not in general be viable. To see this consider the following example. Here we are dealing with a Σ_2 sentence and the idea is that we can get closer and closer to $P_=$ by making the x from $\exists x$ scarcer and scarcer and thus increasing the entropy of our probability function.

Example Let

$$K = \{\, w(\exists x \forall y,\ R(x, y)) = 1 \,\}.$$

Assume that the W-method gives a solution w in this case. So w is a probability function on L and

$$w(\exists x \forall y,\ R(x, y)) = 1.$$

The plan is to show that there is some W on L also satisfying K such that for each N there will be some $r > N$ such that $d_r(W^{(r)}, P_=) < d_r(w^{(r)}, P_=)$.

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 $e_k > 2^{-r}$. Let $w^{(r)}$ be w restricted to $L^{(r)}$ and define $W^{(r)}$ also on $L^{(r)}$ as follows:

For a state description

$$\Theta_{\vec{\epsilon}} = \bigwedge_{i,j=1}^{r} R^{\epsilon_{ij}}(a_i, a_j)$$

set

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

Claim 3 $d_r(W^{(r)}, P_=) < d_r(w^{(r)}, P_=)$.

Proof It is enough to show that

$$-\sum_{\vec{\epsilon}} W^{(r)}(\Theta_{\vec{\epsilon}}) \log W^{(r)}(\Theta_{\vec{\epsilon}}) > -\sum_{\vec{\epsilon}} w^{(r)}(\Theta_{\vec{\epsilon}}) \log w^{(r)}(\Theta_{\vec{\epsilon}}). \tag{4.15}$$

Let δ and τ respectively range over the maps from

$$\{\langle i, j\rangle \mid 1 \le i, j \le r,\ i \neq k\} \to \{0, 1\}$$

and

$$\{\langle k, j\rangle \mid 1 \le j \le r,\ \} \to \{0, 1\}$$

Then (4.15) will be

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

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

$$-\sum_{\vec{\tau}} W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \ge -\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \log w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \tag{4.16}$$

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,

$$W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau_1}}) = W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau_2}}) = 2^{-r}w^{(r)}(\bigvee_{\vec{\tau}}\Theta_{\vec{\delta}\cup\vec{\tau}}). \tag{4.17}$$

Using this (4.16) will be

$$-2^r W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})\log W^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \geq -\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})\log w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}). \tag{4.18}$$

which by (4.17) will be

$$-w^{(r)}(\bigvee_{\vec{\tau}}\Theta_{\vec{\delta}\cup\vec{\tau}})\log(2^{-r}w^{(r)}(\bigvee_{\vec{\tau}}\Theta_{\vec{\delta}\cup\vec{\tau}})) \geq -\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})\log w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})$$

The state descriptions are pairwise disjoint and so this will be

$$-\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})\log(2^{-r}\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})) \geq -\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}})\log w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) \tag{4.19}$$

But $x\log x$, is a convex function and for a convex function f, we have

$$f\left(\frac{\sum_{i=1}^n x_i}{n}\right) \leq \frac{\sum_{i=1}^n f(x_i)}{n}.$$

The number of possible $\vec{\tau}$'s in (4.19) is 2^r and so for the convex function $x\log x$ we should have

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

that is (4.19).

Furthermore the inequality in (4.15) is strict because if we had equality for all such $\vec{\delta}$ then we would have $W^{(r)} = w^{(r)}$. To see that we have this, let ν be the map from $\{\langle k, j\rangle \mid 1 \leq j \leq r, \} \to \{0, 1\}$ taking everything to 1. Then we will have

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

$$2^{-r}\sum_{\vec{\delta}} w^{(r)}(\bigvee_{\vec{\tau}}\Theta_{\vec{\delta}\cup\vec{\tau}}) = 2^{-r}\sum_{\vec{\delta}}\sum_{\vec{\tau}} w^{(r)}(\Theta_{\vec{\delta}\cup\vec{\tau}}) =$$

$$2^{-r}\sum_{\vec{\epsilon}} w^{(r)}(\Theta_{\vec{\epsilon}}) = 2^{-r}$$

This will mean that

$$e_k = w(\forall x R(a_k, x)) \leq w^{(r)}(\bigwedge_{j=1}^{r} R(a_k, a_j)) = 2^{-r}. \tag{4.20}$$

But this is a contradiction because r has been chosen large, so $2^{-r} < e_k$, This finishes the proof of Claim 3.

To construct our required W we now 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, $r_s = k_s - 1,\ s \geq 2$

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

An explanation here is that in forming W we use w but replace a_{k_0} by a 'random element' as in the above construction, replace a_{k_1} in w by a_{k_0}, a_{k_2} in w 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 $W(\exists x\, \forall y R(x,y)) = 1$.

Let w' be defined as follow:

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

then we will have

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

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

and so $d_{r_s}(W^{r_s}, P_=) = d_{r_s}(w'^{r_s}, P_=)$.

Now using the Claim proved above we have that by choosing the r_s sufficiently large,

$$d_{r_s}(w'^{(r_s)}, P_=) < d_{r_s}(w^{(r_s)}, P_=).$$

and so

$$d_{r_s}(W^{(r_s)}, P_=) < d_{r_s}(w^{(r_s)}, P_=).$$

The r_s's are an infinite increasing sequence so for each N we can find $N < r_s$ and we will have

$$d_{r_s}(W^{(r_s)}, P_=) < d_{r_s}(w^{(r_s)}, P_=).$$

as required.

Case 2 There is some g such that $e_k = 0$ for $k \geq g$.

In this case pick an $0 < j$ such that $e_j > 0$ and a 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 $n > g$,

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

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

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

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

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

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

since $w(\forall y R(a_i, y)) = 0$ for $i > g+1$, so

$$W(\exists x \forall y R(x, y)) = 1.$$

To show that

$$d_r(W^{(r)}, P_=) < d_r(w^{(r)}, P_=)$$

it is enough to show that

$$-\sum_{\vec{\epsilon}} W^{(r)}(\Theta_{\vec{\epsilon}}) \log W^{(r)}(\Theta_{\vec{\epsilon}}) > -\sum_{\vec{\epsilon}} w^{(r)}(\Theta_{\vec{\epsilon}}) \log w^{(r)}(\Theta_{\vec{\epsilon}}). \tag{4.21}$$

Now 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. Now if $\sigma(\Theta_{\vec{\epsilon}_1}) = \Theta_{\vec{\epsilon}_2}$ then $\sigma(\Theta_{\vec{\epsilon}_2}) = \Theta_{\vec{\epsilon}_1}$. So to show (4.21) it will be enough to show that for each $\vec{\epsilon}$,

$$W^{(r)}(\Theta_{\vec{\epsilon}})\log W^{(r)}(\Theta_{\vec{\epsilon}})+W^{(r)}(\Theta_{\vec{\epsilon}'})\log W^{(r)}(\Theta_{\vec{\epsilon}'}) \leq w^{(r)}(\Theta_{\vec{\epsilon}})\log w^{(r)}(\Theta_{\vec{\epsilon}})+w^{(r)}(\Theta_{\vec{\epsilon}'})\log w^{(r)}(\Theta_{\vec{\epsilon}'}) \tag{4.22}$$

where $\Theta_{\vec{\epsilon}'} = \sigma(\Theta_{\vec{\epsilon}})$ and that this inequality is strict for some $\Theta_{\vec{\epsilon}}$ eventually.
But (4.22) is:

$$\frac{w^{(r)}(\Theta_{\vec{\epsilon}}) + w^{(r)}(\Theta_{\vec{\epsilon}'})}{2}\log(\frac{w^{(r)}(\Theta_{\vec{\epsilon}}) + w^{(r)}(\Theta_{\vec{\epsilon}'})}{2})+\frac{w^{(r)}(\Theta_{\vec{\epsilon}'}) + w^{(r)}(\Theta_{\vec{\epsilon}})}{2}\log(\frac{w^{(r)}(\Theta_{\vec{\epsilon}'}) + w^{(r)}(\Theta_{\vec{\epsilon}})}{2})$$

$$\leq w^{(r)}(\Theta_{\vec{\epsilon}})\log w^{(r)}(\Theta_{\vec{\epsilon}}) + w^{(r)}(\Theta_{\vec{\epsilon}'})\log w^{(r)}(\Theta_{\vec{\epsilon}'})$$

that is

$$\left(w^{(r)}(\Theta_{\vec{\epsilon}}) + w^{(r)}(\Theta_{\vec{\epsilon}'})\right)\log(\frac{w^{(r)}(\Theta_{\vec{\epsilon}}) + w^{(r)}(\Theta_{\vec{\epsilon}'})}{2}) \leq w^{(r)}(\Theta_{\vec{\epsilon}})\log w^{(r)}(\Theta_{\vec{\epsilon}})+w^{(r)}(\Theta_{\vec{\epsilon}'})\log w^{(r)}(\Theta_{\vec{\epsilon}'}) \tag{4.23}$$

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

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

and

$$w(\forall yR(a_{g+1}, y)) = 0$$

so

$$0 = w(\forall yR(a_{g+1}, y)) < \frac{1}{2}w(\forall yR(a_j, y)) = W(\forall yR(a_{g+1}, y))$$

and $w \neq W$ and so $w^{(r)} \neq W^{(r)}$ for all r eventually that is a contradiction, and so the inequality in 4.23 must be strict.

As seen by the above example there can be situations where the closest probability function in the sense of W-method does not exist and thus the method can not be applied.

We will complete this chapter by investigating some properties of the W-method.

4.4 The W-method And Cloned State Descriptions

In this section we will study some properties of the W-method or more precisely some properties of the probability function chosen by the W-method for a knowledge base K.

Here we will first prove that the W-method is invariant under the permutation of those individuals that do not explicitly appear in the knowledge base K. The idea here is that for a permutation σ, that permutes say a_i,a_j not appearing in K, and a state description Θ we should have

$$w(\Theta) = w(\sigma(\Theta)) \tag{4.24}$$

where $\sigma(\Theta)$ will be the result of transposing a_i and a_j in Θ. The reason for this is that if (4.24) does not hold we can improve the probability function w by setting

$$w'(\Theta) = w'(\sigma(\Theta)) = \frac{w(\Theta) + w(\sigma(\Theta))}{2}$$

which will result in increasing the entropy because of the convexity of negative entropy function.

We will then introduce the notion of cloned state descriptions, that is an idea along the same lines as the notion of *slow* formulae in the previous chapter.

Here the idea of cloning a state description is to extend that state description to include a larger number of individuals where the new individuals are indistinguishable from the ones which have already appeared. In other words we extend a state description by adding new individuals that look exactly like the ones we already have.

We will show that when possible (in accordance with the knowledge base K) the W-method will not put any weight on cloned state descriptions and will divide all the probability amongst those structures that have infinitely many mutually distinguishable individuals.

4.4.1 The W-Method And Permutation of constants

Let σ be a permutation of $a_1, a_2, \ldots$ 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 31 *Let K be a linear knowledge base consisting of linear constraints in the $w(\phi_i)$, $i = 1, \ldots, n$, that is*

$$K = \{\sum_{j=1}^{n} c_{ji} w(\phi_j) = b_i \mid i = 1, \ldots, m\}$$

such that the constant a_i and a_j do not appear in K and let w_0 be the Maximum Entropy solution for K obtained by the W-method. Then

$$w_0(\sigma(\psi)) = w_0(\psi)$$

where $\sigma(\psi)$ is the result of transposing a_i and a_j throughout ψ.

Proof. Let's assume w_0 does not have this property, that is $w_0(\sigma(\psi)) \neq w_0(\psi)$ for some state description ψ and define the probability function w' as follow,

$$w'^{(n)}(\Theta^{(n)}) = 2^{-1}(w_0^{(n)}(\Theta^{(n)}) + w_0^{(n)}(\sigma(\Theta^{(n)}))),$$

where $\Theta^{(n)}$ are the state descriptions over the first n individuals and $\sigma(\Theta^{(n)})$ is the result of transposing a_i and a_j throughout $\Theta^{(n)}$ for $n \geq i, j$.

First of all w' will still satisfy K. To see this notice that since a_i and a_j do not appear in K we have $\sigma(\phi_i) = \phi_i$ and so

$$w'(\phi_i) = \frac{1}{2}(w_0(\phi_i) + w_0(\sigma(\phi_i)))$$

$$= \frac{1}{2}(w_0(\phi_i) + w_0(\phi_i)) \text{ since } \sigma(\phi_i) = \phi_i$$

$$= w_0(\phi_i).$$

Hence w' satisfies K.

Claim 4 *$d_n(w', P_=) < d_n(w_0, P_=)$ for large n eventually.*

Proof. It is enough to show that for $n \geq i, j$

$$\sum_{\Theta_i} w'^{(n)}(\Theta_i) \log(w'^{(n)}(\Theta_i)) < \sum_{\Theta_i} w_0^{(n)}(\Theta_i) \log(w_0^{(n)}(\Theta_i)). \tag{4.25}$$

But we have

$$\sum_{\Theta_i} w'^{(n)}(\Theta_i) \log(w'^{(n)}(\Theta_i)) = \sum_{\Theta_i} 2^{-1} \left(w_0^{(n)}(\Theta_k) + w_0^n(\sigma(\Theta_k))\right) \log(2^{-1}(w_0^{(n)}(\Theta_k) + w_0^n(\sigma(\Theta_k))))$$

For each Θ_i there is exactly one Θ_j such that $\Theta_j = \sigma(\Theta_i)$ and we will have $\Theta_i = \sigma(\Theta_j)$ too. Thus to show (4.25) it is enough to show that

$$2^{-1} \left(w_0^{(n)}(\Theta_i) + w_0^n(\sigma(\Theta_i))\right) \log(2^{-1}(w_0^{(n)}(\Theta_i) + w_0^n(\sigma(\Theta_i))) +$$
$$2^{-1} \left(w_0^{(n)}(\Theta_j) + w_0^n(\sigma(\Theta_j))\right) \log(2^{-1}(w_0^{(n)}(\Theta_j) + w_0^n(\sigma(\Theta_j)))$$

$$\leq w_0^{(n)}(\Theta_i) \log(w_0^{(n)}(\Theta_i)) + w_0^{(n)}(\Theta_j) \log(w_0^{(n)}(\Theta_j)) \tag{4.26}$$

with strict inequality for at least one Θ_i.

That is

$$2. \left(\frac{w_0^{(n)}(\Theta_i) + w_0^n(\Theta_j)}{2} \log(\frac{w_0^{(n)}(\Theta_i) + w_0^n(\Theta_j)}{2}) \right) \leq$$
$$w_0^{(n)}(\Theta_i) \log(w_0^{(n)}(\Theta_i)) + w_0^{(n)}(\Theta_j) \log(w_0^{(n)}(\Theta_j))$$

But $x \log(x)$ is a convex function so

$$2. \left(\frac{w_0^{(n)}(\Theta_i) + w_0^n(\Theta_j)}{2} \log(\frac{w_0^{(n)}(\Theta_i) + w_0^n(\Theta_j)}{2}) \right) \leq$$
$$w_0^{(n)}(\Theta_i) \log(w_0^{(n)}(\Theta_i)) + w_0^{(n)}(\Theta_j) \log(w_0^{(n)}(\Theta_j))$$

and thus

$$\sum_{\Theta_i} w'^{(n)}(\Theta_i) \log(w'^{(n)}(\Theta_i)) \leq \sum_{\Theta_i} w_0^{(n)}(\Theta_i) \log(w_0^{(n)}(\Theta_i))$$

This inequality should be strict eventually otherwise we will have $w' = w_0$ which is a contradiction because we have $w_0(\psi) \neq w_0(\sigma(\psi))$ while $w'(\psi) = w'(\sigma(\psi))$ and this completes the proof of the Claim 4. ■

Claim 4 gives the required contradiction because we assumed w_0 to be the closest probability function to $P_=$ that satisfies K. Hence we should have

$$w_0(\psi) = w_0(\sigma(\psi))$$

and this completes the proof of Theorem 31. ■

Thus the W-method remains invariant under the permutations that permute those individuals that do not appear explicitly in K.

We will next study the behavior of the W-method on cloned state descriptions.

4.4.2 W-Method And Cloned State Descriptions

Take a knowledge base of the form $K = \{ w(\Theta) = 1 \}$, where $\Theta \in SL$ is a consistent Π_1 sentence.

Definition 7 *For $m \leq p$ and Φ, Ψ both consistent with Θ, we say that the state description $\Phi(a_1, \ldots, a_p)$ is a clone of the state description $\Psi(a_1, \ldots, a_m)$ if there is a function τ from p to m such that*

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

Assuming it exists, let w be the probability function chosen for K by the W-method and set $\bigvee \beta_p$ to be the disjunction of those state descriptions $\Phi(a_1, \ldots, a_p)$ which are clones of some state description on $a_1, \ldots, a_m$.

Claim 5 *If there is a state description on $a_1, a_2, \ldots, a_{m+1}$ that is consistent with Θ which is not clone of any state description on $a_1, \ldots, a_m$ then*

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

Proof. Suppose there is $a > 0$ such that $w(\bigvee \beta_p) \geq a$, for all p eventually where a is the largest number with this property. We shall show that for $p > m$ any state description $\Delta(a_1, \ldots, a_p)$ that is consistent with Θ must be clone of *some* state description on $a_1, \ldots, a_m$.

Suppose on the contrary that a state description $\Delta(a_1, \ldots, a_n)$ (where $n > m$) did exist and was consistent with Θ but was not clone of any state description on $a_1, \ldots, a_m$.

We may assume that $a_1, \ldots, a_n$ are all distinguishable in Δ, in other words replacing any a_i in $\Delta(a_1, \ldots, a_n)$ by a_j, $1 \leq j \leq n$, $i \neq j$ gives a contradiction. [For otherwise if say a_n and a_{n-1} were indistinguishable we could replace $\Delta(a_1, \ldots, a_n)$ with $\Delta(a_1, \ldots, a_{n-1}, a_{n-1})$, etc.]

Define for the state description $\Phi(a_1, \ldots, a_p)$ with $p \geq m$,

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

where $\bigvee \beta_r$ is the disjunction of those state descriptions which extend $\Phi(a_1, \ldots, a_p)$ and are clones of some state description on $a_1, \ldots, a_m$. Notice that this limit exists and for $p > m$

$$\sum_{\beta_p} w^c(\beta_p) = a > 0$$

where this time β_p range over the state descriptions on $a_1, \ldots, a_p$.

We define the probability function w as follows. For a state description $\Lambda(a_1, \ldots, a_r)$ where $r \geq n$:

If $\Lambda(a_1, \ldots, a_r)$ extends $\Delta(a_1, \ldots, a_n)$ and is a clone of $\Delta(a_1, \ldots, a_n)$ set

$$\mathrm{w}(\Lambda(a_1, \ldots, a_r)) = w(\Lambda(a_1, \ldots, a_r)) + Q_r^{-1} a$$

where Q_r is the number of clones of $\Delta(a_1, \ldots, a_n)$ on $a_1, \ldots, a_r$;

If $\Lambda(a_1, \ldots, a_r)$ is a clone of some state description $\Psi(a_1, \ldots, a_m)$ set

$$\mathrm{w}(\Lambda(a_1, \ldots, a_r)) = w(\Lambda(a_1, \ldots, a_r)) - w^c(\Lambda(a_1, \ldots, a_r));$$

Otherwise set

$$\mathrm{w}(\Lambda(a_1, \ldots, a_r)) = w(\Lambda(a_1, \ldots, a_r)).$$

Claim 6 w *extends to a probability function on L and is closer to* $P_=$ *than w in the sense defined in the W-method, that is*

$$d_n(\mathrm{w}, P_=) < d_n(w, P_=)$$

for all n eventually.

This provides the required contradiction (by the choice of w), so such a $\Delta(a_1, \ldots, a_n)$ cannot exist and thus for $p > m$ any state description $\Delta(a_1, \ldots, a_p)$ that is consistent with Θ will be a clone of some state description $\Psi(a_1, \ldots, a_m)$.

Before we proceed to prove Claim 6 it might be helpful to mention that the idea here is that because $\Delta(a_1, \ldots, a_n)$ is not a clone of any state description $\Psi(a_1, \ldots, a_m)$, for large $r > n$, $\Delta(a_1, \ldots, a_n)$ has far more clones extending it than there are clones of state descriptions on $a_1, \ldots, a_m$. In the long run then it will be more advantageous in terms of entropy to spread measure uniformly onto these clones of $\Delta(a_1, \ldots, a_n)$ than (possibly non-uniformly) on the clones of state descriptions on $a_1, \ldots, a_m$.

Proof. We will first show that w extends to a probability function on L. To show this it will be enough to show that

$$\sum_{\Lambda_i} \mathrm{w}(\Lambda_i) = 1 \tag{4.27}$$

where Λ_i range over the set of state descriptions on $a_1, \ldots, a_r$, say Γ_r, and that

$$\mathrm{w}(\Lambda(a_1, \ldots, a_r)) = \sum_{\substack{\Lambda_j \in \Gamma_{r+1} \\ \Lambda_j \models \Lambda}} \mathrm{w}(\Lambda_j(a_1, \ldots, a_{r+1})). \tag{4.28}$$

To see (4.27) let Γ_r^1 be those state descriptions in Γ_r that extend $\Delta(a_1, \ldots, a_n)$ and are clones of $\Delta(a_1, \ldots, a_n)$ and Γ_r^2 be those that are clones of some state descriptions on $a_1, \ldots, a_m$. Set $\Gamma_r^3 = \Gamma_r - (\Gamma_r^1 \cup \Gamma_r^2)$. Thus

$$\sum_{\Lambda_i} \mathrm{w}(\Lambda_i) = \sum_{\Lambda_i \in \Gamma_r^1} \mathrm{w}(\Lambda_i) + \sum_{\Lambda_i \in \Gamma_r^2} \mathrm{w}(\Lambda_i) + \sum_{\Lambda_i \in \Gamma_r^3} \mathrm{w}(\Lambda_i)$$

$$= \sum_{\Lambda_i \in \Gamma_r^1} \left(w(\Lambda_i) + Q_r^{-1} a \right) + \sum_{\Lambda_i \in \Gamma_r^2} (w(\Lambda_i) - w^c(\Lambda_i)) + \sum_{\Lambda_i \in \Gamma_r^3} w(\Lambda_i)$$

$$= \sum_{\Lambda \in \Gamma_r} w(\Lambda) - \sum_{\Lambda \in \Gamma_r^2} w^c(\Lambda) + a$$

$$= 1 + a - \sum_{\Lambda \in \Gamma_r^2} w^c(\Lambda).$$

So to show (4.27) it will be enough to show that

$$\sum_{\Lambda \in \Gamma_r^2} w^c(\Lambda) = a. \tag{4.29}$$

To see this notice that $w^c(\Lambda(a_1, \ldots, a_r)) = 0$ for $\Lambda(a_1, \ldots, a_r) \in \Gamma_r^1 \cup \Gamma_r^3$ so

$$\sum_{\Lambda \in \Gamma_r^2} w^c(\Lambda) = \sum_{\Lambda \in \Gamma_r} w^c(\Lambda) = a.$$

To see that w extends correctly to a probability function on L, as in (4.28), we will consider each case separately. For the first case, that is when $\Lambda(a_1, \ldots, a_r)$ extends $\Delta(a_1, \ldots, a_n)$ and is a clone of $\Delta(a_1, \ldots, a_n)$,

$$\mathrm{w}(\Lambda(a_1, \ldots, a_r)) = w(\Lambda(a_1, \ldots, a_r)) + Q_r^{-1} a$$

$$= \left(\sum_{\substack{\Lambda_j \in \Gamma_{r+1} \\ \Lambda_j \vDash \Lambda}} w(\Lambda_j(a_1, \ldots, a_{r+1})) \right) + Q_r^{-1} a. \tag{4.30}$$

Let $\Gamma_{r+1} = \Gamma_{r+1}^{\Delta} \cup \overline{\Gamma_{r+1}^{\Delta}}$ where Γ_{r+1}^{Δ} are those state descriptions in Γ_{r+1} that are clones of $\Delta(a_1, \ldots, a_n)$.

Notice that state descriptions in $\overline{\Gamma_{r+1}^{\Delta}}$ that extend $\Lambda(a_1, \ldots, a_r)$ are *not* clones of any state description on $a_1, \ldots, a_m$ and thus for $\Lambda(a_1, \ldots, a_{r+1}) \in \overline{\Gamma_{r+1}^{\Delta}}$, such that $\Lambda(a_1, \ldots, a_{r+1}) \vDash \Lambda(a_1, \ldots, a_r)$,

$$\mathrm{w}(\Lambda(a_1, \ldots, a_{r+1})) = w(\Lambda(a_1, \ldots, a_{r+1}))$$

and $Q_{r+1} = |\Gamma_{r+1}^{\Delta}| = n.|\Gamma_r^{\Delta}| = n.Q_r$ as every state description in Γ_k^{Δ} has exactly n extensions to state descriptions Γ_{k+1}^{Δ}[3]. So for (4.30) we have,

[3]Notice that here we are using the fact that $a_1, \ldots, a_n$ are all distinguishable in $\Delta(a_1, \ldots, a_n)$ as mentioned earlier.

$$\mathrm{w}(\Lambda(a_1,\ldots,a_r)) = Q_r^{-1}a + \sum_{\substack{\Lambda_j\in\Gamma_{r+1}\\ \Lambda_j\vDash\Lambda}} w(\Lambda_j(a_1,\ldots,a_{r+1}))$$

$$= Q_r^{-1}a + \sum_{\substack{\Lambda_j\in\Gamma^{\Delta}_{r+1}\\ \Lambda_j\vDash\Lambda}} w(\Lambda_j(a_1,\ldots,a_{r+1})) + \sum_{\substack{\Lambda_j\in\overline{\Gamma^{\Delta}_{r+1}}\\ \Lambda_j\vDash\Lambda}} w(\Lambda_j(a_1,\ldots,a_{r+1}))$$

$$= \sum_{\substack{\Lambda_j\in\Gamma^{\Delta}_{r+1}\\ \Lambda_j\vDash\Lambda}} \left(w(\Lambda_j(a_1,\ldots,a_{r+1})) + Q_{r+1}^{-1}a\right) + \sum_{\substack{\Lambda_j\in\overline{\Gamma^{\Delta}_{r+1}}\\ \Lambda_j\vDash\Lambda}} \mathrm{w}(\Lambda_j(a_1,\ldots,a_{r+1}))$$

$$= \sum_{\substack{\Lambda_j\in\Gamma^{\Delta}_{r+1}\\ \Lambda_j\vDash\Lambda}} \mathrm{w}(\Lambda_j(a_1,\ldots,a_{r+1})) + \sum_{\substack{\Lambda_j\in\overline{\Gamma^{\Delta}_{r+1}}\\ \Lambda_j\vDash\Lambda}} \mathrm{w}(\Lambda_j(a_1,\ldots,a_{r+1}))$$

$$= \sum_{\substack{\Lambda_j\in\Gamma_{r+1}\\ \Lambda_j\vDash\Lambda}} \mathrm{w}(\Lambda_j(a_a,\ldots,a_{r+1})). \tag{4.31}$$

We shall now show that w will do better than w in terms of entropy.

To show that w is closer to $P_=$ than w in the sense of W-method it will be enough to show that for r large enough

$$\sum_{\Lambda\in\Gamma_r} \mathrm{w}(\Lambda)\log(\mathrm{w}(\Lambda)) < \sum_{\Lambda\in\Gamma_r} w(\Lambda)\log(w(\Lambda))$$

that is

$$\sum_{\Lambda\in\Gamma_r^1} \mathrm{w}(\Lambda)\log(\mathrm{w}(\Lambda)) + \sum_{\Lambda\in\Gamma_r^2} \mathrm{w}(\Lambda)\log(\mathrm{w}(\Lambda)) + \sum_{\Lambda\in\Gamma_r^3} \mathrm{w}(\Lambda)\log(\mathrm{w}(\Lambda)) < \sum_{\Lambda\in\Gamma_r} w(\Lambda)\log(w(\Lambda)).$$

Expanding the left hand side we have

$$\sum_{\Lambda\in\Gamma_r^1} (w(\Lambda) + Q_r^{-1}a)\log(w(\Lambda) + Q_r^{-1}a) + \sum_{\Lambda\in\Gamma_r^2} (w(\Lambda) - w^c(\Lambda))\log(w(\Lambda) - w^c(\Lambda))$$

$$+ \sum_{\Lambda\in\Gamma_r^3} w(\Lambda)\log(w(\Lambda)) < \sum_{\Lambda\in\Gamma_r} w(\Lambda)\log(w(\Lambda)).$$

Notice that $0 < w(\Lambda) - w^c(\Lambda) \leq 1$ and so $\log(w(\Lambda) - w^c(\Lambda)) \leq 0$ and $\sum_{\Lambda \in \Gamma_r^2}(w(\Lambda) - w^c(\Lambda)) \log(w(\Lambda) - w^c(\Lambda)) \leq 0$ so it will be enough to show that

$$\sum_{\Lambda \in \Gamma_r^1}(w(\Lambda) + Q_r^{-1}a) \log(w(\Lambda) + Q_r^{-1}a) < \sum_{\Lambda \in \Gamma_r^1} w(\Lambda) \log(w(\Lambda)) + \sum_{\Lambda \in \Gamma_r^2} w(\Lambda) \log(w(\Lambda)). \tag{4.32}$$

Expanding the left hand side we have

$$\sum_{\Lambda \in \Gamma_r^1} w(\Lambda) \log(w(\Lambda) + Q_r^{-1}a) + Q_r^{-1}a \sum_{\Lambda \in \Gamma_r^1} \log(w(\Lambda) + Q_r^{-1}a) <$$

$$\sum_{\Lambda \in \Gamma_r^1} w(\Lambda) \log(w(\Lambda)) + \sum_{\Lambda \in \Gamma_r^2} w(\Lambda) \log(w(\Lambda))$$

Rearranging this we will have

$$\sum_{\Lambda \in \Gamma_r^1} w(\Lambda) \log(1 + \frac{a}{Q_r w(\Lambda)}) + Q_r^{-1}a \sum_{\Lambda \in \Gamma_r^1} \log(w(\Lambda) + Q_r^{-1}a) < \sum_{\Lambda \in \Gamma_r^2} w(\Lambda) \log(w(\Lambda)) \tag{4.33}$$

The first thing to notice here is that if $\Lambda_1, \Lambda_2 \in \Gamma_r^1$ then we can assume that w gives them the same probability, otherwise we can define a bijection σ_s between the $\Delta \in \Gamma_s^1$ extending Λ_1 and the $\Delta \in \Gamma_s^1$ that extend Λ_2 for $s \geq r$ such that if $\Delta' \in \Gamma_{s+1}^1$ extends $\Delta \in \Gamma_s^1$ then $\sigma_{s+1}(\Delta') \in \Gamma_{s+1}^1$ extends $\sigma_s(\Delta) \in \Gamma_s^1$. Now defining

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

for $\Delta \in \Gamma_s^1$ extending Λ_1 or Λ_2 and identity on other state descriptions gives a probability function satisfying K that is nearer to $P_=$ than w.

Let $\frac{b}{Q_r}$ be the common value for $w(\Lambda)$ for $\Lambda \in \Gamma_r^1$. Then (4.33) will become

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

Let $\sum_{\Lambda \in \Gamma_r^2} w(\Lambda) = d_r$, so $d_r \to a$ as $r \to \infty$, and notice that since any $\Lambda \in \Gamma_r^2$ is a clone of some state description on $L^{(m)}$ then $|\Gamma_r^2| \leq D.(m)^{r-m} \leq D.(n-1)^r$ where D is the number of atoms of $L^{(m)}$ consistent with K. On the other hand

$$\sum_{\Lambda \in \Gamma_r^2} w(\Lambda) \log(w(\Lambda)) \geq |\Gamma_r^2| \frac{d_r}{|\Gamma_r^2|} \log\left(\frac{d_r}{|\Gamma_r^2|}\right)$$

$$\geq d_r \log\left(\frac{d_r}{D.(n-1)^r}\right)$$

whilst the left hand side is at most

$$c - a\log(Q_r) = c' - a\log(n^r)$$

for some constants c and c' and it will be enough to show that

$$c' - a\log(n^r) < d_r \log(\frac{d_r}{D.(n-1)^r}),$$

for r large enough, that is

$$c' - a\log(n^r) < d_r\log(d_r) - d_r\log(D) - d_r\log((n-1)^r)$$

or

$$c' + d_r(\log(D) - \log(d_r)) < a.r\log(n) - d_r.r\log(n-1). \tag{4.34}$$

We have $\lim_{r\to\infty} d_r(\log(D) - \log(d_r)) = a(\log(D) - \log(a))$ so we can choose r large enough such that

$$|(c' + d_r(\log(D) - \log(d_r))) - (c' + a(\log(D) - \log(a)))| < \frac{\epsilon}{2},$$

and so it will be enough to show that for r large enough

$$c'' + \epsilon < r(a\log(n) - d_r\log(n-1))$$

which holds because as $r \to \infty$,

$$a\log(n) - d_r\log(n-1) \to a\log(\frac{n}{n-1}) > 0$$

and this completes the proof of Claim 6. ■

Thus 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 $a_1, \ldots, a_r$ that are clones of some state description on $a_1, \ldots, a_m$, then eventually every state description consistent with Θ should be clone of some state description $a_1, \ldots, a_m$ which is a contradiction as we have assumed the existence of a state description on $a_1, \ldots, a_{m+1}$ consistent with Θ that is not a clone of any state description on $a_1, \ldots, a_m$. ■

So where K allows it, for any m the limit as $p \to \infty$ of the probability of the state descriptions on $a_1, \ldots, a_p$ that are clones of state descriptions on $a_1, \ldots, a_m$ will tend to zero. In other words w will in the limit put all the probability on the structures in which there are infinitely many explicitly distinct individuals.

Chapter 5

Conclusions

In this thesis we set out to investigate inference processes on first order languages as a means to provide a satisfactory answer to the main question introduced in Chapter 1 -that is assigning probabilities (beliefs) to sentences of a languages on the basis of a set of linear knowledge base.

In Chapter 2 we adopted the BP-method, introduced by Paris and Barnett in [6], to generalize the Minimum Distance and Limiting Centre of Mass inference processes to unary first order languages. Although the results were explicitly stated for MD and CM_∞[1], the same method will work for the spectrum of inference process based on the generalized Renyi Entropies. In particular we argued that in the special case where the knowledge base is equivalent to a finite, consistent set of axioms $\mathcal{T}$ that hold categorically, i.e.

$$K = \{ w(\phi) = 1 \mid \phi \in \mathcal{T} \}$$

all the above inference processes and in fact any inference process satisfying renaming principle will be well defined and will always give the same answer.

In Chapter 3 we focused our attention on the Maximum Entropy inference process as the most commonly accepted inference process. We proved that the BP-method can be applied to generalize the ME inference process to unary first order languages with equality on Π_1 knowledge bases and for this we presented a different machinery to work with the BP-method than the one introduced in [6]. Although we later showed that the BP-method is not applicable in the most general case by providing an example

[1]The same results have been proved for Maximum Entropy inference process in [6].

of a Π_2 knowledge base, we proved that the method is well defined and converges for Σ_1 knowledge bases on general polyadic languages. We conjecture that the same holds for Π_1 knowledge bases and in an attempt to investigate this conjecture, we introduced the notion of *slow* formulae to categorize a subset of Π_1 formulae for which there is a bound on the number of models of any finite size. We proved that the BP-method is well defined and converges for knowledge bases consisting of slow formulae.

In Chapter 4 we studied an alternative generalization of Maximum Entropy to first order languages, the W-method, introduced by Jon Williamson in [34].

Although we showed that the W-method is not a universally well defined method either, we proved that it is well defined on the unary first order languages and for the general polyadic languages with Σ_1 knowledge bases. Furthermore we proved that the two methods give the same answer in these cases. We conjecture that the W-method is also well defined for Π_1 knowledge bases and that the two methods give the same answer in this case too. In the second half of Chapter 4 we investigated some properties of the W-method and in particular we defined the notion of cloned state descriptions as an analogy to the notion of slow formulae to investigate the behavior of the W-method.

The question of finding a suitable generalization of Maximum Entropy that is well defined on a general polyadic language independent of the quantifier complexity of the knowledge base will still remain open. There is still a lot of work needed to settle the case of a Π_1 knowledge base, as mentioned in Chapter 3.

Bibliography

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

[2] Bacchus, F, Grove, A.J., Halpern, J.Y. and Koller, D., From statistical knowledge to degrees of belief, *Artificial Intelligence,* 1996, **87**:75-143.

[3] Balogh, Bollobas and Weinreich, The speed of hereditary properties of graphs, *Journal of Combinatorial Theory*, series B 79, 2000, 131-156.

[4] Balogh, Bollobas and Weinreich, The penultimate range of growth for graph properties in *European Journal of Combinatorics*, v.22 n.3,2001, 277-289.

[5] Ballobas and Thomason, Projections of bodies and hereditary properties of hyper graphs in *Bulletin of the London Mathematical Society,* 27(5), 1995, 417-424.

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

[7] Carnap, R., A basic system of inductive logic, in *Studies in Inductive Logic and Probability*, Volume II, ed. R. C. Jeffrey, University of California Press, 1980, 7-155.

[8] Chang, C.C and Keisler, H.J., *Model Theory*, Studies in Logic and the Foundations of Mathematics, Vol. 73., North Holland Publishing Co., 1973.

[9] Dimitracopoulos, C., Paris, J.B., Vencovská, A. and Wilmers, G.M., A multivariate probability distribution based on the propositional calculus, *Manchester Centre for Pure Mathematics*, University of Manchester, UK, preprint number 1999/6. Also at `http://www.maths.manchester.ac.uk/~jeff/`

[10] F. Fagin, Probabilities on finite models, *Journal of Symbolic Logic*, Vol. 41, No 1, (1976) 50-58.

[11] Fitelson, B., *Inductive Logic*, `http://fitelson.org/il.pdf`

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

[13] A.J. Grove, J.Y. Halpern, D. Koller, Random Worlds and Maximum Entropy, *Journal of Artificial Intelligence Research*, **2** (1994) 33-88.

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

[15] A.J. Grove, J.Y. Halpern, D. Koller, Asymptotic conditional probabilities: the non-unary case. *J. Symbolic Logic,* **6**1(1) (1996) 250-276.

[16] Hintikka, J. and Niiniluoto, I., An axiomatic foundation for the logic of inductive generalization, in *Studies in Inductive Logic and Probability, Volume II*, Ed. R.C.Jeffrey, University of California Press, Berkeley and Los Angeles, 1980, 158-181.

[17] Jaynes, E.T., Information theory and statistical mechanics. The Physical Review, 1957, 106(4):620630.

[18] Jaynes, E.T., *Probability theory: the logic of science.* Cambridge University Press, Cambridge, 2003.

[19] Johnson, W.E., Probability: The deductive and inductive problems, *Mind*, 1932, **41**(164):409-423.

[20] Kuipers, T.A.F., A survey of inductive systems, in *Studies in Inductive Logic and Probability, Volume II*, Ed. R.C.Jeffrey, University of California Press, Berkeley and Los Angeles, 1980, 183-192.

[21] Kuipers, T.A.F., On the generalization of the continuum of inductive methods to universal hypotheses, *Synthese*, 1978, **37**:255-284.

[22] Paris, J.B., *A short course on Inductive Logic, JAIST 2007*, `http://www.maths.manchester.ac.uk/~jeff`

[23] Paris, J.B., On filling-in missing conditional probabilities in causal networks, in *International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems*, **13**(3), 2005, 263-280.

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

[25] Paris, J.B., On the distribution of probability functions in the natural world, in *Probability Theory: Philosophy, Recent History and Relations to Science*, Eds. V.F.Hendricks, S.A.Pedersen & K.F.Jøgensen, Synthese Library Vol.297, 2001, 125-145.

[26] Paris, J.B. and Vencovská, On the applicability of maximum entropy to inexact reasoning, *International Journal of Approximate Reasoning*, 1989, **3**(1), 1-34.

[27] Paris, J.B. and Vencovská, A note on the inevitability of maximum entropy, *International Journal of Approximate Reasoning*, 1990, **4**(3), 183-224.

[28] Paris, J.B. and Vencovská, In defense of the maximum entropy inference process, *International Journal of Approximate Reasoning*, 1997, **17**(1), 77-103.

[29] Paris, J.B. and Vencovská, A., Common sense and stochastic independence, in *Foundations of Bayesianism*, Eds. D.Corfield & J.Williamson, Kluwer Academic Press, 2001, 203-240.

[30] Landes, J., Paris, J.B. and Vencovská, A., A survey of some recent results on spectrum exchangability in polyadic inductive logic, to be submitted to *Knowledge, Rationality and Decision.*

[31] Rosenkrantz, R. D., *Inference, method and decision: towards a Bayesian philosophy of science.* Reidel, Dordrecht, 1977.

[32] Scheinerman E. R. and Zito, J., On the size of hereditary classes of graphs in *Journal Combinatorial Theory* Ser. B 61, 1994, 16-39.

[33] Williamson, J., *Bayesian nets and causality: philosophical and computational foundations.* Oxford University Press, Oxford, 2005.

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