Maximum Entropy Inference with Quantified Knowledge

Barnett, O. W. and Paris, J. B.

2008

MIMS EPrint: **2008.104**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Maximum Entropy Inference with Quantified Knowledge

O. W. Barnett* and J. B. Paris
School of Mathematics
The University of Manchester
Manchester M13 9PL
owen_barnett@hotmail.com jeff.paris@manchester.ac.uk

August 15, 2008

Abstract

We investigate uncertain reasoning with quantified sentences of the predicate calculus treated as the limiting case of maximum entropy inference applied to finite domains.

Motivation and notation

In this modest note we consider one possible approach to the following problem $\mathcal{P}$:

> Suppose that my subjective beliefs in some sentences $\theta_1, \theta_2, \ldots, \theta_m$ of a predicate language are constrained to satisfy a certain set K, say, of linear constraints. In that case what belief should I assign to some other sentence ϕ ?

Following Johnson [14] and Carnap et al [4], [5] we shall limit ourselves to the well studied case where the overlying predicate language $\mathcal{L}$ contains just finitely many *unary* predicate symbols, $P_1, P_2, \ldots, P_t$ and denumerably many constant symbols $a_1, a_2, a_3, \ldots$, the intention here being that these constants are distinct and exhaust the universe. In particular the language does not have equality nor any functions symbols. Then according to ideas of de Finetti [6], Gaifman [10],

*Supported by a UK Engineering and Physical Sciences Research Council (EPSRC) Research Studentship

Scott & Krauss [25], Williamson [27] (amongst others) it can be argued that subjective beliefs such as we have here should satisfy being the values assigned by some *probability function* on $\mathcal{L}$, that is a function Bel from the set $S\mathcal{L}$ of sentences of the language $\mathcal{L}$ into $[0,1]$ satisfying:

For all $\theta, \phi, \exists\psi(x) \in S\mathcal{L}$

(P1) If $\vDash \theta$ then $Bel(\theta) = 1$,

(P2) If $\vDash \neg(\theta \wedge \phi)$ then $Bel(\theta \vee \phi) = Bel(\theta) + Bel(\phi)$,

(P3) $Bel(\exists x\psi(x)) = \lim_{r\to\infty} Bel(\bigvee_{j=1}^{r} \psi(a_j))$.

This last condition, commonly known as Gaifman's condition, may seem a little strange from the point of view of mathematical logic, where one normally does not impose any requirement that every element of the universe has a name. In the philosophical context of induction, where we are thinking of the a_i as observable instances (for example runs of an experiment) however, it would on the contrary seem unnatural to allow the possible existence of 'observable instances' which could not be specifically referred to by name. In consequence, much of the work in this area, which harks back to Carnap's Inductive Logic programme, traditionally assumes this condition.[1]

Within our assumed context then the problem $\mathcal{P}$ of assigning beliefs can be refined to:

> Given that the subjective probabilities $Bel(\theta_1), Bel(\theta_2), \ldots, Bel(\theta_m)$ I assign to $\theta_1, \theta_2, \ldots, \theta_m \in S\mathcal{L}$ must satisfy some set K of linear constraints what subjective probability $Bel(\theta)$ is it *rational* for me to assign to $\theta \in S\mathcal{L}$?

Whilst it may be far from crystal clear what we might mean here by 'rational' we can at least make an attempt on this problem by formulating 'rationality', or common sense, conditions that one's subjective beliefs should satisfy and see to what extent they limit the choices for $Bel(\theta)$. One such requirement surely is that the values given to the θ as θ ranges over $S\mathcal{L}$ should be consistent with each other, and the constraints K, and the fact that Bel is to satisfy (P1-3). If we accept this, as we will, then the above problem amounts to asking what is the *rational choice* of a probability function constrained to satisfy K?

In the analogous case of a finite *propositional language* this problem was extensively studied in [20] and [22] and led to the recommendation that the only common sense choice is the *Maximum Entropy* solution, $ME(K)$, of the set K

[1] Apart from this consideration it also has the simplifying consequence that any function Bel from the quantifier-free sentences of $\mathcal{L}$ to $[0,1]$ satisfying (P1-2) extends uniquely to domain $S\mathcal{L}$ so as to satisfy (P1-3).

of linear constraints.[2] That is, if the finite propositional language has propositional variables $p_1, p_2, \ldots, p_n$ then a probability function (i.e. in the propositional case satisfying just (P1-2)) Bel is determined by its values on the 2^n *atoms* $\alpha_1, \alpha_2, \ldots, \alpha_{2^n}$, i.e. sentences of the form

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

and $ME(K)$ is defined to be that (unique) probability function Bel satisfying K for which the entropy

$$-\sum_{i=1}^{2^n} Bel(\alpha_i) \log Bel(\alpha_i)$$

is maximal. [For a fuller account of ME and other inference processes see [16].]

Unfortunately this approach cannot immediately be applied to our predicate language $\mathcal{L}$, not least because it is infinite. However, in the subjective spirit of what we are trying to capture here we might argue that the infinity of a universe with individuals $a_1, a_2, a_3, \ldots$ is really just a potential infinity of finite structures with individuals $a_1, a_2, a_3, \ldots, a_n$ and that the rational choice for the infinite universe should be the limit of the rational choices for the finite universes, assuming of course such a unique limit exists.

This idea of modeling the universe by a finite, albeit very large, structure is a very attractive one, firstly because it strikes a cord with the way we often seem to reason in practice, by weighing up a finite number of essentially finite possibilities, and secondly because it allows one to use well understood and transparent finite combinatorial arguments.

Not surprisingly then here have been a number papers which have investigated this general approach both without and within the context of assigning beliefs such as we are considering here. In particular, Kemeny's early contribution [15], the paper of Fagin [9] and the study of Zero-One Laws that it helped initiate (for a recent introduction see [7]), the work of Paris and Vencovská [19], [21], [18], of Shastri [26] and in particular the ambitious and wide-ranging development of Grove et al. in [1], [2], [11], [12], [13].

Roughly the idea in these last papers, what Grove et al call the *random-worlds method*, is to initially identify an agent's *objective* knowledge of statistical data $\mathcal{K}$ with the set of structures (i.e. possible worlds) of some large finite cardinality r in which this knowledge is true and identify the belief the agent (should) give to an assertion θ on the basis of $\mathcal{K}$ as the *proportion* of structures in this set in which θ is true. Finally to remove the dependence on r the actual assigned belief is taken as the limit of these belief values as r tends to infinity, reflecting the idea that r is to be taken 'very large'. Whilst there can, of course, be problems here

[2] And even beyond the linear case see [23].

in demonstrating the existence of this limit Grove et al. show (see in particular [2]) that the method works well and gives acceptable answers and insight in both direct inference and default reasoning for a wide range of contexts and forms of knowledge. Furthermore, the method can be extended, see in particular [1], to incorporate also prior beliefs in $\mathcal{K}$.

Unlike their method, in this paper the identification of 'infinite' with 'finite but very large' is carried out at the level of the (subjective) knowledge base rather than at the level of the raw data/statistics. It is the knowledge base that ultimately tends to infinity rather than the size of the random worlds. Furthermore, as already indicated, the approach taken in this paper (see also [3]) to problem $\mathcal{P}$ (where there is no objective knowledge, just finitely many linear constraints on beliefs) derives from a somewhat different perspective. We have seen that solving $\mathcal{P}$ amounts to picking a probability function to satisfy K. We are interested in solving this problem, how to make this choice, by directly imposing 'common sense' principles on the choice process itself. In the case of propositional knowledge bases such a set of principles was formulated in [20] (see also [17] for a less technical account), and it was shown that picking the maximum entropy solution was the unique such choice process or, as we call it, *inference process*, satisfying these principles. From this point of view then, taking $Bel(\theta) = ME(K)(\theta)$, where $ME(K)$ is the maximum entropy solution of K, can be argued to be *the* 'common sense', or rational, belief to assign to θ.

What we show in this short paper is that this methodology can be extended also to (finite, linear) predicate knowledge bases within this restricted language $\mathcal{L}$ by treating them as the limiting case of propositional knowledge bases and applying 'common sense', i.e. maximum entropy, to these. Interestingly, the recommendation to maximize entropy (and minimize cross-entropy) also arise naturally in the context in the random-worlds approach; see section 6 of [2] for an extended discussion of the close relationship between the random-worlds method and maximum entropy (also early results in [19], where belief constraints are treated as if they arose from a large population of random worlds, and a criticism of this assumption in section 4 of [18]). In consequence, the random-worlds method can be applied to give the same solutions as we are prescribing. However the significance of this paper rests on what it implies for *common sense* when reasoning with first-order beliefs rather than providing an algorithm. In short, the interest is as much in the justification for the answers as in the answers themselves.

To give a particular example of what we have in mind here for addressing problem $\mathcal{P}$ suppose that

$$K_0 = \left\{ Bel(\exists x\, P(x)) = \frac{1}{2} \right\}.$$

Then according to (P3),

$$Bel(\exists x\, P(x)) = \lim_{r\to\infty} Bel\left(\bigvee_{i=1}^{r} P(a_i)\right).$$

Based on this axiom, the idea behind our approach is to iteratively replace sentences $\exists x\, \theta(x)$ in our knowledge base K by $\bigvee_{i=1}^{r} \theta(a_i)$ for some large r, to produce a knowledge base $K^{(r)}$. So for example the knowledge base K_0 above becomes

$$K_0^{(r)} = \left\{ Bel\left(\bigvee_{i=1}^{r} P(a_i)\right) = \frac{1}{2} \right\}.$$

In this way our knowledge base over a (unary) predicate language is transformed into a knowledge base over a propositional language with propositional variables $P_j(a_i)$, for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, t$ and we can now apply an inference process such as ME. If we let r tend to infinity we would hope to attain a well defined limit value so that we could set

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

where $\theta^{(r)}$ etc. is the result of iteratively replacing existential quantifiers by disjunctions of $a_1, a_2, \ldots, a_r$ as described above.

We will now set up a formal framework and show that this approach does always yield a probability function Bel on $\mathcal{L}$ satisfying K.

The Existence of the Limit

Let $\mathcal{L}^k$ be the language $\mathcal{L}$ as above but with only constant symbols $a_1, \ldots, a_k$. Let $Q_1, \ldots, Q_J$, where $J = 2^t$, be an enumeration in some fixed order of the formulae of the form $\pm P_1 \wedge \ldots \wedge \pm P_t$. Let L^r be the *propositional* language with propositional variables $P_j(a_i)$, $i = 1, \ldots, r$, $j = 1, \ldots, t$. For $r > k$, define $()^{(r)} : S\mathcal{L}^k \to SL^r$ inductively as follows. For $\phi, \psi, \exists x\, \psi(x)$ sentences of $\mathcal{L}^k$,

$$\begin{aligned}
P_j(a_i)^{(r)} &= P_j(a_i) \\
(\neg\phi)^{(r)} &= \neg\phi^{(r)} \\
(\phi \vee \psi)^{(r)} &= \phi^{(r)} \vee \psi^{(r)} \\
(\phi \wedge \psi)^{(r)} &= \phi^{(r)} \wedge \psi^{(r)} \\
(\exists x \psi(x))^{(r)} &= \bigvee_{i=1}^{r} \psi(a_i)^{(r)}.
\end{aligned}$$

At this point it will be useful to explicitly note the following consequence of this construction.

Lemma 1 *If $\theta, \phi \in S\mathcal{L}^k$, $k \leq r$ and $\theta \equiv \phi$ (in the predicate calculus) then $\theta^{(r)} \equiv \phi^{(r)}$ (in the propositional calculus).*

Proof. If $\theta \equiv \phi$ then θ and ϕ have the same models so in particular they have the same models with universe $\{ a_1, a_2, \ldots, a_r \}$ over the predicate language $\mathcal{L}^r$ with each a_i, $i = 1, 2, \ldots, r$, interpreted as itself. Clearly in these structures $\theta \leftrightarrow \theta^{(r)}$ and $\phi \leftrightarrow \phi^{(r)}$ must hold. Hence $\theta^{(r)} \leftrightarrow \phi^{(r)}$ holds in all such structures so there cannot be a valuation on SL^r that gives them different truth values, otherwise we could simply use that as for the basis for such a structure for $\mathcal{L}^r$ giving them again these different values. Hence $\theta^{(r)} \equiv \phi^{(r)}$ now in the propositional calculus. ■

Continuing again now with our main theme, let K be a finite satisfiable set of linear constraints on a probability function Bel on $\mathcal{L}$, say K is

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

for some sentences $\theta_1, \theta_2, \ldots, \theta_s$ of $\mathcal{L}$, $a_{ij}, b_i \in \mathbb{R}$, and set $K^{(r)}$ to be the knowledge base obtained by replacing every sentence θ in K by $\theta^{(r)}$. Note that since K is finite there is a bound k on the j such that a_j appears in K so $K^{(r)}$ is well defined, in the sense that the $\theta_j^{(r)} \in SL^r$, for large r.

We now state a result that shows the significant advantage of working with a language like $\mathcal{L}^k$ with only unary predicates and constants. In this lemma, and thereafter, ψ^ϵ, for $\epsilon = 0, 1$, is taken to be ψ if $\epsilon = 1$ and $\neg\psi$ if $\epsilon = 0$, whilst the α_i for $i = 1, 2, \ldots, J^k$ enumerate the exhaustive and exclusive set of sentences of the form

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

Lemma 2 *Any sentence θ of $\mathcal{L}^k$ is equivalent to a disjunction of sentences $\phi_{i\vec{\epsilon}}$ of the form*

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

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

The proof of this lemma is a straightforward adaptation of the proof of a similar theorem given in [11].

Before we can show that our limit does indeed exist we first need to check that the $K^{(r)}$ as defined above are actually satisfiable for large r.

Theorem 3 *If K is a finite, satisfiable, set of linear constraints over $\mathcal{L}$ then $K^{(r)}$ is also satisfiable as a set of constraints over L^r for large r.*

Proof. Suppose $Bel : S\mathcal{L} \to [0,1]$ is a probability function satisfying K and let k be an upper bound on the j such that a_j appears in K. It will suffice to show that for each large r there exists a probability function $Bel^{(r)} : SL^r \to [0,1]$ such that, for all sentences θ of $\mathcal{L}^k$,

$$Bel(\theta) = Bel^{(r)}(\theta^{(r)}).$$

To see the idea behind the proof, suppose first of all that we were in the very simple situation in which $Bel(\phi_{i\vec{\epsilon}}) = 1$ for

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

as in Lemma 2, so in particular $\phi_{i\vec{\epsilon}}$ must be consistent. Now pick an atom,

$$\Phi_{i\vec{\epsilon}} = \bigwedge_{i=1}^{r} Q_{m_i}(a_i)$$

of L^r which extends α_i and has the property that for $j = 1, 2, \ldots, t$ j is represented amongst the $m_1, m_2, \ldots, m_r$ if and only if $\exists x Q_j(x)$ appears positively as a conjunct in $\phi_{i\vec{\epsilon}}$ (i.e. $\epsilon_j = 1$). Provided r is large enough, it is possible to construct such an atom, as we shall shortly demonstrate. It now turns out that the translation $()^{(r)} : S\mathcal{L}^k \to SL^r$ is such that if we define $Bel_{i\vec{\epsilon}}$ on SL^r to give this atom $\Phi_{i\vec{\epsilon}}$ probability 1 then

$$Bel(\phi_{j\vec{\delta}}) = Bel_{i\vec{\epsilon}}(\phi^{(r)}_{j\vec{\delta}}) = \begin{cases} 1 \text{ if } \langle j, \vec{\delta}\rangle = \langle i, \vec{\epsilon}\rangle, \\ 0 \text{ otherwise,} \end{cases}$$

and hence by Lemma 2 $Bel(\theta) = Bel_{i\vec{\epsilon}}(\theta^{(r)})$, as required for all $\theta \in S\mathcal{L}^k$.

Of course we cannot expect in general that Bel will give all the probability to just one $\phi_{i\vec{\epsilon}}$. However we can easily get round this by (effectively) conditioning on $\phi_{i\vec{\epsilon}}$ (when $Bel(\phi_{i\vec{\epsilon}}) > 0$) and then subsequently recombining the $Bel_{i\vec{\epsilon}}$ in the obvious way at the end.

In detail, for $\vec{\epsilon} = \epsilon_1, \epsilon_2, \ldots, \epsilon_J \in \{0,1\}$ not all zero let

$$\Phi_{i\vec{\epsilon}} = \alpha_i \wedge \bigwedge_{i=1}^{r-k} Q_{m_{k+i}}(a_{k+i})$$

where for $1 \leq i \leq r-k$

- $m_{k+i} = i$ if $i \leq J$ and $\epsilon_i = 1$,
- $m_{k+i} = \min\{j \mid \epsilon_j = 1\}$ otherwise.

Notice that these $\Phi_{i\vec{\epsilon}}$ are disjoint for distinct $\langle i, \vec{\epsilon}\rangle$. If $i, \vec{\epsilon}$ are such that $Bel(\phi_{i\vec{\epsilon}}) > 0$ then define $Bel_{i\vec{\epsilon}} : SL^r \to [0,1]$ by

$$Bel_{i\vec{\epsilon}}(\Phi_{i\vec{\epsilon}}) = 1$$

and

$$Bel_{i\vec{\epsilon}}\left(\bigwedge_{i=1}^{r} Q_{h_i}(a_i)\right) = 0$$

if

$$\bigwedge_{i=1}^{r} Q_{h_i}(a_i) \neq \Phi_{i\vec{\epsilon}}.$$

If $Bel(\phi_{i\vec{\epsilon}}) = 0$ just choose $Bel_{i\vec{\epsilon}}$ to be any probability function on SL^r (this also covers the case when all the ϵ_j are zero).

Then

$$\alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i)\right)^{\epsilon_j}$$

is equivalent to a disjunction of sentences of the form

$$\alpha_i \wedge \bigwedge_{i=k+1}^{r} Q_{h_{k+i}}(a_{k+i}),$$

of which only $\Phi_{i\vec{\epsilon}}$ has non-zero probability with respect to $Bel_{i\vec{\epsilon}}$, so we have

$$Bel_{i\vec{\epsilon}}\left(\alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i)\right)^{\epsilon_j}\right) = Bel_{i\vec{\epsilon}}(\Phi_{i\vec{\epsilon}}) = 1. \tag{1}$$

Now we will show that

$$Bel_{i\vec{\epsilon}}\left(\alpha_j \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i)\right)^{\delta_j}\right) = 0 \quad \text{if} \quad \langle j, \vec{\delta}\rangle \neq \langle i, \vec{\epsilon}\rangle. \tag{2}$$

This is clear if $i \neq j$ so assume $i = j$. Then for this sentence to have non-zero belief then we must have

$$\Phi_{i\vec{\epsilon}} \models \alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i)\right)^{\delta_j}. \tag{3}$$

However, if $\vec{\delta} \neq \vec{\epsilon}$ then either ($\delta_j = 0$ and $\epsilon_j = 1$) or ($\delta_j = 1$ and $\epsilon_j = 0$) for some $1 \leq j \leq J$. If $\delta_j = 0$ then

$$\alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\delta_j} \models \bigwedge_{i=1}^{r} \neg Q_j(a_i).$$

But if $\epsilon_j = 1$ then

$$\Phi_{i\vec{\epsilon}} \models Q_j(a_{k+j}).$$

Hence (3) cannot hold. A similar argument shows that (3) cannot hold when $\delta_j = 1$ and $\epsilon_j = 0$ so it must be the case that (2) is true.

Now define $Bel^{(r)} : SL^r \to [0,1]$ by

$$Bel^{(r)}(\theta) = \sum_{i,\vec{\epsilon}} Bel(\phi_{i\vec{\epsilon}}) Bel_{i\vec{\epsilon}}(\theta)$$

By Lemmas 1 and 2 it will suffice to show that

$$Bel^{(r)} \left(\alpha_j \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\delta_j} \right) = Bel \left(\alpha_j \wedge \bigwedge_{j=1}^{J} (\exists x Q_j(x))^{\delta_j} \right).$$

We can now complete the proof as follows.

$$\begin{aligned} Bel^{(r)} \left(\alpha_j \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\delta_j} \right) = & \\ = & \sum_{i,\vec{\epsilon}} Bel(\phi_{i\vec{\epsilon}}) Bel_{i\vec{\epsilon}} \left(\alpha_j \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\delta_j} \right) \\ = & Bel(\phi_{j\vec{\delta}}) Bel_{j\delta} \left(\alpha_j \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\delta_j} \right) \quad \text{by (2)} \\ = & Bel(\phi_{j\vec{\delta}}) \qquad \text{by (1)} \end{aligned}$$

as required. ■

We are now ready to prove the main result of this paper.

Theorem 4 *For $\theta \in S\mathcal{L}$,*

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

exists and is a probability function on $\mathcal{L}$.

Proof. By Lemma 2 every sentence $\theta(a_1, \ldots, a_k) \in S\mathcal{L}$ 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

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

so

$$\phi^{(r)}_{i\vec{\epsilon}} = \alpha_i \wedge \bigwedge_{j=1}^{J} \left(\bigvee_{i=1}^{r} Q_j(a_i) \right)^{\epsilon_j}$$

is equivalent to

$$\bigvee_{\substack{m_j \in P_{\vec{\epsilon}} \text{ for } j=k+1,\ldots,r \\ P_{i\vec{\epsilon}} \subseteq \{m_j \,|\, k+1 \le j \le r\}}} \left(\alpha_i \wedge \bigwedge_{j=k+1}^{r} Q_{m_j}(a_j) \right) \tag{4}$$

i.e. the disjunction of all the atoms of $L^{(r)}$ which logically imply $\phi^{(r)}_{i\vec{\epsilon}}$. Note that each atom logically implies precisely one sentence $\phi^{(r)}_{i\vec{\epsilon}}$. If we let

$$p_{\vec{\epsilon}} = |P_{\vec{\epsilon}}|, \qquad p_{i\vec{\epsilon}} = |P_{i\vec{\epsilon}}|$$

then the number of disjuncts in (4) is

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

(see, for example, exercise 4 page 182 of [8]).

Thus if

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

where Bel is a belief function satisfying the Renaming Principle (such as applying the inference process ME would give, see for example [20] or [16]), then the entropy is

$$\begin{aligned}
E(\vec{x}) &= -\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log \left(\frac{x_{i\vec{\epsilon}}}{\sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} (p_{\vec{\epsilon}} - j)^{r-k}} \right) \\
&= -\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log x_{i\vec{\epsilon}} + \sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log \left(p_{\vec{\epsilon}}^{r-k} \sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} \left(1 - \frac{j}{p_{\vec{\epsilon}}}\right)^{r-k} \right) \\
&= -\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log x_{i\vec{\epsilon}} + (r-k) \sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} \\
&\quad + \sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log \left(\sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} \left(1 - \frac{j}{p_{\vec{\epsilon}}}\right)^{r-k} \right).
\end{aligned}$$

Let

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

so that

$$E(\vec{x}) = -\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log x_{i\vec{\epsilon}} + (r-k) \sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} + \delta(\vec{x}, r). \tag{5}$$

Note that $\delta(\vec{x}, r) \to 0$ as $r \to \infty$ since we are summing a finite number of terms and

$$\left(1 - \frac{j}{p_{\vec{\epsilon}}}\right)^{r-k} \to 0 \qquad \text{as} \qquad r \to \infty$$

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

$$\sum_{j=0}^{p_{i\vec{\epsilon}}} (-1)^j \binom{p_{i\vec{\epsilon}}}{j} \left(1 - \frac{j}{p_{\vec{\epsilon}}}\right)^{r-k} \to 1 \quad \text{for all } i, \vec{\epsilon} \text{ with } x_{i,\vec{\epsilon}} > 0.$$

By Lemma 2 each sentence θ in the knowledge base K is logically equivalent to a disjunction of sentences $\phi_{i\vec{\epsilon}}$ so, by Lemma 1, each $\theta^{(r)}$ in the knowledge base $K^{(r)}$ is similarly equivalent to the corresponding disjunction of sentences $\phi^{(r)}_{i\vec{\epsilon}}$. Thus, if $\vec{x}$ is the vector formed by listing the $x_{i\vec{\epsilon}}$ in some fixed order, then $K^{(r)}$ is equivalent to a system of linear equations $\vec{x}A = \vec{b}$ where the matrix A is independent of r. Let

$$S = \{\vec{x} \,|\, \vec{x}A = \vec{b}\}, \qquad T = \{\vec{x} \in S \,|\, \sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} \text{ is maximal}\}.$$

It can easily be shown that S and T are convex. Let $\vec{X}$ be that point in T for which

$$F(\vec{x}) = -\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}} \log x_{i\vec{\epsilon}}$$

is maximal (unique by the convexity of the set T and function F.) Let

$$x^{(r)}_{i\vec{\epsilon}} = ME(K^{(r)})(\phi^{(r)}_{i\vec{\epsilon}})$$

so that $\vec{x}^{(r)}$ is the point in S for which $E(\vec{x})$ is maximal. In particular then, $E(\vec{x}^{(r)}) \geq E(\vec{X})$ so, by (5)

$$\frac{F(\vec{x}^{(r)}) - F(\vec{X}) + \delta(\vec{x}^{(r)}, r) - \delta(\vec{X}, r)}{r-k} \geq \sum_{i,\vec{\epsilon}} X_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} - \sum_{i,\vec{\epsilon}} x^{(r)}_{i\vec{\epsilon}} \log p_{\vec{\epsilon}}. \tag{6}$$

But $\vec{X} \in T$ so by definition

$$\sum_{i,\vec{\epsilon}} X_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} - \sum_{i,\vec{\epsilon}} x^{(r)}_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} \geq 0. \tag{7}$$

Since the LHS of (6) tends to 0 as $r \to \infty$ we have

$$\sum_{i,\vec{\epsilon}} x_{i\vec{\epsilon}}^{(r)} \log p_{\vec{\epsilon}} \to \sum_{i,\vec{\epsilon}} X_{i\vec{\epsilon}} \log p_{\vec{\epsilon}} \quad \text{as} \quad r \to \infty. \tag{8}$$

Now $\vec{x}^{(r)}$ is a bounded sequence so it must have a convergent subsequence. Furthermore, the limit of any convergent subsequence must be in T by (8). But by (6) and (7) we have

$$F(\vec{x}^{(r)}) \geq F(\vec{X}) - \delta(\vec{x}^{(r)}, r) + \delta(\vec{X}, r).$$

Therefore, if $\vec{l}$ is the limit of some convergent subsequence of $\vec{x}^{(r)}$ then $\vec{l} \in T$ and $F(\vec{l}) \geq F(\vec{X})$. But $\vec{X}$ is the unique point in T which maximizes F so $\vec{l} = \vec{X}$. Since this is true for any convergent subsequence of $\vec{x}^{(r)}$ we must have

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

This shows that the limit stated in the Theorem does exist and, since we could chose k arbitrarily large to start with, that Bel satisfies (P1-2) (since $ME(K^{(r)})$ does and by Lemma 1 $\models \theta$ implies $\models \theta^{(r)}$).

To show (P3) suppose that $\exists x\, \psi(x, a_1, a_2, \ldots, a_k) \in S\mathcal{L}$, so this sentence is equivalent to a disjunction of some $\phi_{i\vec{\epsilon}}$ and

$$Bel(\exists x\, \psi(x, a_1, a_2, \ldots, a_k)) = \lim_{r\to\infty} \sum_{i\vec{\epsilon}} ME(K^{(r)})(\phi_{i\vec{\epsilon}}^{(r)}).$$

Let $s \geq k$. Then $\bigvee_{j=1}^{s} \psi(a_j, a_1, a_2, \ldots, a_k)$ is equivalent to a disjunction of some $\phi_{ij\vec{\epsilon}}$ where *either* the initial existentially quantifier $\exists x$ in $\exists x\, \psi(x, a_1, a_2, \ldots, a_k)$ does not appear in $\phi_{i\vec{\epsilon}}$, and $j = 1$ and $\phi_{ij\vec{\epsilon}} = \phi_{i\vec{\epsilon}}$, *or* this existential quantifier does so appear, as part of the conjunct $\exists x Q_h(x)$ say, (so $\epsilon_h = 1$), and $1 \leq j \leq s$ and $\phi_{ij\vec{\epsilon}}$ is the result of replacing this conjunct in $\phi_{i\vec{\epsilon}}$ by $Q_h(a_j)$. With this notation then

$$Bel(\bigvee_{j=1}^{s} \psi(a_j, a_1, a_2, \ldots, a_k)) = \lim_{r\to\infty} \sum_{ij\vec{\epsilon}} ME(K^{(r)})(\phi_{ij\vec{\epsilon}}^{(r)}).$$

But in this expression for a fixed $i\vec{\epsilon}$ the sum

$$\sum_{j} ME(K^{(r)})(\phi_{ij\vec{\epsilon}}^{(r)})$$

is either $ME(K^{(r)})(\phi_{i\vec{\epsilon}}^{(r)})$ or, by Renaming,

$$C_{sr} \times ME(K^{(r)})(\phi_{i\vec{\epsilon}}^{(r)})$$

where C_{sr} is the proportion of $\langle m_{k+1}, m_{k+2}, \ldots, m_r \rangle$ in the summation at (4) for which $m_j = h$ for some $1 \leq j \leq s$. Clearly by choosing s large enough we can make this C_{sr} as close as we want to 1, independently of r. It follows then that

$$Bel(\exists x\, \psi(x, a_1, a_2, \ldots, a_k)) = \lim_{s \to \infty} Bel(\bigvee_{j=1}^{s} \psi(a_j, a_1, a_2, \ldots, a_k)),$$

as required. ■

An Example

Suppose

$$K = \{Bel(\forall x\, P(x)) = b\}$$

for some $0 < b < 1$. Using the above notation then

$$K^{(r)} = \left\{ Bel(\bigwedge_{i=1}^{r} P(a_i)) = b \right\}.$$

Since ME satisfies Renaming,

$$ME(K^{(r)}) \left(\bigwedge_{i=1}^{r} P(a_i)^{\epsilon_i} \right) = \frac{1-b}{2^r - 1}$$

when $\sum_{i=1}^{r} \epsilon_i < r$. For $\sum_{i=1}^{m} \epsilon_i < m < r$,

$$\begin{aligned} ME(K^{(r)}) \left(\bigwedge_{i=1}^{m} P(a_i)^{\epsilon_i} \right) &= \sum_{\vec{\tau}} ME(K^{(r)}) \left(\bigwedge_{i=1}^{m} P(a_i)^{\epsilon_i} \wedge \bigwedge_{i=m+1}^{r} P(a_i)^{\tau_i} \right) \\ &= 2^{r-m} \left(\frac{1-b}{2^r - 1} \right) \\ &= \frac{1-b}{2^m - 2^{m-r}} \\ &\to \frac{1-b}{2^m} \qquad \text{as } r \to \infty. \end{aligned}$$

Similarly

$$\begin{aligned} ME(K^{(r)}) \left(\bigwedge_{i=1}^{m} P(a_i) \right) &= \sum_{\vec{\tau}} ME(K^{(r)} \left(\bigwedge_{i=1}^{m} P(a_i) \wedge \bigwedge_{i=m+1}^{r} P(a_i)^{\tau_i} \right) \\ &= b + (2^{r-m} - 1) \left(\frac{1-b}{2^r - 1} \right) \\ &\to b + \frac{1-b}{2^m} \qquad \text{as } r \to \infty. \end{aligned}$$

Hence

$$\lim_{r\to\infty} ME(K^{(r)}) \left(\bigwedge_{i=1}^{m} P(a_i)^{\epsilon_i} \right) = b \cdot 0^{m-\sum \epsilon_i} + (1-b)\frac{1}{2^m}.$$

This solution can be thought of as saying that there are two possible 'situations' or knowledge bases, K_1, K_2. In the first of these (which occurs with probability b) the $P(a_i)$ are all certain to hold, in other words

$$K_1 = \{Bel(\forall x\, P(x)) = 1\}.$$

In the second (which occurs with probability $1-b$) the $P(a_i)$ and $\neg P(a_i)$ are equally likely to hold (and the $P(a_i), P(a_j)$ stochastically independent) which amounts to taking $K_2 = \emptyset$.

In a way this strict dichotomy is not what one might have originally expected, namely that this strong possibility that $P(a_i)$ holds for all i (assuming b is reasonable far from zero) would have made $P(a_i)$ more probable even when it was known that, say, $P(a_1)$ failed. But this is not the case. Conditioning on $\neg P(a_1)$ immediately puts one into the second situation where the remaining $P(a_i)$ are stochastically independent with probability $1/2$.

Indeed (as is already well known) if we take $K = \emptyset$ i.e., total absence of any prior knowledge at all, then our prescribed method (and that of [2] to give but one other example) will again lead to treating the $P(a_i)$ in this way. In consequence this probability function will give, for example, the same probability $1/2$ to $P(a_{10})$, say, conditioned on $P(a_i)$ for $i = 1, 2, \ldots, 9$ all holding as it would with no such evidence.[3] Is this the parting of the ways then as far as common sense is concerned?

Arguably not, at least within the context as given. The underlying problem here seems to us to be that we bring prejudices to this solution which we have not incorporated into the original knowledge base. Namely, that predicates such as $P(x)$ are not just random, but have some structure that marks out those objects that satisfy them as somehow similar, in short, that they are 'projectible' [24]. Philosophers have long struggled with the problem of satisfactorily capturing this notion, so perhaps we can forgive maximum entropy its failure to conjure it up from the empty knowledge set alone.

Conclusion

In this note we have suggested a technique for transforming a knowledge base K (possibly featuring quantified sentences) over a predicate language with unary

[3]This probability function was called c^* by Carnap [4], and fell out of his favor because of this immunity to induction.

predicates $P_1, \ldots, P_t$, to a knowledge base $K^{(r)}$ over a propositional language with variables $P_j(a_i)$, $i = 1, \ldots, r$, $j = 1, \ldots t$. We have shown that, for r sufficiently large, if K is satisfiable then so is $K^{(r)}$ and that

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

exists and determines a probability function on the predicate language satisfying K.

This 'technique' was motivated by the idea that a countably infinite universe might be an idealization from a finite, but inestimably large, universe and that the rational beliefs assigned to the infinite universe should therefore be the limits of their finite counterparts. Our main result shows that this idea can be sensibly realized and that the answers so obtained are relatively easy to calculate and explain.

The modest results of this paper suggest a number of possibly more substantial problems, most particularly trying to extend these results, or at least the methodology behind these results, to languages with predicates of higher arity rather than limiting to just unary. To what extent this is possible seems highly problematic. Certainly Theorem 3 fails if we allow in even just binary predicates (since then we can essentially define an unbounded linear ordering), whilst Grove et al in [2], [11] are led to conjecture that within their approach maximum entropy is 'inherently inapplicable once we move beyond unary predicates'. On the other hand, if maximum entropy can no longer be identified with 'common sense' beyond the unary then what, if anything, is it to be replaced by?

Acknowledgements

We would like to thank the referees for their useful and constructive comments.

References

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

[2] F. Bacchus, A.J. Grove, J.Y. Halpern, D. Koller, From statistical knowledge to degrees of belief, *Artificial Intelligence,* 87 (1996) 75-143.

[3] O.W. Barnett, *On the Application of Probabilistic Inference Processes to Predicate Knowledge Bases*, Doctorial Thesis, The University of Manchester, 2006. [See `http://www.maths.man.ac.uk/~jeff/#students`]

[4] R. Carnap, *The Continuum of Inductive Methods*, University of Chicago Press, (1952).

[5] R. Carnap, Replies and systematic expositions, in: *The Philosophy of Rudolf Carnap*, ed. P.A.Schilpp, La Salle, Illinois, Open Court, (1963).

[6] B. de Finetti, *Theory of Probability, Vol. 1*, Wiley, New York, (1974).

[7] H-D. Ebbinghaus, J. Flum, *Finite Model Theory,* Springer, (1999).

[8] P.J. Eccles, *An Introduction to Mathematical Reasoning: Numbers, Sets and Functions,* Cambridge University Press, (2004).

[9] R. Fagin, Probabilities on Finite Models, *J. Symbolic Logic,* 41(1) (1976) 50-58.

[10] H. Gaifman, Concerning measures in first order calculi, *Israel J. of Mathematics* 24 (1964) 1-18.

[11] A.J. Grove, J.Y. Halpern, D. Koller, Random Worlds and Maximum Entropy, *Journal of Artificial Intelligence Research*, 2 (1994) 33-88.

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

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

[14] W.E. Johnson, Probability: The deductive and inductive problems, *Mind* 49 (1932) 409-423.

[15] J.G. Kemeny, A logical measure function, *Journal of Symbolic Logic,* 18(4) (1953) 289-308.

[16] J.B. Paris, *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge University Press, (1994).

[17] J.B. Paris, Common sense and maximum entropy, *Synthese*, 117 (1999) 75-93.

[18] J.B. Paris, On filling-in missing conditional probabilities in causal networks, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 13(3) (2005) 263-280.

[19] J.B. Paris, A. Vencovská, On the applicability of maximum entropy to inexact reasoning, *International Journal of Approximate Reasoning,* 3 (1989) 1-34.

[20] J.B. Paris, A. Vencovská, A note on the inevitability of maximum entropy, *International Journal of Approximate Reasoning,* 4 (3) (1990) 183-224.

[21] J.B. Paris, A. Vencovská, A model of belief, *Artificial Intelligence*, 64 (1993) 197-241.

[22] J.B. Paris, A. Vencovská, In defense of the maximum entropy inference process, *International Journal of Approximate Reasoning*, 17(1) (1997) 77-103.

[23] J.B. Paris, A. Vencovská, Common sense and stochastic independence, in *Foundations of Bayesianism,* eds. D.Corfield & j.Williamson, Kluwer Applied Logic Series, (2001).

[24] W.V.O. Quine, *Natural Kinds in Ontological Reality and other essays*, Columbia University Press, (1969).

[25] D. Scott, P. Krauss, Assigning probabilities to logical formulas, in *Aspects of Inductive Logic*, eds. J.Hintikka & P.Suppes, North-Holland, Amsterdam, (1966) 219-264.

[26] L. Shastri, Default reasoning in semantic networks: a formalization of recognition and inheritance, *Artificial Intelligence,* 39 (1989) 285-355.

[27] J. Williamson, Countable additivity and subjective probability, *British J. for the Philosophy of Science,* 50 (1999) 401-416.