Probabilistic Reasoning in Terminological Logics

Manfred Jaeger
Max-Planck-Institut für Informatik,
Im Stadtwald, D-66123 Saarbrücken

Abstract

In this paper a probabilistic extensions for terminological knowledge representation languages is defined. Two kinds of probabilistic statements are introduced: statements about conditional probabilities between concepts and statements expressing uncertain knowledge about a specific object. The usual model-theoretic semantics for terminological logics are extended to define interpretations for the resulting probabilistic language. It is our main objective to find an adequate modelling of the way the two kinds of probabilistic knowledge are combined in commonsense inferences of probabilistic statements. Cross entropy minimization is a technique that turns out to be very well suited for achieving this end.

1 INTRODUCTION

Terminological knowledge representation languages (concept languages, terminological logics) are used to describe hierarchies of concepts. While the expressive power of the various languages that have been defined (e.g. KL-ONE [BS85] $\mathcal{ALC}$ [SSS91]) varies greatly in that they allow for more or less sophisticated concept descriptions, they all have one thing in common: the hierarchies described are purely qualitative, i.e. only inclusion, equality, or disjointness relations between concepts can be expressed.

In this paper we investigate an extension of terminological knowledge representation languages that incorporate quantitative statements.

A hybrid terminological logic that allows to express both general world knowledge about the relationships between concepts, and information about the nature of individual objects, gives rise to two kinds of quantitative statements: terminological (T-box) axioms may be refined by stating graded or partial subsumption relations, and assertions (A-box statements) can be generalized by allowing to express uncertain knowledge.

Let us illustrate the use of quantitative statements by an example. The following is a simple knowledge base that could be formulated in any concept language:

Example 1.1

$$\begin{array}{lrcll} \text{T-box:} & \mathsf{Flying\_bird} & \subseteq & \mathsf{Bird} & (1)\\ & \mathsf{Antarctic\_bird} & \subseteq & \mathsf{Bird} & (2)\\ \text{A-box:} & \mathit{Opus} & \in & \mathsf{Bird} & (3) \end{array}$$

In this purely qualitative description a lot of information we may possess cannot be expressed. The two subconcepts of Bird that are specified, for instance, are very different with regard to the degree by which they exhaust the superconcept. One would like to make this difference explicit by stating relative weights, or conditional probabilities, for concepts in a manner like

$$\mathrm{P}(\mathsf{Flying\_bird}|\mathsf{Bird}) = 0.95 \quad (4)$$
$$\mathrm{P}(\mathsf{Antarctic\_bird}|\mathsf{Bird}) = 0.01 \quad (5)$$

Also, it may be desirable to express a degree by which the two concepts Antarctic_bird and Flying_bird, which stand in no subconcept- superconcept relation, intersect:

$$\mathrm{P}(\mathsf{Flying\_bird}|\mathsf{Antarctic\_bird}) = 0.2 \quad (6)$$

For the A-box, apart from the certain knowledge $\mathit{Opus} \in \mathsf{Bird}$, some uncertain information may be available, that we should be able to express as well. There may be strong evidence, for example, that *Opus* is in fact an antarctic bird. Hence

$$\mathrm{P}(\mathit{Opus} \in \mathsf{Antarctic\_bird}) = 0.9 \quad (7)$$

could be added to our knowledge base.

It is important to realize that these two kinds of probabilistic statements are of a completely different nature. The former codifies *statistical information* that, generally, will be gained by observing a large number of individual objects and checking their membership of the various concepts. The latter expresses a *degree*

of belief in a specific proposition. Its value most often will be justified only by a subjective assessment of "likelihood".

This dual use of the term "probability" has caused a lot of controversy over what the true meaning of probability is: a measure of frequency, or of subjective belief (e.g. [Jay78]). A comprehensive study of both aspects of the term is [Car50]. More recently, Bacchus and Halpern have developed a probabilistic extension of first-order logic that accommodates both notions of probability [Bac90],[Hal90].

Now that we have stressed the differences in assigning a probability to subsets of a general concept on the one hand, and to assertions about an individual object on the other, we are faced with the question of how these two notions of probability interact: how does a body of statistical information affect our beliefs in assertions about an individual?

Among the first to address this problem was Carnap, who formulated the rule of *direct (inductive) inference* [Car50]: if for an object a it is known that it belongs to a class C, and our statistics say that an element of C belongs to another class D with probability p, then our degree of belief in a's membership of D should be just p. Applied to the statements (1),(3) and (4) of our example, direct inference yields a degree of belief of 0.95 in the proposition $Opus \in \mathsf{Flying\_bird}$.

A generalization of direct inference is *Jeffrey's rule* [Jef65]: if all we know about a, is that it belongs to either of finitely many mutually disjoint classes $\mathrm{C}_1, \ldots, \mathrm{C}_n$, and to each possibility we assign a probability p_i ($\sum_{i=1}^{n} p_i = 1$), if furthermore, the statistical probability for D given C_i is q_i, then our degree of belief for a being in D should be given by

$$\sum_{i=1}^{n} p_i q_i.$$

Bacchus et al. have developed a method to derive degrees of belief for sentences in first-order logic on the basis of first-order and statistical information [BGHK92], [BGHK93]. The technique they use is motivated by direct inference, but is of a far more general applicability. However, it does not allow to derive new subjective beliefs given both subjective and statistical information.

In this paper we develop a formal semantical framework that for terminological logics models the influence of statistical, generic information on the assignment of degrees of belief to specific assertions. In order to do this, we will interpret both kinds of probabilistic statements in one common probability space that essentially consists of the set of concept terms that can be formed in the language of the given knowledge base [1]. Defining all the probability measures on the same probability space allows us to compare the measure assigned to an object a with the generic measure defined by the given statistical information. The most reasonable assignment of a probability measure to a, then, is to choose, among all the measures consistent with the constraints known for a, the one that most closely resembles the generic measure. The key question to be answered, therefore, is how resemblance of probability measures should be measured. We argue that minimizing the cross entropy of the two measures is the appropriate way.

Paris and Vencovská, considering probabilistic inferences very similar in nature to ours, use a different semantical interpretation, which, too, leads them to the minimum cross entropy principle [PV90], [PV92].

Previous work on probabilistic extensions of concept languages was done by Heinsohn and Owsnicki-Klewe [HOK88],[Hei91]. Here the emphasis is on computing new conditional probabilities entailed by the given ones. Formal semantics for the interpretation of probabilistic assertions, which are the main contribution of our work, are not given.

2 SYNTAX

In order to facilitate the exposition of our approach we shall use, for the time being, a very restricted, merely propositional, concept language, which we call $\mathcal{PCL}$. In the last section of this paper an explanation will be given of how the formalism can be extended to more expressive concept languages, notably $\mathcal{ALC}$.

The *concept terms* in our language are just propositional expressions built from a finite set of *concept names* $\mathrm{S_C} = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\}$. The set of concept terms is denoted by $\mathrm{T}(\mathrm{S_C})$. *Terminological axioms* have the form

$$\mathsf{A} \subseteq \mathrm{C} \quad \text{or} \quad \mathsf{A} = \mathrm{C}$$

with $\mathsf{A} \in \mathrm{S_C}$ and $\mathrm{C} \in \mathrm{T}(\mathrm{S_C})$. *Probabilistic terminological axioms* are expressions

$$\mathrm{P}(\mathrm{C}|\mathrm{D}) = p,$$

where C and D are concept terms and $p \in]0,1[$. Finally, we have *probabilistic assertions*

$$\mathrm{P}(a \in \mathrm{C}) = p,$$

where a is an element of a finite set of *object names* $\mathrm{S_O}$, and $p \in [0,1]$.

A knowledge base ($\mathcal{KB}$) in $\mathcal{PCL}$ consists of a set of terminological axioms ($\mathcal{T}$), a set of probabilistic terminological axioms ($\mathcal{PT}$) and a set of probabilistic assertions ($\mathcal{P}_a$) for every object name a:

$$\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup \{\mathcal{P}_a | a \in \mathrm{S_O}\}.$$

[1] Different from [Bac90],[Hal90], for instance, where statistical and propositional probabilities are interpreted by probability measures on domains and sets of worlds, respectively

There is a certain asymmetry in our probabilistic treatment of terminological axioms on the one hand, and assertions on the other. While deterministic assertions were completely replaced by probabilistic ones ($a \in$C has to be expressed by $\mathrm{P}(a \in \mathrm{C}) = 1$), deterministic terminological axioms were retained, and not identified with 0,1-valued probabilistic axioms (which, therefore, are not allowed in $\mathcal{PT}$).

There are several reasons for taking this approach: First, our syntax for probabilistic terminological axioms is very general in that conditional probabilities for arbitrary pairs of concept terms may be specified. Terminological axioms, on the other hand, are generally required (as in our definition) to have only a concept name on their left hand side. Also, in order to make the computation of subsumption with respect to a terminology somewhat more tractable, usually additional conditions are imposed on $\mathcal{T}$ (e.g. that it must not contain cycles) that we would not want to have on $\mathcal{PT}$ (it may be very important, for instance, to be able to specify both P(C|D) and P(D|C)). In essence, it can be said that the non-uniformity of our treatment of deterministic and probabilistic terminological axioms results from our intention to define a probabilistic extension for terminological logics that does not affect the scope and efficiency of standard terminological reasoning in the given logics.

Furthermore, it will be seen that even for actual probabilistic reasoning it proves useful to make use of the deterministic information in $\mathcal{T}$ and the probabilistic information in $\mathcal{PT}$ in two different ways, and it would remain to do so, if both kinds of information were encoded uniformly.

3 SEMANTICS

Our approach to formulating semantics for the language $\mathcal{PCL}$ modifies and extends the usual model-theoretic semantics for concept languages. The terminological axioms $\mathcal{T}$ are interpreted by means of a domain $\mathbf{D}$ and an interpretation function I in the usual way. In order to give meaning to the expressions in $\mathcal{PT}$ and the $\mathcal{P}_a$ ($a \in \mathrm{S_O}$), we first have to specify the probability space on which the probability measures described by these expressions shall be defined.

For this probability space we choose the language itself. That is to say, we take the *Lindenbaum algebra*

$$\mathfrak{A}(\mathrm{S_C}) := ([\mathrm{T}(\mathrm{S_C})], \vee, \wedge, \neg, 0, 1)$$

as the underlying probability space. Here, $[\mathrm{T}(\mathrm{S_C})]$ is the set of equivalence classes modulo logical equivalence in $\mathrm{T}(\mathrm{S_C})$. The operations $\vee$, $\wedge$, and $\neg$ are defined by performing disjunction, conjunction, and negation on representatives of the equivalence classes. We shall use letters C,D,... both for concept terms from $\mathrm{T}(\mathrm{S_C})$, and their equivalence class in $[\mathrm{T}(\mathrm{S_C})]$.

An *atom* in a boolean algebra $\mathfrak{A}$ is an element $\mathrm{A} \neq 0$, such that there is no $\mathrm{A}' \notin \{0, \mathrm{A}\}$ with $\mathrm{A}' \subset \mathrm{A}$ (to be read as an abbreviation for $\mathrm{A}' \wedge \neg \mathrm{A} = 0$). The atoms of $\mathfrak{A}(\mathrm{S_C})$ with $\mathrm{S_C} = \{\mathsf{A}_1, \ldots, \mathsf{A}_n\}$ are just the concept terms of the form $\mathrm{B}_1 \wedge \ldots \wedge \mathrm{B}_n$ with $\mathrm{B}_i \in \{\mathsf{A}_i, \neg \mathsf{A}_i\}$ for $i = 1, \ldots, n$. The set of atoms of $\mathfrak{A}(\mathrm{S_C})$ is denoted by $\mathrm{A}(\mathrm{S_C})$.

Every element of $\mathfrak{A}(\mathrm{S_C})$, then, is (in the equivalence class of) a finite disjunction of atoms.

On $\mathfrak{A}(\mathrm{S_C})$ probability measures may be defined. Recall that $\mu : \mathfrak{A}(\mathrm{S_C}) \to [0, 1]$ is a probability measure iff $\mu(1) = 1$, and $\mu(\mathrm{C} \vee \mathrm{D}) = \mu(\mathrm{C}) + \mu(\mathrm{D})$ for all C,D with $\mathrm{C} \wedge \mathrm{D} = 0$. The set of probability measures on $\mathfrak{A}(\mathrm{S_C})$ is denoted by $\Delta\mathfrak{A}(\mathrm{S_C})$. Note that $\mu \in \Delta\mathfrak{A}(\mathrm{S_C})$ is fully specified by the values it takes on the atoms of $\mathfrak{A}(\mathrm{S_C})$.

The general structure of an interpretation for a vocabulary $\mathrm{S} = \mathrm{S_C} \cup \mathrm{S_O}$ can now be described: a standard interpretation ($\mathbf{D}$,I) for $\mathcal{T}$ will be extended to an interpretation $(\mathbf{D}, \mathrm{I}, \mu, (\nu_a)_{a \in \mathrm{S_O}})$, where $\mu \in \Delta\mathfrak{A}(\mathrm{S_C})$ is the *generic* measure used to interpret $\mathcal{PT}$, and $\nu_a \in \Delta\mathfrak{A}(\mathrm{S_C})$ interprets $\mathcal{P}_a$. Hence, we deviate from the standard way interpretations are defined by not mapping $a \in \mathrm{S_O}$ to an element of the domain, but to a probability measure expressing our uncertain knowledge of a.

What conditions should we impose on an interpretation to be a model of a knowledge base? Certainly, the measures μ and ν_a must satisfy the constraints in $\mathcal{PT}$ and $\mathcal{P}_a$. However, somewhat more is required when we intend to model the interaction between the two kinds of probabilistic statements that takes place in "commonsense" reasoning about probabilities.

The general information provided by $\mathcal{PT}$ leads us to assign degrees of belief to assertions about an object a that go beyond what is strictly implied by $\mathcal{P}_a$.

What, then, are the rules governing this reasoning process? The fundamental assumption in assigning a degree of belief to a's belonging to a certain concept C is to view a as a random element of the domain about which some partial information has been obtained, but that, in aspects that no observation has been made about, behaves like a typical representative of the domain, for which our general statistics apply.

In the case that $\mathcal{P}_a$ contains constraints only about mutually exclusive concepts this intuition leads to Jeffrey's rule: If

$$\mathcal{P}_a = \{\mathrm{P}(a \in \mathrm{C}_i) = p_i \mid i = 1, \ldots, n\},$$

where the C_i are mutually exclusive, and, as may be assumed without loss of generality, exhaustive as well, and $\mu \in \Delta\mathfrak{A}(\mathrm{S_C})$ reflects our general statistical knowledge about the domain, then the probability measure

that interprets a should be defined by

$$\nu_a(\mathrm{C}) := \sum_{i=1}^{n} (p_i \times \mu(\mathrm{C} \mid \mathrm{C}_i)) \quad (\mathrm{C} \in \mathfrak{A}(\mathrm{S_C})).$$

For constraints on not necessarily exclusive concepts we need to find a more general definition for a measure "most closely resembling" the given generic measure μ and satisfying the constraints. Formally, we are looking for a function d that maps every pair (μ, ν) of probability measures on a given (finite) probability space to a real number $d(\mu, \nu) \geq 0$, the "distance" of ν to μ:

$$d : \Delta^n \times \Delta^n \rightarrow \mathbf{R}^{\geq 0},$$

where

$$\Delta^n := \{(x_1, \ldots, x_n) \in [0,1]^n \mid \textstyle\sum_{i=1}^{n} x_i = 1\}$$

denotes the set of probability measures on a probability space of size n.

Given such a d, a subset N of Δ^n and a measure μ, we can then define the set of elements of N that have minimal distance to μ:

$$\pi^d_{\mathrm{N}}(\mu) := \{\nu \in \mathrm{N} \mid d(\mu, \nu) = \inf\{d(\mu, \nu') \mid \nu' \in \mathrm{N}\}\} \quad (8)$$

Three requirements are immediate that have to be met by a distance function d in order to be used for defining the belief measure ν_a most closely resembling the generic μ:

(i) If N is defined by a constraint-set $\mathcal{P}_a$, then $\pi^d_{\mathrm{N}}(\mu)$ is a singleton.

(ii) If $\mu \in \mathrm{N}$, then $\pi^d_{\mathrm{N}}(\mu) = \{\mu\}$.

(iii) If N is defined by a set of constraints on disjoint sets, then $\pi^d_{\mathrm{N}}(\mu)$ is the probability measure obtained by Jeffrey's rule applied to μ and these constraints.

We propose to use the *cross entropy* of two probability measures as the appropriate definition for their distance. For probability measures $\mu = (\mu_1, \ldots, \mu_n)$ and $\nu = (\nu_1, \ldots, \nu_n)$ define:

$$CE(\mu, \nu) := \begin{cases} \displaystyle\sum_{\substack{i=1 \\ \mu_i, \nu_i \neq 0}}^{n} \nu_i \ln \frac{\nu_i}{\mu_i} & \text{if for all } i: \\ & \mu_i = 0 \Rightarrow \nu_i = 0, \\ \infty & \text{otherwise.} \end{cases}$$

This slightly generalizes the usual definition of cross entropy by allowing for 0-components in μ and ν.

Cross entropy often is referred to as a "measure of the distance between two probability measures" [DZ82], or a "measure of information dissimilarity for two probability measures" [Sho86]. These interpretations have to be taken cautiously, however. Note in particular that neither is CE symmetric nor does it satisfy the triangle inequality. All that CE has in common with a metric is positivity:

$$CE(\mu, \nu) \geq 0,$$

where equality holds iff $\mu = \nu$. Hence property (ii) holds for CE. It has been shown that cross entropy satisfies (i) (for any closed and convex set N, provided there is at least one $\nu \in$N with $CE(\mu, \nu) < \infty$), and (iii) as well ([SJ80], [Wen88]). Therefore, we may define for closed and convex N$\subseteq \Delta^n$ and $\mu \in \Delta^n$:

$$\pi_{\mathrm{N}}(\mu) := \begin{cases} \text{the unique} & \text{if } CE(\mu, \nu) < \infty \\ \text{element in } \pi^{CE}_{\mathrm{N}}(\mu) & \text{for some } \nu \in \mathrm{N} \\ \text{undefined} & \text{otherwise.} \end{cases}$$

There are several lines of argument that support the use of cross entropy for forming our beliefs about a on the basis of the given generic μ and a set of constraints.

One is to appeal directly to cross entropy's properties as a measure of information discrepancy, and to argue that our beliefs about a should deviate from the generic measure by assuming as little additional information as possible.

Another line of argument does not focus on the properties of cross entropy directly, but investigates fundamental requirements for a procedure that changes a given probability measure μ to a posterior measure ν in a (closed and convex) set N. Shore and Johnson [SJ80], [SJ83] formulate five axioms for such a procedure (the first one being just our uniqueness condition (i)), and prove that when the procedure satisfies the axioms, and is of the form $\nu = \pi^d_{\mathrm{N}}(\mu)$ for some function d, then d must be equivalent to cross entropy (i.e. must have the same minima).

Paris and Vencovská, in a similar vein, have given an axiomatic justification of the *maximum entropy* principle [PV90], which, when applied to knowledge bases expressing the two types of probabilistic statements in a certain way, yields the same results as minimizing cross entropy [PV92].

With cross entropy as the central tool for the interpretation of $\mathcal{P}_a$, we can now give a complete set of definitions for the semantics of $\mathcal{PCL}$.

Definition 3.1 Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup\{\mathcal{P}_a \mid a \in \mathrm{S_C}\}$ a $\mathcal{PCL}$-knowledge base. We define for $\mu \in \Delta\mathfrak{A}(\mathrm{S_C})$:

- μ is *consistent with* $\mathcal{T}$ iff $\mathcal{T} \models \mathrm{C} = 0 \Rightarrow \mu(\mathrm{C}) = 0$;
- μ is *consistent with* $\mathcal{PT}$ iff $\mathrm{P}(\mathrm{C}|\mathrm{D}) = p \in \mathcal{PT} \Rightarrow \mu(\mathrm{C} \wedge \mathrm{D}) = p \times \mu(\mathrm{D})$;
- μ is *consistent with* $\mathcal{P}_a$ iff $\mathrm{P}(a \in \mathrm{C}) = p \in \mathcal{P}_a \Rightarrow \mu(\mathrm{C}) = p$.

For a given $\mathcal{KB}$, we use the following notation:

$\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C}) :=$
$\quad \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T}\},$
$Gen(\mathcal{KB}) :=$
$\quad \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T} \text{ and } \mathcal{PT}\},$
$Bel_a(\mathcal{KB}) :=$
$\quad \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T} \text{ and } \mathcal{P}_a\}.$

When no ambiguities can arise, we also write Gen (the set of possible generic measures) and Bel_a (the set of possible belief measures for a) for short.

Definition 3.2 Let $\mathrm{S} = \mathrm{S_C} \cup \mathrm{S_O}$ be a vocabulary. A *$\mathcal{PCL}$-interpretation for* S is a triple $(\mathbf{D}, \mathrm{I}, \mu)$, where $\mathbf{D}$ is a set,

$$\mathrm{I} : \mathrm{S_C} \to 2^{\mathbf{D}}, \quad \mathrm{I} : \mathrm{S_O} \to \Delta\mathfrak{A}(\mathrm{S_C}),$$

and $\mu \in \Delta\mathfrak{A}(\mathrm{S_C})$. Furthermore, for all concept terms C with $\mathrm{I(C)} = \emptyset$: $\mu(\mathrm{C}) = 0$ and $\mathrm{I}(a)(\mathrm{C}) = 0$ $(a \in \mathrm{S_O})$ must hold. For $\mathrm{I}(a)$ we also write ν_a.

Definition 3.3 Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup\{\mathcal{P}_a \mid a \in \mathrm{S_O}\}$ be a $\mathcal{PCL}$-knowledge base. Let $(\mathbf{D}, \mathrm{I}, \mu)$ be a $\mathcal{PCL}$-interpretation for the language of $\mathcal{KB}$. We define: $(\mathbf{D}, \mathrm{I}, \mu) \models \mathcal{KB}$ ($(\mathbf{D}, \mathrm{I}, \mu)$ is a *model* of $\mathcal{KB}$) iff

(i) $(\mathbf{D}, \mathrm{I}\restriction \mathrm{S_C}) \models \mathcal{T}$ in the usual sense.

(ii) $\mu \in Gen(\mathcal{KB})$.

(iii) For all $a \in \mathrm{S_O}$: $\pi_{Bel_a(\mathcal{KB})}$ is defined for μ, and $\mathrm{I}(a) = \pi_{Bel_a(\mathcal{KB})}(\mu)$.

Definition 3.4 Let $\mathrm{J} \subseteq [0, 1]$. We write

$$\mathcal{KB} \models \mathrm{P(C|D)} \in \mathrm{J}$$

iff for every $(\mathbf{D}, \mathrm{I}, \mu) \models \mathcal{KB}$: $\mu(\mathrm{C} \mid \mathrm{D}) \in \mathrm{J}$ (if $\mu(\mathrm{D}) = 0$, this is considered true for every J). Also, we use the notation

$$\mathcal{KB} \models \mathrm{P(C|D)} = \mathrm{J}$$

iff $\mathcal{KB} \models \mathrm{P(C|D)} \in \mathrm{J}$, and J is the minimal subset of [0,1] with this property. Analogously, we use $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) \in \mathrm{J}$, and $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$.

According to definition 3.2 we are dealing with probability measures on the concept algebra $\mathfrak{A}(\mathrm{S_C})$. An explicit representation of any such measure, i.e. a complete list of the values it takes on $\mathrm{A}(\mathrm{S_C})$, would always be of size $2^{|\mathrm{S_C}|}$. Fortunately, we usually will not have to actually handle such large representations, though. Since all the probability measures we consider for a specific knowledge base $\mathcal{KB}$ are in $\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C})$, the relevant probability space for models of $\mathcal{KB}$ only consists of those atoms in $\mathrm{A}(\mathrm{S_C})$ whose extensions are not necessarily empty in models of $\mathcal{KB}$:

$$\mathrm{A}(\mathcal{T}) := \{\mathrm{C} \in \mathrm{A}(\mathrm{S_C}) \mid \mathcal{T} \not\models \mathrm{C} = 0\}$$

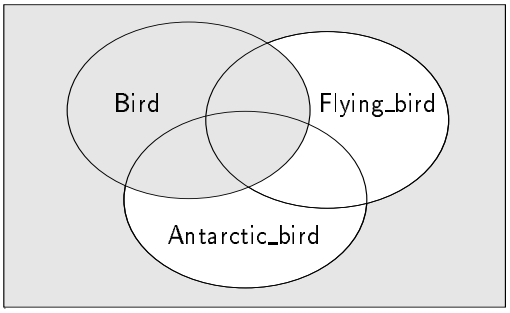

Figure 1: The Algebras $\mathfrak{A}(\mathrm{S_C})$ and $\mathfrak{A}(\mathcal{T})$

Denote the algebra that is generated by these atoms with $\mathfrak{A}(\mathcal{T})$. Technically speaking, $\mathfrak{A}(\mathcal{T})$ is the *relativization* of $\mathfrak{A}(\mathrm{S_C})$ to the element

$$\mathrm{C}(\mathcal{T}) := \bigvee \mathrm{A}(\mathcal{T})$$

of $\mathfrak{A}(\mathrm{S_C})$. Figure 1 shows the structure of $\mathfrak{A}(\mathrm{S_C})$ for the vocabulary of our introductory example. The shaded area represents the element $\mathrm{C}(\mathcal{T})$ for the $\mathcal{T}$ in the example. $\mathfrak{A}(\mathcal{T})$, here, consists of five atoms compared to eight atoms in $\mathfrak{A}(\mathrm{S_C})$.

How much smaller than $\mathfrak{A}(\mathrm{S_C})$ can $\mathfrak{A}(\mathcal{T})$ be expected to be in general? This question obviously is difficult to answer, because it requires a thorough analysis of the structure that $\mathfrak{A}(\mathcal{T})$ is likely to have for real-world instances of $\mathcal{T}$. Here we just mention one property of $\mathcal{T}$ that ensures a non-exponential growth of $|\mathrm{A}(\mathcal{T})|$ when new terminological axioms introducing new concept names are added: call $\mathfrak{A}(\mathcal{T})$ *bounded in depth by k* iff every atom in $\mathrm{A}(\mathcal{T})$ contains at most k non-negated concept names from $\mathrm{S_C}$ as conjuncts. It is easy to see that if $\mathfrak{A}(\mathcal{T})$ is bounded in depth by k, then $|\mathrm{A}(\mathcal{T})|$ will have an order of magnitude of $|\mathrm{S_C}|^k$ at most. Hence, when new axioms are added to $\mathcal{T}$ in such a way that $\mathfrak{A}(\mathcal{T})$ remains bounded in depth by some number k, then the growth of $|\mathrm{A}(\mathcal{T})|$ is polynomial.

The use of the structural information in $\mathcal{T}$ for reducing the underlying probability space from $\mathfrak{A}(\mathrm{S_C})$ to $\mathfrak{A}(\mathcal{T})$ is the second reason for the nonuniform treatment of deterministic and probabilistic terminological axioms that was announced in section 2. If deterministic axioms were treated in precisely the same fashion as probabilistic ones, this would only lead us to handle probability measures all with zeros in the same large set of components, but not to drop these components from our representations in the first place.

Example 3.5 Let $\mathcal{KB}_1$ contain the terminological and probabilistic statements from example 1.1 (the assertion $Opus \in \mathsf{Bird}$ being replaced by $\mathrm{P}(Opus \in \mathsf{Bird}) = 1$). The three statements (4)-(6) in $\mathcal{PT}$ do not determine a unique generic measure μ, but

for every $\mu \in Gen(\mathcal{KB}_1)$

$$\mu(\mathsf{Flying\_bird} \mid \mathsf{Antarctic\_bird}) = 0.2$$
$$\text{and } \mu(\mathsf{Flying\_bird} \mid \mathsf{Bird} \wedge \neg\mathsf{Antarctic\_bird}) = 0.958$$

holds: the first conditional probability is explicitly stated in (6), the second can be derived from (4)-(6) by elementary computations.

Since the constraints in $\mathcal{P}_{Opus}$ are equivalent to $\mathrm{P}(Opus \in \mathsf{Antarctic\_bird}) = 0.9$ and $\mathrm{P}(Opus \in \mathsf{Bird} \wedge \neg\mathsf{Antarctic\_bird}) = 0.1$, and in this case $\pi_{Bel_{Opus}}(\mu)$ is given by Jeffrey's rule,

$$\begin{aligned} \pi_{Bel_{Opus}}(\mu)(\mathsf{Flying\_bird}) &= 0.9 \times 0.2 + 0.1 \times 0.958 \\ &= 0.2758 \end{aligned}$$

holds for every $\mu \in Gen$. Hence

$$\mathcal{KB}_1 \models \mathrm{P}(Opus \in \mathsf{Flying\_bird}) = 0.2758.$$

In the following section we investigate how inferences like these can in general be computed from a $\mathcal{PCL}$-knowledge base.

4 COMPUTING PROBABILITIES

4.1 COMPUTING Gen AND Bel_a

The constraints in $\mathcal{PT}$ and $\mathcal{P}_a$ are linear constraints on $\Delta\mathfrak{A}(\mathrm{S_C})$. When we change the probability space we consider from $\mathfrak{A}(\mathrm{S_C})$ to $\mathfrak{A}(\mathcal{T})$, a constraint of the form $\mathrm{P(C|D)} = p$ is interpreted as

$$\mathrm{P}(\mathrm{C} \wedge \mathrm{C}(\mathcal{T}) | \mathrm{D} \wedge \mathrm{C}(\mathcal{T})) = p.$$

Similarly, $\mathrm{P}(a \in \mathrm{C}) = p$ must be read as $\mathrm{P}(a \in \mathrm{C} \wedge \mathrm{C}(\mathcal{T})) = p$.

If $|\mathrm{A}(\mathcal{T})| = n$, then $\Delta\mathfrak{A}(\mathcal{T})$ is represented by Δ^n. Each of the constraints in $\mathcal{PT}$ or $\mathcal{P}_a$ defines a hyperplane in $\mathbf{R}^n$. $Gen(\mathcal{KB})$ ($Bel_a(\mathcal{KB})$) then, is the intersection of Δ^n with all the hyperplanes defined by constraints in $\mathcal{PT}$ ($\mathcal{P}_a$). Thus, if $\mathcal{PT}$ ($\mathcal{P}_a$) contains k linear independent constraints, $Gen(\mathcal{KB})$ ($Bel_a(\mathcal{KB})$) is a polytope of dimension $\leq n - k$.

Figure 2 shows the intersection of Δ^4 with the two hyperplanes defined by $\{(x_1, x_2, x_3, x_4) \mid x_1 = 0.2(x_1 + x_2)\}$ and $\{(x_1, x_2, x_3, x_4) \mid x_1 = 0.3(x_1 + x_3 + x_4)\}$. The resulting polytope is the line connecting a and b.

A simple algorithm that computes the vertices of the intersection of Δ^n with hyperplanes $\mathrm{H}_1, \ldots, \mathrm{H}_k$ successively computes $\mathrm{P}_i := \Delta^n \cap \mathrm{H}_1 \cap \ldots \cap \mathrm{H}_i$ ($i = 1, \ldots, k$). After each step P_i is given by a list of its vertices. P_{i+1} is obtained by checking for every pair of vertices of P_i, whether they are connected by an edge, and if this is the case, the intersection of this edge with H_{i+1} (if nonempty) is added to the list of vertices of P_{i+1}.

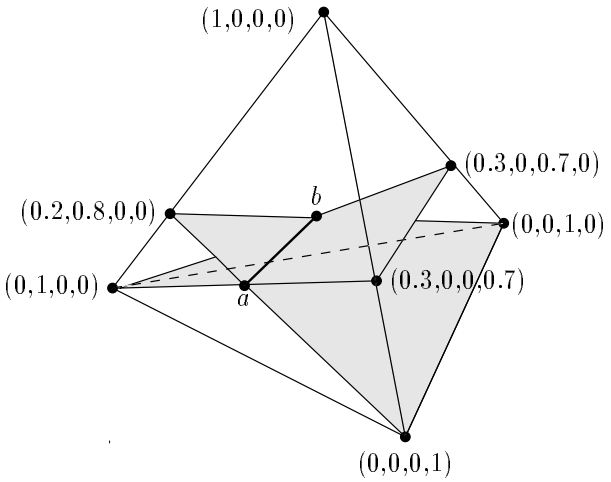

Figure 2: Intersection of Δ^4 With Two Hyperplanes

Example 4.1 The following knowledge base, $\mathcal{KB}_2$, will be used as a running example throughout this section.

$$\mathcal{T}: \quad \mathsf{C} \subseteq \mathsf{A} \wedge \mathsf{B} \tag{9}$$
$$\mathcal{PT}: \quad \mathrm{P}(\mathsf{C}|\mathsf{A}) = 0.1 \tag{10}$$
$$\mathrm{P}(\mathsf{C}|\mathsf{B}) = 0.9 \tag{11}$$
$$\mathcal{P}_a \quad \mathrm{P}(a \in \mathsf{A}) = 0.5 \tag{12}$$
$$\mathrm{P}(a \in \mathsf{B}) = 0.5 \tag{13}$$

The algebra $\mathfrak{A}(\mathcal{T})$ here is generated by the five atoms

$$\mathrm{A}_1 = \neg\mathsf{A} \wedge \neg\mathsf{B} \wedge \neg\mathsf{C}, \quad \mathrm{A}_2 = \neg\mathsf{A} \wedge \mathsf{B} \wedge \neg\mathsf{C},$$
$$\mathrm{A}_3 = \mathsf{A} \wedge \neg\mathsf{B} \wedge \neg\mathsf{C}, \quad \mathrm{A}_4 = \mathsf{A} \wedge \mathsf{B} \wedge \neg\mathsf{C},$$
$$\mathrm{A}_5 = \mathsf{A} \wedge \mathsf{B} \wedge \mathsf{C}.$$

$Gen(\mathcal{KB}_2)$ is the intersection of Δ^5 with the hyperplanes

$$\mathrm{H}_1 = \{(x_1, \ldots, x_5) \mid \tfrac{x_5}{x_3 + x_4 + x_5} = 0.1\} \quad \text{and}$$
$$\mathrm{H}_2 = \{(x_1, \ldots, x_5) \mid \tfrac{x_5}{x_2 + x_4 + x_5} = 0.9\},$$

which is computed to be the convex hull of the three points

$$\mu^0 = (1, 0, 0, 0, 0), \qquad \mu^1 = (0, \tfrac{1}{91}, \tfrac{81}{91}, 0, \tfrac{9}{91}),$$
$$\mu^2 = (0, 0, \tfrac{80}{90}, \tfrac{1}{90}, \tfrac{9}{90}).$$

These probability measures represent the extreme ways in which the partial information in $\mathcal{PT}$ can be completed: μ^0 is the borderline case, where (10) and (11) are vacuously true, because of the probabilities of the conditioning concepts being zero. μ^1 and μ^2, on the contrary, both assign probability 1 to $\mathsf{A} \vee \mathsf{B}$, but represent two opposing hypotheses about the conditional probability of A given B. This probability is 1 for μ^2, standing for the case that B is really a subset of A, and 0.9 for μ^1, representing the possibility that A and B intersect only in C.

The set Bel_a is the convex hull of

$$\nu^0 = (0.5, 0, 0, 0.5, 0), \quad \nu^1 = (0.5, 0, 0, 0, 0.5),$$
$$\nu^2 = (0, 0.5, 0.5, 0, 0).$$

For the remainder of this paper we will assume that Gen and Bel_a are given explicitly by a list of their vertices, because this allows for the easiest formulation of general properties of $\mathcal{PCL}$. Since the number of vertices in Gen and Bel_a can grow very large, it will probably be a more efficient strategy in practice, to just store a suitable normal form of the sets of linear constraints, and to compute specific solutions as needed.

4.2 CONSISTENCY OF A KNOWLEDGE BASE

The first question about a knowledge base $\mathcal{KB}$ that must be asked is the question of consistency: does $\mathcal{KB}$ have a model? Following (i)-(iii) in definition 3.3, we see that $\mathcal{KB}$ is inconsistent iff one of the following statements (a), (b), and (c) holds:

(a) $\mathcal{T}$ is inconsistent.

(b) $Gen(\mathcal{KB}) = \emptyset$.

(c) For all $\mu \in Gen$ there exists $a \in \mathrm{S_O}$ such that $\pi_{Bel_a}(\mu)$ is not defined.

Inconsistency that is due to (a) usually is ruled out by standard restrictions on $\mathcal{T}$: a T-box that does not contain terminological cycles, and in which every concept name appears at most once on the left hand side of a terminological axiom, always has a model. It is trivial to check whether $\mathcal{KB}$ is inconsistent for the reason of $Gen(\mathcal{KB})$ being empty. Also, $\mathcal{KB}$ will be inconsistent if $Bel_a(\mathcal{KB}) = \emptyset$ for some $a \in \mathrm{S_O}$, because in this case $\pi_{Bel_a}(\mu)$ is undefined for every μ.

It remains to dispose of the case where $Gen(\mathcal{KB})$ and all $Bel_a(\mathcal{KB})$ are nonempty, but (c) still holds. By the definition of $\pi_{Bel_a}(\mu)$ this happens iff for all $\mu \in Gen(\mathcal{KB})$ there exists $a \in \mathrm{S_O}$ such that $CE(\mu,\nu) = \infty$ for all $\nu \in Bel_a(\mathcal{KB})$. Since $CE(\mu,\nu)$ is infinite iff for some index i: $\mu_i = 0$ and $\nu_i > 0$, it is the set of 0-components of μ and ν that we must turn our attention to.

Definition 4.2 Let $\mu \in \Delta^n$. Define

$$\mathrm{Z}(\mu) := \{i \in \{1,\ldots,n\} \mid \mu_i = 0\}$$

For a polytope M the notation intM is used for the set of interior points of M; $conv\{\mu^1,\ldots,\mu^k\}$ stands for the convex hull of $\mu^1,\ldots,\mu^k \in \Delta^n$. The next theorem is a trivial observation.

Theorem 4.3 Let $\mathrm{M} \subseteq \Delta^n$ be a polytope and $\mu \in int\mathrm{M}$. Then for every $\mu' \in$M:

$$\mathrm{Z}(\mu) \subseteq \mathrm{Z}(\mu').$$

Particularly, $\mathrm{Z}(\mu')$=$\mathrm{Z}(\mu)$ if $\mu' \in int\mathrm{M}$.

With these provisions we can now formulate a simple test for (c):

Theorem 4.4
Let M=$conv\{\mu^1,\ldots,\mu^k\}$ and N=$conv\{\nu^1,\ldots,\nu^l\}$ be polytopes in Δ^n. Define $\bar{\mu} := \frac{1}{k}(\mu^1 + \ldots + \mu^k)$. Then the following are equivalent:

(i) $\forall \mu \in \mathrm{M}\ \forall \nu \in \mathrm{N}:\ CE(\mu,\nu) = \infty$.

(ii) $\mathrm{Z}(\bar{\mu}) \not\subseteq \mathrm{Z}(\nu^j)$ for $j = 1,\ldots,l$.

Proof: (i) is equivalent to $\mathrm{Z}(\mu) \not\subseteq \mathrm{Z}(\nu)$ for all $\mu \in$M and all $\nu \in$N, which in turn is equivalent to (ii), because by theorem 4.3 $\mathrm{Z}(\bar{\mu})$ is minimal in $\{\mathrm{Z}(\mu) \mid \mu \in \mathrm{M}\}$, and the sets $\mathrm{Z}(\nu^j)$ are maximal in $\{\mathrm{Z}(\nu) \mid \nu \in \mathrm{N}\}$ (i.e. $\forall \nu \in \mathrm{N}\ \exists j \in \{1,\ldots,l\}$ with $\mathrm{Z}(\nu) \subseteq \mathrm{Z}(\nu^j)$). □

Example 4.5 $\mathcal{KB}_2$ is consistent: $\mathcal{T}$ clearly is consistent, $Gen(\mathcal{KB}_2)$ and $Bel_a(\mathcal{KB}_2)$ are nonempty, and $\mathrm{Z}(\bar{\mu}) = \emptyset$ holds for $\bar{\mu} := 1/3(\mu^0 + \mu^1 + \mu^2)$.

4.3 STATISTICAL INFERENCES

Statistical inferences from a knowledge base $\mathcal{KB}$ are computations of sets J for which $\mathcal{KB} \models \mathrm{P(C|D)}$=J.

Definition 4.6 Let $\mathcal{KB}$ be a $\mathcal{PCL}$- knowledge base.

$$Gen^*(\mathcal{KB}) :=$$
$$\{\mu \in Gen(\mathcal{KB}) \mid \forall a \in \mathrm{S_O} : \pi_{Bel_a(\mathcal{KB})}(\mu) \text{ is defined}\}$$

Thus, $Gen^*(\mathcal{KB})$ is the set of generic measures that actually occur in models of $\mathcal{KB}$. $Gen^*(\mathcal{KB})$ is a convex subset of $Gen(\mathcal{KB})$, which, if $\mathcal{KB}$ is consistent, contains at least all the interior points of $Gen(\mathcal{KB})$. If $\mathcal{KB} \models \mathrm{P(C|D)}$=J we then have

$$\mathrm{J} = \{\mu(\mathrm{C} \mid \mathrm{D}) \mid \mu \in Gen^*(\mathcal{KB}),\ \mu(\mathrm{D}) > 0\}.$$

The following theorem, however, states that J can be essentially computed by simply looking at Gen, rather than Gen^*. Essentially here means that the closure of J (clJ) does not depend on the difference $Gen \setminus Gen^*$.

Theorem 4.7
Let $\mathcal{KB}$ be a consistent $\mathcal{PCL}$-knowledge base, C,D $\in \mathrm{T}(\mathrm{S_C})$. Let Gen=conv$\{\mu^1,\ldots,\mu^k\}$, and suppose that $\mathcal{KB} \models \mathrm{P(C|D)} = \mathrm{J}$. Then, either $\mu^i(\mathrm{D}) = 0$ for $i = 1,\ldots,k$ and J=$\emptyset$, or J is a nonempty interval and

$$\inf \mathrm{J} = \min\{\mu^i(\mathrm{C} \mid \mathrm{D}) \mid\ 1 \le i \le k,\ \mu^i(\mathrm{D}) > 0\} \qquad (14)$$

$$\sup \mathrm{J} = \max\{\mu^i(\mathrm{C} \mid \mathrm{D}) \mid 1 \le i \le k,\ \mu^i(\mathrm{D}) > 0\} \qquad (15)$$

Proof: The proof is straightforward. The continuous function $\mu \mapsto \mu(\mathrm{C} \mid \mathrm{D})$ attains its minimal and maximal values at vertices of Gen. From the continuity of this function it follows that for computing the closure of J one can take the minimum and maximum in (14) and (15) over every vertex of Gen, even though they may not all belong to Gen^*. Furthermore, it is easy to see that vertices μ^i with $\mu^i(\mathrm{D}) = 0$ need not be

considered. The details of the proof are spelled out in [Jae94]. □

Applying theorem 4.4 to the face of Gen on which $\mu(\mathrm{C} \mid \mathrm{D}) = \inf \mathrm{J}$ yields a method to decide whether one point of this face is in Gen^*, i.e. whether $\inf \mathrm{J} \in \mathrm{J}$. Analogously for $\sup \mathrm{J}$.

Corollary 4.8 Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT}$ and $\mathcal{KB}' = \mathcal{T}' \cup \mathcal{PT}' \cup \bigcup\{\mathcal{P}'_a \mid a \in \mathrm{S_O}\}$ be two consistent knowledge bases with $\mathcal{T} = \mathcal{T}'$ and $\mathcal{PT} = \mathcal{PT}'$. For C,D$\in$T($\mathrm{S_C}$) let $\mathcal{KB} \models \mathrm{P(C|D)} = \mathrm{J}$ and $\mathcal{KB}' \models \mathrm{P(C|D)} = \mathrm{J}'$. Then $\mathrm{J} = cl\mathrm{J}'$.

By corollary 4.8 the statistical probabilities that can be derived from a consistent knowledge base are essentially independent from the statements about subjective beliefs contained in the knowledge base. The influence of the latter is reduced to possibly removing endpoints from the interval J that would be obtained by considering the given terminological and statistical information only. This is a very reasonable behaviour of the system: generally subjective beliefs held about an individual should not influence our theory about the quantitative relations in the world in general. If, however, we assign a strictly positive degree of belief to an individual's belonging to a set C, then this should preclude models of the world in which C is assigned the probability 0, i.e. C is seen as (practically) impossible. Those are precisely the conditions under which the addition of a set $\mathcal{P}_a$ to a knowledge base will cause the rejection of measures from (the boundary of) Gen for models of $\mathcal{KB}$.

Example 4.9 Suppose we are interested in what $\mathcal{KB}_2$ implies with respect to the conditional probability of C given $\mathsf{A} \wedge \mathsf{B}$, i.e. we want to compute J with

$$\mathcal{KB}_2 \models \mathrm{P}(\mathsf{C} \mid \mathsf{A} \wedge \mathsf{B}) = \mathrm{J}.$$

From $\mu^1(\mathsf{C} \mid \mathsf{A} \wedge \mathsf{B}) = 1$, $\mu^2(\mathsf{C} \mid \mathsf{A} \wedge \mathsf{B}) = 0.9$, and theorem 4.7

$$cl\mathrm{J} = [0.9, 1]$$

immediately follows. Since, furthermore, $\mu^1 \in Gen^*(\mathcal{KB}_2)$, and $\mu(\mathsf{C} \mid \mathsf{A} \wedge \mathsf{B}) = 0.9$ also holds for every $\mu \in int\, conv\{\mu^2, \mu^0\} \subset Gen^*(\mathcal{KB}_2)$, we even have

$$\mathrm{J} = [0.9, 1]. \tag{16}$$

4.4 INFERENCES ABOUT SUBJECTIVE BELIEFS

Probabilistic inferences about subjective beliefs present greater difficulties than those about statistical relations. If $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$, then, by definition 3.3 and 3.4,

$$\mathrm{J} = \{\pi_{Bel_a}(\mu)(\mathrm{C}) \mid \mu \in Gen^*\} =: \pi_{Bel_a}(Gen^*)(\mathrm{C}).$$

Theorem 4.10 If $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$, then J is an interval.

Proof: A simple proof shows that the mapping

$$\pi_{Bel_a} : \Delta^n \to Bel_a$$

is continuous (see [Jae94]). Hence, the codomain $\pi_{Bel}(Gen^*)$ of the connected set Gen^* is connected. Applying another continuous function $\nu \mapsto \nu(\mathrm{C})$ to $\pi_{Bel}(Gen^*)$ yields a subset of [0,1] that again is connected, hence an interval. □

A procedure that computes the sets $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ will certainly have to compute the minimum cross entropy measure $\pi_{Bel_a}(\mu)$ for certain measures μ. This is a nonlinear optimization problem. Generally no closed form solution (like the one given by Jeffrey's rule for the case of constraints on disjoint sets) exists, but an optimization algorithm must be employed to produce a good approximation of the actual solution. There are numerous algorithms available for this problem. See [Wen88] for instance for a C-program, based on an algorithm by Fletcher and Reeves ([FR64]) that implements a nonlinear optimization procedure for cross entropy minimization.

The greatest difficulty encountered when we try to determine $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ does not lie in individual computations of $\pi_{Bel_a}(\mu)$, but in the best choices of μ for which to compute $\pi_{Bel_a}(\mu)$. Unlike the case of statistical inferences, it does not seem possible to give a characterization of $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ in terms of a finite set of values $\pi_{Bel_a}(\mu^i)(\mathrm{C})$ for a distinguished set $\{\mu^1, \ldots, \mu^k\} \subset Gen$.

At present we cannot offer an algorithm for computing $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ any better than by using a search algorithm in Gen^* based on some heuristics, and yielding increasingly good approximations of $\pi_{Bel_a}(Gen^*)(\mathrm{C})$. Such a search might start with elements μ of Gen^* that are themselves maximal (minimal) with respect to $\mu(\mathrm{C})$, and then proceed within Gen^* in a direction in which values of $\pi_{Bel_a}(\cdot)(\mathrm{C})$ have been found to increase (decrease), or which has not been tried yet. The maximal (minimal) values of $\pi_{Bel_a}(\cdot)(\mathrm{C})$ found so far can be used as a current approximation of $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ at any point in the search. The search may stop when a certain number of iterations did not produce any significant increase (decrease) for these current bounds.

Obviously, the complexity of such a search depends on the dimension and the number of vertices of Gen. The cost of a single computation of π_{Bel_a} depends on the size of the probability space $\mathfrak{A}(\mathcal{T})$ and the number of constraints in $\mathcal{P}_a$. In the following we show that the search-space Gen^* can often be reduced to a substantially smaller space.

We show that the interval $\pi_{Bel_a}(Gen^*)(\mathrm{C})$ only depends on the restrictions of the measures in Gen^* and Bel_a to the probability space generated by C and the concepts that appear in $\mathcal{P}_a$.

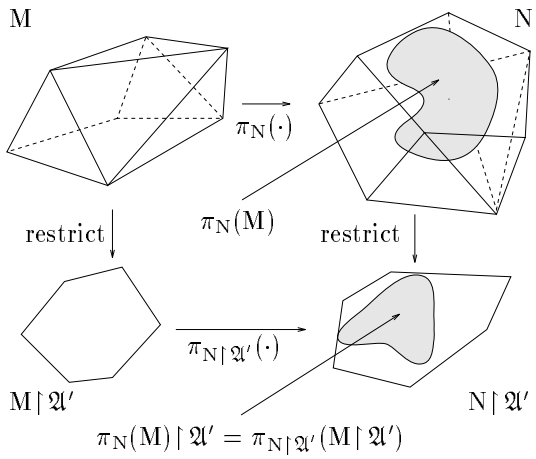

Figure 3: Theorem 4.12

Definition 4.11 Let $\mathfrak{A}$ be an algebra and $\mathfrak{A}'$ a subalgebra of $\mathfrak{A}$. Let $\mu \in \Delta\mathfrak{A}$ and $M \subseteq \Delta\mathfrak{A}$. $\mu\restriction\mathfrak{A}'$ then denotes the restriction of μ to $\mathfrak{A}'$, and $M\restriction\mathfrak{A}'$ is the set $\{\mu\restriction\mathfrak{A}' \mid \mu \in M\}$.

Theorem 4.12 Let $\mathfrak{A}'$ be a subalgebra of the finite algebra $\mathfrak{A}$ generated by a partition $A'_1, \ldots, A'_k$ of $\mathfrak{A}$. Let M,N$\subseteq \Delta\mathfrak{A}$, where N is defined by a set of constraints on $\mathfrak{A}'$, i.e.

$$N = \{\nu \in \Delta\mathfrak{A} \mid \nu(C_i) = p_i,\ i = 1, \ldots, l\}$$

for some $p_i \in [0,1]$ and $C_i \in \mathfrak{A}'$. Then:

$$\pi_N(M)\restriction\mathfrak{A}' = \pi_{N\restriction\mathfrak{A}'}(M\restriction\mathfrak{A}').$$

Furthermore, for every C$\in \mathfrak{A}$ and $\mu \in$ M:

$$\pi_N(\mu)(C) = \sum_{i=1}^{k} \pi_{N\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')(A'_i)\mu(C \mid A'_i).$$

Figure 3 illustrates the first part of the theorem.

Proof: The theorem is contained in [SJ81] in a version for probability measures given by density functions, from which the discrete version can be derived. A direct proof for the discrete case on a more elementary level than the one given in [SJ81] is contained in [Jae94]. □

Theorem 4.12 also plays a vital role in the generalization of $\mathcal{PCL}$ to a probabilistic version of $\mathcal{ALC}$ which we turn to in the next section.

Example 4.13 We conclude our discussion of $\mathcal{KB}_2$ by looking at its implications with respect to $P(a \in \mathsf{C})$. Unlike in our previous example 3.5, the probabilistic information about a in $\mathcal{KB}_2$ does not refer to disjoint concepts, so that here Jeffrey's rule can not be used, and cross entropy minimization in its general form must be put to work.

The information about the likelihood for a being in C is particularly ambiguous: the conditional probabilities of C given the two reference classes A and B, that a may belong to with equal probability, are very dissimilar, thereby providing conflicting default information. Also, the generic probability $\mu(\mathsf{A} \vee \mathsf{B})$ covers the whole range [0,1] for $\mu \in Gen$. Since assigning a value to $P(a \in \mathsf{A} \vee \mathsf{B})$ (which, given the other information in $\mathcal{P}_a$, is equivalent to making up one's mind about $P(a \in \mathsf{A} \wedge \mathsf{B})$) is an important intermediate step for a reasonable estimate of $P(a \in \mathsf{C})$, and the result of this step depends on the prior value $\mu(\mathsf{A} \vee \mathsf{B})$, this is another reason why it is difficult to propose any narrow interval as appropriate for $P(a \in \mathsf{C})$.

It does not come as a surprise, therefore, that no bounds for $P(a \in \mathsf{C})$ can be derived from $\mathcal{KB}_2$ apart from those that directly follow from $\mathcal{P}_a$: from the information in $\mathcal{P}_a$ alone

$$\mathcal{KB}_2 \models P(a \in \mathsf{C}) \in [0, 0.5]$$

is obtained. These bounds can not be substantially improved as computations of $\pi_{Bel_a(\mathcal{KB}_2)}(\mu^\lambda)(\mathsf{C})$ with $\mu^\lambda := \lambda\mu^1 + (1-\lambda)\mu^0$ (with μ^0, μ^1 as in example 4.1) for some $\lambda \in]0,1]$ show. For $\lambda = 1$, $\pi_{Bel_a(\mathcal{KB}_2)}(\mu^1)(\mathsf{C})$ is just $\nu^2(\mathsf{C}) = 0$, ν^2 being the only measure in $Bel_a(\mathcal{KB}_2)$ with finite cross entropy with respect to μ^1. With decreasing λ, $\pi_{Bel_a(\mathcal{KB}_2)}(\mu^\lambda)(\mathsf{C})$ is found to increase, having, for example, the value 0.495 at λ=0.001. Hence, $\mathcal{KB}_2 \models P(a \in \mathsf{C}) = J$ for an interval J with

$$[0, 0.495] \subseteq J \subseteq [0, 0.5].$$

Looking at this result may arouse the suspicion that the whole process of cross entropy minimization really is of little avail, because in the end almost every possible belief measure for a will be in the codomain of $\pi_{Bel_a}(Gen)$. While this can certainly happen, one should not adopt too pessimistic a view based on the current example, where the poor result can really be blamed on the ambiguity of the input. If, for instance, (13) was removed from $\mathcal{KB}_2$, thereby obtaining a smaller knowledge base $\mathcal{KB}'_2$, then the much stronger inference

$$\mathcal{KB}'_2 \models P(a \in \mathsf{C}) = 0.5 \times 0.1 = 0.05$$

could be made. If, on the other hand, $\mathcal{KB}''_2$ is defined by adding

$$P(a \in \mathsf{A} \wedge \mathsf{B}) = 0.25 \qquad (17)$$

to $\mathcal{KB}_2$, then

$$\mathcal{KB}''_2 \models P(a \in \mathsf{C}) = [0.25 \times 0.9, 0.25] = [0.225, 0.25]$$

by our previous result (16).

5 A PROBABILISTIC VERSION OF $\mathcal{ALC}$

5.1 ROLE QUANTIFICATION

The probabilistic concept language $\mathcal{PCL}$ we have described so far does not supply some of the concept-

forming operations that are common to standard concept languages. Most notably, role quantification was not permitted in $\mathcal{PCL}$. In this section we show how the formalism developed in the previous sections can be generalized to yield probabilistic extensions for more expressive languages. Our focus, here, will be on $\mathcal{ALC}$, but the results obtained for this language equally apply to other concept languages.

In $\mathcal{ALC}$ the concept-forming operations of section 2 are augmented by role quantification: the vocabulary now contains a set $\mathrm{S_R} = \{\mathsf{r}, \mathsf{s}, \ldots\}$ of role names in addition to, and disjoint from, $\mathrm{S_C}$ and $\mathrm{S_O}$. New concept terms can be built from a role name r and a concept term C by *role quantification*

$$\forall \mathsf{r} : \mathrm{C} \quad \text{and} \quad \exists \mathsf{r} : \mathrm{C}.$$

The set of concept terms constructible from $\mathrm{S_C}$ and $\mathrm{S_R}$ via the boolean operations and role quantification is denoted $\mathrm{T}(\mathrm{S_C},\mathrm{S_R})$. This augmented set of concept terms together with the syntax rules for terminological axioms, probabilistic terminological axioms, and probabilistic assertions from section 2 yields a probabilistic extension of $\mathcal{ALC}$ which, unsurprisingly, we call $\mathcal{PALC}$. Note that probabilistic assertions of the form $\mathrm{P}((a, b) \in \mathsf{r}) = p$ are not included in our syntax.

Example 5.1 Some pieces of information relating the world of birds and fish are encoded in the following $\mathcal{PALC}$-knowledge base $\mathcal{KB}_3$.

$$\begin{array}{ll} \mathcal{T}: & \mathsf{Herring} \subseteq \mathsf{Fish} \\ & \mathsf{Penguin} \subseteq \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Herring} \\ \mathcal{PT}: & \mathrm{P}(\mathsf{Penguin} | \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Herring}) = 0.2 \\ \mathcal{P}_{Opus}: & \mathrm{P}(\mathit{Opus} \in \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}) = 1 \end{array}$$

The presence of quantification over roles in this knowledge base does not prevent us from forming a subjective degree of belief for the proposition $\mathit{Opus} \in \mathsf{Penguin}$: Since $\forall \mathsf{feeds\_on} : \mathsf{Herring}$ is subsumed by $\forall \mathsf{feeds\_on} : \mathsf{Fish}$, we know that the conditional probability of $\mathsf{Penguin}$ given $\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}$ must lie in the interval [0,0.2], but no better bounds can be derived from $\mathcal{T} \cup \mathcal{PT}$. *Opus* is only known to belong to $\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}$, so that we would conclude that the likelihood for this individual actually being a penguin is in [0,0.2] as well.

This example indicates that probabilistic reasoning within the richer language $\mathcal{PALC}$ works in very much the same way as in $\mathcal{PCL}$. In the following section it is shown how the semantics for $\mathcal{PCL}$ can be generalized to capture this kind of reasoning in $\mathcal{PALC}$.

5.2 PROBABILISTIC SEMANTICS FOR $\mathcal{PALC}$

Central to our semantics for the language $\mathcal{PCL}$ were the concepts of the Lindenbaum algebra $\mathfrak{A}(\mathrm{S_C})$ and of the cross entropy of probability measures on this algebra.

The Lindenbaum algebra for $\mathcal{PALC}$ can be defined in precisely the same manner as was done for $\mathcal{PCL}$. The resulting algebra $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ is quite different from $\mathfrak{A}(\mathrm{S_C})$ however: not only is it infinite, it also is nonatomic, i.e. there are infinite chains $\mathrm{C}_0 \supset \mathrm{C}_1 \supset \ldots$ in $[\mathrm{T}(\mathrm{S_C},\mathrm{S_R})]$ with $\mathrm{C}_i \neq \mathrm{C}_{i+1} \neq 0$ for all i.

The set of probability measures on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ is denoted $\Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$. Probability measures, here, are still required to only satisfy finite additivity. $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ not being closed under infinite disjunctions, there is no need to consider countable additivity. Observe that even though $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ is a countable algebra, probability measures on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ can not be represented by a sequence $(p_i)_{i\in\mathbf{N}}$ of probability values with $\sum_{i\in\mathbf{N}} p_i = 1$ (i.e. a discrete probability measure), because these p_i would have to be the probabilities of the atoms in $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$.

Replacing $\mathfrak{A}(\mathrm{S_C})$ with $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ and $\Delta\mathfrak{A}(\mathrm{S_C})$ with $\Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ definitions 3.1 and 3.2 can now be repeated almost verbatim for $\mathcal{PALC}$ (with the additional provision in definition 3.2 that role names are interpreted by binary relations on $\mathbf{D}$).

So, things work out rather smoothely up to the point where we have to define what it means for a $\mathcal{PALC}$- interpretation to be a model of a $\mathcal{PALC}$ knowledge base. In the corresponding definition for $\mathcal{PCL}$ (definition 3.3) cross entropy played a prominent role. When we try to adopt the same definition for $\mathcal{PALC}$ we are faced with a problem: cross entropy is not defined for probability measures on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$. While we may well define cross entropy for measures that are either discrete, or given by a density function on some common probability space, measures on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ do not fall into either of these categories. Still, in example 5.1 some kind of minimum cross entropy reasoning (in the special form of direct inference) has been employed. This has been possible, because far from considering the whole algebra $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, we only took into account the concept terms mentioned in the knowledge base in order to arrive at our conclusions about $\mathrm{P}(\mathit{Opus} \in \mathsf{Penguin})$. The same principle will apply for any other, more complicated knowledge base: when it only contains the concept terms $\mathrm{C}_1, \ldots, \mathrm{C}_n$, and we want to estimate the probability for $\mathrm{P}(a \in \mathrm{C}_{n+1})$, then we only need to consider probability measures on the finite subalgebra of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ generated by $\{\mathrm{C}_1, \ldots, \mathrm{C}_{n+1}\}$.

The following definition and theorem enables us to recast this principle into formal semantics for $\mathcal{PALC}$.

Definition 5.2 Let $\mathfrak{A}'$ be a finite subalgebra of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ with $\{\mathrm{A}'_1, \ldots, \mathrm{A}'_k\}$ the set of its atoms. Let $\mathrm{N} \subseteq \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be defined by a set of constraints on $\mathfrak{A}'$ (cf. theorem 4.12). Let $\mu \in \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ such that $\pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')$ is defined. For every $\mathrm{C} \in \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ define

$$\pi^*_{\mathrm{N}}(\mu)(\mathrm{C}) := \sum_{i=1}^{k} \pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')(\mathrm{A}'_i)\mu(\mathrm{C} \mid \mathrm{A}'_i).$$

Clearly, $\pi^*_{\mathrm{N}}(\mu)$ is a probability measure on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$. The following theorem shows that $\pi^*_{Bel}(\mu)$ realizes cross entropy minimization for every finite subalgebra of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ containing the concepts used to define Bel.

Theorem 5.3 Let $\mu \in \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, let $Bel \subseteq \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be defined by a finite set of constraints

$$\{\mathrm{P}(\mathrm{C}_i) = p_i \mid p_i \in [0,1],\ \mathrm{C}_i \in \mathfrak{A}(\mathrm{S_C},\mathrm{S_R}),\ i = 1,\ldots,n\}.$$

Let $\mathfrak{A}'$ be the finite subalgebra generated by $\{\mathrm{C}_1,\ldots,\mathrm{C}_n\}$, and assume that $\pi_{Bel\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')$ is defined. Then, for every finite $\mathfrak{A}^* \supseteq \mathfrak{A}'$: $\pi_{Bel\restriction\mathfrak{A}^*}(\mu\restriction\mathfrak{A}^*)$ is defined and equal to $\pi^*_{Bel}(\mu)\restriction\mathfrak{A}^*$.

Proof: Substituting $\mathfrak{A}^*$ for $\mathfrak{A}$, $\{\mu\restriction\mathfrak{A}^*\}$ for M, and $Bel\restriction\mathfrak{A}^*$ for N in theorem 4.12 gives

$$\pi_{Bel\restriction\mathfrak{A}^*}(\mu\restriction\mathfrak{A}^*)(\mathrm{C}) = \sum_{i=1}^{k} \pi_{Bel\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')(\mathrm{A}'_i)\mu(\mathrm{C} \mid \mathrm{A}'_i)$$

for every $\mathrm{C}\in\mathfrak{A}^*$. The right hand side of this equation is just the definition of $\pi^*_{Bel}(\mu)(\mathrm{C})$. □

With $\pi^*_{Bel}(\mu)$ as the measure that, in a generalized way, minimizes cross entropy with respect to μ in Bel, it is now straightforward to define when a $\mathcal{PALC}$-interpretation $(\mathbf{D},\mathrm{I},\mu)$ shall be a model of a $\mathcal{PALC}$-knowledge base $\mathcal{KB}$: just replace $\pi_{Bel_a(\mathcal{KB})}(\mu)$ with $\pi^*_{Bel_a(\mathcal{KB})}(\mu)$ in the corresponding definition for $\mathcal{PCL}$ (definition 3.3).

Probabilistic inferences from a $\mathcal{PALC}$-knowledge base $\mathcal{KB}$ can now be made in basically the same manner as in $\mathcal{PCL}$: to answer a query about a conditional probability P(C|D) for two concepts, consider the algebra generated by C, D, and the concept terms appearing in $\mathcal{PT}$. Call this algebra $\mathfrak{M}_{\mathrm{C,D}}$. The relativized algebra $\mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})$ is defined as above, and $Gen\restriction\mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})$ can be computed as in section 4.1. Theorem 4.7, applied to $Gen\restriction\mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})$ can then be used to compute the J with $\mathcal{KB} \models \mathrm{P}(\mathrm{C}|\mathrm{D}) = \mathrm{J}$.

When J with $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$ shall be computed, the relevant algebra to be considered is generated by C and the concept terms appearing in $\mathcal{P}_a$. Writing $\mathfrak{N}_{a,\mathrm{C}}$ for this algebra,

$$\mathrm{J} = \{\pi_{Bel_a\restriction\mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})}(\mu)(\mathrm{C}) \mid \mu \in Gen\restriction\mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})\}$$

then holds. Note that $Gen\restriction\mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})$ can not be computed directly in the manner described in section 4.1, because Gen will usually be defined by constraints on concepts not all contained in $\mathfrak{N}_{a,\mathrm{C}}$. One way to obtain a representation for $Gen\restriction\mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})$ is to first compute $Gen\restriction\mathfrak{B}(\mathcal{T})$, with $\mathfrak{B}$ the algebra generated by C and the concept terms appearing in either $\mathcal{PT}$ or $\mathcal{P}_a$, and then restrict the result to $\mathfrak{N}_{a,\mathrm{C}}$.

Example 5.4 Suppose we want to determine J_0 with

$$\mathcal{KB}_3 \models \mathrm{P}(\mathsf{Penguin}|\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}) = \mathrm{J}_0.$$

This query and $\mathcal{PT}$ together contain three different concept terms which generate an algebra $\mathfrak{M}$ whose relativization by $\mathcal{T}$ contains just the four atoms

$$\begin{array}{ll} \mathrm{A}_1: & \mathsf{P}, \\ \mathrm{A}_2: & \mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{H} \wedge \neg \mathsf{P}, \\ \mathrm{A}_3: & \mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{F} \wedge \neg \forall \mathsf{f\_o} : \mathsf{H}, \\ \mathrm{A}_4: & \neg(\mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{F}) \end{array}$$

(using suitable abbreviations for the original names). $Gen\restriction\mathfrak{M}(\mathcal{T})$ then is defined by

$$\{(\mu_1,\ldots,\mu_4) \in \Delta^4 \mid \frac{\mu_1}{\mu_1+\mu_2} = 0.2\}.$$

The value for $\mu_1/(\mu_1+\mu_2+\mu_3)$, representing $\mathrm{P}(\mathsf{P}|\mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{F})$, ranges over the interval [0,0.2] in this set, so the answer to our query is the expected interval.

To compute J_1 with

$$\mathcal{KB}_3 \models \mathrm{P}(\mathit{Opus} \in \mathsf{P}) = \mathrm{J}_1,$$

we consider the even smaller algebra $\mathfrak{N}(\mathcal{T})$ consisting of the atoms

$$\begin{array}{lll} \mathrm{B}_1 : \mathsf{P}, & & \mathrm{B}_2 : \neg\mathsf{P} \wedge \mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{F}, \\ \mathrm{B}_3 : \neg(\mathsf{B} \wedge \forall \mathsf{f\_o} : \mathsf{F}). & & \end{array}$$

$Bel_{Opus}\restriction\mathfrak{N}(\mathcal{T})$ then is

$$\{(\nu_1,\nu_2,\nu_3) \in \Delta^3 \mid \nu_1 + \nu_2 = 1\}.$$

It is easy to see that

$$Gen\restriction\mathfrak{N}(\mathcal{T}) = \{(\mu_1,\mu_2,\mu_3) \in \Delta^3 \mid \frac{\mu_1}{\mu_1+\mu_2} \leq 0.2\}.$$

For every $\mu = (\mu_1,\mu_2,\mu_3) \in Gen\restriction\mathfrak{N}(\mathcal{T})$, $(\nu_1,\nu_2,\nu_3) := \pi_{Bel\restriction\mathfrak{N}(\mathcal{T})}(\mu)$ is defined by Jeffrey's rule, so that $\nu_1 = \nu_1/(\nu_1+\nu_2) = \mu_1/(\mu_1+\mu_2)$. Hence,

$$\begin{array}{rcl} \mathrm{J}_1 & = & \{\nu_1 \mid (\nu_1,\nu_2,\nu_3) \in \pi_{Bel\restriction\mathfrak{N}(\mathcal{T})}(Gen\restriction\mathfrak{N}(\mathcal{T}))\} \\ & = & [0,0.2] \end{array}$$

in accordance with our intuitive reasoning in example 5.1.

6 CONCLUDING REMARKS

The semantics we have given to probabilistic extensions of terminological logics are designed for soundness rather than for inferential strength. Allowing any generic measure μ consistent with the constraints to be used in a model is the most cautious approach that can be taken. In cases where it seems more desirable to always derive unique values for probabilities P(C|D) or $\mathrm{P}(a \in \mathrm{C})$ instead of intervals, this approach can be modified by using the maximum entropy measure in

Gen only (as the one most reasonable generic measure).

Generalizations of the formalism here presented are possible in various directions. It could be permitted, for instance, to also state subjective degrees of belief for expressions of the form $(a, b) \in \mathsf{r}$. Since these establish a connection between a and b, it will then no longer be possible to interpret a and b by individual probability measures on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$. Rather, for a language containing object names $\{a_1, \ldots, a_n\}$, a joint probability measure $\nu_{a_1 \ldots a_n}$ on the Lindenbaum algebra of all n-ary expressions constructible from $\mathrm{S_C} \cup \mathrm{S_R}$ will have to be used.

References

[Bac90] F. Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. MIT Press, 1990.

[BGHK92] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller. From statistics to beliefs. In *Proc. of National Conference on Artificial Intelligence (AAAI-92)*, 1992.

[BGHK93] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller. Statistical foundations for default reasoning. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1993.

[BS85] R.J. Brachmann and Schmolze. An overview of the kl-one knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[Car50] R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, 1950.

[DZ82] P. Diaconis and S.L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982.

[FR64] R. Fletcher and C.M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7:149–154, 1964.

[Hal90] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[Hei91] J. Heinsohn. A hybrid approach for modeling uncertainty in terminological logics. In R.Kruse and P.Siegel, editors, *Proceedings of the 1st European Conference on Symbolic an Quantitative Approaches to Uncertainty*, number 548 in Springer Lecture Notes in Computer Science, 1991.

[HOK88] J. Heinsohn and B. Owsnicki-Klewe. Probabilistic inheritance and reasoning in hybrid knowledge representation systems. In W. Hoeppner, editor, *Proceedings of the 12th German Workshop on Artificial Intelligence (GWAI-88)*, 1988.

[Jae94] M. Jaeger. A probabilistic extension of terminological logics. Technical Report MPI-I-94-208, Max-Planck-Institut für Informatik, 1994.

[Jay78] E.T. Jaynes. Where do we stand on maximum entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, 1978.

[Jef65] R.C. Jeffrey. *The Logic of Decision*. McGraw-Hill, 1965.

[PV90] J.B. Paris and A. Vencowská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[PV92] J.B. Paris and A. Vencowská. A method for updating that justifies minimum cross entropy. *International Journal of Approximate Reasoning*, 7:1–18, 1992.

[Sho86] J.E. Shore. Relative entropy, probabilistic inference, and ai. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*. Elsevier, 1986.

[SJ80] J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26(1):26–37, 1980.

[SJ81] J.E. Shore and R.W. Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, IT-27(4):472–482, 1981.

[SJ83] J.E. Shore and R.W. Johnson. Comments on and correction to "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". *IEEE Transactions on Information Theory*, IT-29(6):942–943, 1983.

[SSS91] M. Schmidt-Schauß and G. Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48(1):1–26, 1991.

[Wen88] W.X. Wen. Analytical and numerical methods for minimum cross entropy problems. Technical Report 88/26, Computer Science, University of Melbourne, 1988.