A Probabilistic Extension of Terminological Logics

Manfred Jaeger

MPI–I–94–208 March 1994

Author's Address

Manfred Jaeger

Max-Planck-Institut für Informatik
Im Stadtwald
66123 Saarbrücken
Germany
jaeger@mpi-sb.mpg.de

Abstract

In this report we define a probabilistic extension for a basic terminological knowledge representation languages. 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 modeling of the way the two kinds of probabilistic knowledge are combined in what we call default reasoning about probabilities. Cross entropy minimization is a technique that turns out to be a very promising tool towards achieving this end.

Contents

1 Introduction

Terminological knowledge representation languages (concept languages, terminological logics) are used to describe hierarchies of concepts [WS92]. 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.

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

$$\begin{array}{lrcl} \text{T-box:} & \mathsf{Flying\_bird} & \subseteq & \mathsf{Bird} \\ & \mathsf{Antarctic\_bird} & \subseteq & \mathsf{Bird} \\ \text{A-box:} & \mathit{Opus} & \in & \mathsf{Bird} \end{array}$$

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

$$\mathrm{P}(\mathsf{Flying\_bird}|\mathsf{Bird}) = 0.95,$$
$$\mathrm{P}(\mathsf{Antarctic\_bird}|\mathsf{Bird}) = 0.01.$$

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:

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

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

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

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 uncertainty in our 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 has developed a probabilistic extension of first-order logic that accomodates both notions of probability [Bac90].

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?

Consider our example above: here the given general conditional probabilities enable us to give a reasonable estimate for the likelihood of *Opus* being a flying bird. As *Opus* is an antarctic bird with probability 0.9, and antarctic birds do fly with probability 0.2, we would argue that *Opus* flies with a probability of at least $0.9{\times}0.2{=}0.18$. Reasoning somewhat more accurately, the probability for *Opus* being a flying non-antarctic bird will also be taken into account. The inference principle sketched here is known as *Jeffrey's rule* [Jef65] (a generalization of *direct inference* [Car50]).

This example shows clearly that general statistical knowledge directly influences subjective beliefs in specific propositions. Assigning degrees of belief on the basis of statistical information is one important aspect of what we will call *default reasoning about probabilities*: just as in logical default reasoning (e.g. [McC80],[Rei80]), where the inference rules of mathematical logic are augmented by inference rules that allow to deduce propositions that are “usually true” or “true by default”, inferences validated by probability theory are combined with inferences that are only justified when an assumption of representativeness or typicality of the situation under consideration is made. Like logical default reasoning, default reasoning about probabilities is defeasible: new information added to a knowledge base from which some inference was made may cause the retraction of this inference. The knowledge base of example 1.1, for instance, can be enlarged with $\mathrm{P}(Opus \in \mathsf{Flying\_bird}) = 0$, which will certainly invalidate the inference $\mathrm{P}(Opus \in \mathsf{Flying\_bird}) \geq 0.18$ drawn earlier.

It is necessary to emphasize that default reasoning about probabilities here is not seen as probabilistic extension of logical default reasoning, or as providing probabilistic semantics for logical default reasoning (cf. [Pea89]).

Also one should be very cautious with equating default reasoning about probabilities with *commonsense* reasoning about probabilities, because it is not at all certain even that the latter really exists! Different people's intuitions about which consequences to draw from probabilistic information often are greatly at odds with one another, and some of the more widespread beliefs about how probability works are directly contradicting the rules of probability theory (people in a casino, for instance, will crowd around a roulette wheel on which a long run of red numbers has just occured, in the expectation that here the probability for a black number has somehow increased). Hence, there seem to be few rules of handling uncertainty that are both "common" and "sensible" (see [KST82] for a wealth of material on how statistical information is commonly (mis-) interpreted).

In this paper we shall not attempt to develop a comprehensive theory of default reasoning about probabilities, even for the restricted domain of reasoning in terminologies. Rather, we shall augment probability theory with one additional reasoning principle - cross entropy minimization - that, generalizing the rule of direct inference, seems to go a long way towards an adequate modelling of that part of default reasoning about probabilities that links statistical with uncertain information. Thus, our basic motivation is very similar to the one in [BGHK92], [BGHK93]. The formalism to be developed, however, will be quite different from the one proposed by Bacchus et al., this one being unsatisfactory for our purpose, because it does not make any provisions for using existing degrees of belief in the derivation of new ones.

Other related work includes that of Paris and Vencovská who, 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]. In these works 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 this work, are not given.

In the following section we shall define a probabilistic extension for a very restricted concept language: the propositional fragment of $\mathcal{ALC}$. In section 3 semantical structures for the interpretation of knowledge bases in this language are defined. Section 4 discusses the problem of computing probability values resulting from the given semantics. Finally, in section 5, we investigate how the formalism developed so far can be extended to more expressive concept languages.

2 A probabilistic concept language

For a given finite vocabulary of concept names $\mathrm{S_C} = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\}$ and object names $\mathrm{S_O} = \{a, b, c, \ldots\}$ we define a language $\mathcal{PCL}$ by the following syntax rules:

- Concept terms:

  (i) Every concept name is a concept term.

  (ii) If C and D are concept terms, then $\mathrm{C} \wedge \mathrm{D}$, $\mathrm{C} \vee \mathrm{D}$ and $\neg\mathrm{C}$ are concept terms.

  The set of concept terms that can be constructed from $\mathrm{S_C}$ is designated by $\mathrm{T(S_C)}$.

- Terminological axioms: If A is a concept name and C is a concept term, then

  $$\begin{array}{lcll} \mathsf{A} & \subseteq & \mathrm{C} & \text{(Concept specialization) and} \\ \mathsf{A} & = & \mathrm{C} & \text{(Concept definition)} \end{array}$$

  are terminological axioms.

- Probabilistic terminological axioms: If C and D are concept terms, and p is a real number in $]0,1[$, then

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

  is a probabilistic terminological axiom.

- Probabilistic assertions: If a is an object name, C is a concept term, and p is a real number in [0,1], then

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

  is a probabilistic assertion.

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:

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

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 $\mathrm{P}(\mathrm{C}|\mathrm{D})$ and $\mathrm{P}(\mathrm{D}|\mathrm{C})$). 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

3.1 ...for $\mathcal{T}$

For $\mathcal{T}$ we stay with the standard model-theoretic interpretations used for concept languages.

Such an interpretation is defined as a pair $(\mathbf{D},\mathrm{I})$, where $\mathbf{D}$ is a nonempty set (the domain) and I is a function that assigns a subset of $\mathbf{D}$ to every concept name. (For concept languages that allow for binary relations (roles) to be used in the formation of concept terms, the interpretation I would also have to assign a binary relation on $\mathbf{D}$ to every role name.)

The interpretation function I is extended to concept terms by defining

$$\begin{array}{lll} \mathrm{I(C)} := \mathrm{I(C_1)} \cap \mathrm{I(C_2)} & \text{if} & \mathrm{C} = \mathrm{C_1} \wedge \mathrm{C_2} \\ \mathrm{I(C)} := \mathrm{I(C_1)} \cup \mathrm{I(C_2)} & \text{if} & \mathrm{C} = \mathrm{C_1} \vee \mathrm{C_2} \\ \mathrm{I(C)} := \mathbf{D} \setminus \mathrm{I}(\bar{\mathrm{C}}) & \text{if} & \mathrm{C} = \neg\bar{\mathrm{C}} \end{array}$$

It is obvious, then, what it means for an interpretation $(\mathbf{D},\mathrm{I})$ to be a model of a *terminological statement* (where a terminological statement is a generalization of a terminological axiom, by allowing for an arbitrary concpt term on the left-hand side):

$$\begin{array}{l} (\mathbf{D},\mathrm{I}) \models \mathrm{C} \subseteq \mathrm{D} \quad \text{iff} \quad \mathrm{I(C)} \subseteq \mathrm{I(D)}, \\ (\mathbf{D},\mathrm{I}) \models \mathrm{C} = \mathrm{D} \quad \text{iff} \quad \mathrm{I(C)} = \mathrm{I(D)} \quad (\mathrm{C},\mathrm{D} \in \mathrm{T(S_C)}). \end{array}$$

A terminological statement σ is a *logical consequence* of $\mathcal{T}$, written $\mathcal{T} \models \sigma$, iff for every $(\mathbf{D},\mathrm{I}) \models \mathcal{T}$: $(\mathbf{D},\mathrm{I}) \models \sigma$.

In case that $\mathcal{T} \models \mathrm{C} \subseteq \mathrm{D} \wedge \neg\mathrm{D}$ ($\mathcal{T} \models \mathrm{C} \subseteq \mathrm{D} \vee \neg\mathrm{D}$) for some concept term D, we write $\mathcal{T} \models \mathrm{C} = 0$ ($\mathcal{T} \models \mathrm{C} = 1$) for short.

3.2 ...for $\mathcal{PT}$

As mentioned in the introduction, an axiom of the form $\mathrm{P(C|D)} = p$ shall be viewed as a specification of a conditional probability. As long as we are only concerned with finite domains, we can paraphrase such a conditional probability by saying either "the fraction of elements in D that are also in C is p", or "an element randomly selected from D is in C with probability p", where in the second case we assume that every element of D is equally likely to be selected.

When we want to allow infinite extensions for D as well, there is no way to give such offhand interpretations of a conditional probability. Neither can we talk about a fraction of elements when both D and $D \cap C$ are infinite, nor is it possible (for countably infinite D) to select an element of D so that the probability to be chosen is the same for all elements.

Since we do not want to limit ourselves to interpretations over finite domains, we have to take a different approach. Our interest lying only with subsets of the domain that can be defined by a concept term, we can dispense with probability distributions on the domain altogether, and, instead, assign probabilities to concept terms directly.

This can be done by using the *Lindenbaum algebra* of concept terms

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

as the underlying probability space. Here $[\mathrm{T}(\mathrm{S_C})]$ is the set of equivalence classes in $\mathrm{T}(\mathrm{S_C})$ defined by the equivalence relation

$$\mathrm{C}_1 \equiv \mathrm{C}_2 \ \text{ iff } \ \models \mathrm{C}_1 = \mathrm{C}_2.$$

Hence, for every concept term C, $[\mathrm{C}] := \{\mathrm{D} \in \mathrm{T}(\mathrm{S_C}) \mid \mathrm{C} \equiv \mathrm{D}\}$ is an element of $[\mathrm{T}(\mathrm{S_C})]$. The operations $\vee$, $\wedge$, and $\neg$ are defined by performing disjunction, conjunction, and negation on representatives of the equivalence classes, i.e. $[\mathrm{C}] \vee [\mathrm{D}] := [\mathrm{C} \vee \mathrm{D}]$, $[\mathrm{C}] \wedge [\mathrm{D}] := [\mathrm{C} \wedge \mathrm{D}]$, and $\neg[\mathrm{C}] := [\neg \mathrm{C}]$. Finally, $0 := [\mathsf{A} \wedge \neg \mathsf{A}]$ and $1 := [\mathsf{A} \vee \neg \mathsf{A}]$ for some $\mathsf{A} \in \mathrm{S_C}$. It is easy to see that $\vee$, $\wedge$, and $\neg$ are well-defined, and that the resulting structure is a boolean algebra.

Note that the symbols $\vee$, $\wedge$, $\neg$ on the left hand side of these definitions must formally be distinguished from conjunction, disjunction, and negation of concept terms. A more exacting treatment of the matter would demand the use of a different set of symbols (like $\sqcup$, $\sqcap$, $^{-}$) for the operations in the boolean algebra. It is convenient, however, to gloss over this distinction, and to informally identify the two sets of operations. Also, in the sequel we will not continue to reflect the distinction between C and [C] in our notation, but simply use C for both the concept term and its equivalence class in $\mathfrak{A}(\mathrm{S_C})$.

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 $\mathrm{A}' \subseteq \mathrm{A}$ (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 C of $\mathfrak{A}(S_C)$, then, is (in the equivalence class of) the finite disjunction of the atoms A with $A \subseteq C$.

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

By extending an interpretation (**D**,I) for $\mathcal{T}$ with a probability measure $\mu \in \Delta\mathfrak{A}(S_C)$, semantics for $\mathcal{PT}$ can now be defined:

$$(\mathbf{D}, \mathrm{I}, \mu) \models \mathrm{P}(\mathrm{C}|\mathrm{D}) = p \quad \text{iff} \quad \mu(\mathrm{C} \wedge \mathrm{D}) = p \times \mu(\mathrm{D}).$$

3.3 ...for $\mathcal{P}_a$

Finding adequate semantics for probabilistic assertions turns out to be far more difficult than it was for probabilistic terminological axioms.

While there seem to be few alternatives to interpreting the latter by supplying, in one way or another, a probability measure on concept terms, there are at least two fundamentally different approaches to giving meaning to probabilistic assertions.

The first one is the possible-worlds approach taken (within a much more general framework of probabilistic logic than ours) for example in [Bac90], [GHK92]. Here probabilities for assertions like "$a \in$ C" are given through a probability measure on a set of interpretations (possible worlds). The probability of "$a \in$ C" is just the probability of the subset of interpretations that satisfy "$a \in$ C".

We shall take a different approach, however. Rather than modelling the uncertainty expressed in an axiom of the form $\mathrm{P}(a \in \mathrm{C}) = p$ by an uncertainty about which world we are living in, we shall view a as a "vague object" in a given world: in our interpretations a will not be assigned an element of the domain, but a probability distribution ν_a over the concept algebra:

$$\mathrm{I}: \; a \mapsto \nu_a \in \Delta\mathfrak{A}(S_C). \tag{1}$$

This leads to much simpler semantic structures than would be arrived at by introducing possible worlds semantics at this point, because unlike in those, not every single possibility of what a may be in reality is represented, but only the summary probabilistic information about properties of a that can be formulated in the given language.

Interpreting a by a probability measure on the same algebra as used for the interpretation of the probabilistic terminological axioms enables us to directly connect the beliefs held about an individual object with what, according to our statistical knowledge encoded in $\mathcal{PT}$, should typically be true for an element of the domain. Such a connection lies at the heart of a formalization of default reasoning about probabilities that, following the lead of direct inference and Jeffrey's rule, shall now be proposed.

3.3.1 Default reasoning about probabilities and Jeffrey's rule

The approach that is outlined by (1) is all we would need to provide a $\mathcal{PCL}$-knowledge base with sound semantics. We just might extend a given interpretation function I to $\mathrm{S_O}$ in this manner, and define

$$(\mathbf{D}, \mathrm{I}, \mu) \models \ \mathrm{P}(a \in \mathrm{C}) = p \quad \text{iff} \quad \nu_a(\mathrm{C}) = p.$$

But recall that we set out to do something more than to formulate sound but unrelated semantics for $\mathcal{PT}$ and $\mathcal{P}_a$. It was our intention to also capture the interaction between the two kinds of probabilistic statements that takes place in default reasoning about probabilities.

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

What, then, are the rules governing this reasoning process? In the introduction one such rule has already been used: the direct inference principle. If our statistics say that $\mathrm{P}(\mathrm{C}|\mathrm{D}) = p$, and if all we know about a is $\mathrm{P}(a \in \mathrm{D}) = q$ [1] then our best guess for $\mathrm{P}(a \in \mathrm{C} \wedge \mathrm{D})$ is $p \times q$.

The direct inference principle lends itself to an immediate generalization to the case that $\mathcal{P}_a$ supplies information about a's membership of disjoint concepts. Suppose that

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

[1] Note that the letter P in these expressions just stands as an abbreviation for "probability", not as a symbol for a specific probability distribution. Particularly, using the same letter in both of these expressions does not imply that they describe the same probability distribution.

where the C_i are disjoint. Putting

$$C_{n+1} := \neg \bigvee_{i=1}^{n} C_i \quad \text{and} \quad p_{n+1} := 1 - \sum_{i=1}^{n} p_i$$

we can complete $\mathcal{P}_a$ to a constraint-set for a partition $\{C_1, \ldots, C_{n+1}\}$ of the concept algebra.

Let a probability measure μ on $\mathfrak{A}(S_C)$ be given. We then assign to a a probability measure ν_a on $\mathfrak{A}(S_C)$ defined by

$$\nu_a(C) := \sum_{i=1}^{n+1} (p_i \times \mu(C \mid C_i)) \tag{2}$$

for every concept term C. This definition of a probability measure from a prior distribution μ and a set of constraint-values p_i for a partition of the probability space is known as *Jeffrey's rule* [Jef65].

The rationale for using (2) as a definition for the degree of belief that a belongs to a certain concept C is the assumption that a is a random element of the domain about which some partial information has been obtained, but which, in aspects that no observation has been made about, behaves like a typical representative of those objects that our statistics are grounded on.

3.3.2 Cross entropy

While Jeffrey's rule is an immediately appealing formalization for default reasoning about probabilities in the presence of disjoint constraints, there is no such simple rule to handle the case of general constraints.

Consider, for example, a knowledge base with

$$\begin{aligned}
\mathcal{T} &= \emptyset \\
\mathcal{PT} &= \mathrm{P}(\mathsf{Flying_animal}|\mathsf{Bird}) = 0.9 \\
& \mathrm{P}(\mathsf{Flying_animal}|\mathsf{Tropical_animal}) = 0.3 \\
\mathcal{P}_{Opus} &= \mathrm{P}(Opus \in \mathsf{Bird}) = 0.5 \\
& \mathrm{P}(Opus \in \mathsf{Tropical_animal}) = 0.5.
\end{aligned}$$

The two concepts Bird and Tropical_animal are neither disjoint, nor does one subsume the other (which would make $\mathcal{P}_{Opus}$ equivalent to a set of constraints on disjoint concepts). Therefore, Jeffrey's rule cannot be applied.

How, then, do we evaluate the probability for *Opus* being a flying animal? This evaluation should again be based on the assumption that *Opus* is as unexceptional an individual as possible within the limits drawn by $\mathcal{P}_{Opus}$. More precisely: the probability distribution ν_{Opus} we assign to *Opus* should resemble the generic probability distribution μ that is used in the interpretation of $\mathcal{PT}$ as closely as possible within the set of probability measures that satisfy $\mathcal{P}_{Opus}$.

The notion of "resemblance" of probability measures obviously needs some clarification. 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 prior distribution μ, we can then define the set of elements of N that have minimal distance to μ (i.e. the d-projection of μ onto N):

$$\pi^d_{\mathrm{N}}(\mu) := \{\nu \in \mathrm{N} \mid d(\mu, \nu) = \inf\{d(\mu, \nu') \mid \nu' \in \mathrm{N}\}\} \tag{3}$$

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.

Cross entropy (also called Kullback-Leibler distance) for two probability measures $\mu = (\mu_1, \ldots, \mu_n)$ and $\nu = (\nu_1, \ldots, \nu_n)$ with $\mu_i, \nu_i > 0$ for all i is defined as

$$CE(\mu, \nu) := \sum_{i=1}^n \nu_i \ln \frac{\nu_i}{\mu_i}. \tag{4}$$

As $x \ln \frac{x}{y}$ is not defined when either x or y equals 0, the precondition that all the components of μ and ν be strictly positive is usually imposed. For our purpose, however, it is convenient to extend this definition to probability distributions with 0-components. Since

$$\begin{array}{lll} x \ln \frac{x}{y} & \longrightarrow 0 & \text{for } x \to 0 \text{ and fixed } y > 0 \\ & \longrightarrow \infty & \text{for } y \to 0 \text{ and fixed } x > 0, \end{array}$$

and a pair of components (μ_i, ν_i) with $\mu_i = \nu_i = 0$ should not add to the distance of the two measures μ and ν, we define

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

for arbitrary μ, 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 with a grain of salt, 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. The following shows that conditions (i) and (iii) (with minor qualifications accounting for the possibility of CE being infinite) are satisfied as well.

Theorem 3.1 Let $\mu \in \Delta^n$, let $\mathrm{N} \subseteq \Delta^n$ be closed and convex such that $CE(\mu, \nu) < \infty$ for some $\nu \in \mathrm{N}$. Then $\pi_{\mathrm{N}}^{CE}(\mu)$ is a singleton.

Proof: See [SJ80], Theorem IV. The proof given there for the case of continuous probability distributions can easily be adapted for the finite probability distributions we are dealing with. □

A constraint set $\mathcal{P}$ defines a closed and convex subset of Δ^n (see section 4.2 for the details), so that (i) holds for CE. As a consequence of theorem 3.1 we may define:

Definition 3.2 Let $\mathrm{N} \subseteq \Delta^n$ be closed and convex. We define a partial function on Δ^n by

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

Theorem 3.3 Let $\mu \in \Delta^n$, let $\mathrm{N} \subseteq \Delta^n$ be defined by a set of constraints on disjoint sets such that $CE(\mu,\nu) < \infty$ for some $\nu \in \mathrm{N}$. Then $\pi_{\mathrm{N}}(\mu)$ is the probability measure obtained by applying Jeffrey's rule to μ and the constraint-set.

Proof: A proof can be found in [Wen88]. The theorem can also be obtained as a corollary to theorem 4.18 (page 33). □

There are basically two lines of argument that support the use of cross entropy as the appropriate function to be minimized in an updating procedure for probability measures.

The first one is to appeal directly to cross entropy's properties as a measure of information discrepancy, and to argue that an updating procedure should always choose that posterior distribution that assumes the least amount of additional information.

The second line of argument does not focus on the properties of cross entropy directly, but investigates fundamental requirements for a procedure that updates a given probability distribution μ to a posterior distribution 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

$$\nu = \pi_{\mathrm{N}}^{d}(\mu)$$

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].

3.4 Models of $\mathcal{KB}$

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

Definition 3.4 Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup\{\mathcal{P}_a \mid a \in \mathrm{S_C}\}$ be 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 $\mathrm{P(C|D)} = p \in \mathcal{PT} \Rightarrow \mu(\mathrm{C} \wedge \mathrm{D}) = p \times \mu(\mathrm{D})$;
- 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:

$$\begin{aligned}
\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C}) &:= \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T}\},\\
\mathit{Gen}(\mathcal{KB}) &:= \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T} \text{ and } \mathcal{PT}\},\\
\mathit{Bel}_a(\mathcal{KB}) &:= \{\mu \in \Delta\mathfrak{A}(\mathrm{S_C}) \mid \mu \text{ is consistent with } \mathcal{T} \text{ and } \mathcal{P}_a\}.
\end{aligned}$$

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 3.5 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 3.6 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) $\mu \in \mathit{Gen}(\mathcal{KB})$;

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

Defining μ and $\mathrm{I}(a)$ as probability distributions on $\mathfrak{A}(\mathrm{S_C})$ is just one of several feasible approaches that can be taken. Another one, keeping somewhat closer to the usual notion of an interpretation, is to define an extension of a given classical structure $(\mathbf{D}, \mathrm{I})$ by letting μ and $\mathrm{I}(a)$ be probability measures on the subalgebra $\mathfrak{A}(\mathrm{I}){:=}\{\mathrm{I}(\mathrm{C}) \mid \mathrm{C} \in \mathrm{T}(\mathrm{S_C})\}$ of $2^{\mathbf{D}}$. While up to the point reached by definitions 3.5 and 3.6, this would be somewhat more economical than the approach actually chosen, by making unnecessary the definition of the Lindenbaum algebra and the condition in definition 3.5 that $\mathrm{I}(\mathrm{C}){=}\emptyset$ implies $\mu(\mathrm{C}) = \mathrm{I}(a)(\mathrm{C}) = 0$, it would not save us anything in the long run, because when we want to get an overview of all possible measures occuring in such models, we will be led to representing this class of probability measures on the various algebras $\mathfrak{A}(\mathrm{I})$ by the set of probability measures they induce on the one algebra $\mathfrak{A}(\mathrm{S_C})$ via

$$\mu \in \Delta\mathfrak{A}(\mathrm{I}) \mapsto \tilde{\mu} \in \Delta\mathfrak{A}(\mathrm{S_C}),\ \tilde{\mu}(\mathrm{C}) := \mu(\mathrm{I}(\mathrm{C})).$$

Therefore, apart from the point at which the Lindenbaum algebra $\mathfrak{A}(\mathrm{S_C})$ and the sets *Gen* and *Bel* are introduced, the two approaches are hardly different. Particularly, they both validate the same set of probabilistic inferences from $\mathcal{KB}$.

The following definition provides the necessary notation for probabilistic consequences of $\mathcal{KB}$, and completes the set of definitions for the semantics of $\mathcal{PCL}$.

Definition 3.7 Let $\mathrm{J} \subseteq [0,1]$. We write

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

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

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

iff $\mathcal{KB} \models \mathrm{P}(\mathrm{C}|\mathrm{D}) \in \mathrm{J}$, 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 $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$.

We shall now try to get an overview of the models of our example from the beginning of this section. Since $\mathcal{T} = \emptyset$, we can construct models of $\mathcal{KB}$ where all of the eight atoms of $\mathfrak{A}(\mathrm{S_C})$:

$$\neg\mathsf{Ta} \wedge \neg\mathsf{B} \wedge \neg\mathsf{Fa}, \neg\mathsf{Ta} \wedge \neg\mathsf{B} \wedge \mathsf{Fa}, \ldots, \mathsf{Ta} \wedge \mathsf{B} \wedge \mathsf{Fa}$$

Table 1: Probability distributions consistent with $\mathcal{PT}$

	000	001	010	011	100	101	110	111
μ^1	1	0	0	0	0	0	0	0
μ^2	0	0	0.1	0.9	0	0	0	0
μ^3	0	0	0.032	0	0.677	0	0	0.29
μ^4	0	0	0	0	0.7	0.3	0	0
μ^5	0	0	0	0	0.666	0	0.033	0.3
μ^6	0	0	0	0.863	0	0.041	0.096	0
μ^7	0	0	0	0.857	0	0	0.1	0.042
μ^8	0	1	0	0	0	0	0	0

(using convenient abbreviations for the original concept names) have nonempty interpretations.

$Gen(\mathcal{KB})$ is the convex hull of the eight distributions listed in table 1 (see section 4.2 for how theses distributions are obtained). Every row in this table represents the probability distribution on $\mathfrak{A}(\mathrm{S_C})$ that assigns the given probabilities to the atoms specified at the head of the columns, where "000" abbreviates $\neg\mathsf{Ta} \wedge \neg\mathsf{B} \wedge \neg\mathsf{Fa}$, "001" abbreviates $\neg\mathsf{Ta} \wedge \neg\mathsf{B} \wedge \mathsf{Fa}$, etc.

$Bel_{Opus}(\mathcal{KB})$ is given in table 2. In these tables, as in the remainder of this paper, we distinguish elements of a set of probability measures by different superscripts, subscripts being reserved to mark the individual components of one probability measure. There being no occasion to raise a probability measure or one of its components to a power, this should not cause any ambiguities.

Not every element of Gen is a possible generic distribution μ in a model of $\mathcal{KB}$: neither for μ^1 nor for μ^2 nor for any convex combination $\mu = \lambda\mu^1 + (1-\lambda)\mu^2 = (1-\lambda, 0, 0.1\lambda, 0.9\lambda, 0, 0, 0)$ $(\lambda \in [0,1])$ of these two measures does a distribution $\nu \in Bel_{Opus}$ exist with $CE(\mu, \nu) < \infty$. This is because there is no distribution in Bel_{Opus} that has a 0 in all but the 1.,3. and 4. component (see theorem 4.8 for a formal statement). It stands to reason that generic distributions μ that are in this way incompatible with all the measures in Bel_{Opus} cannot be used in models of $\mathcal{KB}$. Otherwise we would have to assign a positive value to some probability $\mathrm{P}(Opus \in \mathrm{C})$ for a concept C with $\mu(\mathrm{C}) = 0$, i.e. an "impossible" concept.

Table 2: Probability distributions consistent with $\mathcal{P}_{Opus}$

	000	001	010	011	100	101	110	111
ν^1	0	0.5	0	0	0	0	0.5	0
ν^2	0	0.5	0	0	0	0	0	0.5
ν^3	0	0	0.5	0	0.5	0	0	0
ν^4	0	0	0.5	0	0	0.5	0	0
ν^5	0	0	0	0.5	0.5	0	0	0
ν^6	0	0	0	0.5	0	0.5	0	0
ν^7	0.5	0	0	0	0	0	0.5	0
ν^8	0.5	0	0	0	0	0	0	0.5

For μ^3 there is exactly one measure ν in Bel_{Opus} with $CE(\mu^3, \nu) < \infty$:

$$\mu_i^3 = 0 \Rightarrow \nu_i = 0 \quad (i = 1, \ldots, 8)$$

only holds for $\nu = \nu^3$, hence $\pi_{Bel_{Opus}}(\mu^3) = \nu^3$. As $\nu^3(\mathsf{Fa}) = 0$ this shows that no probability greater than 0 can be deduced from $\mathcal{KB}$ for $Opus \in$ Flying_animal. On the other hand we can show analogously that $\pi_{Bel_{Opus}}(\mu^6) = \nu^6$. Since $\nu^6(\mathsf{Fa}) = 1$, P($Opus \in$ Flying_animal) = 1 is consistent with $\mathcal{KB}$ also. What about intermediate values between 0 and 1? Theorem 4.13 will show that for all $p \in [0, 1]$ there are models of $\mathcal{KB}$ with $\nu_{Opus}(\mathsf{Fa}) = p$. Thus, $\mathcal{KB} \models$P($Opus \in$ Flying_animal)=[0,1].

The somewhat surprising result in this example, then, is that no nontrivial bounds for P($Opus \in$ Fa) can be deduced from the given knowledge base. The reason for this is that our semantics force us to consider all the generic distributions consistent with $\mathcal{PT}$. When $\mathcal{PT}$ contains only a very small number of constraints, as it does in our example, this is still a very large set of probability measures, and most of the measures in Bel_a will occur as the *CE*-projection of one of the measures in this set. In typical applications, however, it can be expected that fairly comprehensive information about conditional probabilities of concepts is available, which substantially reduces the space of possible generic distributions. In this case our semantics will always yield quite specific bounds for probabilities P($a \in$ C), whether or not $\mathcal{P}_a$ itself defines strong bounds on ν_a.

A hard boiled proponent of maximum-entropy methods would probably suggest that we always choose the maximum-entropy distribution from the set *Gen* as the one most reasonable hypothesis for the generic distribution μ, and use this measure alone in all models of the given knowledge base.

While this approach would make sure that every knowledge base determines unique measures μ and ν_a ($a \in \mathrm{S_O}$) for models of $\mathcal{KB}$ (thereby tremendously simplifying and strengthening the probabilistic reasoning that can be carried out in this framework), it would also exhibit rather undesirable properties of the maximum entropy principle, like yielding different results when ostensibly irrelevant information is added to the knowledge base (cf. [PV89]).

As long as it is not for some reason absolutely mandatory to assign unique probability values to assertions, a formalism that only yields intervals of reasonable generic and belief probabilities, and thereby reflects the degree of incompleteness of the information in the knowledge base, should be preferred to those that give definite results in all cases, but inevitably lead to inferences far more specific than is warranted by the fragmentary information in $\mathcal{KB}$.

For our example, the maximum-entropy distribution satisfying $\mathcal{PT}$ is

$$\mu^{\mathrm{me}} = (0.169, 0.169, 0.013, 0.225, 0.275, 0.054, 0.021, 0.072);$$

and

$$\nu^{\mathrm{me}}_{Opus} := \pi_{Bel_{Opus}}(\mu^{\mathrm{me}}) = (0.097, 0.097, 0.017, 0.29, 0.256, 0.05, 0.044, 0.15)$$

with $\nu^{\mathrm{me}}_{Opus}(\mathsf{Fa}) = 0.587$ (all the numbers in this example, as in every numerical example to follow, are rounded).

The greater soundness in our reasoning that derives from the cautious approach of taking into consideration all consistent generic μ's comes at a high price. Computing $\mathrm{J} \subseteq [0,1]$ for which $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$ poses a difficult problem for which no efficient solution can be given as yet. This problem along with some other computational aspects of $\mathcal{PCL}$ is discussed in the following section.

4 Computing probabilities

In the last two sections syntax and semantics for $\mathcal{PCL}$ have been defined. In this section we shall explore the problem of computing the probabilistic statements that are entailed by a $\mathcal{PCL}$-knowledge base.

There are basically three stages in which the data given in $\mathcal{T}$, $\mathcal{PT}$, and $\mathcal{P}_a$ will be processed, and queries be answered. First, the structural information in $\mathcal{T}$ allows us to reduce the concept algebra $\mathfrak{A}(\mathrm{S_C})$ to a (typically significantly smaller) algebra, which can be used as the underlying probability space in models of $\mathcal{KB}$. This step is explained in section 4.1. Secondly, the sets *Gen* and Bel_a are computed (section 4.2). This step being accomplished, it is easy to check whether the given knowledge base is consistent, i.e. has any model in the sense of definition 3.6. These first two stages can be regarded as a preprocessing of the given data, since they are performed independently of the specific queries to be put to the knowledge base.

In a third stage we finally compute the probability values we are interested in. Of the two types of queries - $\mathrm{P(C|D)}=?$ and $\mathrm{P}(a \in \mathrm{C})=?$ - we wish to provide an answer for, the first one can be dealt with easily once the preprocessing has taken place (section 4.3). A query of the second kind, in contrast, poses a difficult computational problem that is discussed in section 4.4.

4.1 The algebra $\mathfrak{A}(\mathcal{T})$

According to definition 3.5 we are dealing with probability distributions on the concept algebra $\mathfrak{A}(\mathrm{S_C})$. An explicit representation of any such distribution, i.e. a complete list of the values it takes on $\mathrm{A(S_C)}$ 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 $\mathrm{A(S_C)}$ whose extensions are not necessarily empty in models of $\mathcal{KB}$. Technically speaking, this probability space is a relative algebra of $\mathfrak{A}(\mathrm{S_C})$, for which we now provide a formal definition.

Definition 4.1 For a boolean algebra $\mathfrak{A} = \{\mathrm{A}, \vee^{\mathrm{A}}, \wedge^{\mathrm{A}}, \neg^{\mathrm{A}}, 0^{\mathrm{A}}, 1^{\mathrm{A}}\}$ and $a \in\mathrm{A}$, the *relative algebra* of $\mathfrak{A}$ with respect to a (written $\mathfrak{A} \mid a$) is the boolean algebra $\mathfrak{B} = \{\mathrm{B}, \vee^{\mathrm{B}}, \wedge^{\mathrm{B}}, \neg^{\mathrm{B}}, 0^{\mathrm{B}}, 1^{\mathrm{B}}\}$ with $\mathrm{B}=\{a' \in \mathrm{A} \mid a' \subseteq a\}$, $\vee^{\mathrm{B}} = \vee^{\mathrm{A}}$, $\wedge^{\mathrm{B}} = \wedge^{\mathrm{A}}$, $0^{\mathrm{B}} = 0^{\mathrm{A}}$, $1^{\mathrm{B}} = a$, and $\neg^{\mathrm{B}} a' = a \wedge \neg^{\mathrm{A}} a'$.

The difference between a relative algebra and a subalgebra of $\mathfrak{A}$ should be noted: both kinds of structures are derived from $\mathfrak{A}$ in a canonical way from a subset of A. In the case of a relative algebra $\mathfrak{A} \mid a$ the result is the full local structure of $\mathfrak{A}$ at a, whereas a subalgebra $\mathfrak{A}' \subseteq \mathfrak{A}$ is a coarsening of the global structure of $\mathfrak{A}$.

The element $\mathrm{C} \in \mathfrak{A}(\mathrm{S_C})$ by which we relativize $\mathfrak{A}(\mathrm{S_C})$ is the disjunction of the atoms not equivalent to 0 relative to $\mathcal{T}$:

Definition 4.2 Let $\mathcal{T}$ be a set of terminological axioms in the vocabulary $\mathrm{S_C}$. Define

$$\begin{array}{rcll} \mathrm{A}(\mathcal{T}) & := & \{\mathrm{C} \in \mathrm{A}(\mathrm{S_C}) \mid \mathcal{T} \not\models \mathrm{C} = 0\} & \\ \mathrm{C}(\mathcal{T}) & := & \bigvee \mathrm{A}(\mathcal{T}) & \\ \mathfrak{A}(\mathcal{T}) & := & \mathfrak{A}(\mathrm{S_C}) \mid \mathrm{C}(\mathcal{T}) & \qquad (\text{“}\mathfrak{A}(\mathrm{S_C}) \textit{ relativized by } \mathcal{T}\text{”}) \end{array}$$

The atoms of $\mathfrak{A}(\mathcal{T})$ then, are just the elements of $\mathrm{A}(\mathcal{T})$, so that $\mid \mathrm{A}(\mathcal{T}) \mid$ is the size of probability distributions on $\mathfrak{A}(\mathcal{T})$.

Example 4.3 Let $\mathrm{S_C} = \{\mathsf{A}_0, \mathsf{A}_1, \ldots, \mathsf{A}_n\}$, let

$$\mathcal{T} = \{\mathsf{A}_1 \subseteq \mathsf{A}_0, \mathsf{A}_2 \subseteq \mathsf{A}_0, \ldots, \mathsf{A}_n \subseteq \mathsf{A}_0\}.$$

Then

$$\mathrm{A}(\mathcal{T}) = \{\mathrm{B}_0 \wedge \ldots \wedge \mathrm{B}_n \mid \mathrm{B}_i \in \{\mathsf{A}_i, \neg\mathsf{A}_i\},\ \mathrm{B}_0 = \neg\mathsf{A}_0 \ \Rightarrow\ \forall i\, \mathrm{B}_i = \neg\mathsf{A}_i\}$$

with

$$|\mathrm{A}(\mathcal{T})| = 2^{|\mathrm{S_C}|-1} + 1.$$

Example 4.4 Let $\mathrm{S_C} = \{\mathsf{A}_0, \mathsf{A}_1, \ldots, \mathsf{A}_n\}$, let

$$\mathcal{T} = \{\mathsf{A}_1 \subseteq \mathsf{A}_0, \mathsf{A}_2 \subseteq \mathsf{A}_1, \ldots, \mathsf{A}_n \subseteq \mathsf{A}_{n-1}\}.$$

Here

$$\mathrm{A}(\mathcal{T}) = \{\mathrm{B}_0 \wedge \ldots \wedge \mathrm{B}_n \mid \mathrm{B}_i \in \{\mathsf{A}_i, \neg\mathsf{A}_i\},\ \forall i : \mathrm{B}_i = \neg\mathsf{A}_i \Rightarrow \mathrm{B}_j = \neg\mathsf{A}_j \text{ for all } j, i \leq j \leq n\}$$

and

$$|\mathrm{A}(\mathcal{T})| = |\mathrm{S_C}| + 1.$$

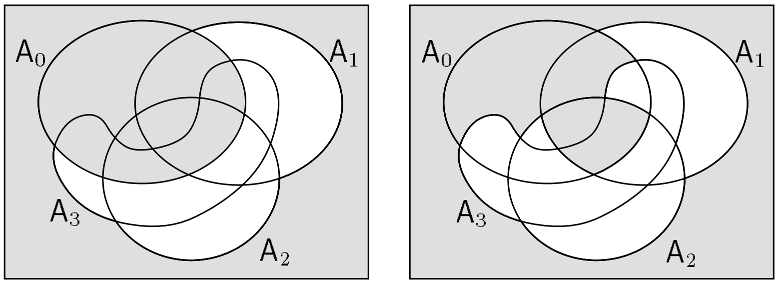

Figure 1: Algebras $\mathfrak{A}(\mathrm{S_C})$ and $\mathfrak{A}(\mathcal{T})$ in examples 4.3 (left) and 4.4 (right)

Figure 1 illustrates the two examples for $n = 3$. The shaded area is the element $\mathrm{C}(\mathcal{T})$ with respect to which the full algebra $\mathfrak{A}(\{\mathsf{A}_0, \ldots, \mathsf{A}_3\})$ is relativized. Examples 4.3 and 4.4 show two extreme cases of one very large algebra and one very small algebra $\mathfrak{A}(\mathcal{T})$, corresponding to terminologies with a very poor and a very rich structure, respectively.

It is not difficult for any given $\mathcal{T}$, to compute a list of the elements of $\mathrm{A}(\mathcal{T})$ in such a way that no more than $2 \mid \mathrm{A}(\mathcal{T}) \mid \mid \mathrm{S_C} \mid$ tests for the consistency of a concept term C with respect to $\mathcal{T}$ have to be made.

Alternatively, $\mathfrak{A}(\mathcal{T})$ could be defined as the Lindenbaum algebra of $\mathcal{T}$, which is defined just as $\mathfrak{A}(\mathrm{S_C})$ above, only the equivalence classes here are taken over the relation

$$\mathrm{C}_1 \equiv_{\mathcal{T}} \mathrm{C}_2 \;\text{ iff }\; \mathcal{T} \models \mathrm{C}_1 = \mathrm{C}_2.$$

The resulting algebra is isomorphic to $\mathfrak{A}(\mathcal{T})$ as defined in 4.2.

4.2 Consistent probability measures and overall consistency

The next step in making deductions from a $\mathcal{PCL}$-knowledge base is to compute $Gen(\mathcal{KB})$ and $Bel_a(\mathcal{KB})$ ($a \in \mathrm{S_O}$).

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 $\mathrm{P}(\mathrm{C}|\mathrm{D}) = p$ 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 $\mid \mathrm{A}(\mathcal{T}) \mid = n$, 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 $\leq n-k$.

Example 4.5 Let

$$\begin{array}{rcl} \mathcal{T} & = & \mathsf{A}_2 \subseteq \mathsf{A}_1 \\ & & \mathsf{A}_3 \subseteq \mathsf{A}_1 \wedge \neg\mathsf{A}_2 \\ \mathcal{PT} & = & \mathrm{P}(\mathsf{A}_2 | \mathsf{A}_2 \vee \mathsf{A}_3) = 0.2 \\ & & \mathrm{P}(\mathsf{A}_2 | \neg\mathsf{A}_3) = 0.3 \\ \mathcal{P}_a & = & \mathrm{P}(a \in \mathsf{A}_2 \vee \mathsf{A}_3) = 0.3. \end{array}$$

$\mathrm{A}(\mathcal{T})$ consists of the four atoms

$$\begin{array}{ll} \mathrm{A}_1 = \neg\mathsf{A}_1 \wedge \neg\mathsf{A}_2 \wedge \neg\mathsf{A}_3, & \mathrm{A}_2 = \mathsf{A}_1 \wedge \neg\mathsf{A}_2 \wedge \neg\mathsf{A}_3, \\ \mathrm{A}_3 = \mathsf{A}_1 \wedge \neg\mathsf{A}_2 \wedge \mathsf{A}_3, & \mathrm{A}_4 = \mathsf{A}_1 \wedge \mathsf{A}_2 \wedge \neg\mathsf{A}_3. \end{array}$$

The subset of $\Delta^4 = \{(x_1, \ldots, x_4) \mid x_i \in [0,1], \sum x_i = 1\}$ that satisfies the constraints in $\mathcal{PT}$ (letting x_i stand for the probability of A_i) is the intersection of Δ^4 with the two hyperplanes defined by the linear equations

$$\begin{array}{rcl} x_4 & = & 0.2(x_3 + x_4) \\ x_4 & = & 0.3(x_1 + x_2 + x_4), \end{array}$$

or, equivalently,

$$\begin{array}{rcl} -0.2x_3 + 0.8x_4 & = & 0 \\ -0.3x_1 - 0.3x_2 + 0.7x_4 & = & 0. \end{array}$$

The result is the line connecting

$$a = (0.318, 0, 0.546, 0.136) \quad \text{and} \quad b = (0, 0.318, 0.546, 0.136)$$

(cf. figure 2).

The constraint in $\mathcal{P}_a$ can be directly translated into the condition $x_3 + x_4 = 0.6$, so that Bel_a is given by the four points

$$(0.4, 0, 0, 0.6), \; (0.4, 0, 0.6, 0), \; (0, 0.4, 0, 0.6), \; \text{and} \; (0, 0.4, 0.6, 0).$$

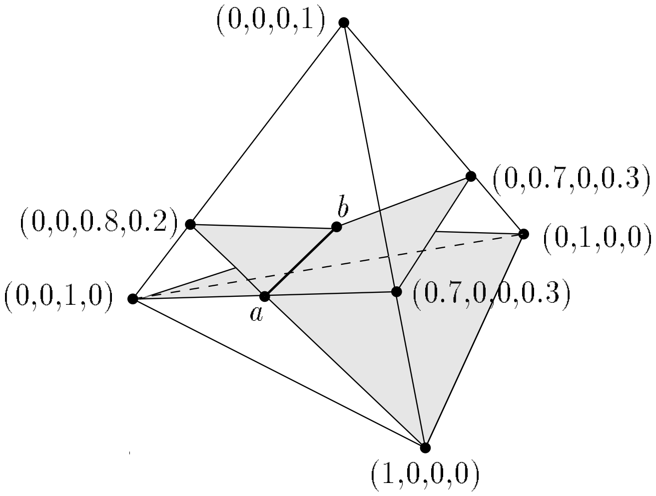

Figure 2: Intersection of Δ^4 with two hyperplanes in example 4.5

Having calculated the sets Gen and Bel_a, a $\mathcal{PCL}$-knowledge base $\mathcal{KB}$ can easily be checked for consistency. Following (i)-(iii) in definition 3.6, 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 4.6 Let $\mu \in \Delta^n$. Define

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

For a polytope M the notation *int* M 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 lemma is a trivial observation.

Lemma 4.7 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 4.8 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 $\mu \in \mathrm{M}$ and all $\nu \in \mathrm{N}$, which in turn is equivalent to (ii), because by lemma 4.7 $\mathrm{Z}(\bar{\mu})$ 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 $\mathrm{Z}(\nu) \subseteq \mathrm{Z}(\nu^j)$). □

By theorem 4.8, **(c)** is equivalent to

(c′) There exists $a \in \mathrm{S_O}$ such that $\pi_{Bel_a}(\bar{\mu})$ is not defined.

and can be tested by a finite number of comparisons of index sets $\mathrm{Z}(\cdot)$.

Definition 4.9 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^* is the set of generic measures that actually occur in models of $\mathcal{KB}$. Gen^* is a convex subset of Gen, which, if $\mathcal{KB}$ is consistent, contains at least all the interior points of Gen (the interior points all having the same minimal set $\mathrm{Z}(\cdot)$).

Example 4.10 To illustrate the various sources of inconsistency for $\mathcal{PCL}$- knowledge bases we consider some examples that are all constructed from the following list of axioms and assertions:

$$\begin{aligned}
\tau_1 &= \mathsf{A}_2 \subseteq \neg\mathsf{A}_1 \\
\tau_2 &= \mathsf{A}_2 \subseteq \mathsf{A}_1 \wedge \neg\mathsf{A}_1 \\[1em]
\rho_1 &= \mathrm{P}(\mathsf{A}_1|\mathsf{A}_2) = 0.3 \\
\rho_2 &= \mathrm{P}(\mathsf{A}_1|\mathsf{A}_2) = 0.7 \\[1em]
\alpha_1 &= \mathrm{P}(a \in \mathsf{A}_2) = 0.5 \\
\alpha_2 &= \mathrm{P}(a \in \mathsf{A}_2) = 0 \\
\alpha_3 &= \mathrm{P}(a \in \mathsf{A}_1) = 0.7
\end{aligned}$$

$\mathcal{KB}_1 := \{\rho_1, \rho_2\} \cup \{\alpha_2\}$ is consistent. The apparent contradiction in ρ_1 and ρ_2 is resolved in the limiting case where $\mu(\mathsf{A}_2) = 0$.

We call $\mathcal{KB}$ *inconsistent for* C iff $\mu(\mathrm{C}) = 0$ in all models of $\mathcal{KB}$. Thus, $\mathcal{KB}_1$ is inconsistent for A_2.

The same is true for $\mathcal{KB}_2 := \{\tau_2\}$. Here, the terminological axiom τ_2 has $\mathrm{I}(\mathsf{A}_2) = \emptyset$ as a consequence, which in turn entails $\mu(\mathsf{A}_2) = 0$ in every model of $\mathcal{KB}_2$.

$\mathcal{KB}_3 := \{\rho_1, \rho_2\} \cup \{\alpha_1\}$ is inconsistent. As before, we have $\mu(\mathsf{A}_2) = 0$ for all μ consistent with $\mathcal{PT}$. In contrast to $\mathcal{KB}_1$ however, there is in $\mathcal{KB}_3$ also an assertion of a positive probability for the inconsistent concept A_2. This makes $\mathcal{KB}_3$ inconsistent: here there is no pair of measures $\mu \in Gen(\mathcal{KB}_3)$ and $\nu_a \in Bel_a(\mathcal{KB}_3)$ such that $\mathrm{Z}(\mu) \subseteq \mathrm{Z}(\nu_a)$.

$\mathcal{KB}_4 := \{\tau_1\} \cup \{\alpha_1, \alpha_3\}$ is inconsistent because there is no measure on $\mathfrak{A}(\{\tau_1\})$ that satisfies α_1 and α_3.

4.3 Answering queries P(C|D)= ?

Once an appropriate representation of *Gen* has been obtained, and overall consistency has been ascertained, it is not much of an effort to answer queries of the form $\mathrm{P}(\mathrm{C}|\mathrm{D}) \in ?$.

Basically, all that has to be done is to compute $\mu(\mathrm{C}|\mathrm{D})$ for every vertex μ of *Gen*. The minimal and maximal values thereby obtained delimit the range of conditional probabilities $\mathrm{P}(\mathrm{C}|\mathrm{D})$ that occur in models of $\mathcal{KB}$.

Theorem 4.11 Let $\mathcal{KB}$ be a consistent $\mathcal{PCL}$-knowledge base, C,D$\in$T(S$_\mathrm{C}$). Let Gen=$conv\{\mu^1,\dots,\mu^k\}$ and suppose that $\mathcal{KB}\models$ P(C|D) = J. Then, either μ^i(D) = 0 for $i=1,\dots,k$ and J=$\emptyset$, or J is a nonempty interval and

$$\begin{aligned} \inf \mathrm{J} &= \min\{\mu^i(\mathrm{C}\mid\mathrm{D}) \mid \mu^i(\mathrm{D})>0,\ i=1,\dots,k\}, && (5)\\ \sup \mathrm{J} &= \max\{\mu^i(\mathrm{C}\mid\mathrm{D}) \mid \mu^i(\mathrm{D})>0,\ i=1,\dots,k\}. && (6) \end{aligned}$$

Proof: For consistent $\mathcal{KB}$ we have $Gen^*(\mathcal{KB})\neq\emptyset$, and

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

Thus $\mathcal{KB}\models$ P(C|D) = $\emptyset$ iff μ(D)=0 for every $\mu\in Gen^*$. Since Gen^* contains at least all the interior points of Gen, this is equivalent to μ^i(D)=0 for $i=1,\dots,k$.

Assume that μ^i(D)>0 for some $i\in\{1,\dots,k\}$. Without loss of generality μ^i(D)>0 for $i=1,\dots,l$, and μ^i(D)=0 for $i=l+1,\dots,k$ with $1\leq l\leq k$.

Then, μ(D)>0 for all the interior points of Gen, which means that J is nonempty.

J is an interval: $\mu\mapsto\mu$(C|D) is a continuous function on the connected set $\{\mu\in Gen^*\mid\mu(\mathrm{D})>0\}$. The codomain of a continuous function on a connected domain is connected; and the connected subsets of **R** are just the intervals.

Furthermore, for any polytope M with $\mu(\mathrm{D})>0$ for all $\mu\in$M, the function $\mu\mapsto\mu$(C|D) attains its maximal (minimal) values at vertices of M (one way to prove this is to consider for any fixed $\mu^1,\ \mu^2\in$ M the function $\lambda\mapsto(\lambda\mu^1+(1-\lambda)\mu^2)(\mathrm{C}|\mathrm{D})$ $(\lambda\in[0,1])$ which is monotone, and, therefore, attains its maximal and minimal values for $\lambda=0$ and $\lambda=1$, i.e. for μ^1 and μ^2 respectively).

Now, equations (5) and (6) can be proven. We give a proof for (5) in two steps. The proof for (6) proceeds analogously.

First step: $\inf\mathrm{J}\geq\min\{\mu^i(\mathrm{C}|\mathrm{D})\mid i=1,\dots,l\}$. Let $\mu^*\in Gen^*$ with μ^*(D)>0. If $\mu^*\in conv\{\mu^1,\dots,\mu^l\}$ then clearly $\mu^*(\mathrm{C}|\mathrm{D})\geq\min\{\mu^i(\mathrm{C}|\mathrm{D})\mid i=1,\dots,l\}$ because of μ(C|D) being minimal on a vertex of $conv\{\mu^1,\dots,\mu^l\}$. Suppose $\mu^*\notin conv\{\mu^1,\dots,\mu^l\}$. As $\mu^*(\mathrm{D})>0$, μ^* is not in the convex hull of $\{\mu^{l+1},\dots,\mu^k\}$ either. Therefore, μ^* is a convex combination of a point $\mu^{*1}\in conv\{\mu^1,\dots,\mu^l\}$, and a point $\mu^{*2}\in conv\{\mu^{l+1},\dots,\mu^k\}$. Then,

$$\mu^*(\mathrm{C}|\mathrm{D})=\mu^{*1}(\mathrm{C}|\mathrm{D})\geq\min\{\mu^i(\mathrm{C}|\mathrm{D})\mid i=1,\dots,l\}.$$

Second step: $\inf\mathrm{J}\leq\min\{\mu^i(\mathrm{C}|\mathrm{D})\mid i=1,\dots,l\}$. Let $\mu^*\in\{\mu^1,\dots,\mu^l\}$ such that $\mu^*(\mathrm{C}|\mathrm{D})=\min\{\mu^i(\mathrm{C}|\mathrm{D})\mid i=1,\dots,l\}$. If $\mu^*\in$ Gen* then $\mu^*(\mathrm{C}|\mathrm{D})\in$ J, and we are done. Suppose $\mu^* \notin \mathrm{Gen}^*$. Since every interior point of *Gen* is an element of Gen^* we can choose a sequence $(\mu^{*i})_{i\geq 1}$ in Gen^* such that $\mu^{*i} \to \mu^* \ (i \to \infty)$. Finally, by the continuity of $\mu \mapsto \mu(\mathrm{C}|\mathrm{D})$:

$$\inf \mathrm{J} \leq \lim_{i\to\infty} \mu^{*i}(\mathrm{C}|\mathrm{D}) = \mu^*(\mathrm{C}|\mathrm{D}).$$

□

Theorem 4.11 provides the means only to compute the closure of the interval we want. While this will be quite sufficient in almost every situation, it should be noted that it is not difficult either to decide whether $\inf \mathrm{J} \in \mathrm{J}$ or $\sup \mathrm{J} \in \mathrm{J}$: using the assumptions in the theorem, let

$$\mathrm{F}^{\mathrm{inf}} := conv\{\mu^i \mid \mu^i(\mathrm{C} \mid \mathrm{D}) = \inf \mathrm{J}\}.$$

$\mathrm{F}^{\mathrm{inf}}$ is a face of *Gen* consisting of all the points in *Gen* on which the conditional probability of C given D has the constant value $\inf \mathrm{J}$. Hence, $\inf \mathrm{J} \in \mathrm{J}$ iff $\mu \in Gen$ for some $\mu \in \mathrm{F}^{\mathrm{inf}}$. Theorem 4.8 can be applied to $\mathrm{F}^{\mathrm{inf}}$ and the polytopes $Bel_a(\mathcal{KB})$ $(a \in \mathrm{S_O})$ in order to decide whether this is the case. Analogously for $\sup \mathrm{J}$.

A simple but noteworthy corollary is the following.

Corollary 4.12 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 $\mathrm{C,D} \in \mathrm{T(S_C)}$ 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 4.12 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 $\mathcal{KB}$.

4.4 Answering queries $\mathrm{P}(a \in \mathrm{C}) \in$?

Having dealt with the preliminary problem of computing the sets Gen and Bel_a, which, as we have seen, comprises a test for the consistency of $\mathcal{KB}$, and allows us to answer queries about conditional probabilities for generic measures, we can now turn to the main issue of this section: how to compute the interval J with $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$. The best solution to this problem would be to determine for given Gen and Bel_a the codomain $\pi_{Bel_a}(Gen^*) \subseteq Bel_a$, which contains just those measures that appear as the interpretation ν_a of a in models of $\mathcal{KB}$, so that $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$ for

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

Given an explicit description of $\pi_{Bel_a}(Gen^*)$ it would therefore be easy to answer the query $\mathrm{P}(a \in \mathrm{C}) =?$ for any $\mathrm{C} \in \mathrm{T}(\mathrm{S_C})$. Unfortunately, such an explicit description seems to be rather difficult to obtain, so that for the time being we shall take the approach of doing separate computations for every query. Before we deal with these computations in greater detail in section 4.4.3, we first show that the sets J are indeed intervals.

4.4.1 Continuity

Before we attempt any computations of $\mathrm{J} \subseteq [0,1]$ with $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$, we have to ascertain that this J is indeed an interval. There certainly would be something wrong with our semantics, if it was possible that for some $p_1, p_2, p_3 \in [0, 1]$, with $p_1 < p_2 < p_3$, both $p_1, p_3 \in \mathrm{J}$ and $p_2 \notin \mathrm{J}$ held true. However, we do not have to worry about such a situation, as a corollary to the following theorem shows.

Theorem 4.13 Let $\mathrm{N} \subseteq \Delta^n$ be closed and convex. Then $\pi_\mathrm{N} : \Delta^n \rightarrow \mathrm{N}$ is continuous.

Proof: The proof only relies on the continuity of cross entropy. Let $\mu_0, \mu_1, \mu_2, \ldots \in \mathrm{domain}(\pi_\mathrm{N})$ such that

$$(\mu_i)_{i \geq 1} \rightarrow \mu_0 \ (i \rightarrow \infty).$$

Let

$$\nu_i := \pi_\mathrm{N}(\mu_i) \ \ (i \geq 0).$$

Assume that

$$(\nu_i)_{i\geq 1} \not\to \nu_0 \ (i \to \infty).$$

Then there is an open neighbourhood U of ν_0 in N and a subsequence $(\nu_{i_j})_{j\geq 1} \subseteq (\nu_i)_{i\geq 1}$ such that $\nu_{i_j} \notin$ U for all $j \geq 1$. Since N is bounded, we can choose a convergent sequence (ν_{i_j}):

$$\nu_{i_j} \to \nu_0' \ (j \to \infty)$$

for some $\nu_0' \notin$ U.

From the continuity of $CE(\cdot,\cdot)$ we have

$$CE(\mu_{i_j}, \nu_0) \to CE(\mu_0, \nu_0) \ (j \to \infty)$$

and

$$CE(\mu_{i_j}, \nu_{i_j}) \to CE(\mu_0, \nu_0') \ (j \to \infty).$$

Also, $CE(\mu_0, \nu_0) < CE(\mu_0, \nu_0')$ by the definition of ν_0. Hence, for suitably large $j: \ CE(\mu_{i_j}, \nu_0) < CE(\mu_{i_j}, \nu_{i_j})$, a contradiction. □

Corollary 4.14 Let N $\subseteq \Delta^n$ be convex; let M $\subseteq \Delta^n$ be connected. Then $\pi_{\mathrm{N}}(\mathrm{M})$ is connected.

Corollary 4.15 Let $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$. Then, J is an interval.

Proof: By the preceding corollary, $\pi_{Bel_a}(Gen^*)$ is connected. The function $\nu \mapsto \nu(\mathrm{C})$ $(\nu \in \Delta\mathfrak{A}(\mathcal{T}))$ is continuous, so that the codomain of $\pi_{Bel_a}(Gen^*)$ under this function is connected. The connected subsets of **R** are just the intervals. □

4.4.2 Computing $\pi_{Bel}(\mu)$

We have not addressed, so far, the question of finding for any given μ and constraints $\mathcal{P}$ the distribution ν that minimizes $CE(\mu, \nu)$ among the measures consistent with $\mathcal{P}$. Only for the special case of $\mathcal{P}$ being a set of constraints on disjoint concepts we know that ν is just the measure obtained by Jeffrey's rule.

In general it is not possible to give such a simple closed-form solution. Instead a nonlinear optimization algorithm must be used to compute an approximate 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.

4.4.3 Computing $\pi_{Bel}(Gen^*)(\mathrm{C})$: An approximation

Finding $\pi_{Bel}(\mu)$ for a single measure μ seems to be the easier part of the inference problem we are faced with. As it turns out, greater difficulties are presented by the question of how to determine the full range of values $\pi_{Bel}(Gen^*)$, given that we have an effective procedure to (approximately) compute π_{Bel}.

It has already been mentioned that apparently we cannot hope to compute an explicit description of the complete set $\pi_{Bel_a}(Gen^*)$, which in general is not even convex. However, all we need to know about $\pi_{Bel_a}(Gen^*)$ is the set of intervals $\{\pi_{Bel_a}(Gen^*)(\mathrm{C}) \mid \mathrm{C} \in \mathrm{T}(\mathrm{S_C})\}$ it defines.

A finite number of points $\{\nu^1, \ldots, \nu^n\} \in \pi_{Bel_a}(Gen^*)$ provides the approximation

$$[\min\{\nu^i(\mathrm{C}) \mid i = 1, \ldots, n\}, \max\{\nu^i(\mathrm{C}) \mid i = 1, \ldots, n\}]$$

for $\pi_{Bel_a}(Gen^*)(\mathrm{C})$. An approximate answer to the query $\mathrm{P}(a \in \mathrm{C}) =?$ therefore can be given by finding $\nu^{\inf}$, $\nu^{\sup} \in \pi_{Bel_a}(Gen^*)$ such that $\nu^{\inf}(\mathrm{C})$ is close to $\inf(\pi_{Bel_a}(Gen^*)(\mathrm{C}))$, and $\nu^{\sup}(\mathrm{C})$ is close to $\sup(\pi_{Bel_a}(Gen^*)(\mathrm{C}))$.

There does not seem to be an easy way of finding elements of $\pi_{Bel_a}(Gen^*)$ that are guaranteed to provide good approximations for $\inf(\pi_{Bel_a}(Gen^*)(\mathrm{C}))$ or $\sup(\pi_{Bel_a}(Gen^*)(\mathrm{C}))$, so that for the time being we shall have to settle for the somewhat unsatisfactory approach of using a search-algorithm based on some heuristics. 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 and 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) change of these current bounds.

Obviously, the complexity of such a search depends on the dimension and the number of vertices of Gen (recall that for consistent $\mathcal{KB}$, $Gen \backslash Gen^*$ consists only of points on the boundary of Gen, so that the searchspace Gen^* basically is just Gen). The cost of a single computation of π_{Bel} depends on the size of the probability space $\mathfrak{A}(\mathcal{T})$ and the number of constraints in $\mathcal{P}$. In the following we show that the search-space Gen^* can often be reduced to a substantially smaller space.

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

Definition 4.16 Let $\mathfrak{A}$=$\{\mathrm{A},\ldots\}$ be an algebra (not necessarily finite). Let $\mathfrak{A}'$, μ, and M be such that either

(i) $\mathfrak{A}'$ is a subalgebra of $\mathfrak{A}$, $\mu \in \Delta\mathfrak{A}$, and $\mathrm{M} \subseteq \Delta\mathfrak{A}$, or

(ii) $\mathfrak{A}'$ is a relative algebra of $\mathfrak{A}$ with respect to some $a \in \mathfrak{A}$, $\mu \in \Delta\mathfrak{A}$ with $\mu(a) = 1$, and $\mathrm{M} \subseteq \Delta\mathfrak{A}$ with $\nu(a) = 1$ for every $\nu \in \mathrm{M}$.

Then $\mu\restriction\mathfrak{A}'$ denotes the restriction of μ to $\mathfrak{A}'$, and $\mathrm{M}\restriction\mathfrak{A}'$ is the set $\{\mu\restriction\mathfrak{A}' \mid \mu \in \mathrm{M}\}$.

The following is a technical lemma that will be needed in the proof of theorem 4.18.

Lemma 4.17 $\forall x_1, y_1, x_2, y_2 > 0$:

$$(x_1 + x_2)\ln\frac{x_1 + x_2}{y_1 + y_2} \leq x_1\ln\frac{x_1}{y_1} + x_2\ln\frac{x_2}{y_2}.$$

Equality holds iff

$$\frac{x_1}{x_1 + x_2} = \frac{y_1}{y_1 + y_2}.$$

Proof: We make the substitutions

$$x := x_1 + x_2,$$
$$y := y_1 + y_2.$$

Then, for suitable $\lambda_x, \lambda_y \in]0, 1[$:

$$x_1 = \lambda_x x, \quad x_2 = (1 - \lambda_x)x, \quad y_1 = \lambda_y y, \quad y_2 = (1 - \lambda)y.$$

The inequality we want to prove now reads:

$$x\ln\frac{x}{y} \leq \lambda_x x\ln\frac{\lambda_x x}{\lambda_y y} + (1 - \lambda_x)x\ln\frac{(1 - \lambda_x)x}{(1 - \lambda_y)y}, \tag{7}$$

where equality holds iff $\lambda_x = \lambda_y$. Let x, y be fixed and consider the right side of (7) as a function $\mathrm{f}(\lambda_x, \lambda_y)$ of λ_x and λ_y.

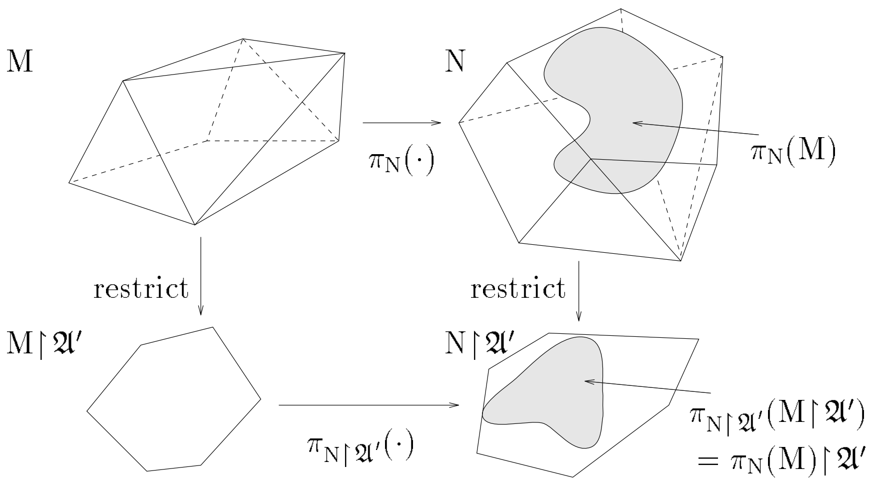

Figure 3: Theorem 4.18

Clearly, $\mathrm{f}(\lambda_x, \lambda_y) = x\ln\frac{x}{y}$ for $\lambda_x = \lambda_y$. To see that this is the minimal value that $\mathrm{f}(\lambda_x, \lambda_y)$ attains for any pair of λ_x, λ_y consider the partial derivative of $\mathrm{f}(\lambda_x, \lambda_y)$ with respect to λ_y:

$$\frac{\partial \mathrm{f}(\lambda_x, \lambda_y)}{\partial \lambda_y} = \left(\frac{(1-\lambda_x)}{(1-\lambda_y)} - \frac{\lambda_x}{\lambda_y} \right) x \tag{8}$$

Equating the right hand side of (8) with zero yields the unique solution $\lambda_x = \lambda_y$. Checking the second derivative $\frac{\partial^2 \mathrm{f}(\lambda_x,\lambda_y)}{\partial \lambda_y^2}$ proves this solution to be indeed a minimum.

Hence, for any fixed λ_x, $\mathrm{f}(\lambda_x, \lambda_y)$ has its unique minimum at $\lambda_y = \lambda_x$ with the value $x\ln\frac{x}{y}$. This proves the lemma. □

The following theorem has been shown in [SJ81] for probability distributions with densities. As it also plays a vital role in the generalization of $\mathcal{PCL}$ to a probabilistic version of $\mathcal{ALC}$, we state it here in a version more suited for the given context, and provide a proof on a somewhat more elementary level than the one given in [SJ81].

Theorem 4.18 Let $\mathfrak{A}'$ be a subalgebra of the finite algebra $\mathfrak{A}$ generated by a partition $\mathrm{A}' = \{\mathrm{A}'_1, \ldots, \mathrm{A}'_k\}$ of $\mathfrak{A}$. Let M,N$\subseteq \Delta\mathfrak{A}$, where N is defined by a set of constraints on $\mathfrak{A}'$, i.e.

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

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

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

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

$$\pi_{\mathrm{N}}(\mu)(\mathrm{C}) = \sum_{j=1}^{k} \pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu \restriction \mathfrak{A}')(\mathrm{A}'_j)\mu(\mathrm{C} \mid \mathrm{A}'_j). \tag{9}$$

Figure 3 illustrates the first part of the theorem.

Proof: We first show that for every $\mu \in \Delta\mathfrak{A}$, $\pi_{\mathrm{N}}(\mu)$ is defined if and only if $\pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu \restriction \mathfrak{A}')$ is and in this case,

$$\pi_{\mathrm{N}}(\mu) \restriction \mathfrak{A}' = \pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu \restriction \mathfrak{A}'), \tag{10}$$

which proves the first part of the theorem.

Let $\mathrm{A} = \{\mathrm{A}_1, \ldots, \mathrm{A}_n\}$ be the set of atoms of $\mathfrak{A}$. Every $\mathrm{A}'_j \in \mathrm{A}'$ is the disjunction of some $\mathrm{A}_i \in \mathrm{A}$. For $i \in \{1, \ldots, n\}$ let $\mathrm{ind}(i) \in \{1, \ldots, k\}$ be such that $\mathrm{A}_i \subseteq \mathrm{A}'_{\mathrm{ind}(i)}$.

Suppose that $\pi_{\mathrm{N}}(\mu)$ is defined. Then $\mathrm{Z}(\mu) \subseteq \mathrm{Z}(\nu)$ for a suitable $\nu \in$N (cf. theorem 4.8). It is easy to see that in this case $\mathrm{Z}(\mu \restriction \mathfrak{A}') \subseteq \mathrm{Z}(\nu \restriction \mathfrak{A}')$ holds, and hence $\pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu \restriction \mathfrak{A}')$ is defined.

Conversely, let $\pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu \restriction \mathfrak{A}')$ be defined; let $\bar{\nu} \in \mathrm{N} \restriction \mathfrak{A}'$ such that $\mathrm{Z}(\mu \restriction \mathfrak{A}') \subseteq \mathrm{Z}(\bar{\nu})$. Since N is defined by a set of constraints on $\mathfrak{A}'$ alone, every extension of $\bar{\nu}$ to the algebra $\mathfrak{A}$ is in N. For every $j \in \{1, \ldots, k\}$ with $\bar{\nu}_j > 0$ there exists an index $\mathrm{inv}(j) \in \{1, \ldots, n\}$ with ind(inv(j))=j, and $\mu_{\mathrm{inv}(j)} > 0$. For $\nu \in \Delta\mathfrak{A}$ with $\nu_i = \bar{\nu}_{\mathrm{ind}(i)}$ if $\bar{\nu}_{\mathrm{ind}(i)} > 0$ and i=inv(ind(i)), $\nu_i = 0$ otherwise, then holds: $\nu \in$ N, and $\mathrm{Z}(\mu) \subseteq \mathrm{Z}(\nu)$. Hence, $\pi_{\mathrm{N}}(\mu)$ is defined.

Let $\mu = (\mu_1, \ldots, \mu_n) \in \Delta\mathfrak{A}$ such that $\pi_{\mathrm{N}}(\mu)$ is defined. Let

$$\nu := \pi_{\mathrm{N}}(\mu),\ \bar{\mu} := \mu \restriction \mathfrak{A}', \text{ and } \bar{\nu} := \nu \restriction \mathfrak{A}'.$$

Suppose that $\bar{\nu} \neq \pi_{\mathrm{N}\restriction\mathfrak{A}'}(\bar{\mu})$. Since $\bar{\nu} \in \mathrm{N} \restriction \mathfrak{A}'$ this means that there is some $\bar{\nu}^* \in \mathrm{N} \restriction \mathfrak{A}'$ with

$$CE(\bar{\mu}, \bar{\nu}^*) < CE(\bar{\mu}, \bar{\nu}). \tag{11}$$

We define an extension ν^* of $\bar{\nu}^*$ to the algebra $\mathfrak{A}$ by

$$\nu_i^* := \bar{\nu}^*_{\mathrm{ind}(i)} \frac{\mu_i}{\bar{\mu}_{\mathrm{ind}(i)}} \qquad (i \in \{1, \ldots, n\}).$$

$\nu^* = (\nu_1^*, \ldots, \nu_n^*)$ is a probability measure in N that extends $\bar{\nu}^*$. In order to prove (10), it remains to show that $CE(\mu, \nu^*) < CE(\mu, \nu)$, a contradiction to the definition of ν:

$$CE(\mu, \nu^*) = \sum_{i=1}^{n} \nu_i^* \ln \frac{\nu_i^*}{\mu_i} = \sum_{j=1}^{k} \sum_{\substack{i \\ \mathrm{ind}(i)=j}} \nu_i^* \ln \frac{\nu_i^*}{\mu_i}$$

$$= \sum_{j=1}^{k} \sum_{\substack{i \\ \mathrm{ind}(i)=j}} \mu_i \frac{\bar{\nu}_j^*}{\bar{\mu}_j} \ln \frac{\bar{\nu}_j^*}{\bar{\mu}_j} = \sum_{j=1}^{k} \bar{\nu}_j^* \ln \frac{\bar{\nu}_j^*}{\bar{\mu}_j}$$

$$= CE(\bar{\mu}, \bar{\nu}^*) < CE(\bar{\mu}, \bar{\nu}) \leq CE(\mu, \nu)$$

The last inequality is obtained by applying lemma 4.17.

Also by lemma 4.17, equality holds in this inequality iff

$$\frac{\nu_i}{\bar{\nu}_{\mathrm{ind}(i)}} = \frac{\mu_i}{\bar{\mu}_{\mathrm{ind}(i)}}$$

for all $i \in \{1, \ldots, n\}$. This, however, just means that the conditional probabilities $\nu(\cdot \mid \mathrm{A}_j')$ and $\mu(\cdot \mid \mathrm{A}_j')$ are the same for $j = 1, \ldots, k$. Thus, given (10), $CE(\mu, \nu)$ is minimal for ν as defined in (9).

□

The following example illustrates the use of theorem 4.18 for reducing the search space in a computation of $\pi_{Bel_a}(Gen^*)(\mathrm{C})$.

Example 4.19

$$\text{Let } \mathcal{T} = \begin{array}{l} \mathsf{A}_2 \subseteq \mathsf{A}_1 \\ \mathsf{A}_3 \subseteq \mathsf{A}_2 \\ \mathsf{A}_4 \subseteq \neg \mathsf{A}_2 \\ \mathsf{A}_5 \subseteq \mathsf{A}_4 \\ \mathsf{A}_6 \subseteq \mathsf{A}_1 \wedge \neg \mathsf{A}_2 \end{array}$$

The atoms of the algebra $\mathfrak{A}(\mathcal{T})$ defined by these axioms are

Table 3: Generic distributions in example 4.19

	μ^1	μ^2	μ^3	μ^4	μ^5	μ^6
$\mu(\mathrm{A}_1)$	0.327	0.327	0	0	0	0
$\mu(\mathrm{A}_2)$	0.073	0.073	0.108	0.108	0.308	0.308
$\mu(\mathrm{A}_3)$	0.055	0.055	0.081	0.081	0.231	0.231
$\mu(\mathrm{A}_4)$	0	0	0.195	0.195	0	0
$\mu(\mathrm{A}_5)$	0.327	0.327	0.292	0.292	0.092	0.092
$\mu(\mathrm{A}_6)$	0	0	0	0	0	0
$\mu(\mathrm{A}_7)$	0	0	0	0	0	0
$\mu(\mathrm{A}_8)$	0	0	0	0	0.185	0.185
$\mu(\mathrm{A}_9)$	0.055	0.055	0.081	0.081	0.046	0.046
$\mu(\mathrm{A}_{10})$	0.164	0	0.243	0	0.138	0
$\mu(\mathrm{A}_{11})$	0	0.164	0	0.243	0	0.138
$\mu(\mathsf{A}_1)$	0.545	0.545	0.811	0.811	0.462	0.462

$$\begin{array}{lll} \mathrm{A}_1 = 000000 & \mathrm{A}_2 = 000100 & \mathrm{A}_3 = 000110 \\ \mathrm{A}_4 = 100000 & \mathrm{A}_5 = 100001 & \mathrm{A}_6 = 100100 \\ \mathrm{A}_7 = 100101 & \mathrm{A}_8 = 100110 & \mathrm{A}_9 = 100111 \\ \mathrm{A}_{10} = 110000 & \mathrm{A}_{11} = 111000 & \end{array}$$

where, as before, 111000 (for example) is an abbreviations for

$$\mathsf{A}_1 \wedge \mathsf{A}_2 \wedge \mathsf{A}_3 \wedge \neg\mathsf{A}_4 \wedge \neg\mathsf{A}_5 \wedge \neg\mathsf{A}_6.$$

Let

$$\begin{aligned} \mathcal{PT} \;=\; & \mathrm{P}(\mathsf{A}_2|\mathsf{A}_1) = 0.3 \\ & \mathrm{P}(\mathsf{A}_5|\mathsf{A}_4) = 0.6 \\ & \mathrm{P}(\mathsf{A}_6 \wedge \mathsf{A}_4|\mathsf{A}_1) = 0.1 \\ & \mathrm{P}(\mathsf{A}_6|\neg\mathsf{A}_4) = 0.4 \\ & \mathrm{P}(\mathsf{A}_1|\mathsf{A}_5) = 0.5 \\ & \mathrm{P}(\mathsf{A}_1|\mathsf{A}_4) = 0.3 \end{aligned}$$

$$\begin{aligned} \mathcal{P}_a = \; & \mathrm{P}(a \in \mathsf{A}_4 \vee \mathsf{A}_6) = 0.8 \\ & \mathrm{P}(a \in \mathsf{A}_5) = 0.1 \end{aligned}$$

Table 3 shows the six probability measures whose convex hull is $Gen(\mathcal{KB})$, which in this example is identical to $Gen^*(\mathcal{KB})$. Suppose that we are interested in

Table 4: A search in *Gen*

μ	$\pi_{Bel_a}(\mu)(\mathsf{A}_1)$
μ^3	0.76
μ^5	0.411
μ^4	0.76
μ^6	0.411
μ^1	0.69
$0.5\mu^3 + 0.5\mu^4$	0.76
$0.5\mu^5 + 0.5\mu^6$	0.411
$0.5\mu^1 + 0.5\mu^3$	0.722
$0.3\mu^1 + 0.3\mu^3 + 0.4\mu^5$	0.581
$0.3\mu^1 + 0.3\mu^2 + 0.4\mu^3$	0.715
$0.15\mu^1 + 0.15\mu^2 + 0.15\mu^3 + 0.15\mu^4 + 0.15\mu^5 + 0.25\mu^6$	0.581

the probability $\mathrm{P}(a \in \mathsf{A}_1)$. The last row in table 3 shows the probability that is assigned to A_1. A search in *Gen* for the points μ that maximize (minimize) $\pi_{Bel_a}(\mu)(\mathsf{A}_1)$ might start with μ^3, μ^5, and then proceed with various convex combinations of the μ^i. An exemplary path for such a search is given in table 4.

Here, in the course of the search, no points are found that yield values for $\pi_{Bel_a}(\mu)(\mathsf{A}_1)$ increasing (lowering) the bounds obtained by μ^3 (μ^5). Hence, our (not necessarily optimal) conclusion in this example is $\mathrm{P}(a \in \mathsf{A}_1) = [0.411, 0.76]$. Notice that the information in $\mathcal{P}_a$ alone allows for the whole range [0,1] of probabilities as $\mathrm{P}(a \in \mathsf{A}_1)$, so that in this case the minimum-cross entropy approach has indeed resulted in a substantial strengthening of our conclusions.

We now show how theorem 4.18 facilitates the search for the bounds of $\mathrm{P}(a \in \mathsf{A}_1)$.

The subalgebra $\mathfrak{A}'$ of $\mathfrak{A}(\mathcal{T})$ generated by the concepts $\mathsf{A}_4 \vee \mathsf{A}_6$, A_5, and A_1 consists of the six atoms

$$\begin{array}{lll} \mathrm{A}'_1 = \mathrm{A}_1, & \mathrm{A}'_2 = \mathrm{A}_4 \vee \mathrm{A}_{10} \vee \mathrm{A}_{11}, & \mathrm{A}'_3 = \mathrm{A}_2, \\ \mathrm{A}'_4 = \mathrm{A}_5 \vee \mathrm{A}_6 \vee \mathrm{A}_7, & \mathrm{A}'_5 = \mathrm{A}_3, & \mathrm{A}'_6 = \mathrm{A}_8 \vee \mathrm{A}_9. \end{array}$$

The restriction $Gen' := Gen \upharpoonright \mathfrak{A}'$ is a polytope with only three vertices shown in table 5.

By theorem 4.18 we know that it is sufficient to determine the range of values

Table 5: The reduced polytope Gen'

	μ'^1	μ'^2	μ'^3
$\mu'(\mathrm{A}'_1)$	0.327	0	0
$\mu'(\mathrm{A}'_2)$	0.164	0.438	0.138
$\mu'(\mathrm{A}'_3)$	0.073	0.108	0.308
$\mu'(\mathrm{A}'_4)$	0.327	0.292	0.092
$\mu'(\mathrm{A}'_5)$	0.055	0.081	0.231
$\mu'(\mathrm{A}'_6)$	0.055	0.081	0.231
$\mu'(\mathsf{A}_1)$	0.545	0.811	0.462

Table 6: A search in Gen'

μ	$\pi_{Bel_a}(\mu)(\mathsf{A}_1)$
μ'^2	0.76
μ'^3	0.411
μ'^1	0.69
$0.5\mu'^1 + 0.5\mu'^2$	0.722
$0.5\mu'^1 + 0.5\mu'^3$	0.513
$0.5\mu'^2 + 0.5\mu'^3$	0.586
$0.3\mu'^1 + 0.3\mu'^2 + 0.4\mu'^3$	0.581

$\pi_{Bel_a \restriction \mathfrak{A}'}(\mu')$ for $\mu' \in Gen'$.

Table 6 shows the result of a search in Gen' which allows us to arrive at the same conclusion as above within a shorter time.

5 Extending probabilistic reasoning to more expressive languages

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.

5.1 Role quantification

In $\mathcal{ALC}$ the concept-forming operations of section 2 are augmented by role quantification: the vocabulary now contains a finite set $\mathrm{S_R} = \{\mathsf{r}, \mathsf{s}, \ldots\}$ of role names in addition to, and disjoint from, $\mathrm{S_C}$ and $\mathrm{S_O}$. A new syntax rule allows for the construction of concept terms from concept terms and role names:

(iii) If C is a concept term and $\mathsf{r} \in \mathrm{S_R}$, then $\forall \mathsf{r} : \mathrm{C}$ and $\exists \mathsf{r} : \mathrm{C}$ are concept terms.

$\mathrm{T}(\mathrm{S_C},\mathrm{S_R})$ denotes the set of concept terms constructible via the rules (i)-(iii).

Adding (iii) to our previous concept-construction rules, leaving the rules for terminological axioms, probabilistic terminological axioms, and probabilistic assertions unchanged, gives us a probabilistic extension of $\mathcal{ALC}$ that, 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.

Role names are interpreted as binary relations over the domain, i.e. the interpretation function I (cf. section 3.1) assigns a subset of $\mathbf{D}\times\mathbf{D}$ to every $\mathsf{r}\in\mathrm{S_R}$. An extension of I to arbitrary concept terms is now given by

$$\begin{array}{ll} \mathrm{I(C)} := \{d \in \mathbf{D} \mid \exists e \in \mathrm{I(C')} \text{such that } (d, e) \in \mathrm{I}(\mathsf{r})\} & \text{if} \quad \mathrm{C} = \exists \mathsf{r} : \mathrm{C'} \\ \mathrm{I(C)} := \{d \in \mathbf{D} \mid \forall e : (d, e) \in \mathrm{I}(\mathsf{r}) \rightarrow e \in \mathrm{I(C')}\} & \text{if} \quad \mathrm{C} = \forall \mathsf{r} : \mathrm{C'} \end{array}$$

in addition to the rules given in 3.1. The relations $(\mathbf{D},\mathrm{I})\models \sigma$ and $\mathcal{T} \models \sigma$ for a terminological statement σ are defined as before.

5.2 Probabilistic semantics for $\mathcal{PALC}$

Probabilistic information formulated in $\mathcal{PALC}$ can often be handled according to the same intuitive rules as described for $\mathcal{PCL}$. As an example consider the following knowledge base.

$$\begin{aligned}\mathcal{KB}_1 = \quad & \mathrm{P}(\mathsf{Penguin}|\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}) = 0.2 \\ & \mathrm{P}(\mathit{Opus} \in \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}) = 1\end{aligned}$$

$\mathcal{KB}_1$ states that a bird whose food consists exclusively of fish is a penguin with probability 0.2, and that Opus is a bird that meets this description. The presence of role quantification in this example does not present any problem for using direct inference in order to obtain the estimate $\mathrm{P}(\mathit{Opus} \in \mathsf{Penguin}) = 0.2$. Application of direct inference in this example is particularly simple, because $\mathcal{KB}_1$ is basically propositional: the same information could be expressed by substituting a new concept name for the term $\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}$, thereby obtaining a simple $\mathcal{PCL}$-knowledge base. Matters are different in the following example.

$$\begin{aligned}\mathcal{KB}_2 = \quad & \mathsf{Herring} \subseteq \mathsf{Fish} \\ & \mathsf{Penguin} \subseteq \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Herring} \\ & \mathrm{P}(\mathsf{Penguin}|\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Herring}) = 0.2 \\ & \mathrm{P}(\mathit{Opus} \in \mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish}) = 1\end{aligned}$$

Here, $\forall \mathsf{feeds\_on} : \mathsf{Fish}$ and $\forall \mathsf{feeds\_on} : \mathsf{Herring}$ cannot be simply treated as two distinct propositional variables, because this would mean to ignore the fact that $\forall \mathsf{feeds\_on} : \mathsf{Herring}$ is subsumed by $\forall \mathsf{feeds\_on} : \mathsf{Fish}$ (i.e. $\mathcal{KB}_2 \models \forall \mathsf{feeds\_on} : \mathsf{Herring} \subseteq \forall \mathsf{feeds\_on} : \mathsf{Fish}$). This information, however, enables us to give [0,0.2] as an estimate for $\mathrm{P}(\mathit{Opus} \in \mathsf{Penguin})$ (since $\mathrm{P}(\mathsf{Penguin}|\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Fish})$ could be any value between 0 and $0.2 = \mathrm{P}(\mathsf{Penguin}|\mathsf{Bird} \wedge \forall \mathsf{feeds\_on} : \mathsf{Herring})$).

We shall now generalize the semantical definitions given in section 3 to give semantics to $\mathcal{PALC}$-knowledge bases in such a way that inferences of the kind in the examples above become sound.

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 distributions on this algebra.

The Lindenbaum algebra for $\mathcal{PALC}$ can be defined in precisely the same manner as was done for $\mathcal{PCL}$ (cf. page 8). 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 decreasing chains $C_0 \supseteq C_1 \supseteq C_2 \supseteq \ldots$ in $[T(S_C, S_R)]$ with $C_i \neq C_{i+1} \neq 0$ for all i.

The set of probability measures on $\mathfrak{A}(S_C,S_R)$ is denoted $\Delta\mathfrak{A}(S_C,S_R)$. Probability measures, here, are still required to only satisfy finite additivity. $\mathfrak{A}(S_C,S_R)$ not being closed under infinite disjunctions, it is only of limited interest to consider countable additivity. Observe that even though $\mathfrak{A}(S_C,S_R)$ is a countable algebra, probability measures on $\mathfrak{A}(S_C,S_R)$ 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 $\mathfrak{A}(S_C,S_R)$.

The following two examples are meant to give an impression of what probability measures on $\mathfrak{A}(S_C,S_R)$ may look like.

Example 5.1 In this example we construct a probability measure as the limit of a sequence of probability measures on finite subalgebras of $\mathfrak{A}(S_C,S_R)$. As $T(S_C,S_R)$ is countable, so is $\mathfrak{A}(S_C,S_R)$. Therefore, we can write

$$\mathfrak{A}(S_C,S_R) = C_0, C_1, C_2, \ldots$$

Without loss of generality, $C_0 = 0$ and $C_1 = 1$. Let $\mathfrak{A}_i$ be the subalgebra of $\mathfrak{A}(S_C,S_R)$ generated by $\{C_0, C_1, C_2, \ldots, C_i\}$ $(i \geq 1)$ with $\{A_{i,1}, A_{i,2}, \ldots, A_{i,n_i}\}$ the set of its atoms. In the i^{th} step of our construction we define a probability measure μ_i on $\mathfrak{A}_i$. For $\mathfrak{A}_1$ there is little choice, but to define

$$\mu_1(C_0) := 0, \quad \text{and } \mu_1(C_1) := 1.$$

Assume that μ_i has been defined. In order to define μ_{i+1}, every atom $A_{i+1,j}$ $(j \leq n_{i+1})$ must be assigned a probability value extending the definition of μ_i.

$A_{i+1,j}$ is either $A_{i,k} \wedge C_{i+1}$ or $A_{i,k} \wedge \neg C_{i+1}$ for some $A_{i,k}$ $(k \leq n_i)$. Let

$$\overline{A_{i+1,j}} = \begin{cases} A_{i,k} \wedge \neg C_{i+1} & \text{if } A_{i+1,j} = A_{i,k} \wedge C_{i+1} \\ A_{i,k} \wedge C_{i+1} & \text{if } A_{i+1,j} = A_{i,k} \wedge \neg C_{i+1}. \end{cases}$$

We put

$$\mu_{i+1}(A_{i+1,j}) := \begin{cases} \mu_i(A_{i,k}) & \text{if } \overline{A_{i+1,j}} = 0 \\ \frac{1}{2}\mu_i(A_{i,k}) & \text{if } \overline{A_{i+1,j}} \neq 0. \end{cases}$$

μ_i then is a probability measure on $\mathfrak{A}_{i+1}$ that extends μ_i. Hence, for every $C_i \in \mathfrak{A}(S_C,S_R)$ we can define:

$$\mu(C_i) := \mu_i(C_i),$$

which gives us a probability measure on $\mathfrak{A}(S_C,S_R)$.

Example 5.2 A very natural way to obtain a probability distribution on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ is to let it be induced by a probability measure on the domain $\mathbf{D}$ of a classical interpretation for $\mathrm{S_C} \cup \mathrm{S_R}$. Suppose, therefore, that $(\mathbf{D},\mathrm{I})$ is an interpretation for $\mathrm{S_C} \cup \mathrm{S_R}$, where $\mathbf{D}$ is equipped with a $\sigma-$algebra $\mathfrak{D}$ and a probability measure δ.[2] Furthermore, assume that each $\mathrm{C} \in \mathrm{T}(\mathrm{S_C}, \mathrm{S_R})$ is interpreted by an element of $\mathfrak{D}$. Since by definition $\mathrm{C}_1 \equiv \mathrm{C}_2 \Rightarrow \mathrm{I}(\mathrm{C}_1) = \mathrm{I}(\mathrm{C}_2)$, I also unambiguously assigns an element of $\mathfrak{D}$ to every $\mathrm{C} \in \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$(recall our convention not to distinguish notationally the concept term C from the equivalence class of C in $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$). Then,

$$\mu(\mathrm{C}) := \delta(\mathrm{I}(\mathrm{C}))$$

is well defined and routinely checked to be a probability measure on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$.

With $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ and $\Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ at our disposal, definition 3.6 can now be repeated almost verbatim for $\mathcal{PALC}$.

Definition 5.3 Let $\mathrm{S} = \mathrm{S_C} \cup \mathrm{S_R} \cup \mathrm{S_O}$ be a vocabulary. A *$\mathcal{PALC}$-interpretation for* S is a triple $(\mathbf{D}, \mathrm{I}, \mu)$ where $\mathbf{D}$ is a set,

$$\begin{aligned} &\mathrm{I} : \mathrm{S_C} \to 2^{\mathbf{D}}, \\ &\mathrm{I} : \mathrm{S_R} \to 2^{\mathbf{D}\times\mathbf{D}}, \\ &\mathrm{I} : \mathrm{S_O} \to \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R}), \end{aligned}$$

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

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 3.6) 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 the two examples at the beginning of this section some kind of minimum cross entropy reasoning (in the special form of direct inference) has been employed.

[2] See [Hal50], for example, for the basic measure theory. For countable $\mathbf{D}$, $\mathfrak{D}$ may be assumed to be the power set $2^{\mathbf{D}}$.

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 $\mathrm{P}(Opus \in \mathsf{Penguin})$. 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 distributions 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 5.4 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 4.18). Let $\mu \in \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be 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 $\pi^*_{\mathrm{N}}(\mu)$ realizes cross entropy minimization for every finite subalgebra of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ containing the concepts used to define N.

Theorem 5.5 Let $\mu \in \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, let $\mathrm{N} \subseteq \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ 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 $\pi_{\mathrm{N}\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')$ is defined. Then, for every finite $\mathfrak{A}^* \supseteq \mathfrak{A}'$: $\pi_{\mathrm{N}\restriction\mathfrak{A}^*}(\mu\restriction\mathfrak{A}^*)$ is defined and equal to $\pi^*_{\mathrm{N}}(\mu)\restriction\mathfrak{A}^*$.

Proof: First observe that $\mathrm{N}\restriction\mathfrak{A}^*$ is just the subset

$$\{\nu \in \Delta\mathfrak{A}^* \mid \nu(\mathrm{C}_i) = p_i,\ i=1,\ldots,n\} \subseteq \Delta\mathfrak{A}^*,$$

because every measure in this set can be extended to a measure in N by a construction along the same lines as in example 5.1 (see also lemma 5.7 below). Hence we may substitute $\mathfrak{A}^*$, $\mathrm{N}\restriction\mathfrak{A}^*$, and $\{\mu\restriction\mathfrak{A}^*\}$ for $\mathfrak{A}$, N, and M, respectively, in theorem 4.18, which then yields

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

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

$\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, $Gen(\mathcal{KB})$ and $Bel_a(\mathcal{KB})$ are defined for $\mathcal{PALC}$ as in definition 3.4. We can now generalize definition 3.6 to the language $\mathcal{PALC}$.

Definition 5.6 Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup\{\mathcal{P}_a \mid a \in \mathrm{S_O}\}$ be a $\mathcal{PALC}$-knowledge base. Let $(\mathbf{D}, \mathrm{I}, \mu)$ be a $\mathcal{PALC}$-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) $(\mathbf{D},\mathrm{I}\restriction \mathrm{S_C}\cup\mathrm{S_R})\models \mathcal{T}$ 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 $\mathrm{I}(a) = \pi^*_{Bel_a(\mathcal{KB})}(\mu)$.

Note that definitions 5.3, 5.4 and 5.6 are proper extensions of definitions 3.5, 3.2 and 3.6 respectively, i.e. when applied to a $\mathcal{PALC}$ knowledge base in a language without role names the two sets of definitions are equivalent.

5.3 Probability distributions on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$

We now turn to the question of how the semantical definitions given in the previous section give rise to effective ways of computing the consequences entailed by a $\mathcal{PALC}$- knowledge base.

Basically, the same procedures as for $\mathcal{PCL}$ will be used: having only to consider the finite subalgebras of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ generated by the concept terms actually appearing either in the knowledge base, or in the queries put to the system, we can handle probability distributions on $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ in very much the same way as we did for $\mathfrak{A}(\mathrm{S_C})$.

In section 4, for the most part, we assumed all the probability distributions we worked with to be defined on one common algebra $\mathfrak{A}(\mathcal{T})$. Only in section 4.4 it was noted that we may restrict these distributions to adequate smaller algebras in order to facillitate the computation of answers to specific queries. For $\mathcal{PALC}$ it proves useful to take the more refined approach right from the beginning, and to define the convex sets Gen, Bel_a ($a \in \mathrm{S_O}$) of consistent probability measures each on its own appropriate finite algebra. For Gen (Bel_a), this will just be the algebra generated by the concept terms in $\mathcal{PT}$ ($\mathcal{P}_a$), relativized by $\mathcal{T}$. For a finite subalgebra $\mathfrak{M} \subseteq \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, the relative algebra $\mathfrak{M}(\mathcal{T})$ is defined as in definition 4.2.

In the following $\mathfrak{M}$ and $\mathfrak{N}$ always denote finite subalgebras of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$. Think of $\mathfrak{M}$ ($\mathfrak{N}$) as the subalgebra generated by the concept terms in $\mathcal{PT}$ ($\mathcal{P}_a$). First we note a lemma that formally states a simple fact, a special case of which has already been used in the proof of theorem 5.5.

Lemma 5.7 Let $\mathfrak{M}\subset\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be finite. Let $Gen \subseteq \Delta\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be defined by a set of terminological axioms $\mathcal{T}$ and $\mathcal{PT} = \{\mathrm{P}(\mathrm{C}_i|\mathrm{D}_i) = p_i \mid i = 1,\ldots,n\}$ with $\mathrm{C}_i,\ \mathrm{D}_i \in \mathfrak{M}$ $(i = 1,\ldots,n)$. Then,

$$Gen\!\upharpoonright\!\mathfrak{M}(\mathcal{T}) = \{\mu \in \Delta\mathfrak{M}(\mathcal{T}) \mid \mu(\mathrm{C}_i \wedge \mathrm{C}(\mathcal{T}) \mid \mathrm{D}_i \wedge \mathrm{C}(\mathcal{T})) = p_i,\ i = 1,\ldots,n\},$$

where $\mathrm{C}(\mathcal{T})$ as in definition 4.2. Analogously for Bel and $Bel\!\upharpoonright\!\mathfrak{N}(\mathcal{T})$.

Proof: Only the inclusion '$\supseteq$' is not completely trivial. We must verify that every $\mu \in \Delta\mathfrak{M}(\mathcal{T})$ satisfying the given constraints can be extended to a measure in Gen. First, μ is extended to the subalgebra $\mathfrak{M}$ by defining $\mu(\neg\mathrm{C}(\mathcal{T})) := 0$.

Starting with $\mu_0 := \mu$ defined on $\mathfrak{M}$ we then proceed as in example 5.1. In order to construct a measure in $\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ we make sure that $\mu_i \in \Delta_{\mathcal{T}}\mathfrak{A}_i$ for every $i \geq 0$. This can be done by modifying the definition of μ_{i+1} from example 5.1 as follows:

$$\mu_{i+1}(\mathrm{A}_{i+1,j}) := \begin{cases} 0 & \text{if } \ \mathcal{T} \models \mathrm{A}_{i+1,j} = 0 \\ \mu_i(\mathrm{A}_{i,k}) & \text{if } \ \mathcal{T} \models \overline{\mathrm{A}_{i+1,j}} = 0 \\ \frac{1}{2}\mu_i(\mathrm{A}_{i,k}) & \text{if } \ \mathcal{T} \not\models \mathrm{A}_{i+1,j} = 0 \text{ and } \mathcal{T} \not\models \overline{\mathrm{A}_{i+1,j}} = 0. \end{cases}$$

If $\mathcal{T} \models \mathrm{A}_{i+1,j} = 0$ and $\mathcal{T} \models \overline{\mathrm{A}_{i+1,j}} = 0$, then $\mathcal{T} \models \mathrm{A}_{i,k} = 0$, and (by induction hypothesis) $\mu_i(\mathrm{A}_{i,k}) = 0$, so that μ_{i+1} is well-defined. □

When we use representations of Gen and Bel on different algebras $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}(\mathcal{T})$, we need to be able to determine what set of probability distributions is induced on $\mathfrak{N}(\mathcal{T})$ by Gen: in order to compute $\pi^*_{Bel}(\mu)$ with $\mu \in Gen$, for instance, we must have $\mu\!\upharpoonright\!\mathfrak{N}(\mathcal{T})$, which is not given directly by a representation of $Gen\!\upharpoonright\!\mathfrak{M}(\mathcal{T})$. More generally, given $Gen\!\upharpoonright\!\mathfrak{M}(\mathcal{T})$, we shall have to find $Gen\!\upharpoonright\!\mathfrak{N}(\mathcal{T})$, which is just $\mathrm{T}(Gen, \mathfrak{N}(\mathcal{T}))$ in the terminology introduced by the following definition.

Definition 5.8 Let $\mathcal{T}$ be a set of terminological axioms, $\mathrm{M} \subseteq \Delta\mathfrak{M}(\mathcal{T})$. We define the *transformation of* M *for* $\mathfrak{N}(\mathcal{T})$:

$$\begin{aligned}&\mathrm{T}(\mathrm{M}, \mathfrak{N}(\mathcal{T})) := \\ &\quad \{\nu \in \Delta\mathfrak{N}(\mathcal{T}) \mid \exists \mu \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R}) : \ \mu\restriction\mathfrak{M}(\mathcal{T}) \in \mathrm{M}, \ \mu\restriction\mathfrak{N}(\mathcal{T}) = \nu\}\end{aligned}$$

Specifically, for $\mathfrak{M}(\mathcal{T}) \subseteq \mathfrak{N}(\mathcal{T})$, $\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$ is the set of all extensions of measures in M to $\mathfrak{N}(\mathcal{T})$, and for $\mathfrak{M}(\mathcal{T}) \supseteq \mathfrak{N}(\mathcal{T})$, $\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$ is just $\mathrm{M}\restriction\mathfrak{N}(\mathcal{T})$. We write $\mathrm{T}(\mu,\mathfrak{N}(\mathcal{T}))$ for $\mathrm{T}(\{\mu\},\mathfrak{N}(\mathcal{T}))$.

In order to determine whether a pair of measures $\mu \in \Delta\mathfrak{M}(\mathcal{T})$, $\nu \in \Delta\mathfrak{N}(\mathcal{T})$ have a common extension to $\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, it is in fact sufficient to consider common extensions of μ and ν to only the (finite) algebra $\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T})$ generated by $\mathfrak{M}$ and $\mathfrak{N}$, relativized by $\mathcal{T}$:

Theorem 5.9

$$\begin{aligned}&\mathrm{T}(\mathrm{M}, \mathfrak{N}(\mathcal{T})) = \\ &\quad \{\nu \in \Delta\mathfrak{N}(\mathcal{T}) \mid \exists \mu \in \Delta\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T}) : \ \mu\restriction\mathfrak{M}(\mathcal{T}) \in \mathrm{M}, \ \mu\restriction\mathfrak{N}(\mathcal{T}) = \nu\}\end{aligned}$$

Proof: Any $\mu \in \Delta\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T})$ can be extended to $\Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ in the way described in the proof of lemma 5.7. □

Example 5.10 Consider subalgebras $\mathfrak{M}$ and $\mathfrak{N}$ as illustrated in figure 4. The structure of $\mathfrak{M}$ here is represented by the shapes drawn with solid lines comprising the four atoms $\mathrm{A}_1, \ldots, \mathrm{A}_4$. $\mathfrak{N}$ is depicted by dotted lines and consists of only two atoms, B_1 and B_2. The algebra $\mathfrak{A}(\mathfrak{M},\mathfrak{N})$ generated by $\mathfrak{M}$ and $\mathfrak{N}$ has the six atoms $\mathrm{C}_1, \ldots, \mathrm{C}_6$. Let

$$\mu^0 = (0, 0, 0.5, 0.5) \quad \text{and} \quad \mu^1 = (0.25, 0.25, 0.25, 0.25)$$

be probability measures on $\mathfrak{M}$, and $\mathrm{M}{:=}\mathit{conv}\,\{\mu^0, \mu^1\}$. The only extension of μ^0 to $\mathfrak{A}(\mathfrak{M},\mathfrak{N})$ is

$$\alpha^0 = (0, 0, 0, 0.5, 0.5, 0).$$

Since $\alpha^0\restriction\mathfrak{N} = (0, 1)$, we have $\mathrm{T}(\mu^0, \mathfrak{N})=\{(0, 1)\}$.

μ^1 allows for various extensions to $\mathfrak{A}(\mathfrak{M},\mathfrak{N})$. Among them are

$$\begin{aligned}\alpha^1 &= (0.25, 0.25, 0, 0.25, 0.25, 0), \quad & \alpha^2 &= (0, 0, 0.25, 0.25, 0.25, 0.25),\\ \alpha^3 &= (0, 0.25, 0, 0.25, 0.25, 0.25), \quad & \alpha^4 &= (0.25, 0, 0.25, 0.25, 0.25, 0),\end{aligned}$$

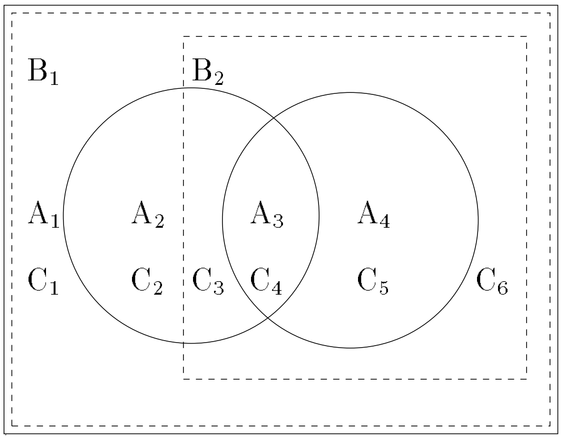

Figure 4: Finite subalgebras in example 5.10

with $\alpha^1 \upharpoonright \mathfrak{N} = (0.5, 0.5)$, $\alpha^2 \upharpoonright \mathfrak{N} = (0, 1)$, and $\alpha^3 \upharpoonright \mathfrak{N} = \alpha^4 \upharpoonright \mathfrak{N} = (0.25, 0.75)$. Thus,

$$\{(0.5, 0.5), (0.25, 0.75), (0, 1)\} \subseteq \mathrm{T}(\mathrm{M}, \mathfrak{N}).$$

Intuitively it seems clear that in fact $\mathrm{T}(\mathrm{M}, \mathfrak{N}) = \mathit{conv}\,\{(0.5, 0.5), (0, 1)\}$, because these measures were obtained by extending the vertices of M in "extreme" ways to measures on $\mathfrak{A}(\mathfrak{M},\mathfrak{N})$. Theorem 5.14 below, the preparation for which we now turn to, provides a formal result verifying this intuition.

Lemma 5.11 Let $\mathfrak{M}, \mathfrak{N} \subset \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be finite, $\mathcal{T}$ a set of terminological axioms. Let $\mu^0, \mu^1 \in \Delta\mathfrak{M}(\mathcal{T})$, $\lambda \in [0, 1]$. Then,

$$\mathrm{T}(\lambda\mu^0 + (1-\lambda)\mu^1, \mathfrak{N}(\mathcal{T})) = \lambda\mathrm{T}(\mu^0, \mathfrak{N}(\mathcal{T})) + (1-\lambda)\mathrm{T}(\mu^1, \mathfrak{N}(\mathcal{T})).$$

The expression on the right side of this equation naturally stands for the set

$$\{\lambda\nu^0 + (1-\lambda)\nu^1 \mid \nu^0 \in \mathrm{T}(\mu^0, \mathfrak{N}(\mathcal{T})),\ \nu^1 \in \mathrm{T}(\mu^1, \mathfrak{N}(\mathcal{T}))\}.$$

Proof: The inclusion from right to left is easy to prove:
let $\nu \in \lambda\mathrm{T}(\mu^0, \mathfrak{N}(\mathcal{T})) + (1-\lambda)\mathrm{T}(\mu^1, \mathfrak{N}(\mathcal{T}))$. By definition this means that for some $\alpha^0,\ \alpha^1 \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R}):\ \alpha^i \upharpoonright \mathfrak{M}(\mathcal{T}) = \mu^i$ $(i = 0, 1)$, and

$$\lambda(\alpha^0 \upharpoonright \mathfrak{N}(\mathcal{T})) + (1-\lambda)(\alpha^1 \upharpoonright \mathfrak{N}(\mathcal{T})) = \nu.$$

For $\alpha := \lambda\alpha^0 + (1-\lambda)\alpha^1$ we then have $\alpha \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, $\alpha\restriction\mathfrak{M}(\mathcal{T}) = \lambda\mu^0 + (1-\lambda)\mu^1$, and $\alpha\restriction\mathfrak{N}(\mathcal{T}) = \nu$. Therefore, $\nu \in \mathrm{T}(\lambda\mu^0 + (1-\lambda)\mu^1, \mathfrak{N}(\mathcal{T}))$.

For the inclusion from left to right assume that $\nu \in \mathrm{T}(\lambda\mu^0 + (1-\lambda)\mu^1, \mathfrak{N}(\mathcal{T}))$, i.e. there is an $\alpha \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ such that $\alpha\restriction\mathfrak{M}(\mathcal{T}) = \lambda\mu^0 + (1-\lambda)\mu^1$, and $\alpha\restriction\mathfrak{N}(\mathcal{T}) = \nu$.

We have to show that for suitable $\alpha^0, \alpha^1 \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, $\alpha^i\restriction\mathfrak{M}(\mathcal{T}) = \mu^i$ $(i = 0, 1)$, and $\lambda\alpha^0 + (1-\lambda)\alpha^1 = \alpha$ holds. Then

$$\begin{aligned} \nu = \alpha\restriction\mathfrak{N}(\mathcal{T}) &= \lambda\alpha^0\restriction\mathfrak{N}(\mathcal{T}) + (1-\lambda)\alpha^1\restriction\mathfrak{N}(\mathcal{T}) \\ &\in \lambda\mathrm{T}(\mu^0, \mathfrak{N}(\mathcal{T})) + (1-\lambda)\mathrm{T}(\mu^1, \mathfrak{N}(\mathcal{T})). \end{aligned}$$

The α^i are obtained by applying Jeffrey's rule to α and the (disjoint) constraints $\alpha^i(\mathrm{A}_j) = \mu^i(\mathrm{A}_j)$ $(i = 0, 1;\ j = 1, \ldots, n)$, where $\mathrm{A}_1, \ldots, \mathrm{A}_n$ are the atoms of $\mathfrak{M}(\mathcal{T})$. Thus, for every $\mathrm{C}\in\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, and $i \in \{0, 1\}$:

$$\alpha^i(\mathrm{C}) = \sum_{j=1}^{n} \mu^i(\mathrm{A}_j)\alpha(\mathrm{C} \mid \mathrm{A}_j).$$

Clearly, $\alpha^i \in \Delta_{\mathcal{T}}\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$, and $\alpha^i\restriction\mathfrak{M}(\mathcal{T}) = \mu^i$ holds. Also, for $\mathrm{C}\in \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$

$$\begin{aligned} (\lambda\alpha^0 + (1-\lambda)\alpha^1)(\mathrm{C}) &= \sum_{j=1}^{n} \Big(\lambda\mu^0(\mathrm{A}_j)\alpha(\mathrm{C} \mid \mathrm{A}_j) + (1-\lambda)\mu^1(\mathrm{A}_j)\alpha(\mathrm{C} \mid \mathrm{A}_j)\Big) \\ &= \sum_{j=1}^{n} (\lambda\mu^0(\mathrm{A}_j) + (1-\lambda)\mu^1(\mathrm{A}_j))\alpha(\mathrm{C} \mid \mathrm{A}_j) \\ &= \sum_{j=1}^{n} \alpha(\mathrm{A}_j)\alpha(\mathrm{C} \mid \mathrm{A}_j) \\ &= \alpha(\mathrm{C}). \end{aligned}$$

□

Theorem 5.12 Let $\mathfrak{M}, \mathfrak{N} \subset \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be finite, $\mathcal{T}$ a set of terminological axioms. Let $\mathrm{M}=conv\{\mu^1, \ldots, \mu^m\} \subseteq \Delta\mathfrak{M}(\mathcal{T})$. Then,

$$\mathrm{T}(\mathrm{M}, \mathfrak{N}(\mathcal{T})) = conv\{\mathrm{T}(\mu^i, \mathfrak{N}(\mathcal{T})) \mid\ i = 1, \ldots, m\}.$$

Proof: By induction on m: For $m = 1$ the theorem trivially holds. Let $m > 1$, and assume that the theorem holds for $\mathrm{M}^* := conv\{\mu^1, \ldots, \mu^{m-1}\}$.

$$\begin{aligned}
\mathrm{T}(\mathrm{M}, \mathfrak{N}(\mathcal{T})) &= \mathrm{T}(conv\{\mu^m, \mathrm{M}^*\}, \mathfrak{N}(\mathcal{T})) \\
&= \{\mathrm{T}(\lambda\mu^m + (1-\lambda)\mu^*, \mathfrak{N}(\mathcal{T})) \mid \mu^* \in \mathrm{M}^*,\ \lambda \in [0,1]\} \\
&= \{\lambda\mathrm{T}(\mu^m, \mathfrak{N}(\mathcal{T})) + (1-\lambda)\mathrm{T}(\mu^*, \mathfrak{N}(\mathcal{T})) \mid \mu^* \in \mathrm{M}^*,\ \lambda \in [0,1]\} \\
&\quad \text{(by lemma 5.11)} \\
&= conv\{\mathrm{T}(\mu^m, \mathfrak{N}(\mathcal{T})), \mathrm{T}(\mathrm{M}^*, \mathfrak{N}(\mathcal{T}))\} \\
&= conv\{\mathrm{T}(\mu^i, \mathfrak{N}(\mathcal{T})) \mid i = 1, \ldots, m\} \\
&\quad \text{(by the induction hypothesis)}
\end{aligned}$$

□

Next, we show that $\mathrm{T}(\mu,\mathfrak{N}(\mathcal{T}))$ is the convex hull of finitely many points, and how these points are obtained from μ. Following the lead of example 5.10, we first show that $\mathrm{T}(\mu, \mathfrak{A}(\mathfrak{M}, \mathfrak{N})(\mathcal{T}))$ is the convex hull of the "extreme" extensions of μ to $\mathfrak{A}(\mathfrak{M}, \mathfrak{N})(\mathcal{T})$.

Let $\{\mathrm{A}_1, \ldots, \mathrm{A}_n\}$ and $\{\mathrm{A}'_1, \ldots, \mathrm{A}'_k\}$ be the sets of atoms of $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}(\mathcal{T})$, respectively. The atoms of $\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T})$, then, are the conjunctions $\mathrm{A}_i \wedge \mathrm{A}'_j$ with $\mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_j = 0$.

For every $i \in \{1, \ldots, n\}$ there exists at least one $j \in \{1, \ldots, k\}$ such that $\mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_j = 0$ (otherwise $\mathcal{T} \models \mathrm{A}_i \subseteq \neg\mathrm{A}'_1 \wedge \ldots \wedge \neg\mathrm{A}'_k$, $\mathcal{T} \models \neg\mathrm{A}'_1 \wedge \ldots \wedge \neg\mathrm{A}'_k = 0$, hence $\mathcal{T} \models \mathrm{A}_i = 0$, a contradiction). Let

$$\begin{aligned}
\mathrm{I} &:= \{(i,j) \in \{1, \ldots, n\} \times \{1, \ldots, k\} \mid \mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_j = 0\}, \\
\mathscr{J} &:= \{J : \{1, \ldots, n\} \to \{1, \ldots, k\} \mid \forall i\, (i, J(i)) \in \mathrm{I}\}.
\end{aligned}$$

Every $J \in \mathscr{J}$ defines an extension $\mu(J)$ of μ to $\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T})$ via

$$\mu(J)(\mathrm{A}_i \wedge \mathrm{A}'_j) = \begin{cases} \mu(\mathrm{A}_i) & \text{if } J(i) = j, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In example 5.10 α^1 is $\mu(J)$ for $J : 1 \mapsto 1$, $2 \mapsto 2$, $3 \mapsto 4$, and $4 \mapsto 3$.

Lemma 5.13 $\mathrm{T}(\mu,\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T}))=conv\{\mu(J) \mid J \in \mathscr{J}\}$

Proof: The inclusion from right to left is trivial: every $\mu(J)$ lies in the convex set $\mathrm{T}(\mu,\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T}))$.

For the inclusion from left to right let $\mu^* \in \mathrm{T}(\mu, \mathfrak{A}(\mathfrak{M}, \mathfrak{N})(\mathcal{T}))$. μ^* then is of the form

$$\mu^*(\mathrm{A}_i \wedge \mathrm{A}'_j) = \lambda_{i,j}\mu(\mathrm{A}_i) \quad ((i,j) \in \mathrm{I})$$

with

$$\sum_{\{j|(i,j)\in \mathrm{I}\}} \lambda_{i,j} = 1$$

for all $i \in \{1,\ldots,n\}$. In order to prove that $\mu^* \in conv\{\mu(J) \mid J \in \mathscr{J}\}$ we show that

$$\mu^* = \sum_{J\in\mathscr{J}} \left(\prod_{i=1}^{n} \lambda_{i,J(i)}\right) \mu(J). \tag{13}$$

By a simple induction on n it is readily verified that

$$\begin{aligned} \sum_{J\in\mathscr{J}} \left(\prod_{i=1}^{n} \lambda_{i,J(i)}\right) &= 1, \qquad \text{and} \\ \sum_{J\in\mathscr{J},J(i)=j} \left(\prod_{i=1}^{n} \lambda_{i,J(i)}\right) &= \lambda_{i,j} \quad \text{for all } (i,j)\in \mathrm{I}. \end{aligned}$$

Thus, the right hand side of (13) is a convex combination $\bar{\mu}$ of the $\mu(J)$ $(J \in \mathscr{J})$, and

$$\bar{\mu}(\mathrm{A}_i \wedge \mathrm{A}'_j) = \sum_{J\in\mathscr{J},J(i)=j} \left(\prod_{i=1}^{n} \lambda_{i,J(i)}\right) \mu(\mathrm{A}_i) = \lambda_{i,j}\mu(\mathrm{A}_i).$$

Hence, $\bar{\mu} = \mu^*$, which proves (13). □

From $\mathrm{T}(\mu,\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T}))$ it is only a small step to $\mathrm{T}(\mu,\mathfrak{N}(\mathcal{T}))$:

$$\begin{aligned} \mathrm{T}(\mu,\mathfrak{N}(\mathcal{T})) &= \mathrm{T}(\mu,\mathfrak{A}(\mathfrak{M},\mathfrak{N})(\mathcal{T}))\restriction\mathfrak{N}(\mathcal{T}) \\ &= conv\{\mu(J) \mid J\in\mathscr{J}\}\restriction\mathfrak{N}(\mathcal{T}) \\ &= conv\{\mu(J)\restriction\mathfrak{N}(\mathcal{T}) \mid J\in\mathscr{J}\} \end{aligned}$$

with

$$\mu(J)\restriction\mathfrak{N}(\mathcal{T})(\mathrm{A}'_j) = \sum_{\{i|J(i)=j\}} \mu(\mathrm{A}_i).$$

Collecting our results we get

Theorem 5.14 Let $\mathfrak{M}$, $\mathfrak{N} \subset \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ be finite, $\mathcal{T}$ a set of terminological axioms, $\{\mathrm{A}_1,\ldots,\mathrm{A}_n\}$ and $\{\mathrm{A}'_1,\ldots,\mathrm{A}'_k\}$ the sets of atoms of $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}(\mathcal{T})$. Let $\mathrm{M}=conv\{\mu^1,\ldots,\mu^m\} \subseteq \Delta\mathfrak{M}(\mathcal{T})$.

Then $\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$ is the convex hull of measures ν on $\mathfrak{N}(\mathcal{T})$ that are defined by

$$\nu(\mathrm{A}'_j) = \sum_{\{i|J(i)=j\}} \mu^l(\mathrm{A}_i), \tag{14}$$

where $l \in \{1,\ldots,m\}$, and $J : \{1,\ldots,n\} \to \{1,\ldots,k\}$ is any function with $\mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_{J(i)} = 0$ for all $i \in \{1,\ldots,n\}$.

Note that not every ν defined by (14) for some μ^l and $J(\cdot)$ needs to really be a vertex of $\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$. Only, every vertex of $\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$ is of this form.

For future reference we note one more lemma:

Lemma 5.15 Let $\mathfrak{M}$, $\mathfrak{N}$, and M as in theorem 5.14. Then

$$int\,\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T})) \subseteq \mathrm{T}(int\,\mathrm{M},\mathfrak{N}(\mathcal{T})).$$

Proof: By theorem 5.14 there exist measures $\nu^{i,j}$ such that

$$\mathrm{T}(\mu^i,\mathfrak{N}(\mathcal{T})) = conv\{\nu^{i,j} \mid j = 1,\ldots,l_i\} \quad (i = 1,\ldots,m).$$

Let $\mu \in int\,\mathrm{T}(\mathrm{M},\mathfrak{N}(\mathcal{T}))$. For μ exists a representation

$$\mu = \sum_{\substack{i=1,\ldots,m \\ j=1,\ldots,l_i}} \lambda_{i,j}\nu^{i,j}$$

with $\lambda_{i,j} > 0$ for all i,j. Hence

$$\mu \in \sum_{i=1,\ldots,m} \lambda_i \mathrm{T}(\mu^i,\mathfrak{N}(\mathcal{T}))$$

for $\lambda_i := \sum_{j=1,\ldots,l_i} \lambda_{i,j} > 0$, and by lemma 5.11 (extended to convex combinations of arbirtrary length):

$$\mu \in \mathrm{T}(\sum_{i=1,\ldots,m} \lambda_i\mu^i,\mathfrak{N}(\mathcal{T}))$$

with $\sum_{i=1,\ldots,m} \lambda_i\mu^i \in int\,\mathrm{M}$. □

5.4 Probabilistic inferences in $\mathcal{PALC}$

In the previous section the relationship between polytopes of probability measures defined on different finite subalgebras of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ has been explored. In the present section we show that all the probabilistic consequences from a $\mathcal{PALC}$ knowledge base can be inferred by only working with such polytopes.

Let $\mathcal{KB} = \mathcal{T} \cup \mathcal{PT} \cup \bigcup\{\mathcal{P}_a \mid a \in \mathrm{S_O}\}$ be a $\mathcal{PALC}$-knowledge base. Let $\mathfrak{M}$, $\mathfrak{N}_a$ be the subalgebras of $\mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ generated by the concept terms occuring in $\mathcal{PT}$ and $\mathcal{P}_a$, respectively.

First, we turn to the question of consistency of $\mathcal{KB}$. Just as for $\mathcal{PCL}$, an inconsistency of $\mathcal{KB}$ can only be due to one of the following three causes.

a* $\mathcal{T}$ is inconsistent.

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

c* For all $\mu \in Gen(\mathcal{KB})$ there exists $a \in \mathrm{S_O}$ such that $\pi^*_{Bel_a(\mathcal{KB})}(\mu)$ is not defined.

(cf. page 24).

Lemma 5.7 tells us that

$$Gen = \emptyset \ \text{ iff } \ \{\mu \in \Delta\mathfrak{M}(\mathcal{T}) \mid \ \mu \text{ consistent with } \mathcal{PT}\} = \emptyset,$$

so that we can check whether $Gen = \emptyset$ by only considering probability measures satisfying the given constraints on the finite algebra $\mathfrak{M}(\mathcal{T})$ in precisely the same manner as we did for $\mathcal{PCL}$. Analogously for Bel_a.

By the following theorem, testing for an inconsistency caused by **c*** can be reduced to the finite case treated in theorem 4.8.

Theorem 5.16 The following are equivalent:

(i) $\forall\mu \in Gen\ \exists a \in \mathrm{S_O} : \pi^*_{Bel_a}(\mu)$ is undefined.

(ii) With $Gen\restriction\mathfrak{M}(\mathcal{T}) = conv\{\mu^1, \ldots, \mu^k\}$ and $\bar{\mu} := \frac{1}{k}(\mu^1 + \ldots + \mu^k)$:
$\exists a \forall \mu \in \mathrm{T}(\bar{\mu}, \mathfrak{N}_a(\mathcal{T}))$: $\pi_{Bel_a\restriction\mathfrak{N}_a(\mathcal{T})}(\mu)$ is undefined.

Proof: By the definition of π^*, (i) is equivalent to: $\forall\mu \in Gen\, \exists a \in \mathrm{S_O}$ such that $\pi_{Bel_a\restriction\mathfrak{N}_a(\mathcal{T})}(\mu\restriction\mathfrak{N}_a(\mathcal{T}))$ is not defined.

Since $\mathrm{Z}(\bar{\mu})$ is minimal in $Gen\restriction\mathfrak{M}(\mathcal{T})$, $\bar{\mu}$ can be extended to $\mu^* \in Gen$ such that $\mathrm{Z}(\mu^*\restriction\mathfrak{N}_a(\mathcal{T}))$ is minimal in $Gen\restriction\mathfrak{N}_a(\mathcal{T})$ simultaneously for all a. Therefore, given such a μ^*, (i) is equivalent to: $\exists a \in \mathrm{S_O}$ such that $\pi_{Bel_a\restriction\mathfrak{N}_a(\mathcal{T})}(\mu^*\restriction\mathfrak{N}_a(\mathcal{T}))$ is not defined, which, in turn, is equivalent to (ii), because $\mu^*\restriction\mathfrak{N}_a(\mathcal{T})$ is minimal with respect to $\mathrm{Z}(\cdot)$ in $\mathrm{T}(\bar{\mu}, \mathfrak{N}_a(\mathcal{T}))$. □

For every $a \in \mathrm{S_O}$, a test for (ii) in theorem 5.16 is now given by theorem 4.8 with M=$\mathrm{T}(\bar{\mu}, \mathfrak{N}_a(\mathcal{T}))$ and N=$Bel_a\restriction\mathfrak{N}_a(\mathcal{T})$.

Theorem 5.17 Let $\mathcal{KB}$ be consistent, and $\mathcal{KB} \models \mathrm{P(C|D) = J}$. Let $\mathfrak{M}_{\mathrm{C,D}}$ be the algebra generated by $\mathfrak{M}$, C and D. Let

$$\mathrm{T}(Gen\restriction\mathfrak{M}(\mathcal{T}), \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})) = conv\{\mu^1, \ldots, \mu^k\}.$$

Then, either $\mu^i(\mathrm{D}) = 0$ for $i = 1, \ldots, k$ and J=$\emptyset$, or J is an interval and

$$\begin{aligned} \inf \mathrm{J} &= \min\{\mu^i(\mathrm{C} \mid \mathrm{D}) \mid \mu^i(\mathrm{D}) > 0,\ i = 1, \ldots, k\}, \\ \sup \mathrm{J} &= \max\{\mu^i(\mathrm{C} \mid \mathrm{D}) \mid \mu^i(\mathrm{D}) > 0,\ i = 1, \ldots, k\}. \end{aligned}$$

Proof: By definition,

$$\mathrm{J} = Gen^* \upharpoonright \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})(\mathrm{C} \mid \mathrm{D}).$$

$Gen^* \upharpoonright \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T})$ differs from the polytope $\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T}))$ at most by points on the boundary of $\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T}))$, because by the same argument used in the proof of theorem 5.16, every point in $int\, \mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{M}_{\mathrm{C,D}}(\mathcal{T}))$ can be extended to an element of Gen^*. From this point onwards, we can argue just as in the proof of theorem 4.11. □

Theorem 5.18 Let $\mathcal{KB} \models \mathrm{P}(a \in \mathrm{C}) = \mathrm{J}$. Then

$$\mathrm{J} \subseteq \pi_{Bel_a \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})}(\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})))(\mathrm{C}) \subseteq cl\, \mathrm{J}$$

with $\mathfrak{N}_{a,\mathrm{C}}$ the algebra generated by $\mathfrak{N}_a$ and C.

Proof: By definition of J and π^*,

$$\begin{aligned} \mathrm{J} &= \pi^*_{Bel_a(\mathcal{KB})}(Gen^*)(\mathrm{C}) \\ &= \pi_{Bel_a(\mathcal{KB}) \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})}(Gen^* \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}))(\mathrm{C}). \end{aligned} \tag{15}$$

Since

$$Gen^* \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}) \subseteq Gen \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}) = \mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})),$$

this already proves the first inclusion stated in the theorem. For the second inclusion, we show that

$$\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})) \subseteq cl\,(Gen^* \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})), \tag{16}$$

which in conjunction with (15) yields the desired result, because for the continuous function $\pi_{\ldots}(\cdot)(\mathrm{C})$ and every $\mathrm{M} \subseteq \Delta\mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})$

$$\pi_{\ldots}(cl\, \mathrm{M})(\mathrm{C}) \subseteq cl\, \pi_{\ldots}(\mathrm{M})(\mathrm{C})$$

holds. By an application of lemma 5.15 and the fact that every element in $int\, Gen \upharpoonright \mathfrak{M}(\mathcal{T})$ can be extended to an element of Gen^*, we have

$$int\, \mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})) \subseteq \mathrm{T}(int\, Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})) \subseteq Gen^* \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}).$$

Thus,

$$cl\, int\, \mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T})) \subseteq cl\, Gen^* \upharpoonright \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}),$$

which shows (16) because $\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}))$, being a polytope, is the same as $cl\, int\, \mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{a,\mathrm{C}}(\mathcal{T}))$. □

5.5 An example

This section is dedicated to the discussion of an example that demonstrates the application of the results from sections 5.3 and 5.4.

The $\mathcal{PALC}$-knowledge base given in table 7 describes some features of a blocks world, of ornithology, and of a penguin Opus who visits the blocks world in order to select one block as nesting material.

The axioms in $\mathcal{T}$ are fairly self-explanatory. The clauses *disjoint*(...) are used to express the disjointness in pairs of the concepts in their argument with a little more clarity, than would be achieved if this was done within the strict syntax of $\mathcal{PALC}$. $\mathcal{PT}$ expresses some statistical information about the colours of free blocks, about the flying abilities of birds, and about the conditional probability of an object that is only red to be in fact purely light red.

Note that it is impossible to express in $\mathcal{PALC}$ that an object has exactly one colour, and that the concept $\forall\mathsf{has\_colour}\colon \mathsf{Red}$ is consistent with $\forall\mathsf{has\_colour}\colon \neg\mathsf{Red}$, because the conjunction of these two just describes objects that do not have any colour at all.

The two objects mentioned in $\mathcal{KB}$ are known to be a penguin and a block respectively. The latter one also being a free block with a probability of 0.8.

Suppose, now, that we want to determine the probability p that *Opus_choice* is a block that is either blue or light red (for a block, the two concepts $\forall\mathsf{has\_colour}\colon \mathsf{Blue}$ and $\forall\mathsf{has\_colour}\colon \mathsf{Red}$ are inconsistent, because a block is defined to have at least one colour, and the two concepts Blue and Red are mutually exclusive).

Intuitively we may reason as follows: a lower bound for the probability p is obtained by assuming that none of the blocks that are not free are blue or light red, and that all the free blocks that are red are not light red. This last

Table 7: A knowledge base

$$
\begin{array}{lll}
\mathcal{T} = & \mathsf{Block} & \subseteq \exists \mathsf{has\_colour}\colon \mathsf{Colour} \wedge \exists\ \mathsf{has\_position}\colon \mathsf{Position} \\
& \mathsf{Free\_block} & = \mathsf{Block} \wedge \neg\exists\ \mathsf{below}\colon \mathsf{Block} \\
& \mathsf{Blue} & \subseteq \mathsf{Colour} \\
& \mathsf{Red} & \subseteq \mathsf{Colour} \\
& \mathsf{Light\_red} & \subseteq \mathsf{Red} \\
& \mathsf{Bird} & \subseteq \exists\ \mathsf{has\_colour}\colon \mathsf{Colour} \\
& \mathsf{Flying\_bird} & \subseteq \mathsf{Bird} \\
& \mathsf{Penguin} & \subseteq \mathsf{Bird} \wedge \neg \mathsf{Flying\_bird} \\
& \multicolumn{2}{l}{disjoint(\mathsf{Block},\ \mathsf{Colour},\ \mathsf{Position},\ \mathsf{Bird})} \\
& \multicolumn{2}{l}{disjoint(\mathsf{Blue},\ \mathsf{Red})} \\
\mathcal{PT} = & \multicolumn{2}{l}{\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Red} | \mathsf{Free\_block}) = 0.5} \\
& \multicolumn{2}{l}{\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Blue} | \mathsf{Free\_block}) = 0.01} \\
& \multicolumn{2}{l}{\mathrm{P}(\mathsf{Flying\_bird} | \mathsf{Bird}) = 0.95} \\
& \multicolumn{2}{l}{\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Light\_red} | \forall \mathsf{has\_colour} : \mathsf{Red}) = 0.5} \\
\mathcal{P}_{Opus} = & \multicolumn{2}{l}{\mathrm{P}(Opus \in \mathsf{Penguin}) = 1} \\
\mathcal{P}_{Opus\_choice} = & \multicolumn{2}{l}{\mathrm{P}(Opus\_choice \in \mathsf{Block}) = 1} \\
& \multicolumn{2}{l}{\mathrm{P}(Opus\_choice \in \mathsf{Free\_block}) = 0.8}
\end{array}
$$

possibility has to be taken into account, because we do not know whether the general information $\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Light\_red} | \forall \mathsf{has\_colour} : \mathsf{Red}) = 0.5$ also applies to free blocks. Hence, the lower bound for p is given by 0.8×0.01 that is identified as the probability that *Opus_choice* is a blue free block.

An upper bound for p is obtained by assuming conversely that every block that is not free is either blue or light red, and that every red free block is in fact of a light red. In this case p is estimated as the sum of 0.2 (the probability that $Opus\_choice \in \mathsf{Block} \wedge \neg \mathsf{Free\_block}$) and $0.8(0.5 + 0.01) = 0.408$.

Interpreting the information given in $\mathcal{PT}$ cautiously, we therefore arrive at the interval [0.008,0.608] as a plausible range for p.

Being asked for a rather more specific estimate, we would probably argue that a good assumption to make is that

$$\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Light\_red} | \forall \mathsf{has\_colour} : \mathsf{Red} \wedge \ \mathsf{Free\_block}) = \mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Light\_red} | \forall \mathsf{has\_colour} : \mathsf{Red}) \quad (17)$$

which implies $\mathrm{P}(\forall \mathsf{has\_colour} : \mathsf{Light\_red} | \mathsf{Free\_block}) = 0.25$, and the lower and upper bound for p would change to 0.8(0.25+0.01)=0.208 and 0.2+0.208=0.408 respectively.

We now demonstrate how inferences from $\mathcal{KB}$ are made using our semantics for $\mathcal{PALC}$ and the effective procedures outlined in the previous sections.

Using the abbreviations BLr for $\forall \mathsf{has\_colour} : (\mathsf{Blue} \vee \mathsf{Light\_red})$ and Oc for *Opus_choice*, we have to answer the query

$$\mathrm{P}(Oc \in \mathsf{BLr}) = ?$$

By theorem 5.18 we know that this query is essentially answered by computing the interval

$$\mathrm{J}_0 := \pi_{Bel_{Oc}\upharpoonright \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})}(\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))(\mathsf{BLr}),$$

where $\mathfrak{N}_{Oc,\mathsf{BLr}}$ is the algebra generated by $\{\mathsf{Block}, \ \mathsf{Free\_block}, \ \mathsf{BLr}\}$, and $\mathfrak{M}$ is the algebra generated by the concept terms appearing in $\mathcal{PT}$.

Figure 5 illustrates both the structure of $\mathfrak{M}(\mathcal{T})$, represented by the shapes drawn with solid lines, and the structure of $\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$ represented by shapes with dotted lines. $\mathfrak{M}(\mathcal{T})$ has 17 atoms $\{\mathrm{A}_1, \ldots, \mathrm{A}_{17}\}$ (out of the $2^6 = 64$ atoms of $\mathfrak{M}$). $\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$ has 6 atoms $\{\mathrm{A}'_{\mathrm{I}}, \ldots, \mathrm{A}'_{\mathrm{VI}}\}$ (out of 8). The atoms of $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$ are indicated by their indices in figure 5.

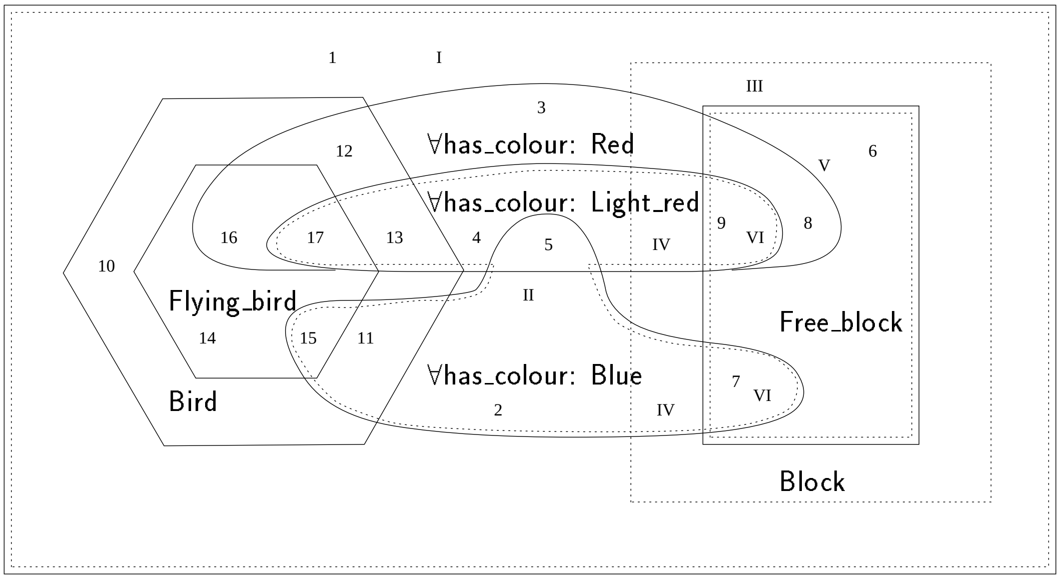

Figure 5: The algebras $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$

The sets $Gen \upharpoonright \mathfrak{M}(\mathcal{T})$, $Bel_{Opus} \upharpoonright \mathfrak{N}_{Opus}(\mathcal{T})$ and $Bel_{Oc} \upharpoonright \mathfrak{N}_{oc}(\mathcal{T})$ are nonempty. $Gen \upharpoonright \mathfrak{M}(\mathcal{T})$ is computed to be a polytope with 52 vertices, and $\mathrm{Z}(\bar{\mu})=\emptyset$ for interior points $\bar{\mu}$. For such $\bar{\mu}$ and for every finite subalgebra $\mathfrak{A} \subset \mathfrak{A}(\mathrm{S_C},\mathrm{S_R})$ there is an element $\bar{\mu}' \in \mathrm{T}(\bar{\mu}, \mathfrak{A}(\mathcal{T}))$ with $\mathrm{Z}(\bar{\mu}') = \emptyset$. Hence, $\pi_{Bel_{Opus} \upharpoonright \mathfrak{N}_{Opus}(\mathcal{T})}(\bar{\mu}')$ and $\pi_{Bel_{Oc} \upharpoonright \mathfrak{N}_{Oc}(\mathcal{T})}(\bar{\mu}'')$ are defined for suitable $\bar{\mu}' \in \mathrm{T}(\bar{\mu}, \mathfrak{N}_{Opus})$, $\bar{\mu}'' \in \mathrm{T}(\bar{\mu}, \mathfrak{N}_{Oc})$, and $\mathcal{KB}$ is consistent.

After these preliminaries, we can actually start computing J_0. Since $\mathcal{P}_{Oc}$ is equivalent to the disjoint constraints

$$\mathrm{P}(Oc \in \mathsf{Block} \wedge \neg \mathsf{Free\_block}) = 0.2$$
$$\mathrm{P}(Oc \in \mathsf{Block} \wedge \mathsf{Free\_block}) = 0.8,$$

$\pi_{Bel_{Oc} \upharpoonright \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})}(\mu')(\mathsf{BLr})$ is given by Jeffrey's rule for every $\mu' \in \Delta\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$, i.e.

$$\begin{aligned} \pi_{\ldots}(\mu')(\mathsf{BLr}) &= 0.2\mu'(\mathsf{BLr} \mid \mathsf{Block} \wedge \neg \mathsf{Free\_block}) + \\ &\quad 0.8\mu'(\mathsf{BLr} \mid \mathsf{Block} \wedge \mathsf{Free\_block}) \\ &= 0.2\frac{\mu'_{\mathrm{IV}}}{\mu'_{\mathrm{III}} + \mu'_{\mathrm{IV}}} + 0.8\frac{\mu'_{\mathrm{VI}}}{\mu'_{\mathrm{V}} + \mu'_{\mathrm{VI}}}. \end{aligned} \tag{18}$$

Table 8: Consistent atoms of $\mathfrak{M}(\mathcal{T})$ and $\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$

$i \mid \{j \mid \mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_j = 0\}$

i	j	i	j	i	j	i	j
1	I,III	6	V	11	II	16	I
2	II,IV	7	VI	12	I	17	II
3	I,III	8	V	13	II		
4	II,IV	9	VI	14	I		
5	II	10	I	15	II		

In this example we do not have to rely on a heuristic search in $\mathrm{T}(Gen \upharpoonright \mathfrak{M}(\mathcal{T}), \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ for elements that maximize or minimize (18). A probability measure that maximizes (18) is obtained as follows: we first look at the two terms in (18) separately, and find elements μ^1, μ^2 in $Gen \upharpoonright \mathfrak{M}(\mathcal{T})$ such that these terms become maximal for suitable $\mu'^i \in \mathrm{T}(\mu^i, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ $(i = 1, 2)$. It then turns out that a convex combination of μ'^1 and μ'^2 maximizes both terms in (18) simultaneously.

Table 8 lists for each $i \in \{1, \ldots, 17\}$ the $j \in \{\mathrm{I}, \ldots, \mathrm{VI}\}$ for which $\mathcal{T} \not\models \mathrm{A}_i \wedge \mathrm{A}'_j = 0$. According to table 8 there exists for every $\mu \in Gen$ with

$$\mu_2 + \mu_4 > 0 \tag{19}$$

an element $\mu' \in \mathrm{T}(\mu, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ with

$$\frac{\mu'_{\mathrm{IV}}}{\mu'_{\mathrm{III}} + \mu'_{\mathrm{IV}}} = 1 : \tag{20}$$

choose $J \in \mathscr{J}$ (cf. page 49) with

$$J : 1 \mapsto \mathrm{I},\ 2 \mapsto \mathrm{IV},\ 3 \mapsto \mathrm{I},\ \text{and } 4 \mapsto \mathrm{IV},$$

then $\mu(J)(\mathrm{A}'_{\mathrm{IV}}) > 0$ and $\mu(J)(\mathrm{A}'_{\mathrm{III}}) = 0$, hence (20) holds for

$$\mu' = \mu(J) \upharpoonright \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}).$$

Checking the vertices of $Gen \upharpoonright \mathfrak{M}(\mathcal{T})$, we find the element

$$\mu^1 = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),$$

which satisfies (19), so that (20) holds for

$$\mu'^1 := \mu^1(J)\upharpoonright\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}) = (0,0,0,1,0,0).$$

For the second term in (18) we find by table 8 that for every $\mu \in Gen$ and every $\mu' \in \mathrm{T}(\mu, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$

$$\frac{\mu'_{\mathrm{VI}}}{\mu'_{\mathrm{V}} + \mu'_{\mathrm{VI}}} = \frac{\mu_7 + \mu_9}{\mu_6 + \mu_7 + \mu_8 + \mu_9}$$

holds. We therefore look for a vertex of $Gen\upharpoonright\mathfrak{M}(\mathcal{T})$ that is maximal with respect to $(\mu_7 + \mu_9)/(\mu_6 + \mu_7 + \mu_8 + \mu_9)$ and find that this is the case at (among others) the vertex

$$\mu^2 = (0,0,0,0,0,0.321,0.00655,0,0.32759,0.01724,0,0,0,0,0,0.32759,0)$$

with $(\mu_7^2 + \mu_9^2)/(\mu_6^2 + \mu_7^2 + \mu_8^2 + \mu_9^2) = 0.51$. $\mathrm{T}(\mu^2, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ contains the one element

$$\mu'^2 = (0.3448, 0, 0, 0, 0.321, 0.334).$$

For

$$\mu' := 0.5\mu'^1 + 0.5\mu'^2$$

we then have $\mu' \in Gen\upharpoonright\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ (by lemma 5.11), and

$$\begin{aligned} \pi_{Bel_{Oc}\upharpoonright\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})}(\mu')(\mathsf{BLr}) &= 0.2\tfrac{1}{0+1} + 0.8\tfrac{0.334}{0.321+0.334} \\ &= 0.2 + 0.8 \times 0.51 = 0.608, \end{aligned}$$

so that $\sup \mathrm{J}_0$=0.608. In fact, it is not difficult to see that $\mu' \in Gen^*\upharpoonright\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})$, and therefore 0.608=$\max \mathrm{J}_0$.

Similarly, it can be shown that (18) is minimized in $\mathrm{T}(Gen\upharpoonright\mathfrak{M}(\mathcal{T}), \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$ by

$$\mu'' = (0.00862, 0.16379, 0.5, 0, 0.32431, 0.003275)$$

with

$$0.2\frac{\mu''_{\mathrm{IV}}}{\mu''_{\mathrm{III}} + \mu''_{\mathrm{IV}}} + 0.8\frac{\mu''_{\mathrm{VI}}}{\mu''_{\mathrm{V}} + \mu''_{\mathrm{VI}}} = 0.008.$$

Hence, in the formal framework the same result is obtained as was argued for intuitively.

To conclude the discussion of this example, we examine the result of strengthening our approach by only looking at the maximum entropy distribution in $Gen\restriction\mathfrak{M}(\mathcal{T})$, which is

$$\begin{aligned}\mu^{\mathrm{me}} = (&0.0783, 0.0783, 0.0904, 0.0678, 0.0678, 0.1146, 0.0025, 0.0667, 0.05,\\ &0.0047, 0.0047, 0.0055, 0.0041, 0.0907, 0.0907, 0.1048, 0.0786).\end{aligned}$$

For every $\mu' \in \mathrm{T}(\mu^{\mathrm{me}}, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))$:

$$\frac{\mu'_{\mathrm{VI}}}{\mu'_{\mathrm{V}} + \mu'_{\mathrm{VI}}} = \frac{\mu_7^{\mathrm{me}} + \mu_9^{\mathrm{me}}}{\mu_6^{\mathrm{me}} + \mu_7^{\mathrm{me}} + \mu_8^{\mathrm{me}} + \mu_9^{\mathrm{me}}} = 0.2238,$$

while $\frac{\mu'_{\mathrm{IV}}}{\mu'_{\mathrm{III}}+\mu'_{\mathrm{IV}}}$ can attain any value in [0,1]. Hence

$$\begin{aligned}\pi_{Bel_{Oc}\restriction\mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T})}(\mathrm{T}(\mu^{\mathrm{me}}, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))(\mathsf{BLr}) &\in [0.8 \times 0.2238, 0.2 + 0.8 \times 0.2238]\\ &= [0.179, 0.379],\end{aligned}$$

which slightly differs from the interval [0.208,0.408] that was received by assuming (17). The discrepancy is caused by the somewhat pathological atom

$$\mathrm{A}_5 = \forall\mathsf{has\_colour} : \mathsf{Blue} \wedge \forall\mathsf{has\_colour} : \mathsf{Light\_red}$$

that causes the maximum entropy approach to prefer a probability distribution with

$$\begin{aligned}&\mathrm{P}(\forall\mathsf{has\_colour} : \mathsf{Light\_red}|\forall\mathsf{has\_colour} : \mathsf{Red} \wedge \neg(\mathsf{Bird} \vee \mathsf{Free\_block})) >\\ &\quad\mathrm{P}(\forall\mathsf{has\_colour} : \mathsf{Light\_red}|\forall\mathsf{has\_colour} : \mathsf{Red} \wedge \mathsf{Free\_block}),\end{aligned}$$

because $\forall\mathsf{has\_colour} : \mathsf{Light\_red} \wedge \neg(\mathsf{Bird} \vee \mathsf{Free\_block})$ consists of the two atoms A_4 and A_5, while $\forall\mathsf{has\_colour} : \mathsf{Light\_red} \wedge \mathsf{Free\_block}$ has only the atom A_9. If $\mathrm{P}(\mathrm{A}_5)$ was constrained to be 0, then (17) would hold for μ^{me}, and the result

$$\pi_{Bel_{Oc}\restriction\mathfrak{A}_{Oc,\mathsf{BLr}}(\mathcal{T})}(\mathrm{T}(\mu^{\mathrm{me}}, \mathfrak{N}_{Oc,\mathsf{BLr}}(\mathcal{T}))(\mathsf{BLr}) \in [0.208, 0.408]$$

be obtained.

6 Conclusion

In this work we presented a probabilistic extension for the terminological knowledge representation language $\mathcal{ALC}$. For the resulting language $\mathcal{PALC}$, semantics were defined that model default reasoning about probabilities which takes place when uncertain information about an object is partially completed by taking it to be, to the greatest possible extent, a typical representative of the domain of discourse.

It was shown that for a $\mathcal{PALC}$- knowledge base consistency can be decided, and the generic conditional probabilities it entails be computed. The situation is more difficult with respect to the resulting subjective probabilities assigned to an object. Here, we have to rely, as yet, on search algorithms producing a sequence of intervals converging to the desired interval, but about whose rate of convergence no statement has been made.

Further work should be directed towards the question about how the probability intervals entailed for an object can be computed explicitly, or, at least, be approximated in such a way that the distance between their endpoints and the bounds computed so far is known to substantially decrease with each iteration.

Another open problem that will have to be addressed is to clarify the relation between the semantic modeling as developed in this work, and the possible-worlds approach taken in [Bac90], [Bac91], [GHK92].

Finally, quite generally, it is desirable to further improve our insight into the nature of the probability measures defined by minimizing cross entropy, and their adequacy for modeling default reasoning about probabilities.

References

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

[Bac91] F. Bacchus. Default reasoning from statistics. In *Proc. National Conference on Artificial Intelligence (AAAI-91)*, pages 392–398, 1991.

[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.

[GHK92] A.J. Grove, J.Y. Halpern, and D. Koller. Random worlds and maximum entropy. In *Proc. 7th IEEE Symp. on Logic in Computer Science*, 1992.

[Hal50] Paul R. Halmos. *Measure Theory.* Van Nostrad Reinhold Company, 1950.

[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.

[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.

[KST82] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press, 1982.

[McC80] J. McCarthy. Circumscription - a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[Pea89] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 505–516, 1989.

[PV89] J.B. Paris and A. Vencovská. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1989.

[PV90] J.B. Paris and A. Vencovská. 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.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[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.

[WS92] W.A. Woods and J.G. Schmolze. The kl-one family. *Computers Math. Applic.*, 23(2-5):133–177, 1992.