Default Inferences From Statistical Knowledge

Fahiem Bacchus*
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L–3G1
fbacchus@watdragon.waterloo.edu

July 1st, 1991

Abstract

There are two common and distinct uses of probabilities: probabilities used as degrees of belief and probabilities used as statistical measures. Probabilities used as statistical measures can represent various assertions about the objective statistical state of the world, while probabilities used as degrees of belief can represent various assertions about the subjective state of an agent's beliefs. In this paper we examine how an agent who knows certain statistical facts about the world might infer certain probabilistic degrees of beliefs in other assertions based on these statistics. For example, an agent who knows that most birds fly (a statistical fact) may have a degree of belief greater than 0.5 in the assertion that `Tweety` flies given that `Tweety` is a bird. This inference of degrees of belief from statistical facts is know as direct inference. We develop a formal logical mechanism for performing direct inference, and demonstrate how this mechanism can be applied to the problem of making default inferences as studied in AI.

1 Direct Inference

There are at least two distinct uses of probabilities: probabilities used to represent statistical assertions which we will refer to as statistical probabilities, and probabilities used to represent degrees of belief which we will refer to as propositional probabilities.[1] The key difference between these

*This work was supported by NSERC grant #OGP0041848

[1] Probabilities used to model degrees of belief have often been called called subjective probabilities. However, we avoid this terminology because traditionally proponents of subjective probabilities have denied the kind of *direct* connection between statistical and subjective probabilities that we wish to propose. The name propositional probabilities is used as these are probabilities assigned to particular propositions.

two types of probabilities is that statements of statistical probability represent assertions about the objective statistical state of the world, while statements of propositional probability represent assertions about the subjective state of an agent's beliefs.

For example, the statement `More than 75% of all birds fly` is a statistical assertion about the proportion of fliers among the set of birds; its truth is determined by the objective state of the world. On the other hand the statement `The probability that Tweety flies is greater than 0.75` is an assertion about a degree of belief; its truth is determined by the subjective state of the agent who made the statement. In the world `Tweety` either flies or doesn't fly; there is no probability involved.

The interesting question becomes: What is the relationship between the agent's subjective degrees of belief and objective statistical features of the world? The answer that we will develop in this paper is that the agent's subjective degrees of belief are *determined* by objective statistical knowledge though a mechanism of *direct inference*.[2] The mechanism of direct inference has been studied for a long time in philosophy, e.g., [1, 2, 3, 4, 5], but has received scant attention in AI (a notable is exception is Loui [6]).

Direct inference is the inference of propositional probabilities describing the agent's subjective state from statistical probabilities describing the world's objective state. Simply stated direct inference is based on the claim (posited by various philosophers) that the propositional probability of a particular individual `c` possessing a property, should be equal to the proportion (i.e., statistical probability) of individuals that have that property from among some class of individuals that `c` is known to belong to. For example, if `Tweety` is a bird and more than 75% of birds fly, then direct inference would sanction the conclusion that the agent's degree of belief in the proposition `Tweety flies` should be greater than 0.75.

Direct inference is not only of philosophical interest, but is in fact used in many real situations. For example, most of actuarial science is based on the principles of direct inference. When an insurance companies quotes life insurance rates to particular individuals they compute those rates by equating the probability of a particular individual's death with the proportion of deaths among some set of similar individuals (e.g., similar in sex, age, job hazard, etc.). Another example comes from weather forecasts where the probability of rain on a particular day is obtained by examining the proportion of rainy days among a set of days with similar meteorological features.

In this paper we will outline a formal mechanism for performing direct inference. This mechanism is logically based, and we use a logic which can represent and reason with statistical and propositional probabilities. By reasoning inside of this logic we can reason about the consequences of our mechanism of direct inference. We will then demonstrate how our mechanism of direct inference can be applied to perform default reasoning, and will discuss the advantages of doing

[2] In the formalism we will develop, the agent's degrees of belief will be determined by *his* statistical knowledge, and of course the agent's knowledge is simply another part of the agent's subjective state. We will not, however, explore the issue of how the objective state of the world influences the agent's knowledge: this is part of the problem of knowledge acquisition and perception. Instead, we will assume (like other work in knowledge representation) that we are given a specification of the agent's knowledge base, and will explore the question of what can be further inferred about the agent's subjective state from this knowledge.

default reasoning in this manner.

2 The Probability Logic

Prior to the construction of a mechanism of direct inference we will need a formalism capable of representing and reasoning with both propositional and statistical probabilities. Halpern [7] has recently developed a logic which combines a statistical logic developed by Bacchus [8] and his own propositional probability logic. This combined probability logic suits our needs: it is capable of representing and reasoning with both types of probabilities. Below we present a brief outline of this logic (termed a type III probability logic by Halpern). The interested reader is encouraged to see Halpern [7] for some more details, or for a more extensive exposition with many examples, Bacchus [9]. The reader is also referred to the latter reference for a more detailed exposition, with all of the proofs, of the system of direct inference we present here.

The combined logic is most similar semantically to a two-sorted quantified modal logic, and syntactically to a two-sorted first-order logic. There is a sort of *objects* and a sort of *numbers*. The sort of objects consists of function and predicate symbols suitable for describing the domain of interest, while the numeric sort is used to make numeric assertions. In particular, the numeric sort includes the constants 1, -1, and 0; the functions $+$ and $\times$; and the predicates $=$ and $<$. Additional inequality predicates and numeric constants can easily be added by definition, and we will use them freely. The formulas of the language are generated by applying the standard first-order formula formation rules, and in addition we have two rules that generate new numeric terms (specifically probability terms) from already generated formulas.

1. If α is a formula and $\vec{x}$ is a vector of n distinct object variables, then $[\alpha]_{\vec{x}}$ is a numeric term. In particular, it is a *statistical probability term*

2. If α is a formula, then $\mathsf{prob}(\alpha)$ is a numeric term. In particular, it is a *propositional probability term*.

Since these constructs are terms, they can in turn be used in the generation of additional new formulas.

We extend the language by definition to include conditional probability terms (of both types): $[\alpha|\beta]_{\vec{x}} =_{df} [\alpha \wedge \beta]_{\vec{x}}/[\beta]_{\vec{x}}$, and $\mathsf{prob}(\alpha|\beta) = \mathsf{prob}(\alpha \wedge \beta)/\mathsf{prob}(\beta)$.[3]

Semantically the language is interpreted using structures of the form

$$M = \langle \mathcal{O}, S, \vartheta, \mu_{\mathcal{O}}, \mu_S \rangle^4$$

Where:

[3]For ease of exposition we will ignore the technicalities of dealing with division by zero. See Bacchus [9] for details.

[4]These structures are Halpern's type III probability structures.

1. $\mathcal{O}$ is the domain of objects (i.e., the domain of discourse). S is a set of states or possible worlds. ϑ is a world dependent interpretation of the symbols. The numeric symbols are interpreted as relations and function of the reals $I\!R$. The distinguished numeric functions and predicates, $+$, $\times$, 1, -1, 0, $<$ and $=$, are given their normal interpretation in every state. Hence, these symbols are *rigid*; i.e., their interpretation is the same in every world $s \in S$.

2. $\mu_{\mathcal{O}}$ is a discrete probability measure over $\mathcal{O}$. That is, for every $A \subseteq \mathcal{O}$, $\mu_{\mathcal{O}}(A) = \sum_{o \in A} \mu_{\mathcal{O}}(o)$ and $\sum_{o \in \mathcal{O}} \mu_{\mathcal{O}}(o) = 1$.

3. μ_S is a discrete probability measure over S, such that for $S' \subseteq S$, $\mu_S(S') = \sum_{s \in S'} \mu_S(s)$ and $\sum_{s \in S} \mu_S(s) = 1$.

The formulas of the language are interpreted with respect to this semantic structure in a manner standard for modal languages. In particular, the interpretation of a formula will depend on a structure, M, a current world $s \in S$, and a variable assignment function v. The probability terms have the following interpretation:

1. $([\alpha]_{\vec{x}})^{(M,s,v)} = \mu^n_{\mathcal{O}}\{\vec{a} : (M, s, v[\vec{x}/\vec{a}]) \models \alpha\}$, where $v[\vec{x}/\vec{a}]$ is the variable assignment function identical to v expect that $v(x_i) = a_i$, and $\mu^n_{\mathcal{O}}$ is the n-fold product measure formed from $\mu_{\mathcal{O}}$.

2. $(\mathsf{prob}(\alpha))^{(M,s,v)} = \mu_S\{s' : (M, s', v) \models \alpha\}$.

So we see that $[\alpha]_{\vec{x}}$ denotes the measure of the set of satisfying instantiations of $\vec{x}$ in α (the rational for using the product measure is given in [9]). And $\mathsf{prob}(\alpha)$ denotes the measure of the set of worlds that satisfy α. Another way of interpreting the statistical probability terms is to consider them as representing the probability that a *randomly selected*[5] vector of individuals $\vec{x}$ will satisfy α. Hence, we can call the variables, $\vec{x}$, bound by the statistical terms *random designators*.

A few of examples will help the reader understand the language and its semantics.

1. $[\mathtt{fly(x)}|\mathtt{bird(x)}]_{\mathtt{x}} > 0.75$. A particular triple (M, s, v) satisfies this formula if the probability that a randomly selected bird flies is greater than 0.75. That is, if more that 75% of the birds in the world 's' fly.

2. $\mathsf{prob}(\mathtt{fly(Tweety)}) > \mathsf{prob}(\mathtt{swim(Tweety)})$. This formula is satisfied if the measure of the set of worlds in which `Tweety` flies is greater than the measure of the set of worlds in which she swims.

3. $\mathsf{prob}\Big([\mathtt{grey(x)}|\mathtt{elephant(x)}]_{\mathtt{x}} > .75\Big) > 0.9$. This formula is satisfied if the measure of the set of worlds in which a randomly selected elephant has a greater that 75% probability of

[5]That is, the probability of selecting any particular vector of individuals $\vec{o} \in \mathcal{O}^n$ is $\mu^n_{\mathcal{O}}\{\vec{o}\}$.

being grey, is greater than 0.9.[6]

A powerful proof theory for this language can be constructed which is complete for various special cases [7, 9]. We simply note that all of the examples of reasoning we give here can be performed within this proof theory.

We will need one more feature in our language: the ability to take the expected values of numeric terms. For this purpose we add a "expectation" operator to the language.

1. If t is a *rigid term* or a *statistical probability term*, then $\mathsf{expet}(t)$ is a new numeric term.

2. Semantically this operator is interpreted according to the following definition:

$$(\mathsf{expet}(t))^{(M,s,v)} = \sum_{s' \in S} \mu_S(s') \times t^{(M,s',v)}.$$

   That is, the expected value of a term is the weighted (by μ_S) average of the term's denotation over the set of possible worlds.

2.1 Some Important Properties

The following lemmas describe some important properties of the statistical and expectation terms. Here we use the standard notation $\models \alpha$ to indicate that α is valid, i.e, true for every triple (M, s, v). We also use the notation $\mathsf{cert}(\alpha)$ as an abbreviation for $\mathsf{prob}(\alpha) = 1$.

Lemma 1 (Properties of the Statistical Terms)

1. If no $x_i \in \vec{x}$ which appears in $\alpha \wedge \beta$ is free in λ[7] then $\models \mathsf{cert}\Big([\beta \wedge \lambda]_{\vec{x}} \neq 0 \rightarrow [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}\Big)$.

2. $\models \mathsf{cert}\Big((\forall \vec{x}.\beta \rightarrow \lambda) \rightarrow [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}\Big)$.

3. $\models \mathsf{cert}\Big(\Big[\alpha\Big|[\alpha|\beta]_{\vec{x}} \in (r_1, r_2) \wedge \beta\Big]_{\langle\vec{x},\vec{y}\rangle} \in (r_1, r_2)\Big)$, where $t \in (r_1, r_2)$ is an abbreviation for $r_1 \leq t \leq r_2$.

[6] Halpern's paper [7] contains the erroneous claim that such sentences could not be represented in type III structures. However, it is easy to see that the statistical probability terms are not rigid. Consider the term $[\mathtt{P(x)}]_{\mathtt{x}}$. The set of satisfying instantiations of `x` in `P(x)` in any world corresponds to the denotation of the predicate symbol `P` in that world. Since this denotation varies across worlds, this term is denoting the measure of a different set of individuals in different worlds: its value changes from world to world. Therefore, an assertion like $[\mathtt{grey(x)}|\mathtt{elephant(x)}]_{\mathtt{x}} > .75$ can be true in some worlds and false in others; i.e., we can find a model in which the probability μ_S of the worlds in which it is true is greater than 0.9.

[7] Freedom and bondage is extended in our language to include the clause that the random designators are bound. Also it is not difficult to see that we can rename the random designators without changing the values of the statistical terms.

The last item needs some explanation. Say that our domain consists of a set of boys and girls and that some of those boys have the property that they are popular. That is, they have the statistical property that more that 50% of the girls like them. If we randomly select a girl and a *popular* boy we would expect that there is a greater than 50% probability that the randomly selected girl will like the randomly selected popular boy, since every popular boy is liked by more than 50% of the girls. This last item is a generalization of this reasoning. It says that if we have two randomly selected vectors of individuals $\vec{x}$ and $\vec{y}$ that stand in the relation β, and furthermore the $\vec{y}$ has the property that a randomly selected $\vec{x}$ standing in β with it has probability between r_1 and r_2 of also standing in α with it, then the probability that the $\vec{x}$ and $\vec{y}$ selected stand in the relation α will also be in the bounds (r_1, r_2).

These equalities between the statistical terms are certain; i.e., they are true in every world, and thus they have propositional probability 1.

Because both the statistical terms and rigid terms are bounded above and below in all worlds (the statistical terms are bounded by 0 and 1, the rigid terms are constant), the limit of the infinite sum giving the denotation of the expectation terms exists. Hence, the following properties of the expectation terms can easily be demonstrated by appeal to the properties of limits of real number sums.

Lemma 2 (Properties of the Expectation Terms)

1. If r is rigid, then $\mathsf{expet}(r) = r$.
2. $\mathsf{cert}(t = t') \rightarrow \mathsf{expet}(t) = \mathsf{expet}(t')$.
3. $\mathsf{cert}(t < t') \rightarrow \mathsf{expet}(t) < \mathsf{expet}(t')$. An identical implication holds for all the inequality relations: $>$, $\leq$, and $\geq$.
4. If t is a statistical term, then $\mathsf{expet}(t) = 0 \equiv \mathsf{cert}(t = 0)$.
5. If t is a statistical term then $\mathsf{expet}(t) = 1 \equiv \mathsf{cert}(t = 1)$.
6. If r is rigid, then $\mathsf{expet}(r \times t) = r \times \mathsf{expet}(t)$.

3 The Epistemological Framework

Our combined probability language can be viewed as a generalized logic of belief. To be precise, it is not difficult to show that formulas with propositional probability one satisfy the axioms of *KD45* [10] (a.k.a. weak *S5* $+$ consistency). For example, $\mathsf{cert}(\alpha) = 1 \rightarrow \mathsf{cert}(\mathsf{cert}(\alpha))$ (the axiom of positive introspection) is valid. This means that the formulas that the agent is certain of (i.e., have degree of belief one) behave in the same manner as formulas that are fully believed under ordinary Hintikka style models of belief.

The generalization over Hintikka style models of belief comes along two dimensions. First, through the use of non-unit propositional probabilities the probability logic can represent graded assertions of belief, and by using numeric predicates it can represent various *qualitative* assertions about the strength of beliefs. The second generalization comes from the fact that the logic can represent statistical assertions. Hence, the set of assertions to which the agent can ascribe a degree of belief is richer: the logic can represent an agent's degrees of belief in statistical assertions as well as ordinary first-order assertions.

Let us call the formulas of the probability logic which do not contain any prob or expet operators *objective formulas*. We will assume that there is some finite set of objective formulas that the agent has accepted, i.e., has assigned degree of belief one to. This base of accepted beliefs acts as the agent's knowledge base.[8] The formalism is explicit about the chance of error in the knowledge base. That is, the agent is *certain* of these assertions, but some of them may in fact be false. We will develop a model whereby we can reason *in the combined probability logic* about the degrees of belief the agent should assign to other objective assertions given this base of accepted beliefs.

4 The Reference Class

As we have mentioned, direct inference is performed by assuming that the probability that a particular vector of individuals $\vec{c}$ have a particular property α is equal to the proportion of individuals that have that property from among some class of individuals that $\vec{c}$ is known to belong to. The essential problem in direct inference is choosing the correct class of individuals from which to derive the probability. The class of individuals chosen is called the *reference class*, and the problem of choosing the correct reference class is known as the reference class problem [11]. The approach that we take here is to use the most specific reference class that we know $\vec{c}$ to belong to. This amounts to taking into consideration the entire belief base of the agent (i.e., all of the objective assertions the agent has accepted). There is no philosophical difficulty with this choice; we should, if we can, take into account all of $\vec{c}$'s known properties. The problem that occurs is that the agent may not know any useful statistical information about this possibly very narrow reference class. This problem lead Reichenbach [1] to proclaim that one should choose the smallest class for which one had "reliable statistics."

Since the agent's statistical knowledge is limited it would seem that our choice of the most specific reference class would often be useless. The insight that allows us to avoid this difficulty is that Reichenbach's maxim, and indeed other systems of direct inference like Kyburg's [2], choose the reference class based on what statistics are known. Equivalently this could be viewed as being inferences based on what statistics are *not known*, a kind of inference that has been much studied in AI. This is the manner in which we solve the reference class problem: if we do not have knowledge that a property of $\vec{c}$ has any influence on the probability of α, then we will assume non-monotonically that it does not.[9]

[8] It would be more accurate to refer to these formulas as being a belief base. However, most knowledge bases in AI are in fact belief bases; i.e., there is always some chance of error.

[9] In this work we distinguish between default reasoning, the inference of a plausible conclusion from less than sure knowledge, and non-monotonic reasoning from lack of knowledge (cf. Moore [12]).

5 The Formalism

Let $\mathtt{KB}$ be the *conjunction* of the *complete finite set* of objective formulas accepted by the agent. $\mathtt{KB}$ will usually include information about particular individuals, e.g., bird(Tweety), general logical relationships between properties, e.g., $\forall \mathtt{x}.\mathtt{bird}(\mathtt{x}) \rightarrow \mathtt{animal}(\mathtt{x})$, and *statistical information*, e.g., $[\mathtt{fly}(\mathtt{x})|\mathtt{bird}(\mathtt{x})]_{\mathtt{x}} > .75$.

Definition 1 (Randomization) Let α be an objective formula. If $\langle c_1, \ldots, c_n \rangle$ are *all* the n distinct object constants that appear in $\alpha \wedge \mathtt{KB}$ and $\langle v_1, \ldots, v_n \rangle$ are n distinct object variables that do not occur in $\alpha \wedge \mathtt{KB}$, then let $\mathtt{KB}^v$ denote the new formula which results from textually substituting c_i by v_i in $\mathtt{KB}$, for all i. Similarly for α^v.

Definition 2 (Direct Inference Principle) If α is an *objective* formula, the agent's degree of belief in α should be determined by the equality

$$\mathsf{prob}(\alpha) = \mathsf{expet}\Big([\alpha^v|\mathtt{KB}^v]_{\vec{v}}\Big).^{10}$$

In addition, the agent must fully believe that $[\mathtt{KB}^v]_{\vec{v}} > 0$, i.e., $\mathsf{cert}([\mathtt{KB}^v]_{\vec{v}} > 0)$.[11]

We note a couple of important properties of this principle.

Theorem 1 *The degrees of belief given by the direct inference principle are in fact probabilities.*

Theorem 2 $\mathsf{prob}(\alpha) = 1$ *is a logical consequence of the direct inference principle if and only if* $\models \mathtt{KB} \rightarrow \alpha$.

Definition 3 (The Base Theory) Let D_0 be the set of formulas consisting of $\mathsf{cert}([\mathtt{KB}^v]_{\vec{v}} > 0)$ union all instances of the direct inference principle. Let T_0 denote the closure of D_0 under logical consequence.

Theorem 3 *If* $\mathtt{KB} \wedge [\mathtt{KB}^v]_{\vec{v}} > 0$ *is satisfiable, then so is* T_0.

This last theorem implies that if the agent starts off with a consistent knowledge base, then his base theory T_0 generated by direct inference will also be consistent.

As an example, say that

$$\mathtt{KB} = \mathtt{bird}(\mathtt{Tweety}) \wedge [\mathtt{fly}(\mathtt{x})|\mathtt{bird}(\mathtt{x})]_{\mathtt{x}} > c,$$

[10] The reason that we use the expected value of the statistical term is explained in [9]. One reason we can cite here is that the expected value is rigid while the ordinary statistical term is not, but there are other good reasons for this choice.

[11] There is a reasonable justification for this assumption [9], but here it is sufficient to note that it allows us to avoid division by zero.

where c is a rigid numeric constant with any value greater than 0.5. Then we have that $\mathsf{prob}(\mathtt{fly}(\mathtt{Tweety})) > c \in T_0$. That is, by reasoning from the direct inference principle the agent will draw the intuitive conclusion. We have that

$$\mathsf{prob}(\mathtt{fly}(\mathtt{Tweety})) = \mathsf{expet}\Big([\mathtt{fly}(\mathtt{v})|\mathtt{bird}(\mathtt{v}) \wedge ([\mathtt{fly}(\mathtt{x})|\mathtt{bird}(\mathtt{x})]_{\mathtt{x}} > c)]_{\mathtt{v}}\Big)$$

is an instance of the direct inference principle, and hence it is in T_0. Lemma 1.1 allows us to remove the conjunct $[\mathtt{fly}(\mathtt{x})|\mathtt{bird}(\mathtt{x})]_{\mathtt{x}} > c$ from the conditioning formula;[12] Theorem 2 demonstrates that $\mathsf{cert}([\mathtt{fly}(\mathtt{v})|\mathtt{bird}(\mathtt{v})]_{\mathtt{v}} > c) \in T_0$;[13] the fact that T_0 is closed under logical consequence and properties of the expectation terms demonstrates that $\mathsf{prob}(\mathtt{fly}(\mathtt{Tweety})) > c \in T_0$.

6 Non-monotonic Reasoning about Statistics

Often the base theory T_0 will not contain any useful information about the agent's degree of belief. For example, say that we added the assertion $\mathtt{yellow}(\mathtt{Tweety})$ to the above KB. Now T_0 will contain

$$\mathsf{prob}(\mathtt{fly}(\mathtt{Tweety})) = \mathsf{expet}\Big([\mathtt{fly}(\mathtt{v})|\mathtt{yellow}(\mathtt{v}) \wedge \mathtt{bird}(\mathtt{v})]_{\mathtt{v}}\Big),$$

where we have already used Lemma 1.1 to remove irrelevant conjuncts. Although T_0 contains useful information about the proportion of birds that fly, it contains no information about the proportion of yellows birds that fly: the former statistic does not constrain the latter.[14]

As we noted above we address this issue by allowing the agent to make non-monotonic assumptions to extend his base theory.

Definition 4 (Non-Monotonic Assumptions) If $\mathsf{cert}(\forall\vec{x}.\beta_0 \rightarrow \beta_1) \in T_0$, then $\mathsf{expet}\big([\alpha|\beta_0]_{\vec{x}}\big) = \mathsf{expet}\big([\alpha|\beta_1]_{\vec{x}}\big)$ is a legitimate non- monotonic assumption, where α, β_0 and β_1 are all objective formulas.

This definition allows the agent to inherit (expected) statistics from less specific reference classes for which he may possess better statistical knowledge.

For example, the agent can make the non-monotonic assumption

$$\mathsf{expet}\Big([\mathtt{fly}(\mathtt{v})|\mathtt{yellow}(\mathtt{v}) \wedge \mathtt{bird}(\mathtt{v})]_{\mathtt{v}}\Big) = \mathsf{expet}\Big([\mathtt{fly}(\mathtt{v})|\mathtt{bird}(\mathtt{v})]_{\mathtt{v}}\Big)$$

With this assumption added to T_0 he can now infer $\mathsf{prob}(\mathtt{fly}(\mathtt{Tweety})) > c$.

---

[12] In general, Lemma 1.1 and the properties of the expectation terms allows us to remove those parts of KB that are not related to the constants appearing in α from the conditioning formula. Furthermore, this excision is determined by the semantic properties of the statistical terms, not by a syntactic criterion.

[13] It is easy to see that the random designators can be renamed without affecting the value of the statistical terms.

[14] This is a property shared by other default reasoning formalism (e.g., ϵ-probabilities [13] and conditional logics [14]).

We do not have space to present the full details of the non-monotonic reasoning component of our system, but we note here that its main intent is not to determine uniquely the agent's final theory. Rather its intent is to allow the agent to extend his base theory to generate useful conclusions. We agree with the argument that the agent's final decision as to which theory to use (i.e., which set of non-monotonic assumptions to make) will depend on meta-logical pragmatic considerations (Konolige and Myers [15]). The best a system can do is to eliminate certain counter-intuitive theories, and as our examples will demonstrate, this is what is accomplished.

7 Examples

In the proofs below we use the following abbreviations (1) "expt. prop." to indicate properties of the expectation operator (Lemma 2), (2) "dir. inf." indicating our initial application of the direct inference principle (usually with irrelevant conjuncts removed), and (3) "stat. prop." to indicate properties of the statistical probability terms (Lemma 1).

Example 1 (Opus does not fly.) Let

$$\begin{aligned} \mathtt{KB} = \ & \mathtt{bird(Opus)} \quad \wedge \quad \mathtt{penguin(Opus)} \\ & \wedge \quad [\mathtt{fly(x)}|\mathtt{bird(x)}]_{\mathtt{x}} > c \quad \wedge \quad \forall \mathtt{x}.\mathtt{penguin(x)} \rightarrow \mathtt{bird(x)} \\ & \wedge \quad [\mathtt{fly(x)}|\mathtt{penguin(x)}]_{\mathtt{x}} < 1 - c. \end{aligned}$$

$\mathsf{prob}(\mathtt{fly(Opus)}) = \mathsf{expet}\Big([\mathtt{fly(v)}|\mathtt{bird(v)} \wedge \mathtt{penguin(v)}]_{\mathtt{v}}\Big)$	dir. inf.
$\mathsf{cert}\Big([\mathtt{fly(v)}|\mathtt{bird(v)} \wedge \mathtt{penguin(v)}]_{\mathtt{v}} = [\mathtt{fly(v)}|\mathtt{penguin(v)}]_{\mathtt{v}}\Big)$	Lem. 1.2
$\mathsf{cert}\Big([\mathtt{fly(v)}|\mathtt{penguin(v)}]_{\mathtt{v}} < 1 - c\Big)$	Thm. 2
$\mathsf{expet}\Big([\mathtt{fly(v)}|\mathtt{bird(v)} \wedge \mathtt{penguin(v)}]_{\mathtt{v}}\Big) < 1 - c$	expt. prop.
$\mathsf{prob}(\neg\mathtt{fly(Opus)}) < 1 - c$	

The proof hinges on the fact that because penguins are known to be a subset of birds (i.e., $\mathsf{cert}(\forall \mathtt{x}.\mathtt{penguin(x)} \rightarrow \mathtt{bird(x)}) \in T_0$ by Theorem 2) we have by Lemma 1.2 that

$$\mathsf{cert}\Big([\mathtt{fly(v)}|\mathtt{bird(v)} \wedge \mathtt{penguin(v)}]_{\mathtt{v}} = [\mathtt{fly(v)}|\mathtt{penguin(v)}]_{\mathtt{v}}\Big).$$

This example shows that the agent's base theory T_0 has a natural subset preference as a consequence of the properties of the statistical terms. That is, there is no need for a meta-logical criterion.

Example 2 (Clyde likes Tony but not Fred.) If

$$\begin{aligned} \mathtt{KB} = \ & \mathtt{elephant(Clyde)} \quad \wedge \quad \mathtt{zookeeper(Tony)} \\ & \wedge \quad \mathtt{zookeeper(Fred)} \\ & \wedge \quad [\mathtt{likes(x,y)}|\mathtt{elephant(x)} \wedge \mathtt{zookeeper(y)}]_{\langle \mathtt{x},\mathtt{y}\rangle} > c \\ & \wedge \quad [\mathtt{likes(x,Fred)}|\mathtt{elephant(x)}]_{\mathtt{x}} < 1 - c \end{aligned}$$

It is not difficult to see that $\mathsf{prob}(\mathtt{likes}(\mathtt{Clyde},\mathtt{Tony})) > c \in T_0$. The more interesting case is that $\mathsf{prob}(\mathtt{likes}(\mathtt{Clyde},\mathtt{Fred})) < 1 - c \in T_0$.

$$\begin{array}{ll} \mathsf{prob}(\mathtt{likes}(\mathtt{Clyde},\mathtt{Fred})) = & \\ \qquad \mathsf{expet}\Big(\Big[\ \mathtt{likes}(\mathtt{u},\mathtt{v}) \ \Big| \begin{array}{l} \mathtt{elephant}(\mathtt{u}) \wedge \mathtt{zookeeper}(\mathtt{v}) \\ \wedge \quad [\mathtt{likes}(\mathtt{x},\mathtt{v})|\mathtt{elephant}(\mathtt{x})]_{\mathtt{x}} < 1 - c \end{array} \ \Big]_{\langle \mathtt{u},\mathtt{v}\rangle} \Big) & \text{dir. inf.} \\ \mathsf{cert}\Big(\Big[\ \mathtt{likes}(\mathtt{u},\mathtt{v}) \ \Big| \begin{array}{l} \mathtt{elephant}(\mathtt{u}) \wedge \mathtt{zookeeper}(\mathtt{v}) \\ \wedge \quad [\mathtt{likes}(\mathtt{x},\mathtt{v})|\mathtt{elephant}(\mathtt{x})]_{\mathtt{x}} < 1 - c \end{array} \ \Big]_{\langle \mathtt{u},\mathtt{v}\rangle} \ < 1 - \mathtt{c}\Big) & \text{stat. prop.} \\ \mathsf{prob}(\mathtt{likes}(\mathtt{Clyde},\mathtt{Fred})) > 1 - c & \end{array}$$

This proof hinges on the fact that

$$\Big[\mathtt{likes}(\mathtt{u},\mathtt{v})\Big|\mathtt{elephant}(\mathtt{u}) \wedge \mathtt{zookeeper}(\mathtt{v}) \wedge [\mathtt{likes}(\mathtt{x},\mathtt{v})|\mathtt{elephant}(\mathtt{x})]_{\mathtt{x}} < 1 - c\Big]_{\langle \mathtt{u},\mathtt{v}\rangle} < 1 - c.$$

This equality follows from Lemma 1.3. What happens here is that the statistical language is powerful enough to express not only first-order properties of individuals, but also their statistical properties. In this case `Fred` is differentiated from `Tony` by the fact that he generally unpopular with the elephants. This is a special statistical property that `Fred` possesses. When we condition on all that is known we also condition on this statistical property, and it influences the inferred degree of belief.

In default logic this would correspond to knowing a special default particular to the individual `Fred` and a more general default applicable to the class of zoo-keepers. Intuition would mandate a preference for the more specific default. This preference is another built in feature of our system of direct inference: again it follows directly from the properties of the statistical terms.

Example 3 (University Students) Let

$$\begin{array}{ll} \mathtt{KB} = & [\mathtt{adult}(\mathtt{x})|\mathtt{u\text{-}student}(\mathtt{x})]_{\mathtt{x}} > c \quad \wedge \quad [\mathtt{employed}(\mathtt{x})|\mathtt{adult}(\mathtt{x})]_{\mathtt{x}} > c \\ & \wedge \quad [\mathtt{employed}(\mathtt{x})|\mathtt{u\text{-}student}(\mathtt{x})]_{\mathtt{x}} < 1 - c \quad \wedge \quad \mathtt{u\text{-}student}(\mathtt{John}) \quad \wedge \quad \mathtt{adult}(\mathtt{John}) \end{array}$$

We have that

$$\mathsf{prob}(\mathtt{employed}(\mathtt{John})) = \mathsf{expet}\Big([\mathtt{employed}(\mathtt{v})|\mathtt{u\text{-}student}(\mathtt{v}) \wedge \mathtt{adult}(\mathtt{v})]_{\mathtt{v}}\Big)$$

is in T_0. Clearly we need to make some non-monotonic assumptions to generate any useful conclusions. We could form the theory T_1 by assuming that the expected proportion of employed among the adult university students is the same as among the university students, or we could form the theory T_2 by assuming that this proportion is the same as among the adults. In T_1 we will have that `John` is probably unemployed, while in T_2 he is probably employed. T_2 is an unintuitive theory, and we can show that under some reasonable constraints on c T_2 is in fact contradicted by the base theory T_0. That is, T_0 contains the assertion that the expected proportion of employed among adult university students is *not* the same as the expected proportion of employed among adults, contradicting the assumption that lead to T_2.

All that we need is for c to be greater than $(\sqrt{5}-1)/2 \approx 0.62$, a slightly stronger condition than $c > 0.5$. We have by statistical reasoning from KB that

$$\begin{aligned}[\texttt{employed(v)}|\texttt{u-student(v)} \wedge \texttt{adult(v)}]_{\texttt{v}} &= \frac{[\texttt{employed(v)} \wedge \texttt{adult(v)}|\texttt{u-student(v)}]_{\texttt{v}}}{[\texttt{adult(v)}|\texttt{u-student(v)}]_{\texttt{v}}} \\ &< \frac{[\texttt{employed(v)}|\texttt{u-student(v)}]_{\texttt{v}}}{[\texttt{adult(v)}|\texttt{u-student(v)}]_{\texttt{v}}} < \frac{1-c}{c} < c\end{aligned}$$

The last inequality holds when c is greater than the positive root of the equation $(1-c)/c = c$. This root has value $(\sqrt{5}-1)/2$. Since this inequality is a logical consequence of KB it is certain in T_0. It then follows from the properties of the expectation terms that the non-monotonic assumption

$$\mathsf{expet}\Big([\texttt{employed(v)}|\texttt{u-student(v)} \wedge \texttt{adult(v)}]_{\texttt{v}}\Big) = \mathsf{expet}\Big([\texttt{employed(v)}|\texttt{adult(v)}]_{\texttt{v}}\Big)$$

is contradicted by T_0, as in T_0 the expectation term on the left is less than c while the term on the right is greater than c. This example shows how reasoning with the statistical information can be used to eliminate unreasonable theories.

8 Discussion

We have provided a formal mechanism for performing direct inference, and as our examples demonstrate this mechanism also has applications to default reasoning. Since this mechanism depends on both direct inference and non-monotonic reasoning from lack of knowledge, the reader may question its validity: why not just use non-monotonic reasoning from lack of knowledge, as is done in most other default reasoning formalisms? There are a number of advantages that accrue from being explicit about the statistical knowledge involved in default inferences.

It is clear that there must be an empirical basis for much of default reasoning. For example, the view of defaults as conventions does not explain how an agent is to reason non-monotonically when he interacts with the world: we cannot expect the world to follow our conventions. Our approach encodes much of its information as qualitative statistical assertions. The empirical content of these assertions is intuitive and transparent, and they represent information that can be learned by an agent through his experience. This can be contrasted with the approach of ϵ-probabilities [13], where the empirical content of probabilities infinitesimally removed from one is less than clear. Of course the interpretation of the empirical component of defaults as being statistical is controversial: the best evidence for it lies in the intuitive correctness of the answers it produces.

By adopting a statistical interpretation for assertions that have have been traditionally been encoded as “defaults,” we have gained the ability to perform an extensive amount of reasoning with this information. For example, under this interpretation we can infer the default “Canaries are typically not green” from “Canaries are typically yellow” and “Yellow objects are not green.”

Of all the alternative formalisms for default reasoning the formalism that comes closest to our approach in its ability to reason with the defaults is conditional logic [14]. However, by reasoning with the statistical semantics we can eliminate unintuitive theories and capture the natural subset and specificity preferences. These types of inferences are not possible with conditional logics.

Another important advantage of our dual approach to default reasoning is that it generates *graded* default conclusions. Hence, we have a distinction *in the formalism* between defeasible conclusions and deductive conclusions. This can be important when critical decisions must be made (see Langlotz and Shortliffe [16]). Furthermore, the mechanism is quite capable of reasoning with quantitative statistical information. Hence, the formalism does not promote a schism between commonsense reasoning and "scientific" reasoning. Dependent on the quality of his information the agent can progress from qualitative commonsense reasoning to more exact quantitative evidential reasoning: the same "logic" is used, it is just the amount and quality of information that changes.

The fact that the conclusions are graded also means that the formalism avoids the lottery paradox [17, 18]. For example, we may know that birds typically have properties P_1 through P_n, but also that no bird possesses all of these properties. Other mechanisms of default reasoning are paralyzed by this situation, as if they infer each of the P_i they will also infer the conjunction of the P_i. In our mechanism, however, the agent's degree of belief in each of the P_i will be high, but his degree of belief in conjunctions of the P_i will be lower. This seems perfectly reasonable, as each default conclusion is a source of possible error: the agent should become less certain as he is required to commit to more default conclusions.

References

[1] Hans Reichenbach. *Theory of Probability*. University of California Press, Berkeley, 1949.

[2] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. D. Reidel, Dordrecht, Netherlands, 1974.

[3] John L. Pollock. A theory of direct inference. *Theory and Decision*, 15:29–96, 1983.

[4] John L. Pollock. Foundations for direct inference. *Theory and Decision*, 17:221–256, 1984.

[5] Isaac Levi. *The Enterprise of Knowledge*. MIT-Press, Cambridge, Massachusetts, Cambridge, Massachusetts, 1980.

[6] Ronald P. Loui. *Theory and Computation of Uncertain Inference and Decision*. PhD thesis, The University of Rochester, September 1987.

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

[8] Fahiem Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. PhD thesis, The University of Alberta, 1988.

[9] Fahiem Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. MIT-Press, Cambridge, Massachusetts, 1990.

[10] Joseph Y. Halpern and Yorham Moses. A guide to the modal logics of knowledge and belief. In *Procceedings of the International Joint Conference on Artifical Intelligence (IJCAI)*, pages 480–490, 1985.

[11] Henry E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50(3):374–397, September 1983.

[12] Robert C. Moore. Semantical considerations on nonmonotonic logic. In Matthew L. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*, pages 127–136. Morgan Kaufmann, San Mateo, California, 1987.

[13] Hector Geffner and Judea Pearl. A framework for reasoning with defaults. In Henry E. Kyburg, Jr., Ronald Loui, and Greg Carlson, editors, *Knowledge Representation and Defeasible Inference*. Kluwer Academic Press, London, 1989.

[14] James P. Delgrande. An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, 36:63–90, 1988.

[15] Kurt Konolige and Karen Myers. Representing defaults with epistemic concepts. *Computational Intelligence*, 5(1):32–44, 1989.

[16] Curt Langlotz and Edward H. Shortliffe. Logical and decision-theoretic methods for planning under uncertainty. *AI Magazine*, 10(1):39–48, 1989.

[17] Henry E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, Connecticut, 1961.

[18] David Poole. What the lottery paradox tells us about default reasoning. In Ronald J. Brachman, Hector J. Levesque, and Raymond Reiter, editors, *Proceedings of the First Conference on Principles of Knowledge Representation and Reasoning*, pages 333–340. Morgan Kaufmann, San Mateo, California, 1989.