Lp, A Logic for Representing and Reasoning with Statistical Knowledge

Fahiem Bacchus*
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L–3G1
519–888–4670
fbacchus@logos.uwaterloo.ca

Appears in *Computational Intelligence*, Vol 6, 1990, pp. 209–231

*This research was supported by grants from the University of Alberta and the University of Waterloo.

Abstract

This paper presents a logical formalism for representing and reasoning with statistical knowledge. One of the key features of the formalism is its ability to deal with qualitative statistical information. It is argued that statistical knowledge, especially that of a qualitative nature, is an important component of our world knowledge and that such knowledge is used in many different reasoning tasks. The work is further motivated by the observation that previous formalisms for representing probabilistic information are inadequate for representing statistical knowledge.

The representation mechanism takes the form of a logic that is capable of representing a wide variety of statistical knowledge, and that possesses an intuitive formal semantics based on the simple notions of sets of objects and probabilities defined over those sets. Furthermore, a proof theory is developed and is shown to be sound and complete.

The formalism offers a perspicuous and powerful representational tool for statistical knowledge, and a proof theory which provides a formal specification for a wide class of deductive inferences. The specification provided by the proof theory subsumes most probabilistic inference procedures previously developed in AI. The formalism also subsumes ordinary first-order logic, offering a smooth integration of logical and statistical knowledge.

Keywords Probability Logic, Knowledge Representation, Statistical Knowledge.

1 Introduction

1.1 Why Statistical Knowledge?

One of the primary benefits of using first-order logic to represent knowledge lies in its well defined and intuitive semantics. The semantics of first-order logic is based on the very simple notions of distinguishing individual objects and grouping them into sets. The sets are collections of objects which share some common properties, for example, the set of red objects or the set of cylindrical objects. The notion of sets of individual objects can be easily extended to sets of vectors of objects. Such sets can be used to represent groups of objects which stand in some relation with each other; for example, the set of pairs of objects $\langle x, y \rangle$, where 'x' is the father of 'y', can be used to represent the relation of male parentage.

The intuitive nature of first-order semantics can be explained by our natural ability to recognize various properties objects possess and various relationships which exist between objects. For example, we can generally recognize if an object is red, or if it is cylindrical, or if one object is on top of another object. The ability to classify objects according to their properties seems to be fundamental to our perception of the world around us, and indeed much of our knowledge of the world is knowledge of the properties of various objects and the relationships between various objects.

Another feature of the world around us that we perceive naturally are statistical relationships of relative frequency. The vast amount of statistical knowledge we possess about the world is an indication of our natural ability to accumulate such information. Usually, however, we are not able to precisely quantify these relative frequencies, at least not without the expenditure of extra resources.[1] This knowledge is often qualitative in nature; it is in the form of empirical

[1] For example, we may conduct surveys, or in limited domains we may actually count the numbers of objects with various properties.

generalizations. These generalizations may come from our personal experience, e.g., realizing that among computer scientists males outnumber females, or it may come from information from other sources, e.g., the large amounts of statistical information presented in the popular media.

One of the motivations of this work was this observation that we possess so much statistical knowledge. It is reasonable to infer that this knowledge must be useful to us in dealing with the world, and therefore it is reasonable to assume that an AI system would also find such knowledge useful. In fact, it is not difficult to find areas in AI where such knowledge is already used or can be used. For example, some of the work on uncertainty in AI (see Kanal and Lemmer [1986; 1987] for a sample) is concerned with statistical knowledge in applications that range from expert systems to vision systems. Statistical knowledge is also useful in non-monotonic reasoning [Bacchus, 1990] where defaults can be expressed as empirical generalizations, and in learning, e.g., [Etzioni, 1988]. The aim of this work is to provide a representation formalism that is general enough to support extended uses of statistical knowledge in areas of AI that already use statistics and to support new applications based on statistical knowledge.

1.2 Interpretations of Probability

The axiomatic formulation of probability functions supports a number of different intuitive interpretations. A rough division can be made between these interpretations into those that use probabilities to model notions of proportion or relative frequency and those that use probabilities to model the epistemic concept of degrees of belief. Before we discuss the differences between these two interpretations, let us review the axiomatic specification of probabilities (see, e.g., [Lindley, 1965a; Chung, 1974; Feller, 1968] for more details).

1.2.1 Axiomatic Definition of Probabilities

The modern axiomatic theory of probability is due to Kolmogorov [1950] who based probability theory on measure theory. Even though the intuitive interpretation of probability varies, his axiomatic specification is almost universally accepted.

Under this specification probability functions are real-valued functions defined over a field of subsets (or algebra of subsets). Formally, we have some *sample space*, S, and a field of subsets of S, Π. Π is a field of subsets of S if and only if it satisfies the following conditions:

1. S is a member of Π.

2. Π is closed under complementation; i.e., $A \in \Pi$ implies that $S - A \in \Pi$, where $S - A$ are all those elements of S that are not in A. We will write $\neg A$ for $S - A$.

3. Π is closed under finite unions; i.e., $A \in \Pi$ and $B \in \Pi$ implies that $A \cup B \in \Pi$.

Normally we also require, as does Kolmogorov, that Π be a *sigma-field*, in which case it is closed not only under finite unions, but more generally under *countable* unions. That is, instead of **3**, Π will satisfy the more general condition:

3*. If $A_1, A_2, \ldots$ are a countable collection of sets in Π, i.e., $\forall i. A_i \in \Pi$, then $\bigcup_{i=1}^{\infty} A_i \in \Pi$.

The subsets in Π are called *measurable sets*. Not every subset of S need be in Π; i.e., not every set need be measurable. Note also that the power set of S, i.e., the set of all subsets of S, 2^S, will always be a sigma-field of subsets.

Given a sample space S and a suitable field of subsets Π, probability functions pr (or probability measures) are defined to those functions which map Π to the reals[2] and which satisfy the following axioms (known as the *Kolmogorov* axioms for probability):

P1. $pr(A) \geq 0$ for all $A \in \Pi$.

P2. $pr(S) = 1$.

P3. If $A \cap B = \emptyset$ then $pr(A \cup B) = pr(A) + pr(B)$.

The last property is know as *finite additivity*. If Π is a sigma-field, then we require a more general condition of *countable additivity*:

P3*. $pr(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} pr(A_i)$, if the A_i's are pairwise disjoint.

1.2.2 Statistics vs. Degrees of Belief

Axiomatic specifications generally support a number of different intuitive interpretations. For example, the axioms which specify the algebraic structure of groups supports many different types of groups. The set of permutations of a finite vector of elements forms a group as does the set of rotations of a rigid body in n-space. Similarly, the Kolmogorov axioms support different interpretations of probabilities.

In particular, there are two quite distinct interpretations of probability functions: probabilities acting as statistical measures of proportion or relative frequency, and probabilities acting as measures, or degrees, of belief.

If we have a collection of individuals that are grouped into different classes dependent on their properties, then we can make various assertions about the proportion of individuals in the different classes. For example, of the 10 provinces in Canada, 2 are west of Saskatchewan. Hence, we can say that 20% of the provinces in Canada are west of Saskatchewan. Proportions satisfy the axioms of probability; hence, from the previous assertion we can infer that 80% of the provinces in Canada are not west of Saskatchewan. Similarly, 50% of the provinces are east of Ontario, and since Ontario is east of Saskatchewan, this set is disjoint from the set of provinces west of Saskatchewan; hence, we can infer from the probability axioms that 70% of the provinces are either east of Ontario or west of Saskatchewan. More generally, we can consider the probabilities to be specifying assertions of relative frequency. In this more general case we have a possibly infinite sequence of events and each event has different properties. We can then talk about the relative frequency of various properties in the sequence of events. For example, we could have a sequence of tosses of a coin, and we can make assertions about the relative frequency of tosses that landed heads among this sequence.

The view of probabilities as statistical assertions about proportion or relative frequency is known as the empirical interpretation of probabilities, and it is one of the oldest interpretations. Many

[2] More general probability functions which are not required to be real-valued have been investigated, e.g., [Koopman, 1940; Aleliunas, 1986], and as the reader will see we will use generalized field-valued probability functions in this work.

writers on probability theory have chosen to take this interpretation of probability, e.g., [Venn, 1866; Neyman, 1950; Reichenbach, 1949; Salmon, 1967; von Mises, 1957; Popper, 1959].

The other interpretation of probability is to view probability as being a degree of belief. Under this interpretation probability becomes an epistemic concept, related to an agent's beliefs, instead of an empirical property related to relative frequencies. In this case the sample space S becomes a set of propositions, usually expressed in a logical language; the complementation of a proposition becomes its negation; and the union of two propositions becomes the proposition formed from their disjunction. Hence, probabilities are assigned to propositions. Each proposition is an assertion about the world; thus, intuitively the probabilities of these propositions are viewed as being an agent's degree of belief in the truth of the assertion made by the proposition. For example, the sample space might consist of the two propositions "Tweety is a bird" and "Tweety can fly." Hence, we could have $pr(\text{Tweety is a bird}) = 0.7$ and $pr(\text{Tweety can fly}) = 0.5$ which is to be interpreted as representing that the agent has a degree of belief of 0.7 that Tweety is a bird and a degree of belief of 0.5 that he can fly.

There are many variations of the view of probabilities as degrees of belief. Mainly these variations are distinguished by the manner in degrees of belief are assigned prior to the agent having any information about the state of the world, i.e., the initial probability function, and the manner in which the agent modifies his degrees of belief in the face of new information or evidence about the state of the world. For example, under the view of logical probabilities (Carnap [1962], Hintikka [1966]) there is a logically determined initial probability function, and the agent can change his degrees of belief by conditioning this prior probability with his new information. Alternately, on the Bayesian or subjective view (De Finetti [1964], Savage [1964], Lindley [1965a; 1965b], Jeffrey [1983], Cheeseman [1988]) there is no such thing as a logically determined prior probability, instead the agent has a personal prior probability distribution that can be anything the agent wishes.

My aim here is not to argue about what interpretation of probabilities is better. Rather, I would claim that both interpretations are equally important in AI, and both have distinct uses. What is important, however, is to be clear about the difference between the two interpretations, and to maintain the clarity of the distinction in our schemes for representing probabilities.

The difference between the two types of probabilities can be better appreciated by examining the following pair of assertions "More than 75% of all birds fly" and "The probability that Tweety can fly is greater than 0.75." The first assertion is a statistical assertion about the proportion of fliers among the set of birds. The truth of this assertion is determined by the objective state of the world: in the world this statistical assertion is either true or false. The second assertion is an assertion about the subjective state of some agent. The truth of this assertion is determined by the state of the agent's beliefs: either the agent assigns a degree of belief greater than 0.75 to the assertion "Tweety flies" or he does not.

Note that statistical assertions are more like ordinary logical assertions than like assertions about degrees of belief. For example, the assertion "More than 75% of all birds fly" is much like the assertion "Tweety flies:" both are assertions about the state of the world. Such assertions can be the subject of an agent's beliefs, e.g., an agent may believe or not believe either assertion, or he may assign a degree of belief to either assertion. So an agent may think that it is highly likely that Tweety flies and he may think that it is highly likely that more than 75% of all birds fly. In fact, one can view the study of statistical inference, particularly the Bayesian view of statistical

inference, as being the study of mechanisms for assigning degrees of belief to statistical assertions. For example, the agent starts with degrees of belief in assertions of the form "The percentage of birds that fly is x," for $x \in [0, 100]$, accumulates information from various samples of birds, and generates new degrees of belief in these assertions. For example, he may come to strongly believe that the percentage of flying birds is between 75% and 95%. Lindley [1965a, Chapter 1] gives a useful discussion of this point.

1.3 The Contribution of This Work

Degree of belief probabilities have received a great deal of attention in AI and elsewhere, and various representation formalisms have been developed for this type of probability. These formalisms are useful for representing the uncertainty attached to various assertions. For example, some of the information in a knowledge base may be from questionable sources, so we may want to attach a degree of belief to it. Such degrees of belief can be useful when making decisions about what actions to execute. However, this is not the only type of probability that is important for AI, probabilities used to represent statistical assertions are also of importance. A lot of the knowledge that we wish to represent in AI programs is in the form of *statistical* claims about the world. These range from imprecisely quantified generalizations, like "Most birds fly," or, "People with a runny nose usually have a cold," to more precisely quantified statistical statements, like those found in medical diagnosis systems (e.g., the MYCIN system [Shortliffe and Buchanan, 1975][3]).

Unfortunately, this type of probability has not received much attention, and furthermore the representation formalisms that have been developed for degree of belief probabilities are not well suited for representing statistical assertions.

The contribution of this work is to provide a representation formalism for statistical assertions. This formalism takes the form of a logic that is capable of representing and reasoning with a wide variety of statistical assertions. This is accomplished through an empirical probabilistic component in the semantics. The logic is an extension of ordinary first-order logic, and its development is complete; that is, not only are the syntax and semantics specified, but also a sound and complete *deductive* proof theory is provided.[4] The proof theory is capable of reasoning with the statistical knowledge, as well as with sentences of first-order logic.

Statistical assertions obey the axioms of probability. Therefore, we can reason with them by reasoning with the probability axioms. For example, from the assertion "Most birds fly" along with "All penguins are birds" and "No penguins fly," it is possible to deduce that "Most birds are not Penguins." When more precisely quantified statistics are available, sophisticated Bayesian analysis can be performed. In fact, the proof theory captures the inferences that can be produced from most of the deductive probabilistic reasoning systems (i.e., systems based on the axioms of probability) previously developed.

It should be noted, however, that while the proof theory captures the deductive inferences possible from a collection of statistical information, it does not capture non-deductive inferences.

[3] In expert systems the knowledge is gathered from an expert and has often been called subjective probabilities (degrees of belief) [Duda *et al.*, 1981], but in many cases, especially in the medical domain, we are actually gathering the expert's estimate of underlying statistical data.

[4] The proof theory is complete with respect to the field-valued probabilities that we use in our models. We will discuss the motivation and justifications for using field-valued instead of real-valued probabilities later.

Many of the most interesting uses of statistical information comes from various non-deductive inferences. For example, the study of statistics is primarily concerned with *statistical inference*. This is the *non-deductive* inference of statistical assertions from sample data. Our deductive proof theory does not provide for this kind of statistical inference. Rather, it provides a logic for representing and reasoning with *the results of* statistical inference, i.e., the statistical assertions. Similarly, another use of statistical information is to assign degrees of belief to assertions about particular individuals. For example, if one knows the statistical information "More than 75% of all birds fly" and all one knows about Tweety is that he is a bird, then a reasonable *non-deductive* inference is to believe with degree greater than 0.75 that Tweety can fly. Again the proof theory we provide here is not capable of making these kinds of inference. However, I would argue that the formalism presented here is a necessary first step in developing systems for capturing these kinds of inferences: such formalisms will require mechanisms for representing statistical assertions and for reasoning deductively with them. In fact, a system for infering degrees of belief in assertions about particular individuals has already been developed [Bacchus, 1990], and it depends heavily on the formalism presented here to represent and reason with the statistical assertions that are used in this kind of inference.

There is one more point that should be raised. Just as previous formalisms for degree of belief probabilities are not naturally suited for representing statistical assertions, it turns out that the formalism presented here is not naturally suited for representing degree of belief probabilities. This means that the formalism cannot efficiently deal with uncertainty in one's knowledge. That is, it can represent logical and statistical assertions, but it cannot represent uncertainty about these assertions. Clearly, representing uncertainty in these assertions is important for certain applications. However, since knowledge bases without uncertainty measures have many applications in AI, it is to be expected that knowledge bases containing a richer variety of information (i.e., statistical information as well as logical information) will also be useful.

Subsequent to this work, however, it has been demonstrated how the formalism developed here can be combined with formalisms for degree of belief probabilities to produce a formalism that can deal with both types of probabilities simultaneously [Halpern, 1989]. The important part of Halpern's work is that it provides a formalism that can represent degrees of belief assigned to statistical assertions, and this is important if one wants to study statistical inference. In support of our claim, the work presented here was an necessary first step in this development.

1.4 Outline of the Presentation

The next section provides a discussion of previous work on the problem of representing probabilistic information. Most of this work is aimed at dealing with degree of belief probabilities, and hence these formalisms are not naturally suited for representing statistical knowledge. We then discuss the logic **Lp** which has been developed to solve this problem, pointing out some of the types of statistical knowledge that the formalism can represent along with the logic's key features.

Section 4 starts into the formal results of the paper, first presenting the syntax and semantics of **Lp**, and then giving examples of the types of knowledge representable with this logic.

The deductive proof theory is presented next. The proof theory is shown to be both sound and complete. We also present some examples of the types of reasoning possible with this proof theory. Nilsson [1986] has developed a form of probabilistic reasoning he calls probabilistic entailment,

and Pearl [1986a] has developed a mechanism of probability reasoning using Bayesian networks. Both Nilsson and Pearl developed their mechanisms for degree of belief probabilities. Since degree of belief probabilities and statistical assertions satisfy the same axioms these mechanism have analogues for statistical probabilities. It will be show that the proof theory for**Lp** subsumes the statistical analogues of these mechanisms.

The final section summarizes and discusses the results and contributions of the work presented.

2 Previous Work

2.1 Probabilities over Possible Worlds

There has been an extensive quantity of work on representing and reasoning with degree of belief probabilities, for example, [Carnap, 1962; Gaifman, 1964; Scott and Krauss, 1966; Nilsson, 1986; Bundy, 1985; Field, 1977; van Fraassen, 1981; LeBlanc, 1983; Morgan, 1984; Fagin *et al.*, 1988].

These works have investigated the assignment of probabilities to sentences of a logical language, either first-order or propositional. One way of formalizing the attachment of probabilities to logical sentences is to consider a probability distribution over a set of possible worlds. The probability of any sentence then becomes the probability of the subset of possible worlds in which it is true. Alternately, one can place a probability distribution over the Lindenbaum-Tarski algebra generated by the sentences of the logic. This algebra is composed of equivalent classes of sentences that are defined by the relation of provable equivalence,[5] (see Bell and Machover [1977]). The probability of a sentence then becomes the probability of the equivalence class it is a member of. These two approaches for attaching probabilities to sentences are essentially equivalent (Bacchus [1988a; 1990]).

The sentences of such languages represent assertions about the world and probabilities assigned to them act as degree of belief in these assertions. Instead of either asserting a sentence or its negation, as in ordinary logic, one can attach some intermediate degree to it, a degree of belief. So, for example, one could represent a degree of belief of greater than 0.9 in the assertion "Tweety can fly" by assigning the sentence $Fly(Tweety)$ a probability greater than 0.9. These probabilities have the properties that one would expect; e.g., the probability of the negation of a sentence is 1 minus the probability of the sentence. Furthermore, when the probabilities are all 0 and 1 the probability logic reduces to ordinary logic.

Despite these advantages, however, it is not easy to represent statistical information, e.g., the assertion "More than 90% of all birds fly."[6]

We can first note that propositional languages do not possess sufficient power to represent these kinds of statements. This particular statistical statement is an assertion which indicates a

[5] That is, if $\alpha \equiv \beta$ is valid, then α and β are in the same equivalence class.

[6] It is well known that first-order logic is in some sense universally expressive. In particular, we can represent this assertion in ordinary first-order logic by formalizing set theory, and then building up sufficient mathematics inside the language to represent statements of this form (as does Kyburg [1974]). This can be done in first-order logic; so it can certainly be done in first-order logic generalized to have probabilities attached to the sentences. However, this is not an efficient representation, nor will there be any direct reflection in the semantics of the statistical information (i.e., the assertion will be buried in a complex construction of sets). We are concerned here with efficient representations and intuitive semantics. Furthermore, if one chooses the route of using set theory one gives up any hope of finding a reasonable proof theory.

relationship between the properties of being a bird and being able to fly, but it is not an assertion about any particular bird. This indicates that some sort of variable is required. Propositional languages do not have variables, and so are inadequate for this task even when they are generalized to take on probabilities instead of just truth values.

When we move to first-order languages we do get access to variables, variables which can range over the domain of individuals. A seemingly reasonable way to represent this statement is to consider the probabilistic generalization of the universal sentence $\forall x.Bird(x) \Rightarrow Fly(x)$. The universal in first-order logic says that all birds fly, so perhaps if we attach a probability of > 0.9 to it we will get what we need. Unfortunately, this is not the case. If there is *single* bird who is thought to be unable to fly, the universal will be forced to have a probability close to zero. That is, the probability of the universal must be $1 - pr[\exists x.Bird(x) \wedge \neg Fly(x)]$, where we use the notation '$pr[\alpha]$' to denote the probability of a sentence 'α'. Hence, if one believes to degree greater that 0.1 that a non-flying bird exists, the probability of the universal must be < 0.9.

This difficulty can also be understood in terms of probabilities over possible worlds. It seems quite reasonable to believe that the statement "More than 90% of all birds fly" is true in most of the worlds one believes possible, while at the same time believing that $\exists x.Bird(x) \wedge \neg Fly(x)$ is also true in most of these worlds.

The simplest representation using universal quantification, then, fails to do the job. Another possibility, however, is to use conditional probabilities. We have probabilities attached to sentences hence with two sentences we can form conditional probabilities. It has been suggested (Cheeseman [1988]) that meta-level statements of the following form can be used to capture statistical statements, in particular for the statement about birds:

$$\forall x.pr[Fly(x)|Bird(x)] > 0.9.$$

The reason that this is a meta-level quantification is that the universal quantifier is quantifying over a formula $pr[Fly(x)|Bird(x)]$ which is not a formula of a first-order language since it contains a 'pr' operator. However, it is not difficult to formalize a language in which such formulas are well-formed; i.e., we can reduce the meta-level quantification to object-level quantification in a richer language (see [Bacchus, 1990]).

This statement is intended to assert that for every term t in the first-order language the conditional probability $pr[Fly(t)|Bird(t)]$, with the variable x substituted by the term t, is > 0.9.

This representation cannot be considered to be a representation of the statistical assertion. Rather, it is most reasonably interpreted as a prescription that determines how statistical information can be used to specify how an agent should modify his degrees of belief. Consider the following collection of assertions.

1. 80% of the provinces in Canada are east of Alberta.

2. British Columbia is a province of Canada.

3. British Columbia is not east of Alberta.

I am certain that these assertions are simultaneously true. If we could represent the statistical assertion as claimed by Cheeseman, my current beliefs would be described by the following representation.

1. $\forall x.pr[East\_Alberta(x)|province(x)] = 0.8.$
2. $pr[province(B.C.)] = 1.$
3. $pr[\neg East\_Alberta(B.C.)] = 1.$

But it is not difficult to see that this representation is inconsistent. In particular, since $pr[\neg\alpha] = 1 - pr[\alpha]$, we have that $pr[East\_Alberta(B.C.)] = 0$. However, the first item says that

$$\begin{aligned} \texttt{0.8} &= \texttt{pr}\Big[\texttt{East\_Alberta(B.C.)}|\texttt{province(B.C.)}\Big] \\ &= \frac{\texttt{pr}\Big[\texttt{East\_Alberta(B.C.)} \wedge \texttt{province(B.C.)}\Big]}{\texttt{pr}\Big[\texttt{province(B.C.)}\Big]} \\ &= \frac{\texttt{pr}\Big[\texttt{East\_Alberta(B.C.)}\Big]}{\texttt{1}} \\ &= \texttt{pr}\Big[\texttt{East\_Alberta(B.C.)}\Big]. \end{aligned}$$

Hence, using this proposed representation of the statistical information we generate an inconsistency. That is, under this representation I cannot be in an state of belief where I hold all three beliefs *simultaneously*! There is no single probability function over my beliefs that will satisfy all three of these constraints. But clearly this is unreasonable as I really do believe, as do millions of other Canadians, that the three assertions are true simultaneously. Cheeseman has argued that in these examples we need to distinguish between prior probabilities and posterior probabilities. It is true that I would not be faced with a contradiction if I held the statisical belief in a prior state and the beliefs about B.C. in a posterior state, but this argument misses the point; I hold all three of these beliefs simultaneously in my current state of beliefs. Note also that the argument does not depend on me having a degree of belief of 1 in the last two assertions, any reasonably high degree of belief will also yield the contradiction.

One of the important uses of statistical information is its influence on an agent's degrees of beliefs, and although Cheeseman's proposal fails to provide a representation of the statistical information, it does provide a specification of how an agent's degrees of beliefs can be determined. I will argue later, however, that this specification is flawed, but first we can examine the manner in which it works.

Consider the formula

$$\forall \texttt{x.pr}[\texttt{East\_Alberta(x)}|\texttt{province(x)}] = \texttt{0.8}.$$

this formula can be used determine an agent's beliefs about the properties of a particular provience. If the agent has an initial state of beliefs in which his degrees of belief are described *solely* by quantified conditional probabilities of this form, then he can condition this initial state with facts about particular individuals and generate reasonable inferences in his posterior state of beliefs. For example, only the above formula is satisfied by my initial state and I then condition on the new evidence `province(B.C.)`, then my posterior state of beliefs will assign a degree of belief of 0.8 to the assertion `East_Alberta(B.C.)`. On the other hand if I condition my initial state with the new evidence

$\texttt{province(B.C.)} \wedge \neg\texttt{East\_Alberta(B.C.)}$ I will have a degree of belief of 0 in $\texttt{East\_Alberta(B.C.)}$ in my posterior state of beliefs, as the conditional

$$\texttt{pr}[\texttt{East\_Alberta(B.C.)}|\texttt{province(B.C.)}]$$

does not determine the conditional

$$\texttt{pr}[\texttt{East\_Alberta(B.C.)}|\texttt{province(B.C.)} \wedge \neg\texttt{East\_Alberta(B.C.)}].$$

In general, given such an initial state of beliefs, if you know nothing about a province other than it is a province you will have a degree of belief of 0.8 that it is east of Alberta; i.e., you will have a degree of belief that is determined by the proportion of provinces that are east of Alberta.

Geffner and Pearl [1988] use a representation like this in their system of default reasoning. In their system the agent's prior beliefs are described by a background context which contains universally quantified conditional probabilities. Assertions about particular individuals *are not allowed* in the background context. As we have shown such assertions can easily contradict the universal assertions that are present. When reasoning about particular individuals the agent moves to a posterior set of beliefs by conditioning on information specific to these individuals. For example, to reason about the bird *Tweety* the agent will condition his background context with the formula *bird(Tweety)*. As our example demonstrates, if the agent's background context contains the formula $\texttt{pr}[\texttt{fly(x)}|\texttt{bird(x)}] > \texttt{0.9}$ the agent's posterior beliefs will assign a high degree of belief to *fly(Tweety)*. If, on the other hand, the agent knows that *Tweety* is a penguin, then when he conditions his background context, his degree of belief in *fly(Tweety)* will be different.

Although this representation supports a relationship between statistical informaton and an agent's degrees of belief, thus supporting a probabilistic approach to default reasoning, it has a number of flaws:

1. To avoid inconsistency the approach requires an unnatural division in the agent's beliefs. I have lots of statistical beliefs which I hold simultaneously with beliefs about particular individuals, and it seems to be quite unnatural to suppose that I actually have two disparate sets of beliefs: a background context containing only statistical assertions and a evidence corpus where I have all my information about particular individuals. However, such a split is exactly what this representation requires.

2. After I have conditioned the background context with my evidence my posterior state of beliefs no longer satisfies the universally quantified conditionals in the background context. Hence, as we learn new facts we always have to return to the initial prior and condition that with the sum total of our evidence. For example, if we first learn that `Oscar` is a penguin and we condition on this information, then we will not be able to incrementally condition on the new information that `Tweety` is a bird. We will have to return to the background context and condition on $\texttt{penguin(Oscar)} \wedge \texttt{bird(Tweety)}$. Such an approach is certainly at variance with the normal Bayesian epistemology. Under that view today's posterior becomes tomorrow's prior. That is, one never returns to a fixed prior, rather one's state of beliefs continues to evolve by conditioning on new evidence.

3. The "representation" of the statistical information in the background context cannot be used to support other tasks. For example, we cannot use this representation to support the learning of statistical information. Under known approaches to learning statistics (statistical inference), statistical information is generated from samples which contain information about particular individuals. One cannot start from a state of beliefs which contains beliefs about provinces that are not east of Alberta and move to a state of beliefs where the formula $\forall \mathtt{x}.\mathtt{pr}[\mathtt{East\_Alberta(x)}|\mathtt{province(x)}] = \mathtt{0.8}$ is true. That is, one cannot attain the state described by the background context from information about particular individuals.

   Nor can this representation support degrees of belief in statistical assertions without a resort to complexities like second-order probabilities. Hence, we cannot represent and reason with assertions like "I think that it is very likely that more than 75% of all birds fly."

4. Such a complex representation for capturing the effect of statistical information on an agent's degrees of belief is unnecessary. In [Bacchus, 1990] a system for accomplishing this kind of inference is developed. The system makes use of the formalism presented here and as a result is able to provide a much more natural representation for the knowledge that underwrites inferences of this type. In particular, there is no split in the agent's knowledge into background context and evidence; rather, the agent's statistical knowledge co-exists with his knowledge about particular individuals in a single knowlege base represented using the formalism presented here—no inconsistency is generated. Furthermore, the formalism supports an extensive amount of reasoning with the statistical assertions, something that is not easy with the more complex represention using the universally quantified conditional probabilities.

The problematic nature of these attempts to represent statistical information give some evidence to our claim that degree of belief probabilities are not naturally suited for representing statistical assertions. The most convincing argument, however, will come in when we present our own formalism for representing such information. The formalism, a logic called **Lp**, offers a natural representation for statistical assertions through a semantic structure that is markedly different from the assignment of probabilities to possible worlds that is used for degree of belief probabilities.

Lp does not have a probability distribution over a set of possible worlds; instead, it has a probability distribution over the domain of discourse. In the logic statistical assertions are expressed through probability terms which contain open formulas (i.e., formulas with free variables). For example, the statement "More than 50% of all dogs bark" can be expressed with the **Lp** sentence $[Bark(x)|Dog(x)]_x > 0.5$. This sentence is formed from the '$>$' predicate symbol, the constant '0.5', and a probability term which contains two open formulas, $Bark(x)$ and $Dog(x)$. The square brackets are used to form probability terms. These terms are formed by binding some of the free variables of the open formula. In this case the free variable 'x' is bound by the probability term.

Intuitively, the probability term represents the probability that a randomly selected dog, x, will be able to bark. Equivalently, it can be viewed as representing the proportion of objects, x, that bark among those that are dogs. These probability terms have a completely different semantics from the semantics of universally quantified sentences, and can be used to express a wide variety of statistical knowledge.

In **Lp**, however, closed formulas can only have probability one or zero. That is, in **Lp** a closed formula like $Bark(Fido)$ is either true or false; no intermediate value is possible. This result,

which will be proved later, indicates there are definite formal differences in the logics suitable for representing statistical assertions and the logics suitable for representing degrees of belief, a difference that reflects the fundamental difference in the meaning of these two types of probabilities.

2.2 Probabilities over the Domain of Discourse

As mentioned above, the logic **Lp** uses a probability distribution over the domain of discourse. There has been some previous work using this idea.

Most closely related to **Lp** is a probability logic developed by Keisler [1985]. His work lays some useful foundations for probability logics where the probability distribution is defined over the domain of discourse. The aim of his work is, however, to develop a logic for expressing mathematical notions where uncountable domains of discourse are common. Keisler has shown that when the domain is uncountable there are problems in developing a logic that possesses both probabilities over the domain of discourse and universal quantification.[7] Uncountable domains are not, however, of paramount importance in AI, where we are primarily concerned with statements about the 'ordinary objects' of human experience. We have restricted the class of admissible models to be at most countably infinite in cardinality which has enabled the creation of a probabilistic logic that retains universal quantification. In fact, **Lp**, unlike Keisler's logic, is an *extension of ordinary first-order logic*; thus it can represent all statements expressible in first-order logic, as well as probabilistic knowledge.

Another key difference between **Lp** and Keisler's logic is that Keisler uses a device he calls probability quantifiers (P-quantifiers). This device is similar to the so called J-operators which are standard in many-valued logics (see Rosser and Turquette [1952]). The intent of this device is to give access in the object language (syntax) to the probabilities that exist in the semantics. For example, one can write the sentence $(Px = 0.5)\theta$ to indicate that 50% of the objects in the domain satisfy the formula θ (i.e., θ is true when the variable x in θ is interpreted as that object). These P-quantifiers become part of the fixed logical symbols of the language. With these P-quantifiers, however, the numerical values of the probabilities remain outside the main part of the logic. That is, the numbers (like 0.5) which appear inside the P-quantifiers cannot be referred to outside of the P-quantifiers. Nor can we use variables instead of numbers inside of the P-quantifiers, since we would not be able to quantify over these variables.

There are two consequences of this restriction. First, arithmetic relationships between probabilities cannot be expressed. For example, it is not possible to express the statement: "More politicians are lawyer than engineers" using P-quantifiers. This statement cannot be expressed without a commitment to the values of the probabilities of both cases. That is, we could say something like $(Px = .8)(Lawyer(x)|Politician(x))$ and $(Px = .4)(Engineer(x)|Politician(x))$ but not something like

$$(Px)(Lawyer(x)|Politician(x)) > (Px)(Engineer(x)|Politician(x)),$$

as this is not a valid form of the P-quantifier. Hence, we have no way of representing qualitative notions of probability, which, we will argue later, are crucial for AI applications.

[7] This limitation arises from the fact that the projection sets generated through universal quantification may not, in general, be in the domain of any probability function; i.e., they might not be measurable sets.

The second consequence is that since the probabilities are hidden inside the P-quantifiers Keisler's logic cannot be used to reason with the probabilities themselves. Again, this severely limits its usefulness for AI applications.

There has also been some work in the philosophy of language which has used probabilities over the domain of discourse. Åqvist et al. [1980] give a semantic analysis of adverbs of frequency (e.g., always, sometimes, often). Their semantic model is essentially a first-order logic with a probability distribution over the domain of discourse. They, however, restrict themselves to finite models and a less expressive logic. They have concentrated on representing adverbs of frequency, and their logic is otherwise quite limited in its expressive power. Also they have not addressed any proof theoretic issues; so, like Keisler, their logic cannot be used to reason with the probabilities.

3 Description of the Formalism

3.1 Types of Statistical Knowledge

One of the criticisms of the use of probabilities in AI was stated in an influential article by McCarthy and Hayes [1969], in which they observed:

> *The information necessary to assign numerical probabilities is not ordinarily available. Therefore, a formalism that required numerical probabilities would be epistemologically inadequate.*

This has been an on-going and valid criticism of the use of probabilities for general knowledge representation.

Perhaps this is also the reason why the area in which probabilities have had their major impact has been in specialized expert systems. In such domains numbers (of some degree of accuracy) are often available, and are obtained by interviewing domain experts. The development of methods for structuring probabilities into causal networks (Pearl [1986a]) has further increased the popularity of using probabilities in expert systems, especially for medical diagnosis where probabilities seem to have some definite advantages (for example, see the arguments presented by Horvitz et al. [1986], Heckerman et al. [1987] and Schachter et al. [1987]).

However, the impact of probabilities in general knowledge representation tasks remains limited. These approaches still require a significant amount of numerical data, which makes them unsuitable for general knowledge. Furthermore, all of this work has been based on propositional languages, and such languages are inadequate for general representation tasks.

This work meets the objection of McCarthy and Hayes by developing a logic which is capable of expressing a wide range of non-numerical probabilistic knowledge; furthermore, if numbers are available they too can be represented.[8] It is useful to note some of the qualitative varieties of statistical information:

Relative: Statistical knowledge may be strictly comparative. For example, while most would agree that more politicians were trained as lawyers than as engineers, few would be able to assign values to these probabilities.

[8] Clearly we do not want a representation of probabilities that is incapable of representing the numerical information that we might have.

Interval: Even when we can give numeric values to the statistics in question these values may only be in the form of intervals; e.g., the proportion of politicians who are lawyers may be in the range 0.6–0.9. This form of statistical information is particularly important since it is exactly the form produced by statistical inference.

Functional: We might also know functional determiners of statistics. For example, it would seem that the weight of a bird is a factor in its ability to fly. It is clear that given the weight of a bird we cannot deduce its ability to fly, nor its inability to fly (except perhaps in extreme cases). This knowledge could, however, be expressed as "A greater proportion of birds will be fliers among a set of lighter birds than among a set of heavier birds." This kind or functional information is common in medical domains.

Conditional: In a general knowledge base we may have totally unrelated sets of statistical knowledge, e.g., "Most dogs bark" along with "Most old cars need repairs." Conditional probabilities can be used to represent the fact that these pieces of information are only applicable to certain types of individuals.[9] For example, if we construe 'Most' as meaning more than 50%, these statements could be represented as the **Lp** sentences "$[Bark(x)|Dog(x)]_x > 0.5$" and "$[Needs\_Repairs(x)|Old\_Car(x)]_x > 0.5$," where the square brackets are used to indicate probability and x can be considered to be a random member of the predicates' denotations. In the first sentence, for example, no assertion is being made about the probability of a randomly selected object, x, barking unless x is known to be a dog.

Independence: Knowledge of independence is also a type of statistical knowledge that people possess. For example, most doctors would agree that the color of their patients' shoes has no influence on their illness. Work by Pearl and his associates has demonstrated the importance of this kind of knowledge ([Pearl, 1986b; Pearl and Paz, 1986; Pearl and Verma, 1987]).

The logic developed in this work can represent and reason with all of these different types of statistical information.

Along with all of this statistical knowledge it is also clear that a lot of the knowledge that we need to use is purely logical or taxonomic (for example, see the arguments put forward by Schubert in [1988]). **Lp** is an extension of first-order logic, so logical knowledge of this nature can also be expressed. Furthermore, the proof theory smoothly integrates logical reasoning with probabilistic reasoning.

3.2 The Field of Numbers

A key innovation that enabled us to solve the problem of epistemological adequacy, was to make the logic two-sorted, by including a totally ordered field of numbers in the semantic model. One sort of entity in the logic is a set of objects, $\mathcal{O}$, and the other sort is a field of numbers, $\mathcal{F}$. The

[9] Hempel [1962, page 136] makes a cogent argument that *all* probabilities are in fact conditional probabilities. Indeed, in the logic constructed unconditional probabilities make very little sense, except perhaps, in very circumscribed domains.

intention is that the set of objects consists of the things of interest (e.g., cars, people, kinds of cars), while the field of numbers consists of ordinary real numbers.[10]

With numbers as part of the object language the numeric values of the probability terms become accessible at the syntactic and proof theoretic level, accomplishing what Keisler accomplished with his P-quantifiers, but doing it more flexibly. With the numeric values of the probability terms as part of the object language, it becomes possible to express relationships between them without specifying the actual numbers. For example, it is possible to express the previous statement "More politicians are lawyers than engineers" in **Lp** as

$$[Lawyer(x)|Politician(x)]_x > [Engineer(x)|Politician(x)]_x.$$

No commitment is made to a specific value for either of the probability terms mentioned, i.e., no value is asserted for either $[Lawyer(x)|Politician(x)]_x$ or $[Engineer(x)|Politician(x)]_x$. Furthermore, from the distinguished symbols '1', '0' and '-1', that are also part of the object language, it becomes possible to build up by definition terms which denote any rational number that we may wish to refer to. Using these terms it becomes possible to express intervals. For example, if the terms '0.7' and '0.9' are added to the language by definition, statements like "The proportion of politicians that are lawyers is between 70% and 90%" can be represented by the **Lp** sentence:

$$[Lawyer(x)|Politician(x)]_x \in (0.7, 0.9).$$

Of course one need not make direct use of axioms to reason about such rational number terms. One can compute expressions containing only rational number terms using standard arithmetic hardware. Any computations performed in this manner will be sound inferences which could have been duplicated by the proof theory.

Once a field of numbers was added to the logic it also becomes possible to include 'measuring' functions in the logic. These measuring functions map the set of objects to the field of numbers. Using the measuring functions it is possible to express a statement like "Jack's weight is 80 kilograms." This can be expressed with the **Lp** sentence $Weight\_in\_Kgs(Jack) = 80$, where *Weight_in_Kgs* is defined to be a measuring function and *Jack* and *80* are constants (object and field constants respectively). The ability to express functional probabilistic knowledge was gained by allowing sentences to be constructed recursively from these types of symbols. For example, the statement "Heavier birds are less likely to be able to fly" can be expressed in **Lp** using a measuring function symbol.[11]

[10] Using real numbers was the intention; however, a reasonable proof theory can only be developed for totally ordered fields (see Section 3.3).

[11] These measuring functions are very similar to Hayes's [1985] quality spaces. In particular, we could have defined "*Weight*" to be a function from individuals to an abstract numeric quantity of weight and further defined "*kilograms*," "*pounds*," etc. as field functions from the abstract numeric quantities to numbers which give the weight in particular units. The only difference is that we would be assuming that the abstract quantities are elements of a totally ordered field. Hayes, on the other hand, simply assumes that the abstract quantities are closed under addition. Furthermore, we have chosen to avoid the name "measure" function used by Hayes to avoid confusion with the mathematical concept of a measure function, of which probability functions are instances.

3.3 Finitely-additive, Field-valued Probabilities

Mathematically standard probabilities are *real-valued* measures that are typically required to be *sigma-additive* (Section 1.2.1).

The probabilities used in this work differ in both respects. First, their range of values is only required to be a totally ordered field instead of being restricted to a particular totally ordered field, the reals. Second, they are only required to be finitely-additive instead of being restricted to be sigma-additive. We have chosen to weaken these two requirements as we desire an expressive logic that still retains some reasonable proof theoretic properties, in particular, that possesses a complete proof theory.

Complete proof theories possess two very important properties. First, they allow semantic proofs of valid deductions. Completeness implies that every semantic entailment will be mirrored by a deductive proof; hence, with completeness one need only demonstrate semantic entailment and one will have the guarantee that the syntactic proof also exists. It may sometimes be useful to give the detailed syntactic proof, but often this is very tedious. Completeness allows one to explore the theoretical reasoning power of the logic by examining the semantics instead of by manipulating symbols.

The second advantage of a complete proof theory is it gives a formal specification for the set of valid inferences which can be made within a logic. This formal specification can be used to guide the construction of automated reasoners (which may or may not have any surface resemblance to the formal proof theory). It can also be used to analyze the power of tractable subsets of the logic, i.e., to answer the important question "What inferences are these tractable subsets giving up to attain their tractability?"

Recently Abadi and Halpern [1988] have demonstrated that the set of valid formulas of a first-order probability logic with real-valued sigma-additive probabilities over the domain of discourse is not recursively enumerable, except when the domain is bounded in size. In particular, this means that one cannot give a complete proof theory for a logic like **Lp** under the requirement that the probability functions be sigma-additive and real-valued: complete proof theories *do not exist* for this case.

There are two aspects to this problem: the requirement for sigma-additivity and the requirement for real values.[12] A trivial way to deal with the problem of sigma-additivity is to restrict the domain of discourse to be finite. For finite domains, finite additivity trivially corresponds to sigma additivity. This is essentially the route taken by Fagin et al. [1988], they restrict their attention to propositional languages where it is impossible to refer to an infinite set (since sentences, being finite in length, can only refer to a finite collection of atomic symbols). First-order languages, however, allow one to refer to an infinite collection of objects in a sentence of finite length through quantification.

One can still trivialize the problem of sigma-additivity in first-order logics by restricting the domain of discourse to be finite, and perhaps one can argue that finite domains are sufficient for AI. Unfortunately one cannot give a complete proof theory even for finite domains, as demonstrated by

[12] It can also be pointed out that these two requirements on probability functions have been questioned before. Koopman [1940] and more recently Aleliunas [1988] have investigated non-real-valued probability functions, and Savage [1964] used finitely-additive probability functions in his development of the Bayesian approach to statistical inference, instead of sigma-additive function.

Abadi and Halpern. To have a complete proof theory one must make the stronger restriction that the domain is bounded in size by some finite number 'N'. This is much more difficult to justify even for AI domains. The domain may be finite, but where do we find a fixed upper bound, N, on its size: one can always imagine one more individual in the domain. Furthermore, there are many domains which we may want an AI program to reason about that are not finite, e.g., commonsense reasoning about arithmetic [Simmons, 1986].

By just requiring the probabilities to be finitely-additive, not sigma-additive, we are able to attain a logic with a complete proof theory while still allowing infinite domains (we still restrict ourselves to countably infinite domains, however). One implication is that the probability functions that we use will not have all of the properties that standard probability functions have. However, all of the properties that our more relaxed probability functions possess will also be properties of sigma-additive probability functions: standard sigma-additive probability functions *are* finitely-additive, but not vice-versa. The extra properties possessed by sigma-additive probability functions show up when considering limit properties of the probability functions; for normal finite computations our finitely-additive probabilities will behave just like sigma-additive ones.

The other relaxation in the probability functions we use is that fact that they are field-valued not necessarily real-valued. Fagin et al. [1988] and also Halpern [1989] are able to axiomatize real-valued probabilities by using the theory of real-closed fields (see Shoenfeld [1967]). Tarski [1951] has shown that this theory is complete for the reals. That is, any sentence in the theory of real-closed fields is provable if and only if it is true of the reals. However, this completeness comes at the price of expressiveness. It is well know that there is no first-order axiomatization that characterize the reals up to isomorphism (Barwise [1977]). What the theory of real-closed fields does is to carve out a characterizable *piece* of the theory of the reals. In particular, it restricts the language so that only a limited set of the sentences can be written in the theory. It is for this limited set of sentences that the completeness result cited above holds.

The result is that the theory of real-closed fields is not very expressive. Its main weakness is that it allows no functions other that addition and multiplication. Also it does not allow constants other than those that can be built up from 0, 1, and -1 (i.e., built up under arithmetic as solutions to polynomials). In particular, it does not allow "indefinite" numeric constants, i.e., numeric constants whose denotation is defined by something other than a particular number; e.g., one could not represent a constant whose denotation is the number of members of AAAI, if one did not know *exactly* what this number was.

This means that the 'measuring' functions used here, which play a key role in extending the expressiveness of **Lp**, would not be allowed. Also it would not be possible to make statements that assert that probabilities are functions of other values. So, for example, one could not assert that certain quantities are normally distributed.

Again, the desire for expressiveness in conjunction with a reasonable proof theory mandated considering a more relaxed version of probability functions. By allowing our probabilities to be field-valued instead of just real-valued we are able to include the expressive features, like measuring functions, without giving up a reasonable proof theory.[13]

[13] It should be noted that Fagin et al. [1988] and Halpern [1989] are concerned with various applications in theoretical computer science. For these applications it is important to have standard probabilities, and thus they are justifiably concerned with forcing the probabilities to be real-valued and sigma-additive, even at the cost of expressiveness.

In terms of behavior field-valued probabilities present less of a problem than do non-sigma-additive probabilities. The reals are a totally ordered field; thus anything true of field-valued probabilities will also be true of real-valued ones. It is also known that every totally ordered field contains the rationals as a subfield, which means that the probabilities in **Lp** can be assigned any rational values (between 0 and 1) that one wishes. This means that field-valued probabilities will be sufficient for computational purposes.

As with sigma-additive probabilities, real-valued probabilities possess certain limit properties that go beyond the properties of the field-valued probabilities used here. However, every property of the field-valued probabilities will also be a property of real-valued probabilities: real-valued probabilities are field-valued, but not vice-versa.

It will be shown later that many interesting statements true of standard probabilities are also provable of the probabilities used here. In fact, existent work in AI has only used simple properties of standard probabilities, properties that are also provable of the probabilities used here.

It is important to comment on what a completeness result for finitely-additive, field-valued probabilities means. All sentences valid for these probabilities will be deducible, and *all of these sentences will also be valid for standard probabilities.* That is, the proof theory will be sound with respect to standard probabilities. However, there will be some sentences valid for standard probabilities which will not be deducible. That is, the proof theory will not be complete for *standard* probabilities (indeed, as Abadi and Halpern have shown, no complete proof theory exists). However, we will still have the advantages of a complete proof theory cited above. That is, we will be able to examine the power of the proof theory through the semantics, as long as we ensure that we do not depend on sigma-additivity or real values, and only use those properties of probabilities that are true of our finitely-additive and field-valued probabilities. Furthermore, we can still use the proof theory as a formal specification for a wide set of valid inferences. The semantics already tells us exactly what properties of standard probabilities this formal specification captures, i.e., their finitely-additive and field-valued properties. In other words we start off with a specification which captures a well defined subset of the valid inferences[14], and use this specification to guide the construction of automated reasoners which, most probably, will only capture subsets of this subset.

4 The Logic Lp

This section presents the syntax and semantics of the logic **Lp**. The formalization of **Lp** follows the standard steps used in the development of ordinary first-order logic (see for example [Bell and Machover, 1977]). First, the set of allowed symbols is defined. Then rules are given which specify the strings of symbols that are the well-formed formulas. This defines the syntax of **Lp**. Next, the semantics of **Lp** are given, by first defining the set of admissible models, the **Lp**-structures, then a correspondence between truth in the models and the well formed formulas. In the next section a deductive proof theory is presented which provides a correspondence between truth in the model and a syntactic manipulation of the formulas. The deductive proof theory is shown to be both sound and complete with respect to the **Lp**-structures.

[14] One could argue that this subset is the most useful for AI, since we are primarily concerned with objects of ordinary experience, which must be finitary.

The letters 'n' and 'm' are used as meta-variables rangeing over the natural numbers.

4.1 Symbols

We start with a collection of symbols chosen by the user to be suitable for describing the domain of interest. This collection includes a set of constant symbols $(a, b, c, \ldots)$; a set of function symbols $(f, g, h, \ldots)$; a set of predicate symbols $(P, Q, R, \ldots)$; and a set of measuring function symbols, e.g., (*Weight*, *Size*, μ, ν, ρ, $\ldots$).

These symbols can be of two sorts, object symbols and field symbols. We will write the field symbols in a **bold** font when there is a danger of confusion.

Along with the user defined symbols we have a collection of distinguished symbols that are always part of the language regardless of the user's choice of vocabulary. The distinguished symbols include a set of variables $(x, y, z, \ldots)$, which also come in two sorts; the binary object predicate symbol $=$; the field constant symbols 1, 0 and -1; the field binary predicate symbols $\geq$ and $=$[15]; the field binary function symbols $+$, $\times$, and $-$; the logical connectives $\wedge$ and $\neg$; the quantifier $\forall$; and the probability term formers [, and].

4.2 Formulas

From these symbols we generate the formulas of **Lp** in a manner standard for a sorted first-order logic:

T0) A single object variable or constant is an *o-term*; a single field variable or constant is an *f-term.*

T1) If f is an n-ary object function symbol and $\vec{n}t$ are o-terms, then $ft_1 \ldots t_n$ is an *o-term.* If $\mathbf{f}$ is an n-ary field function symbol and $\vec{n}\mathbf{t}$ are f-terms then, $\mathbf{ft}_1 \ldots \mathbf{t}_n$ is an *f-term.* If ν is an n-ary measuring function symbol and $\vec{n}t$ are o-terms, then $\nu t_1 \ldots t_n$ is an *f-term.*

F1) If P is an n-ary object predicate symbol and $\vec{n}t$ are o-terms, then $Pt_1 \ldots t_n$ is a *formula.* If $\mathbf{P}$ is an n-ary field predicate symbol and $\vec{n}\mathbf{t}$ are f-terms, then $\mathbf{Pt}_1 \ldots \mathbf{t}_n$ is a *formula.*

F2) If α is a formula then so is $\neg\alpha$.

F3) If α and β are formulas then so is $\alpha \wedge \beta$.

F4) If α is a formula and x is a variable (of either type), then $\forall x.\alpha$ is a *formula.*

T2) If α is a formula and $\vec{x}$ is a vector of n object variables $\langle \vec{n}x \rangle$, then $[\alpha]_{\vec{x}}$ is an *f-term.* In particular, terms of this type are probability terms. We will call the variables $\vec{x}$ used in the probability terms *random designators.*

This definition of formulas is different from the standard first-order definition; the last rule of formation allows terms to be constructed from formulas.

The connectives $\vee$, $\Rightarrow$ and $\equiv$, and the quantifier $\exists$ are defined in the standard manner from the given primitives. The predicate symbols $=$ and $\geq$ as well as the function symbols $+$, $\times$, and

[15] Note, '$=$' is used both as a field and as an object equality symbol. However, this should not cause any confusion.

– are written in the more readable infix form. Furthermore, standard conventions of scope and precedence are used to limit the use of parentheses. It is also convenient to introduce the following abbreviations to express inequalities between field terms.

Definition 4.1 a) $x \leq y =_{df} y \geq x$ b) $x \in (y, z) =_{df} y \leq x \wedge x \leq z$
c) $x < y =_{df} \neg(x \geq y)$ d) $x > y =_{df} \neg(y \geq x)$

We also need to extend our language to include conditional probability terms.

Definition 4.2 [Axiom of Conditional Probabilities]

$$[\beta]_{\vec{x}} \neq 0 \Rightarrow [\alpha|\beta]_{\vec{x}} \times [\beta]_{\vec{x}} = [\alpha \wedge \beta]_{\vec{x}} \quad \wedge \quad [\beta]_{\vec{x}} = 0 \Rightarrow [\alpha|\beta]_{\vec{x}} = 0.$$

The important feature of this definition is that conditional probabilities are defined to be equal to zero if the conditioning formula has zero probability. Normally such conditional probabilities remain undefined, but in a logical language there is no way of duplicating this. In the syntax we have no access to the denotation of the terms; hence, we have no way of determining in the syntax if a probability term is equal to zero. If we wish to generate conditional probability terms in the syntax we must make provision for those terms formed by conditioning on formulas with zero probability. We have chosen here to make these terms equal to zero.

With this definition we can always rewrite a formula containing a conditional probability term into an equivalent formula containing only standard probability terms. With this observation it can be demonstrated that any formula of the extended language, containing conditional probability terms, will be provable from a proof theory augmented by the above definitional axiom if and only if its equivalent formula in the unextended language is provable from the unaugmented proof theory.[16] These results allow us to use conditional probability terms in our language and at the same time not worry about them in the formal development of the language.

We will also freely extend our language to include constants denoting any non-transcendental real number, i.e., any number which is the root of a rational coefficient polynomial. We can capture the behavior of such constants by adding defining axioms. For example, if we wished to add the new constant '0.5' and have it behave in the proper manner (e.g., we want $0.5 + 0.5 = 1$ to be valid in the extended language) we could add the new constant along with the axiom $0.5 \times (1 + 1) = 1$. Similarly, we could add the constant $\sqrt{2}$ by adding the axiom $\sqrt{2} \times \sqrt{2} = (1 + 1)$. As with the conditional probabilities it can be shown that such extensions of the language do not change its proof-theoretic properties [Shoenfield, 1967]. That is, we can always rewrite the formulas containing the new constants to equivalent formulas containing just the initial constants -1, 1 and 0.

4.3 Semantic Model

Definition 4.3 [The Model] An **Lp**-Structure is defined to be the tuple

$$\mathcal{M} = \left\langle \mathcal{O}, \mathcal{F}, \{ \Pi_n, \mu_n \mid n = 1, 2 \ldots \} \right\rangle$$

[16] See Shoenfield [1967, p. 59] for a description of how definitional extensions like this one can be used to rewrite formulas in the extended language to equivalent formulas of the unextended language, and a proof that such extensions do not change the proof-theoretic properties of the language.

Where:

a) $\mathcal{O}$ represents a *countable* set of individual objects $\mathcal{O}$.

b) $\mathcal{F}$ represents a totally ordered field of numbers $\mathcal{F}$. Like every totally ordered field, $\mathcal{F}$ contains two distinguished elements that are the units of addition and multiplication. In the field of real numbers these units are called zero and one, and the same names will be used to refer to the units of $\mathcal{F}$.

c) Each Π_n is a field of subsets of $\mathcal{O}^n$. This field contains all singleton sets of $\mathcal{O}^n$, i.e., every singleton n-tuple. It also contains all subsets of $\mathcal{O}^n$ defined by the formulas of **Lp** (later the semantic definition of the formulas will give a more precise characterization of these subsets). Π_n is the domain of μ_n.

e) $\{\mu_n \mid n = 1, 2, \ldots\}$ is a sequence of probability functions. The domain of each μ_n is Π_n, and it range is $\mathcal{F}$.

The sequence of probability functions is subject to some further constraints which ensure that the probability terms behave coherently. The implications of these constraints are discussed in the next section.

1. The sequence of probability functions is a sequence of product measures. That is, for any two sets $A \in \mathcal{O}^n$ and $B \in \mathcal{O}^m$, and their Cartesian product $A \times B \in \mathcal{O}^{n+m}$, if $A \in \Pi_n$ and $B \in \Pi_m$, then

$$A \times B \in \Pi_{n+m} \quad \text{and} \quad \mu_{n+m}(A \times B) = \mu_n(A) \times \mu_m(B).^{17}$$

For models where the domain of discourse is *finite* the product measure constraint is sufficient. However, there are many natural notions which involve countably infinite sequences of events or individuals. For example, infinite sequences of trials are often referred to in the study of statistics. Usually, to deal with infinite domains the probability functions are required to be sigma-additive.

As noted previously, making this restriction presents a difficulty when developing a proof theory. In particular, it is not possible to produce a proof theory that is complete for sigma-additive probabilities. To avoid this problem a weaker constraint is placed on the probability functions. When $\mathcal{O}^n$ is finite this weaker constraint is already satisfied—it can be deduced directly from the condition of finiteness and the facts that (a) every singleton set in $\mathcal{O}^n$ is μ_n measurable, and (b) the μ_n are product measures. When $\mathcal{O}^n$ is countably infinite the constraint is not guaranteed, but it is weaker than the condition of sigma-additivity: the constraint is a consequence of sigma-additivity and the use of product measures, but not vice-versa.

The weaker constraint has the advantage (over sigma-additivity) of being specifiable as an axiom in the logic, and it ensures that the probability terms in **Lp** have a necessary coherence property. The additional constraint is quite simple:

[17] The reader familiar with measure theory may wonder why we did not just define μ_n to be $\mu_1 \times \ldots \times \mu_1$ (n times), as does Halpern [1989]. The reason is that this simpler definition will not work for probabilities which are not sigma-additive.

2. Each μ_n is invariant under permutations. That is, for every permutation π of $\{1, \ldots, n\}$ and $S \in \Pi_n$ if

$$\pi S = \{(a_{\pi(1)}, \ldots, a_{\pi(n)}) : (\vec{n}a) \in S\},$$

then

$$\pi S \in \Pi_n \quad \text{and} \quad \mu_n(\pi S) = \mu_n(S).$$

4.4 The Effect of The Product Measure Constraint

Constraining the sequence of probability functions to be a sequence of product measures insures that distinct variables bound by the probability term formers behave in an independent manner. This is similar to the independence of distinct universally quantified variables in first-order logic, e.g., the sentence $\forall \mathtt{x} \forall \mathtt{y}.\mathtt{P}(\mathtt{x}) \wedge \mathtt{Q}(\mathtt{y})$ can be decomposed into two independent sentences, i.e., $\forall \mathtt{x}.\mathtt{P}(\mathtt{x})$ and $\forall \mathtt{y}.\mathtt{Q}(\mathtt{y})$. Since y and x are distinct variables bound by separate quantifiers, their meanings are independent of each other.

With this independence we have, for example, that the probability terms are unaffected by tautologies, e.g., $[\mathtt{P}(\mathtt{x}) \wedge (\mathtt{R}(\mathtt{y}) \vee \neg \mathtt{R}(\mathtt{y}))]_{\langle \mathtt{x},\mathtt{y} \rangle} = [\mathtt{P}(\mathtt{x})]_{\langle \mathtt{x} \rangle}$. This is a result of the fact that the random designator 'y' is independent of the designator 'x'.

It should be noted that this constraint on the probability functions does not make any implicit assumptions of independence of the form commonly found in probabilistic inference engines (e.g., the independence assumptions of the Prospector system [Duda *et al.*, 1981], see Johnson [1986]). This constraint affects the values of probability terms with distinct variables, also, complex probability terms, e.g., $[[\alpha]_x = \mathbf{z}]_y$. (This can be seen from axiom (P7), presented in the next section, which expresses the constraint.) The constraint does not, however, make any presumptions concerning the independence of formulas containing the same set of probability variables. That is, in general, $[\alpha \wedge \beta]_{\vec{x}} \neq [\alpha]_{\vec{x}} \times [\beta]_{\vec{x}}$.

In fact, the probabilistic knowledge that we wish to express in **Lp** normally makes some claim of correlation between the properties possessed by the same object (or tuple of objects), e.g., the correlation between the properties of being a bird and being able to fly. In this example, the correlation can be expressed by the probability term $[Fly(x)|Bird(x)]_x$ where the same variable appears in both formulas. This probability term expresses the proportion of flying birds among birds. This can be contrasted with the probability term $[Fly(y)|Bird(x)]_{\langle x,y \rangle}$. In this term the variables are distinct, and its semantic meaning is that we have chosen pairs of objects and are expressing the proportion of the pairs in which the first object is a bird while the second object can fly to the pairs in which the first object is a bird irrespective of the properties of the second object. Since we are referring to different objects, there is no reason for there to be any correlation between their properties.

Correlations between the properties of a particular *tuple* of objects can be expressed through the use of n-place predicates. For example, the probability term

$$[(Boy(\mathtt{x}) \wedge Girl(y)) \vee (Girl(\mathtt{x}) \wedge Boy(y))|Loves(\mathtt{x}, y)]_{\langle x,y \rangle}$$

is not, in general, equal to the product any simpler probability terms.

The property ensured by the second constraint is that the probability terms are invariant under permutation of the variables. That is, the order of the variables cited in the probability terms

makes no difference, e.g., $[\alpha]_{\langle x,y\rangle} = [\alpha]_{\langle y,x\rangle}$. Universal quantification also displays this property, e.g., $\forall x \forall y.\alpha \equiv \forall y \forall x.\alpha$.

A further coherence property of the probability terms, which is not due to either of the constraints on the probability measures, is that they are invariant under variable name changes, e.g., $[\mathtt{P(x)}]_{\mathtt{x}} = [\mathtt{P(y)}]_{\mathtt{y}}$. This behavior comes from the manner in which the semantics of the formulas is defined.

4.5 Semantics of Formulas

Meaning is given to the formulas of **Lp** by defining a correspondence between the formulas and an **Lp**-Structure, $\mathcal{M}$, augmented by the truth values $\top$ and $\bot$ (true and false). Such a correspondence is called an interpretation. An interpretation assigns to every object constant symbol an element of $\mathcal{O}$, to every n-ary object function symbol a function from $\mathcal{O}^n$ to $\mathcal{O}$, and to every n-ary object predicate symbol a subset of $\mathcal{O}^n$. It maps the distinguished predicate symbol '=' to the equality relation, i.e., the set $\{\langle x, y\rangle : \langle x, y\rangle \in \mathcal{O}^2 \text{ and } x = y\}$. Similarly, it maps the field constant, function and predicate symbols to elements of $\mathcal{F}$, functions from $\mathcal{F}^n$ to $\mathcal{F}$, and subsets of $\mathcal{F}^n$ respectively, mapping the distinguished symbols 1, 0, -1, $+$, $\times$, $-$, $\geq$, and $=$, to the expected constants, operations, and relations. It maps each n-ary measuring function symbol to a function from $\mathcal{O}^n$ to $\mathcal{F}$. Finally, it assigns to each object variable x an element of $\mathcal{O}$ and to each field variable $\mathbf{x}$ an element of $\mathcal{F}$.

These assignments serve as the inductive basis for an interpretation of the formulas. Two interpretations σ and τ are said to **agree** on a given symbol θ if $\theta^\sigma = \theta^\tau$, where θ^σ denotes the interpretation of θ under σ. Also, σ and τ are said to have the same **underlying structure** if they agree on all constant, predicate, and function symbols (of all types). Let $\sigma(x/o)$ denote a new interpretation which is identical to σ except that it assigns the individual o to the variable x (types must match). More generally, let $\sigma(\vec{x}/\vec{a})$, where $\vec{a} = \langle \vec{n}a\rangle$ and $\vec{x} = \langle \vec{n}x\rangle$ are vectors of individuals and variables (of matching type), denote a new interpretation identical to σ except that $(x_i)^{\sigma(\vec{x}/\vec{a})} = a_{i\ (i=1,\ldots,n)}$. An interpretation σ is extended to a truth value interpretation of the formulas of **Lp** in the following recursive manner:

T0) If x is a variable or constant (of either type) then x^σ is already defined.

T1) If f is an n-ary function symbol (of either type) and $\vec{n}t$ are terms of the same type, or if f is an n-ary measuring function symbol and $\vec{n}t$ are o-terms, then

$$(ft_1 \ldots t_n)^\sigma = f^\sigma(t_1^\sigma \ldots t_n^\sigma).$$

F1) If P is an n-ary predicate symbol (of either type) and $\vec{n}t$ are terms of the same type then

$$(Pt_1 \ldots t_n)^\sigma = \begin{cases} \top & \text{if } \langle t_1^\sigma, \ldots, t_n^\sigma\rangle \in P^\sigma, \\ \bot & \text{otherwise.} \end{cases}$$

F2) For every formula α,

$$(\neg\alpha)^\sigma = \begin{cases} \top & \text{if } \alpha^\sigma = \bot, \\ \bot & \text{otherwise.} \end{cases}$$

F3) For every pair of formulas α and β,

$$(\alpha \wedge \beta)^\sigma = \begin{cases} \top & \text{if } \alpha^\sigma = \top \text{ and } \beta^\sigma = \top, \\ \bot & \text{otherwise.} \end{cases}$$

F4a) For every formula α and object variable x,

$$(\forall x.\alpha)^\sigma = \begin{cases} \top & \text{if } \alpha^{\sigma(x/a)} = \top \text{ for every } a \in \mathcal{O}, \\ \bot & \text{otherwise.} \end{cases}$$

F4b) For every formula α and field variable $\mathbf{x}$,

$$(\forall \mathbf{x}.\alpha)^\sigma = \begin{cases} \top & \text{if } \alpha^{\sigma(\mathbf{x}/u)} = \top \text{ for every } u \in \mathcal{F}, \\ \bot & \text{otherwise.} \end{cases}$$

T2) For every formula α the f-term created by the probability term former, $[\alpha]_{\vec{x}}$, is given the interpretation,

$$\left([\alpha]_{\vec{x}}\right)^\sigma = \mu_n\{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Since μ_n is a probability function which maps to the field of numbers $\mathcal{F}$, it is clear that $[\alpha]_{\vec{x}}$ denotes an element of $\mathcal{F}$ under any interpretation σ; thus, it is a valid f-term. As mentioned before, Π_n, the domain of μ_n, is a field of subsets of $\mathcal{O}^n$ which includes those subsets defined by the formulas of **Lp**. Hence, the above set is in Π_n.

5 Examples of Knowledge Representable in Lp

Now we present some examples of what can be represented in **Lp**.

Example 1 [Notions of typicality] "Most birds can fly:"

$$[\mathit{fly}(x) | \mathit{bird}(x)]_x > 0.5,$$

where '> 0.5' is the least presumptive reading of 'Most'.

Example 2 [Functional probabilistic relations] "Heavier birds are less likely to be able to fly:"

$$\forall r.\Big([\mathit{weight}(x) > r | \mathit{bird}(x)]_x > 0 \wedge [\mathit{weight}(x) < r | \mathit{bird}(x)]_x > 0 \Rightarrow$$
$$[\mathit{fly}(x) | \mathit{bird}(x) \wedge \mathit{weight}(x) < r]_x > [\mathit{fly}(x) | \mathit{bird}(x) \wedge \mathit{weight}(x) > r]_x\Big)$$

That is, for every number $\mathbf{r}$ as long are there are some birds with weight greater than $\mathbf{r}$ and some birds with weight less than $\mathbf{r}$, the proportion of flying birds among the set of birds lighter than $\mathbf{r}$ is greater than the proportion of flying birds among the set of birds heavier than $\mathbf{r}$.

Example 3 [Mixing universal quantification and probabilities] "The probability of finding a given type of animal at a zoo is given by a function, **f**, of the expense of acquiring and maintaining that type of animal:"

$$\forall x.\Big(animal\_type(x) \Rightarrow [at(x,y)|zoo(y)]_y = \mathbf{f}(expense(x))\Big),$$

where *expense* is a measuring function symbol and **f** is a field function symbol. Here *expense* might be a function that can be calculated through other information in the knowledge base, e.g.,

$$\forall x.(expense(x) = weight(x) \times 100 + initial\_cost(x)).$$

Also, **f** could be declared to be non-decreasing:

$$\forall r_1 r_2.(r_1 > r_2 \Rightarrow \mathbf{f}(r_1) > \mathbf{f}(r_2)).$$

Example 4 [Knowledge of independence] The canonical tri-functional expression of independence (see [Pearl, 1986b]) "The properties P and Q are independent given R" can be expressed in **Lp**.

$$[P(x) \wedge Q(x)|R(x)]_x = [P(x)|R(x)]_x \times [Q(x)|R(x)]_x.$$

In most systems of probabilistic reasoning knowledge of property independence can only be captured at a meta-level.

Example 5 [Notions from Statistics]

1. "A sequence of ten tosses of a fair coin will land heads with a frequency between 45–55% with greater than 95% probability:"

   $$[frequency\_heads(x) \in (0.45, 0.55)|sequence\_of\_tosses(x)]_x > 0.95.$$

   Here the domain contains a set of objects, *sequence_of_tosses(x)*, each member of which represents a sequence of ten coin tosses of a fair coin, and a measuring function, *frequency_heads*, that maps each sequence of tosses to a number in the closed interval [0,1], a number which represents the relative frequency of heads in that sequence.

2. "The height of adult males (humans) is normally distributed with mean 177cm and standard deviation 13cm:"

   $$\forall \mathbf{xy}.\Big([height(z) \in (\mathbf{x},\mathbf{y})|Adult\_male(z)]_z = \mathbf{normal}(\mathbf{x},\mathbf{y},177,13)\Big)$$

   Here **normal** is a field function which, given an interval $(\mathbf{x},\mathbf{y})$, a mean, and a standard deviation, returns an approximation of the integral over the given interval of a normal distribution with the specified mean and standard deviation over the given interval. (The result is an approximation since we only have access to rational number approximations of the real values).

6 Deductive Proof Theory

This section provides a deductive proof theory for **Lp**. The proof theory consists of a set of axioms and rules of inference, and is shown to be both sound and complete. The proof theory for **Lp** is very similar to the proof theory for ordinary first-order logic,[18] the major change being in the set of axioms. Two new sets of axioms must be introduced, one set to deal with the logic of the probability function, and another set to define the logic of the field $\mathcal{F}$. There are, however, a few technical difficulties arising from the probability function.

One technicality arises from the fact that when probability terms are formed by rule **T2** (Section 4.2) all of the variables $x_i \in \vec{x}$, which appear in the formula α, are bound by the probability term former. That is, their semantic interpretation is altered, as specified by the rule of interpretation **T2** (Section 4.5). This creates a difficulty with those formulas which also contain other quantifiers, a difficulty that is similar to the difficulty arising from nested quantifiers in ordinary first-order logic.

One of the rules of inference in first-order logic allows terms to be substituted for the variable bound by the universal quantifier. For example, in first-order logic it is valid to infer the sentence $\mathit{Man}(\mathit{Socrates}) \Rightarrow \mathit{Mortal}(\mathit{Socrates})$ given the sentence $\forall x.(\mathit{Man}(x) \Rightarrow \mathit{Mortal}(x))$. Here the term *Socrates* has been substituted for the bound variable x. When first-order quantifiers are nested, care must be used to avoid invalid conclusions. For example, in the formula $\forall x.P(x) \Rightarrow \exists x.Q(x)$ a term t substituted for the first (universally) quantified x cannot be substituted for the second x; the second x is in the scope of a distinct quantifier. Such a substitution would lead to the erroneous conclusion $P(t) \Rightarrow Q(t)$. In general, if a term $\mathtt{t}$ is to be substituted for the universally quantified variable $\mathtt{x}$ in $\forall \mathtt{x}.\alpha$, $\mathtt{t}$ can only be substituted for the *free occurrences* of $\mathtt{x}$ in α.

Another technicality arises from the fact that the term t may itself contain variables (especially in **Lp**, where the probability terms can contain arbitrary formulas). When such a term is substituted into a formula its variables may be accidently captured by other quantifiers in the formula. For example, in the formula $\forall x \exists y.P(x) \wedge Q(y)$ if the term $f(y)$ is substituted for the variable x the formula $\exists y.P(f(y)) \wedge Q(y)$ results, where the y in $f(y)$ has been captured by the existential quantifier. This formula cannot be validly inferred from the previous formula.

Since the probability terms bind variables, these two difficulties arise in the interaction of the probability terms with the ordinary quantifiers $\forall$ and $\exists$. These difficulties are dealt with, as in first-order logic, by definitions which specify when a given variable is free in a given formula. Substitution of terms for variables is then defined in such a way that only free variables are affected. The problem of accidental capture is overcome by developing rules for renaming quantified variables. These rules transform formulas to new formulas which are identical in their semantic meaning and in which there is no possibility of accidently capturing any of the variables in the term to be substituted in.

The final technicality is that the probability function generates terms from formulas. Most of the theorems of first-order logic are proved by induction on the formulas of the logic. With **Lp** these theorems must be proved by simultaneous induction on both the formulas and the terms.

The development of the proof theory consists of two parts. First we discuss how the notions of substitution can be extended to suit the requirements of **Lp**. After this, we present the axioms and

[18] A reference for all discussions of first-order logic in this section is the textbook *A Course in Mathematical Logic* by Bell and Machover [1977].

rules of inference which make up the deductive proof theory of **Lp**. This proof theory is shown to be sound and complete.

6.1 Substitution

In this section α and β are used to refer *either* to terms or formulas of **Lp**. We can extend the standard first-order definitions which relate to substitution to handle the variables bound by the probability terms, and we can prove various theorems to demonstrate that the behavior of substitution in **Lp** is a natural extension of its behavior in first-order logic. Here we simply discuss these matters as a mostly intuitive level. A rigorous development of substitution in **Lp** is contained in [Bacchus, 1988b], which also contains the proofs of the theorems cited.

As usual we start with the notion of free and bound occurrences of variables in formulas. In **Lp**, just as in first-order logic, a particular occurrence is free if and only if it is not bound. However, not only are the universally quantified variables bound, but also those variables which appear in the probability term.

Definition 6.1 If $\alpha = [\beta]_{\vec{x}}$, and $x \in \vec{x}$ (i.e., $x = x_i$ for some i), then every occurrence of x in α is **bound** in α. Otherwise, a given occurrence of x in α is **free** in α iff that occurrence is free in β.

We say x is **free in** α if x has at least one free occurrence in α. The **free variables** of α are all those variables that are free in α. The next theorem shows that it is only the free variables of a formula or term which can alter its meaning, once we have fixed on a specific **Lp**-Structure.

Theorem 6.2 *Let σ and τ be interpretations with the same underlying structure $\mathcal{M}$ and which agree on every free variable of α. Then*

$$\alpha^{\sigma} = \alpha^{\tau}.$$

A formula α which has no free variables is called a **sentence** or a **closed** formula. This theorem implies that the truth value of a sentence, α^{σ}, depends only on the underlying structure $\mathcal{M}$. This allows a definition of structure (model) satisfaction.

Definition 6.3 An **Lp**-structure $\mathcal{M}$ **satisfies** a sentence α, written $\mathcal{M} \models \alpha$, if $\alpha^{\sigma} = \top$ for all interpretations, σ, whose underlying structure is $\mathcal{M}$. More generally, an interpretation σ **satisfies** a formula α (set of formulas Φ) if $\alpha^{\sigma} = \top$ ($\beta^{\sigma} = \top$ for every $\beta \in \Phi$), written $\sigma \models \alpha$ ($\sigma \models \Phi$). Finally, a set of formulas Φ **entails** a formula α (written $\Phi \models \alpha$) if every interpretation which satisfies Φ also satisfies α.

To deal with substitution when an accidental capture might occur it is necessary to rename some of the bound variables in the formula. In first-order logic this is accomplished by defining rules which generate **variants**. Variants are new formulas that contain different variable names, but preserve the semantics of the original formula. First-order logic gives rules which allow one to recursively change the names of universally quantified variables. By renaming the variables in this manner we avoid the problem of accidental capture. In **Lp** we can extend the notion of variants to the probability variables. For example, if we wanted to substitute the term '*f(y)*' for the variable '*x*' in the formula $\forall x.P(x) \Rightarrow [R(x,y)]_y > 0.5$ we would get $\forall x.P(x) \Rightarrow [R(f(y),y)]_y > 0.5$, in which

the y has been accidentally captured by the binding of the probability term. It is necessary to produce a variant of the original formula in which the probability term has a renamed variable, e.g., the variant $\forall x.P(x) \Rightarrow [R(x,z)]_z > 0.5$. Substitution with the variant formula can then proceed without problem. For our purposes we simply note that among the set of variants of a formula are the following variants defined by variant probability terms:

Definition 6.4 [Variant probability terms] If $[\beta]_{\vec{x}}$ is a probability term then $[\beta(\vec{x}/\vec{y})]_{\vec{y}}$, is a **variant probability term**, where $\vec{y}$ is a vector of object variables of the same size as the vector $\vec{x}$, and we have that either $y_i = x_i$ (i.e., no variable name change for the i-th probability variable) or y_i is a new variable which does not occur in β. Furthermore, $\beta(\vec{x}/\vec{y})$ indicates the new formula which results from substituting all free occurrences of x_i in β by y_i, for all i.

Definition 6.5 [Variant formulas generated by variant probability terms] If α is a formula which contains a probability term $[\beta]_{\vec{x}}$ and $[\beta']_{\vec{y}}$ is a variant of $[\beta]_{\vec{x}}$, then the new formula $\alpha([\beta]_{\vec{x}}/[\beta']_{\vec{y}})$, which is the result of substituting the probability term by its variant, is a **variant** (formula) of α.

These variants are in addition to the variants generated by renaming the universally quantified variables.

Variants have the property that they preserve meaning, and the following theorem shows that this property is preserved by the extended set of variants defined in **Lp**.

Theorem 6.6 *If α' is a variant of α then for every interpretation function σ*

$$\alpha'^{\sigma} = \alpha^{\sigma}.$$

Furthermore, the underlying sets defined by two variant probability terms are identical. That is, if $\alpha' = [\beta']_{\vec{y}}$ and $\alpha = [\beta]_{\vec{x}}$, then

$$\{\vec{a} | \beta'^{\sigma(\vec{y}/\vec{a})} = \top\} = \{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

With the concept of a variant we can give a specification of how any term t can be substituted for any variable x in any formula α. We first form a variant of α, α', such that α' does not have any variables in common with t. Then we proceed to substitute t for all *free occurances* of x in α'. We denote the new formula formed by the substitution process $\alpha(x/t)$. For example, to substitute the term $\mathtt{f(y)}$ for $\mathtt{x}$ in the formula $[\mathtt{R(x,y)}]_{\mathtt{y}} > \mathtt{0.5}$ we first form a variant; $[\mathtt{R(x,z)}]_{\mathtt{z}} > \mathtt{0.5}$ will do. Then we preform the substitution to produce $[\mathtt{R(f(y),z)}]_{\mathtt{z}} > \mathtt{0.5}$. That is, this formula is $([\mathtt{R(x,y)}]_{\mathtt{y}} > \mathtt{0.5})(\mathtt{x/f(y)})$.

The next theorem shows that substitution in **Lp** behaves semantically in the same manner as in first-order logic.

Theorem 6.7 *For all α, t, x and interpretations σ*

$$\alpha(x/t)^{\sigma} = \alpha^{\sigma(x/t')} \qquad \text{where } t' = t^{\sigma}.$$

The results of this section allow us to prove that the subsets of $\mathcal{O}^n$ defined by the formulas of **Lp** forms a field of subsets. This fact will be used later in the proof of completeness. We include the proof here since it gives a flavor of the semantic behavior of the probability terms.

Theorem 6.8 *The set of subsets of $\mathcal{O}^n$ defined by the formulas of* **Lp**, *is a field of subsets.*

Proof: Let A and B be two subsets of $\mathcal{O}^n$ defined by formulas of **Lp**, i.e., $A = \{\vec{a}|\alpha^{\sigma(\vec{x}/\vec{a})} = \top\}$ and $B = \{\vec{b}|\beta^{\sigma(\vec{y}/\vec{b})} = \top\}$. By Definition 6.4, there exists two variants of $[\alpha]_{\vec{x}}$ and $[\beta]_{\vec{y}}$, $[\alpha']_{\vec{z}}$ and $[\beta']_{\vec{z}}$, formed by substituting all the variables $x_i \in \vec{x}$ in α and all the variables $y_i \in \vec{y}$ in β by a new set of variables $\langle \vec{n}z \rangle$ which do not appear in α or β. By Theorem 6.6, $A' = \{\vec{c}\,|{\alpha'}^{\sigma(\vec{z}/\vec{c})} = \top\} = A$, and $B' = \{\vec{c}\,|{\beta'}^{\sigma(\vec{z}/\vec{c})} = \top\} = B$; thus, $A \cap B = A' \cap B' = \{\vec{c}\,|(\beta' \wedge \alpha')^{\sigma(\vec{z}/\vec{c})} = \top\}$. That is, the intersection of A and B is definable by a formula of **Lp**. Similarly, for A, as defined above, by the semantic definition, $\alpha^{\sigma(\vec{x}/\vec{a})} = \top$ iff $\neg\alpha^{\sigma(\vec{x}/\vec{a})} = \bot$. Thus, $\vec{a} \in A$ iff $\vec{a} \notin A' = \{\vec{a}|\neg\alpha^{\sigma(\vec{x}/\vec{a})} = \top\}$. That is, A' is the complement of A with respect to $\mathcal{O}^n$, and is definable by a formula of **Lp**. Finally, if we take the term $[\alpha \vee \neg\alpha]_{\vec{x}}$ the set $A = \{\vec{a}|(\alpha \vee \neg\alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is equal to $\mathcal{O}^n$; thus the universal set is definable by a formula of **Lp**. Hence, the set of subsets of $\mathcal{O}^n$ definable by formulas of **Lp** is closed under intersections and complementations, and it contains $\mathcal{O}^n$. ■

6.2 Proof Theory

This section gives a proof theory for **Lp**. The proof theory consists of a set of axioms and rules of inference, and it is shown to be both sound and complete. There are, in addition to the normal first-order axioms, two new sets of axioms. One set of axioms defines the logic of the probability terms, and the other set defines the logic of the field $\mathcal{F}$.

In this subsection α, β, etc., will usually be used to represent formulas, not formulas or terms, as was the common usage in the previous subsection. It will be explicitly stated when they may also refer to terms.

6.2.1 Axioms and Rules of Inference

First the axioms and rules of inference (actually there is only one) for the proof theory are presented.

If α is a formula of **Lp** then a *generalization* of α is any formula of the form $\forall x_1 \ldots \forall x_n.\alpha$, where $\{\vec{n}x\}$ is a set variables of any type.

First-order Axioms All the axioms of the Predicate Calculus with equality.

PC1a) $\alpha \Rightarrow \beta \Rightarrow \alpha$.

PC1b) $(\alpha \Rightarrow \beta \Rightarrow \delta) \Rightarrow (\alpha \Rightarrow \beta) \Rightarrow \alpha \Rightarrow \delta$.

PC1c) $(\neg\alpha \Rightarrow \beta) \Rightarrow (\neg\alpha \Rightarrow \neg\beta) \Rightarrow \alpha$.

PC2) $\forall x.(\alpha \Rightarrow \beta) \Rightarrow \forall x.\alpha \Rightarrow \forall x.\beta$.

PC3) $\alpha \Rightarrow \forall x.\alpha$,
where x is not free in α.

PC4) $\forall x.\alpha \Rightarrow \alpha(x/t)$,
where t is any term, of the same type as x.

EQ5) $t = t$,
where t is any term.

EQ6) $t_1{=}t_{n+1} \Rightarrow \cdots \Rightarrow t_n{=}t_{2n} \Rightarrow ft_1 \ldots t_n{=}ft_{n+1} \ldots t_{2n}$,
where f is any n-ary function symbol and $t_1, \ldots, t_{2n}$ are terms of a compatible type.

EQ7) $t_1{=}t_{n+1} \Rightarrow \cdots \Rightarrow t_n{=}t_{2n} \Rightarrow Pt_1 \ldots t_n{\Rightarrow}Pt_{n+1} \ldots t_{2n}$,
where P is any n-ary predicate symbol and $t_1, \ldots, t_{2n}$ are terms of the same type.

Note that the axioms **PC1** are axioms which can generate all tautologies of the propositional calculus.

Field Axioms All of the axioms of a totally ordered field (see [MacLane and Birkhoff, 1968]). Here all variables are field variables and they are all *universally quantified*, unless the existential quantifier is used.

F1) $x + (y + z) = (x + y) + z$.

F2) $x + 0 = x$.

F3) $\exists y.(x + y = 0)$.

F4) $x + y = y + x$.

F5) $x \times 1 = x$.

F6) $x \times (y \times z) = (x \times y) \times z$.

F7) $x \times y = y \times x$.

F8) $x \times (y + z) = (x \times y) + (x \times z)$.

F9) $1 \geq 0 \wedge \neg(1 = 0)$.

F10) $x \neq 0 \Rightarrow \exists y.(y \times x = 1)$.

F11) $(x \geq y \wedge y \geq z) \Rightarrow x \geq z$.

F12) $(x \geq y \wedge y \geq x) \Rightarrow x = y$.

F13) $x \geq x$.

F14) $x \geq y \vee y \geq x$.

F15) $x \geq y \Rightarrow x + z \geq y + z$.

F16) $(x \geq y \wedge z \geq 0) \Rightarrow x \times z \geq y \times z$.

Probability Function Axioms

P1) $\forall x_1 \ldots \forall x_n.\alpha \Rightarrow [\alpha]_{\vec{x}} = 1$,
where $\vec{x} = \langle \vec{n}x \rangle$ and every x_i is an object variable.
This axiom says that if the set of satisfying instances of a formula are all the vectors of $\mathcal{O}^n$ then the probability of this set will be one. That is, $\mathcal{O}^n$ has probability one.

P2) $[\alpha]_{\vec{x}} \geq 0$.

P3) $[\alpha]_{\vec{x}} + [\neg\alpha]_{\vec{x}} = 1$.

P4) $[\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq [\alpha \vee \beta]_{\vec{x}}$.

P5) $[\alpha \wedge \beta]_{\vec{x}} = 0 \Rightarrow [\alpha]_{\vec{x}} + [\beta]_{\vec{x}} = [\alpha \vee \beta]_{\vec{x}}$.
These four axioms express the normal behavior of probabilities, i.e., that they are non-negative, the entire domain has probability one, and they are finitely-additive.

P6) $[\alpha]_{\vec{x}} = [\alpha(x_i/z)]_{\vec{x}(x_i/z)}$,
where z is an object variable that does not occur in α, and $\vec{x}(x_i/z)$ is the new vector of object variables: $\langle x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n \rangle$.
This axiom captures the fact that variant probability terms are equal (Theorem 6.6).

P7) $\forall \mathbf{z}_1 \mathbf{z}_2 . [[\alpha]_{\vec{x}} = \mathbf{z}_1]_{\vec{y}} = \mathbf{z}_2 \Rightarrow ([\alpha]_{\langle \vec{x}, \vec{y} \rangle} \geq \mathbf{z}_1 \times \mathbf{z}_2)$.
This is a technical axiom which enforces the product measure constraint. The completeness proof will demonstrate how it accomplishes this.

P8) $[\alpha]_{\vec{x}} = [\alpha]_{\pi(\vec{x})}$,
where π is any permutation of $\{1, \ldots, n\}$, and $\pi(\vec{x})$ is the permuted vector $\vec{x}$, i.e., $\pi(\vec{x}) = \langle x_{\pi(1)}, \ldots, x_{\pi(n)} \rangle$.
This axiom enforces the constraint that the probabilities are invariant under permutations.

Generalization

G1) All generalizations of the preceding axioms.

Rule of inference The only rule of inference is *modus ponens*, i.e.,

R1) From $\{\alpha, \alpha \Rightarrow \beta\}$ infer β.

6.2.2 Deductions

Here we review the standard notion of a special sequence of formulas called a deduction, and notes some of its properties.

Definition 6.9 Let Φ be a set of **Lp** formulas. A **deduction Φ in Lp** is a finite non-empty sequence of formulas $\vec{n}\phi$ such that for each k $(1 \leq k \leq n)$ ϕ_k is an axiom of **Lp**, or $\phi_k \in \Phi$, or ϕ_k is obtained by modus ponens from earlier formulas in the same sequence (i.e., there exists $i, j \leq k$ such that $\phi_j = \phi_i \Rightarrow \phi_k$). The set Φ is called the **set of hypotheses**. If Φ is empty the deduction is called a **proof**, i.e., a proof is a deduction which just uses the axioms of **Lp**. A deduction whose last formula is α is called a **deduction of** α. The symbol '$\vdash$' is used to indicate deducibility, i.e., '$\Phi \vdash \alpha$' means there is a deduction of α from Φ, and '$\vdash \alpha$' means that there is a proof of α.

Theorem 6.10 (Deduction Theorem) *Given a deduction of β from $\{\Phi, \alpha\}$, a deduction of $\alpha \Rightarrow \beta$ can be constructed from Φ.*

Proof: Standard first-order proof holds. ■

To demonstrate some deductions in **Lp** we prove the following results:

Lemma 6.11 *The following are provable in* **Lp**.

a) $=$ is an equivalence relation. That is, for any terms t_1, t_2, and t_3 of **Lp** we have:

(i) $\vdash t_1{=}t_1$,

(ii) $\vdash t_1{=}t_2 \Rightarrow t_2{=}t_1$,

(iii) $\vdash t_1{=}t_2 \wedge t_2{=}t_3 \Rightarrow t_1{=}t_3$.

b) $\vdash ([\alpha \Rightarrow \beta]_{\vec{x}} = 1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}} = 1) \Rightarrow [\alpha]_{\vec{x}} = [\beta]_{\vec{x}}$.

Proof: The first proposition will be used in the completeness proof, and the second is a handy fact about the probability terms which is often used. The proofs demonstrate the nature of symbolic reasoning with the various axioms using the formal proof theory. The proof of **(a)** is a standard first-order proof.

a) $t_1{=}t_1$ is an instance of axiom EQ5. With the predicate symbol P taken to be the equality predicate symbol '=' we have $t_1{=}t_2 \Rightarrow t_1{=}t_1 \Rightarrow t_2{=}t_1$ is an instance of axiom EQ7. So we have $t_1{=}t_2 \vdash t_2{=}t_1$. And, by the deduction theorem, $\vdash t_1{=}t_2 \Rightarrow t_2{=}t_1$. Also, $t_2{=}t_1 \Rightarrow t_2{=}t_3 \Rightarrow t_2{=}t_2 \Rightarrow t_1{=}t_3$ is another instance of EQ7. Since $t_1{=}t_2 \wedge t_2{=}t_3 \vdash t_1{=}t_2$ by tautologies and $t_1{=}t_2 \vdash t_2{=}t_1$, we have, through applications of modus ponens, $t_1{=}t_2 \wedge t_2{=}t_3 \vdash t_1{=}t_3$. Thus, $\vdash t_1{=}t_2 \wedge t_2{=}t_3 \Rightarrow t_1{=}t_3$, by the deduction theorem.

b) We construct a deduction of $[\alpha]_{\vec{x}} = [\beta]_{\vec{x}}$ from $[\alpha \Rightarrow \beta]_{\vec{x}}{=}1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}}{=}1$ (The axiom or rule of inference used in each step is specified at the right).

$\Big([\alpha \Rightarrow \beta]_{\vec{x}} = 1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}} = 1\Big) \Rightarrow [\alpha \Rightarrow \beta]_{\vec{x}} = 1$	(PC1)
$[\alpha \Rightarrow \beta]_{\vec{x}} = 1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}} = 1$	(Hyp.)
$[\alpha \Rightarrow \beta]_{\vec{x}} = 1$	(m.p.)
$[\neg\alpha \vee \beta]_{\vec{x}} = 1$	
$[\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq [\neg\alpha \vee \beta]_{\vec{x}}$	(P4)
$[\neg\alpha \vee \beta]_{\vec{x}} = 1 \Rightarrow [\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq [\neg\alpha \vee \beta]_{\vec{x}} \Rightarrow [\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq 1$	(EQ7)
$[\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq 1$	(m.p.)
$[\neg\alpha]_{\vec{x}} + [\alpha]_{\vec{x}} = 1$	(P3)
$[\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq [\neg\alpha]_{\vec{x}} + [\alpha]_{\vec{x}}$	(EQ7)
$[\neg\alpha]_{\vec{x}} + (-)[\neg\alpha]_{\vec{x}} = 0$	(F3)
$[\neg\alpha]_{\vec{x}} + [\beta]_{\vec{x}} + (-)[\neg\alpha]_{\vec{x}} \geq [\alpha]_{\vec{x}} + [\alpha]_{\vec{x}} + (-)[\neg\alpha]_{\vec{x}}$	(F15, m.p)
$[\beta]_{\vec{x}} + 0 \geq [\alpha]_{\vec{x}} + 0$	(F4, EQ7)
$[\beta]_{\vec{x}} \geq [\alpha]_{\vec{x}}$.	(F2, EQ7)
Similarly from $[\neg\beta \vee \alpha]_{\vec{x}} = 1$ we derive	
$[\alpha]_{\vec{x}} \geq [\beta]_{\vec{x}}$	
thus	
$[\beta]_{\vec{x}} = [\alpha]_{\vec{x}}$.	(F12, m.p.)

So $[\alpha \Rightarrow \beta]_{\vec{x}}{=}1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}}{=}1 \vdash [\alpha]_{\vec{x}}{=}[\beta]_{\vec{x}}$; thus by the deduction theorem,

$$\vdash ([\alpha \Rightarrow \beta]_{\vec{x}}{=}1 \wedge [\beta \Rightarrow \alpha]_{\vec{x}}{=}1) \Rightarrow [\alpha]_{\vec{x}}{=}[\beta]_{\vec{x}}.$$

■

6.2.3 Soundness and Completeness of the Proof Theory

The proof theory given above is sound and complete with respect to the class of **Lp**-structures defined in Section 4.3. Completeness is proved by way of a Henkin construction and is given in the appendix, along with the proof of soundness.

Let Φ be a set of **Lp** sentences, and let α an **Lp** sentence.

Theorem 6.12 (Completeness) *If* $\Phi \models \alpha$, *then* $\Phi \vdash \alpha$.

This theorem says that if the sentence α is true in every **Lp**-structure which is a model for Φ (i.e., every sentence of Φ is true in the structure) then a syntactic proof of α from Φ will exist using the proof theory supplied. In other words, syntactic proofs exist for every semantic entailment.

Theorem 6.13 (Soundness) *If* $\Phi \vdash \alpha$, *then* $\Phi \models \alpha$.

This theorem says that if α is provable from Φ using the proof theory then α must be true in every **Lp**-structure which is a model of Φ. In other words, syntactic proofs only generate valid semantic entailments.

6.3 Properties of the Probability Terms

This section demonstrates some of the properties of the probability terms. The existence of a completeness proof allows a proof of these lemmas from the semantics; the corresponding syntactic proof is guaranteed to exist. In these cases, a proof from the semantics is much simpler, as it just requires using some notions from set theory and probability theory, whereas, a syntactic proof would involve extensive symbolic manipulation, a task more suited to an automatic theorem prover. It should be noted that the semantic proofs only use those properties of probabilities that are true of the finitely-additive, field-valued probabilities used in the **Lp**-structures. The guarantee that a syntactic proof will exist does not hold if we use any of the special properties of standard probabilities (i.e., if we use sigma-additivity or any special properties of the reals).

Lemma 6.14 *The following are provable in* **Lp**.

a) $[\alpha]_{\vec{x}} \leq 1$.

b) $[\alpha \wedge \beta]_{\vec{x}} \leq [\alpha]_{\vec{x}}$ and $[\alpha \wedge \beta]_{\vec{x}} \leq [\beta]_{\vec{x}}$.

c) $[\alpha \vee \beta]_{\vec{x}} \geq [\alpha]_{\vec{x}}$ and $[\alpha \vee \beta]_{\vec{x}} \geq [\beta]_{\vec{x}}$.

d) $[\alpha \vee \beta]_{\vec{x}} = [\alpha]_{\vec{x}} + [\beta]_{\vec{x}} - [\alpha \wedge \beta]_{\vec{x}}$.

Proof: All of these results can be simply deduced from the fact that semantically the probability terms represent assignments of probability. That is, each probability term represents the probability of a corresponding set of objects in $\mathcal{O}^n$. Hence, all of these results follow from the properties of the probability functions μ_n. Equivalently they can be deduced from the probability and field axioms, in a manner similar to the proof of Lemma 6.11. ■

That these results are provable in **Lp** is an important point. They indicate that the probability functions have the familiar properties of standard probability functions, even though they assume values in an arbitrary totally ordered field instead of in the field of real numbers.

The advantage of having the field axioms arises in those situation where numeric probabilities are not available. In this case the field axioms allow one to reason with whatever information is available. For example, if the knowledge base contained the set of statements $\{[\mathtt{P(x)}]_{\mathtt{x}} > [\mathtt{Q(x)}]_{\mathtt{x}}$, $[\mathtt{Q(x)}]_{\mathtt{x}} > [\mathtt{R(x)}]_{\mathtt{x}}\}$, then it would be possible, using axiom F11, to infer $[\mathtt{P(x)}]_{\mathtt{x}} > [\mathtt{R(x)}]_{\mathtt{x}}$, even though no numeric values were available. The axioms also allow one to combine numeric and qualitative reasoning. For example if the knowledge base contained $\{[\mathtt{P(x)}]_{\mathtt{x}} > [\mathtt{Q(x)}]_{\mathtt{x}}$, $[\mathtt{Q(x)}]_{\mathtt{x}} = [\mathtt{R(x)}]_{\mathtt{x}} + [\mathtt{S(x)}]_{\mathtt{x}}$, $[\mathtt{R(x)}]_{\mathtt{x}} = .7$, $[\mathtt{S(x)}]_{\mathtt{x}} = .2\}$, the field axioms, the axioms of equality, and some numeric computation could be used to infer that $[\mathtt{P(x)}]_{\mathtt{x}} > 0.9$.

Lemma 6.15 (Bayes's Theorem) *Using Definition 4.2, the following is provable in* **Lp***:*

$$([\alpha]_{\vec{x}} \neq 0 \wedge [\beta]_{\vec{x}} \neq 0) \Rightarrow [\beta|\alpha]_{\vec{x}} = [\alpha|\beta]_{\vec{x}} \times \frac{[\beta]_{\vec{x}}}{[\alpha]_{\vec{x}}}.$$

This theorem shows that the powerful mechanisms of Bayesian inference are also valid in **Lp**. Bayesian analysis is useful when numeric probabilities are available. It requires a certain minimum amount of probabilistic information (although, as Pearl has shown [1986a], the information requirements can be made reasonable if knowledge of dependencies are also available). Inference engines formally based on Bayes's theorem and the laws of probability can be used on numeric probabilities expressed in **Lp**. Since both the probability axioms and Bayes's theorem are valid in **Lp**, the conclusions obtained from such inference engines will be valid deductions in **Lp**.

Let Γ be any set of **Lp** sentences, including possibly the empty set. The next lemma shows that when $\Gamma \vdash \beta \Rightarrow \lambda$, λ does not affect the conditional probability.

Lemma 6.16 *If* $\Gamma \vdash \forall x_1, \ldots, x_n.\beta \Rightarrow \lambda$ *then* $\Gamma \vdash [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}$.

Proof: If $[\beta]_{\vec{x}} = 0$ then, by the definition of the conditional probabilities, we will have that both probability terms are equal to zero.

Let $\mathcal{M} \models \Gamma$ be a model of Γ and let σ be any interpretation whose underlying structure is $\mathcal{M}$. If $[\beta]_{\vec{x}} > 0$ then $\{\vec{a}|(\beta)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is not empty. Let $\vec{c}$ be a member of this set. By the soundness theorem $\Gamma \models \forall x_1, \ldots, x_n.\beta \Rightarrow \lambda$. Hence, since $\beta^{\sigma(\vec{x}/\vec{c})} = \top$ we have that $\lambda^{\sigma(\vec{x}/\vec{c})} = \top$, and, by the semantic definition, $\vec{c} \in \{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Therefore, we have $\{\vec{a}|(\beta)^{\sigma(\vec{x}/\vec{a})} = \top\} \subset \{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Clearly, the opposite containment also holds, hence, the two sets are equal. By the semantic definition we have $[\beta]_{\vec{x}} = [\beta \wedge \lambda]_{\vec{x}}$, and it is easy to show that $[\beta \wedge \alpha]_{\vec{x}} = [\beta \wedge \lambda \wedge \alpha]_{\vec{x}}$ also. The lemma follows from the definition of conditional probabilities and the completeness theorem. ■

We can also show that deductive consequences always have greater conditional probability.

Lemma 6.17 *If* $\Gamma \vdash \forall x_1 \ldots x_n.(\beta \Rightarrow \lambda)$ *then* $\Gamma \vdash [\lambda|\alpha]_{\vec{x}} \geq [\beta|\alpha]_{\vec{x}}$.

Proof: Again let σ be any interpretation whose underlying structure is a model of Γ. Using the soundness theorem it is easy to show that $\{\vec{a}|(\beta \wedge \alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is a subset of $\{\vec{a}|(\lambda \wedge \alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Since μ_n is a probability function $[\lambda \wedge \alpha]_{\vec{x}} \geq [\beta \wedge \alpha]_{\vec{x}}$, and the result follows from the definition of conditionals. ■

Finally we show that closed formulas (sentences) in **Lp** can only have probability 0 or 1. This result shows that **Lp** is not naturally suited for representing degree of belief probabilities. It should be noted, however, that there are technical encodings which permit **Lp** to represent degrees of belief [Abadi and Halpern, 1988].

Lemma 6.18 *If* α *is a closed formula then* $[\alpha]_{\vec{x}} = 0$ *or* 1.

Proof: By the semantic definition, for any interpretation σ:

$$([\alpha]_{\vec{x}})^{\sigma} = \mu_n\{\vec{a}|\alpha^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Since α has no free variables, σ and $\sigma(\vec{x}/\vec{a})$ will agree on all the free variables of α, for any $\vec{a}$. Hence, by Theorem 6.2, $\alpha^{\sigma} = \alpha^{\sigma(\vec{x}/\vec{a})}$. Either $\alpha^{\sigma} = \top$ or $\alpha^{\sigma} = \bot$, since σ is an interpretation and α is a formula. Thus, the above set of $\vec{a}$ is either all of $\mathcal{O}^n$ or the empty set, and, for any μ_n, the probability is either 0 or 1. ■

6.4 Examples of Reasoning with the Statistical Knowledge

Example 6 [Nilsson's Probabilistic Entailment] Nilsson [1986] developed a probability logic based on the possible worlds approach. He shows how the probabilities of sentences in the logic are constrained by known probabilities, i.e., constrained by the probabilities of a base set of sentences. For example, if $pr[P \wedge Q] = 0.5$, then the values of $pr[P]$ and $pr[Q]$ are both constrained to be ≥ 0.5. Nilsson demonstrates how the implied constraints of a base set of sentences can be represented in a canonical manner, as a set of linear equations. These linear equations can be used to identify the strongest constraints on the probability of a new sentence, i.e., the tightest bounds on its probability. These constraints are, in Nilsson's terms, probabilistic entailments.

Nilsson gives some approximate methods for calculating these entailments, as well as noting that the methods of linear programming can give exact solutions. The important point, however, is that these bounds are simply consequences of the laws of probability.

The probabilities that Nilsson works with are degree of belief probabilities, while **Lp** deals with statistical probabilities. However, since both types of probabilities satisfy the same set of axioms there is a statistical analogue to Nilsson's probabilistic entailment. Furthermore, the proof theory we have developed is sufficiently powerful to capture the statistical version of probabilistic entailment.

For example, if the base set in Nilsson's logic is $\{pr[P]=0.6,\ pr[P \Rightarrow Q]=0.8\}$, probabilistic entailment gives the conclusion $0.4 \leq pr[Q] \leq 0.8$. If we write the symbols P and Q as one

place predicates, then in **Lp** we could write a statistical analogue of this base set as the formulas $[P(x)]_x = 0.6$, and $[P(x) \Rightarrow Q(x)]_x = 0.8$.

From this knowledge it is easy to deduce the bounds $[0.4, 0.8]$ on the probability term $[Q(x)]_x$. More generally, any bounds that can be computed using Nilsson's notion of probabilistic entailment can be deduced from **Lp**'s proof theory.

Example 7 [Reasoning with empirical generalizations (defaults)]

1. If the statement "P's are typically Q's" is given the statistical interpretation that the proportion of `P`'s that are `Q`'s is greater than $\mathbf{c}$, where $\mathbf{c}$ is some number close to 1, then the opposite conclusion, that "P's are typically not Q's," can be proved to be false.[19] That is,

$$[Q(x)|P(x)]_x > \mathbf{c} \wedge \mathbf{c} > 0.5 \vdash \neg\Big([\neg Q(x)|P(x)]_x > \mathbf{c}\Big).$$

   The derivation follows from axiom P3.

2. Similarly, the if the statement "P's are Q's" is asserted then the statement "P's are typically not Q's," can be proved to be false. For example, "Penguins are birds" implies that "Penguins are typically not birds" is false.

$$\forall x.penguin(x) \Rightarrow bird(x) \vdash \neg\Big([\neg bird(x)|penguin(x)]_x > \mathbf{c}\Big).$$

   The derivation follows from axioms P1 and P3.

3. The knowledge, "Most ravens are black" along with "Black objects are not white," can be used to deduce that "Most ravens are not white."

$$\begin{aligned}&\{[black(x)|raven(x)]_x > \mathbf{c},\\ &\quad \forall x.black(x) \Rightarrow \neg white(x)\}\\ &\vdash [\neg white(x)|raven(x)]_x > \mathbf{c}.\end{aligned}$$

   This can be shown with an argument similar to Lemma 6.17.

4. The knowledge, "Most birds fly" along with "Penguins do not fly", can be used to deduce that "Most birds are not penguins."

$$\begin{aligned}&\{[fly(x)|bird(x)]_x > \mathbf{c},\\ &\quad \forall x.penguin(x) \Rightarrow \neg fly(x)\}\\ &\vdash [\neg penguin(x)|bird(x)]_x > \mathbf{c}.\end{aligned}$$

   These examples demonstrate that if defaults are treated as qualitative statistical assertions we can use **Lp** to represent and reason with them. Using **Lp** in this way is the basis for a statistical approach to default reasoning [Bacchus, 1990; Bacchus, 1989].

[19] The fact that many non-monotonic formalisms allow both of these statements to be asserted without contradiction has been noted, and cited as a weakness, by both Touretzky et al. [1987] and Delgrande [1987].

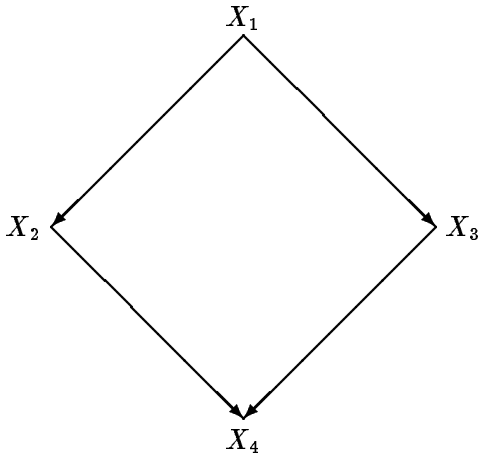


Figure 1: A Bayes net

Example 8 [Bayesian Networks]

Consider the Bayes net in Figure 1. If all of the variables X_1–X_4 are propositional (binary) variables one could write them as one place predicates in **Lp**. Hence, the open formula '$X_1(x)$', for example, would denote the set of individuals with property X_1. The Bayes net gives a graphical device for specifying a product form for the joint distribution of the properties X_i [Pearl, 1986a]. In this case the distribution represented by the Bayes net in the figure could also be specified by the **Lp** sentence

$$\begin{aligned}&[X_1(x) \wedge X_2(x) \wedge X_3(x) \wedge X_4(x)]_x \\ &\quad = \quad [X_4(x)|X_3(x) \wedge X_2(x)]_x \times [X_3(x)|X_1(x)]_x \times [X_2(x)|X_1(x)]_x \times [X_1(x)]_x\end{aligned}$$

It can easily be demonstrated that any probability distribution which satisfies this equation will also satisfy every equation of the same form with any number of the predicates negated (uniformly). For example, the equation

$$\begin{aligned}&[X_1(x) \wedge \neg X_2(x) \wedge X_3(x) \wedge \neg X_4(x)]_x \\ &\quad = \quad [\neg X_4(x)|X_3(x) \wedge \neg X_2(x)]_x \times [X_3(x)|X_1(x)]_x \times [\neg X_2(x)|X_1(x)]_x \times [X_1(x)]_x\end{aligned}$$

will be satisfied by every probability distribution which satisfies the first equation. Furthermore, the proof depends only on finite properties of the probability function, i.e., only on properties true of the probabilities used in the **Lp**-structures. Hence, by the completeness result, all such equations will be provable from the first via **Lp**'s proof theory.

This means that the behavior of the Bayes net is captured by the first **Lp** sentence. That is, the fact that this product decomposition holds for every instantiation of the properties X_i is captured by the proof theory.

In addition to the structural decomposition, Bayes nets must provide a quantification of the links. This means the conditional probabilities in the product must be specified. In this example if we add the **Lp** sentences $\{[X_1(x)]_x = 0.5, [X_2(x)|X_1(x)]_x = .75, [X_3(x)|X_1(x)]_x = .4, [X_4(x)|X_2(x) \wedge$

$X_3(x)]_x = .3\}$, we can then determine the proportion of individuals with any particular property among the set of individuals with any particular collection of properties, e.g., the values of terms like $[X_1(x)|X_2(x) \wedge \neg X_4(x)]_x$. Again these probabilities will be semantically entailed by the product decomposition and by the link conditional probabilities. Thus, the new probability values will be provable from the proof theory. Note that the probabilities represented in **Lp** are statistical. However, this is the appropriate interpretation for many applications of Bayes nets, particularly in expert system applications.

Of course the proof theory has none of the computational advantages of the Bayes net. However, what is important is that **Lp** provides a declarative representation of the net. The structure embedded in the net is represented in a form that can be reasoned with and can be easily changed. There is also the possibility of automatically compiling Bayes net structures from declarative **Lp** sentences. Furthermore, the proof theory offers the possibility of integrating Bayes net reasoning with more general logical and qualitative statistical reasoning.

7 Conclusions and Future Research

We have presented a logical framework for representing and reasoning with a very wide variety of statistical and logical information. The logical framework includes a proof theory which gives a formal specification for the set of valid inferences that can be made from a knowledge base of statistical and logical information. The interesting feature of the proof theory is that it captures the interaction between qualitative statistical reasoning, quantitative statistical reasoning, and first-order logical reasoning. Examples have been provided which demonstrate the framework's representational and reasoning power.

Given the ability to represent statistical information provided by the formalism, many interesting applications of statistical information become possible.

Mechanisms for integrating **Lp** with formalisms for representing degree of belief probabilities have already been developed by Halpern [1989], who has built on the logic developed here (see [Bacchus, 1990] for a detailed description). This work has produced a formalism that can deal with both types of probabilities and can represent and agent's degrees of belief in *statistical assertions*. This opens the door to further work using statistical knowledge.

One such area is the problem of statistical inference: how can an agent infer degrees of belief in statistical assertions from knowledge of individual cases. Such work would have an impact on learning research.

Another application of **Lp**'s ability to represent statistical knowledge comes in non-monotonic reasoning. A system of non-monotonic reasoning has been developed [Bacchus, 1990] which bases its default conclusions on statistical information. As one of our examples demonstrated, using **Lp** gives one the ability to reason extensively with the defaults.

Lp has much in common with ordinary first-order logic, so it would be useful to review some classical applications of first-order logic in AI with an eye to incorporating input from statistical sources. In particular, planning and diagnosis could both be viewed from a more general point of view once one has a general mechanism for handling statistical information.

Finally, there is the problem of automated reasoning within this framework. It is clear that a naive application of automated theorem proving techniques would not be sufficient; reasoning

about the field terms would probably prove to be very difficult. Much more promising and interesting would be the investigation of hybrid techniques using existent probabilistic reasoners, like Bayes nets, in conjunction with slower but more general ATP techniques. The use of such hybrid techniques has already proved to be very useful in speeding up deductive inference [Schubert *et al.*, 1987]. The important point here is that the logic and its proof theory provides formal tools which can be used to analyze the completeness and soundness of particular implementations.

In sum, the work presented here is a necessary first step towards greater use of statistical information in AI, and opens up the possibility for many interesting applications of such information.

Acknowledgment

This work comes from my thesis, and my supervisor Len Schubert deserves a great deal of the credit for its generality. His suggestions usually turned out to be possible, and always profitable. My external examiner Joe Halpern went over the work very carefully and helped clean up a number of mathematical rough edges. In particular, through our discussions I gained a much greater understanding of the tradeoffs and subtleties involved in giving up sigma-additive, real-valued probabilities. I also got very useful input from Henry Kyburg, Jeff Pelletier, Randy Goebel, Rene Elio, Mohan Mathan, Teddy Seidenfeld, Hector Levesque and Ray Reiter.

References

[Abadi and Halpern, 1988] Martin Abadi and Joseph Y. Halpern. Decidability and expressiveness of first-order logics of probability. Technical Report RJ. 7220 (67987) 12/18/89, IBM Research, Almaden Research Center, 650 Harry Road, San Jose, California, 95120–6099, 1988.

[Aleliunas, 1986] Romas Aleliunas. Models of reasoning based on formal deductive probability theories. Technical report, University of Waterloo, 1986.

[Aleliunas, 1988] Romas Aleliunas. A new normative theory of probability logic. In *Proceedings of the Canadian Artificial Intelligence Conference*, pages 67–74. Morgan Kaufmann, San Mateo, California, 1988.

[Åqvist *et al.*, 1980] Lennard Åqvist, Jaap Hoepelman, and Christian Rohrer. Adverbs of frequency. In Christian Rohrer, editor, *Time, Tense and Quantifiers: Proceedings of the Stuttgart Conference on the Logic of Tense and Quantification.* M. Niemeyer, Tubingen, 1980.

[Bacchus, 1988a] Fahiem Bacchus. On probability distributions over possible worlds. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 15–21, 1988.

[Bacchus, 1988b] Fahiem Bacchus. *Representing and Reasoning With Probabilistic Knowledge.* PhD thesis, The University of Alberta, 1988. Available as University of Waterloo Research Report CS-88-31, Department of Computer Science, Waterloo, Ontario, Canada, N2L 3G1. pp. 1-135.

[Bacchus, 1989] Fahiem Bacchus. A modest, but semantically well founded, inheritance reasoner. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, 1989.

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

[Barwise, 1977] Jon Barwise. An introduction to first-order logic. In Jon Barwise, editor, *Handbook of Mathematical Logic*, chapter A.1. North-Holland, Amsterdam, 1977.

[Bell and Machover, 1977] John Bell and Moshé Machover. *A Course in Mathematical Logic.* Elsevier, Amsterdam, 1977.

[Bundy, 1985] Alan Bundy. Incidence calculus: A mechanism for probabilistic reasoning. *Journal of Automated Reasoning*, 1:263–283, 1985.

[Carnap, 1962] Rudolf Carnap. *Logical Foundations of Probability.* University of Chicago Press, Chicago, 1962.

[Cheeseman, 1988] Peter Cheeseman. An inquiry into computer understanding. *Computational Intelligence*, 4(1), February 1988.

[Chung, 1974] Kai Lai Chung. *A Course in Probability Theory.* Academic Press, New York, 1974.

[De Finetti, 1964] Bruno De Finetti. Foresight: Its logical laws, its subjective sources. In Henry E. Kyburg, Jr. and H. Smokler, editors, *Studies in Subjective Probability.* John Wiley and Sons, New York, 1964.

[Delgrande, 1987] James P. Delgrande. A first-order conditional logic for prototypical properties. *Artificial Intelligence*, 33:105–130, 1987.

[Duda *et al.*, 1981] Richard O. Duda, Peter E. Hart, and Nils J. Nilsson. Subjective Bayesian methods for rule-based inference systems. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 192–199. Morgan Kaufmann, San Mateo, California, 1981.

[Etzioni, 1988] Oren Etzioni. Hypothesis filtering: A practical approach to reliable learning. In *Proceedings of the Fifth International Conference on Machine Learning*, 1988.

[Fagin *et al.*, 1988] Ronald Fagin, Joseph Y. Halpern, and Nimrod Megiddo. A logic for reasoning about probabilities. Technical Report RJ 6190 4/88, IBM Research, Almaden Research Center, 650 Harry Road, San Jose, California, 95120–6099, 1988.

[Feller, 1968] William Feller. *An Introduction to Probability Theory and Its Applications: Volume 1.* John Wiley and Sons, New York, 1968.

[Field, 1977] Hartley Field. Logic, meaning, and conceptual role. *Journal of Philosophy*, 77:374–409, 1977.

[Gaifman, 1964] Haim Gaifman. Concerning measures in first-order calculi. *Israel Journal of Mathematics*, 2:1–18, 1964.

[Geffner and Pearl, 1988] Hector Geffner and Judea Pearl. A framework for reasoning with defaults. Technical Report 870058 (R-94), Cognitive Systems Laboratory, U.C.L.A., Los Angeles, CA. 90024–1596, U.S.A., 1988.

[Halpern, 1989] Joseph Y. Halpern. An analysis of first-order logics of probability. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1375–1381, 1989.

[Hayes, 1985] Patrick J. Hayes. The second naive physics manifesto. In J. R. Hobbs and R. C. Moore, editors, *Formal Theories of The Commonsense World*, pages 71–107. Ablex Publishing, 1985.

[Heckerman and Horvitz, 1987] David Heckerman and Eric J. Horvitz. On the expressiveness of rule-based systems for reasoning with uncertainty. In *Proc. AAAI National Conference*, pages 121–126, 1987.

[Hempel, 1962] Carl G. Hempel. Deductive-nomological vs. statistical explanation. In Herbert Feigland and Grover Maxwell, editors, *Minnesota Studies in the Philosophy of Science Vol III*, pages 98–169. University of Minnesota Press, Minneapolis, 1962.

[Hintikka, 1966] Jaakko Hintikka. A two-dimensional continuum of inductive methods. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*. North-Holland, Amsterdam, 1966.

[Horvitz *et al.*, 1986] Eric J. Horvitz, David Heckerman, and Carl P. Langlotz. A framework for comparing alternative formalisms for plausible reasoning. In *Proc. AAAI National Conference*, pages 210–214, 1986.

[Jeffrey, 1983] Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, 1983.

[Johnson, 1986] R. W. Johnson. Independence and Bayesian updating methods. *Artificial Intelligence*, 29:217–222, 1986.

[Kanal and Lemmer, 1986] L. N. Kanal and J. F. Lemmer, editors. *Uncertainty in Artificial Intelligence Vol I*. North-Holland, Amsterdam, 1986.

[Kanal and Lemmer, 1987] L. N. Kanal and J. F. Lemmer, editors. *Uncertainty in Artificial Intelligence Vol II*. North-Holland, Amsterdam, 1987.

[Keisler, 1985] H. J. Keisler. Probability quantifiers. In J. Barwise and S. Feferman, editors, *Model Theoretic Logics*, chapter XIV. Springer-Verlag, New York, 1985.

[Kolmogorov, 1950] A. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, 1950.

[Koopman, 1940] Bernard O. Koopman. The axioms and algebra of intuitive probability. *Annals of Mathematics*, 41(2):269–292, April 1940.

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

[LeBlanc, 1983] Hughes LeBlanc. Alternatives to standard first-order semantics. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic. Vol II*, pages 225–258. D. Reidel, Dordrecht, Netherlands, 1983.

[Lindley, 1965a] D. V. Lindley. *Introduction to Probability and Statistics: Part 1: Probability.* Cambridge University Press, Cambridge, 1965.

[Lindley, 1965b] D. V. Lindley. *Introduction to Probability and Statistics: Part 2: Statistics.* Cambridge University Press, Cambridge, 1965.

[MacLane and Birkhoff, 1968] S. MacLane and G. Birkhoff. *Algebra.* Macmillan, London, 1968.

[McCarthy and Hayes, 1969] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[Morgan, 1984] Charles G. Morgan. Weak conditional comparative probability as a formal semantic theory. *Zeit. fur Math. Log.*, 30:199–212, 1984.

[Neyman, 1950] J. Neyman. *First Course in Probability and Statistics.* Holt, New York, 1950.

[Nilsson, 1986] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[Pearl and Paz, 1986] Judea Pearl and Azaria Paz. On the logic of representing dependencies by graphs. In *Proceedings of the Canadian Artificial Intelligence Conference*, pages 94–98. Morgan Kaufmann, San Mateo, California, 1986.

[Pearl and Verma, 1987] Judea Pearl and Thomas Verma. The logic of representing dependencies by directed graphs. In *Proc. AAAI National Conference*, pages 374–379, 1987.

[Pearl, 1986a] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[Pearl, 1986b] Judea Pearl. On the logic of probabilistic dependencies. In *Proc. AAAI National Conference*, pages 339–343, 1986.

[Popper, 1959] K. R. Popper. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42, 1959.

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

[Rosser and Turquette, 1952] J. B. Rosser and A. R. Turquette. *Many-Valued Logic.* North-Holland, Amsterdam, 1952.

[Salmon, 1967] Wesley Salmon. *The Foundations of Scientific Inference.* University of Pittsburgh Press, Pittsburgh, 1967.

[Savage, 1964] Leonard J. Savage. *The Foundations of Statistics.* Dover, New York, 1964.

[Schachter and Heckerman, 1987] Ross D. Schachter and David Heckerman. A backwards view for assessment. *AI Magazine*, 6(3):55–62, 1987.

[Schubert *et al.*, 1987] L. K. Schubert, M. A. Papalaskaris, and J. Taugher. Accelerating deductive inference: Special methods for taxonomies, colours, and times. In N. J. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 187–220. Springer-Verlag, New York, 1987.

[Schubert, 1988] L. K. Schubert. Cheeseman: a travesty of truth. *Computational Intelligence*, 4(1), February 1988.

[Scott and Krauss, 1966] Dana Scott and Peter Krauss. Assigning probabilities to logical formulas. In Jaakko Hintikka and Patrick Suppes, editors, *Aspects of Inductive Logic*. North-Holland, Amsterdam, 1966.

[Shoenfield, 1967] Joseph R. Shoenfield. *Mathematical Logic*. Addison-Wesley, London, 1967.

[Shortliffe and Buchanan, 1975] Edward H. Shortliffe and Bruce G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379, 1975.

[Simmons, 1986] Reid G. Simmons. "commonsense" arithmetic reasoning. In *Proc. AAAI National Conference*, pages 118–124, 1986.

[Tarski, 1951] A. Tarski. *A Decision Method for Elementary Algebra and Geometry, 2nd Edition.* University of California Press, Berkeley, 1951.

[Touretzky *et al.*, 1987] David S. Touretzky, John F. Horty, and Richmond H. Thomason. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, pages 476–482, 1987.

[van Fraassen, 1981] B. van Fraassen. Probabilistic semantics objectified. *Journal of Philosophic Logic*, 10:371–394, 1981.

[Venn, 1866] John Venn. *The Logic of Chance.* Chelsea Publishing Company, London, 1866.

[von Mises, 1957] Richard von Mises. *Probability, Statistics, and Truth.* Dover, New York, 1957.

A Appendix: Proof of Soundness and Completeness Theorems

Theorem (Soundness) **6.13** If $\Phi \vdash \alpha$, then $\Phi \models \alpha$.

Proof: Let $\vec{n}\phi$ be a deduction of α from Φ, i.e. $\phi_n = \alpha$. We show by induction on $k = 1, \ldots n$ that $\Phi \models \phi_k$. If ϕ_k is an axiom then we claim that ϕ_k is satisfied by every interpretation. Thus, $\Phi \models \phi_k$. If $\phi_k \in \Phi$ then it is clear that $\Phi \models \phi_k$. The last case is if for some $i, j < k$ we have $\phi_j = \phi_i \Rightarrow \phi_k$. By induction $\Phi \models \phi_i$ and $\Phi \models \phi_j$, so, from the semantic definition and the definition of '$\Rightarrow$', it follows that $\Phi \models \phi_k$. Now, all that remains is to prove the claim that the axioms of **Lp** are satisfied by every interpretation. The first-order axioms pose no problem, since **Lp** is an extension of first-order logic and has almost the same model structure. The standard proof of the soundness theorem for first-order logic suffices to show that these axioms are valid (satisfied by every interpretation). Since in the **Lp**-Structure $\mathcal{F}$ is defined to be an ordered field, it is clear that all of the field axioms are valid. Finally, since each μ_n is defined to be a probability function in the **Lp**-Structure, we can use the semantic definition of the probability terms $[\alpha]_{\vec{x}}$ to see that axioms P1-P5 are valid. Theorem 6.6 shows that axiom P6 is valid. The fact that the sequence of probability functions is a sequence of product measures yields the validity of axiom P7. The additional constraint **(2)** ensures that axiom P8 is valid. ■

We will give a rough proof of the completeness theorem. A rigorous treatment is contained in [Bacchus, 1988b]. The proof is accomplished by way of a Henkin construction, and the main complication is ensuring that all of the standard results about provable equivalence and maximal consistent sets of formulas that are true in first-order logic continue to hold in **Lp**.

A.1 Provable Equivalence

The notion of provable equivalence is important for the Henkin construction. It has already been shown that the terms in the language are divided into equivalence classes by the '=' relation (Lemma 6.11). We also need to show that the extended set of variants in **Lp** possess the property of provable equivalence.

Definition A.1 Two formulas α and β are said to be **provably equivalent** if $\alpha \vdash \beta$ and $\beta \vdash \alpha$. Two *terms* α and β are said to be **provably equivalent** if $\vdash \alpha = \beta$.

Theorem A.2 *If α' is a variant of α then α' and α are provably equivalent, where α can be* either *a term or a formula.*

A.2 Maximal Consistency

To prove the completeness theorem we need the notion of *maximal consistent* sets of formulas. It can be shown that the standard properties of maximal consistent sets of formulas continue to hold in **Lp**.

Definition A.3 A set of **Lp** formulas Φ is **inconsistent** if for some α both $\Phi \vdash \alpha$ and $\Phi \vdash \neg\alpha$, otherwise, Φ is **consistent**. Φ is **maximal consistent** if Φ is consistent and is not a proper subset of any other consistent set of formulas.

The important results about these notions are given by the following theorems.

Theorem A.4 *For any Φ and α,*

a) *$\Phi, \neg\alpha$ is inconsistent iff $\Phi \vdash \alpha$,*

b) *Φ, α is inconsistent iff $\Phi \vdash \neg\alpha$.*

Theorem A.5 *A set Φ is maximal consistent iff both of the following conditions are satisfied:*

a) *Φ is consistent.*

b) *For every formula α, $\alpha \in \Phi$ or $\neg\alpha \in \Phi$.*

Theorem A.6 *If Φ is maximal consistent and $\Phi \vdash \alpha$ then $\alpha \in \Phi$.*

A.3 Completeness of the Proof Theory

The completeness proof is a direct consequence of the following model existence proof which uses a standard Henkin construction with modifications to deal with the definition of the probability function and to handle the two sorted universe.

Theorem A.7 (Existence of a Model) *If Ω is a consistent set of* **Lp** *formulas then there exists an interpretation σ, with underlying* **Lp**-*Structure $\mathcal{M}$, which satisfies Ω. That is, $\beta^\sigma = \top$ for all $\beta \in \Omega$.*

Proof: First, we extend **Lp** to a new language **LP(C)** by adding a denumerable set of new constants $\{c_i | i = 0, 1, \ldots\}$. Clearly, if Ω is a consistent set of **Lp** formulas then it will also be a consistent set of **LP(C)** formulas. Next, we extend Ω to a maximal consistent set of **LP(C)** formulas Φ which has witnesses, i.e. if $\neg\forall x.\alpha \in \Phi$ then for some constant c, $\neg\alpha(x/c) \in \Phi$. If a witness for $\exists x.\alpha$ does not already exist in **Lp** we use one of the new constants as a witness.

Now we construct an **Lp**-Structure and an interpretation which satisfies the maximal set Φ. For each term t we define $[\![t]\!] = \{s | s = t \in \Phi\}$. By Lemma 6.11 it is deducible that '=' defines an equivalence relation and since Φ, being maximal consistent, is closed under deduction it follows that these are equivalence classes of terms in **Lp(C)**. Since these are equivalence classes, it is clear that $[\![t]\!] = [\![s]\!]$ iff $s = t \in \Phi$.

Lemma A.8 *Let $t_1, \ldots, t_{2n}$ be terms of identical type such that $[\![t_i]\!] = [\![t_{i+n}]\!]$ $(i = 1, \ldots, n)$ then:*

(a) *For any n-ary function symbol, f, of type compatible with the terms $t_1, \ldots, t_{2n}$*

$$[\![f\vec{n}t]\!] = [\![f t_{n+1} \ldots t_{2n}]\!],$$

(b) *For any n-ary predicate symbol, P, of the same type as the terms $t_1, \ldots, t_{2n}$*

$$\textit{if} \quad P\vec{n}t \in \Phi \qquad \textit{then} \quad P t_{n+1} \ldots t_{2n} \in \Phi.$$

For each variable, x, we put $x^\sigma = [\![x]\!]$, where x can be a variable of either type.

Finally we define the sequence of probability functions μ_n on any set of $\mathcal{O}^n$ defined by a formula α by

$$\mu_n\Big\{\langle [\![a_1]\!], \ldots, [\![a_n]\!]\rangle \colon \alpha^{\sigma(\vec{x}/\langle [\![a_1]\!],\ldots,[\![a_n]\!]\rangle)} = \top\Big\} = [\![[\alpha]_{\vec{x}}]\!].$$

Lemma A.9 *For any n the probability function μ_n is well defined. That is, if A is a set of tuples in $\mathcal{O}^n$ defined by two different formulas α and β then $\mu_n(A)$ is independent of which formula is used.*

Proof: By assumption, $A = \{\langle [\![a_1]\!], \ldots, [\![a_n]\!]\rangle \colon \alpha^{\sigma(\vec{x}/\langle [\![a_1]\!],\ldots,[\![a_n]\!]\rangle)} = \top\}$ and also $A = \{\langle [\![b_1]\!], \ldots, [\![b_n]\!]\rangle \colon \beta^{\sigma(\vec{y}/\langle [\![b_1]\!],\ldots,[\![b_n]\!]\rangle)} = \top\}$. By definition, $\mu_n(A) = [\![[\alpha]_{\vec{x}}]\!]$ also $\mu_n(A) = [\![[\beta]_{\vec{y}}]\!]$. The claim of the lemma is that $[\![[\alpha]_{\vec{x}}]\!] = [\![[\beta]_{\vec{y}}]\!]$. Let $\vec{z} = \langle \vec{n} z\rangle$ be a new set of object variables which do not appear in either α or β. There exists two variants α and β, called α' and β' respectively, formed by substituting all the variables $x_i \in \vec{x}$ in α and all the variables $y_i \in \vec{y}$ in β by the new variables $z_i \in \vec{z}$. By Theorem 6.6, the sets A' and B' (of tuples of $\mathcal{O}^n$) defined by these variants is the same as the set A. Further, by Theorem A.2, it is provable that $[\alpha']_{\vec{z}} = [\alpha]_{\vec{x}}$ also $[\beta']_{\vec{z}} = [\beta]_{\vec{y}}$. So by Theorem A.6 we have $[\![[\alpha']_{\vec{z}}]\!] = [\![[\alpha]_{\vec{x}}]\!]$ also $[\![[\beta']_{\vec{z}}]\!] = [\![[\beta]_{\vec{y}}]\!]$. Hence, the claim can be reduced to proving that $[\![[\alpha']_{\vec{z}}]\!] = [\![[\beta']_{\vec{z}}]\!]$.

Since

$$\{\langle [\![c_1]\!], \ldots, [\![c_n]\!]\rangle \colon \alpha'^{\sigma(\vec{z}/\langle [\![c_1]\!],\ldots,[\![c_n]\!]\rangle)} = \top\} = \{\langle [\![c_1]\!], \ldots, [\![c_n]\!]\rangle \colon \beta'^{\sigma(\vec{z}/\langle [\![c_1]\!],\ldots,[\![c_n]\!]\rangle)} = \top\}$$

it must be the case that the formulas $\forall z_i \ldots \forall z_n.(\alpha' \Rightarrow \beta')$ and $\forall z_i \ldots \forall z_n.(\beta' \Rightarrow \alpha')$ are in Φ. As Φ is maximal consistent, either these formulas or their negations must be in Φ. If their negations are in Φ it is easy to see, using the witness property of Φ, that the two sets A' and B' cannot be equal. Using axiom P1 and Theorem A.6, the formulas $[\alpha' \Rightarrow \beta']_{\vec{z}} = 1$ and $[\beta' \Rightarrow \alpha']_{\vec{z}} = 1$ must be in Φ. Thus, by Lemma 6.11, $[\alpha']_{\vec{z}} = [\beta']_{\vec{z}} \in \Phi$. Hence, by definition, $[\![[\alpha']_{\vec{z}}]\!] = [\![[\beta']_{\vec{z}}]\!]$. ■

This defines each μ_n on all subsets of $\mathcal{O}^n$ defined by formulas of **Lp**. It should also be clear from the construction of $\mathcal{O}$ that μ_n is also defined on each singleton set of $\mathcal{O}^n$, since the formula $[x_1 = t_1 \wedge \ldots \wedge x_n = t_n]_{\vec{x}}$ defines the singleton set $\{\langle [\![t_1]\!], \ldots, [\![t_n]\!]\rangle\}$. In an **Lp**-Structure each μ_n is defined on a field of subsets of $\mathcal{O}^n$, Π_n. However, Theorem 6.8 shows that the set of subsets defined by the formulas of **Lp** is itself a field of subsets. Hence, μ_n is already defined over a field of subsets which includes all singleton sets as well as all subsets defined by the formulas of **Lp**. That is, Π_n can be taken to be the field of subsets over which μ_n is already defined.

Lemma A.10 *For each term t $t^\sigma = [\![t]\!]$.*

Now we can prove that Φ is in fact satisfied by σ. We prove by induction (on the length of a formula β) that

(a) if $\beta \in \Phi$ then $\beta^\sigma = \top$, and

(b) if $\neg\beta \in \Phi$ then $\beta^\sigma = \bot$ (hence $\neg\beta^\sigma = \top$).

See [Bacchus, 1988b] for the details.

Thus $\beta^\sigma = \top$ for all $\beta \in \Phi$. Since Φ is maximal consistent it contains all instances of all axioms. Thus the structure and interpretation constructed satisfies all of these axioms. In particular, since all of the field axioms are true it is clear that $\mathcal{F}$ has the structure of a field. Further, since all of the probability axioms are true it is the case that the functions μ_n are in fact probability functions.

The sequence of probability functions is a sequence of product measures, since every instance of axiom P7 is true. Let $A = \{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\} \subset \mathcal{O}^n$ and $B = \{\vec{b} | \beta^{\sigma(\vec{y}/\vec{b})} = \top\} \subset \mathcal{O}^m$ be two sets in the domain of μ_n and μ_m respectively, with $\mu_n(A) = z_1$ and $\mu_m(B) = z_2$. It can be seen that the equivalence class of the probability term $[\alpha \wedge \beta]_{(\vec{x},\vec{y})}$ is equal to the probability of their Cartesian product. We have $[[\alpha \wedge \beta]_{\vec{x}} = z_1]_{\vec{y}} = z_2$ is true, so must be in Φ. Hence, by axiom P7 the probability of the Cartesian product is greater than or equal to $z_1 \times z_2$. It must be shown that it is in fact equal. This can be done by considering the complement of the Cartesian product. This set is not a product set, but it is equal to the union of two product sets. That is, it is equal to the (disjoint) union of $\neg A \times \mathcal{O}^m$ and $A \times \neg B$. Using P7 again, we see that the complement is greater than equal to $1 - z_1 + z_1 \times (1 - z_2)$, which is $1 - z_1 \times z_2$. The result now follows from axiom P3.

Axiom P8 insures that the probability functions satisfy the constraint of invariance under permutations.

Hence, the structure constructed is a valid **Lp**-Structure.

Since Ω is contained in Φ, it is obvious that σ satisfies Ω. That is, $\alpha^\sigma = \top$ for all $\alpha \in \Omega$ as claimed. ■

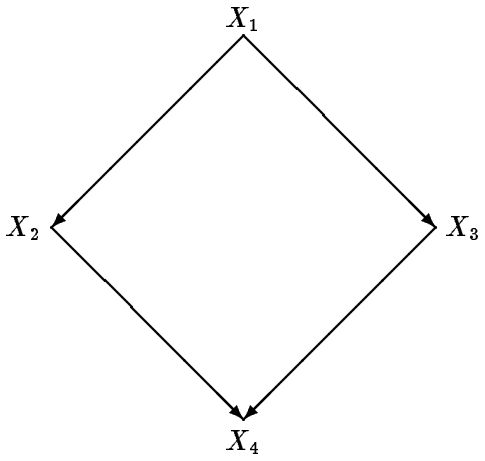
From the model existence theorem it is easy to prove completeness.

Theorem (Completeness) **6.12** If $\Phi \models \alpha$, then $\Phi \vdash \alpha$.

Proof: If $\Phi \models \alpha$ then no interpretation satisfies $\{\Phi, \neg\alpha\}$. Hence, by the Existence Theorem, $\{\Phi, \neg\alpha\}$ is inconsistent. Thus, by Theorem A.4, $\Phi \vdash \alpha$. ■

Figure Legends

Figure 1: A Bayes net.


X_1
X_2
X_3
X_4