Maximum Entropy Probabilistic Logic

Mark A. Paskin
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
paskin@cs.berkeley.edu

Abstract

Recent research has shown there are two types of uncertainty that can be expressed in first-order logic—propositional and statistical uncertainty—and that both types can be represented in terms of probability spaces. However, these efforts have fallen short of providing a general account of how to design probability measures for these spaces; as a result, we lack a crucial component of any system that reasons under these types of uncertainty. In this paper, we describe an automatic procedure for defining such measures in terms of a probabilistic knowledge base. In particular, we employ the principle of maximum entropy to select measures that are consistent with our knowledge and that make the fewest assumptions in doing so. This approach yields models of first-order uncertainty that are principled, intuitive, and economical in their representation.

Introduction

Integrating representations of uncertainty with the expressive semantics of first-order logic is the theme of much research in Artificial Intelligence. Recent work has shown that there are two types of uncertainty that can be expressed in first-order logic: *propositional uncertainty*, where we are uncertain of the truth of logical sentences, and *statistical uncertainty*, where we are uncertain of the distribution of properties across objects (Bacchus 1990). This work also shows that both types of uncertainty can be represented in terms of probability spaces (Halpern 1990). However, these efforts have fallen short of providing a general account of how to design and represent probability measures for these spaces; as a result, we lack a crucial component of any system that reasons under propositional and statistical uncertainty.

In this paper, we describe an automatic procedure for defining these measures in terms of a probabilistic knowledge base that contains certain and uncertain first-order knowledge. In general, our knowledge will be insufficient to determine the measures uniquely, and so we adopt the following strategy: we view the probabilistic knowledge base as a set of constraints, and of the measures that satisfy the constraints, we choose the one with maximum entropy (Jaynes 1979). We show that this choice leads to models of propositional and statistical uncertainty that are principled, intuitive, and economical in their representation.

We begin by reviewing the basic concepts underlying propositional uncertainty and then discuss its connection to the principle of maximum entropy. Next, we show how recent algorithmic advances, which provide general-purpose machinery to implement the maximum entropy principle, may be applied to yield principled and compact representations of propositional uncertainty. We then extend the approach to include statistical uncertainty, and show how uncertain knowledge of each type may be used to inform inferences of the other. We conclude with a discussion of some important issues and a summary of related work.

Degrees of Belief and Random Worlds

Nilsson (1986) was among the first to consider the problem of representing propositional uncertainty, i.e., uncertainty regarding the truth of logical sentences. For example, an agent may be unsure of the truth of such sentences as $\mathit{flies}(\mathit{Tweety})$ or $\forall x(\mathit{bird}(x) \to \mathit{flies}(x))$, and may ascribe a *degree of belief* (or probability of truth) to each. In this section, we describe one approach, which has become known as the *random worlds* formulation.

The truth of logical sentences is defined in terms of *possible worlds*. Let L be a finite first-order logic language (i.e., a collection of finitely many relation, function, and constant symbols, along with the usual variable symbols, connectives, quantifiers, and the equality symbol); let $\mathcal{S}_L$ be the set of sentences of L.[1] A *possible world* (or *structure*) w for L consists of: a set of objects $\mathcal{O}_w$ (called the *domain* of w); a set of relations over the domain, each corresponding to a relation symbol in L; and, a set of functions over the domain, each corresponding to a function symbol in L. (As usual, constant symbols can be treated as function symbols of zero arity.) The *universe of* L, denoted $\mathcal{U}_L$, is the set of all possible worlds for L.

We can represent the semantics of first-order logic by a deterministic *valuation function* $V : \mathcal{S}_L \times \mathcal{U}_L \to \{\mathrm{T}, \mathrm{F}\}$: a sentence is either true (T) or false (F) in each possible world. Thus, if our agent knew which of the possible worlds is the *actual world*, then it could apply the valuation function to

[1] While we focus on first-order logic languages, the framework trivially admits propositional logic languages as a special case.

infer the truth or falsehood of every sentence with certainty. We can therefore interpret its uncertainty regarding the truth of sentences as derivative of an underlying uncertainty regarding which of the possible worlds is the actual world.

The probabilistic way to model this uncertainty is with a *random world* W, i.e., a random variable that ranges over the possible worlds in $\mathcal{U}_L$; W is governed by a distribution (or measure) $\mathbf{P}_W$, called a *world model*. An agent's world model expresses its degree of belief that any possible world is the actual world, and can be used to compute the degree of belief (*sentence probability*) of a sentence α as:

$$\mathbf{P}_W\{\alpha\} \triangleq \mathbf{P}_W(\mathcal{U}_L^\alpha) = \sum_{w \in \mathcal{U}_L^\alpha} \mathbf{P}_W(w) \tag{1}$$

where $\mathcal{U}_L^\alpha \triangleq \{w \in \mathcal{U}_L : V(\alpha, w) = \mathrm{T}\}$ is the set of *models* of α. This notation generalizes naturally to sets of sentences, allowing us to express conditional degrees of belief.

Several intuitive properties follow immediately from the random worlds formulation. For example, for all world models $\mathbf{P}_W$: if $\alpha \equiv \beta$ then $\mathbf{P}_W\{\alpha\} = \mathbf{P}_W\{\beta\}$; if α is valid, then $\mathbf{P}_W\{\alpha\} = 1$; if α is unsatisfiable, then $\mathbf{P}_W\{\alpha\} = 0$; $\mathbf{P}_W\{\alpha\} + \mathbf{P}_W\{\neg\alpha\} = 1$ for all α; and if $\alpha \models \beta$, then $\mathbf{P}_W\{\alpha\} \leq \mathbf{P}_W\{\beta\}$ and $\mathbf{P}_W\{\beta \mid \alpha\} = 1$.

Random Worlds and Maximum Entropy

Given a probabilistic knowledge base that expresses our propositional uncertainty, we would like to compute degrees of belief for new sentences. Specifically, we will assume a probabilistic knowledge base consisting of a set of *facts* $A = \{\alpha_1, \ldots, \alpha_m\}$ and a set of *beliefs* $B = \{\beta_1, \ldots, \beta_n\}$; each fact α_i is a sentence that is known to be true with certainty, and each belief β_i is a sentence accompanied by a corresponding degree of belief $\mathbf{P}_W\{\beta_i\}$ (that is not 0 or 1).

The random worlds formulation allows us to reason under propositional uncertainty, given a world model. Thus, we view the task of reasoning from a probabilistic knowledge base as essentially that of building this measure. However, we are immediately faced with a problem of identifiability: in general, our probabilistic knowledge base can be compatible with infinitely many possible world models. We can either accept this indeterminacy (and perhaps compute bounds on degrees of belief), or introduce an additional criterion that eliminates it. We pursue the latter approach.

The principle of maximum entropy (Jaynes 1979) represents one method of selecting a unique measure. In this framework, we view our knowledge base as a set of constraints that must be satisfied by the world model: it must assign zero probability to worlds inconsistent with the facts, and it must agree with the sentence probability of each belief. Of all such measures, we select the one with maximum entropy; in information-theoretic terms, this corresponds to selecting the measure that makes the fewest assumptions necessary to be consistent with our knowledge base.

The maximum entropy principle is a general technique for selecting a unique measure in underdetermined problems, but we can give it further justification in the context of modeling random worlds. Paris (1999) and others have shown not only that maximum entropy world models are consistent with several common sense reasoning principles—such as insensitivity to renaming, indifference to irrelevant information, and the assumption of independence in the absence of explicit information to the contrary—but that they are determined by them; that is, any process that translates a probabilistic knowledge base into a world model and is consistent with these common sense reasoning principles must yield a maximum entropy world model.

The Automatic Construction of Maximum Entropy World Models

While the maximum entropy approach was suggested in Nilsson's original paper, no general purpose algorithm to implement it was provided.[2] Recent theoretical and algorithmic advances in the Statistics and Machine Learning communities provide us with the necessary tools to give a general solution. We first give a brief introduction to the framework, and then show how it may be applied to the problem of constructing maximum entropy world models.

Maximum Entropy Probability Models

Suppose we wish to model a random variable Y that ranges over some finite set of values $\mathcal{Y}$. We have access to a reference distribution $\mathbf{P}_Y$ (for example, an empirical distribution), and we wish to summarize this distribution by another, simpler distribution $\mathbf{Q}_Y$ that models Y in terms of a number of features. In particular, we select a set of feature functions $F = \{f_1, f_2, \ldots, f_n\}$, where each feature function $f_i : \mathcal{Y} \to \mathbb{R}$ maps possible values of Y to real numbers; intuitively, a feature function's output indicates the degree to which the corresponding feature is present.

We can summarize the reference distribution $\mathbf{P}_Y$ by applying the principle of maximum entropy, and defining $\mathbf{Q}_Y$ as the solution to the following optimization problem:

$$\begin{array}{ll} \text{maximize:} & H(\mathbf{Q}_Y) \\ \text{subject to:} & \mathbb{E}_{\mathbf{Q}_Y}[f_i] = \mathbb{E}_{\mathbf{P}_Y}[f_i] \quad (\forall f_i \in F) \end{array} \tag{2}$$

where

$$H(\mathbf{Q}_Y) = -\sum_{y \in \mathcal{Y}} \mathbf{Q}_Y(y) \log \mathbf{Q}_Y(y)$$

is the entropy of the distribution $\mathbf{Q}_Y$. That is, we seek a distribution that agrees with the reference distribution on the expectations of the feature functions and that has maximum entropy. If the reference expectations $\mathbb{E}_{\mathbf{P}_Y}[f_i]$ are our only knowledge of $\mathbf{P}_Y$, then the result of this optimization is a distribution that agrees with our knowledge and makes the fewest assumptions in doing so.

The maximum entropy optimization (2) must be performed over all possible distributions $\mathbf{Q}_Y$, which seems a daunting task. Consider a simpler method of summarizing $\mathbf{P}_Y$, in which we perform maximum likelihood optimization over the Gibbs (or log-linear) distribution

$$\mathbf{Q}_Y(y) \triangleq \frac{1}{Z(\Lambda)} \exp\left\{\sum_{i=1}^{n} \lambda_i f_i(y)\right\}$$

[2] Although an algorithm had been developed in the Statistics community (Darroch & Ratcliff 1972), it seems Nilsson was only aware of a somewhat less general formulation (Cheeseman 1983).

where $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the parameters and

$$Z(\Lambda) \triangleq \sum_{y \in \mathcal{Y}} \exp\left\{\sum_{i=1}^{n} \lambda_i f_i(y)\right\}$$

is a normalization factor. A surprising result proves not only that the solutions to these optimizations are unique, but that they are identical (Della Pietra, Della Pietra, & Lafferty 1997). Thus, we can calculate the maximum entropy distribution under expectation constraints by solving the maximum likelihood problem for Gibbs distributions.

Moreover, the maximum likelihood optimization for Gibbs distributions has several nice properties. First, the reference expectations are sufficient statistics, and so the reference distribution itself is not necessary to carry out the optimization. Second, maximum likelihood for Gibbs distributions is a convex optimization problem, and therefore can be solved by hill-climbing algorithms like Generalized Iterative Scaling (Darroch & Ratcliff 1972), Improved Iterative Scaling (Della Pietra, Della Pietra, & Lafferty 1997), or a number of more generic convex optimization algorithms.

Maximum Entropy World Models

The framework above has a natural application to the problem of modeling random worlds, but we face an immediate obstacle: it requires the modeled variable's range to be finite, but the number of possible worlds can be infinite (because possible worlds can have arbitrarily many objects).[3]A simple way to sidestep this difficulty is to impose a bound on the number of objects in each possible world; for a fixed, finite language, the set of such possible worlds is always finite. Thus, we will assume that one of the facts in our knowledge base is a sentence that enforces this constraint.[4]

As stated above, we seek a world model $\mathbf{Q}_W$ with three properties: (1) it should give positive probability only to worlds consistent with the facts A; (2) it should be consistent with the sentence probability $\mathbf{P}_W\{\beta_i\}$ of each belief $\beta_i \in B$; and (3) of all world models with these properties, $\mathbf{Q}_W$ should have maximum entropy.

To obtain $\mathbf{Q}_W$, we apply the maximum entropy framework described above. First, we choose the range of the random world to be $\mathcal{U}_L^A$, the set of worlds consistent with the facts; this ensures the first property. Next, we choose the appropriate set of features; for each belief β_i we include a feature function $f_{\beta_i} : \mathcal{U}_L^A \to \{0, 1\}$ defined by

$$f_{\beta_i}(w) \triangleq \begin{cases} 1 & \text{if } V(\beta_i, w) = \mathrm{T} \\ 0 & \text{otherwise} \end{cases}$$

Recall that the result of the maximum entropy optimization is a distribution $\mathbf{Q}_W$ that agrees with $\mathbf{P}_W$ on the feature expectations. Given our choice of feature functions, we have

$$\mathbb{E}_{\mathbf{Q}_W}[f_{\beta_i}(W)] = \mathbf{Q}_W(\mathcal{U}_L^{\beta_i}) = \mathbf{Q}_W\{\beta_i\}$$

Therefore, we can view the maximum entropy optimization (2) as solving the following problem:

$$\begin{array}{ll} \text{maximize:} & H(\mathbf{Q}_W) \\ \text{subject to:} & \mathbf{Q}_W\{\beta_i\} = \mathbf{P}_W\{\beta_i\} \quad (\forall \beta_i \in B) \end{array}$$

Thus, applying the maximum entropy framework ensures the second and third properties. The result of this process is a *maximum entropy world model* of the form

$$\mathbf{Q}_W(w) \triangleq \frac{1\{w \in \mathcal{U}_L^A\}}{Z(\Lambda)} \exp\left\{\sum_{\beta \in B} \lambda_\beta f_\beta(w)\right\} \tag{3}$$

where $Z(\Lambda)$ is the normalization constant.

A Simple Example

We illustrate the approach with the following example. Let our language have four constants (*Alice*, *Bob*, *Chris* and *David*), a unary relation symbol (*male*), and a binary relation symbol (*married*). Then our probabilistic knowledge base could contain the following facts:

$$\begin{array}{ll} \alpha_1: & \forall x(\neg \mathit{married}(x, x)) \\ \alpha_2: & \forall x, y(\mathit{married}(y, x) \leftrightarrow \mathit{married}(x, y)) \\ \alpha_3: & \forall x, y, z(\mathit{married}(x, y) \wedge \mathit{married}(x, z) \to (y = z)) \\ \alpha_4: & \forall x, y(\mathit{married}(x, y) \to \neg(\mathit{male}(x) \leftrightarrow \mathit{male}(y))) \\ \alpha_5: & \neg \mathit{male}(\mathit{Alice}) \wedge \mathit{male}(\mathit{Bob}) \wedge \mathit{male}(\mathit{David}) \end{array}$$

as well as the following beliefs:

$$\begin{array}{llll} \beta_1: & \mathbf{P}_W\{\mathit{male}(\mathit{Chris})\} & = & 0.85 \\ \beta_2: & \mathbf{P}_W\{\exists x(\mathit{married}(\mathit{Alice}, x))\} & = & 0.75 \end{array}$$

Given this probabilistic knowledge base, we would like to construct a world model so that we can compute new degrees of belief. First, we bound the number of objects we wish to reason about; in this example we assume *Alice*, *Bob*, *Chris* and *David* refer to four distinct objects. Next, we build the corresponding maximum entropy world model $\mathbf{Q}_W$; its parameters are $\lambda_{\beta_1} = 2.14$ and $\lambda_{\beta_2} = 0.11$. Finally, using this world model, we can compute new degrees of belief; for example, we find that $\mathbf{Q}_W\{\mathit{married}(\mathit{Alice}, \mathit{Bob})\} = 0.28$, but $\mathbf{Q}_W\{\mathit{married}(\mathit{Alice}, \mathit{Chris})\} = 0.19$.

Statistical Uncertainty and Random Objects

So far, we have restricted our attention to representing and reasoning under propositional uncertainty. However, as Bacchus (1990) and others have noted, there is another type of uncertainty that can be expressed in first-order logic languages. Compare the following two statements:

With 90% probability, all birds fly. (4)

90% of all birds fly. (5)

Statement (4) expresses propositional uncertainty: a probability is ascribed to the proposition $\forall x(\mathit{bird}(x) \to \mathit{flies}(x))$; the proposition is either true or false, and the probability represents our degree of belief that the proposition is true. In

[3]The requirement that the variable's range be finite stems from the fact that every value is assigned a probability that is bounded away from zero, yielding a finite sum only when the range is finite.

[4]We view this restriction largely as a technicality, since in many domains, the number of objects we wish to reason about can be bounded. However, it does limit representation power: there are first-order knowledge bases that are inconsistent in all worlds with finite domains, but are consistent in worlds with infinite domains.

contrast, the truth of statement (5) is not in doubt; rather, uncertainty is expressed regarding how properties are distributed across objects in the domain. In particular, (5) states that were we to "sample" objects from the domain, 90% of those satisfying *bird*(x) would also satisfy *flies*(x). This sort of uncertainty is called *statistical uncertainty*.

We have shown that representing propositional uncertainty can be reduced to representing uncertainty about which of the possible worlds is the actual world. In contrast, when representing statistical uncertainty, we will (initially) assume we know which of the possible worlds is the actual world. We consider a process in which we sample objects randomly from the actual world and represent uncertainty regarding the sorts of properties such objects will have. As with degree of belief, this notion of randomness can be formalized in terms of a probability space; we begin by introducing some helpful notation.

We will view a formula with k free variables as a k-ary predicate. Let ϕ be a k-ary formula with free variables $x_1, \ldots, x_k$ and let w be a possible world; then we define the *support of* ϕ *in* w as

$$\mathcal{O}_w^\phi \triangleq \left\{ \langle o_1, \ldots, o_k \rangle \in \mathcal{O}_w^k : V^{[x_1/o_1, \ldots, x_k/o_k]}(\phi, w) = \mathrm{T} \right\}$$

where $V^{[x_1/o_1, \ldots, x_k/o_k]}$ is the valuation function altered such that each variable symbol x_i is interpreted as a reference to the corresponding object o_i. In other words, $\mathcal{O}_w^\phi$ is the set of instantiations of ϕ's free variables that satisfy ϕ in w. If Φ is a set of k-ary formulas, then the *support of* Φ *in* w is $\mathcal{O}_w^\Phi \triangleq \cap_{\phi \in \Phi} \mathcal{O}_w^\phi$. Because an n-ary formula can trivially be considered a k-ary formula for $n \leq k$, this definition extends to the case where some formulas in Φ have a natural arity less than Φ.

As mentioned above, statistical uncertainty is based upon a notion of sampling objects from the domain. Let us assume that w is the actual world. We define a *random object of* w to be a random variable O ranging over $\mathcal{O}_w$, the domain of w; O is governed by a distribution $\mathbf{P}_O^w$ called an *object model for* w. We will also require a notion of sampling tuples of objects from the domain; to do this, we can extend an object model to ascribe probability to k-tuples of objects via the product construction:

$$\mathbf{P}_O^w(\langle o_1, o_2, \ldots, o_k \rangle) \triangleq \prod_{i=1}^{k} \mathbf{P}_O^w(o_i)$$

This reflects an assumption that objects are sampled independently and identically distributed from the domain.[5]

Of course, we are not really interested in sampling particular objects. Rather, we are interested in the properties they possess in the world; these properties are expressed using formulas. For a fixed world w, we can view a k-ary formula ϕ as denoting $\mathcal{O}_w^\phi$, the set of k-tuples of domain objects for which ϕ is satisfied. This leads us to define the *formula probability of* ϕ *in* w as

$$\mathbf{P}_O^w\{\phi\} \triangleq \mathbf{P}_O^w(\mathcal{O}_w^\phi) = \sum_{\langle o_1, \ldots, o_k \rangle \in \mathcal{O}_w^\phi} \mathbf{P}_O^w(\langle o_1, \ldots, o_k \rangle)$$

[5]This choice is not motivated purely by simplicity; there are more subtle reasons as well (Bacchus 1990, 89–90).

That is, the formula probability of ϕ in a world w is the probability of drawing an instantiation of ϕ's free variables from the domain of w so that ϕ is true in w. We can extend this notation to sets of formulas, allowing us to express conditional formula probabilities: if Φ and Ψ are sets of k-ary formulas (with $\mathbf{P}_O^w\{\Psi\} > 0$), then $\mathbf{P}_O^w\{\Phi \mid \Psi\} \triangleq \mathbf{P}_O^w\{\Phi \cup \Psi\}/\mathbf{P}_O^w\{\Psi\}$ is the conditional probability Φ holds given Ψ holds (under $\mathbf{P}_O^w$). Thus, the quantity referred to in statement (5) is exactly $\mathbf{P}_O^w\{\mathit{flies}(x) \mid \mathit{bird}(x)\}$.

Formalizing statement (5) in these terms highlights the fact that an agent's concept of the actual world and its object model represent independent explanations for the statistical predications it makes. The speaker reports the fact that $\mathbf{P}_O^w\{\mathit{flies}(x) \mid \mathit{bird}(x)\} = 0.9$; this may be a result of the composition of the actual world w, the way in which the agent samples the world, or some mixture of the two. For example, it may be that there is only one flightless bird in the actual world, but the agent samples it 10% of the time. To rule out this indeterminacy, we will assume *uniform object models*—each object is equally likely to be drawn from a world. This assumption is valid if the agent makes objective observations, but can also be justified if our goal is to model the agent's statistical knowledge of the world rather than its true composition (Bacchus 1990, 114–117).

Reasoning With Statistical Uncertainty Under Maximum Entropy

Propositional and statistical uncertainty are complementary, and there are two ways in which we can incorporate statistical uncertainty into the framework we have developed thus far. In the first, we can use degrees of belief to generate statistical knowledge; in the second combination, we can use statistical knowledge to alter our degrees of belief. We treat each of these combinations in turn.

From Degrees of Belief to Statistical Knowledge

In the first combination, we allow our statistical estimates to be influenced by our beliefs regarding the truth of sentences. For example, if an agent is mostly certain that all birds fly (a proposition), it should also think it likely that a randomly-sampled bird would fly. Thus far, we have assumed knowledge of the actual world w when performing statistical inference; the natural way to effect this combination is to replace the actual world w with a random world W, governed by a world model $\mathbf{P}_W$ (Halpern 1990).

Let $\mathbf{P}_W$ be a world model, and for each possible world w let $\mathbf{U}_O^w$ be a uniform object model for w. We define the *probability of a formula* ϕ (over all worlds) to be

$$\mathbf{P}_O\{\phi\} \triangleq \sum_{w \in \mathcal{U}_L} \mathbf{U}_O^w\{\phi\}\mathbf{P}_W\{w\} \tag{6}$$

We can define joint and conditional formula probabilities similarly. Thus, we can represent uncertainty regarding the objects of a fixed possible world using an object model $\mathbf{P}_O^w$; but given a world model $\mathbf{P}_W$ and the assumption of uniform object models, we can express statistical information about the properties of objects across many possible domains.[6] Returning to our example: an agent that strongly believes all birds fly has a world model in which worlds with flightless birds are improbable; when this agent samples a bird object from the mixture of these worlds, the result will probably come from a world in which all birds fly.

From Statistical Knowledge to Degrees of Belief

In the second combination, we would like to inform our degrees of belief with statistical knowledge. For example, an agent that knows most birds fly (a statistical fact) should assign zero probability to all worlds in which this is not true. Doing so will raise its degree of belief $\mathbf{P}_W\{\mathit{flies}(\mathit{Tweety}) \mid \mathit{bird}(\mathit{Tweety})\}$, since the majority of bird objects that the constant symbol *Tweety* could refer to will also fly. Thus, this composition allows us to perform *direct inference*, i.e., to reason from statistical knowledge about populations to beliefs about individuals.

To make this possible, we follow Bacchus (1990) in extending the syntax of first-order logic so that we can make *statistical predications*, i.e., logical sentences that are true whenever certain statistical properties hold. We begin by including a new unit of syntax which is distinct from formulas and terms, called a *proportion expression*.[7] Rational numbers are proportion expressions, as are *statistical quantifications* of the form

$$[\,\phi\,]_{\{x_1,x_2,\ldots,x_k\}} \qquad \text{and} \qquad [\,\phi \mid \psi\,]_{\{x_1,x_2,\ldots,x_k\}}$$

where ϕ and ψ are formulas. Such expressions are called statistical quantifications because the variables $x_1, x_2, \ldots, x_k$ are bound within ϕ and ψ. Proportion expressions are closed under arithmetic, and may be combined with relational operators (e.g., $<$, $=$, etc.) to form statistical predications.

We augment the semantics of the logic by interpreting the arithmetic and relational operators in the usual way. Finally, we define the interpretation of $[\,\phi\,]_{\{x_1,x_2,\ldots,x_k\}}$ in the possible world w to be $\mathbf{U}_O^w\{\phi\}$, the fraction of instantiations of $x_1, x_2, \ldots, x_k$ that make ϕ true in w. Similarly, we define the interpretation of $[\,\phi \mid \psi\,]_{\{x_1,x_2,\ldots,x_k\}}$ in w to be $\mathbf{U}_O^w\{\phi \mid \psi\}$ (or zero if $\mathbf{U}_O^w\{\phi \mid \psi\}$ is undefined).

Returning to our first example, statement (5) can be encoded by the following statistical predication:

$$[\,\mathit{flies}(x) \mid \mathit{bird}(x)\,]_{\{x\}} = 0.9$$

In most worlds w, $\mathbf{U}_O^w\{\mathit{flies}(x) \mid \mathit{bird}(x)\}$ will not be exactly 0.9, in which case this sentence is false. (In fact, this can only be true in worlds where the number of objects satisfying $\mathit{bird}(x)$ is a multiple of 10.) The statement can be weakened by using an interval, e.g., $[\,\mathit{flies}(x) \mid \mathit{bird}(x)\,]_{\{x\}} \in [0.85, 0.95]$.

[6]Interestingly, sentence probabilities (Equation (1)) may be seen as a special case of formula probabilities (Equation (6)), since sentences are zero-arity formulas. Let α be a sentence and w a possible world. If $V(\alpha, w) = \mathrm{T}$, then $\mathcal{O}_w^\alpha = \{\langle\rangle\}$ is the set containing the empty tuple; if $V(\alpha, w) = \mathrm{F}$, then $\mathcal{O}_w^\alpha = \emptyset$.

[7]Unlike Bacchus (1990) and Halpern (1990), who define logics of probability, we do not treat proportion expressions as first-class objects; they are simply a new type of ground term.

Extending the logical language to admit statistical predications gives us two ways to inform our degrees of belief with statistical knowledge. In the first, we condition on statistical predications when calculating degrees of belief; for example, we can compute

$$\mathbf{P}_W\{\mathit{flies}(\mathit{Tweety}) \mid \mathit{bird}(\mathit{Tweety}),\; [\,\mathit{flies}(x) \mid \mathit{bird}(x)\,]_{\{x\}} > 0.9\},$$

the probability that Tweety flies given he is a bird and over 90% of birds fly. In the second, we include statistical predications in the probabilistic knowledge base used to construct the world model; they can be included as facts or as beliefs, allowing us to leverage certain and uncertain statistical knowledge in the construction of our world model.

Extending the Example

Let us add a unary predicate *dem* to our "marriage" language to indicate whether a person is a democrat or not. We now add the following fact and beliefs to our knowledge base:

$$\begin{array}{lll}
\alpha_6: & [\,\mathit{dem}(x) \leftrightarrow \mathit{dem}(y) \mid \mathit{married}(x,y)\,]_{\{x,y\}} & \geq 0.5 \\
\beta_3: & \mathbf{P}_W\{\mathit{dem}(\mathit{Chris})\} & = \; 0.7 \\
\beta_4: & \mathbf{P}_W\{[\,\mathit{male}(x) \mid \mathit{dem}(x)\,]_{\{x\}} \geq 0.5\} & = \; 0.9
\end{array}$$

That is: 50% or more marriages pair people with the same party affiliation; with 70% probability Chris is a democrat; and, with 90% probability, 50% or more democrats are male.

This extended knowledge base yields a maximum entropy world model with the parameters $\lambda_{\beta_1} = 1.27$, $\lambda_{\beta_2} = 1.05$, $\lambda_{\beta_3} = 2.65$ and $\lambda_{\beta_4} = 4.45$. Using it, we can calculate our new degree of belief $\mathbf{Q}_W\{\mathit{married}(\mathit{Alice}, \mathit{Chris})\}$ to be 0.4. (Chris is probably a democrat, which increases his chances of being male and therefore of being married to Alice.) We can also calculate the statistical quantity $\mathbf{Q}_O\{\exists y(\mathit{married}(x,y)) \mid \mathit{dem}(x)\}$ (the probability a democrat is married) to be 0.57.

Discussion

Our proposal raises a number of interesting and important issues, which we now briefly discuss.

Applying the principle of maximum entropy to random worlds provides several nice inferential properties, but also gives rise to some subtle issues. For example, encoding the same problem domain with different languages can lead to divergent predictions (Halpern & Koller 1995). Also, certain kinds of knowledge must be encoded with care to obtain the desired semantics, e.g., causal knowledge (Hunter 1989). Finally, evidence and knowledge behave differently under maximum entropy; i.e., conditioning on a sentence as evidence and including it as a fact in the knowledge base can yield different degrees of belief.[8] It is therefore important to

[8](I thank Andrew Ng for pointing this out.) As an example, let L be a propositional language with two symbols p and q, and let our probabilistic knowledge base consist of the belief that $\mathbf{P}_W\{p \wedge q\} = 0.4$. Then we find $\mathbf{Q}_W\{p \mid q\} = 2/3$ under the corresponding maximum entropy world model $\mathbf{Q}_W$. However, if we include q as a fact in this knowledge base, we then find $\mathbf{Q}'_W\{p\} = 0.4$.

distinguish between sentences that are observed to be true in a particular context, and those which are true in all contexts.

Another important issue is that of knowledge acquisition: we have have made no mention of how the probabilistic knowledge base we have assumed is to be obtained. In particular, the issue of obtaining statistical predications can be troublesome; we refer the reader to (Bacchus *et al.* 1996, Section 7.3) for a good discussion.

Finally, our proposal solves a knowledge representation problem, but we are left with a formidable computational problem: in general, it is intractable to compute sentence and formula probabilities. In fact, for just a propositional language, exact inference in this model is $\#P$-complete and approximate inference is NP-hard (Roth 1996). In practice, this intractability stems from the enormous size of the sample space $\mathcal{U}_L$; it is exponential in the size of the language, and doubly-exponential in the maximum domain size.

The situation here is much like that of probabilistic inference in graphical models: while inference in arbitrary graphical models is intractable, exact (or approximate) inference becomes tractable in models with sparse (or weak) dependence structure. As we have discussed, maximum entropy world models assume independence in the absence of explicit information to the contrary, and therefore can exhibit significant independence structure. In current work, we are examining how this independence structure can be leveraged to speed exact and approximate inference.

Related Work

Grove, Halpern, & Koller (1994) were the first to present a computational approach to reasoning from a knowledge base of statistical information; in the special case where the language consists only of unary predicates, they show that degrees of belief can be approximated by maximum entropy computations. Bacchus *et al.* (1994) extend their framework so that the knowledge base may include beliefs as well as statistical predications. Bacchus *et al.* (1996) discuss the problem of direct inference, and show that their framework exhibits several nice properties (many of which are shared by the current proposal). However, the formalism cannot be used with arbitrary first-order logic languages, as ours can.

There are now several works that extend logic programs to represent propositional uncertainty. Probabilistic Logic Programs (Lukasiewicz 1998) represent one such approach, and have been extended to make use of maximum entropy techniques when the world model is underdetermined (Lukasiewicz & Kern-Isberner 1999). Sato & Kameya (2001) present another extension of logic programs that expresses uncertainty regarding facts (but not rules) in the knowledge base, and whose parameters can be learned. Stochastic Logic Programs (Muggleton 2002) also extend logic programs to represent propositional uncertainty, but the underlying measure is constructed over resolution proofs rather than possible worlds.

Acknowledgements. I thank Stuart Russell, Kevin Murphy, David Andre, and Sekhar Tatikonda for comments on an earlier draft of this work; all remaining errors are my own. This research was supported by NSF grant IRI-96-34215.

References

Bacchus, F.; Grove, A. J.; Halpern, J. Y.; and Koller, D. 1994. Generating new beliefs from old. In de Mantaras, R. L., and Poole, D., eds., *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 37–45. San Francisco, CA: Morgan Kaufmann.

Bacchus, F.; Grove, A. J.; Halpern, J. Y.; and Koller, D. 1996. From statistical knowledge bases to degrees of belief. *Artificial Intelligence* 87(1-2):75–143.

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities*. Cambridge, MA.: MIT Press.

Cheeseman, P. 1983. A method of computing generalized bayesian probability values for expert systems. In Bundy, A., ed., *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 198–202. Los Altos, CA: William Kaufmann.

Darroch, J. N., and Ratcliff, D. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics* 43:1470–1480.

Della Pietra, S.; Della Pietra, V.; and Lafferty, J. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(4):380–393.

Grove, A. J.; Halpern, J. Y.; and Koller, D. 1994. Random worlds and maximum entropy. *Journal of Artificial Intelligence Research* 2:33–88.

Halpern, J. Y., and Koller, D. 1995. Representation dependence in probabilistic inference. In Mellish, C., ed., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1853–1860. San Francisco, CA: Morgan Kaufmann.

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

Hunter, D. 1989. Causality and maximum entropy updating. *International Journal of Approximate Reasoning* 3(1):87–114.

Jaynes, E. T. 1979. Where do we stand on maximum entropy? In Levine, R. D., and Tribus, M., eds., *The Maximum Entropy Formalism*. Cambridge, MA.: MIT Press. 15–118.

Lukasiewicz, T., and Kern-Isberner, G. 1999. Probabilistic logic programming under maximum entropy. In Hunter, A., and Parsons, S., eds., *Proceedings of the Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 279–292. London, UK: Springer.

Lukasiewicz, T. 1998. Probabilistic logic programming. In Prade, H., ed., *Proceedings of the Thirteenth European Conference on Artificial Intelligence*, 388–392. Brighton, UK: J. Wiley & Sons.

Muggleton, S. 2002. Stochastic logic programs. *Journal of Logic Programming*. Forthcoming.

Nilsson, N. 1986. Probabilistic logic. *Artificial Intelligence* 28(1):71–87.

Paris, J. B. 1999. Common sense and maximum entropy. *Synthese* 117(1):75–93.

Roth, D. 1996. On the hardness of approximate reasoning. *Artificial Intelligence* 82:273–302.

Sato, T., and Kameya, Y. 2001. Parameter learning of logic programs for symbolic-statistical modeling. *Journal of Artificial Intelligence Research* 15:391–454.