Representation Dependence in Probabilistic Inference *

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com

Daphne Koller
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
daphne@cs.berkeley.edu

Abstract

Non-deductive reasoning systems are often *representation dependent*: representing the same situation in two different ways may cause such a system to return two different answers. This is generally viewed as a significant problem. For example, the *principle of maximum entropy* has been subjected to much criticism due to its representation dependence. There has, however, been almost no work investigating representation dependence. In this paper, we formalize this notion and show that it is not a problem specific to maximum entropy. In fact, we show that any probabilistic inference system that sanctions certain important patterns of reasoning, such as a minimal default assumption of independence, must suffer from representation dependence. We then show that invariance under a restricted class of representation changes can form a reasonable compromise between representation independence and other desiderata.

1 Introduction

It is well known that the way a problem is represented can have a significant impact on the ease with which people solve it, and on the complexity of an algorithm for solving it. We are interested in what is arguably an even more fundamental issue: the extent to which the *answers* that we get depend on how our input is represented.

To make our discussion more concrete, we discuss this issue in one particular context: probabilistic inference. We focus on probabilistic inference both because of the recent interest in using probability for knowledge representation (e.g., [Pearl, 1988]) and because it has been the source of many of the concerns expressed regarding representation. However, our approach should be applicable far more generally.

Suppose we have a procedure for making inferences from a probabilistic knowledge base. How sensitive is it to the way knowledge is represented? Consider the following examples, which use perhaps the best-known non-deductive notion of probabilistic inference, *maximum entropy* [Jaynes, 1978].[1]

Example 1.1: Suppose we have no information whatsoever. What probability should we assign to the proposition *colorful*? Symmetry arguments might suggest $1/2$: Since we have no information, it seems that an object should be just as likely to be colorful as non-colorful. This is also the conclusion reached by maximum entropy. But now suppose we consider a more refined view of the world where we have colors, and by *colorful* we actually mean $\mathit{red} \vee \mathit{blue} \vee \mathit{green}$. In this case, maximum entropy dictates that the probability of $\mathit{red} \vee \mathit{blue} \vee \mathit{green}$ is $7/8$. Note that, in both cases, the only conclusion that follows from our constraints is the trivial one: that the probability of the query is somewhere between 0 and 1. ∎

Example 1.2: Suppose we are told that half of the birds fly. There are two reasonable ways to represent this information. One is to have propositions *bird* and *fly*, and use a knowledge base $KB_1^{\mathit{fly}} =_{\mathrm{def}} [\Pr(\mathit{fly} \mid \mathit{bird}) = 1/2]$. A second might be to have as basic predicates *bird* and *flying-bird*, and use a knowledge base $KB_2^{\mathit{fly}} =_{\mathrm{def}} [(\mathit{flying\text{-}bird} \Rightarrow \mathit{bird}) \wedge \Pr(\mathit{flying\text{-}bird} \mid \mathit{bird}) = 1/2]$. Although the first representation may appear more natural, it seems that both representations are intuitively equivalent insofar as representing the information that we have been given. But if we use an inference method such as maximum entropy, the first representation leads us to infer $\Pr(\mathit{bird}) = 1/2$, while the second leads us to infer $\Pr(\mathit{bird}) = 3/4$. ∎

Examples such as these are the basis for the frequent criticisms of maximum entropy on the grounds of representation dependence. But other than pointing out these examples, there has been little work on this problem. In fact, other than the work of Salmon [Salmon, 1961; Salmon, 1963], there seems to have been no work on formalizing the notion of representation dependence. One might say that the consensus was: "whatever representation independence is, it is not a property enjoyed by maximum entropy." But are there any other inference procedures that have it? In this paper we attempt to understand the notion of representation dependence, and to study the extent to which it is achievable.

*Research sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080, and by a University of California President's Postdoctoral Fellowship.

[1] Although much of our discussion is motivated by the representation dependence problem encountered by maximum entropy, an understanding of maximum entropy and how it works is not essential for understanding our discussion.

To study representation dependence, we must first understand what we mean by a "representation". The real world is complex. In any reasoning process, we must focus on certain details and ignore others. At a semantic level, the relevant distinctions are captured by using a space X of possible alternatives or states. In Example 1.1, our first representation focused on the single attribute *colorful*. In this case, we have only two states in the state space, corresponding to *colorful* being true and false, respectively. The second representation, using *red*, *blue*, and *green*, has a richer state space. Clearly, there are other distinctions that we could make. At a syntactic level, we often capture relevant distinctions using some formal language. For example, if we use propositional logic as our basic knowledge representation language, our choice of primitive propositions characterizes the distinctions that we have chosen to make. In this case, we can take the states to be truth assignments to these propositions. Similarly, if we use *belief networks* [Pearl, 1988] as our knowledge representation language, we must choose some set of relevant variables. The states are then then possible assignments of values to these variables.

What does it mean to shift from a representation (i.e., state space) X to another representation Y? For us, this amounts to associating subsets of X with subsets of Y. Thus, for example, the state where *colorful* holds can be associated with the set of states where either *red*, *blue*, or *green* holds. We capture the notion of representation shift formally by *embeddings*. An *embedding* f from X to Y maps subsets of X to subsets of Y in a way that preserves complementation and intersection. An embedding is just the semantic version of the standard logical notion of *interpretation* [Enderton, 1972, pp. 157–162], which has also been used in the recent literature on *abstraction* [Giunchiglia and Walsh, 1992; Nayak and Levy, 1994]. Essentially, an interpretation maps formulas in a vocabulary $\mathbf{\Phi}$ to formulas in a different vocabulary $\mathbf{\Psi}$ by mapping the primitive propositions in $\mathbf{\Phi}$ (e.g., *colorful*) to formulas over $\mathbf{\Psi}$ (e.g., *red* $\vee$ *blue* $\vee$ *green*) and then extending to complex formulas in the obvious way. The representation shift in Example 1.2 can also be captured in terms of an interpretation, this one taking *flying-bird* to *fly* $\wedge$ *bird*.

When doing probabilistic reasoning, we are actually interested in the probability of the various states in X. We therefore assume that a user's knowledge base KB consists of a set of probabilistic assertions, such as Pr(*fly* | *bird*) $= 1/2$, that place constraints on distributions over X. A representation shift from X to Y induces a corresponding shift >from information about X to information about Y. More formally, an embedding f from X to Y can be extended to a mapping f^* from constraints on distributions over X to constraints over distributions over Y in a straightforward way. For example, if the embedding f maps *colorful* to *red* $\vee$ *blue* $\vee$ *green*, then f^*(Pr(*colorful*) $\geq 2/3$) is Pr(*red* $\vee$ *blue* $\vee$ *green*) $\geq 2/3$.

A *probabilistic inference procedure* $\mid\!\sim$ takes a probabilistic knowledge base and uses it to reach conclusions about the probability of various events over the space X. Such a procedure is said to be *invariant under* f if f does not change the conclusions that we make; that is, if for any KB and θ, $KB \mid\!\sim \theta$ iff $f(KB) \mid\!\sim f(\theta)$. Roughly speaking, $\mid\!\sim$ is *representation independent* if it is invariant under all embeddings. This captures the intuition that $\mid\!\sim$ gives us the same answers no matter how we shift representations. Of course, not all embeddings count as legitimate representation shifts. For example, consider an embedding f defined in terms of an interpretation that maps both the propositions p and q to the proposition r. Then the process of changing representations using f gives us the information that p and q are equivalent, information that we might not have had originally. Intuitively, f gives us new information if it tells us that certain situations—e.g., those where $p \wedge \neg q$ holds—are not possible. Formally, f is said to be *faithful* if $f(\{x\}) \neq \emptyset$ for all $x \in X$. We show that f is faithful if and only if for any KB and θ, the assertion θ necessarily follows from KB if and only if $f(\theta)$ follows from $f(KB)$. That is, faithful embeddings are precisely those that give us no new information.

At first glance, representation independence seems like a reasonable desideratum. However, as we show in Section 3, it has some rather unfortunate consequences. In particular, we show that any representation independent inference procedure must act essentially like logical entailment for a knowledge base with only non-probabilistic information. In fact, if we also require that our inference procedure ignore blatantly irrelevant information, then it must act like logical entailment for *every* knowledge base. Finally, representation independence is completely incompatible with even the simplest default assumption of independence: Even if we are told nothing about the basic propositions p and q, representation independence does not allow us to jump to the conclusion that p and q are independent.

This seems to put us in a rather awkward situation: It seems we must either give up on representation independence, or make do with an inference procedure that is essentially incapable of even minimal inductive reasoning (jumping to conclusions). But things are not quite as bleak as they seem. In practice, we would claim that the choice of language does carry a great deal of information. That information is what gives us the intuition that certain embeddings are legitimate, while others that have the same abstract structure are not. For example, suppose that certain propositions represent colors while others represent birds. While we may be willing to transform *colorful* to *red* $\vee$ *blue* $\vee$ *green*, we may not be willing to transform *red* to *fly*. There is no reason to demand that an inference procedure behave the same way if we suddenly shift to a wildly inappropriate representation, where the symbols mean something completely different. Given a class of "appropriate" embeddings (where the notion of appropriate might be application-dependent), we may well be able to get an interesting notion of representation independence with respect to that class.

In Section 5, we provide a general approach to constructing inference procedures that are invariant under a specific class of embeddings. We assume that the user starts with some set of initial *prior probability distributions* that characterize his beliefs in the absence of information. We show that if the prior distributions are chosen appropriately, so that they are invariant under the class of embeddings of interest, then we can "bootstrap" up to obtain a general inference procedure that is invariant under the same class of embeddings, by using *cross-entropy* [Kullback and Leibler, 1951], a well-known generalization of probabilistic conditioning. This result can be used in a number of ways. For example, it shows us how to construct an inference procedure that is invariant under a given set of embeddings: we simply choose a class of priors appropriately. It also allows us to combine some degree of representation independence with certain non-deductive properties that we want of our inference procedure. We simply choose a class of priors that has the desired property, and determine the set of embeddings under which this class of priors is invariant. We demonstrate this process by presenting an inference method that supports a default assumption of independence, and yet is invariant under a natural class of embeddings.

2 Probabilistic Inference

We begin by defining probabilistic inference procedures. Such a procedure takes as input a probabilistic knowledge base and returns a probabilistic conclusion. We express this in a semantic framework that we hope can be understood by both logicians and probabilists.

We take both the knowledge base and the conclusion to be probabilistic assertions about the probabilities of events over some state space X. Formally, these can be viewed as statements (or constraints) on distributions over X. For example, if A is a subset of X, a statement $\Pr(A) \geq 2/3$ holds only for distributions where A has probability at least $2/3$. Therefore, if Δ_X is the set of all probability distributions on X, we can view a knowledge base as a set of constraints over Δ_X. We place very few restrictions on the language used to express the constraints. All that we require is that the language is closed under conjunction and negation, so that if KB and KB' are knowledge bases expressing constraints, then so are $KB \wedge KB'$ and $\neg KB$. Given a knowledge base KB placing constraints on Δ_X, we write $\mu \models KB$ if μ is a distribution in Δ_X that satisfies the constraints in KB, and we let $[\![KB]\!]_X$ denote all the distributions satisfying these constraints. We say that KB is *consistent* if $[\![KB]\!]_X \neq \emptyset$, i.e., if the constraints are satisfiable. Finally, we say that KB *entails* θ (where θ is another set of constraints on Δ_X), written $KB \models_X \theta$, if $[\![KB]\!]_X \subseteq [\![\theta]\!]_X$, i.e., if every distribution that satisfies KB also satisfies θ. We write $\models_X \theta$ if θ is satisfied by every distribution in Δ_X. We omit the subscript X from $\models$ if it is clear from context.

Entailment is well-known to be to be a very weak method of drawing conclusions from a knowledge base. For example, it is unable to ignore irrelevant information. Consider the knowledge base $\Pr(\mathit{fly} \mid \mathit{bird}) \geq 0.9$. Even though we know nothing to suggest that *red* is at all relevant, entailment will not allow us to reach any nontrivial conclusion about $\Pr(\mathit{fly} \mid \mathit{bird} \wedge \mathit{red})$. One way to get more powerful conclusions is to consider, not all the distributions that satisfy KB, but a subset of them. Intuitively, given a knowledge base KB, an inference procedure picks a subset of the distributions satisfying KB, and infers θ if θ holds in this subset.

Definition 2.1: An *X-inference procedure* is a function $I_X : 2^{\Delta_X} \mapsto 2^{\Delta_X}$ such that $I_X(A) \subseteq A$ for $A \subseteq \Delta_X$ and $I_X(A) = \emptyset$ iff $A = \emptyset$. We write $KB \mathrel{|\!\!\sim}_{I_X} \theta$ if $I_X([\![KB]\!]_X) \subseteq [\![\theta]\!]_X$. ∎

We are typically interested in X-inference procedures not just for one space X, but for a family $\mathcal{X}$ of spaces. In cases where $X \in \mathcal{X}$ is clear from context, we write $KB \mathrel{|\!\!\sim}_I \theta$, omitting the subscript X. Clearly entailment is an X-inference procedure for any X, where I_X is simply the identity function. Another well-known inference procedure is *maximum entropy*, which picks out of $[\![KB]\!]_X$ the subset of the distributions having the maximum entropy.

Example 2.2: Given a distribution μ over a finite space X, its *entropy* $H(\mu)$ is defined as $-\sum_{x \in X} \mu(x) \log \mu(x)$. Given a set A of distributions in Δ_X, let $I_X^{me}(A)$ consist of the distributions in A that have the highest entropy. I_X^{me} clearly defines an inference procedure, which we denote $\mathrel{|\!\!\sim}_{me}$. Thus, $KB \mathrel{|\!\!\sim}_{me} \theta$ if θ holds in all the distributions of maximum entropy satisfying KB. ∎

There are, of course, many other inference procedures. In fact, as the following proposition shows, any binary relation $\mathrel{|\!\!\sim}$ satisfying certain reasonable properties is an inference procedure of this type.

Proposition 2.3: *I is an X-inference procedure if and only if the following properties hold for every KB, KB', φ, ψ over X:*

- Reflexivity: $KB \mathrel{|\!\!\sim}_I KB$.
- Left Logical Equivalence: *if KB is logically equivalent to KB', i.e., if $\models KB \Leftrightarrow KB'$, then $KB \mathrel{|\!\!\sim}_I \theta$ iff $KB' \mathrel{|\!\!\sim}_I \theta$.*
- Right Weakening: *if $KB \mathrel{|\!\!\sim}_I \theta$ and $\models \theta \Rightarrow \psi$ then $KB \mathrel{|\!\!\sim}_I \psi$.*
- And: *if $KB \mathrel{|\!\!\sim}_I \theta$ and $KB \mathrel{|\!\!\sim}_I \psi$, then $KB \mathrel{|\!\!\sim}_I \theta \wedge \psi$.*
- Consistency: *if KB is consistent then $KB \not\mathrel{|\!\!\sim}_I$ false.* ∎

Interestingly, these properties are commonly viewed as part of a core of reasonable properties for a nonmonotonic inference relation [Kraus *et al.*, 1990].

Although our basic framework puts no constraints on the state space X and very few constraints on the language used to describe constraints, in practice, we often use a logical language to describe the possible states, and this language then determines the state space. Typical languages include propositional logic, first-order logic, or a language describing the values for some set of random variables. In general, a base logic $\mathcal{L}$ defines a set of formulas $\mathcal{L}(\Phi)$ for a given vocabulary Φ. In propositional logic, the vocabulary Φ is simply a set of propositional symbols. In probability theory, the vocabulary can consist of a set of random variables. In first-order logic, the vocabulary is a set of constant symbols, function symbols, and predicate symbols. For simplicity, we assume that for each base logic all the vocabularies are finite subsets of one fixed infinite vocabulary Φ^*. Each state in our state space defines an interpretation for the symbols in Φ. Hence, in the case of propositional logic, a state can be viewed as a model over Φ: a truth assignment to the primitive propositions; for first-order logic, a state consists of a domain and an interpretation of the symbols in Φ; in the probabilistic setting, a state is an assignment of values to the random variables. In this case, we often take our state space to be $\mathcal{W}(\Phi)$, the set of all models (or assignments) over the vocabulary Φ, or some subset of it. Note that the truth of any formula φ in $\mathcal{L}(\Phi)$ is determined by a state. If φ is true in some state w, we write $w \models \varphi$. We call formulas in the base language *objective*. Objective formulas can be thought of as describing events.

The *probabilistic extension* $\mathcal{L}^{pr}(\Phi)$ of a base logic $\mathcal{L}(\Phi)$ is simply the set of probability formulas over $\mathcal{L}(\Phi)$. Formally, for each $\varphi \in \mathcal{L}(\Phi)$, $\Pr(\varphi)$ is a numeric term. The formulas in $\mathcal{L}^{pr}(\Phi)$ are defined to be all the Boolean combinations of arithmetic expressions involving numeric terms. For example, $\Pr(\mathit{fly} \mid \mathit{bird}) \geq 1/2$ is a formula in $\mathcal{L}^{pr}(\{\mathit{fly}, \mathit{bird}\})$ (where we interpret a conditional probability expression $\Pr(\varphi \mid \psi)$ as $\Pr(\varphi \wedge \psi) / \Pr(\psi)$ and then multiply to clear the denominator). We ascribe semantics to $\mathcal{L}^{pr}(\Phi)$ via a probability distribution μ over $\mathcal{W}(\Phi)$. We interpret the numeric term $\Pr(\varphi)$ as $\mu(\{w \in \mathcal{W}(\Phi) : w \models \varphi\})$. Since an objective formula φ describes an event over the space X, a formula θ in $\mathcal{L}^{pr}(\Phi)$ is clearly a constraint on distributions over X. We write $\mu \models \theta$ if the distribution μ satisfies the formula θ. Note that $\mathcal{L}^{pr}$ does not contain objective formulas; we interpret the notation $\mu \models \varphi$ for an objective formula φ as an abbreviation for $\mu \models \Pr(\varphi) = 1$. Thus, an objective formula corresponds to a constraint of a special form, that says a particular event is certain.

3 Representation independence

As we discussed in the introduction, we capture the idea of a representation shift using the notion of an embedding. If X and Y are two different representations, then an embedding from X to Y reflects the correspondence of the states in X and the states in Y. This embedding should respect the logical structure of events. Formally, we require that it be a homomorphism with respect to conjunction and negation.

Definition 3.1: An *embedding* f from X to Y is a function $f : 2^X \mapsto 2^Y$ such that $f(A \cap B) = f(A) \cap f(B)$ and $f(\overline{A}) = \overline{f(A)}$ for all $A, B \subseteq X$. ∎

Clearly, an embedding f induces a map $f^* : 2^{\Delta_X} \mapsto 2^{\Delta_Y}$ defined as follows: $f^*(\mu) = \{\nu \in \Delta_Y : \nu(f(S)) = \mu(S)$ for all $S \subseteq X\}$ and $f^*(A) = \cup_{\mu \in A} f^*(\mu)$.[2]

Example 3.2 : In Example 1.1, we might have $X = \{\mathit{colorful}, \overline{\mathit{colorful}}\}$ and $Y = \{\mathit{red}, \mathit{blue}, \mathit{green}, \overline{\mathit{colorful}}\}$. In this case, we might have $f(\mathit{colorful}) = \{\mathit{red}, \mathit{blue}, \mathit{green}\}$ and $f(\overline{\mathit{colorful}}) = \{\overline{\mathit{colorful}}\}$. Consider the distribution $\mu \in \Delta_X$ such that $\mu(\mathit{colorful}) = 0.7$ and $\mu(\overline{\mathit{colorful}}) = 0.3$. Then $f^*(\mu)$ is the set of distributions ν such that the total probability assigned to the set of states $\{\mathit{red}, \mathit{blue}, \mathit{green}\}$ by ν is 0.7. Note that there are uncountably many such distributions. ∎

Embeddings can be viewed as the semantic analogue to the notion of interpretation defined in [Enderton, 1972].

Definition 3.3: Let Φ and Ψ be two vocabularies. In the propositional case, a *interpretation of* Φ *into* Ψ is a function i that associates with every propositional symbol $p \in \Phi$ a formula $i(p) \in \mathcal{L}(\Psi)$. A more complex definition in the same spirit applies to first-order vocabularies. For example, if R is a k-ary predicate, then $i(R)$ is a formula with k free variables. ∎

Given an interpretation i, we get a syntactic translation from formulas in $\mathcal{L}(\Phi)$ to formulas in $\mathcal{L}(\Psi)$ using i in the obvious way; for example, $i((p \wedge \neg q) \vee r) = (i(p) \wedge \neg i(q)) \vee i(r)$ (see [Enderton, 1972] for the details). Clearly an interpretation i from Φ to Ψ induces an embedding from $\mathcal{W}(\Phi)$ to $\mathcal{W}(\Psi)$ (at least, from the sets of models definable by formulas): we map $[\![\varphi]\!]_\Phi$—the set of states in $\mathcal{W}(\Phi)$ satisfying φ—to $[\![i(\varphi)]\!]_\Psi$—the set of states in $\mathcal{W}(\Psi)$ satisfying $i(\varphi)$.

If the embedding f is a "reasonable" representation shift, we would like an inference procedures to return the same answers if we shift representations using f.

Definition 3.4: We say that I is *invariant* under f if for all constraints KB and θ, we have $\mathit{KB} \mathrel{|\!\!\sim}_I \theta$ iff $f^*(\mathit{KB}) \mathrel{|\!\!\sim}_I f^*(\theta)$. ∎

As we argued in the introduction, not every embedding is appropriate as a representation shift. The process of changing representation should not give us any new information. When does a shift give us new information? One obvious situation is when the shift makes impossible something which we considered to be possible. In our example from the introduction, we had an interpretation i with the property that $i(p) = i(q)$ for two propositions p and q. Clearly, this interpretation gives us new information: that p and q are equivalent. Semantically, the associated embedding f has the following undesirable property: it maps the set of states satisfying $p \wedge \neg q$ to the empty set. This means a state where $p \wedge \neg q$ holds does not have an analogue in the new representation. We want to disallow such embeddings.

Definition 3.5: We say that an embedding $f : X \mapsto Y$ is *faithful* if for any $x \in X$, $f(x) \neq \emptyset$. ∎

This has the desired consequence of not giving us new information:

Lemma 3.6: *An embedding $f : X \mapsto Y$ is faithful if and only if for all constraints KB and θ, we have KB $\models \theta$ iff $f^*(\mathit{KB}) \models f^*(\theta)$.*

It is clear that our embedding from Example 3.2 is faithful: $f(\mathit{colorful}) = \{\mathit{red}, \mathit{blue}, \mathit{green}\}$ and $f(\overline{\mathit{colorful}}) = \overline{\mathit{colorful}}$. Therefore, an inference procedure which is invariant under all faithful embeddings would return the same answers for $\Pr(\mathit{colorful})$ as for $\Pr(\mathit{red} \vee \mathit{blue} \vee \mathit{green})$.

The issue is somewhat more subtle for Example 1.2. There, we would like to have an embedding f generated by the interpretation $i(\mathit{flying\text{-}bird}) = \mathit{fly} \wedge \mathit{bird}$ and $i(\mathit{bird}) = \mathit{bird}$. This is not a faithful embedding, since $\mathit{flying\text{-}bird} \Rightarrow \mathit{bird}$ is not a valid formula, while $i(\mathit{flying\text{-}bird} \Rightarrow \mathit{bird})$ is $(\mathit{fly} \wedge \mathit{bird}) \Rightarrow \mathit{bird}$ which is valid. Looking at this problem semantically, we see that the state corresponding to the model where $\mathit{flying\text{-}bird} \wedge \neg \mathit{bird}$ holds is mapped to $\emptyset$. But this is clearly the source of the problem. According to our linguistic intuitions for this domain, this is not a "legitimate" state. Rather than considering all the states in $\mathcal{W}(\{\mathit{flying\text{-}bird}, \mathit{bird}\})$, it is perhaps more appropriate to consider the subset X consisting of the truth assignments characterized by the formulas $\{\mathit{flying\text{-}bird} \wedge \mathit{bird}, \neg \mathit{flying\text{-}bird} \wedge \mathit{bird}, \neg \mathit{flying\text{-}bird} \wedge \neg \mathit{bird}\}$. If we now use i to embed X into $\mathcal{W}(\{\mathit{fly}, \mathit{bird}\})$, the resulting embedding is indeed faithful. So, as for the previous example, invariance under this embedding would guarantee that we get the same answers under both representations.

Definition 3.7: We say that $\mathrel{|\!\!\sim}_I$ is *representation independent* if $\mathrel{|\!\!\sim}_I$ is invariant under all faithful embeddings. ∎

Are there any representation independent inference procedures? It follows trivially from Lemma 3.6 that entailment is representation independent. Are there others? As we now show, any such inference procedures are unlikely to be interesting.

We say that an inference procedure I is *essentially entailment* for KB if $\mathit{KB} \mathrel{|\!\!\sim}_I \alpha < \Pr(\varphi) < \beta$ implies $\mathit{KB} \models \alpha \leq \Pr(\varphi) \leq \beta$. Thus, when entailment lets us conclude $\Pr(\varphi) \in [\alpha, \beta]$, an inference procedure that is essentially entailment lets us draw only the slightly stronger conclusion $\Pr(\varphi) \in (\alpha, \beta)$.

[2] Strictly speaking, we should write $f(\{\mu\})$, not $f(\mu)$. However, here and elsewhere we omit set braces around function arguments that are singleton sets.

Theorem 3.8: *Every representation independent inference procedure is essentially entailment for every objective KB.*

This result tells us that from an objective knowledge base KB—one asserting only that certain events are known to be true—we can reach only three possible conclusions about another event φ. If φ is entailed by the KB, we conclude $\Pr(\varphi) = 1$; if $\neg\varphi$ is entailed by the KB, we conclude $\Pr(\varphi) = 0$; and if both φ and $\neg\varphi$ are consistent with the KB, the *strongest* conclusion we can make about $\Pr(\varphi)$ is that it is somewhere between 0 and 1.

This result implies that various inference procedures cannot be representation independent. In particular, since *true* $\mid\!\sim_{me}$ $\Pr(p) = 1/2$ for a primitive proposition p, it follows that $\mid\!\sim_{me}$ is not essentially entailment. It follows that maximum entropy is not representation independent.

It is consistent with this theorem that there are representation independent inference procedures that are not almost entailment for probabilistic knowledge bases. For example, as we show in the full paper, there is a representation independent inference procedure I such that $(\Pr(p) > 1/4) \mid\!\sim_I \Pr(p) > 1/2$. However, as we now show, such an inference procedure is unlikely to be an interesting one. In particular, we show that if adding "irrelevant" information to the knowledge base does not affect our inferences, then every representation independent inference procedure is essentially entailment. We define "irrelevant" syntactically here; in the full paper we give the semantic analogue of the definitions. Syntactically, "irrelevant" means "in a disjoint vocabulary".

Definition 3.9: We say that $\mid\!\sim_I$ enforces *minimal irrelevance* if, whenever Φ and Ψ are disjoint vocabularies, $KB, \theta \in \mathcal{L}^{pr}(\Phi)$ and $KB' \in \mathcal{L}^{pr}(\Psi)$, then $KB \mid\!\sim_I \theta$ iff $KB \wedge KB' \mid\!\sim_I \theta$. ∎

Minimal irrelevance certainly seems like an innocuous and reasonable property. Adding information about symbols that do not appear in either KB or θ should not affect whether we can infer θ from KB. Logical entailment, being a monotonic inference procedure, clearly enforces minimal irrelevance. It is not hard to show that maximum entropy does too (see [Paris, 1994] for a proof). Unfortunately, minimal irrelevance combined with representation independence forces us to inference procedures that are essentially entailment.

Theorem 3.10: *Any representation independent inference procedure that enforces minimal irrelevance is essentially entailment.*

In the full paper, we show that other desirable properties are completely inconsistent with representation independence. As one example, we show that representation independence is inconsistent with a default assumption of independence. Again, we define minimal default independence syntactically here, deferring a semantic definition to the full paper.

Definition 3.11: We say that $\mid\!\sim_I$ enforces *minimal default independence* if, whenever Φ and Ψ are disjoint vocabularies, $KB \in \mathcal{L}^{pr}(\Phi)$, $\varphi \in \mathcal{L}(\Phi)$, and $\psi \in \mathcal{L}(\Psi)$, then $KB \mid\!\sim_I \varphi$ iff $KB \mid\!\sim_I \Pr(\varphi|\psi) = \Pr(\varphi)$. ∎

Clearly entailment does not satisfy minimal default independence. Maximum entropy, however, does. Indeed, a semantic property that implies minimal default independence is used in [Shore and Johnson, 1980] as one of the axioms in an axiomatic characterization of maximum-entropy.

Theorem 3.12: *Any inference procedure that enforces minimal default independence cannot be representation independent.*

4 Discussion

These results suggest that any type of representation independence is hard to come by. They also raise the concern that perhaps our definitions were not quite right. We can provide what seems to be even more support for the latter point.

Example 4.1: Let P be a unary predicate and $c_1, \ldots, c_{100}, d$ be constant symbols. Suppose that we have two vocabularies $\Phi = \{P, d\}$ and $\Psi = \{P, c_1, \ldots, c_{100}, d\}$. Consider the interpretation i from Φ to Ψ for which $i(d) = d$ and $i(P(x)) = P(x) \wedge P(c_1) \wedge \ldots P(c_{100})$. It is fairly straightforward to verify that the embedding f corresponding to i is faithful. Intuitively, since all the c_i's may refer to the same domain element, the only conclusion we can make with certainty from $P(c_1) \wedge \ldots \wedge P(c_{100})$ is that there exists at least one P in the domain. But we can draw this conclusion from $f^*(KB)$ only if $P(x)$ appears positively in *KB*, in which case we already know that there is at least one P. But it does not seem unreasonable that an inference procedure should assign different degrees of belief to $P(d)$ given $\exists x P(x)$ on the one hand and $\exists x(P(x) \wedge P(c_1) \wedge \ldots \wedge P(c_{100}))$ on the other,[3] particularly if the domain is small. In fact, many inductive reasoning systems explicitly adopt a *unique names assumption*, which would clearly force different conclusions in these two situations. ∎

This example suggests that, at least in the first-order case, even faithful embeddings do not always match out our intuition for a "reasonable" representation shift. One might therefore think that perhaps the problem is with our definition even in the propositional case. Maybe there is a totally different definition of representation independence that avoids these problems. While this is possible, we do not believe it to be the case. The techniques we used to prove Theorem 3.10 and 3.12 seem to apply to any reasonable notion of representation independence.[4] To give the flavor of the type of arguments used to prove these theorems, consider Example 1.1, and assume that *true* $\mid\!\sim_I \Pr(\textit{colorful}) = \alpha$ for $\alpha \in (0,1)$.[5] Using an embedding g such that $g(\textit{colorful}) = \textit{red}$, we conclude that *true* $\mid\!\sim_I \Pr(\textit{red}) = \alpha$. Similarly, we can conclude $\Pr(\textit{blue}) = \alpha$ and $\Pr(\textit{green}) = \alpha$. But in order for $\mid\!\sim_I$ to be invariant under our original embedding, we must have *true* $\mid\!\sim_I \Pr(\textit{red} \vee \textit{blue} \vee \textit{green}) = \alpha$, which is completely inconsistent with our previous conclusions. But the embeddings we use in this argument are very natural ones; we would not *want* a definition of representation independence that disallowed them.

But if we cannot avoid our negative results by moving to a better definition, where does that leave us? One approach is to declare that representation dependence is justified; the choice of an appropriate representation is indeed a significant one, which does encode some of the information at our disposal. In

[3] Actually, $i(P(d)) = P(d) \wedge P(c_1) \wedge \ldots P(c_{100})$, but the latter is equivalent to $P(d)$ given our knowledge base.

[4] They applied to all of the many definitions that we tried.

[5] In fact, it suffices to assume that *true* $\mid\!\sim_I \Pr(\textit{colorful}) \in [\alpha, \beta]$, as long as $\alpha > 0$ or $\beta < 1$.

particular, it encodes the bias of the knowledge-base designer about the world. Researchers in machine learning have long realized that bias is an inevitable component of effective inductive reasoning. So we should not be completely surprised if it turns out that other types of leaping to conclusions (as in our context) also depend on the bias.

Bias and representation independence are two extremes in a spectrum. If we accept that the knowledge base encodes the user's bias, there is no obligation to be invariant under any representation shifts at all. On the other hand, if we assume the representation used carries no information, coherence requires that our inference procedure give the same answers for all "equivalent" representations. We believe that the right answer lies somewhere in between. There are typically a number of reasonable ways in which we can represent our information, and we might want our inference procedure to return the same conclusions no matter which of these we choose. It thus makes sense to require that our inference procedure be invariant under embeddings that take us from one reasonable representation to another. But it does not follow that it must be invariant under *all* embeddings, or even all embeddings that are syntactically similar to the ones we wish to allow. We may be willing to refine *colorful* to *red* $\vee$ *blue* $\vee$ *green* or to define *flying-bird* as *fly* $\wedge$ *bird*, but not to transform *red* to *fly*. In the next section, we show how to construct inference procedures that are representation independent under a limited class of representation shifts.

5 Selective invariance

As discussed above, we want to construct an inference procedure I that is invariant only under certain embeddings. In order to do this, it is important to understand the conditions under which I is invariant under a specific embedding f from X to Y.

When do we conclude θ from $KB \subseteq \Delta_X$? Recall that an inference procedure I picks a subset $\mathcal{D}_X = I_X(KB)$, and concludes θ iff θ holds for every distribution in $\mathcal{D}_X$. Similarly, when applied to $f^*(KB) \subseteq \Delta_Y$, I picks a subset $\mathcal{D}_Y = I_Y(f^*(KB))$. For I to be invariant under f with respect to KB, there has to be a tight connection between $\mathcal{D}_X$ and $\mathcal{D}_Y$. To understand this connection, first consider a pair of distributions μ over X and ν over Y. When do these distributions generate the same probabilities for corresponding events? That is, when do we have that $\mu \models \theta$ iff $\nu \models f^*(\theta)$ for every θ? Clearly, this happens if and only if $\mu(A) = \nu(f(A))$ for any event $A \subseteq X$, i.e., if $\nu \in f^*(\mu)$. In this case, we say that μ and ν *correspond*. In order to understand how to apply this idea to sets $\mathcal{D}_X$ and $\mathcal{D}_Y$, consider the following example:

Example 5.1: Consider our embedding f of Example 3.2, and let $\mathcal{D}_X = \{\mu, \mu'\}$ where $\mu(\textit{colorful}) = 0.7$ as in Example 3.2, while $\mu'(\textit{colorful}) = 0.6$. How do we guarantee that we reach the corresponding conclusions from $\mathcal{D}_X$ and $\mathcal{D}_Y$? Assume, for example, that $\mathcal{D}_Y$ contains some distribution ν that does not correspond to either μ or μ', e.g., the distribution that assigns probability $1/4$ to all four states. In this case, the conclusion $\Pr(\textit{colorful}) \leq 0.7$ holds in $\mathcal{D}_X$, because it holds for both these distributions; but the corresponding conclusion $\Pr(\textit{red} \vee \textit{blue} \vee \textit{green}) \leq 0.7$ does not hold in $\mathcal{D}_Y$. Therefore, every distribution in $\mathcal{D}_Y$ must correspond to some distribution in $\mathcal{D}_X$. Conversely, every distribution in $\mathcal{D}_X$ must correspond to a distribution in $\mathcal{D}_Y$. For suppose that there is no distribution $\nu \in \mathcal{D}_Y$ corresponding to μ. Then we get the conclusion $\Pr(\textit{blue} \vee \textit{red} \vee \textit{green}) \neq 0.7$ from $\mathcal{D}_Y$, but the corresponding conclusion $\Pr(\textit{colorful}) \neq 0.7$ does not follow from $\mathcal{D}_X$. Note that these two conditions do *not* imply that $\mathcal{D}_Y$ must be precisely the set of distributions corresponding to distributions in $\mathcal{D}_X$. In particular, we might have $\mathcal{D}_Y$ containing only a single distribution ν corresponding to μ (and at least one corresponding to μ'), e.g., one with $\nu(\textit{red}) = 0.5$, $\nu(\textit{blue}) = 0$, $\nu(\textit{green}) = 0.2$, and $\nu(\overline{\textit{colorful}}) = 0.3$. ∎

We say that $\mathcal{D}_X$ and $\mathcal{D}_Y$ *correspond* under f if for all for all $\nu \in \mathcal{D}_Y$ there exists a corresponding $\mu \in \mathcal{D}_X$, and for all $\mu \in \mathcal{D}_X$, there exists a corresponding distribution $\nu \in \mathcal{D}_Y$.

Proposition 5.2: *Let f be an embedding of X into Y, and consider $\mathcal{D}_X \subseteq \Delta_X$ and $\mathcal{D}_Y \subseteq \Delta_Y$. Then $\mathcal{D}_X \models \theta$ iff $\mathcal{D}_Y \models f^*(\theta)$ for all θ exactly when $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f.*

We say that I is *invariant under f with respect to KB* if $KB \mathrel{|\!\!\sim}_I \theta$ iff $f^*(KB) \mathrel{|\!\!\sim}_I f^*(\theta)$ for all constraints θ. By definition, I is invariant under f iff it is invariant under f with respect to every KB. By Proposition 5.2, in order for I to be invariant under f, we must have a correspondence between $I_X(KB)$ and $I_Y(f^*(KB))$, for each KB. At first glance, it seems rather difficult to guarantee correspondence for every knowledge base. It turns out that the situation is not that bad. In this section, we show how, starting with a correspondence for the knowledge base *true*—that is, starting with a correspondence between $I_X(\Delta_X)$ and $I_Y(\Delta_Y)$—we can bootstrap to a correspondence for all KB's, using standard probabilistic updating procedures.

Consider first the problem of updating with objective information. The standard way of doing this update is via *conditioning*. For a distribution $\mu \in \Delta_X$ and an event $B \subseteq X$, define $\mu|B$ to be the distribution that assigns probability $\mu(w)/\mu(B)$ to every $w \in B$, and zero to all other states. For a set of distributions $\mathcal{D}_X \subseteq \Delta_X$, define $\mathcal{D}_X|B$ to be $\{\mu|B : \mu \in \mathcal{D}_X\}$.

Proposition 5.3: *Let $B \subseteq X$ be an event. If $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f, then $D_X|B$ and $D_Y|f(B)$ also correspond under f.*

What if we want to update on a constraint which is not objective? The standard extension of conditioning to this case is via *cross-entropy* [Kullback and Leibler, 1951].

Definition 5.4: The *cross-entropy of μ' relative to μ*, denoted $C(\mu', \mu)$, is defined as $\sum_{w \in \mathcal{W}} \mu'(w) \log(\mu'(w)/\mu(w))$. For a distribution μ over X and a constraint θ, let $\mu|\theta$ denote the set of distributions μ' satisfying θ for which $C(\mu', \mu)$ is minimal. ∎

Intuitively, $C(\mu', \mu)$ measures the "distance" from μ to μ'. The distribution μ' satisfying θ for which $C(\mu', \mu)$ is minimal can be thought of as the "closest" distribution to μ that satisfies θ. If θ denotes an objective constraint, then the unique distribution satisfying θ for which $C(\mu', \mu)$ is minimal is the conditional distribution $\mu|\theta$. That is why we have deliberately used the same notation here as for conditioning. We also define $D_X|\theta$ as we did for conditioning.

We can now apply a well-known result (see, e.g., [Seidenfeld, 1987]) to generalize Proposition 5.3 to the case of cross-entropy.

Theorem 5.5: *Let θ be an arbitrary constraint over Δ_X. If $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f, then $D_X|B$ and $D_Y|f(B)$ also correspond under f.*

As we now show, Theorem 5.5 gives us a way to "bootstrap" invariance. We construct an inference procedure that uses cross-entropy starting from some set of *prior probability distributions*. Intuitively, these encode the user's prior beliefs about the domain. As information comes in, these distributions are updated using cross-entropy. If we design our priors so that certain invariances hold, Theorem 5.5 guarantees that these invariances continue to hold throughout the process.

Formally, a *prior function* $\mathcal{P}$ takes a space X and returns a set of probability distributions in Δ_X. We now define an inference procedure $I^{\mathcal{P}}$ as follows. We define $I_X^{\mathcal{P}}(KB) = \{\mu|KB : \mu \in \mathcal{P}(X)\}$. Note that $I_X^{\mathcal{P}}(\mathit{true}) = \mathcal{P}(X)$, so that when we have no constraints at all, we use $\mathcal{P}(X)$ as the basis for our inference. Most of the standard inference procedures are of the form $I^{\mathcal{P}}$ for some prior function $\mathcal{P}$. It is fairly straightforward to verify, for example, that entailment is $\mid\!\sim_{\mathcal{P}}$ for $\mathcal{P}(X) = \Delta_X$. Standard Bayesian conditioning is of this form (at least for objective knowledge bases), where we take $\mathcal{P}(X)$ to be a single distribution for each space X. More interestingly, it is well-known [Kullback and Leibler, 1951] that maximum entropy is $I^{\mathcal{P}_u}$ where $\mathcal{P}_u(X)$ is the singleton set containing only the uniform prior on X.

So what can we say about the robustness of $I^{\mathcal{P}}$ to representation shifts? By Theorem 5.5 and Proposition 5.2, we obtain the following corollary:

Corollary 5.6: *$I^{\mathcal{P}}$ is invariant under the embedding f from X to Y iff $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ correspond under f.*

Thus, Corollary 5.6 tells us that if we want $\mathcal{I}^{\mathcal{P}}$ to be invariant under some set $\mathcal{F}$ of embeddings, then we must ensure that our prior function has the right correspondence property.

This result sheds some light on the maximum entropy inference procedure. As we mentioned, $\mid\!\sim_{me}$ is precisely the inference procedure based on the prior function $\mathcal{P}_u$. The corollary asserts that $\mid\!\sim_{me}$ is invariant under f precisely when the uniform priors on X and Y correspond under f. This shows that maximum entropy's lack of representation independence is an immediate consequence of the identical problem for a uniform prior. Is there a class $\mathcal{F}$ of embeddings under which maximum entropy is invariant? Clearly, the answer is yes. It is easy to see that any embedding that takes the elements of X to (disjoint) sets of equal cardinality has the correspondence property required by Corollary 5.6. It follows that maximum entropy is invariant under all such embeddings. In fact, the requirement that maximum entropy be invariant under a subset of these embeddings is one of the axioms in a well-known axiomatic characterization of maximum-entropy [Shore and Johnson, 1980].

If we do not like the behavior of maximum entropy under representation shifts, Theorem 5.6 provides a solution. We should simply start out with a different prior function. Of course, if we want to maintain invariance under all representation shifts, we are forced to use the class of all priors, which gives us entailment as an inference procedure. If, however, we have prior knowledge as to which embeddings encode "reasonable" representation shifts, we can often make do with a smaller class of priors, resulting in an inference procedure that is more prone to leap to conclusions. Given a class of "reasonable" embeddings $\mathcal{F}$, we can often find a prior function $\mathcal{P}$ which is "closed" under each $f \in \mathcal{F}$. That is, for each distribution $\mu \in \mathcal{P}(X)$ and each embedding $f \in F$ from X to Y, we make sure that there is a corresponding distribution $\nu \in \mathcal{P}(Y)$, and vice versa. Therefore, we can guarantee that $\mathcal{P}$ has the appropriate structure using a process of closing off under each f in $\mathcal{F}$.

Of course, we can also execute this process in reverse. Say we want to support a certain reasoning pattern that requires leaping to conclusions. The classical example of such a reasoning pattern is, of course, a default assumption of independence. What is the "most" representation independence that we can get without losing this reasoning pattern? As we now show, Theorem 5.6 gives us the answer.

An *independence structure* Π over Φ^* (our fixed infinite vocabulary) is a partition of the symbols in Φ^* into a collection of disjoint sets or *cells* $\Phi_1, \Phi_2, \ldots$. A distribution μ on $\mathcal{W}(\Phi)$ *respects the independence structure* Π if, for any formulas $\varphi_i \in \mathcal{L}(\Phi_i \cap \Phi)$ and $\varphi_j \in \mathcal{L}(\Phi_j \cap \Phi)$ with $i \neq j$, we have $\mu(\varphi_i \wedge \varphi_j) = \mu(\varphi_i)\mu(\varphi_j)$. Thus, μ makes the denotations of the symbols in different cells independent. Let $\mathcal{P}^{\Pi}(\Phi)$ be the class of all distributions μ over $\mathcal{W}(\Phi)$ that respect Π. We can prove that $\mid\!\sim_{\mathcal{P}_\Pi}$ enforces minimal default independence for symbols in different cells. In fact, it satisfies a somewhat stronger property.

Theorem 5.7: *Let Ψ_1 and Ψ_2 be disjoint vocabularies each of which is the union of cells in Π. If $KB_1 \in \mathcal{L}^{pr}(\Psi_1)$, $\theta_1 \in \mathcal{L}(\Psi_1)$, $KB_2 \in \mathcal{L}^{pr}(\Psi_2)$, $\theta_2 \in \mathcal{L}(\Psi_2)$, then $KB_1 \wedge KB_2 \mid\!\sim_{\mathcal{P}_\Pi} \Pr(\theta_1|\theta_2) = \Pr(\theta_1)$.*

Theorem 3.12 shows that $\mid\!\sim_{\mathcal{P}_\Pi}$ cannot be invariant under all embeddings. Theorem 5.6 tells us that it is invariant under precisely those embeddings for which $\mathcal{P}_\Pi$ is invariant. To characterize these embeddings, suppose that p is in one cell and q and r are in another. Since p and q are in different cells, we have $\mathit{true} \mid\!\sim_{\mathcal{P}} \Pr(p|q) = \Pr(p)$. However, since q and r are in the same cell, we do not have $\mathit{true} \mid\!\sim_{\mathcal{P}} \Pr(r|q) = \Pr(r)$. Hence, $\mathcal{P}_\Pi$ is not invariant under an embedding f that maps p to r. Intuitively, the problem is that f is "crossing cell boundaries". If we restrict to embeddings f that do not cross cell boundaries, i.e., those that for any $p \in \Phi_i$ have $f(p) \in \mathcal{L}(\Phi_i)$, then we avoid this problem.

Proposition 5.8: *The inference procedure $\mid\!\sim_{\mathcal{P}_\Pi}$ is invariant under any embedding f that is faithful and does not cross cell boundaries.*

Theorem 5.6 allows us to define an inference procedure $\mid\!\sim_{\mathcal{P}}$ that enforces minimal default independence (for formulas in different cells), and at the same time is invariant under a large and natural class of embeddings. Given our negative result in Theorem 3.12, this is the best that we could possibly hope for. In general, Theorem 5.6 allows us to understand the tradeoffs between inductive reasoning patterns and invariance under representation shifts.

6 Related Work

Given the importance of representation in reasoning, particularly inductive reasoning, and the fact that one of the main criticisms of maximum entropy has been its sensitivity to representation shifts, it is surprising how little work there has been on the problem of representation dependence. Indeed, to the best of our knowledge, the only work on representation independence in the logical sense that we have considered here is that of Salmon. Salmon [Salmon, 1961] defined a *criterion of linguistic invariance*, which seems essentially equivalent to our notion of representation independence. He tried to use this criterion to defend one particular method of inductive inference but, as pointed out by Barker in the commentary at the end of [Salmon, 1961], his preferred method does not satisfy his criterion either. Salmon then tried to find a modified inductive inference method that did satisfy his criterion [Salmon, 1963], but it is not clear that it does; in any case, our results show that his modified method certainly cannot be representation independent in our sense.

Although statisticians have not considered representation independence in the sense we have defined it here, Bayesian statisticians have been very concerned with related issue of invariance under certain transformations of parameters. For example, we would expect that our beliefs about a person's height should be invariant under a transformation from feet to meters. Their hope is that once we specify the transformation under which we want a distribution to be invariant, the distribution will be uniquely determined [Jaynes, 1968; Kass and Wasserman, 1993]. In this case, the argument goes, the uniquely determined distribution is perforce the "right" one. This idea of picking a distribution using its invariance properties is in the same spirit as the approach we take in Section 5. But unlike the standard Bayesian approach, we do not feel compelled to choose a unique distribution. This enables us to explore a wider spectrum of inference procedures.

Another line of research that is relevant to representation independence is the work on *abstraction* [Giunchiglia and Walsh, 1992; Nayak and Levy, 1994]. Although the goal of this work is again to make connections between two different ways of representing the same situation, there are significant differences in focus. In the work on abstraction, the two ways of representing the situation are not expected to be equivalent. Rather, one representation typically abstracts away irrelevant details that are present in the other. On the other hand, their treatment of the issues is in terms of deductive entailment, not in terms of general inference procedures. It would be interesting to combine these two lines of work.

7 Conclusions

This paper takes a first step towards understanding the issue of representation dependence in probabilistic reasoning, by defining notions of invariance and representation independence, showing that representation independence is incompatible with most types of inductive inference, and defining limited notions of invariance that might that allow a compromise between the desiderata of inductive reasoning and representation independence. Our focus here has been on inference in probabilistic logic, but the notion of representation independence is just as important in many other contexts. Our definitions can clearly be extended to non-probabilistic logics. It is interesting to see in what circumstances our results also carry over. Are there *any* non-deductive logics that are representation independent? We intend to examine this question in future work.

References

[Enderton, 1972] H. B. Enderton. *A Mathematical Introduction to Logic*. Academic Press, New York, 1972.

[Giunchiglia and Walsh, 1992] F. Giunchiglia and T. Walsh. A theory of abstraction. *Artificial Intelligence*, 56(2–3):323–390, 1992.

[Jaynes, 1968] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **SSC**-4:227–241, 1968.

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

[Kass and Wasserman, 1993] R. E. Kass and L. Wasserman. Formal rules for selecting prior distributions: A review and annotated bibliography. Technical Report Technical Report #583, Dept. of Statistics, Carnegie Mellon University, 1993.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Kullback and Leibler, 1951] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

[Nayak and Levy, 1994] P. P. Nayak and A. Y. Levy. A semantic theory of abstractions. 1994.

[Paris, 1994] J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press, 1994.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, Calif., 1988.

[Salmon, 1961] W. Salmon. Vindication of induction. In H. Feigl and G. Maxwell, editors, *Current Issues in the Philosophy of Science*, pages 245–264. Holt, Rinehart, and Winston, New York, 1961.

[Salmon, 1963] W. Salmon. On vindicating induction. In H. E. Kyburg and E. Nagel, editors, *Induction: Some Current Issues*, pages 27–54. Wesleyan University Press, Middletown, Conn., 1963.

[Seidenfeld, 1987] T. Seidenfeld. Entropy and uncertainty. In I. B. MacNeill and G. J. Umphrey, editors, *Foundations of Statistical Inferences*, pages 259–287. 1987.

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