Journal of Artificial Intelligence Research 21 (2004) 319–356 Submitted 11/02; published 03/04

Representation Dependence in Probabilistic Inference

Joseph Y. Halpern HALPERN@CS.CORNELL.EDU
Cornell University, Computer Science Department
Ithaca, NY 14853
http://www.cs.cornell.edu/home/halpern

Daphne Koller KOLLER@CS.STANFORD.EDU
Stanford University, Computer Science Department
Stanford, CA 94035
http://www.cs.stanford.edu/ koller

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. Some have viewed this 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 representation-independent probabilistic inference procedure that ignores irrelevant information is essentially entailment, in a precise sense. Moreover, we show that representation independence is incompatible with even a weak default assumption of independence. We then show that invariance under a restricted class of representation changes can form a reasonable compromise between representation independence and other desiderata, and provide a construction of a family of inference procedures that provides such restricted representation independence, using relative entropy.

1. Introduction

It is well known that the way in which 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. Here too, there is well known work, particularly by Tversky and Kahneman (see, for example, (Kahneman, Slovic, & Tversky, 1982)), showing that the answers given by people can vary significantly (and in systematic ways) depending on how a question is framed. This phenomenon is often viewed as indicating a problem with human information processing; the implicit assumption is that although people do make mistakes of this sort, they shouldn't. On the other hand, there is a competing intuition that suggests that representation does (and *should*) matter; representation dependence is just a natural consequence of this fact.

Here we consider one type of reasoning, probabilistic inference, and examine the extent to which answers depend on the representation. The issue of representation dependence is of particular interest in this context because of the interest in using probability for knowledge representation (e.g., (Pearl, 1988)) and because probabilistic inference has been the source of many of the concerns expressed regarding representation. However, our approach should be applicable far more generally.

We begin by noting that the notion of "probabilistic inference" has two quite different interpretations. In one interpretation, which forms the basis for the Bayesian paradigm, probabilitic inference consists basically of conditioning: We start out with a prior distribution over some event space, and then condition on whatever observations are obtained. In the other interpretation, we are given only a set of probabilistic assertions, and our goal is to reach conclusions about the probabilities of various events. For most of this paper, we focus on the latter interpretation, although we discuss the relationship to the Bayesian approach in Section 7.2.

Suppose that 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 that we have no information whatsoever regarding whether an object is colorful. 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 provided that the language has only the proposition *colorful*. But now suppose we know about the colors red, blue, and green, and have propositions corresponding to each of these colors. Moreover, by *colorful* we actually mean $red \vee blue \vee green$. In this case, maximum entropy dictates that the probability of $red \vee blue \vee green$ 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 that 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 $KB_1^{fly} =_{\rm def} [\Pr(fly \mid bird) = 1/2]$. A second might be to have as basic predicates *bird* and *flying-bird*, and use a knowledge base $KB_2^{fly} =_{\rm def} [(flying\text{-}bird \Rightarrow bird) \wedge \Pr(flying\text{-}bird \mid bird) = 1/2]$. Although the first representation may appear more natural, it seems that both representations are intuitively adequate 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 $\Pr(bird) = 1/2$, while the second leads us to infer $\Pr(bird) = 2/3$. ∎

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 (1961, 1963) and Paris (1994), 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

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.

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.

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 (possible worlds). 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.

We can also interpret a representation as a syntactic entity. In this case, we typically 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. We can then take the states to be truth assignments to these propositions. Similarly, if we use a probabilistic representation language such as *belief networks* (Pearl, 1988) as our knowledge representation language, we must choose some set of relevant random 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? Roughly speaking, we want to capture at the level of the state space a shift from, say, feet to meters. Thus, in X distances might be described in terms of feet where in Y they might be described in terms of meters. We would expect there to be a constraint relating feet to meters. This constraint would not give any extra information about X; it would just relate worlds in X to worlds in Y. Thus, we first attempt to capture representation independence somewhat indirectly, by requiring that adding constraints relating X to Y that place no constraints on X itself should not result in different conclusions about X. The resulting notion, called *robustness*, turns out to be surprisingly strong. We can show that every robust inference procedure must behave essentially like logical entailment.

We then try to define representation independence more directly, by using a mapping f from one representation to another. For example, f could map a world where an individual is 6 feet tall to the corresponding world where the individual is 1.83 meters tall. Some obvious constraints on f are necessary to ensure that it corresponds to our intuition of a representation shift. We can then define a *representation-independent* inference procedure as one that preserves inferences under every legitimate mapping f; i.e., for any KB and θ, $KB \mid\!\sim \theta$ iff $f(KB) \mid\!\sim f(\theta)$.

This definition turns out to be somewhat more reasonable than our first attempt, in that there exist nontrivial representation-independent inference procedures. However, it is still a strong notion. In particular, any representation-independent inference procedure must act essentially like logical entailment for a knowledge base with only *objective* information (i.e., essentially non-probabilistic information). Moreover, we show that representation independence is incompatible with even the simplest default assumption of independence. Even if we are told nothing about the propositions p and q, representation independence does not allow us to jump to the conclusion that p and q are independent.

These results suggest that if we want inference procedures that are capable of jumping to nontrivial conclusions, then we must accept at least some degree of representation dependence. They add support to the claim that the choice of language does carry a great deal of information, and that complete representation independence is too much to expect. On a more positive note, we show that we can use the intuition that the choice of language carries information to get limited forms of representation independence. The idea is that the language should put further constraints on what counts as an "appropriate" representation shift. 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 *sparrow*. 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. We provide a general approach to constructing inference procedures that are invariant under a specific class of representation shifts. This construction allows us to combine some degree of representation independence with certain non-deductive properties that we want of our inference procedure. In particular, we present an inference method that supports a default assumption of independence, and yet is invariant under a natural class of representation shifts.

The rest of this paper is organized as follows. In Section 2, we define probabilistic inference procedures and characterize them. In Section 3, we define robust inference procedures and show that every robust inference procedure is essentially entailment. In Section 4, we define representation independence, and show that representation independence is a very strong requirement. In particular, we show that a representation-independent inference procedure essentially acts like logical entailment on objective knowledge bases and that representation independence is incompatible with a default assumption of independence. Section 5 contains some general discussion of the notion of representation independence and how reasonable it is to assume that the choice of language should affect inference. While it may indeed seem reasonable to assume that the choice of language should affect inference, we point out that this assumption has some consequences that some might view as unfortunate. In Section 6, we discuss how limited forms of representation independence can be achieved. We discuss related work in Section 7, and conclude in Section 8.

2. Probabilistic Inference

We begin by defining probabilistic inference procedures. As we discussed in the introduction, there are two quite different ways in which this term is used. In one, we are given a prior distribution over some probability space; our "knowledge" then typically consists of events in that space, which can be used to condition that distribution and obtain a posterior. In the other, which is the focus of our work, a probabilistic inference procedure takes as input a probabilistic knowledge base and returns a probabilistic conclusion.

We take both the knowledge base and the conclusion to be assertions about the probabilities of events in some *measurable space* $(X, \mathcal{F}_X)$, where a measurable space consists of a set X and an algebra $\mathcal{F}_X$ of subsets of X (that is, $\mathcal{F}_X$ is a set of subsets of X closed under union and complementation, containing X itself).[2] Formally, these assertions can be viewed as statements about (or constraints on) probability measures on $(X, \mathcal{F}_X)$. For example, if

2. If X is infinite, we may want to consider countably-additive probability measures and take $\mathcal{F}_X$ to be closed under countable unions. This issue does not play significant role in this paper. For simplicity, we restrict to finite additivity and require only that $\mathcal{F}_X$ be closed under finite unions.

$S \in \mathcal{F}_X$, a statement $\Pr(S) \geq 2/3$ holds only for distributions where S has probability at least 2/3. Therefore, if $\Delta_{(X,\mathcal{F}_X)}$ is the set of all probability measures on $(X, \mathcal{F}_X)$ (that is, all probability measures with domain $\mathcal{F}_X$), we can view a knowledge base as a set of constraints on $\Delta_{(X,\mathcal{F}_X)}$. When $\mathcal{F}_X$ is clear from context, we often omit it from the notation, writing Δ_X rather than $\Delta_{(X,\mathcal{F}_X)}$.

We place very few restrictions on the language used to express the constraints. We assume that it includes assertions of the form $\Pr(S) \geq \alpha$ for all subsets $S \in \mathcal{F}_X$ and rational $\alpha \in [0,1]$, and that it is closed under conjunction and negation, so that if KB and KB' are knowledge bases expressing constraints, then so are $KB \wedge KB'$ and $\neg KB$. (However, the langauge could include many assertions besides those obtained by starting with assertions of the form $\Pr(S) \geq \alpha$ and closing off under conjunction and negation.) Since the language puts constraints on probability measures, we cannot directly say that $S \in \mathcal{F}_X$ must hold. The closest approximation in the language is the assertion $\Pr(S) = 1$. Thus, we call such constraints *objective*. A knowledge base consisting of only objective constraints is called an *objective knowledge base*. Since $\Pr(T_1) = 1 \wedge \Pr(T_2) = 1$ is equivalent to $\Pr(T_1 \cap T_2) = 1$, without loss of generality, an objective knowledge base consists of a single constraint of the form $\Pr(T) = 1$. Given a knowledge base KB placing constraints on Δ_X, we write $\mu \models KB$ if μ is a measure in Δ_X that satisfies the constraints in KB. We use $[\![KB]\!]_X$ to denote all the measures satisfying these constraints.

In practice, our knowledge is typically represented syntactically, using some logical language to describe the possible states. 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. To facilitate comparison between vocabularies, we assume that for each base logic all the vocabularies are finite subsets of one fixed infinite vocabulary Φ^*.

When working with a language, we assume that each state in the state space defines an interpretation for the symbols in Φ and hence for the formulas in $\mathcal{L}(\Phi)$. In the case of propositional logic, we thus assume that we can associate with each state a truth assignment to the primitive propositions in Φ. For first-order logic, we assume that we can associate with each state a domain and an interpretation of the symbols in Φ. In the probabilistic setting, we assume that we can associate with each state an assignment of values to the random variables. It is often convenient to assume that the state space is in fact some subset W of $\mathcal{W}(\Phi)$, the set of all interpretations for (or assignments to) the vocabulary Φ. 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 $w \models \varphi$.

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). By analogy with constraints, a formula of the form $\Pr(\varphi) = 1$ is called an *objective formula*.

Given a set $W \subseteq \mathcal{W}(\Phi)$, assume that $\mathcal{F}_W$ is the algebra consisting of all sets of the form $[\![\varphi]\!]_W = \{w : w \models \varphi\}$, for $\varphi \in \mathcal{L}(\Phi)$. (In the case of propositional logic, where Φ consists of a finite set of primitive propositions, $\mathcal{F}_W = 2^W$. In the case of first-order logic, not all sets are necessarily definable by formulas, so $\mathcal{F}_W$ may be a strict subset of 2^W.) Let μ be a probability measure on $(W, \mathcal{F}_W)$. We can then ascribe semantics to $\mathcal{L}^{pr}(\Phi)$ in the probability space $(W, \mathcal{F}_W, \mu)$ in a straightforward way. In particular, we interpret the numeric term $\Pr(\varphi)$ as $\mu(\{w \in W : w \models \varphi\})$. Since a formula $\varphi \in \mathcal{L}(\Phi)$ describes an event in the space W, a formula θ in $\mathcal{L}^{pr}(\Phi)$ is clearly a constraint on measures on W. We write $\mu \models \theta$ if the measure $\mu \in \Delta_W$ satisfies the formula θ.

A syntactic knowledge base $KB \in \mathcal{L}^{pr}(\Phi)$ can be viewed as a constraint on Δ_W in an obvious way. Formally, KB represents the set of probability measures $[\![KB]\!]_\Phi \subseteq \Delta_W$, which consists of all measures μ on W such that $\mu \models KB$.

We say that KB (whether syntactic or semantic) 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 measure that satisfies KB also satisfies θ. We write $\models_X \theta$ if θ is satisfied by every measure 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, in particular with respect to its treatment of irrelevant information. Consider the knowledge base consisting only of the constraint $\Pr(\mathit{fly} \mid \mathit{bird}) \geq 0.9$. 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 measures that satisfy KB, but a subset of them. Intuitively, given a knowledge base KB, an inference procedure picks a subset of the measures satisfying KB, and infers θ if θ holds in this subset. Clearly, more conclusions hold for every measure in the subset than hold for every measure in the entire set.

Definition 2.1: An $(X, \mathcal{F}_X)$-*inference procedure* is a partial function $I : 2^{\Delta_{(X,\mathcal{F}_X)}} \mapsto 2^{\Delta_{(X,\mathcal{F}_X)}}$ such that $I(A) \subseteq A$ for $A \subseteq \Delta_{(X,\mathcal{F}_X)}$ and $I(A) = \emptyset$ iff $A = \emptyset$ for all $A \in 2^{\Delta_{(X,\mathcal{F}_X)}}$ in the domain of I (i.e., for all A for which I is defined). We write $KB \mathrel{|\!\sim}_I \theta$ if $I([\![KB]\!]_X) \subseteq [\![\theta]\!]_X$. ∎

When $\mathcal{F}_X$ is clear from context or irrelevant, we often speak of X-inference procedures. We remark that Paris (1994) considers what he calls *inference processes*. These are just inference procedures as we have defined them that, given a a set A of probability measures, return a unique probability measure in A (rather than an arbitrary subset of A). Paris gives a number of examples of inference processes. He also considers various properties that an inference process might have. Some of these are closely related to various properties of representation indepedence that we consider. We discuss Paris's work in Section 7.

Entailment is the X-inference procedure defined on all sets determined by taking I to be the identity. Maximum entropy is also an inference procedure in this sense.

Definition 2.2: Given a probability measure μ on a finite space X (where all sets are measurable), its *entropy* $H(\mu)$ is defined as $-\sum_{x \in X} \mu(x) \log \mu(x)$. (The log is taken to the base 2 here.) Given a set A of measures in Δ_X, let $I_X^{me}(A)$ consist of the measures in A

that have the highest entropy if there are measures in A whose entropy is at least as high as that of any measure in A; if there are no such measures, $\inf^{me}(A)$ is undefined. ∎

It is easy to see that $\inf^{me}(A)$ is defined if A is closed (in the topological sense; i.e., if μ_n is a sequence of probability measures in A and μ_n converges to μ, then $\mu \in A$). Thus, we could take the domain of $\inf_X^{me}$ to consist only of the closed sets of measures in Δ_X. There are also open sets A for which $\inf^{me}(A)$ is defined, although it is not defined for all open sets A. For example, suppose $X = \{x_1, x_2\}$ and let $A = \{\mu : \mu(x_1) < 1/2\}$. Let μ_0 be such that $\mu_0(x_1) = 1/2$. It is easy to check that $H(\mu_0) = 1$, and $H(\mu) < 1$ for $\mu \in A$. However, for all ϵ, there is some $\mu \in A$ such that $H(\mu) > 1 - \epsilon$. It follows that there is no measure in A whose entropy is higher than that of any other measure in A, so $\inf^{me}(A)$ is undefined. On the other hand, if $A' = \{\mu : \mu(x_1) < 2/3\}$, then there is a measure whose entropy is maximum in the open set A', namely the measure μ_0.

There are, of course, many inference procedures besides entailment and maximum entropy that can be defined on a measurable space. In fact, as the following proposition shows, any binary relation $\mid\!\sim$ satisfying certain reasonable properties is an inference procedure of this type.

Proposition 2.3: *If I is an X-inference procedure then the following properties hold for every KB, KB', θ, ψ over X such that KB is in the domain of I.*

- Reflexivity: $KB \mid\!\sim_I KB$.
- Left Logical Equivalence: *if KB is logically equivalent to KB', i.e., if $\models KB \Leftrightarrow KB'$, then for every θ $KB \mid\!\sim_I \theta$ iff $KB' \mid\!\sim_I \theta$.*
- Right Weakening*: if $KB \mid\!\sim_I \theta$ and $\models \theta \Rightarrow \psi$ then $KB \mid\!\sim_I \psi$.*
- And: *if $KB \mid\!\sim_I \theta$ and $KB \mid\!\sim_I \psi$, then $KB \mid\!\sim_I \theta \wedge \psi$.*
- Consistency*: if KB is consistent then $KB \not\mid\!\sim_I$ false.* ∎

Proof: Straightforward from the definitions. ∎

Interestingly, these properties are commonly viewed as part of a core of reasonable properties for a nonmonotonic inference relation (Kraus, Lehmann, & Magidor, 1990).

We would like to also prove a converse, showing that any relation $\mid\!\sim$ over probabilistic constraints on some space X that satisfies the five properties above must have the form $\mid\!\sim_{I_X}$. This is not quite true, as the following example shows.

Example 2.4: Fix a measurable space $(X, \mathcal{F}_X)$. Let the language consist of all (finite) Boolean combination of statements of the form $\Pr(S) \geq \alpha$, where $S \in \mathcal{F}_X$. Now fix one nonempty strict subset S_0 of X, and let φ_n be the statement $\Pr(S_0) \leq 1/n$. Define an inference procedure $\mid\!\sim$ as follows. If KB is not equivalent to *true* (i.e, if $[\![KB]\!]_X \neq \Delta_X$), then $KB \mid\!\sim \theta$ iff $KB \models \theta$. On the other hand, *true* $\mid\!\sim \theta$ iff $\varphi_n \models \theta$ for all sufficiently large n. That is, *true* $\mid\!\sim \theta$ if there exists an N such that for all $n \geq N$, we have $\varphi_n \models \theta$. It is easy to check that all five properties in Proposition 2.3 hold for $\mid\!\sim$. However, $\mid\!\sim$ is not $\mid\!\sim_I$ for an X-inference procedure I. For suppose it were. Note that $\varphi_n \models \varphi_m$ for all $n \geq m$, so $true \mathrel{|\!\sim} \varphi_m$ for all m. Thus, we must have $I(\Delta_X) \subseteq [\![\varphi_n]\!]_X$ for all n. It follows that $I_X(\Delta_X) \subseteq [\![\Pr(S_0) = 0]\!]_X$, and so $true \mathrel{|\!\sim}_I \Pr(S_0) = 0$. However, $\varphi_n \not\models \Pr(S) = 0$ for any n, so we do not have $true \mathrel{|\!\sim} \Pr(S) = 0$. This contradicts the assumption that $\mathrel{|\!\sim} \; = \; \mathrel{|\!\sim}_I$. ∎

Essentially what we need to get the converse to Proposition 2.3 is an infinitary version of the And Rule, which would say that if $KB \mathrel{|\!\sim}_I \theta_i$ for all i, then $KB \mathrel{|\!\sim}_I \bigwedge_i \theta_i$. If the language were closed under infinite conjunctions, then this rule would in fact be just what we need. Since we have not assumed that the language is closed under infinite conjunctions, we use a variant of this rule.

- *Infinitary And:* For any set Σ of statements, if $KB \mathrel{|\!\sim}_I \theta$ for all $\theta \in \Sigma$ and $\Sigma \models \psi$, then $KB \mathrel{|\!\sim}_I \psi$.

Proposition 2.5: *Let $\mathrel{|\!\sim}$ be a relation over probabilistic constraints on X for which the properties Reflexivity, Left Logical Equivalence, Right Weakening, Infinitary And, and Consistency hold for all KB in the domain of $\mathrel{|\!\sim}$. (That is, if KB is in the domain of $\mathrel{|\!\sim}$ in that $KB \mathrel{|\!\sim} \theta$ for some θ, then $KB \mathrel{|\!\sim} KB$, and so on.) Then $\mathrel{|\!\sim}$ is $\mathrel{|\!\sim}_I$ for some X-inference procedure I.*

Proof: See Appendix A.1. ∎

We are typically interested not just in an inference procedure defined on one space X, but in a family of related inference procedures, defined on a number of spaces. For example, entailment is an inference procedure that is defined on all spaces X; maximum entropy is defined on all finite measurable spaces $(X, 2^X)$.

Definition 2.6: If $\mathcal{X}$ is a set of measurable spaces, an $\mathcal{X}$*-inference procedure* is a set $\{I_{(X,\mathcal{F}_X)} : (X, \mathcal{F}_X) \in \mathcal{X}\}$, where $I_{(X,\mathcal{F}_X)}$ is an $(X, \mathcal{F}_X)$-inference procedure for $(X, \mathcal{F}_X) \in \mathcal{X}$. ∎

We sometimes talk about an $\mathcal{X}$-inference procedure I, and write $KB \mathrel{|\!\sim}_I \theta$ when $(X, \mathcal{F}_X) \in \mathcal{X}$ is clear from context. However, it should be stressed that, formally, an $\mathcal{X}$-inference procedure is a really a set of inference procedures (typically related in some natural way).

Clearly entailment is an $\mathcal{X}$-inference procedure for any $\mathcal{X}$, where I_X is simply the identity function for $X \in \mathcal{X}$. If $\mathcal{X}$ consists of finite measurable spaces where all sets are measurable, then maximum entropy is an $\mathcal{X}$-inference procedure. We typically denote this inference procedure $\mathrel{|\!\sim}_{me}$. Thus, $KB \mathrel{|\!\sim}_{me} \theta$ if θ holds for all the probability measures of maximum entropy satisfying KB.

Important assumptions: For the remainder of this paper, we deal only with $\mathcal{X}$-inference procedures I for which $\mathcal{X}$ satisfies two richness assumptions. These assumptions hold for all the standard inference procedures that have been considered.

- We assume that $\mathcal{X}$ is *closed under crossproducts*, so that if $(X, \mathcal{F}_X), (Y, \mathcal{F}_Y) \in \mathcal{X}$, then $(X \times Y, \mathcal{F}_{X \times Y}) \in \mathcal{X}$, where $\mathcal{F}_{X \times Y}$ is the algebra formed by taking finite unions of disjoint sets of the form $S \times T$, for $S \in \mathcal{F}_X$ and $T \in \mathcal{F}_Y$. It is easy to see that this is an algebra, since $\overline{S \times T} = \overline{S} \times T \cup S \times \overline{T} \cup \overline{S} \times \overline{T}$ and $(S \times T) \cap (S' \times T') = (S \cap S') \times (T \cap T')$ (from which it also follows that any union of such sets can be written as a disjoint union). Note that if X and Y are finite sets, $\mathcal{F}_X = 2^X$, and $\mathcal{F}_Y = 2^Y$, then $\mathcal{F}_X \times \mathcal{F}_Y = 2^{X \times Y}$. As we shall see, having $(X \times Y, \mathcal{F}_{X \times Y}) \in \mathcal{X}$ if each of $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ is in $\mathcal{X}$ allows us to relate constraints on X to constraints on Y in a natural way.

- We assume that $\mathcal{X}$ contain sets of all finite cardinalities; more precisely, for all $n \geq 2$, there exists a set $(X, \mathcal{F}_X) \in \mathcal{X}$ such that $|X| = n$ and $\mathcal{F}_X = 2^X$. This assumption is not actually needed for any of our results, since the assumption that $\mathcal{X}$ is closed under crossproducts already implies that, for any finite n, there exists a measurable space $(X, \mathcal{F}_X) \in \mathcal{X}$ such that $|X| \geq n$; this already suffices to prove all the results of the paper. However, assuming that $\mathcal{X}$ has sets of all cardinalities does make the proofs easier.

We also want the domain of I to satisfy certain assumptions, but we defer stating these assumptions until we have introduced some additional definitions and notation.

3. Robustness

In order to define robustness to representation shifts, we must first define the notion of a representation shift. Our first attempt at this definition is based on the idea of using constraints that specify the relationship between the two vocabularies. For example, in Example 1.1, we might have $X = \{\mathit{colorful}, \mathit{colorless}\}$ and $Y = \{\mathit{red}, \mathit{blue}, \mathit{green}, \mathit{colorless}\}$. We can specify the relationship between X and Y via a constraint that asserts that $\mathit{colorful} \Leftrightarrow (\mathit{red} \vee \mathit{blue} \vee \mathit{green})$.

Of course, not every constraint is a legitimate mapping between representations. For example, a formula that asserted $\neg \mathit{colorful}$ is obviously not a legitimate representation shift. At a minimum, we must assume that the constraint does not give any additional information about X as far as logical inference goes. At a syntactic level, we can use the following definition. Given a knowledge base $KB \in \mathcal{L}^{pr}(\Phi)$, we say that $\psi \in \mathcal{L}^{pr}(\Phi \cup \Phi')$ is Φ-*conservative* over KB if, for all formulas $\varphi \in \mathcal{L}^{pr}(\Phi)$, we have $KB \models \varphi$ iff $KB \wedge \psi \models \varphi$. Thus, adding ψ to the knowledge base does not permit any additional logical inferences in the vocabulary Φ. An inference procedure I is *robust* if it is unaffected by conservative extensions; that is, if $KB, \varphi \in \mathcal{L}^{pr}(\Phi)$, then $KB \mid\!\sim_I \varphi$ iff $KB \wedge \psi \mid\!\sim_I \varphi$ for all ψ that are Φ-conservative over KB. Roughly speaking, this says that getting new information that is uninformative as far as logical inference goes does not affect default conclusions.

The formal definition of robustness, which uses semantic rather than syntactic concepts, extends these intuitions to arbitrary constraints on measures (not just ones that can be expressed in the language $\mathcal{L}^{pr}$).

Definition 3.1: For $\mu \in \Delta_{X_1 \times \ldots \times X_n}$, define $\mu_{X_i} \in \Delta_{X_i}$ by taking $\mu_{X_i}(A) = \mu(X_1 \times \cdots \times X_{i-1} \times A \times X_{i+1} \times \cdots \times X_n)$. A constraint φ on Δ_{X_i} can be viewed as a constraint on $\Delta_{X_1 \times \ldots \times X_n}$ by taking $[\![\varphi]\!]_{X_1 \times \ldots \times X_n} = \{\mu \in \Delta_{X_1 \times \ldots \times X_n} : \mu_{X_i} \models \varphi\}$. We frequently identify constraints on X_i with constraints on $X_1 \times \ldots \times X_n$ in this way. For $B \subseteq \Delta_{X_1 \times \cdots \times X_n}$, define $\mathit{proj}_{X_i}(B) = \{\mu_{X_i} : \mu \in B\}$. A constraint ψ on $\Delta_{X_1 \times \cdots \times X_n}$ is said to be X_i-*conservative over the constraint* KB *on* Δ_{X_i} if $\mathit{proj}_{X_i}([\![KB \wedge \psi]\!]_{X_1 \times \cdots \times X_n}) = [\![KB]\!]_{X_i}$. ∎

To see that this definition generalizes the earlier language-oriented definition, note that if φ and KB are constraints on Δ_X and ψ is a constraint on $\Delta_{X\times Y}$, then $KB \wedge \psi \models \varphi$ iff $proj_1([\![KB \wedge \psi]\!]_{X\times Y}) \subseteq [\![\varphi]\!]_X$, while $KB \models \varphi$ iff $[\![KB]\!]_X \subseteq [\![\varphi]\!]_X$.

Definition 3.2: $\{I_X : X \in \mathcal{X}\}$ is a *robust* $\mathcal{X}$-inference procedure if for all spaces $X, Y \in \mathcal{X}$, constraints KB and φ on Δ_X, and constraints ψ on $\Delta_{X\times Y}$ that are X-conservative over KB, we have $KB \mathrel{|\!\!\sim}_{I_X} \varphi$ iff $KB \wedge \psi \mathrel{|\!\!\sim}_{I_{X\times Y}} \varphi$. (Note that this definition implicitly assumes that $X \times Y \in \mathcal{X}$ if $X, Y \in \mathcal{X}$, an assumption we made explicit earlier.) ∎

At first glance, robustness might seem like a reasonable desideratum. After all, why should adding a constraint on $\Delta_{X\times Y}$ that places no restrictions on Δ_X change the conclusions that we might reach about X? Unfortunately, it turns out that this definition is deceptively strong, and disallows any "interesting" inference procedures. In particular, one property we may hope for in an inference procedure is to draw nontrivial conclusions about probabilities of events, that is, conclusions that do not follow from entailment. For example, maximum entropy (or any inference procedure based on symmetry) will conclude $\Pr(p) = 1/2$ from the empty knowledge base. We can show that inference procedures that are robust do not really allow much in the way of nontrivial conclusions about the probabilities of events.

Definition 3.3: An $(X, \mathcal{F}_X)$-inference procedure I is *essentially entailment for the knowledge base* $KB \subseteq \Delta_X$ if for all $S \in \mathcal{F}_X$, if $KB \mathrel{|\!\!\sim}_I \alpha < \Pr(S) < \beta$ then $KB \models \alpha \le \Pr(S) \le \beta$. I is *essentially entailment for* X if it is essentially entailment for all knowledge bases KB in the domain of I_X. ∎

Thus, when entailment lets us conclude $\Pr(S) \in [\alpha, \beta]$, an inference procedure that is essentially entailment lets us draw only the slightly stronger conclusion $\Pr(S) \in (\alpha, \beta)$. To prove this, we need to make three assumptions about the domain of I. (For other results, we need other assumptions about the domain of I.)

DI1. $\alpha \le \Pr(S) \le \beta$ is in the domain of $I_{(X,\mathcal{F}_X)}$ for all $S \in \mathcal{F}_X$, $\alpha, \beta \in \mathbb{R}$.

DI2. If KB is in the domain of I_X, then it is also in the domain of $I_{X\times Y}$ (when KB is viewed as a constraint on $\Delta_{X\times Y}$.)

DI3. If KB_1 and KB_2 are in the domain of I_X, then so is $KB_1 \wedge KB_2$.

Note that sets of the form $\alpha \le \Pr(S) \le \beta$ are closed sets. It certainly seems reasonable to require that such sets be in the domain of an inference procedure; they correspond to the most basic observations. DI2 seems quite innocuous; as observed earlier, we do want to be able to view constraints on Δ_X as constraints on $\Delta_{X\times Y}$, and doing so should not prevent them from being in the domain of I. DI3 also seems to be a reasonable assumption, since if KB_1 and KB_2 correspond to possible observations, we want to be able to draw conclusions from combining the observations. DI3 holds if the domain of I consists of closed sets. But note that it does not hold for I^{me} if we take its domain to consist of all sets that have a measure whose entropy is maximum. For example, if $X = \{x_1, x_2\}$, $A = \{\mu_0\} \cup \{\mu : \mu(x_1) > 3/4\}$, and $B = \{\mu : \mu(x_1) \ge 2/3\}$, where $\mu_0(x_0) = 1/2$, then each of A and B have a measure whose entropy is maximum, but $A \cap B$ does not have a measure whose entropy is maximum.

Theorem 3.4: *If $\{I_X : X \in \mathcal{X}\}$ is a robust $\mathcal{X}$-inference procedure that satisfies DI1, DI2, and DI3, then I_X is essentially entailment for all $X \in \mathcal{X}$.*

Proof: See Appendix A.2. ∎

It is possible to construct robust inference procedures that are almost but not quite entailment, simply by "strengthening" some conclusions from $\Pr(S) \in [\alpha, \beta]$ to $\Pr(S) \in (\alpha, \beta)$. Clearly, however, any robust inference procedure is extremely limited in its ability to jump to conclusions. In the next section, we look at a definition that seems closer to the intuitive notion of representation independence, and has somewhat more reasonable consequences.

4. Representation Independence

4.1 Representation shifts

If X and Y are two different representations of the same phenomena then, intuitively, there should be a way of relating states in X to corresponding states in Y. We want this correspondence to respect the logical structure of events. Formally, we require that it be a homomorphism with respect to complementation and intersection.

Definition 4.1: An $(X, \mathcal{F}_X)$-$(Y, \mathcal{F}_Y)$ *embedding* f is a function $f : \mathcal{F}_X \mapsto \mathcal{F}_Y$ such that $f(S \cup T) = f(S) \cup f(T)$ and $f(\overline{S}) = \overline{f(S)}$ for all $S, T \in \mathcal{F}_X$. ∎

As elsewhere, we talk about X-Y embeddings rather than $(X, \mathcal{F}_X)$-$(Y, \mathcal{F}_Y)$ embeddings if $\mathcal{F}_X$ and $\mathcal{F}_Y$ do not play a significant role.

Our goal is to consider the effect of a transformation on probabilistic formulas. Hence, we are interested in sets of states and their probabilities.

Definition 4.2: If f is an X-Y embedding, $\mu \in \Delta_X$, and $\nu \in \Delta_Y$, then μ and ν *correspond under* f if $\mu(S) = \nu(f(S))$ for all events $S \in \mathcal{F}_X$. We define a mapping $f^* : 2^{\Delta_X} \mapsto 2^{\Delta_Y}$ as follows. We first define f^* on singleton sets (except that, for convenience, we write $f^*(\mu)$ rather than $f^*(\{\mu\})$ by taking $f^*(\mu) = \{\nu \in \Delta_Y : \nu(f(S)) = \mu(S) \text{ for all } S \in \mathcal{F}_X\}$. Thus, $f^*(\mu)$ consists of all measures in Δ_Y that correspond to μ under f. If $\mathcal{D}$ is an arbitrary subset of 2^{Δ_X}, define $f^*(\mathcal{D}) = \cup_{\mu \in \mathcal{D}} f^*(\mu)$ for $\mathcal{D} \subseteq \Delta_X$. ∎

If θ is a constraint on Δ_X expressed in some language, we typically write $f^*(\theta)$ rather than $f^*([\![\theta]\!]_X)$. We implicitly assume that the language is such that the constraint $f^*(\theta)$ is also expressible. It is not hard to see that $f^*(\theta)$ is the constraint that results by replacing every set $S \in \mathcal{F}_X$ that appears in θ by $f(S)$.

Example 4.3: In Example 1.1, we might have $X = \{\textit{colorful}, \textit{colorless}\}$ and $Y = \{\textit{red}, \textit{blue}, \textit{green}, \textit{colorless}\}$. In this case, we might have $f(\textit{colorful}) = \{\textit{red}, \textit{blue}, \textit{green}\}$ and $f(\textit{colorless}) = \{\textit{colorless}\}$. Consider the measure $\mu \in \Delta_X$ such that $\mu(\textit{colorful}) = 0.7$ and $\mu(\textit{colorless}) = 0.3$. Then $f^*(\mu)$ is the set of measures ν such that the total probability assigned to the set of states $\{\textit{red}, \textit{blue}, \textit{green}\}$ by ν is 0.7. Note that there are uncountably many such measures. It is easy to check that if θ is a constraint on Δ_X such as $\Pr(\textit{colorful}) > 3/4$, then $f^*(\theta)$ is $\Pr(\{\textit{red}, \textit{blue}, \textit{green}\}) > 3/4$. ∎

Embeddings can be viewed as the semantic analogue to the syntactic notion of *interpretation* defined in (Enderton, 1972, pp. 157–162), which has also been used in the recent literature on *abstraction* (Giunchiglia & Walsh, 1992; Nayak & Levy, 1995). Essentially, an interpretation maps formulas in a vocabulary Φ to formulas in a different vocabulary Ψ by mapping the primitive propositions in Φ (e.g., *colorful*) to formulas over Ψ (e.g., $\mathit{red} \vee \mathit{blue} \vee \mathit{green}$) 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 $\mathit{fly} \wedge \mathit{bird}$.

Definition 4.4: 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 primitive proposition $p \in \Phi$ 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 f from $W_1 \subseteq \mathcal{W}(\Phi)$ to $W_2 \subseteq \mathcal{W}(\Psi)$: we map $[\![\varphi]\!]_{W_1}$ to $[\![i(\varphi)]\!]_{W_2}$.

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 by telling us that a certain situation—that where $p \wedge \neg q$ holds—is not possible. More formally, the 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 4.5: An X-Y embedding f is *faithful* if, for all $S, T \in \mathcal{F}_X$, we have $S \subseteq T$ iff $f(S) \subseteq f(T)$. ∎

This definition has the desired consequence of disallowing embeddings that give new information as far as logical consequence goes.

Lemma 4.6: *An X-Y embedding f is faithful if and only if for all constraints KB and θ, we have $KB \models \theta$ iff $f^*(KB) \models f^*(\theta)$.*

Proof: See Appendix A.3. ∎

It is clear that our embedding from Example 4.3 is faithful: $f(\mathit{colorful}) = \{\mathit{red}, \mathit{blue}, \mathit{green}\}$ and $f(\mathit{colorless}) = \mathit{colorless}$. The following proposition gives further insight into faithful embeddings.

Proposition 4.7: *Let f be a faithful X-Y embedding. Then the following statements are equivalent:*

(a) μ and ν correspond under f;

(b) for all formulas θ, $\mu \models \theta$ iff $\nu \models f^*(\theta)$.

Proof: See Appendix A.3. ∎

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

Definition 4.8: If $X, Y \in \mathcal{X}$, then the $\mathcal{X}$-inference procedure $\{I_X : X \in \mathcal{X}\}$ is *invariant* under the X-Y embedding f if for all constraints KB and θ on Δ_X, we have $KB \mathrel{|\!\!\sim}_{I_X} \theta$ iff $f^*(KB) \mathrel{|\!\!\sim}_{I_Y} f^*(\theta)$. (Note that, in particular, this means that KB is in the domain of $\mathrel{|\!\!\sim}_{I_X}$ iff $f^*(KB)$ is in the domain of $\mathrel{|\!\!\sim}_{I_Y}$.) ∎

Definition 4.9: The $\mathcal{X}$-inference procedure $\{I_X : X \in \mathcal{X}\}$ is *representation independent* if it is invariant under all faithful X-Y embeddings for all $X, Y \in \mathcal{X}$. ∎

Since the embedding for Example 4.3 is faithful, any representation-independent inference procedure 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.

4.2 Representation-independent inference procedures

Although the definition of representation independence seems natural, so did the definition of robustness. How do the two definitions relate to each other? First, we show that representation independence is a weaker notion than robustness. For this result, we need to consider inference procedures that satisfy two further assumptions.

DI4. If f is a faithful X-Y embedding, then KB is in the domain of I_X iff $f^*(KB)$ is in the domain of I_Y.

DI5. If KB is in the domain of $I_{X \times Y}$, f is a faithful X-Y embedding, and φ_1 is a constraint on Δ_X, then $KB \wedge (\varphi_1 \Leftrightarrow f^*(\varphi_1))$ is in the domain of $I_{X \times Y}$.

DI4 is very natural and is satisfied by all the standard inference procedures. It is easy to check that if KB is closed iff $f^*(KB)$ is closed. While DI5 may not appear so natural, it does hold for domains consisting of closed sets, since it is not hard to check that $\varphi \Leftrightarrow f^*(\varphi_1)$ is closed. DI5 would follow from DI3 and the assumption that $\varphi \Leftrightarrow f^*(\varphi_1)$ is in the domain of $I_{X \times Y}$, but it is actually weaker than the combination of these two assumptions. In particular, it holds for the domain consisting of all sets on which there is a measure of maximum entropy.

Theorem 4.10: *If an $\mathcal{X}$-inference procedure is robust that satisfies DI2, DI4, and DI5, then it is representation independent.*

Proof: See Appendix A.3. ∎

We have already shown that any robust inference procedure must be almost trivial. Are there any interesting representation-independent inference procedures? As we shall see, the answer is mixed. There are nontrivial representation-independent inference procedures, but they are not very interesting.

Our first result shows that representation independence, like robustness, trivializes the inference procedure, but only for some knowledge bases.

Theorem 4.11: *If $\{I_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure then, for all $X \in \mathcal{X}$, I_X is essentially entailment for all objective knowledge bases in its domain.*[3]

Proof: See Appendix A.3. ∎

Corollary 4.12: *If $\{I_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure, KB is objective, and $KB \mid\!\sim_I \alpha < \Pr(S) < \beta$ for some $\alpha \geq 0$ and $\beta \leq 1$, then $\alpha = 0$ and $\beta = 1$.*

This result tells us that from an objective knowledge base $\Pr(T) = 1$, we can reach only three possible conclusions about a set S. If $T \subseteq S$, then we can conclude that $\Pr(S) = 1$; if $T \subseteq \overline{S}$, then we can conclude that $\Pr(S) = 0$; otherwise, the *strongest* conclusion we can make about $\Pr(S)$ is that is somewhere between 0 and 1.

We can construct a representation-independent inference procedure that is not entailment and has precisely this behavior if we restrict attention to countable state spaces. Suppose that X is countable. Given an objective knowledge base KB of the form $\Pr(T) = 1$, where $T \in \mathcal{F}_X$, let KB^+ consist of all formulas of the form $0 < \Pr(S) < 1$ for for all nonempty strict subsets S of T in $\mathcal{F}_X$.[4] We now define an X-inference procedure I^0_X as follows: If KB is equivalent to an objective knowledge base, then $KB \mid\!\sim_{I^0} \varphi$ if $KB \wedge KB^+ \models \varphi$; if KB is not equivalent to an objective knowledge base, then $KB \mid\!\sim_{I^0} \varphi$ if $KB \models \varphi$. It follows easily from Proposition 2.5 that I^0_X is indeed an inference procedure. Moreover, it is not equivalent to the standard notion of entailment; for example, we have $true \mid\!\sim_{I^0} 0 < \Pr(p) < 1$, while $\not\models 0 < \Pr(p) < 1$. Nevertheless, we can prove that I^0 is representation independent.

Lemma 4.13: *Let $\mathcal{X}$ consist of only countable sets. Then $\{I^0_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure.*

3. In an earlier version of this paper (Halpern & Koller, 1995), we claimed that any representation-independent inference procedure that satisfied a minimal irrelevance property (implied by robustness, but not equivalent to it) is essentially entailment for all knowledge bases. As Jaeger (1996) shows, an inference procedure along the lines of I^1 described below can be constructed to show that this result is not correct. We seem to need the full strength of robustness.

4. The requirement that X be countable is necessary here. If X is uncountable and every singleton is in $\mathcal{F}_X$, then KB^+ is inconsistent if both T and $\overline{T}$ are uncountable. It is impossible that each of an uncountable collection of points has positive measure.

Proof: See Appendix A.3. ∎

While objective knowledge bases may not appear so interesting if we restrict to propositional languages, for languages that include first-order and statistical information they become quite interesting. Indeed, as shown in (Bacchus, 1990; Bacchus, Grove, Halpern, & Koller, 1996), knowledge bases with first-order and (objective) statistical information allow us to express a great deal of the information that we naturally encounter. For example, we can express the fact that "90% of birds fly" as an objective statement about the number of flying birds in our domain relative to the overall number of birds. Of course, Theorem 4.11 applies immediately to such knowledge bases.

Theorem 4.11 also implies that various inference procedures cannot be representation independent. In particular, since $true \mathrel{|\!\sim}_{me} \Pr(p) = 1/2$ for a primitive proposition p, it follows that maximum entropy is not essentially entailment. This observation provides another proof that maximum entropy is not representation independent.

It is consistent with Theorem 4.11 that there are representation-independent inference procedures that are not almost entailment for probabilistic knowledge bases. For example, consider the X-inference procedure I^1_X defined as follows. Given $A \subseteq \Delta_X$, if there exists some $S \in \mathcal{F}_X$ such that $A = \{\mu \in \Delta_X : \mu(S) \geq 1/4\}$, then $I^1_X(A) = \{\mu \in \Delta_X : \mu(S) \geq 1/3\}$; otherwise, $I^1_X(A) = A$. Thus, $\Pr(S) \geq 1/4 \mathrel{|\!\sim}_{I^1} \Pr(S) \geq 1/3$. Clearly, I^1_X is not essentially entailment. Yet, we can prove the following result.

Lemma 4.14: *Suppose that $\mathcal{X}$ consists only of measure spaces of the form $(X, 2^X)$, where X is finite. Then $\{I^1_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure.*

Proof: See Appendix A.3. ∎

Note that it follows from Theorem 3.4 that I^1 cannot be robust. Thus, we have shown that representation independence is a strictly weaker notion than robustness.

This example might lead us to believe that there are representation-independent inference procedures that are "interesting" for probabilistic knowledge bases. However, as we now show, a representation-independent inference procedure cannot satisfy one key desideratum: the ability to conclude independence by default. For example, an important feature of the maximum-entropy approach to nonmononotic reasoning (Goldszmidt, Morris, & Pearl, 1993) has been its ability to ignore "irrelevant" information, by implicitly assuming independence. Of course, maximum entropy does not satisfy representation independence. Our result shows that no approach to probabilistic reasoning can simultaneously assure representation independence and a default assumption of independence.

We do not try to give a general notion of "default assumption of independence" here, since we do not need it for our result. Rather, we give a minimal property that we would hope an inference procedure might have, and show that this property is sufficient to preclude representation independence. Syntactically, the property we want is that if Φ and Ψ are disjoint vocabularies, $KB \in \mathcal{L}^{pr}(\Phi)$, $\varphi \in \mathcal{L}(\Phi)$, and $\psi \in \mathcal{L}(\Psi)$, then $KB \mathrel{|\!\sim}_I \Pr(\varphi \wedge \psi) = \Pr(\varphi) \times \Pr(\psi)$.

Definition 4.15: An $\mathcal{X}$-inference procedure $\{I_X : X \in \mathcal{X}\}$ enforces *minimal default independence* if, whenever X and Y are in $\mathcal{X}$, KB is a constraint on Δ_X in the domain of $\mid\!\sim_{I_X}$, $S \in \mathcal{F}_X$, and $T \in \mathcal{F}_Y$, then $KB \mid\!\sim_{I_{X\times Y}} \Pr(S \times T) = \Pr(S) \times \Pr(T)$.[5] ∎

This definition clearly generalizes the syntactic definition.

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

Theorem 4.16: *If $\{I_X : X \in \mathcal{X}\}$ is an $\mathcal{X}$-inference procedure that enforces minimal default independence and satisfies DI1, then I_X is not representation independent.*

Proof: See Appendix A.3. ∎

This result is very interesting as far as irrelevance is concerned. We might hope that learning irrelevant information does not affect our conclusions. While we do not attempt to define irrelevance here, certainly we would expect that if KB' is in a vocabulary disjoint from KB and φ, then, for example, $KB \mid\!\sim_I \Pr(\varphi) = \alpha$ iff $KB \wedge KB' \mid\!\sim_I \Pr(\varphi) = \alpha$. If KB' is objective, then the standard probabilistic approach would be to identify learning KB' with conditioning on KB'. Suppose that we restrict to inference procedures that do indeed condition on objective information (as is the case for the class of inference procedures we consider in Section 6). Then $KB \wedge KB' \mid\!\sim_I \Pr(\varphi) = \alpha$ exactly if $KB \mid\!\sim_I \Pr(\varphi \mid KB') = \alpha$. Thus, Theorem 4.16 tells us that inference procedures that condition on new (objective) information cannot both be representation independent and ignore irrelevant information.

Thus, although representation independence, unlike robustness, does not force us to use entirely trivial inference procedures, it does prevent us from using procedures that have certain highly desirable properties.

5. 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 5.1: Let Q be a unary predicate and $c_1, \ldots, c_{100}, d$ be constant symbols. Suppose that we have two vocabularies $\Phi = \{Q, d\}$ and $\Psi = \{Q, c_1, \ldots, c_{100}, d\}$. Consider the interpretation i from Φ to Ψ for which $i(d) = d$ and $i(Q(x)) = Q(x) \wedge Q(c_1) \wedge \ldots \wedge Q(c_{100})$. Now, consider $KB = \exists x Q(x)$. In this case, $i(KB) = \exists x (Q(x) \wedge Q(c_1) \wedge \ldots \wedge Q(c_{100}))$. Intuitively, since all the c_i's may refer to the same domain element, the only conclusion we can make with certainty from $Q(c_1) \wedge \ldots \wedge Q(c_{100})$ is that there exists at least one Q in the domain, which gives us no additional information beyond KB. We can convert this example into a general argument that the embedding f corresponding to i is faithful. Intuitively, for

5. Since we are working in the space $X \times Y$, KB should be viewed as a constraint on $\Delta_{X\times Y}$ here, $\Pr(S)$ should be understood as $\Pr(S \times Y)$, while $\Pr(T)$ should be understood as $\Pr(X \times T)$. Recall that, by assumption, $X \times Y \in \mathcal{X}$.

any KB, we can only get the conclusion $Q(c_1) \wedge \ldots \wedge Q(c_{100})$ from $f^*(KB)$ if $Q(x)$ appears positively in KB; but, in this case, we already know that there is at least one Q, so we gain no new information from the embedding. But it does not seem unreasonable that an inference procedure should assign different degrees of belief to $Q(d)$ given $KB = \exists x Q(x)$ on the one hand and given $i(KB) = \exists x(Q(x) \wedge Q(c_1) \wedge \ldots \wedge Q(c_{100}))$ on the other,[6] particularly if the domain is small. In fact, many 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 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 that we used to prove Theorem 4.16 and 3.4 seem to apply to any reasonable notion of representation independence.[7] To give the flavor of the type of argument used to prove these theorems, consider Example 1.1, and assume that $\mathit{true} \mid\!\sim_I \Pr(\mathit{colorful}) = \alpha$ for $\alpha \in (0,1)$.[8] Using an embedding g such that $g(\mathit{colorful}) = \mathit{red}$, we conclude that $\mathit{true} \mid\!\sim_I \Pr(\mathit{red}) = \alpha$. Similarly, we can conclude $\Pr(\mathit{blue}) = \alpha$ and $\Pr(\mathit{green}) = \alpha$. But in order for $\mid\!\sim_I$ to be invariant under our original embedding, we must have $\mathit{true} \mid\!\sim_I \Pr(\mathit{red} \vee \mathit{blue} \vee \mathit{green}) = \alpha$, 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.

These results can be viewed as support for the position that representation dependence is justified; the choice of an appropriate representation encodes significant information. In 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 (i.e., learning from evidence). 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.

But we need to be a little careful here. For example, in some cases we can identify the vocabulary (and hence, the representation) with the sensors that an agent has at its disposal. It may not seem that unreasonable that an agent with a temperature sensor and a motion sensor might carve up the world differently from an agent with a color sensor and a distance sensor. But consider two agents with different sensors who have not yet made any observations. Suppose that both of them can talk about the distance to a tree. Is it reasonable that the two agents should reach different conclusions about the distance just because they have different sensors (and thus use different vocabularies), although they have not made any observations? It would then follow that the agents should change their conclusions if they switched sensors, despite not having made any observations. This does not seem so reasonable!

Bias and representation independence can be viewed as two extremes in a spectrum. If we accept that the knowledge base encodes the user's bias, there is no obligation to be

6. Actually, $i(Q(d)) = Q(d) \wedge Q(c_1) \wedge \ldots \wedge Q(c_{100})$, but the latter is equivalent to $Q(d)$ given KB.
7. They certainly applied to all of the many definitions that we tried!
8. In fact, it suffices to assume that $\mathit{true} \mid\!\sim_I \Pr(\mathit{colorful}) \in [\alpha, \beta]$, as long as $\alpha > 0$ or $\beta < 1$.

invariant under any representation shifts at all. On the other hand, if we assume that 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 *sparrow*. In the next section, we show how to construct inference procedures that are representation independent under a limited class of representation shifts.

6. Selective invariance

As discussed above, we want to construct an inference procedure I that is invariant only under certain embeddings. For the purposes of this section, we restrict attention to finite spaces, where all sets are measurable. That is, we focus on $\mathcal{X}$-inference procedures where $\mathcal{X}$ consists only of measure spaces of the form $(X, 2^X)$, where X is finite.

Our first step is to understand the conditions under which an $\mathcal{X}$-inference procedure I is invariant under a specific X-Y embedding f. When do we conclude θ from $KB \subseteq \Delta_X$? Recall that an inference procedure I_X picks a subset $\mathcal{D}_X = I_X(KB)$, and concludes θ iff θ holds for every measure in $\mathcal{D}_X$. Similarly, when applied to $f^*(KB) \subseteq \Delta_Y$, I_Y 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 measures μ on X and ν on Y. Recall from Proposition 4.7 that μ and ν correspond under f iff, for all formulas θ, we have $\mu \models \theta$ iff $\nu \models f^*(\theta)$. To understand how the correspondence extends to sets of probability measures, consider the following example:

Example 6.1: Consider the embedding f of Example 4.3, and let $\mathcal{D}_X = \{\mu, \mu'\}$ where μ is as above, and $\mu'(\textit{colorful}) = 0.6$. 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 measure ν that does not correspond to either μ or μ', e.g., the measure 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 measures; 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 probability measure in $\mathcal{D}_Y$ must correspond to some measure in $\mathcal{D}_X$. Conversely, every measure in $\mathcal{D}_X$ must correspond to a measure in $\mathcal{D}_Y$. For suppose that there is no measure $\nu \in \mathcal{D}_Y$ 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 measures corresponding to measures in $\mathcal{D}_X$. In particular, we might have $\mathcal{D}_Y$ containing only a single measure 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 $\nu(\textit{colorless}) = 0.3$. ∎

Based on this example, we use the following extension to our definition of correspondence.

Definition 6.2: We say that $\mathcal{D}_X$ and $\mathcal{D}_Y$ *correspond* under f if for all $\nu \in \mathcal{D}_Y$, there exists a corresponding $\mu \in \mathcal{D}_X$ (so that $\mu(S) = \nu(f(S))$ for all $S \subseteq X$), and for all $\mu \in \mathcal{D}_X$, there exists a corresponding $\nu \in \mathcal{D}_Y$. ∎

Proposition 6.3: *Suppose that f is a faithful X-Y embedding, $\mathcal{D}_X \subseteq \Delta_X$, and $\mathcal{D}_Y \subseteq \Delta_Y$. The following two conditions are equivalent:*

(a) $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f;

(b) for all θ, $\mathcal{D}_X \models \theta$ iff $\mathcal{D}_Y \models f^*(\theta)$.[9]

Proof: See Appendix A.4. ∎

To produce an inference procedure that is invariant under some X-Y embedding f, we must ensure that *for every KB*, $I_X(KB)$ and $I_Y(KB)$ correspond. 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 the remainder of 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.

First consider the problem of updating with objective information. The standard way of doing this update is via *conditioning*. For a measure $\mu \in \Delta_X$ and an event $S \subseteq X$, define $\mu|S$ to be the measure that assigns probability $\mu(w)/\mu(S)$ to every $w \in S$, and zero to all other states. For a set of measures $\mathcal{D}_X \subseteq \Delta_X$, define $\mathcal{D}_X|S$ to be $\{\mu|S \ : \ \mu \in \mathcal{D}_X\}$. The following result is easy to show.

Proposition 6.4: *Let $S \subseteq X$ be an event and let f be a faithful X-Y embedding. If μ and ν correspond under f, then $\mu|S$ and $\nu|f(S)$ also correspond under f.*

Proof: Almost immediate from the definitions; left to the reader. (In any case, note that this result follows from Theorem 6.4 below.) ∎

Clearly, the result extends to sets of measures.

Corollary 6.5: *If f is a faithful X-Y embedding, and $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f, then $\mathcal{D}_X|S$ and $\mathcal{D}_Y|f(S)$ also correspond under f.*

What if we want to update on a constraint that is not objective? The standard extension of conditioning to this case is via *relative entropy* or *KL-divergence* (Kullback & Leibler, 1951).

9. While (a) implies (b) for arbitrary spaces, the implication from (b) to (a) depends on the restriction to finite spaces made in this section. For suppose that X is the natural numbers N, f is the identity, $\mathcal{D}_X$ consists of all probability measures on N, and $\mathcal{D}_Y$ consists of all measures but that measure μ_0 such that $\mu_0(n) = 1/2^{n+1}$. If the language consists of finite Boolean combinations of assertions of the form $\Pr(S) \geq \alpha$, for $S \subseteq N$, then it is easy to see that $\mathcal{D}_X \models \theta$ iff $\mathcal{D}_Y \models \theta$ for all formulas θ, but clearly $\mathcal{D}_X$ and $\mathcal{D}_Y$ do not correspond under the identity map.

Definition 6.6: If μ and μ' are measures on X, the *relative entropy between* μ' *and* μ, denoted $KL_X(\mu'\|\mu)$, is defined as $\sum_{x\in X}\mu'(x)\log(\mu'(x)/\mu(x))$. For a measure μ on X and a constraint θ, let $\mu|\theta$ denote the set of measures μ' satisfying θ for which $KL_X(\mu'\|\mu)$ is minimal. ∎

Intuitively, the KL-divergence measures the "distance" from μ' to μ. A measure μ' satisfying θ for which $KL_X(\mu'\|\mu)$ is minimal can be thought of as the "closest" measure to μ that satisfies θ. If θ denotes an objective constraint, then the unique measure satisfying θ for which $KL_X(\mu'\|\mu)$ is minimal is the conditional measure $\mu|\theta$ (Kullback & Leibler, 1951). (That is why we have deliberately used the same notation here as for conditioning.) Moreover, it is easy to show that $KL_X(\mu'\|\mu) = 0$ iff $\mu' = \mu$. It follows that if $\mu \in \theta$, then $\mu|\theta = \mu$.

Given a set of measure $\mathcal{D}_X \subseteq \Delta_X$ and a constraint θ on Δ_X, define $\mathcal{D}_X|\theta$ to be $\cup_{\mu\in\mathcal{D}_X}\mu|\theta$.

We can now apply a well-known result (see, e.g., (Seidenfeld, 1987)) to generalize Proposition 6.4 to the case of relative entropy.

Theorem 6.7: *Let θ be an arbitrary constraint on Δ_X. If f is a faithful X-Y embedding and μ and ν correspond under f, then $\mu|\theta$ and $\nu|f^*(\theta)$ also correspond under f.*

Proof: See Appendix A.4. ∎

Again, this result clearly extends to sets of measures.

Corollary 6.8: *If f is a faithful X-Y embedding, and $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f, then $\mathcal{D}_X|\theta$ and $\mathcal{D}_Y|f^*(\theta)$ also correspond under f.*

These results give us a way to "bootstrap" invariance. We construct an inference procedure that uses relative entropy starting from some set of *prior probability measures*. Intuitively, these encode the user's prior beliefs about the domain. As information comes in, these measures are updated using cross-entropy. If we design the priors so that certain invariances hold, Corollary 6.8 guarantees that these invariances continue to hold throughout the process.

Formally, a *prior function* $\mathcal{P}$ on $\mathcal{X}$ maps $X \in \mathcal{X}$ to a set $\mathcal{P}(X)$ of probability measures in Δ_X. Define an inference procedure $I^{\mathcal{P}}$ by taking $I^{\mathcal{P}}_X(KB) = \mathcal{P}(X)|KB$. Note that $I^{\mathcal{P}}_X(\mathit{true}) = \mathcal{P}(X)$, 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 $I^{\mathcal{P}}$ for $\mathcal{P}(X) = \Delta_X$. (This is because, as observed earlier, $\mu|KB = \mu$ if $\mu \in KB$.) Standard Bayesian conditioning (defined for objective knowledge bases) is of this form, where we take $\mathcal{P}(X)$ to be a single measure for each space X. More interestingly, it is well known (Kullback & 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? Using Proposition 6.3 and Corollary 6.5, it is easy to show that if we want $\mathcal{I}^{\mathcal{P}}$ to be invariant under some set $\mathcal{F}$ of embeddings, then we must ensure that the prior function has the right correspondence property.

Theorem 6.9: *If f is a faithful X-Y embedding, then $I^{\mathcal{P}}$ is invariant under f iff $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ correspond under f.*

Proof: See Appendix A.4. ∎

Theorem 6.9 sheds some light on the maximum entropy inference procedure. As we mentioned, $|\!\sim_{me}$ is precisely the inference procedure based on the prior function $\mathcal{P}_u$. The corollary asserts that $|\!\sim_{me}$ 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 the 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 Theorem 6.9. 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 Shore and Johnson's (1980) axiomatic characterization of maximum-entropy. (We remark that Paris (1994, Theorem 7.10) proves that maximum entropy satisfies a variant of his atomicity principle; his invariance result is essentially a special case of Theorem 6.9.)

If we do not like the behavior of maximum entropy under representation shifts, Theorem 6.9 provides a solution: we should simply start out with a different prior function. If we want to maintain invariance under all representation shifts, $\mathcal{P}(X)$ must include all non-extreme priors (i.e., all the measures μ on X such that $\mu(A) \notin \{0, 1\}$ for all A such that $A \notin \{\emptyset, X\}$). This set of priors gives essential 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 $\mathcal{P}$ that is "closed" under each $f \in \mathcal{F}$, i.e., for each measure $\mu \in \mathcal{P}(X)$ and each X-Y embedding $f \in F$ we make sure that there is a corresponding measure $\nu \in \mathcal{P}(Y)$, and vice versa. Thus, 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. Suppose that 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 6.9 gives us the answer.

We begin by providing one plausible formulation of the desired reasoning pattern. For a finite space X, we say that $X_1 \times \cdots \times X_n$ is the *product decomposition* of X if $X = X_1 \times \cdots \times X_n$ and n is the largest number for which X can be written as a product in this way. (It is easy to see that if X is finite, then this "maximal" product decomposition is unique.) A measure $\mu \in \Delta_X$ is a *product measure on* X if $X_1 \times \cdots \times X_n$ is the product decomposition of X and there exist measures $\mu_i \in \Delta_{X_i}$ for $i = 1, \ldots, n$ such that $\mu = \mu_1 \times \cdots \times \mu_n$, that is, $\mu(U_1 \times \cdots \times U_n) = \prod_{i=1}^{n} \mu_i(U_i)$, if $U_i \subseteq X_i$, $i = 1, \ldots, n$. Let $\mathcal{P}_\Pi$ be the set of all product measures on X. If $\mathcal{P}_\Pi$ is the prior and the relative entropy rule is used to update the prior given a knowledge base, then $|\!\sim_{\mathcal{P}_\Pi}$ satisfies a form of minimal default independence. In fact, it is easy to show that it satisfies the following stronger property.

Proposition 6.10: *Suppose that $X_1 \times \cdots \times X_n$ is the product decomposition on X and, for each $i = 1, \ldots, n$, KB_i is a constraint on X_i, and S_i is a subset of X_i. Then*

$$\bigwedge_{i=1}^{n} KB_i \mathrel{|\!\!\sim}_{I_{\mathcal{P}_\Pi}} \Pr(S_1 \wedge \ldots \wedge S_n) = \prod_{i=1}^{n} \Pr(S_i).$$

Proof: See Appendix A.4. ∎

Theorem 4.16 shows that $\mathrel{|\!\!\sim}_{\mathcal{P}_\Pi}$ cannot be invariant under all embeddings. Theorem 6.9 tells us that it is invariant under precisely those embeddings for which $\mathcal{P}_\Pi$ is invariant. These embeddings can be characterized syntactically in a natural way. Suppose that $\Phi_1, \ldots, \Phi_n$ is a partition of a finite set Φ of primitive propositions. Note that a truth assignment to the primitive propositions in Φ can be viewed as a "crossproduct" of truth assignments to the primitive propositions in $\Phi_1, \ldots, \Phi_n$. Under this identification, suppose that a set X of truth assignments to Φ is decomposed as $X_1 \times \cdots \times X_n$, where X_i consists of truth assignments to Φ_i. In that case, if $p \in \Phi_j$ and $q, r \in \Phi_k$ for some $j \neq k$, then *true* $\mathrel{|\!\!\sim}_{\mathcal{P}} \Pr(p \wedge q) = \Pr(p) \times \Pr(q)$, but since since q and r are in the same subset, we do not have *true* $\mathrel{|\!\!\sim}_{\mathcal{P}} \Pr(r \wedge q) = \Pr(r) \times \Pr(q)$. Hence, $\mathcal{P}_\Pi$ is not invariant under an interpretation i that maps p to r and maps q to itself. Intuitively, the problem is that i is "crossing subset boundaries"; it is mapping primitive propositions that are in different subsets to the same subset. If we restrict to interpretations that"preserve subset boundaries", then we avoid this problem.

We can get a semantic characterization of this as follows. If the product decomposition of X is $X_1 \times \cdots \times X_n$ and the product decomposition of Y is $Y_1 \times \cdots \times Y_n$, then f is an X-Y *product embedding* if f is an X-Y embedding and there are X_i-Y_i embeddings f_i, $i = 1, \ldots, n$, and $f(\langle x_1, \ldots, x_n \rangle) = f_1(x_1) \times \cdots \times f_n(x_n)$. Product embeddings capture the intuition of preserving subset boundaries; elements in a given subset X_i remain in the same subset (Y_i) after the embedding. However, the notion of product embedding is somewhat restrictive; it requires that elements in the ith subset of X map to elements in the ith component of Y, for $i = 1, \ldots, n$. We can still preserve default independence if the components of a product are permuted. An g is a *permutation embedding* if there exists a permutation π of $\{1, \ldots, n\}$ such that $g(\langle x_1, \ldots, x_n \rangle) = \langle x_{\pi(1)}, \ldots, x_{\pi(n)} \rangle$.

Theorem 6.11: *The inference procedure $I_{\mathcal{P}_\Pi}$ is invariant under faithful product embeddings and under permutation embeddings.*

Theorem 6.9 thus provides us with the basic tools to easily define an inference procedure that enforces minimal default independence for constraints involving disjoint parts of the language, while at the same time guaranteeing invariance under a large and natural class of embeddings. Given our negative result in Theorem 4.16, this type of result is the best that we could possibly hope for. In general, Theorem 6.9 provides us with a principled framework for controlling the tradeoff between the strength of the conclusions that can be reached by an inference procedure and invariance under representation shifts.

7. Related Work

As we mentioned earlier, there are two types of probabilistic inference. We partition our discussion of related work along those lines.

7.1 Probabilistic Inference from a Knowledge Base

Given the importance of representation in 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 that has focused on representation independence in the logical sense that we have considered here prior to ours is that of Salmon and Paris.

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 (1963) then attempted to define a modified inductive inference method that would satisfy his criterion but it is not clear that this attempt succeeded. In any case, our results show that his modified method certainly cannot be representation independent in our sense.

As we said earlier, Paris (1994) considers inference processes, which given a constraint on Δ_X, choose a unique measure satisfying the constraint. He then considers various properties that an inference process might have. Several of these are closely related to properties that we have considered here. (In describing these notions, we have made some inessential changes so as to be able to express them in our notation.)

- An $\mathcal{X}$-inference process I is *language invariant* if all $X, Y \in \mathcal{X}$ and all constraints KB and φ on Δ_X, we have that $KB \mid\!\sim_{I_X} \varphi$ iff $KB \mid\!\sim_{I_{X\times Y}} \varphi$. Clearly language invariance is a special case of robustness. Paris shows that a *center of mass* inference process (that, given a set $A \subseteq \Delta_X$, chooses the measure that is the center of mass of A) is not language invariant; on the other hand, it is well known that maximum entropy is language invariant.

- An $\mathcal{X}$-inference process I satisfies the *principle of irrelevant information* if for all spaces $X, Y \in \mathcal{X}$, constraints KB and φ on Δ_X, and constraints ψ on Δ_Y, we have $KB \mid\!\sim_{I_X} \varphi$ iff $KB \wedge \psi \mid\!\sim_{I_{X\times Y}} \varphi$. Again, this is a special case of robustness, since a constraint ψ on Δ_Y must be X-conservative. Paris shows that maximum entropy satisfies this principle. (He restricts the domain of the maximum entropy process to closed convex sets, so that there is always a unique probability measure that maximizes entropy.)

- An $\mathcal{X}$-inference process I satisfies the *renaming principle* if, whenever X and Y are finite spaces, $g : X \to Y$ is an isomorphism, and $f : 2^X \to 2^Y$ is the faithful embedding based on g (in that $f(S) = \{g(s) : s \in S\}$), then for all constraints KB and θ on Δ_X, we have $KB \mid\!\sim_{I_X} \theta$ iff $f^*(KB) \mid\!\sim_{I_Y} f^*(\theta)$. Clearly, the renaming principle is a special case of representation independence. Paris shows that a number of inference processes (including maximum entropy) satisfy the renaming principle.

- An $\mathcal{X}$-inference process I satisfies the *principle of independence* if, whenever X, Y, and Z are in $\mathcal{X}$, $S \in \mathcal{F}_X$, $T \in \mathcal{F}_Y$, $U \in \mathcal{F}_Z$, and KB is the constraint $\Pr(U) = a \wedge \Pr(S|U) = b \wedge \Pr(T|U) = c$, where $a > 0$, then $KB \mid\!\sim \Pr(S \times T|U) = bc$. Ignoring the conditional probabilities, this is clearly a special case of minimal default independence. Paris and Vencovska (1990) show that maximum entropy is the unique inference process satisfying a number of principles, including renaming, irrelevant information, and independence.

- An $\mathcal{X}$-inference process I satisfies the *atomicity principle* if, for all $X, Y_1, \ldots, Y_n$ in $\mathcal{X}$, whenever f' is an embedding from $\{0,1\}$ to X, and f is the obvious extension of f' to an embedding from to $\{0,1\} \times Y_1 \times \ldots \times Y_n$ to $X \times Y_1 \times \ldots \times Y_n$, then for all constraints KB and θ on $\Delta_{\{0,1\} \times Y_1 \times \ldots \times Y_n}$, we have $KB \mathrel{|\!\!\sim}_{I_X} \theta$ iff $f^*(KB) \mathrel{|\!\!\sim}_{I_Y} f^*(\theta)$. Clearly atomicity is a special case of representation independence. Paris shows that there is no inference process that satisfies atomicity. The argument is similar in spirit to that used to prove Theorems 4.11 and 4.16, but much simpler, since inference processes return a unique probability measure, not a set of them.

More recently, Jaeger (1996), building on our definitions, has examined representation independence for general nonmonotonic logics. He considers representation independence with respect to a collection of transformations, and proves results about the degree to which certain nonmonotonic formalisms, such as *rational closure* (Lehmann & Magidor, 1992), satisfy representation independence.

Another line of research that is relevant to representation independence is the work on *abstraction* (Giunchiglia & Walsh, 1992; Nayak & Levy, 1995). 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.

7.2 Bayesian Probabilistic Inference

Bayesian statistics takes a very different perspective on the issues we discuss in this paper. As we discussed, the Bayesian approach generally assumes that we construct a prior, and use standard probabilistic conditioning to update that prior as new information is obtained. In this approach, the representation of the knowledge obtained has no effect on the conclusions. Two pieces of information that are semantically equivalent (denote the same event) will have precisely the same effect when used to condition a distribution.

In this paradigm, our analysis is more directly related to the step that precedes the probabilistic conditioning—the selection of the prior. When we have very specific beliefs that we want to encode in a prior distribution (as we do, for example, when constructing a Bayesian network), we design our prior to reflect these beliefs in terms of the vocabulary used. For example, if we have a particular distribution in mind over the location of an object, we will encode it one way when representing the space in terms of Cartesian coordinates, and in another way when using polar coordinates. In effect, we can view the representation transformation as an embedding f, and the two priors as corresponding under f, in the sense of Definition 4.2. Thus, the design of the prior already takes the representation into account.

On the other hand, when we are trying to construct an "uninformed" prior for some class of problems, the issue of representation independence becomes directly relevant. Indeed, most of the standard problems with maximum entropy arise even in the simple case when we simply do Bayesian conditioning starting with a uniform prior over our space.

A standard approach in Bayesian statistics is to use the invariance under certain transformations in order to define an appropriate uninformed prior. For example, we might want a prior over images that is invariant to rotation and translation. In certain cases, once we specify the transformation under which we want a measure to be invariant, the measure is uniquely determined (Jaynes, 1968; Kass & Wasserman, 1993). In this case, the argument goes, the uniquely determined measure is perforce the "right" one. This idea of picking a prior using its invariance properties is in the same spirit as the approach we take in Section 6. Indeed, as this approach simply uses standard probabilistic conditioning for objective information (such as observations), the Bayesian approach with an uninformed prior invariant to a set of embeddings is, in a sense, a special case. However, our approach does not force us to choose a unique prior. Rather, we allow the use of a set of prior distributions, allowing us to explore a wider spectrum of inference procedures.

This approach is also related to the work of Walley (1996), who observes that representation independence is an important desideratum in certain statistical applications involving multinomial data. Walley proposes the use of sets of Dirichlet densities to encode ignorance about a prior, and shows that this approach is representation independent in its domain of application.

8. 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 drawing many standard default conclusions, and defining limited notions of invariance that might that allow a compromise between the desiderata of being able to draw interesting conclusions (not already entailed by the evidence) 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. As we mentioned, Jaeger (1996) has obtained some results on representation independence in a more general setting, but there is clearly much more that can be done. More generally, it would be of interest to understand better the tension between representation independence and the strength of conclusions that can be drawn from an inference procedure.

Acknowledgments

Thanks to Ed Perkins for pointing us to (Keisler & Tarski, 1964) and, in particular, the result that a countably additive probability measure defined on a subalgebra of an algebra $\mathcal{F}$ could not necessarily be extended to a countably additive probability measure on $\mathcal{F}$. Thanks to the reviewers of the paper for their perceptive comments and for pointing out (Horn & Tarski, 1948). Much of Halpern's work on the paper was done while he was at the IBM Almaden Research Center. His recent work has been supported by NSF under grant IRI-96-25901 and IIS-0090145 and ONR under grant N00014-01-1-0795. Some of Koller's work was done at U.C. Berkeley. Her research was 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. Daphne Koller's later work on the paper was supported through the generosity of the Powell foundation, and by ONR grant N00014-96-1-0718. A preliminary version of this appears in *Proceedings of IJCAI '95*, pp. 1853–1860.

Appendix A. Proofs

A.1 Proofs for Section 2

Proposition 2.5: *Let $\mathrel{|\!\sim}$ be a relation on probabilistic constraints on X for which the properties Reflexivity, Left Logical Equivalence, Right Weakening, Infinitary And, and Consistency hold for all KB in the domain of $\mathrel{|\!\sim}$. (That is, if KB is in the domain of $\mathrel{|\!\sim}$, in that $KB \mathrel{|\!\sim} \theta$ for some θ, then $KB \mathrel{|\!\sim} KB$, and so on.) Then $\mathrel{|\!\sim}$ is $\mathrel{|\!\sim}_I$ for some X-inference procedure I.*

Proof: Define I as follows. If $A \subseteq \Delta_X$, KB is in the domain of $\mathrel{|\!\sim}$, and $A = [\![KB]\!]_X$ for some statement KB, then A is in the domain of I and $I(A) = \cap\{[\![\theta]\!]_X : KB \mathrel{|\!\sim} \theta\}$. Note that by Left Logical Equivalence, this is well defined, since if $A = [\![KB']\!]_X$, then $\cap\{[\![\theta]\!]_X : KB \mathrel{|\!\sim} \theta\} = \cap\{[\![\theta]\!]_X : KB' \mathrel{|\!\sim} \theta\}$. If $A \neq [\![KB]\!]_X$ for some statement KB, then A is not in the domain of I. It remains to check that I is an X-inference procedure (i.e., that $I(A) \subseteq A$ and that $I(A) = \emptyset$ iff $A = \emptyset$ for all A in the domain of I), and that $\mathrel{|\!\sim} \; = \; \mathrel{|\!\sim}_I$. To check that I is an X-inference procedure, suppose that A is in the domain of I. Thus, $A = [\![KB]\!]_X$ By Reflexivity, it easily follows that $I([\![KB]\!]_X) \subseteq [\![KB]\!]_X$. Next suppose that $I([\![KB]\!]_X) = \emptyset$. It follows that $\cap\{[\![\theta]\!]_X : KB \mathrel{|\!\sim} \theta\} = \emptyset$. Thus, $\{\theta : KB \mathrel{|\!\sim} \theta\} \models \mathit{false}$. By the Infinitary AND rule, we must have $KB \mathrel{|\!\sim}_I \mathit{false}$. By the Consistency Rule, it follows that $[\![KB]\!]_X = \emptyset$. Thus, I is indeed an X-inference procedure. Finally, note that if $KB \mathrel{|\!\sim} \psi$ then, by definition of I, $I([\![KB]\!]_X) \subseteq [\![\psi]\!]_X$, so $KB \mathrel{|\!\sim}_I \psi$. For the opposite inclusion, note that if $KB \mathrel{|\!\sim}_I \psi$, then $\{\theta : KB \mathrel{|\!\sim} \theta\} \models \psi$. Thus, by the Infinitary And rule, it follows that $KB \mathrel{|\!\sim}_I \psi$. ∎

A.2 Proofs for Section 3

To prove Theorem 3.4, we need the following lemma.

Lemma A.1: *Given two spaces X_0 and X_1, measures $\mu^0 \in \Delta_{(X_0, \mathcal{F}_{X_0})}$ and $\mu^1 \in \Delta_{(X_1, \mathcal{F}_{X_1})}$, and subsets $S_0 \in \mathcal{F}_{X_0}$ and $S_1 \in \mathcal{F}_{X_1}$ such that $\mu^0(S_0) = \mu^1(S_1)$, there exists a measure $\mu^2 \in \Delta_{(X_0 \times X_1, \mathcal{F}_{X_0 \times X_1})}$ such that $\mu^2_{X_i} = \mu^i$, for $i = 1, 2$, and $\mu^2(S_0 \Leftrightarrow S_1) = 1$.*[10]

Proof: For $A \times B \in \mathcal{F}_{X_0} \times \mathcal{F}_{X_1}$, define

$$\mu^2(A \times B) = (\mu^0(A \cap S_0)\mu^1(B \cap S_1)/\mu^1(S_1)) + (\mu^0(A \cap \overline{S_0})\mu^1(B \cap \overline{S_1})/\mu^1(\overline{S_1})),$$

where we take $\mu^0(A \cap S_0)\mu^1(B \cap S_1)/\mu^1(S_1) = 0$ if $\mu^1(S_1) = 0$ and take $\mu^0(A \cap \overline{S_0})\mu^1(B \cap \overline{S_1})/\mu^1(\overline{S_1}) = 0$ if $\mu^1(\overline{S_1}) = 0$. Extend to disjoint unions of such sets by additivity. Since

10. If A and B are sets, we use the notation $A \Leftrightarrow B$ to denote the set $(A \cap B) \cup (\overline{A} \cap \overline{B})$.

all sets in $\mathcal{F}_{X_0 \times X_1}$ can be written as disjoint unions of sets of the form $A \times B \in \mathcal{F}_{X_0} \times \mathcal{F}_{X_1}$, this suffices to define μ^2. To see that μ^2 is actually a measure, note that $\mu^2(X \times Y) = \mu^0(S_0) + \mu^0(\overline{S_0}) = 1$. Additivity is clearly enforced by the definition. Finally, to see that μ^2 has the desired properties, suppose that $\mu^1(S_1) \neq 0$ and $\mu^1(\overline{S_1}) \neq 0$. (The argument is easier if this is not the case; we leave details to the reader.) Then

$$\begin{aligned} \mu^2_{X_0}(A) &= \mu^2(A \times Y) = \mu^0(A \cap S_0)\mu^1(S_1)/\mu^1(S_1) + \mu^0(A \cap \overline{S_0}\mu^1(\overline{S_1})/\mu^1(\overline{S_1}) \\ &= \mu^0(A \cap S_0) + \mu^0(A \cap \overline{S_0}) = \mu^0(A). \end{aligned}$$

Since $\mu^0(S_0) = \mu^1(S_1)$ by assumption (and so $\mu^0(\overline{S_0}) = \mu^1(\overline{S_1})$),

$$\begin{aligned} \mu^2_{X_1}(B) &= \mu^2(X \times B) = \mu^0(S_0)\mu^1(B \cap S_1)/\mu^1(S_1) + \mu^0(\overline{S_0})\mu^1(B \cap \overline{S_1})/\mu^1\overline{S_1}) \\ &= \mu^1(B \cap S_1) + \mu^1(B \cap \overline{S_1}) = \mu^1(B). \end{aligned}$$

This completes the proof. ∎

Theorem 3.4: *If $\{I_X : X \in \mathcal{X}\}$ is a robust $\mathcal{X}$-inference procedure that satisfies DI1, DI2, and DI3, then I_X is essentially entailment for all $X \in \mathcal{X}$.*

Proof: Suppose that $\{I_X : X \in \mathcal{X}\}$ is robust and I_X is not essentially entailment for $X \in \mathcal{X}$. Then there must be a constraint KB on Δ_X and a set $S \in \mathcal{F}_X$ such that $KB \mathrel{|\!\!\sim}_I \alpha < \Pr(S) < \beta$ and $KB \not\models \alpha \leq \Pr(S) \leq \beta$. Thus, there must be some $\gamma \notin [\alpha, \beta]$ such that $KB \wedge \Pr(S) = \gamma$ is consistent. We can assume without loss of generality that $\gamma < \alpha$ (otherwise we can replace S by $\overline{S}$).

We first construct a space $Y_0 \in \mathcal{X}$ that has subsets $U_1, \ldots, U_n$ with the following properties:

(a) There is no measure $\mu \in \Delta_{Y_0}$ such that $\mu(U_i) > \alpha$, for all $i = 1, ..., n$.

(b) For each i, there is some measure $\mu'_i \in \Delta_{Y_0}$ such that $\mu'_i(U_i) = 1$ and $\mu'_i(U_j) > \gamma$ for all $j \neq i$.

We proceed as follows. Choose n and d such that $\gamma < (d-1)/(n-1) < d/n < \alpha$. By assumption, there exists a $Y_0 \in \mathcal{X}$ such that $|Y_0| = n!/(n-d)!$. Without loss of generality, we can assume that Y_0 consists of all tuples of the form $(a_1, \ldots, a_d)$, where the a_i's are all distinct, and between 1 and n. Let U_i be consist of all the tuples in Y_0 that have i somewhere in the subscript; it is easy to see that there are $d(n-1)!/(n-d)!$ such tuples. Suppose that μ is a probability measure in Δ_{Y_0}. It is easy to see that $\mu(U_1) + \cdots + \mu(U_n) = d$, since each tuple in Y_0 is in exactly d of the U_i's and so gets counted exactly d times, and the sum of the probabilities of the tuples is 1. Thus, we cannot have $\mu(U_i) > d/n$ for all i (and, *a fortiori*, we cannot have $\mu(U_i) > \alpha$ for all i). This takes care of the first requirement. Next, consider a probability distribution μ'_i that makes all the tuples making up U_i equally probable, and gives all the other tuples probability 0. Then it is easy to see that $\mu'_i(U_i) = 1$. Moreover, since it is straightforward to check that there are exactly $d(d-1)(n-2)!/(n-d)!$ tuples in $U_i \cap U_j$ for $j \neq i$, we have $\mu'_i(U_j) = [d(d-1)(n-2)!/(n-d)!]/[d(n-1)!/(n-d)!] = (d-1)/(n-1)$. This takes care of the second requirement.

By assumption, there is also a measurable space $Y \in \mathcal{X}$ such that $|Y| = 2$. Suppose that $Y = \{y, y'\}$. Let $Z = X^n \times Y_0 \times Y^n$, where the n is the same as the n chosen in the construction of Y_0. Again, by assumption, $Z \in \mathcal{X}$. For $i = 1, \ldots, n$,

- if $A \subseteq X$, let $A_i = X^{i-1} \times A \times X^{n-i} \times Y_0 \times Y^n \subseteq Z$.
- let $KB_i = \{\mu \in \Delta_Z : \mu_{X_i} \in KB\}$;
- let Y_i be the subset of Y^n where the ith copy of Y is replaced by $\{y\}$;
- let V_i be the subset of Z of the form $X^n \times U_i \times Y_i$ (where $U_1, \ldots, U_n$ are the subsets of Y_i constructed above).

Let σ be the following constraint on Δ_Z:

$$KB_1 \wedge \ldots \wedge KB_n \wedge \Pr(S_1 \Leftrightarrow V_1) = 1 \wedge \ldots \wedge \Pr(S_n \Leftrightarrow V_n) = 1.$$

Let X_i denote the ith copy of X in Z. That is, for ease of exposition, we view Z as being of the form $X_1 \times \cdots \times X_n \times Y_0 \times Y^m$, although all the X_i's are identical, since it is helpful to be able to refer to a specific X_i. We claim that σ is X_i-conservative over KB, for $i = 1, \ldots, n$. Thus, we must show that $\mathit{proj}_{X_i}([\![KB_i \wedge \sigma]\!]_Z) = [\![KB]\!]_X$. It is immediate that $\mathit{proj}_{X_i}([\![KB_i \wedge \sigma]\!]_Z) \subseteq [\![KB]\!]_X$. For the opposite inclusion, suppose that $\nu \in [\![KB]\!]_X$. We must show that there exists some $\mu \in [\![KB_i \wedge \sigma]\!]_Z$ such that $\mu_{X_i} = \nu$. We proceed as follows.

Let μ'_0 be a measure in Δ_{Y_0} such that $\mu'_0(U_i) = 1$ and $\mu'_0(U_j) > \gamma$, for $j \neq i$. By construction of the U_j's, such a measure must exist. For $j \in \{1, \ldots, n\}$, let μ'_j be the measure in Δ_Y such that $\mu'_i(y) = \nu(S)$ and if $j \neq i$, then $\mu'_j(y) = \gamma/\mu'_0(U_j)$ (and $\mu'_j(y') = 1 - \mu'_j(y)$). Let μ' be the measure on $Y_0 \times Y^n$ that is the "crossproduct" of $\mu'_0, \mu'_1, \ldots, \mu'_n$. That is, $\mu'(T_0 \times \cdots \times T_n) = \mu'_0(T_0) \times \cdots \times \mu'_n(T_n)$. By construction, $\mu'(V_j) = \gamma$ for $j \neq i$ and $\mu'(V_i) = \nu(S)$.

By assumption, there is a measure $\nu_0 \in \Delta_X$ such that $\nu_0 \models KB \wedge \Pr(S) = \gamma$. We now proceed inductively to define a measure $\mu^k \in \Delta_{X^k \times Y_0 \times Y^n}$ such that (a) $\Pr((S_1 \Leftrightarrow V_1) \cap \ldots \cap (S_k \Leftrightarrow V_k)) = 1$, (b) $\mu^j_Y = \mu'$ and $\mu^j_{X_j} = \nu$ for $j = 1, \ldots, k$. We define $\mu^0 = \mu'$. For the inductive step, we simply apply Lemma A.1. Finally, we take μ to be μ^n. Our construction guarantees that $\mu_{X^j} = \nu$, hence that $\mu \models KB_j$. In addition, the construction guarantees that $\mu \models \Pr(S_1 \Leftrightarrow V_1) = 1 \wedge \ldots \wedge \Pr(S_n \Leftrightarrow V_n) = 1$. Hence $\mu \models \sigma$, as desired.

It follows from DI1, DI2, and DI3 that σ is in the domain of I_Z. Since $KB_i \wedge \sigma$ is equivalent to σ, it follows that $KB_i \wedge \sigma$ is also in the domain of I_Z. Now, by robustness, for any constraint φ on Δ_{X_i}, we have $KB_i \wedge \sigma \mathrel{|\!\sim}_I \varphi$ iff $KB_i \mathrel{|\!\sim}_I \varphi$. Since $KB_i \mathrel{|\!\sim}_I \Pr(S_i) > \alpha$ and $KB_i \wedge \sigma$ is equivalent to σ, it follows that $\sigma \mathrel{|\!\sim}_I \Pr(S_i) > \alpha$ for $i = 1, \ldots, n$. By the And rule (Proposition 2.3), it follows that $\sigma \mathrel{|\!\sim}_I \Pr(S_1) > \alpha \wedge \ldots \wedge \Pr(S_n) > \alpha$. Since $\sigma \models \Pr((S_1 \Leftrightarrow V_1) \cap (S_n \Leftrightarrow V_n)) = 1$, it easily follows that $\sigma \mathrel{|\!\sim}_I \Pr(U_1) > \alpha \wedge \ldots \wedge \Pr(U_n) > \alpha$. But our construction guarantees that $\Pr(U_1) > \alpha \wedge \ldots \wedge \Pr(U_n) > \alpha$ is inconsistent. Thus, $\sigma \mathrel{|\!\sim}_I \mathit{false}$. By robustness, it follows that $KB_i \mathrel{|\!\sim}_I \mathit{false}$. But this can happen only if $KB \models \mathit{false}$, which implies that $KB \models \alpha \leq \Pr(S) \leq \beta$, contradicting our original assumption. ∎

A.3 Proofs for Section 4

To prove Lemma 4.6, it is useful to first prove two additional results:

Lemma A.2: *If f is an X-Y embedding, then $f(X) = Y$ and $f(\emptyset) = \emptyset$.*

Proof: Suppose that f is an X-Y embedding. We first show that $f(\emptyset) = \emptyset$. From the definition of embedding, it follows that $f(\emptyset) = f(X \cap \emptyset) = f(X) \cap f(\emptyset)$. Thus, $f(\emptyset) \subseteq f(X)$. But the definition of embedding also implies that $f(\emptyset) = f(\overline{X}) = \overline{f(X)}$. Thus, we have $\overline{f(X)} \subseteq f(X)$. This can happen only if $f(X) = Y$ and $f(\emptyset) = \overline{f(X)} = \emptyset$. ∎

Lemma A.3: *If f is a faithful X-Y embedding, then*

(a) for any $\mu \in \Delta_X$, there is a measure $\nu \in \Delta_Y$ such that ν corresponds to μ;

(b) for any $\nu \in \Delta_Y$, there is a measure $\mu \in \Delta_X$ such that μ corresponds to ν.

Proof: To prove (a), consider the algebra of subsets of Y the form $f(S)$, for $S \in \mathcal{F}_X$. Define a function ν' on the algebra via $\nu'(f(S)) = \mu(S)$. This mapping is well defined, for if $f(S) = f(T)$, then faithfulness guarantees that $S = T$. Moreover, ν' is a probability measure on the algebra. To see this, note by Lemma A.2 that $\nu'(Y) = \nu'(f(X)) = \mu(X) = 1$. Moreover, if $f(S) \cap f(T) = \emptyset$, then (by definition of embedding) $f(S \cap T) = \emptyset$ and so, since f is faithful, $S \cap T = \emptyset$ (for otherwise $f(S \cap T) = f(\emptyset)$ by Lemma A.2, but $S \cap T \neq \emptyset$). Thus,

$$\nu'(f(S) \cup f(T)) = \nu'(f(S \cup T)) = \mu(S \cup T) = \mu(S) + \mu(T) = \nu'(f(S)) + \nu'(f(T)).$$

As shown by Horn and Tarski (1948), it is possible to extend ν' to a probability measure ν on $\mathcal{F}_Y$.[11] By construction, we have that ν corresponds to μ.

To prove (b), we use a very similar process. Define a function μ on the algebra of sets $S \subseteq X$ via $\mu(S) = \nu(f(S))$. It is easy to see that μ is already a probability measure in Δ_X, which by construction corresponds to ν. ∎

We can now prove Lemma 4.6.

Lemma 4.6: *An X-Y embedding f is faithful if and only if for all constraints KB and θ, we have $KB \models \theta$ iff $f^*(KB) \models f^*(\theta)$.*

Proof: Suppose that f is faithful. To show that $KB \models \theta$ iff $f^*(KB) \models f^*(\theta)$, we must show that $[\![KB]\!]_X \subseteq [\![\theta]\!]_X$ iff $[\![f^*(KB)]\!]_Y \subseteq [\![f^*(\theta]\!]_Y$. The "only if" direction is immediate from the definition of f^*. To prove the "if" direction, suppose not. Then there must exist some $\mu \in [\![KB]\!]_X - [\![\theta]\!]_X$ such that $f^*(\mu) \subseteq [\![f^*(\theta)]\!]_Y$. Let ν be some probability measure that corresponds to μ. Since $\nu \in f^*(\mu) \subseteq f^*(\theta)$, there must be some $\mu' \in [\![\theta]\!]_X$ such that $\nu \in f^*(\mu')$. Since $\mu' \neq \mu$, there must be some $S \in \mathcal{F}_X$ such that $\mu'(S) \neq \mu(S)$. Since $\nu \in f^*(\mu) \cap f^*(\mu')$, we must have both $\nu(f(S)) = \mu(S)$ and $\nu(f(S)) = \mu'(S)$. But this is a contradiction. This completes the proof of the "if" direction.

For the converse, suppose we have $KB \models \theta$ iff $f^*(KB) \models f^*(\theta)$ for all KB and θ. Given $S, T \in \mathcal{F}_X$, we have the following chain of equivalences:

11. It is critical for this result that we are working with finitely additive measures. There may not be a countably additive measure ν extending ν', even if ν' is countably additive. For example, take $\mathcal{F}'_Y$ to be the Borel sets on $[0, 1]$ and take $\mathcal{F}_Y$ to be all subsets of $[0, 1]$. Let ν' be Lebesgue measure. It is known that, under the continuum hypothesis, there is no countably additive measure extending ν' defined on all subsets of $[0, 1]$ (Ulam, 1930) (see (Keisler & Tarski, 1964) for further discussion).

$S \subseteq T$
iff $\Pr(S) = 1 \models \Pr(T) = 1$
iff $f^*(\Pr(S) = 1) \models f^*(\Pr(T) = 1)$ (by assumption)
iff $\Pr(f(S)) = 1 \models \Pr(f(T)) = 1$ (by definition of f^*)
iff $f(S) \subseteq f(T)$.

Thus, f is faithful. ∎

Proposition 4.7: *Let f be a faithful X-Y embedding. Then the following statements are equivalent:*

(a) μ and ν correspond under f;

(b) for all formulas θ, $\mu \models \theta$ iff $\nu \models f^*(\theta)$.

Proof: We first show that (a) implies (b). So suppose that μ and ν correspond under f. The only if direction of (b) is trivial: If $\mu \models \theta$ then $\nu \in f^*(\mu) \subseteq f^*(\theta)$, since f is faithful. For the if direction, we proceed much as in the proof of Lemma 4.6. Assume that $\nu \models f^*(\theta)$ but that $\mu \not\models \theta$. Since $\nu \in f^*(\theta)$, by definition of f^* there must be some $\mu' \in [\![\theta]\!]_X$ such that $\nu \in f^*(\mu')$. Since $\mu' \models \theta$ whereas $\mu \not\models \theta$, we must have $\mu \neq \mu'$. Hence, there must be some $S \in \mathcal{F}_X$ such that $\mu(S) \neq \mu'(S)$. Since $\nu \in f^*(\mu) \cap f^*(\mu')$, it follows that $\nu(f(S)) = \mu(S)$ and that $\nu(f(S)) = \mu'(S)$, which gives the desired contradiction.

We now show that (b) implies (a). Assume by contradiction that μ and ν do not correspond under f. Then there must be some event $S \in \mathcal{F}_X$ such that $\mu(S) \neq \nu(f(S))$. Let $p = \mu(S)$ and let θ be the constraint $\Pr(S) = p$. Then $\mu \models \theta$, whereas $\nu \not\models f^*(\theta)$, providing the desired contradiction. ∎

Theorem 4.10: *If an $\mathcal{X}$-inference procedure is robust that satisfies DI2, DI4, and DI5, then it is representation independent.*

Proof: Suppose that $\{I_X : X \in \mathcal{X}\}$ is a robust $\mathcal{X}$-inference procedure. We want to show that it is representation independent. So suppose KB, φ are constraints on Δ_X and f is an X-Y embedding, for some $X, Y \in \mathcal{X}$. We want to show that $KB \mathrel{|\!\sim}_{I_X} \varphi$ iff $f^*(KB) \mathrel{|\!\sim}_{I_Y} f^*(\varphi)$.

Let ψ be the following constraint on $\Delta_{X \times Y}$:

$$(\varphi \Leftrightarrow f^*(\varphi)) \wedge (KB \Leftrightarrow f^*(KB)).$$

We claim that ψ is X-conservative over KB and Y-conservative over $f^*(KB)$. Thus, we must show that $proj_X([\![KB \wedge \psi]\!]_{X \times Y}) = [\![KB]\!]_X$ and $proj_Y([\![f^*(KB) \wedge \psi]\!]_{X \times Y}) = [\![f^*(KB)]\!]_Y$. We show that $proj_X([\![KB \wedge \psi]\!]_{X \times Y}) = [\![KB]\!]_X$ here; the argument that $proj_Y([\![f^*(KB) \wedge \psi]\!]_{X \times Y}) = [\![f^*(KB)]\!]_Y$ is almost identical.

Clearly if $\mu \in [\![KB \wedge \psi]\!]_{X \times Y}$ then $\mu_X \in [\![KB]\!]_X$, so $proj_X([\![KB \wedge \psi]\!]_{X \times Y}) \subseteq [\![KB]\!]_X$. For the opposite inclusion, suppose that $\nu \in [\![KB]\!]_X$. We want to find a measure $\nu' \in [\![KB \wedge \psi]\!]_{X \times Y})$ such that $\nu'_X = \nu$. Let ν'' be any measure in $f^*(\mu)$ and let $\nu' \in \Delta_{X \times Y}$ be the "crossproduct" of ν and ν''; that is, $\nu'(A \times B) = \nu(A)\nu''(B)$. Clearly $\nu'_X = \nu$. To see that $\nu' \in [\![KB \wedge \psi]\!]_{X \times Y})$, it clearly suffices to show that $\nu' \models \psi$. But since ν and ν'' correspond under f, it is immediate from Proposition 4.7 that $\nu \models KB$ iff $\nu'' \models f^*(KB)$ and $\nu \models \varphi$ iff $\nu'' \models f^*(\varphi)$. Thus, $\nu \models \psi$, as desired.

Now suppose that $KB \mathrel{|\!\sim}_{I_X} \varphi$. By DI2 and DI5, $KB \wedge \psi$ is in the domain of $I_{X\times Y}$. By robustness, $KB \wedge \psi \mathrel{|\!\sim}_{I_{X\times Y}} \varphi$. Thus, $I([\![KB \wedge \psi]\!]_{X\times Y}) \subseteq [\![\varphi]\!]_{X\times Y}$. Since $I([\![KB \wedge \psi]\!]_{X\times Y}) \subseteq [\![KB \wedge \psi]\!]_{X\times Y} \subseteq [\![\varphi \Leftrightarrow f^*(\varphi)]\!]_{X\times Y}$, it follows that $I([\![KB \wedge \psi]\!]_{X\times Y}) \subseteq [\![f^*(\varphi)]\!]_{X\times Y}$. Moreover, $KB \wedge \psi$ is equivalent to $f^*(KB) \wedge \psi$, so $I([\![f^*(KB) \wedge \psi]\!]_{X\times Y}) \subseteq [\![f^*(\varphi)]\!]_{X\times Y}$, i.e., $f^*(KB) \wedge \psi \mathrel{|\!\sim}_{I_{X\times Y}} f^*(\varphi)$. By DI4, $f^*(KB)$ is in the domain of I_Y. Since ψ is Y-conservative over $f^*(KB)$, the robustness of $\{I_X : X \in \mathcal{X}\}$ implies that $f^*(KB) \mathrel{|\!\sim}_{I_Y} f^*(\varphi)$. The opposite implication (if $f^*(KB) \mathrel{|\!\sim}_{I_Y} f^*(\varphi)$ then $KB \mathrel{|\!\sim}_{I_X} \varphi$) goes the same way. Thus, $\{I_X : X \in \mathcal{X}\}$ is representation independent. ∎

Next, we turn our attention to Theorems 4.11 and 4.16. Both of these results follow in a relatively straightforward way from one key proposition. Before we state it, we need some definitions.

Definition A.4: We say that a constraint KB on Δ_X *depends only on* $S_1, \ldots, S_k \in \mathcal{F}_X$ (the sets $S_1, \ldots, S_k$ are not necessarily disjoint) if, whenever $\mu, \mu' \in \Delta_X$ agree on $S_1, \ldots, S_k$, then $\mu \models KB$ iff $\mu' \models KB$. ∎

For example, if KB has the form $\Pr(S_1) > 1/3 \wedge \Pr(S_2) \leq 3/4$, then KB depends only on S_1 and S_2. Similarly, if KB has the form $\Pr(S_1 \mid S_2) > 3/4$, then KB depends only on S_1 and S_2.

Definition A.5: Given $S_1, \ldots, S_k \in \mathcal{F}_X$, an *atom* over $S_1, \ldots, S_k$ is a set of the form $T_1 \cap \ldots \cap T_k$, where T_i is either S_i or $\overline{S_i}$. ∎

Proposition A.6: *Suppose that $\{I_X : X \in \mathcal{X}\}$ is an $\mathcal{X}$-inference procedure and, for some $X \in \mathcal{X}$, there exist $S, S_1, \ldots, S_K \in \mathcal{F}_X$ and a consistent constraint KB on Δ_X that depends only on $S_1, \ldots, S_k$, such that the following two conditions are satisfied:*

- *both $T \cap S$ and $T \cap \overline{S}$ are nonempty for every nonempty atom T over $S_1, \ldots, S_k$,*
- *$KB \mathrel{|\!\sim}_{I_X} \alpha < \Pr(S) < \beta$, where either $\alpha > 0$ or $\beta < 1$.*

Then $\{I_X : X \in \mathcal{X}\}$ is not representation independent.

Proof: Suppose, by way of contradiction, that $\{I_X : X \in \mathcal{X}\}$ is a representation-independent inference procedure but nevertheless, for some $X \in \mathcal{X}$, there exists sets $S, S_1, \ldots, S_k \in \mathcal{F}_X$ and a knowledge base KB that satisfies the conditions above, for some α, β. Assume that $\alpha > 0$ (a similar argument can be used to deal with the case that $\beta < 1$).

Let $T_1, \ldots, T_M$ be the nonempty atoms over $S_1, \ldots, S_k$. Choose N such that $1/N < \alpha$. Our goal is to find a collection $f_1, \ldots, f_N$ of embeddings of X into some $Y \in \mathcal{X}$ such that each of these embeddings has the same effect on KB, but such that the sets $f_j(S)$ are disjoint. Since $KB \mathrel{|\!\sim}_{I_X} \Pr(f_j(S)) > \alpha$ for $j = 1, \ldots, N$, and $f_j^*(KB) = f^*(KB)$ for $j = 1, \ldots, N$, it will follow that $f^*(KB) \mathrel{|\!\sim}_{I_Y} \Pr(f_j(S)) > \alpha$ for $j = 1, \ldots, N$, a contradiction. We proceed as follows.

By assumption, there exists a set Z in $\mathcal{X}$ such that $|Z| = MN$. Let $Y = X \times Z$. Since $\mathcal{X}$ is closed under crossproducts, $Y \in \mathcal{X}$. Suppose that $Z = \{z_1, \ldots, z_{MN}\}$, and let $Z_i = \{z_{N(i-1)+1}, \ldots, z_{Ni}\}$, for $i = 1, \ldots, M$. Thus, the Z_is partition Z into M disjoint sets, each of cardinality N. Let $B_i = X \times Z_i$, and let $B_{ij} = X \times \{z_{N(i-1)+j}\}$, for $j = 1, \ldots, N$. It is easy to see that we can find faithful X-Y embeddings $f_1, \ldots, f_N$ such that

1. $f_j(T_i) = B_i$, for $i = 1, \ldots, M$, $j = 1, \ldots, N$,
2. $f_j(T_i \cap S) = B_{ij}$, for $i = 1, \ldots, M$, $j = 1, \ldots, N$.

Notice that we need the assumption that both $T_i \cap S$ and $T_i \cap \overline{S}$ are nonempty for $T_1, \ldots, T_M$ (that is, for each nonempty atom over $S_1, \ldots, S_k$) to guarantee that we can find such faithful embeddings. For if $T_i \cap S = \emptyset$, then since f_j is an embedding, $f(T_i \cap S) = \emptyset \neq B_i$; and if $T_i \cap \overline{S} = \emptyset$, then $f_j(T_i \cap \overline{S}) = f_j(T_i) - f(T_i \cap S)) = \emptyset$, which means that $B_i = B_{ij}$, again inconsistent with the construction.

It is easy to check that, since KB depends only on $S_1, \ldots, S_k$, $f_j^*(KB)$ depends only on $f_j(S_1), \ldots, f_j(S_k)$, for $j = 1, \ldots, N$. We next show that $f_j(S_i)$ is independent of j; that is, $f_j(S_i) = f_{j'}(S_i)$ for $1 \leq j, j' \leq N$. Notice that for $h = 1, \ldots, k$, we have that $f_j(S_h) = \cup_{T_i \subseteq S_h} f_j(T_i) = \cup_{\{i : T_i \subseteq S_h\}} B_i$. Thus, $f_j(S_h)$ is independent of j, as desired. Since $f_j^*(KB)$ depends only on $f_j(S_1), \ldots, f_j(S_k)$, it too must be independent of j. Let KB^* be $f_1^*(KB)$ (which, as we have just observed, is identical to $f_2^*(KB), \ldots, f_k^*(KB)$).

Since, by assumption, $\{I_X : X \in \mathcal{X}\}$ is representation independent, and $KB \mathrel{|\!\!\sim}_{I_X} Pr(S) > \alpha$, we have that $KB^* \mathrel{|\!\!\sim}_{I_Y} \Pr(f_j(S)) > \alpha$, for $j = 1, \ldots, N$. Thus, $KB^* \mathrel{|\!\!\sim}_{I_Y} \Pr(f_1(S)) > \alpha \wedge \ldots \wedge \Pr(f_N(S)) > \alpha$. But note that, by construction, $f_j(S) = \cup_{\{i : T_i \cap S \neq \emptyset\}} B_{ij}$. Thus, the sets $f_j(S)$ are pairwise disjoint. Since $\alpha > 1/N$, we cannot have N disjoint sets each with probability greater than α. Thus, $KB^* \mathrel{|\!\!\sim}_{I_Y} \mathit{false}$. But KB is consistent, so $KB^* = f_j(KB)$ must be as well. Thus, $I_Y(KB^*) \neq \emptyset$, by assumption. But this contradicts our conclusion that $KB^* \mathrel{|\!\!\sim}_{I_Y} \mathit{false}$. Thus, $\{I_X : X \in \mathcal{X}\}$ cannot be representation independent. ∎

We can use Proposition A.6 to help prove Theorem 4.11.

Theorem 4.11: *If $\{I_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure then, for all $X \in \mathcal{X}$, I_X is essentially entailment for all objective knowledge bases in its domain.*

Proof: Suppose, by way of contradiction, that $\{I_X : X \in \mathcal{X}\}$ is representation independent but I_X is not essentially entailment for some $X \in \mathcal{X}$ and objective knowledge base KB. Then there must be some set $S \in \mathcal{F}_X$ such that $KB \mathrel{|\!\!\sim}_{I_X} \alpha < \Pr(S) < \beta$ and $KB \not\models \alpha \leq \Pr(S) \leq \beta$. Without loss of generality, we can assume that KB has the form $\Pr(T) = 1$ for some $T \in \mathcal{F}_X$. Moreover, we can assume that if $\overline{T} \neq \emptyset$, then $\overline{T}$ has a nonempty, measurable strict subset. (For otherwise, choose $Y = \{y, y'\} \in \mathcal{X}$ and consider the space $X' = X \times Y$. By assumption, $X' \in \mathcal{X}$. Let f be the X-Y embedding that maps $U \in \mathcal{F}_X$ to $U \times Y$. Since I is representation independent, we have that $\Pr(T \times Y) = 1 \mathrel{|\!\!\sim}_I \alpha < \Pr(S \times Y) < \beta$, and $\overline{T} \times \{y\} \subset \overline{T} \times Y$.)

If $\overline{T}$ is nonempty, let Z be any nonempty, measurable strict subset of $\overline{T}$ (which exists by assumption); otherwise let Z be the empty set. Let U be the set $(T \cap S) \cup (\overline{T} \cap Z)$. Notice that $S \cap T = U \cap T$. Moreover, since, for any set V, $\Pr(T) = 1 \Rightarrow \Pr(V) = \Pr(V \cap T)$ is valid, it follows from Reflexivity and Right Weakening that $KB \mathrel{|\!\!\sim}_{I_X} \Pr(V) = \Pr(V \cap T)$. Thus, $KB \mathrel{|\!\!\sim}_{I_X} \Pr(S) = \Pr(S \cap T) = \Pr(U \cap T) = \Pr(U)$. It follows that $KB \mathrel{|\!\!\sim}_{I_X} \alpha < \Pr(U) < \beta$.

We now want to apply Proposition A.6. Note that KB depends only on T. Thus, we must show that $T \cap U$ and $T \cap \overline{U}$ are nonempty, and if $\overline{T}$ is nonempty, then $\overline{T} \cap U$ and $\overline{T} \cap \overline{U}$ are as well. As we observed above, $T \cap U = T \cap S$. Thus, if $T \cap U = \emptyset$, then $T \subseteq \overline{S}$, contradicting our assumption that $KB \mathrel{|\!\!\sim}_I \Pr(S) > 0$. It is easy to see that $T \cap \overline{U} = T \cap \overline{S}$. Again, we cannot have $T \cap \overline{U} = \emptyset$, for then $T \subseteq S$, contradicting our assumption that

$KB \mathrel{|\!\sim}_I \Pr(S) < 1$. By construction, $\overline{T} \cap U = \overline{T} \cap Z = Z$. By assumption, if $\overline{T} \neq \emptyset$, then $Z \neq \emptyset$. Finally, $\overline{T} \cap \overline{U} = \overline{T} \cap \overline{Z}$; again, by construction, this is a nonempty set if $\overline{T} \neq \emptyset$. It now follows from Proposition A.6 that $\{I_X : X \in \mathcal{X}\}$ is not representation independent. ∎

Corollary 4.12: *If $\{I_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure, then for all $X \in \mathcal{X}$, if KB is an objective knowledge base putting constraints on Δ_X, and $KB \mathrel{|\!\sim}_{I_X} \alpha < \Pr(S) < \beta$ for some $\alpha \geq 0$ and $\beta \leq 1$, then $\alpha = 0$ and $\beta = 1$.*

Proof: Assume the hypotheses of the corollary hold. Since KB is objective, it is of the form $\Pr(T) = 1$ for some $T \in \mathcal{F}_X$. Then there are three possibilities. Either (1) $T \subseteq S$, (2) $T \subseteq \overline{S}$, or (3) both $T \cap S$ and $T \cap \overline{S}$ are nonempty. If (1) holds, we have $KB \models \Pr(S) = 1$, while if (2) holds, we have $KB \models \Pr(S) = 0$. Thus, both (1) and (2) are incompatible with $KB \mathrel{|\!\sim}_{I_X} \alpha < \Pr(S) < \beta$. On the other hand, if (3) holds, it is easy to see that for all γ, $\Pr(S) = \gamma$ is consistent with KB (since there is a probability measure that assigns probability γ to $T \cap S$ and probability $1 - \gamma$ to $T \cap \overline{S}$). Since $KB \mathrel{|\!\sim}_{I_X} \alpha < \Pr(S) < \beta$, by Theorem 4.11, we must have $KB \models \alpha \leq \Pr(S) \leq \beta$. It follows that the only choices of α and β for which this can be true are $\alpha = 0$ and $\beta = 1$. ∎

Theorem 4.16: *If $\{I_X : X \in \mathcal{X}\}$ is an $\mathcal{X}$-inference procedure that enforces minimal default independence and satisfies DI1, then I_X is not representation independent.*

Proof: Suppose that $\{I_X : X \in \mathcal{X}\}$ is an $\mathcal{X}$-inference procedure that enforces minimal default independence and satisfies DI1. Choose $X = \{x, x'\} \in \mathcal{X}$ and let KB be $1/3 \leq \Pr(x) \leq 2/3$. By assumption, $X \times X \in \mathcal{X}$. We can view KB as a constraint on $\Delta_{X \times X}$; in this case, it should be interpreted as $1/3 \leq \Pr(\{x\} \times X) \leq 2/3$. by DI1, KB is is the domain of $I_{X \times X}$. Note that KB is equivalent to the constraint $1/3 \leq \Pr(x') \leq 2/3$. By minimal default independence, we have that $KB \mathrel{|\!\sim}_{I_{X \times X}} \Pr((x, x)) > \Pr(x \times X)/3$ and that $KB \mathrel{|\!\sim}_{I_{X \times X}} \Pr((x', x')) > \Pr(x' \times X)/3$. Applying straightforward probabilistic reasoning, we get that $KB \mathrel{|\!\sim}_{I_{X \times X}} \Pr(\{(x, x), (x', x')\}) > 1/3$. We now apply Proposition A.6, taking S to be $\{(x, x), (x', x')\}$ and S' to be $\{(x, x), (x, x')\}$. Note that KB depends only on S'. It is almost immediate from the definition of S and S' that all of $S \cap S'$, $\overline{S} \cap S'$, $S \cap \overline{S'}$, and $\overline{S} \cap \overline{S'}$ are nonempty. Thus, by Proposition A.6, $\{I_X : X \in \mathcal{X}\}$ is not representation independent. ∎

Lemma 4.13: *Let $\mathcal{X}$ consist of only countable sets. Then $\{I^0_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure.*

Proof: As we said in the main part of the text, it easily follows from Proposition 2.5 that I^0_X is an inference procedure for all $X \in \mathcal{X}$, since it is easily seen to have the five properties described in the proposition. To see that I^0 is representation independent, suppose that f is a faithful X-Y embedding, for $X, Y \in \mathcal{X}$. Clearly KB is objective if and only if $f^*(KB)$ is objective. If KB is not objective, then it is easy to see that $KB \mathrel{|\!\sim}_{I^0} \theta$ iff $f^*(KB) \mathrel{|\!\sim}_{I^0} f^*(\theta)$, since $\mathrel{|\!\sim}_{I^0}$ reduces to entailment in this case. So suppose that KB is objective and has the form $\Pr(T) = 1$, for some $T \in \mathcal{F}_X$. Then $KB \mathrel{|\!\sim}_{I^0} \theta$ iff $KB \wedge KB^+ \models \theta$. By Lemma 4.6, this holds iff $f^*(KB) \wedge f^*(KB^+) \models f^*(\theta)$. On the other hand, $f^*(KB) \mathrel{|\!\sim}_{I^0} f^*(\theta)$ iff $f^*(KB) \wedge (f^*(KB))^+ \models f^*(\theta)$ Thus, it suffices to show that $f^*(KB) \wedge f^*(KB^+) \models f^*(\theta)$ iff $f^*(KB) \wedge (f^*(KB))^+ \models f^*(\theta)$. It is easy to show that $(f^*(KB))^+$ implies $f^*(KB^+)$, so that if $f^*(KB) \wedge f^*(KB^+) \models f^*(\theta)$ then $f^*(KB) \wedge (f^*(KB))^+ \models f^*(\theta)$. It is not necessarily the case that $f^*(KB^+)$ implies $(f^*(KB))^+$. For example, consider the embedding described in Example 4.3. In that case, if KB is the objective knowledge base $\Pr(\mathit{colorful}) = 1$, KB^+ is empty, and hence so is $f^*(KB^+)$, while $(f^*(KB))^+$ includes constraints such as $0 < \Pr(\mathit{green}) < 1$. Nevertheless, suppose that $f^*(KB) \wedge (f^*(KB))^+ \models f^*(\theta)$ and, by way of contradiction, there is some ν such that $\nu \models f^*(KB) \wedge f^*(KB^+) \wedge \neg f^*(\theta)$. Choose μ such that $\nu \in f^*(\mu)$. Then μ and ν correspond, so $\mu \models KB \wedge KB^+ \wedge \neg\theta$. It is easy to show that there exists $\nu' \in f^*(\mu)$ such that $0 < \nu'(S) < 1$ for all nonempty subsets of S of $f(T)$. To see this, note that if $\mu(x) \neq 0$, then it suffices to ensure that $\nu'(f(x)) = \mu(x)$ and $\nu'(y) \neq 0$ for all y in $f(x)$. Since Y is countable, this is straightforward. Since μ and ν' correspond, we must have that $\nu' \models \neg f^*(\theta) \wedge f^*(KB)$. By construction, $\nu' \models (f^*(KB))^+$. This contradicts the assumption that $f^*(KB) \wedge (f^*(KB))^+ \models f^*(\theta)$. ∎

Lemma 4.14: *Suppose that $\mathcal{X}$ consists only of measure spaces of the form $(X, 2^X)$, where X is finite. Then $\{I^1_X : X \in \mathcal{X}\}$ is a representation-independent $\mathcal{X}$-inference procedure.*

Proof: Suppose that $X, Y \in \mathcal{X}$, KB and φ are constraints on Δ_X, and f is an X-Y embedding. We must show that $KB \mathrel{|\!\sim}_{I^1_X} \varphi$ iff $f^*(KB) \mathrel{|\!\sim}_{I^1_Y} f^*(\varphi)$. For the purposes of this proof, we say that a subset A of Δ_X is *interesting* if there exists some $S \in \mathcal{F}_X$ such that $A = \{\mu \in \Delta_X : \mu(S) \geq 1/4\}$. It is easy to see that if KB is interesting then $f^*(KB)$ is interesting. The converse is also true, given our assumption that $\mathcal{X}$ consists of only finite spaces where all sets are measurable. For suppose that $f^*(KB)$ is interesting. Then there is a set $T \subseteq Y$ such that $f^*(KB) = \{\nu \in \Delta_Y : \nu(T) \geq 1/4\}$. Let $\mathcal{A} = \{S' \subseteq X : f(S') \supseteq T\}$. Since X is finite, so is $\mathcal{A}$; it easily follows that $S = \cap\mathcal{A} \in \mathcal{A}$.[12] Clearly if $\mu(S) \geq 1/4$, then $f^*(\mu) \subseteq f^*(KB)$, so $\mu \in [\![KB]\!]_X$. Thus, $[\![KB]\!]_X \supseteq \{\mu \in \Delta_X : \mu(S) \geq 1/4\}$. On the other hand, if $\mu \in KB$, then $f^*(\mu) \subseteq f^*(KB)$. Thus, if $\nu \in f^*(\mu)$, since $S \in \mathcal{A}$, it must be the case that $\mu(S) = \nu(f(S)) \geq \nu(T) \geq 1/4$. Thus, $[\![KB]\!]_X \subseteq \{\mu \in \Delta_X : \mu(S) \geq 1/4\}$. It follows that KB is equivalent to $\Pr(S) \geq 1/4$, and so must be interesting. (We must also have $T = f(S)$, although this is not needed for the result.)

If KB is not interesting, then $KB \mathrel{|\!\sim}_{I^1_X} \varphi$ iff $KB \models \varphi$ iff $f^*(KB) \models f^*(\varphi)$ (since entailment is representation independent) iff $f^*(KB) \mathrel{|\!\sim}_{I^1_Y} \varphi$. On the other hand, if KB is interesting, then KB is equivalent to $\Pr(S) \geq 1/4$ for some $S \subseteq X$, and $f^*(KB)$ is equivalent to $\Pr(f(S)) \geq 1/4$. Moreover, $KB \mathrel{|\!\sim}_{I^1_X} \varphi$ iff $\Pr(S) \geq 1/3 \models \varphi$ iff $\Pr(f(S)) \geq 1/3 \models f^*(\varphi)$ iff $f^*(KB) \mathrel{|\!\sim}_{I^1_Y} \varphi$. Thus, we get representation independence, as desired. ∎

A.4 Proofs for Section 6

Proposition 6.3: *Suppose that f is a faithful X-Y embedding, $\mathcal{D}_X \subseteq \Delta_X$, and $\mathcal{D}_Y \subseteq \Delta_Y$. The following two conditions are equivalent:*

(a) $\mathcal{D}_X$ and $\mathcal{D}_Y$ correspond under f;

(b) for all θ, $\mathcal{D}_X \models \theta$ iff $\mathcal{D}_Y \models f^*(\theta)$.

12. This is not in general true if X is infinite without the additional requirement that $f(\cup_i A_i) = \cup_i f(A_i)$ for arbitrary unions.

Proof: To prove that (a) implies (b), assume by way of contradiction that, for some θ, $\mathcal{D}_X \models \theta$ but $\mathcal{D}_Y \not\models f^*(\theta)$. Then there is some $\nu \in \mathcal{D}_Y$ such that $\nu \not\models f^*(\theta)$. Let $\mu \in \mathcal{D}_X$ be a measure corresponding to μ. Then, by Proposition 4.7, we have that $\mu \not\models \theta$, the desired contradiction. The proof for the other direction of (a) is identical.

To prove that (b) implies (a), first consider a measure $\mu \in \mathcal{D}_X$. We must find a $\nu \in \mathcal{D}_Y$ such that ν corresponds to μ. Suppose that $X = \{x_1, \ldots, x_n\}$ (recall that we are restricting to finite spaces in Section 6) and that $\mu(x_i) = a_i$, $i = 1, \ldots, n$. Let θ be the constraint $\wedge_{i=1}^n \Pr(\{x_i\}) = a_i$. By our assumptions about the language, this constraint is in the language. Clearly $[\![\theta]\!]_X = \{\mu\}$. Since $\mu \in \mathcal{D}_X$, we know that $\mathcal{D}_X \not\models \neg\theta$. Hence, $\mathcal{D}_Y \not\models f^*(\neg\theta)$, so that there exists $\nu \in \mathcal{D}_Y$ such that $\nu \notin f^*(\neg\theta)$. Hence $\nu \in f^*(\theta) = f^*(\{\mu\})$. By definition of f^*, ν corresponds to μ.

Now consider a measure $\nu \in \mathcal{D}_Y$, and let μ be the measure in Δ_X that corresponds to ν. Assume by way of contradiction that $\mu \notin \mathcal{D}_X$. Taking θ as above, it follows that $\mathcal{D}_X \models \neg\theta$ and, therefore, by assumption, $\mathcal{D}_Y \models f^*(\neg\theta)$. Thus, $\nu \models f^*(\neg\theta)$. But $\mu \models \theta$ and, by assumption, μ and ν correspond. This contradicts Proposition 4.7. ∎

Theorem 6.7: *Let θ be an arbitrary constraint on Δ_X. If f is a faithful X-Y embedding and μ and ν correspond under f, then $\mu|\theta$ and $\nu|f^*(\theta)$ also correspond under f.*

Proof: Assume that μ and ν correspond under f. Recall that we are assuming in this section that X is a finite space; let $X = \{x_1, \ldots, x_n\}$. Let $Y_i = f(x_i)$. Given any distribution $\nu'' \in \Delta_Y$, define $\nu''_i = \nu''|Y_i$ and let $(f^*)^{-1}(\nu'')$ denote the unique $\mu'' \in \Delta_X$ such that $\nu'' \in f^*(\mu'')$.

Now suppose that $\mu' \in \mu|\theta$. Define $\nu' \in \Delta_Y$ to be the measure such that

$$\nu'(y) = \mu'(x_i) \cdot \nu_i(y),$$

where i is the index such that $y \in Y_i$. Since $\nu_i = \nu|Y_i$, it follows that $\nu_i(Y_i) = 1$. Thus, $\nu'(Y_i) = \mu(x_i)$, and ν' leaves the relative probabilities of elements within each Y_i the same as in ν. It is easy to verify that ν' and μ' correspond. Hence, by Proposition 4.7, $\nu' \models f^*(\theta)$. We claim that $\nu' \in \nu|f^*(\theta)$. To show that, we need show only that $KL_Y(\nu'\|\nu)$ is minimal among all $KL_Y(\nu''\|\nu)$ such that $\nu'' \models f^*(\theta)$. It follows from standard properties of relative entropy (Cover & Thomas, 1991, Theorem 2.5.3) that for all $\nu'' \in \Delta_Y$, we have

$$KL_Y(\nu''\|\nu) = KL_X((f^*)^{-1}(\nu'')\|(f^*)^{-1}(\nu)) + \sum_{i=1}^{n} KL_Y(\nu''_i\|\nu_i). \tag{1}$$

Note that $\nu_i = \nu'_i$, so $KL_Y(\nu'_i\|\nu_i) = 0$, for $i = 1, \ldots, n$. Thus, it follows from (1) that $KL_Y(\nu'\|\nu) = KL_X(\mu'\|\mu)$.

Now, let $\nu'' \in \Delta_Y$ be such that $\nu'' \models f^*(\theta)$ and let $\mu'' = (f^*)^{-1}(\mu'')$. Since ν'' and μ'' correspond under f, it follows from Proposition 4.7 that $\mu'' \models \theta$. Using (1) once again, we have that

$$\begin{aligned} KL_Y(\nu''\|\nu) &= KL_X(\mu''\|\mu) + \sum_{i=1}^{n} KL_Y(\nu''_i\|\nu_i) \\ &\geq KL_X(\mu''\|\mu). \end{aligned}$$

But since $\mu' \in \mu|\theta$, we know that $KL_X(\mu'\|\mu) \leq KL_X(\mu''\|\mu)$. Hence we conclude that

$$KL_Y(\nu''\|\nu) \geq KL_Y(\nu'\|\nu),$$

so that $\nu' \in \nu|f^*(\theta)$. ∎

Theorem 6.9: *If f is a faithful X-Y embedding, then $I^{\mathcal{P}}$ is invariant under f iff $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ correspond under f.*

Proof: Suppose that f is a faithful X-Y embedding. By definition, $I^{\mathcal{P}}$ is invariant under f iff, for all KB, θ, we have

$$KB \mathrel{|\!\sim}_{I^{\mathcal{P}}} \theta \text{ iff } f^*(KB) \mathrel{|\!\sim}_{I^{\mathcal{P}}} f^*(\theta). \tag{2}$$

By definition of $I^{\mathcal{P}}$, (2) holds iff

$$\mathcal{P}(X)|KB \subseteq [\![\theta]\!]_X \text{ iff } \mathcal{P}(Y)|f^*(KB) \subseteq [\![f^*(\theta)]\!]_Y \text{ for all } KB,\, \theta. \tag{3}$$

By Proposition 6.3, (3) holds iff $\mathcal{P}(X)|KB$ and $\mathcal{P}(Y)|f^*(KB)$ correspond for all KB. By Corollary 6.5, if $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ correspond, then $\mathcal{P}(X)|KB$ and $\mathcal{P}(Y)|f^*(KB)$ correspond for all KB. On the other hand, if $\mathcal{P}(X)|KB$ and $\mathcal{P}(Y)|f^*(KB)$ correspond for all KB, then $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ must correspond: simply take $KB = \mathit{true}$ and observe that $\mathcal{P}(X)|KB) = \mathcal{P}(X)$ and $\mathcal{P}(Y)|f^*(KB) = \mathcal{P}(Y)$. ∎

Proposition 6.10: *Suppose that $X_1 \times \cdots \times X_n$ is the product decomposition on X and, for each $i = 1, \ldots, n$, KB_i is a constraint on X_i, and S_i is a subset of X_i. Then*

$$\bigwedge_{i=1}^{n} KB_i \mathrel{|\!\sim}_{I_{\mathcal{P}_\Pi}} \Pr(S_1 \wedge \ldots \wedge S_n) = \prod_{i=1}^{n} \Pr(S_i).$$

Proof: If KB_i is a satisfiable constraint on Δ_{X_i}, for $i = 1, \ldots, n$, then there exist product measures on X satisfying the constraints $\bigwedge_{i=1}^n KB_i$. These product measures are precisely the measures in $\mathcal{P}_\Pi|(\bigwedge_{i=1}^n KB_i)$. Since each of these measures satisfies $\Pr(S_1 \wedge \ldots \wedge S_n) = \prod_{i=1}^n \Pr(S_i)$ by assumption, the conclusion holds in this case. If any constraint KB_i is not satisfiable, then the result trivially holds. ∎

Theorem 6.11: *The inference procedure $I_{\mathcal{P}_\Pi}$ is invariant under faithful product embeddings and under permutation embeddings.*

Proof: Suppose that f is a faithful X-Y product embedding, $X_1 \times \cdots \times X_n$ is the product decomposition of X, and $Y_1 \times \cdots \times Y_n$ is the product decomposition of Y. To show that $\mathcal{P}_\Pi$ is invariant under f, it suffices to show that $\mathcal{P}_\Pi(X)$ and $\mathcal{P}_\Pi(Y)$ correspond under f. Supposethat $\mu \in \mathcal{P}_\Pi(Y)$. Then $\mu = \mu_1 \times \cdots \times \mu_n$, where μ_i is a measure on X_i, $i = 1, \ldots, n$. Moreover, since f is a product embedding, there exist $f_1, \ldots, f_n$ such that $f = f_1 \times \cdots \times f_n$. Let $\nu_i \in f_i^*(\mu_i)$, for $i = 1, \ldots, n$. It is easy to check that $\nu = \nu_1 \times \cdots \times \nu_n \in f^*(\mu)$.

Conversely, suppose that $\nu \in \mathcal{P}_\Pi(Y)$. Then $\nu = \nu_1 \times \cdots \times \nu_n$, where $\nu_i \in \Delta_{Y_i}$ for $i = 1, \ldots, n$. Define $\mu \in \Delta_{X_i}$ by setting $\mu_i(S) = \nu_i(f_i(S))$. Since f_i is a faithful X_i-Y_i

embedding, is easy to check that $\mu_i \in \Delta_{X_i}$ and that $\nu_i \in f_i^*(\mu_i)$. Thus, $\nu \in f^*(\mu)$. This completes the proof that $\mathcal{P}_\Pi$ is invariant under faithful X-Y product embeddings.

The argument that $\mathcal{P}_\Pi$ is invariant under faithful X-X permutation embeddings is similar (and easier). We leave details to the reader. ∎

References

Bacchus, F. (1990). *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge, Mass.

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

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

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

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

Goldszmidt, M., Morris, P., & Pearl, J. (1993). A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, *15*(3), 220–232.

Halpern, J. Y., & Koller, D. (1995). Representation dependence in probabilistic inference. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pp. 1853–1860.

Horn, A., & Tarski, A. (1948). Measures in Boolean algebras. *Transactions of the AMS*, *64*(1), 467–497.

Jaeger, M. (1996). Representation independence of nonmonotonic inference relations. In *Principles of Knowledge Representation and Reasoning: Proc. Fifth International Conference (KR '96)*, pp. 461–472.

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

Jaynes, E. T. (1978). Where do we stand on maximum entropy?. In Levine, R. D., & Tribus, M. (Eds.), *The Maximum Entropy Formalism*, pp. 15–118. MIT Press, Cambridge, Mass.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge/New York.

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

Keisler, J., & Tarski, A. (1964). From accessible to inaccessible cardinals. *Fundamenta Mathematica*, *53*, 225–308.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, *44*, 167–207.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 76–86.

Lehmann, D., & Magidor, M. (1992). What does a conditional knowledge base entail?. *Artificial Intelligence*, *55*, 1–60.

Nayak, P. P., & Levy, A. Y. (1995). A semantic theory of abstractions. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pp. 196–203.

Paris, J. B. (1994). *The Uncertain Reasoner's Companion*. Cambridge University Press, Cambridge, U.K.

Paris, J., & Vencovská, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, *4*(3), 183–224.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco.

Salmon, W. (1961). Vindication of induction. In Feigl, H., & Maxwell, G. (Eds.), *Current Issues in the Philosophy of Science*, pp. 245–264. Holt, Rinehart, and Winston, New York.

Salmon, W. (1963). On vindicating induction. In Kyburg, H. E., & Nagel, E. (Eds.), *Induction: Some Current Issues*, pp. 27–54. Wesleyan University Press, Middletown, Conn.

Seidenfeld, T. (1987). Entropy and uncertainty. In MacNeill, I. B., & Umphrey, G. J. (Eds.), *Foundations of Statistical Inferences*, pp. 259–287. Reidel, Dordrecht, Netherlands.

Shore, J. E., & Johnson, R. W. (1980). 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.

Ulam, S. (1930). Zur masstheorie in der allgemeinen mengenlehre. *Fundamenta Mathematicae*, *16*, 140–150.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, *58*(1), 3–34. Discussion of the paper by various commentators appears on pp. 34–57.