# Statistical Foundations for Default Reasoning*

**Fahiem Bacchus**
Computer Science Dept.
University of Waterloo
Waterloo, Ontario
Canada, N2L 3G1
fbacchus@logos.uwaterloo.ca

**Adam J. Grove**
NEC Research Inst.
4 Independence Way
Princeton, NJ 08540
grove@research.nj.nec.com

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

**Daphne Koller**
Computer Science Dept.
Stanford University
Stanford, CA 94305
daphne@cs.stanford.edu

## Abstract

We describe a new approach to default reasoning, based on a principle of indifference among possible worlds. We interpret default rules as extreme statistical statements, thus obtaining a knowledge base KB comprised of statistical and first-order statements. We then assign equal probability to all worlds consistent with KB in order to assign a *degree of belief* to a statement $\varphi$. The degree of belief can be used to decide whether to defeasibly conclude $\varphi$. Various natural patterns of reasoning, such as a preference for more specific defaults, indifference to irrelevant information, and the ability to combine independent pieces of evidence, turn out to follow naturally from this technique. Furthermore, our approach is not restricted to default reasoning; it supports a spectrum of reasoning, from quantitative to qualitative. It is also related to other systems for default reasoning. In particular, we show that the work of [Goldszmidt *et al.*, 1990], which applies maximum entropy ideas to $\epsilon$-semantics, can be embedded in our framework.

## 1 Introduction

Default reasoning, i.e., reasoning to plausible but deductively invalid conclusions, has been an important research area in AI for over a decade. Work in this area shows us how, given that we accept the default "Birds typically fly" and the fact "Tweety is a bird", we can arrive at the reasonable (although possibly incorrect) belief that "Tweety flies".

*The work of Fahiem Bacchus was supported by NSERC under their operating grants program and by IRIS. The work of Adam Grove, Joseph Halpern, and Daphne Koller was sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080. Adam Grove's work was performed while at Stanford University, where he was also supported by an IBM Graduate Fellowship. Some of this research was performed while Adam Grove and Daphne Koller were at IBM Almaden Research Center. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

A different reasoning paradigm, which has been studied for an even longer period, is *direct inference.* Direct inference is concerned with reasoning to conclusions about particular individuals from general statistical knowledge. For example, from a knowledge base consisting of the statistical information "90% of birds fly" and the fact "Tweety is a bird", theories of direct inference would allow us to conclude that our *degree of belief* in "Tweety flies" should be 0.9. Different systems for direct inference that have been suggested include [Bacchus, 1990; Bacchus *et al.*, 1992; Kyburg, 1974; Levi, 1980; Pollock, 1990; Reichenbach, 1949; Salmon, 1971].

Direct inference and default reasoning share a number of important characteristics. First, neither is a logically sound inference system. Neither statistical knowledge nor defaults about the class of all birds permit us to deduce anything for certain about a particular bird such as Tweety: Both "Tweety flies" and "Tweety does not fly" are logically consistent with "90% of birds fly" or "birds typically fly". Second, both direct inference and default reasoning are nonmonotonic. If we learn that penguins do not fly, and that Tweety is a penguin, direct inference would generate a different degree of belief in Tweety flying. Similarly, default reasoning systems would retract the conclusion that Tweety flies. And third, various properties, such as ignoring irrelevant information and preferring more specific information, are considered to be desirable in both default reasoning and direct inference.

So how deep is the connection between direct inference and AI default reasoning? Some applications of defaults seem to have little to do with statistics [McCarthy, 1986]. But equally often, defaults have *some* basis in statistics. For example, the default "Birds typically fly" appears to have as one of its justifications the statistical fact that most birds fly. Thus, it seems reasonable to adapt techniques from direct inference to reason with defaults of this type. Our theme in this paper is that this plausible connection between direct inference and default reasoning can be made precise. In particular, we show in Section 3 that a new method for direct inference, first introduced in [Bacchus *et al.*, 1992; Grove *et al.*, 1992b], can provide many of the features considered desirable in default reasoning. Among other things, it provides a preference for more specific defaults as well as the ability to ignore irrelevant information.

This is particularly important as there is a tension between these two requirements. Most theories of default reasoning fail to capture both of them simultaneously (see, for example, [Geffner and Pearl, 1990; Lehmann and Magidor, 1992; Pearl, 1990; Reiter, 1980]). What is even more important is that for us these properties follow directly from an *independently motivated* semantics; they are not the result of adopting an *ad hoc* theory of irrelevance.

In our method, we presume that there is a knowledge base that consists of information about the world, in the form of first-order statements (such as "All penguins are birds") and statistical information. The statistical information might be quantitative, e.g., "90% of birds fly", or it might be in the form of qualitative default information. We interpret a default statement such as "Birds typically fly" as the statistical assertion "Almost all birds fly", which is given a precise semantic interpretation within our formalism. This interpretation of defaults has a number of benefits. The first is simply that we understand what our knowledge base means. Many default theories will tell us how to reason with "Birds typically fly". But, as pointed out by [Neufeld, 1989], there is far less work telling us when we should adopt this default in the first place. Specifically, what is there about the world that makes this a good default? For us, the true proportion of flying birds offers a guide to how reasonable our approximation "Almost all birds fly" really is. In addition, the semantics imposes natural constraints on the defaults. For example, in our formalism the default "Birds typically fly" is inconsistent with both the default "Birds typically do not fly" and the logical assertion "No bird flies".

A major advantage of our approach is that it allows for rich knowledge bases, with arbitrary first-order information and statistical information. Thus, it can support both quantitative and qualitative reasoning. In Section 4, we demonstrate the advantages of being able to perform both types of reasoning in a unified framework, by considering both the Lottery Paradox and the Nixon Diamond example.

We are certainly not the first to apply a probabilistic semantics to nonmonotonic logic (see [Pearl, 1989] for an overview). However, while all the other probabilistic approaches we are aware of use the statistical interpretation as a motivation for using probabilities, none make explicit use of statistical assertions. Nevertheless, there are close technical connections between our approach and $\epsilon$-semantics [Adams, 1975; Geffner and Pearl, 1990]. In particular, we show in Section 5 that the approach of Goldszmidt, Morris, and Pearl [1990], which extends $\epsilon$-semantics by applying ideas of maximum entropy, can be embedded in our framework. Besides providing further justification for the use of maximum entropy in [Goldszmidt *et al.*, 1990], this embedding allows us to use the algorithms they have developed to calculate degrees of belief for formulas in a fragment of our full language.

## 2 The Formalism

We assume that the knowledge base consists of sentences written in a formal language that allows us to express both statistical information and first-order information. We use the probability logic presented in [Grove *et al.*, 1992b], which is a variant of logics developed in [Bacchus, 1990; Halpern, 1990].

This logic augments first-order logic by allowing *proportion expressions* of the form $||\psi(x)||_x$. This term denotes the proportion of domain elements satisfying $\psi$. We actually allow an arbitrary set of variables in the subscript. Thus, for example, $||\mathsf{Son}(x,y)||_x$ describes, for a fixed $y$, the proportion of domain elements that are sons of $y$; $||\mathsf{Son}(x,y)||_y$ describes, for a fixed $x$, the proportion of domain elements whose son is $x$; and $||\mathsf{Son}(x,y)||_{\{x,y\}}$ describes the proportion of pairs of domain elements that are in the son relation. We also allow conditional proportion expressions of the form $||\psi(x)|\theta(x)||_x$, which denotes the proportion of domain elements satisfying $\psi$ from among those elements satisfying $\theta$.[1] A rational number is also a proportion expression, and the set of proportion expressions is closed under addition, subtraction, and multiplication.

One important difference between our syntax and that of [Bacchus, 1990] is the use of *approximate equality* to compare proportion expressions. It is not hard to see that exact comparisons are sometimes inappropriate. Consider a statement such as "90% of birds fly". If this statement appears in a database, it is almost certainly there as a summary of a large pool of data. It is clear that we do not mean that *exactly* 90% of all birds fly. Among other things, this would imply that the number of birds is a multiple of ten, an implication that is surely not intended. We therefore use the approach described in [Grove *et al.*, 1992b; Koller and Halpern, 1992], and compare proportion expressions using (instead of $=$ and $\leq$) one of an infinite family of connectives $\approx_i$ and $\preceq_i$, for $i = 1, 2, 3\ldots$ ("$i$-approximately equal" or "$i$-approximately less than or equal").[2] For example, we can express the statement "90% of birds fly" by the *proportion formula* $||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_1 0.9$. The intuition behind the semantics of approximate equality is that each comparison should be interpreted using some small tolerance factor to account for measurement error, sample variations, and so on. The appropriate tolerance will differ for various pieces of information, so our logic allows different subscripts on the "approximately equals" connectives. A formula such as $||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_1 1 \wedge ||\mathsf{Fly}(x)|\mathsf{Bat}(x)||_x \approx_2 1$ says that both $||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x$ and $||\mathsf{Fly}(x)|\mathsf{Bat}(x)||_x$ are approximately 1, but the notion of "approximately" may be different in each case.

We now briefly sketch the semantics of novel features of the logic. We evaluate the truth of a formula with respect to a triple $(M, \vec{\tau}, V)$, where $M$ is a finite first-order structure, $\vec{\tau} = \langle\tau_1, \tau_2, \ldots\rangle$, $\tau_i > 0$, is a *tolerance*

[1] We discuss the issue of conditioning on an event with probability zero in the full paper.

[2] In [Bacchus *et al.*, 1992] the use of approximate equality was suppressed in order to highlight other issues.

*vector*, used to give semantics to the connectives $\approx_i$ and $\preceq_i$, and $V$ is a valuation, which interprets the free variables as elements of the domain in structure $M$. For each proportion expression $\xi$, we can define a rational number $[\xi]_{M,V}$ which is the interpretation of $\xi$ in structure $M$ under valuation $V$. For example, $[||\mathsf{Son}(x,y)||_x]_{M,V}$ is the fraction of domain elements $x$ which are sons of $V(y)$. Proportion expressions are dealt with using $\vec{\tau}$: $(M,\vec{\tau},V) \models \xi \approx_i \xi'$ if $|[\xi]_{M,V} - [\xi']_{M,V}| \leq \tau_i$. We write $\models \varphi$ if $(M,\vec{\tau},V) \models \varphi$ for all $(M,\vec{\tau},V)$.

We want the agent to use the information in the knowledge base to assign *degrees of belief* to various assertions. Following [Halpern, 1990], we give semantics to degrees of belief in terms of a set of finite first-order models or *possible worlds*, together with a probability distribution over this set. The degree of belief in a sentence $\varphi$ is just the probability of the set of worlds where $\varphi$ is true.[3] In particular, given a knowledge base $\mathsf{KB}$ and domain size $N$, we consider all the worlds of size $N$ consistent with $\mathsf{KB}$. Furthermore, since we assume that $\mathsf{KB}$ is "all the agent knows", we view each of these possible worlds as equally likely; after all, the knowledge base does not give us any reason to prefer one world over any other. This is essentially an application of the *principle of indifference* due to Laplace [1820]. This method, which we call the *random-worlds method*, was investigated in some detail by Johnson [1932] and Carnap [1950, 1952].

Formally, given a vocabulary $\Phi$, a domain size $N$, and a tolerance vector $\vec{\tau}$, we define

$$\Pr^w_{N,\vec{\tau}}(\varphi|\mathsf{KB}) = \frac{\#worlds^{\vec{\tau}}_N(\varphi \wedge \mathsf{KB})}{\#worlds^{\vec{\tau}}_N(\mathsf{KB})},$$

where $\#worlds^{\vec{\tau}}_N(\psi)$ is the number of first-order structures $M$ over the domain $\{1,\ldots,N\}$ such that $(M,\vec{\tau}) \models \psi$.

Typically, we know neither $N$ nor $\vec{\tau}$ exactly. All we know is that $N$ is "large" and that $\vec{\tau}$ is "small". Thus, we would like to take our *degree of belief* in $\varphi$ given $\mathsf{KB}$ to be $\lim_{\vec{\tau}\to\vec{0}} \lim_{N\to\infty} \Pr^w_{N,\vec{\tau}}(\varphi|\mathsf{KB})$. However, there is no guarantee that this limit exists. A necessary condition for the limit to exist is that the knowledge base $\mathsf{KB}$ be *eventually consistent*: that is, for all sufficiently small $\vec{\tau}$ and sufficiently large $N$, $\#worlds^{\vec{\tau}}_N(\mathsf{KB}) > 0$. Essentially, eventual consistency says that not only is the $\mathsf{KB}$ consistent, but that there is nothing in the $\mathsf{KB}$ that limits the domain size (for example, a formula saying 7 domain elements"). For the remainder of the paper, we assume that all knowledge bases are eventually consistent. Even if $\mathsf{KB}$ is eventually consistent, the limit may not exist. In many cases, the nonexistence of a limit can be intuitively justified, and is sometimes related to the issue of multiple extensions. (See Section 4 and [Grove *et al.*, 1992b].) However, there are cases where the limit does not exist for what seem to be the "wrong" reasons. For example, if $\Pr^w_{N,\vec{\tau}}(\varphi|\mathsf{KB})$ oscillates between $\alpha + \tau_i$ and $\alpha - \tau_i$ for some $i$ as $N$ gets large, then the limit will not exist, although it "should" be $\alpha$, since the oscillations about $\alpha$ go to 0 as $\vec{\tau}$ gets small. We avoid such problems by considering the lim sup and lim inf, rather than the limit.[4]

**Definition 2.1 :** If $\lim_{\vec{\tau}\to\vec{0}} \liminf_{N\to\infty} \Pr^w_{N,\vec{\tau}}(\varphi|\mathsf{KB})$ and $\lim_{\vec{\tau}\to\vec{0}} \limsup_{N\to\infty} \Pr^w_{N,\vec{\tau}}(\varphi|\mathsf{KB})$ both exist and are equal, then the *degree of belief in $\varphi$ given* $\mathsf{KB}$, written $\Pr^w_\infty(\varphi|\mathsf{KB})$, is defined as the common limit; otherwise $\Pr^w_\infty(\varphi|\mathsf{KB})$ does not exist.

## 3 Default reasoning

As we mentioned above, we interpret default sentences such as "Birds typically fly" as statistical statements meaning "Almost all birds fly". Our formalism gives us a straightforward way to represent such a default, by writing $||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_i 1$.[5] Note, however, that if the agent has more exact information about the proportion of flying birds, then this information can also be expressed and used during reasoning (see Section 4).

We now review (and slightly extend) results from [Bacchus *et al.*, 1992] showing that this type of translation does in fact capture several important features of default reasoning. We stress that all the results in this section hold for our language in its full generality: the formulas can contain arbitrary non-unary predicates, and have nested quantifiers and proportion statements.

**Proposition 3.1 :** *The set* $D(\mathsf{KB}) = \{\varphi : \Pr^w_\infty(\varphi|\mathsf{KB}) = 1\}$ *contains* $\mathsf{KB}$ *and is closed under* valid implication *(i.e., if* $\models \theta \Rightarrow \psi$ *and* $\theta \in D(\mathsf{KB})$*, then* $\psi \in D(\mathsf{KB})$*).*

Hence, our system satisfies the minimal requirement for default reasoning, that it subsume standard deductive reasoning.

The next proposition shows that our approach goes beyond deductive inference to capture simple default inferences. In the following propositions, let $\vec{x} = \{x_1,\ldots,x_k\}$ and $\vec{c} = \{c_1,\ldots,c_k\}$ be sets of distinct variables and distinct constants, respectively.

**Proposition 3.2:** *Let* $\varphi(\vec{x}), \psi(\vec{x})$ *be formulas, where no constant in* $\vec{c}$ *appears in* $\varphi(\vec{x})$ *or* $\psi(\vec{x})$*. Then*

$$\Pr^w_\infty(\varphi(\vec{c})|\,\psi(\vec{c}) \wedge ||\varphi(\vec{x})|\psi(\vec{x})||_{\vec{x}} \approx_i \alpha) = \alpha.$$

For example, $\Pr^w_\infty(\mathsf{Fly}(\mathsf{Tweety})|\,||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_i 1 \wedge \mathsf{Bird}(\mathsf{Tweety})) = 1$. That is, we can make the standard inference about Tweety. Note that the proposition also holds when we have quantitative information, i.e., it holds for arbitrary $\alpha$.

Going beyond simple default reasoning, one sought-after property we obtain is a preference for more specific defaults.

[3] Note that we define degrees of belief only when $\varphi$ and $\mathsf{KB}$ are sentences, i.e., closed formulas.

[4] For any set $S \subset I\!R$, the infimum of $S$, $\inf S$, is the greatest lower bound of $S$. The *lim inf* of a sequence is the limit of the infimums; that is, $\liminf_{N\to\infty} a_N = \lim_{N\to\infty} \inf\{a_i\}_{i>N}$. The lim inf exists for any sequence bounded from below, even if the limit does not. The *lim sup* is defined analogously, where $\sup S$ denotes the least upper bound of $S$. If $\lim_{N\to\infty} a_N$ does exist, then $\lim_{N\to\infty} a_N = \liminf_{N\to\infty} a_N = \limsup_{N\to\infty} a_N$.

[5] We remark that, here and below, the actual choice of subscript for $\approx$ is unimportant. Typically, however, we capture different defaults by using different subscripts. Intuitively, the different subscripts correspond to different measurements or defaults of different strengths.

**Proposition 3.3:** *Suppose* KB *has the form* $\psi_1(\vec{c}) \wedge (||\varphi(\vec{x})|\psi_1(\vec{x})||_{\vec{x}} \approx_i \alpha) \wedge (||\varphi(\vec{x})|\psi_2(\vec{x})||_{\vec{x}} \approx_j \beta) \wedge \forall \vec{x}(\psi_1(\vec{x}) \Rightarrow \psi_2(\vec{x}))$, *where no constant in* $\vec{c}$ *appears in* $\varphi$, $\psi_1$, *or* $\psi_2$. *Then* $\Pr^w_\infty(\varphi(\vec{c})|\mathsf{KB}) = \alpha$.

For example, if $\mathsf{KB}^*$ is

$$\mathsf{Penguin(Opus)} \wedge ||\mathsf{Fly}(x)|\mathsf{Penguin}(x)||_x \approx_i 0 \wedge ||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_j 1 \wedge \forall x(\mathsf{Penguin}(x) \Rightarrow \mathsf{Bird}(x))$$

then $\Pr^w_\infty(\mathsf{Fly(Opus)}|\mathsf{KB}^*) = 0$. That is, we conclude that Opus the penguin does not fly, even though he is also a bird and birds generally do fly.

Another important property of our approach lies in its ability to treat as irrelevant some information that there is no reason to believe is relevant. This is particularly interesting given the fact it is notoriously hard to achieve specificity and a correct treatment of irrelevant information in the same framework (see the discussion and references in Section 5). There are a number of situations in which information will be treated as irrelevant by our approach, but here we restrict our attention to one special case:

**Proposition 3.4:** *Suppose* KB *has the form* $\psi(\vec{c}) \wedge (||\varphi(\vec{x})|\psi(\vec{x})||_{\vec{x}} \approx_i \alpha) \wedge \mathsf{KB}'$, *where no constant in* $\vec{c}$ *appears in* $\psi(\vec{x})$ *or* $\varphi(\vec{x})$, *and neither* $\psi$ *nor* $\mathsf{KB}'$ *mention any symbol in* $\varphi$. *Then* $\Pr^w_\infty(\varphi(\vec{c})|\mathsf{KB}) = \alpha$.

For example,

$$\Pr^w_\infty\left(\mathsf{Fly(Tweety)} \;\middle|\; \begin{array}{l}\mathsf{Bird(Tweety)} \wedge \\ ||\mathsf{Fly}(x)|\mathsf{Bird}(x)||_x \approx_i 1 \wedge \\ \mathsf{Yellow(Tweety)}\end{array}\right) = 1 .$$

That is, Tweety the yellow bird continues to fly. There is no information in KB about any correlation between the properties "yellow" and "fly"; hence Tweety's yellowness is treated as being irrelevant to his flying ability. Proposition 3.4 also shows that relevance is relative to a particular assertion $\varphi$. A property that is relevant to one assertion will not necessarily be relevant to another. For example, if we know that birds typically have beaks, we can conclude that Opus, a penguin, also has a beak, even though penguins typically do not fly (while birds typically do). More precisely, for the knowledge base $\mathsf{KB}^*$ above (relating to penguins, birds, and flying), we have:

$$\Pr^w_\infty\left(\mathsf{Beaked(Opus)} \;\middle|\; \begin{array}{l}\mathsf{KB}^* \wedge \mathsf{Bird(Opus)} \wedge \\ ||\mathsf{Beaked}(x)|\mathsf{Bird}(x)||_x \approx_i 1\end{array}\right) = 1 .$$

That is, Penguins is an exceptional subclass of birds with respect to flying but not with respect to having beaks. Proposition 3.4 also allows the agent to ignore those parts of KB that do not concern Tweety at all.

## 4 Qualitative versus Quantitative

Systems of direct inference frequently cannot use qualitative information such as "birds typically fly", whereas standard default logics generally cannot use quantitative information such as "90% of birds fly." Nevertheless, we often have both kinds of information available. One significant advantage of our approach is that it can use any combination of qualitative and quantitative information, supporting an entire spectrum of reasoning. Furthermore, in those cases where qualitative defaults are insufficient, our approach can often pinpoint the extra information required to reach a definite conclusion. To demonstrate, we examine two examples that are well-known to be problematic for pure default reasoning: the Lottery Paradox [Kyburg, 1961] and the Nixon Diamond [Reiter and Criscuolo, 1981].

In the Lottery Paradox, the assumption is that a large number of people buy tickets to a lottery in which there is only one winner. The standard assumption is that for any particular person $c$ we would like to conclude by default that $c$ does not win the lottery. This, however, seems to contradict the fact that someone must win it. In order to describe the problem in our framework, we assume for simplicity that the domain consists only of lottery ticket holders. Our knowledge base KB will consist of the single statement $\exists! x\, \mathsf{Winner}(x)$ (i.e., there is a unique winner). If we know the size of the lottery, say $N$, our degree of belief that the individual denoted by a particular constant $c$ wins the lottery is $\Pr^w_{N,\vec{\tau}}(\mathsf{Winner}(c)|\mathsf{KB}) = \frac{1}{N}$. Our degree of belief that *someone* wins will obviously be 1. These answers are clearly the "right" ones given our information. If, however, we do not know the exact number of ticket holders, but have only the qualitative information that this number is "large", then our degree of belief that $c$ wins the lottery is $\Pr^w_\infty(\mathsf{Winner}(c)|\mathsf{KB}) = 0$, although, as before, $\Pr^w_\infty(\exists x\, \mathsf{Winner}(x)|\mathsf{KB}) = 1$. Thus, we conclude by default that $c$ does not win the lottery for any constant $c$, although we still believe with full confidence that someone does win.

A major difficulty with using defaults is that they do not always provide sufficient information to reach a conclusion. A classical example is the problem of conflicting defaults, as demonstrated by the well-known *Nixon Diamond.* Suppose we have the following information:

$$\begin{array}{l}||\mathsf{Pacifist}(x)|\mathsf{Quaker}(x)||_x \approx_1 1 \wedge \\ ||\mathsf{Pacifist}(x)|\mathsf{Republican}(x)||_x \approx_2 0 \wedge \\ \mathsf{Quaker(Nixon)} \wedge \mathsf{Republican(Nixon)} .\end{array}$$

To simplify matters, we further assume that there is a unique individual who is both a Quaker and a Republican. We capture this by taking $\mathsf{KB}_1$ to consist of the above conjunction together with the formula $\exists! x\,(\mathsf{Quaker}(x) \wedge \mathsf{Republican}(x))$. Let $\varphi$ be the formula $\mathsf{Pacifist(Nixon)}$.

What should be our degree of belief in $\mathsf{Pacifist(Nixon)}$; that is, what is $\Pr^w_\infty(\varphi|\mathsf{KB}_1)$? It turns out that this limiting probability does not exist. This is because the limit is *non-robust*—its value depends on the way in which $\vec{\tau}$ goes to 0: if $\tau_1 \ll \tau_2$, so that the "almost all" in the first conjunct is much closer to "all" than the "almost none" in the second conjunct is closer to "none", then the limit is 1. Intuitively, in this case the information in the first conjunct is more precise and hence should be taken more seriously than the information in the second conjunct. Symmetrically, if $\tau_1 \gg \tau_2$, then the limit is 0. On the other hand, if $\tau_1 = \tau_2$, then the limit is 1/2.

The nonexistence of this limit is not simply a technical artifact of our approach. Rather, the fact that the limit fails to exist provides important information about the underlying incompleteness of our knowledge. It shows that in the presence of conflicting defaults, we often need more information about the precise nature of "almost all" and "almost none" to resolve the conflict; our approach pinpoints the type of information that would suffice to reach a decision. Note that our formalism does give us an explicit way to state that the defaults have equal strength, if we wish; namely, we can use $\approx_1$ to capture both default statements, rather than using $\approx_1$ and $\approx_2$. In this case, we get the answer 1/2, as expected. However, it is not always appropriate to conclude that the defaults have equal strength. We can easily extend our formalism to allow the user to prioritize defaults, by defining the relative size of the components $\tau_i$ of the tolerance vector.

If we have more quantitative information, then we can use it. For example, we may have information regarding the precise proportion of Quakers (resp., Republicans) who are pacifists. For example, assume that $\mathsf{KB}_2$ is

$$\begin{array}{l} \|\mathsf{Pacifist}(x)|\mathsf{Quaker}(x)\|_x \approx_1 \alpha \wedge \\ \|\mathsf{Pacifist}(x)|\mathsf{Republican}(x)\|_x \approx_2 \beta \wedge \\ \mathsf{Quaker}(\mathsf{Nixon}) \wedge \mathsf{Republican}(\mathsf{Nixon}) \wedge \\ \exists! x\,(\mathsf{Quaker}(x) \wedge \mathsf{Republican}(x)) \;, \end{array}$$

where $\{\alpha, \beta\} \neq \{0, 1\}$. In this case the limit does exist; we get $\Pr_\infty^w(\varphi|\mathsf{KB}_2) = \frac{\alpha\beta}{\alpha\beta+\bar{\alpha}\bar{\beta}}$, where $\bar{\alpha} = 1-\alpha$ and $\bar{\beta} = 1-\beta$.

Readers familiar with Dempster's rule of combination [Shafer, 1976] will note that this formula is precisely the result of combining the two probability functions that give probability $\alpha$ and $\beta$, respectively, to Nixon being a Pacifist. If we view the fact that Nixon is a Quaker as giving evidence of degree $\alpha$ in favor of Nixon being a pacifist, and the fact that Nixon is a Republican as giving evidence of degree $\beta$ in favor of Nixon being a pacifist, then our technique can be viewed as combining these two pieces of evidence. In the full paper, we show that our approach also captures more general instances of Dempster's rule of combination, and discuss why the appearance of the rule here is not coincidental.

Returning to the formula, notice that if $\alpha = 1$ and $\beta > 0$, then the limiting probability is 1, while if $\beta = 0$ and $\alpha < 1$, then the limiting probability is 0. That is, as expected, an "extreme" value will dominate. If $\alpha = \beta$, then the limiting probability is $\frac{\alpha^2}{\alpha^2+\bar{\alpha}^2}$. Thus, if 80% of Quakers are pacifists and 80% of Republicans are pacifists, the value of the limit would be around 0.94. This has a reasonable explanation: if we have two independent bodies of evidence, both supporting $\varphi$ quite strongly, when we combine them we should get even more support for $\varphi$.

## 5 Maximum entropy

In this section, we show how the approach of [Goldszmidt *et al.*, 1990] can be embedded in our framework. We begin by outlining $\epsilon$-semantics [Geffner and Pearl, 1990], on which the framework of [Goldszmidt *et al.*, 1990] is based. Consider a language consisting of propositional formulas (over some finite set of propositional variables $p_1, \ldots, p_k$) and default rules of the form $B \rightarrow C$ (read "$B$'s are typically $C$'s"), where $B$ and $C$ are propositional formulas. Let $\Omega$ be the set of $2^k$ *propositional worlds*, corresponding to the possible truth assignments to these variables. Given a probability distribution $\mu$ on $\Omega$, we define $\mu(B)$ to be the probability of the set of worlds where $B$ is true. We say that a distribution $\mu$ *$\epsilon$-satisfies* a default rule $B \rightarrow C$ if $\mu(C|B) \geq 1-\epsilon$.

A *parameterized probability distribution* (PPD) is a collection $\{\mu_\epsilon\}_{\epsilon>0}$ of probability distributions over $\Omega$, parameterized by $\epsilon$. A PPD $\{\mu_\epsilon\}_{\epsilon>0}$ $\epsilon$-satisfies a set $\mathcal{R}$ of default rules if for every $\epsilon$, $\mu_\epsilon$ $\epsilon$-satisfies every rule $r \in \mathcal{R}$. A set $\mathcal{R}$ of default rules $\epsilon$-entails $B \rightarrow C$ if for every PPD that $\epsilon$-satisfies $\mathcal{R}$, $\lim_{\epsilon\to 0} \mu_\epsilon(C|B) = 1$.

As shown in [Geffner and Pearl, 1990], $\epsilon$-entailment possesses a number of reasonable properties typically associated with default reasoning, including a preference for more specific information. However, there are a number of desirable properties that it does not have. Among other things, irrelevant information is not ignored. Pearl's notion of *1-entailment* [Pearl, 1990] strengthens $\epsilon$-entailment by allowing it to ignore irrelevant information in certain cases. However, it suffers from the problem that subclasses that are exceptional in one aspect are deemed exceptional in all aspects. In particular, using 1-entailment, we cannot conclude that Opus the penguin has a beak. Since $\epsilon$-entailment is equivalent to Lehmann and Magidor's *preferential entailment* [Lehmann and Magidor, 1992], and 1-entailment is equivalent to their *rational closure*, these approaches to default reasoning all suffer from difficulties when trying to combine specificity with irrelevance. We showed above that our approach does not suffer from this problem.

In order to obtain additional desirable properties, $\epsilon$-semantics is extended in [Goldszmidt *et al.*, 1990] by an application of the *maximum entropy principle* [Jaynes, 1957]. Instead of considering all possible PPD's, as above, only the PPD $\{\mu^*_{\epsilon,\mathcal{R}}\}_{\epsilon>0}$ of maximum entropy is considered (see [Goldszmidt *et al.*, 1990] for precise definitions and technical details). A rule $B \rightarrow C$ is an *ME-plausible consequence* of $\mathcal{R}$ if $\lim_{\epsilon\to 0} \mu^*_{\epsilon,\mathcal{R}}(C|B) = 1$. The notion of ME-plausible consequence is analyzed in detail in [Goldszmidt *et al.*, 1990], where it is shown to inherit all the nice properties of $\epsilon$-entailment (such as the preference for more specific information), while successfully ignoring irrelevant information. Equally importantly, algorithms are provided for computing the ME-plausible consequences of a set of rules in certain cases.

Although no explicit use is made of maximum entropy in our framework, there is a close connection between the random-worlds approach and maximum entropy *provided that the language consists only of unary predicates and constants*, as shown in [Grove *et al.*, 1992b]. These results can be extended to show that the approach of [Goldszmidt *et al.*, 1990] can be embedded in our framework in a straightforward manner. We simply convert all default rules $r$ of the form $B \rightarrow C$ into formulas of the form $\theta_r =_{\rm def} \|\psi_C(x)|\psi_B(x)\|_x \approx_1 1$, where $\psi_B$ is the formula obtained by replacing each occurrence of the propositional variable $p_i$ in $B$ with $P_i(x)$. Note that the formulas that arise under this conversion all use the same approximately equals relation $\approx_1$, since the approach of [Goldszmidt *et al.*, 1990] uses the same $\epsilon$ for all default rules. Moreover, they all involve only unary predicates. Under this translation, we can prove the following theorem, using techniques similar to [Grove *et al.*, 1992b].

**Theorem 5.1:** *Let c be a constant symbol. Using the translation described above, for any set $\mathcal{R}$ of defeasible rules, $B \rightarrow C$ is an ME-plausible consequence of $\mathcal{R}$ iff* $\Pr^w_\infty(\psi_C(c) | \bigwedge_{r\in\mathcal{R}} \theta_r \wedge \psi_B(c)) = 1$.

Thus, all the computational techniques and results described in [Goldszmidt *et al.*, 1990] carry over to this special case of our approach.

It is very encouraging that the results of [Goldszmidt *et al.*, 1990] can be arrived at in two quite different ways. Our result formalizes a connection between entropy and indifference, well known in other contexts like statistical thermodynamics, in the context of an agent reasoning by default. It shows that if one feels that it is reasonable for an agent to be indifferent between possibilities left open by its knowledge, then one has an independent reason for accepting the theory of irrelevance generated by maximum entropy.

It should also be noted that our approach, which does not appeal to entropy maximization directly, has the advantage of being much more general. Most importantly, it can deal sensibly with languages that have predicates of arbitrary arity. It is unlikely that an approach that uses entropy directly could be extended to deal such languages. Once we have even a single binary predicate in the language, all connection between our approach and maximum entropy disappears. As discussed in [Grove *et al.*, 1992b], we cannot even find a suitable probability space to take entropy over. Results of [Grove *et al.*, 1992a] showing that, with a binary predicate in the language, degrees of belief are in general uncomputable support the conjecture that there is none to be found.

## 6 Discussion and conclusions

We have shown that a logic that allows statistical and first-order assertions, together with a principled approach for obtaining degrees of belief from a knowledge base expressed in this logic, can give a general approach for capturing many aspects of default reasoning. Our framework has the added advantage of being able to deal with both default (qualititative) and statistical (quantitative) information. Our results demonstrate the close connection between default reasoning and direct inference.

We close by briefly discussing two criticisms that have been made of entropy-based reasoning systems: language and syntax dependence, and the treatment of causality [Pearl, 1988]. While the random-worlds method is *not* entropy-based, the relationship we observed in Section 5 suggests that similar problems may arise.

With regard to causality, [Goldszmidt *et al.*, 1990; Pearl, 1988] and [Hunter, 1989] have observed that knowledge about causal relationships greatly affects our intuitions concerning the "right" answers to various problems, and that the naive maximum entropy approaches do not take this causal information into consideration. We would argue that this only shows that this information is not properly captured by the straightforward encoding of defaults, and that we may therefore have to include information about causality when expressing defaults in the knowledge base. [Hunter, 1989] presents one possibility for encoding causal information *within* the maximum entropy approach. In [Bacchus *et al.*, 1993], we present a more general approach within the random-worlds framework, and show that it deals with many of the problematic aspects of causal reasoning.

The language problem is more subtle. Maximum-entropy methods can draw different conclusions from knowledge bases that seem to reflect the same information about the world. This is a serious issue, because the choice of the "right" representation of our information is not always clear. In general, we believe that the form in which our information is written down encodes knowledge—it reveals our biases and expectations. It is perfectly reasonable that our bias should affect inductive reasoning. In certain cases, our bias is sufficiently clear that the choice of representation becomes obvious. In physics, for example, the choice of language is sometimes based on the criterion of *time-invariance*. Moreover, in physics and in many other applications of maximum entropy, there is an objective "reality check"—we can compare the answers given by the formalism to reality, and thus independently verify the reasonableness of our representation. In many AI applications, however, there might not be an obvious representation, nor an appropriate reality check. In these cases, we will have to formulate criteria for choosing the right formal knowledge base, given a natural-language specification of our knowledge. This is an important research problem, which we intend to investigate. The fact that our approach can deal with causality leads us to hope that it will be able to deal with the language problem as well in a satisfactory way.

**Acknowledgements**

We would like to thank Moshe Vardi for helpful comments.

## References

[Adams, 1975] E. Adams. *The Logic of Conditionals.* D. Reidel, Dordrecht, Netherlands, 1975.

[Bacchus *et al.*, 1992] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistics to belief. In *Proc. National Conference on Artificial Intelligence (AAAI-92)*, pages 602–608, 1992.

[Bacchus *et al.*, 1993] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Forming beliefs about a changing world. In preparation, 1993.

[Bacchus, 1990] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge.* MIT Press, Cambridge, MA, 1990.

[Carnap, 1950] R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, Chicago, 1950.

[Carnap, 1952] R. Carnap. *The Continuum of Inductive Methods.* University of Chicago Press, Chicago, 1952.

[Geffner and Pearl, 1990] H. Geffner and J. Pearl. A framework for reasoning with defaults. In H. E. Kyburg, Jr., R. Loui, and G. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*. Kluwer Academic Press, Dordrecht, Netherlands, 1990.

[Goldszmidt *et al.*, 1990] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. In *Proc. National Conference on Artificial Intelligence (AAAI-90)*, pages 646–652, 1990.

[Grove *et al.*, 1992a] A. J. Grove, J. Y. Halpern, and D. Koller. Asymptotic conditional probabilities for first-order logic. In *Proc. 24th ACM Symp. on Theory of Computing*, pages 294–305, 1992.

[Grove *et al.*, 1992b] A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. In *Proc. 7th IEEE Symp. on Logic in Computer Science*, pages 22–33, 1992.

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

[Hunter, 1989] D. Hunter. Causality and maximum entropy updating. *International Journal of Approximate Reasoning*, 3(1):379–406, 1989.

[Jaynes, 1957] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[Johnson, 1932] W. E. Johnson. Probability: The deductive and inductive problems. *Mind*, 41(164):409–423, 1932.

[Koller and Halpern, 1992] D. Koller and J. Y. Halpern. A logic for approximate reasoning. In *Proc. Third International Conference on Principles of Knowledge Representation and Reasoning (KR '92)*, pages 153–164, 1992.

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

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

[Laplace, 1820] P. S. de Laplace. *Essai Philosophique sur les Probabilités.* 1820. English translation is *Philosophical Essay on Probabilities*, Dover Publications, New York, 1951.

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

[Levi, 1980] I. Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, MA, 1980.

[McCarthy, 1986] J. McCarthy. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28:86–116, 1986.

[Neufeld, 1989] E. Neufeld. Defaults and probabilities; extensions and coherence. In *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 312–323, 1989.

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

[Pearl, 1989] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 505–516, 1989. Reprinted in *Readings in Uncertain Reasoning*, G. Shafer and J. Pearl (eds.), Morgan Kaufmann, San Mateo, CA, 1990, pp. 699–710.

[Pearl, 1990] J. Pearl. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In M. Vardi, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, pages 121–135. Morgan Kaufmann, San Mateo, CA, 1990.

[Pollock, 1990] J. L. Pollock. *Nomic Probabilities and the Foundations of Induction.* Oxford University Press, Oxford, U.K., 1990.

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

[Reiter and Criscuolo, 1981] R. Reiter and G. Criscuolo. On interacting defaults. In *Proc. Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, pages 270–276, 1981.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Salmon, 1971] W. Salmon. *Statistical Explanation and Statistical Relevance.* University of Pittsburgh Press, Pittsburgh, 1971.

[Shafer, 1976] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.