Generating Degrees of Belief from Statistical Information: An Overview*

Fahiem Bacchus[1], Adam J. Grove[2], Joseph Y. Halpern[3], Daphne Koller[4]

[1] Dept. of Computer Science, University of Waterloo, Waterloo, Canada, N2L 3G1
[2] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
[3] IBM Almaden Research Center, San Jose, CA 95120
[4] Dept. of Computer Science, Stanford University, Stanford, CA 94305

Abstract. Consider an agent (or expert system) with a knowledge base KB that includes statistical information (such as "90% of patients with jaundice have hepatitis"), first-order information ("all patients with hepatitis have jaundice"), and default information ("patients with jaundice typically have a fever"). A doctor with such a KB may want to assign a degree of belief to an assertion φ such as "Eric has hepatitis". Since the actions the doctor takes may depend crucially on this degree of belief, we would like to specify a mechanism by which she can use her knowledge base to assign a degree of belief to φ in a principled manner. We have been investigating a number of techniques for doing so; in this paper we give an overview of one of them. The method, which we call the *random worlds* method, is a natural one: For any given domain size N, we consider the fraction of models satisfying φ among models of size N satisfying KB. If we do not know the domain size N, but know that it is large, we can approximate the degree of belief in φ given KB by taking the limit of this fraction as N goes to infinity. As we show, this approach has many desirable features. In particular, in many cases that arise in practice, the answers we get using this method provably match heuristic assumptions made in many standard AI systems.

Consider an agent (or expert system) with a knowledge base KB that includes statistical information (such as "90% of patients with jaundice have hepatitis"), first-order information ("all patients with hepatitis have jaundice"), and default information ("patients with jaundice typically have a fever"). A doctor with such a KB may want to assign a degree of belief to an assertion φ such as "Eric has hepatitis". Since the actions the doctor takes may depend crucially on this degree of belief, we would like to specify a mechanism by which she can use her knowledge base to assign degrees of belief to φ in a principled manner. We have

* 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. During this work, Grove was at Stanford University, and was supported by an IBM Graduate Fellowship. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

been investigating a number of techniques for doing so. In this paper we give an overview of one of them, and give pointers to the literature for the reader interested in further details.

Our method, which we call the *random worlds* method, can be viewed as a particular realization of the *principle of insufficient reason* [Kri86] or *the principle of indifference* [Key21]. This principle states that all possibilities should be given equal probability, and was regarded as one of the basic principles of probability theory by the earliest workers on probability theory, such as Bernoulli and Laplace [Hac75]. We use this idea to assign equal degrees of belief to all basic "situations" consistent with the known facts. The question is, what is a situation?

In many applications, including the one of most interest to us, it makes sense to consider finite domains only. Assume, without loss of generality, that the domain is $\{1, \ldots, N\}$ for some natural number N. Then, in the random-worlds method, we consider the situations of interest to be first-order models over this domain. Using the principle of indifference, we assume that all these models, or worlds, are equally likely. To assign a degree of belief to a sentence φ given a knowledge base KB, we simply consider the fraction of worlds satisfying KB that also satisfy φ. In many respects, what we are doing here can be viewed as an instance of the general paradigm of Bayesian reasoning, in which one assumes a prior probability distribution over a space of possibilities and calculates a posterior distribution by conditioning on what is known. We take the probability space to consist of all worlds with domain $\{1, \ldots, N\}$, use a prior that assigns them equal probability, and then condition on KB.

One problem with the approach as stated so far is that, in general, we do not know the domain size N. Typically, however, N is known to be large. We therefore approximate the degree of belief for the true, but unknown, N, by computing the value of this degree of belief as N grows large. We note that this method is related to earlier work of [Joh32] and Carnap [Car50, Car52].

We said earlier that we expect the knowledge base to contain statistical information and default rules, as well as first-order facts. This suggests that the language we intend to use is richer than first-order logic. This is indeed the case.

We define a statistical language $\mathcal{L}^{\approx}$, which is a variant of a language designed by Bacchus [Bac90]. $\mathcal{L}^{\approx}$ augments standard first-order logic with a form of statistical quantifier. For a formula $\psi(x)$, the term $||\psi(x)||_x$ is a *proportion expression*. It will be interpreted as a rational number between 0 and 1 that represents the proportion of domain elements satisfying $\psi(x)$. We actually allow an arbitrary set of variables in the subscript and in the formula ψ. Thus, for example, $||Child(x, y)||_x$ describes, for a fixed y, the proportion of domain elements that are children of y; $||Child(x, y)||_y$ describes, for a fixed x, the proportion of domain elements whose child is y; and $||Child(x, y)||_{\{x,y\}}$ describes the proportion of pairs of domain elements that are in the child relation. We also allow proportion expressions of the form $||\psi(x)|\theta(x)||_x$, which we call *conditional proportion expressions*. Such an expression is intended to denote the proportion of domain elements satisfying ψ from among those elements satisfying θ. Finally, any rational number is also considered to be a proportion expression, and the set of proportion expressions is closed under addition and multiplication.

One important difference between our syntax and that of [Bac90] 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, which is surely not an intended implication. We therefore use the approach described in [GHK92b, KH92], 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"). For example, we can express the statement "90% of birds fly" by the *proportion formula* $||Fly(x)|Bird(x)||_x \approx_1 0.9$. 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 $||Fly(x)|Bird(x)||_x \approx_1 1 \wedge ||Fly(x)|Bat(x)||_x \approx_2 1$ says that both $||Fly(x)|Bird(x)||_x$ and $||Fly(x)|Bat(x)||_x$ are approximately 1, but the notion of "approximately" may be different in each case.

The use of approximate equality has another significant advantage: it lets us express default information. We give a statement such as "Birds typically fly" a statistical interpretation, viewing it as saying "Almost all birds fly". Our formalism gives us a straightforward way to represent such a default, by writing $||Fly(x)|Bird(x)||_x \approx_i 1$ for some i. (This interpretation is closely related to other approaches applying probabilistic semantics to nonmonotonic logic; see Pearl [Pea89] for an overview. However, all these other approaches are essentially propositional in nature. While they use the statistical interpretation as a motivation for using probabilities, none make explicit use of statistical assertions.)

We give semantics to these formulas with respect to a pair $(\mathcal{W}, \overline{\tau})$, where $\mathcal{W}$ is a *world*, or first-order model, with domain $\{1, \ldots, N\}$ for some N, and $\overline{\tau}$ is a tolerance vector of the form $\langle \tau_1, \tau_2, \ldots \rangle$, with $\tau_i > 0$. Intuitively, we use $\overline{\tau}$ to interpret approximate equality statements. Thus, $(\mathcal{W}, \overline{\tau}) \models \zeta \approx_i \zeta'$ if the values of ζ and ζ' are within τ_i of each other. For further details of the semantics see [BGHK93b, BGHK93a].

Given N and $\overline{\tau}$, we define $\#worlds_N^{\overline{\tau}}(\chi)$ to be the number of worlds $\mathcal{W}$ over the domain $\{1, \ldots, N\}$ such that $(\mathcal{W}, \overline{\tau}) \models \chi$. Since we want to view each world as equally likely, the degree of belief in φ given KB over worlds of size N given $\overline{\tau}$ becomes:

$$\Pr^w_{N,\overline{\tau}}(\varphi|KB) = \frac{\#worlds_N^{\overline{\tau}}(\varphi \wedge KB)}{\#worlds_N^{\overline{\tau}}(KB)}.$$

If $\#worlds_N^{\overline{\tau}}(KB) = 0$, this degree of belief is not well-defined.

As we said earlier, typically, we know neither N nor $\overline{\tau}$ exactly. All we know is that N is "large" and that $\overline{\tau}$ is "small". Thus, we would like to take our *degree of belief* in φ given KB to be $\lim_{\overline{\tau}\to\overline{0}} \lim_{N\to\infty} \Pr^w_{N,\overline{\tau}}(\varphi|KB)$.[5] This definition, however, is not sufficient; the limit may not exist. We observed above that $\Pr^w_{N,\overline{\tau}}(\varphi|KB)$ is not always well-defined. In particular, it may be the case that for certain values of $\overline{\tau}$, $\Pr^w_{N,\overline{\tau}}(\varphi|KB)$ is not well-defined for arbitrarily large N. To deal with this issue, we assume for the remainder of this paper that the KB is *eventually consistent*: for all sufficiently small $\overline{\tau}$ and sufficiently large N, $\#worlds^{\overline{\tau}}_N(KB) > 0$. Among other things, eventual consistency implies that the KB is satisfiable in finite domains of arbitrarily large size. For example, a KB stating that "there are exactly 7 domain elements" is not eventually consistent.

Even if KB is eventually consistent, the limit may not exist. For example, it may be the case that $\Pr^w_{N,\overline{\tau}}(\varphi|KB)$ oscillates between $\alpha+\tau_i$ and $\alpha-\tau_i$ for some i as N gets large. In this case, for any particular $\overline{\tau}$, the limit as N grows will not exist. However, it seems as if the limit as $\overline{\tau}$ grows small "should", in this case, be α, since the oscillations about α go to 0. We avoid such problems by considering the *lim sup* and *lim inf*, rather than the limit.[6] Thus we have

Definition 1. If

$$\lim_{\overline{\tau}\to\overline{0}} \liminf_{N\to\infty} \Pr^w_{N,\overline{\tau}}(\varphi|KB) \text{ and } \lim_{\overline{\tau}\to\overline{0}} \limsup_{N\to\infty} \Pr^w_{N,\overline{\tau}}(\varphi|KB)$$

both exist and are equal, then the *degree of belief in φ given KB*, written $\Pr^w_\infty(\varphi|KB)$, is defined as the common limit; otherwise $\Pr^w_\infty(\varphi|KB)$ does not exist.

We call this method of computing degrees of belief the *random-worlds method*. What do we gain from using the random worlds method? For one thing, many of the properties that have been intuitively viewed as desirable when computing degrees of belief and in doing default reasoning follow as theorems. We briefly describe a few of these properties here; a more thorough discussion can be found in [BGHK93b, BGHK93a].

First, suppose that given a knowledge base KB, we believe an assertion φ with probability 1. We can interpret this as saying that φ is a default conclusion from KB. Suppose we write $KB \mathrel{|\!\!\approx} \varphi$ if $\Pr^w_\infty(\varphi|KB) = 1$. Then $\mathrel{|\!\!\approx}$ is a nonmonotonic entailment relation. For example, it can be shown that if KB is

$$||Fly(x)|Bird(x)||_x \approx_1 1 \wedge Bird(Tweety),$$

[5] Notice that the order of the two limits over $\overline{\tau}$ and N is important. If the limit $\lim_{\overline{\tau}\to\overline{0}}$ appeared last, then we would gain nothing by using approximate equality, since the result would be equivalent to treating approximate equality as exact equality.

[6] 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, using least upper bounds. Notice that if $\lim_{N\to\infty} a_N$ exists, then $\lim_{N\to\infty} a_N = \liminf_{N\to\infty} a_N = \limsup_{N\to\infty} a_N$.

then $KB \mathrel{|\!\!\approx} \mathit{Fly}(\mathit{Tweety})$. But if KB' is

$$KB \wedge ||\mathit{Fly}(x)|\mathit{Penguin}(x)||_x \approx_2 0 \wedge \mathit{Penguin}(\mathit{Tweety}),$$

then $KB' \mathrel{|\!\!\approx} \neg \mathit{Fly}(\mathit{Tweety})$. Thus, if all we know is that birds typically fly and that Tweety is a bird, then we believe (with degree of belief 1) that Tweety flies. On the other hand, if we get the extra information that Tweety is a penguin and that penguins do not typically fly, then we have degree of belief 1 that Tweety does not fly. This example also shows how the random worlds method prefers more *specific* information: the more specific information about penguins automatically overrides the information about birds.

Although $\mathrel{|\!\!\approx}$ is nonmonotonic, it does satisfy a number of important properties. For example, if $KB \models \varphi$ then $KB \mathrel{|\!\!\approx} \varphi$, so we have degree of belief 1 in all the logical consequences of the knowledge base. In fact, $\mathrel{|\!\!\approx}$ satisfies the properties postulated by Kraus, Lehmann, and Magidor [KLM90, Leh89] to be appropriate for a nonmonotonic consequence relation.

As we saw above, the random-worlds method can reason from statistical information (which in the examples above express default rules) to conclusions about particular individuals. This is an important advantage of this framework. A number of basic criteria for this type of reasoning are generally agreed upon. The first, and least controversial, is basic *direct inference*, where we have a single *reference class* in KB that is precisely the "right one". Formally, assume that (a) we are interested in the query $\varphi(c)$, (b) all we know about the individual c is $\psi(c)$ and (c) we have statistics about the proportion of ψ's that satisfy φ. Then direct inference says that we should use this statistical information to generate a degree of belief in $\varphi(c)$. For example, assume that all we know about Eric is that he exhibits jaundice and ψ represents the class of patients with jaundice. If we know that 80% of patients with jaundice exhibit hepatitis, then basic direct inference will dictate a degree of belief of 0.8 in Eric having hepatitis. We would, in fact, like this to hold regardless of other information we might have in the knowledge base. For example, we may know the proportion of hepatitis among patients in general, or that patients with jaundice and fever typically have hepatitis. But if all we know about Eric is that he has jaundice, we would still like to use the statistics for this class, regardless of this additional information.

This intuition is formalized in the following theorem, which generalizes the principle to properties and classes dealing with more than one individual at a time. In the following let $\overline{x} = \{x_1, \ldots, x_k\}$ and $\overline{c} = \{c_1, \ldots, c_k\}$ be sets of distinct variables and distinct constants, respectively.

Theorem 2. *Let KB be a knowledge base of the form $\psi(\overline{c}) \wedge KB'$, and assume that for all sufficiently small tolerance vectors $\overline{\tau}$:*

$$KB[\overline{\tau}] \models ||\varphi(\overline{x})|\psi(\overline{x})||_{\overline{x}} \in [\alpha, \beta].$$

If no constant in $\overline{c}$ appears in KB', in $\varphi(\overline{x})$, or in $\psi(\overline{x})$, then $\Pr^w_\infty(\varphi(\overline{c})|KB) \in [\alpha, \beta]$.

We remark that the preference for more specific information alluded to above follows immediately from this theorem.

Other results in this spirit are also proved in [BGHK93b, BGHK93a]. For example, it is shown that we can often ignore seemingly irrelevant information. In particular, even if we do not have exactly the "right" reference class, we can often use the smallest reference class that is applicable. So if we know that Eric has brown hair as well as having jaundice, we will be able to disregard Eric's hair color. (Notice that knowing about Eric's hair color prevents us from using Theorem 2, which is why we need to use results about irrelevance in this case.)

Random worlds also deals well with situations where the statistical information from more than one reference class applies to a query. For example, in the well-known *Nixon Diamond* problem [RC81], we have statistics for the occurrence of pacifists in the class of Republicans, and very different statistics for the occurrence of pacifists within the class of Quakers. We are interested in assigning a degree of belief to $Pacifist(Nixon)$, where we know that Nixon is both a Republican and a Quaker. We can show that random worlds essentially treats the two pieces of statistical information as independent pieces of evidence; in fact, the degree of belief resulting obtained by random worlds is essentially equivalent to combining evidence using Dempster's rule of combination [Sha76].

We should point out that the random-worlds method does not give the uncontroversially most intuitive answer in every example. Indeed, it is unlikely any method could, since peoples' intuitions often disagree. For example, it can be shown (using a maximum entropy computation as described below) that if the KB is $0 \leq \|P(x)\|_x \leq 0.6$, then $\Pr^w_\infty(P(c)|KB) = .5$. While one can give some strong arguments to support this answer, it also seems that the value 0.3 is at least as reasonable. One might also argue that an interval valued degree of belief, $[0, 0.6]$, is appropriate. We are currently investigating other methods for assigning degrees of belief, also based on the principle of indifference, that address these issues. (See [BGHK92] for some discussion.)

Given all the nice properties of the random-worlds method, it is reasonable to ask how hard it is to compute degrees of belief. As is shown in [GHK92a], all questions regarding degrees of belief are undecidable in general, even if the KB is purely first-order (and so does not include any statistical information). However, if we assume that KB mentions only *unary* predicates and constants, then the situation becomes much better. Indeed, as shown in [GHK92b], we can typically compute degrees of belief using a maximum entropy computation. Interestingly, it seems that all connection to maximum entropy is lost once our knowledge base contains even a single binary predicate.

While the restriction to unary predicates may seem severe, it is not so unreasonable in practice. The properties *Penguin*, *Bird*, *Fly*, symptoms, and diseases are all unary. Indeed, a case can be made that all of statistics is mainly concerned with unary predicates over *basic units* such as individuals or households. Moreover, results such as Theorem 2 (which hold for the full language—the formulas involved can have predicates of arbitrary arity) give us reason to hope that for many knowledge bases that arise in practice, even ones that go beyond unary predicates, computing degrees of belief for assertions of interest may be tractable.

To summarize, we believe that the random-worlds method is a principled—and very powerful—approach for assigning degrees of belief with a number of attractive features. These include:

- It generalizes reasoning paradigms such as probabilistic logic [Nil86], ϵ-semantics [Pea89], and maximum entropy [Jay78].
- It provably satisfies many well-known reasoning heuristics, such as preference for specific information, indifference to irrelevant information, and Dempster's rule for combining evidence. (See [BGHK93b] for more details of these properties, and their relevance to default reasoning.)
- There is a computational technique for computing degrees of belief with unary KB's (using maximum entropy).
- It provides a rich framework, that can handle quantitative, qualitative, and default information.
- Unlike maximum entropy, the random-worlds method continues to make sense and give reasonable answers in the nonunary case.

References

[Bac90] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge.* MIT Press, Cambridge, MA, 1990.

[BGHK92] 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.

[BGHK93a] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Generating degrees of belief from statistical information. In preparation, 1993.

[BGHK93b] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Statistical foundations for default reasoning. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, 1993.

[Car50] R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, Chicago, 1950.

[Car52] R. Carnap. *The Continuum of Inductive Methods.* University of Chicago Press, Chicago, 1952.

[GHK92a] 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.

[GHK92b] 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.

[Hac75] I. Hacking. *The Emergence of Probability.* Cambridge University Press, Cambridge, UK, 1975.

[Jay78] E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, MA, 1978.

[Joh32] W. E. Johnson. Probability: The deductive and inductive problems. *Mind*, 41(164):409–423, 1932.

[Key21] J. M. Keynes. *A Treatise on Probability.* Macmillan, London, 1921.

[KH92] D. Koller and J. Y. Halpern. A logic for approximate reasoning. In B. Nebel, C. Rich, and W. Swartout, editors, *Proc. Third International Conference on Principles of Knowledge Representation and Reasoning (KR '92)*, pages 153–164, 1992.

[KLM90] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Kri86] J. von Kries. *Die Principien der Wahrscheinlichkeitsrechnung und Rational Expectation.* Freiburg, 1886.

[Leh89] D. Lehmann. What does a conditional knowledge base entail? In *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 212–222, 1989.

[Nil86] N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[Pea89] 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.

[RC81] R. Reiter and G. Criscuolo. On interacting defaults. In *Proc. Seventh International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 270–276, 1981.

[Sha76] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.