Report 84-42
Stanford -- KSL

Does Probability Have a Place in Non-monotonic Reasoning?
Matthew L. Ginsberg,
Dec 1984

card 1 of 1

Stanford Heuristic Programming Project
Working paper No. HPP-84-42

December 1984

Does Probability Have a Place in Non-monotonic Reasoning?

Matthew L. Ginsberg

COMPUTER SCIENCE DEPARTMENT
Stanford University
Stanford, California 94305

Does Probability Have a Place in Non-monotonic Reasoning?

Abstract

Arguments are presented in favor of the answer "yes". The intuitive appeal (or lack thereof) of probabilities is considered briefly. The theoretical adequacies of probabilistic methods are investigated by considering them in light of McCarthy's "typology of uses of non-monotonic reasoning." A quantitative approach which overcomes the usual need for *a priori* probabilities is presented. Some of the practical advantages of using probabilities in a production system are described.

§1. Introduction

The panel on uncertain reasoning at AAAI-84 discussed the question of whether or not implementations of non-monotonic reasoning should be probabilistic. A variety of (generally unsupported) claims were made to the effect that probabilities are unintuitive, that the numbers needed are unavailable, and that the method generally is inappropriate. The counterclaims that probabilities are intuitive, available and appropriate were similarly unsupported.

My intention here is to present some results that deal with these questions. Let me stress that it is *precisely* the question posed in that last paragraph that interests me: Should probabilities be used to implement non-monotonic reasoning systems? The easier question of whether probabilities *can* be used to implement some types of non-monotonic reasoning has been answered rather conclusively by MYCIN and its offspring; more difficult questions involving the nature or definition of probability itself have been grappled with by philosophers for centuries, and I am content to leave them to it.

I will attempt to address the issues of whether the numbers required by a probabilistic theory can in general be made available to a reasoning system, and whether or not probabilistic methods are effective. The first of these is principally a theoretical issue, while the second is more one of pragmatics.

§2. Intuitive appeal

This is a philosophical matter more than anything else, and as such is outside the proper scope of this paper. The most common argument for probabilities is that they provide a convenient method for combining conflicting evidence; it is argued on the other hand that non-numeric methods such as rule ordering bear more resemblance to those we use ourselves when considering conflicting default rules.

This is not convincing. The fact that we have no access to probabilistic information regarding the merits (or lack thereof) of various rules of inference is at best weak evidence that we do not represent the information in this fashion at some level. There is a great deal of low-level information involved in our thinking processes to which we have no direct access.

In any event—whether probabilities are fundamentally intuitive for us or not—they provide a numerical formalism that *is* intuitive for a computer. The manipulations needed to combine probabilistic evidences are precisely those that machines are good at, and I can see no reason not to exploit these abilities.

§3. Theoretical adequacy

McCarthy [McCarthy 1984] has suggested seven distinct uses for a non-monotonic inference procedure. My initial intention is to discuss whether or not probabilistic methods are capable of dealing with each of these.

(1) "As a very streamlined expression of probabilistic information when numerical probabilities ... are unobtainable." One suspects that probabilities will not suffice for this purpose.

(2) Auto-epistemic reasoning: "If I had a brother, I'd know it." Perhaps probabilities can be used effectively here; in any event, the general nature of auto-epistemic reasoning is sufficiently unclear that more fundamental questions about its nature need to be answered before any specific non-monotonic implementation can be applied to it. Moore is working on this [Moore 1984].

(3) As a representation of policy: "The meeting will be on Wednesday unless another decision is explicitly made." This example is due to Doyle.

(4) In commonsense science: "An object will continue in a straight line if nothing interferes with it."

These last two usages are concerned with necessity of marking a conclusion as tentative in the sense that subsequent information may reverse it. Probabilities associated to all statements naturally provide such a marking scheme, although it is possible that the extra information they convey may be inappropriate. In at least one of McCarthy's examples, however, this is not the case—probabilities can encode for us the reliability of the individual scheduling the meeting.

(5) As a communications convention: there appears to be a general understanding that if a default rule is violated, and this information is relevant to a communication act, the information should be explicitly included in the communication.

(6) As a database convention: the closed world assumption is a typical example. If a statement of a certain type is not a logical consequence of the information in the database, its negation is assumed.

These two usages are also similar. They each involve situations where it is necessary to encode default rules without necessarily including quantitative information about the reliability of these rules. Again, probabilities provide a natural mechanism for recording the fact that a given rule is true only by default. McCarthy does point out that non-numerical methods allow the introduction of default rules which are generally *false*; he does not, however, give any examples of situations where the ability to eventually assume something that is likely to be incorrect is a useful one. Note that this latter ability is very different from the ability to realize that a conclusion is tentative—if there is a dangerous flaw in the design of a nuclear reactor, it is important that we realize that the flaw is there, but *not* necessarily that we assume the reactor will have an accident.

(7) As a rule of conjecture. Here, probabilities come into their own. They lead to analyses of the form, "What if Tweety can fly?" and have the additional property that they give us a natural way to order such conjectures. There may be reasons to prefer the investigation

of less likely conjectures over that of more likely ones, but the likelihood information, if available, will generally be extremely useful.

§4. A priori probabilities

Another standing objection to the use of probabilities in AI systems corresponds to the question, "Where do the numbers come from?" Bayesian methods require the existence of initial estimates for the probabilities in question, and it seems impossible to arrive at these estimates without a great deal of knowledge about the domain being considered.

I have a great deal of sympathy with this objection. It has been pointed out, however, that by considering *ranges* of probabilities instead of specific values, it is possible to encode information not only about the strength of our belief in a given proposition, but also about our confidence in our estimate of that strength [Dempster 1968 or Shafer 1976]. A precise formulation of this observation will be the principle result of this section.

When we think of a statement as corresponding not to a precise probability but to a range $[x, y]$, we can think of $y - x$ as corresponding to the uncertainty we have in our probabilistic estimate. Thus a specific range $\{x\} = [x, x]$ implies complete confidence in our probabilistic knowledge, while the maximal range $[0, 1]$ corresponds to complete ignorance—the statement that a certain probability p lies in the range $0 \leq p \leq 1$ has no informational content at all.

More generally, a probability range $[x, y]$ with $x \neq y$ corresponds to partial knowledge. Furthermore, it is possible to use Dempster-Shafer theory to combine probability ranges of this sort; an application of this to semantic nets is described in [Ginsberg 1984a].

In order to see how to obtain the ranges from observational data (or the lack thereof), suppose that the probability of some specific default rule is p, although this value need not be known to us. Now fix some "gullibility" $g \in [0, 1]$, and suppose that we test the default rule experimentally t times. Then there is some $p_{\min}(p, t, g)$ such that the probability of our observing no more than $tp_{\min}$ successful applications of the default rule among the t trials is equal to g. Intuitively, if the "real" probability is p, we require that the chance that the *observed* probability be at least $p_{\min}$ be at least g. Thus if $g = 0$, we get the extremely cautious approximation $p_{\min} = 0$.

We can define $p_{\max}(p, t, g)$ similarly. Having done so, if some default rule D has been tested t times with s successes, we can approximate the overall probability to be assigned to the rule by s/t, and consider the probabilistic range $[p_{\min}(s/t, t, g), p_{\max}(s/t, t, g)]$. Conversely, given a probability range $[x, y]$, we can use this expression to recover s and t (for g fixed).

The details of the calculation require us to solve a familiar problem from probability theory: Given a series of t trials in an experiment where the probability of success on each trial is p, what is the probability that the *observed* probability of success will be in the range $[p_{\min}, p_{\max}]$? This problem is discussed in [von Mises 1964], among other places; there is no exact solution in closed form, but results can be obtained either by using Tchebychev's approximation or by approximating the relevant binomial distribution with a Gaussian. Tchebychev's approximation gives

$$1 - \frac{p_{\min}(1 - p_{\min})}{t(p_{\min} - s/t)^2} = 1 - g,$$

and identically for $p_{\max}$. We therefore need to solve

$$(p - \frac{s}{t})^2 = \frac{p(1 - p)}{tg}, \quad \text{or} \tag{1}$$

$$p^2 t(1 + tg) - pt(1 + 2sg) + s^2 g = 0, \tag{2}$$

leading to

$$p = \frac{1 + 2sg \pm \sqrt{1 + 4sg(1 - \frac{s}{t})}}{2(1 + tg)}. \tag{3}$$

If $g = 0$, we get a probability range of $[0, 1]$ independent of s and t (not very gullible at all!), as we do if $s = t = 0$. In the large t limit, we get the singleton s/t as expected—our confidence in our estimate increases as the amount of data does.

Alternatively, we can use the Gaussian approximation, so that we need to solve

$$2\Phi\left(\frac{|pt - s|}{\sqrt{tp(1 - p)}}\right) = 1 - g, \tag{4}$$

where $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt,$$

with $\Phi(0) = 0$ and $\Phi(\infty) = .5$. If we set $\tilde{g} \equiv \Phi^{-1}((1-g)/2)$, (4) becomes

$$(p - \frac{s}{t})^2 = \frac{\tilde{g}^2 p(1-p)}{t}.$$

This is identical to (1) with g replaced with $1/[\Phi^{-1}((1-g)/2)]^2$.

To solve the inverse problem (in either case), suppose we are given a probability range $[x, y]$. Set $\alpha = x + y$ and $\beta = xy$. Then from (2) we have

$$\alpha = \frac{1+2sg}{1+tg}, \qquad \beta = \frac{s^2 g}{t(1+tg)}$$

$$\frac{\alpha}{\beta} = \frac{(1+2sg)t}{s^2 g}, \qquad \text{so}$$

$$t = \frac{s^2 \alpha g}{(1+2sg)\beta}, \qquad \text{and}$$

$$gs^2 = \beta t(1+tg) = \frac{\beta s^2 \alpha g}{(1+2sg)\beta}\left[1 + \frac{s^2 \alpha g^2}{(1+2sg)\beta}\right].$$

This leads to

$$(1+2sg)^2 = \alpha(1+2sg) + s^2\alpha^2 g^2/\beta, \text{ so that}$$

$$s = \frac{\beta(\alpha-2) - \alpha\sqrt{\beta(1+\beta-\alpha)}}{g(4\beta-\alpha^2)} \qquad \text{and} \qquad t = \frac{2sg + (1-\alpha)}{\alpha g}. \tag{5}$$

If we set $\bar{x} = 1 - x$ and $\bar{y} = 1 - y$, this becomes

$$\begin{aligned} s &= \frac{xy(\bar{x}+\bar{y}) + (x+y)\sqrt{xy\bar{x}\bar{y}}}{g(x-y)^2} \\ t &= \frac{(x\bar{x} + y\bar{y}) + 2\sqrt{xy\bar{x}\bar{y}}}{g(x-y)^2}, \end{aligned} \tag{6}$$

where we should replace g with $\tilde{g} = 1/[\Phi^{-1}((1-g)/2)]^2$ in the Gaussian case. Note that for $g = 0$, $1/[\Phi^{-1}((1-g)/2)]^2 = 0$; for $g = 1$, $1/[\Phi^{-1}((1-g)/2)]^2 = \infty$. Thus the only difference in substance between the two approximations is in the fact that g in (3) or (6) runs over the range $[0,1]$ while $\tilde{g}$ in the Gaussian versions can take any value in $[0,\infty]$.

As an example, suppose $s = 3$, $t = 4$ and $g = 5/12$. Then we get a probability range $[3/8, 15/16]$ from (3). Now (6) reconstructs $s = 3$, $t = 4$. If the next trial produces success, the range becomes $[.45, .95]$; if failure, $[.29, .85]$.

In the Gaussian approximation, we have $1/[\Phi^{-1}((1-g)/2)]^2 = 1.5$, leading to an initial range of $[.55, .88]$, and subsequent ranges of $[.62, .95]$ (success) or $[.41, .76]$ (failure).

§5. Implementation issues

Existing formalisms of non-monotonic reasoning generally proceed by attempting to determine whether or not a default inference will be valid before drawing it. Thus, before concluding that the bird Tweety can fly, we first try to prove that he can't; if the proof fails, we draw the inference that he can.

There are well known difficulties with this. The first is that the problem of proving that Tweety can't fly is only semi-decidable, and implementations of this scheme therefore tend to be painstakingly slow (at best!). The second is that the need to use the appearance of a new datum, such as the fact that Tweety is an ostrich, to reverse an earlier conclusion requires the introduction of a new formalism, such as truth maintenance [Doyle 1979]. Probabilities provide a way around both of these difficulties by marking the conclusion of a proof to indicate that it may be subsequently reversed in the presence of stronger contradictory evidence. We are not claiming here that they can replace a truth maintenance system; it will still be necessary to store information regarding either the use to which information has been put (in a forward-chaining system) or the source from which information was obtained (in a backward-chaining one).

In the presence of an adequate rule for probabilistic combination, many of the attractive properties of a reason maintenance system can be incorporated into a probabilistic one. When the truth value of some conclusion changes as a result of the appearance of new evidence, earlier inferences made using this conclusion can be repeated, with the change in probability therefore propagating to the results that were derived from it.

5.1 Tags

Suppose that we are in fact considering ranges of probabilities instead of specific values, and let P be the set of all closed subintervals of $[0,1]$. Then there are six natural mappings from P to $[0,1]$, given by:

$$\begin{aligned} t &: [x,y] \to x \\ nil &: [x,y] \to 1-y \\ unc &: [x,y] \to y-x \\ mass &: [x,y] \to 1-(y-x) \\ poss &: [x,y] \to y \\ poss\text{-}not &: [x,y] \to 1-x \end{aligned}$$

Intuitively, *t* corresponds to the extent to which a given statement is confirmed by the available evidence, and *nil* to the extent to which it is disconfirmed. *mass* reflects the completeness of our probabilistic information, and *unc* the incompleteness of it. Finally, *poss* and *poss-not* correspond to the degrees to which the statement *might be* true or false respectively.

We will refer to these six functions as *tags*; they provide a natural and uniform framework in which to consider either the truth or falsity of any given proposition, or the extent of our knowledge about it.

5.2 Use of probability to limit inference

Non-numerical inference techniques must of necessity run to completion; there seems to be no way to use qualitative information to terminate the inference process. This can be avoided if quantitative methods are used.

There are two ways in which a probabilistic inference can be shortened. Suppose that we are trying to prove some proposition p; the first cutoff can be implemented by not including in our analysis any inferences which will affect the eventual probability of p by less than some small value c_1. For example, it never rains in southern California (or at least only very rarely) [Hammond 1972]; if we are trying to show that our beach party will be a success, we do not need to consider rain as a reason for it not to be.

A second and independent way to shorten a probabilistic inference is to assume that if the probability exceeds some value c_2 (alternatively, if the result of applying some tag to the probability *range* exceeds c_2), the inference is assumed complete. If the All-Star game is being played in Los Angeles on the same day as our beach party and we have a friend who is giving away tickets to it, then we are probably better off picking another day for the party than looking for an esoteric proof that it will be successful after all.

It is worth considering the effects on the inference procedure if we select extremal values for c_1 or c_2. Taking $c_1 = 0$ allows allows all relevant information to be considered, while $c_2 = 1$ ensures that the entire deduction will not be stopped early. This combination therefore results in all attempted derivations running to their eventual conclusions as described at the beginning of this section. (And as such, is no more efficient than any of the more conventional techniques for non-monotonic reasoning.) If we select $c_1 = 1$, then only monotonic inferences will be considered, while $c_2 = 0$ results in the rather preemptive strategy of considering only the first bit of applicalbe information. Finally, the combination $c_1 = c_2 = 1$ allows us to perform standard monotonic reasoning using a probabilistic database.

5.3 Probabilistic resolution

The inference technique of resolution can be extended to deal with probabilistic information. Consider the derivation of *flies*(Tweety) from $bird(x) \rightarrow flies(x)$ and *bird*(Tweety):

$$\neg bird(x) \vee flies(x)$$

$$bird(\text{Tweety})$$

Unifying the above two expressions by substituting Tweety for x and resolving the results, we obtain *flies*(Tweety).

In general, from $p \vee r$ and $q \vee \neg r$ we can derive $p \vee q$. The implication

$$(p \vee r) \wedge (q \vee \neg r) \rightarrow (p \vee q) \qquad (7)$$

is of course tautologous; it follows that the probability of $p \vee q$ is at least that of $(p \vee r) \wedge (q \vee \neg r)$.

The situation is complicated somewhat in the non-monotonic case by the need to treat negation in a uniform fashion. For example, the probability range corresponding to some proposition p contains information not only about p itself, but about its negation $\neg p$ as well. If the truth value of the original proposition is $[x, y]$, it translates into a *pair* of disjuncts corresponding to

$$p \qquad [x, 1]$$

and

$$\neg p \qquad [1 - y, 1]$$

respectively. Since it will be possible to resolve these two expressions, we should assume that the conjuncts in (7) are not independent. Denoting by p not only a proposition but also the probability range attached to it, we should therefore evaluate (7) using

$$(p \vee r) \wedge (q \vee \neg r) = \max\big(0, (p \vee r) + (q \vee \neg r) - 1\big).$$

It is also important that we be able to resolve $(p \vee r)$ not only with $(q \vee \neg r)$, but with $(q \vee r)$ as well. This process is controlled by the tautologies

$$\begin{gathered} \neg(p \vee r) \wedge \neg(q \vee r) \rightarrow \neg(p \vee q) \qquad \text{and} \\ [(p \vee r) \wedge \neg(q \vee r)] \vee [\neg(p \vee r) \wedge (q \vee r)] \rightarrow (p \vee q). \end{gathered} \tag{8}$$

Here, the facts being resolved will generally be independent.

All of these inferences can be described easily in terms of the tags introduced in section 4.1. We have:

$$\begin{aligned} t(p \vee q) &= \max\big(0, t(p \vee r) + t(q \vee \neg r) - 1\big) \\ nil(p \vee q) &= 0 \end{aligned} \tag{9}$$

and

$$\begin{aligned} t(p \vee q) &= t(p \vee r)\,nil(q \vee r) + nil(p \vee r)\,t(q \vee r) \\ nil(p \vee q) &= nil(p \vee r)\,nil(q \vee r). \end{aligned} \tag{10}$$

(9) corresponds to the usual sort of resolution, while (10) is a consequence of the tautologies (8) and corresponds to resolving $p \vee r$ with $q \vee r$ as opposed to $q \vee \neg r$. In either case, specification of t and *nil* is enough to uniquely determine the new contribution to the probability range assigned to $p \vee q$.

5.4 Implementation results

The ideas described in this paper have been implemented in the expert system-building tool MRS at Stanford. We will conclude by describing some of the details of this implementation. Additional details can be found in [Ginsberg 1984b].

MRS [Genesereth 1982] is a logic-based expert system-building tool. It supports a variety of inference methods, including forward- and backward-chaining. Information is currently entered into MRS on two "levels". The meta-level is used to store information regarding control of inference or procedural attachments for the various MRS primitives (a demon is a procedural attachment to the primitive that stashes an item in the database, for example). The base level is used to store more conventional expert system-type information about the domain in question. Although the inference methods for the two levels are distinct, all of the information is stored in a single database.

The probabilistic implementation associates to each fact in the database a pair $(c\ .\ d)$ corresponding to the probability range $[c, 1-d]$. The probability ranges are thought of as the "truth values" of the propositions, and are combined using Dempster's rule as described in [Ginsberg 1984a].

Tags are used to reduce the probability ranges to specific values, as described in section 4.1. This has the immediate advantage of unifying the treatment of negation within MRS itself—where the two propositions `(not (ostrich fred))` and `(ostrich fred)` had previously been considered to be unrelated, they are now simply differing apects of the same object, and interact more conveniently with, for example, `(known (ostrich fred))` or `(unknown (ostrich fred))`.

Reason maintanence facilities have been implemented in the forward chainer only.

When a rule of inference is invoked, the truth value of the instantiated version of the premise is stored, along with information concerning the instantiation itself. The next time the rule is invoked, if the mass of the difference between the previous truth value and the current one is no greater than the inference cutoff c_1, no action is taken. The effect of this is to avoid propagating a change in the database to a point where it will have no significant effect on the probabilities of the statements involved.

The backward chainer has been implemented using the pair of cutoffs described in the previous section. Timing tests done with $c_1 = c_2 = 1$ (standard monotonic inference only) indicate that the incorporation of the probabilistic facilities has at most a small effect (perhaps 5%) on the system's monotonic performance.

The most important experiment has not yet been done. What is needed are comparable implementations of a large-scale non-monotonic problem using both probabilistic and non-probabilistic methods. It is only when such comparisons can be made that it will be possible to draw secure conclusions.

§6. Conclusion

The efficacy of using probabilities in a non-monotonic inference system is both a theoretical and an experimental question, and we have attempted to address both issues in this paper. Our theoretical arguments dealt with the possibility of using probability ranges and Dempster-Shafer theory to sidestep the Bayesian need for *a priori* probabilities.

The experimental question may well be more interesting, but cannot be settled until a great deal more work is done on full-size non-monotonic systems that do and do not use probabilistic inference methods. The work we have completed at Stanford seems to support the arguments we have presented, but no hard and fast conclusion can be drawn without a great deal more experimental evidence.

Acknowledgement

This research has been supported by the Office of Naval Research under grant number N00014-81-K-0004. The author would like to thank Mike Genesereth and John McCarthy

for many illuminating (if occasionally heated) discussions, and Russ Greiner for the care with which he examined an earlier version of this paper.

References

[Dempster 1968] Dempster, A.P., "A generalization of Bayesian inference," J. Roy. Stat. Soc. *B* **30** (1968) 205-247

[Doyle 1979] Doyle, J., "A truth maintenance system," Artificial Intelligence **12** (1979) 231–272

[Genesereth 1982] Genesereth, M.R., "An overview of MRS for AI experts," HPP working paper 82-27 (1982)

[Ginsberg 1984a] Ginsberg, M.L., "Non-monotonic reasoning using Dempster's rule," AAAI-84, Austin, Texas, 126–129

[Ginsberg 1984b] Ginsberg, M.L., "Implementing probabilistic reasoning," HPP working paper 84-31 (1984)

[Hammond 1972] Hammond, A., "It never rains in southern California"

[McCarthy 1984] McCarthy, J., "Applications of circumscription to formalizing common sense knowledge," *Non-monotonic Reasoning Workshop*, American Association for Artificial Intelligence (1984) 295-324

[Moore 1981] Moore, R.C., "A formal theory of knowledge and action," to appear in *Formal Theories of the Commonsense World*, Hobbs, J.R., and Moore, R.C. (eds.), Ablex Publishing Co. (1984)

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

[von Mises 1964] von Mises, R., *Mathematical Theory of Probability and Statistics*, Academic Press, New York (1964)