JEFFREY'S RULE AND THE PROBLEM OF AUTONOMOUS INFERENCE AGENTS

Judea Pearl

June 1986
CSD-860099

TECHNICAL REPORT
R-62
CSD-860099
June 1986

JEFFREY'S RULE AND THE PROBLEM OF AUTONOMOUS INFERENCE AGENTS*

Judea Pearl
Cognitive Systems Laboratory
UCLA Computer Science Department
Los Angeles, California 90024
(judea@LOCUS.UCLA.EDU)

ABSTRACT

Jeffrey's rule of belief revision was devised by philosophers to replace Bayes conditioning in cases where the evidence cannot be articulated propositionally. This paper shows that unqualified application of this rule often leads to paradoxical conclusions, and that to determine whether or not the rule is valid in any specific case, one must first have topological knowledge about one's belief structure. However, if such topological knowledge is, indeed, available, belief updating can be done by traditional Bayes conditioning; thus, arises the question of whether it is *ever* necessary to use Jeffrey's rule in formalizing belief revision.

*This work was supported in part by the National Science Foundation, Grants #DCR 83-13875 & 85-01234.

JEFFREY'S RULE AND THE PROBLEM OF AUTONOMOUS INFERENCE AGENTS

Judea Pearl

The three-prisoners puzzle shows that, when a new fact is added to our knowledge base, its implications depend critically on the process by which the fact was learned and, in particular, on the collection of all other facts that could have possibly been gathered in that process. Such detailed knowledge may not always be available; we are often required to respond to new information without having the slightest idea of how it was collected. Situations of this kind occur when the task of gathering information is delegated to autonomous agents, each using its own covert procedures which, for various reasons, cannot be explicated in full detail.

Richard Jeffrey (*The Logic of Decisions,* McGraw-Hill, 1965, Chapter 11) was the first to recognize the importance of this problem, and he devised a rule for handling it. The autonomous agents used in Jeffrey's original example are our sensory organs, as described in the following quotation from Jeffrey's book:

Observation by Candlelight

> "The agent inspects a piece of cloth by candlelight and gets the impression that it is green, although he concedes that it might be blue or, even (but very improbably), violet. If G, B and V are the propositions that the cloth is green, blue and violet, respectively, then the outcome of the observation might be that, whereas originally his degrees of belief in G, B and V were .30, .30 and .40, his degrees of belief in those same propositions after the observation are .70, .25 and .05. If there were a proposition E in his preference ranking (i.e., knowledge framework) which described the precise quality of his visual experience in looking at the cloth, one would say that what the agent learned from the observation was that E is true. If his original subjective probability assignment was *prob*, his new assignment should then be $prob_E$, and we would have
>
> $$prob\ G = .30 \quad prob\ B = .30 \quad prob\ V = .40$$
>
> representing his opinions about the color of the cloth before the observation, but would have
>
> $$prob(G|E) = .70 \quad prob(B|E) = .25 \quad prob(V|E) = .05$$
>
> representing his opinions about the color of the cloth after the observation"...."When the agent looks at the piece of cloth by candlelight there is a particular complex pattern of physical stimulation of his retina, on the basis of which his beliefs about the possible colors of the cloth change in the indicated ways. However, the pattern of stimulation need not be describable in the language he speaks; and even if it is, there is every reason to suppose that the agent is quite unaware of what that pattern is, and is quite incapable of uttering or identifying a correct description of it. Thus, a complete description of the

> pattern of stimulation includes a record of the firing times of all the rods and cones in the outer layer of retinal neurons during the period of the observation. Even if the agent is an expert physiologist, he will be unable to produce or recognize a correct record of this sort on the basis of his experience during the observation.''

With this story in mind, Jeffrey then poses the question of how the new information should be used to influence other propositions which depend on the color of the cloth. In his words: ''Then the problem is this: Given that a passage of experience has led the agent to change his degrees of belief in certain propositions $B_1, B_2, .., B_n$ from their original values,

$$prob\ B_1, prob\ B_2, ..., prob\ B_n$$

to new values,

$$PROB\ B_1, PROB\ B_2, ..., PROB\ B_n,$$

how should these changes be propagated over the rest of the structure of his beliefs? If the original probability measure was *prob*, and the new one is *PROB*, and if A is a proposition in the agent's preference ranking but is not one of the n propositions whose probabilities were directly affected by the passage of experience, how shall $PROB\ A$ be determined?''

Jeffrey's solution is based on the critical assumption that ''while the observation changed the agent's degree of belief in B and in certain other propositions, it did not change the *conditional degree of belief* in any propositions on the evidence B or on the evidence $\bar{B}$.'' Thus, if $B_1, B_2 \cdots B_n$ are a set of exhaustive and mutually exclusive propositions (like the colors 'green,' 'blue' and 'violet' in the candlelight example), Jeffrey maintains that for every proposition A we should write:

$$PROB\ (A \mid B_i) = prob\ (A \mid B_i) \quad i = 1,2,...,n \tag{1}$$

This, together with the additivity of *PROB*, leads directly to

$$PROB\,(A) = \sum_i prob\ (A \mid B_i)\, PROB\,(B_i), \tag{2}$$

a formula that became known as *Jeffrey's Rule* of updating probabilities. The convenience of the rule is enticing; we need not know a thing about the process by which the updating from $prob\,(B_i)$ to $PROB\,(B_i)$ took place -- only the net result matters. We simply take $PROB\,(B_i)$ as a new set of priors, and we apply to them the textbook formula (2). A strict probabilistic analysis will, of course, question the universal validity of (1); for, if we denote by e the evidence actually observed and identify $PROB\,(A)$ with $prob\,(A \mid e)$, we then get

$$prob\,(A \mid e) = \sum_i prob\,(A \mid B_i, e)\, prob\,(B_i \mid e) \tag{3}$$

which coincides with (2) only when A and e are conditionally independent, given B_i.

To demonstrate the rationale behind Jeffrey's Rule and some of its weaknesses, let us return to the candlelight example and examine three cases where different meanings are assigned to A.

Case (a)

Assume that the proposition A in (2) stands for the statement "The cloth will be sold the next day," and that we know that the chances of selling the cloth depend solely on its color, via:

$$P(A \mid green) = .40 \quad P(A \mid blue) = .40 \quad P(A \mid violet) = .80$$

Eq. (2), then, permits us to calculate the updated belief in the salability of the cloth, based only on the conclusion of the test process. Whereas, prior to the test, our belief in selling the cloth measured

$$P(A) = (.4)(.3) + (.4)(.3) + (.8)(.4) = .56\,,$$

once the test results become known, our belief should change to:

$$P(A \mid e) = (.4)(.7) + (.4)(.25) + (.8)(.05) = .412\,.$$

This reasoning would pass the scrutiny of even the strictest Bayesian because stating that the salability of the cloth depends only on its color amounts to asserting the conditional independence of A and e in (3)

$$P(A \mid color, e) = P(A \mid color),$$

which legitimizes Jeffrey's assumption

$$PROB\ (A \mid B_i) = prob\ (A \mid B_i)$$

and, hence, his updating rule.

To demonstrate the volatility of this assumption, let us examine an extreme example where it is obviously violated.

Case (b)

Imagine that the main interest of our candlelight observer lies not in the color of the cloth but rather in the chemical composition of the candle's wax. Let A be the proposition that the candle's wax belongs to one notoriously cheap brand, known to produce flames deficient in violet content. Under these circumstances, are we justified in using Jeffrey's Rule? Now the situation is completely reversed; the actual colors of the cloth (B_i) are of no relevance to A prior to the observation; so, $prob(A \mid B_i) = prob\ A$. If we blindly apply Jeffrey's Rule (3) to this situation, we obtain a paradoxical result:

$$PROB\ (A) = \sum_i prob(A)\ PROB\ (B_i) = prob\ (A)\ . \tag{4}$$

No matter how violet or greenish the cloth looks under the candlelight, the observer's belief regarding the spectral content of the flame ought to remain unaltered.

The lesson here is that, even though we lack the knowledge required for precise description of the measurement process, our common-sense understanding of the process is sufficient to alert us to the falsehood of $P(A \mid B_i, e) = P(A \mid B_i)$ and thus protect us from drawing a wrong conclusion as in (4). Qualitatively speaking, we normally summarize the difference between the two situations above by saying that in *case (a),* the color of the cloth "stood between" the evidence and the proposition A (the salability of the cloth), while in *case (b),* it was the evidence which mediated between the colors and proposition A (the brand of wax), as in Figure 1. Before giving these notions precise definitions and formal graphical representations, let us consider a third case (corresponding to Fig. 1 (c)) where, again, Jeffrey's Rule leads to false conclusions.

Case (c)

Imagine that the candlelight observer is color-blind but can judge the colors (B_i) by closely examining the *texture* of the cloth, knowing that all green cloths turn out *coarse* in texture, while 64.2857% of all blue cloths and 94.64% of all violet cloths are *fine* textured. Initially, the agent's opinions about the color of the cloth, $prob(B_i) = (.30, .30, .70)$, were based purely on frequency information. From this information the agent may infer that the proposition A: "the cloth is coarse" deserves a belief

$$prob\ (A) = \sum_i prob(A \mid B_i)\ prob(B_i) = (1)(.30) + (.35714)(.30) + (.0536)(.40) = .4286$$

After examining the texture of the cloth by candlelight, the agent becomes absolutely sure of its coarseness (i.e., $PROB\ (A) = 1$), from which, using some covert mental process, he infers and reports $PROB\ (B_i) = (.70, .25, .05)$. The question is, can we recover the concealed value of $PROB\ (A)$ from the reported values of $PROB\ (B_i)$, using Jeffrey's Rule?

Applying the rule to this situation, gives:

$$PROB(A) = \sum prob(A \mid B_i)\ PROB(B_i) \tag{5}$$

$$= (1)(.70) + (.35714)(.25) + (.0536)(.05) = .7919$$

The correct analysis should yield $PROB(A) = 1$, not .7919 as in (5). In general, the correct updating formula for this case is more complicated than Eq. (5), but it can be obtained using Bayes Rule if only we are given only the parameters $prob(A \mid B_i)$, $prob(B_i)$

and $PROB\,(B_i)$.

We can use this example to demonstrate another difficulty associated with Jeffrey's Rule -- Why can't we apply the rule again on the updated probability $PROB\,(A)$ to obtain a doubly-revised value for $PROB\,(B_i)$.

Imagine that we have two agents: one color-blind who conducts coarseness tests and one color-sensitive who conducts color tests. The first agent examines the cloth and reveals to us $PROB_1(A)$. Agent 2 combines this revelation with the available frequency information $prob\,(A \mid B_i)$ and forms an opinion $prob_2(B_i)$. He then conducts a color test and revises his opinion $prob_2(B_i)$ to $PROB_2(B_i)$. How can we use this new information to update the belief in A ,the coarseness of the cloth? Brute-force application of Jeffrey's Rule:

$$PROB_2(A) = \sum_i prob\,(A \mid B_i)\, PROB_2(B_i) \tag{6}$$

would neglect the fact that part of the belief represented by $PROB_2(B_i)$ actually originated from A and should not be counted twice. A more careful approach will take this into account by also updating the conditional probability in (6) from its original value of $prob\,(A \mid B_i)$ to

$$PROB_1(A \mid B_i) = P\,(A \mid B_i, e) = prob\,(A \mid B_i)\,\frac{PROB_1(A)}{prob\,(A)}\,\alpha$$

where α is a normalizing constant. However, such a procedure would mean that each time new evidence arrives we need to update the conditional probabilities for all pairs of propositions that may be of interest in the future. It is a difficult computational task and is certainly not representative of the way humans deal with multiple evidence. It is possible to show* that Jeffrey's Rule can be modified to handle multiple updatings without changing the conditional probabilities $prob\,(A \mid B_i)$. The modification required is:

$$PROB_2(A) = \sum_i prob\,\,(A \mid B_i)\left[\frac{PROB_2(B_i)}{prob\,\,(A \mid B_i) + \lambda(1 - prob\,\,(A \mid B_i))}\right] \tag{7}$$

where λ is the solution to:

$$PROB_1\,(A) = \sum_i prob\,(A \mid B_i)\left[\frac{PROB_1\,(B_i)}{prob\,(A \mid B_i) + \lambda(1 - prob\,\,(A \mid B_i)}\right]. \tag{8}$$

Thus, in our previous example we had $PROB_1\,(A) = 1$, which yields $\lambda = 0$ in (8) and, when substituted in (7), maintains $PROB_2(A) = 1$, regardless of $PROB_2(B_i)$. This result supports our intuition that, once the coarseness is established by direct methods, it ought not to be challenged by indirect information, such as color examinations.

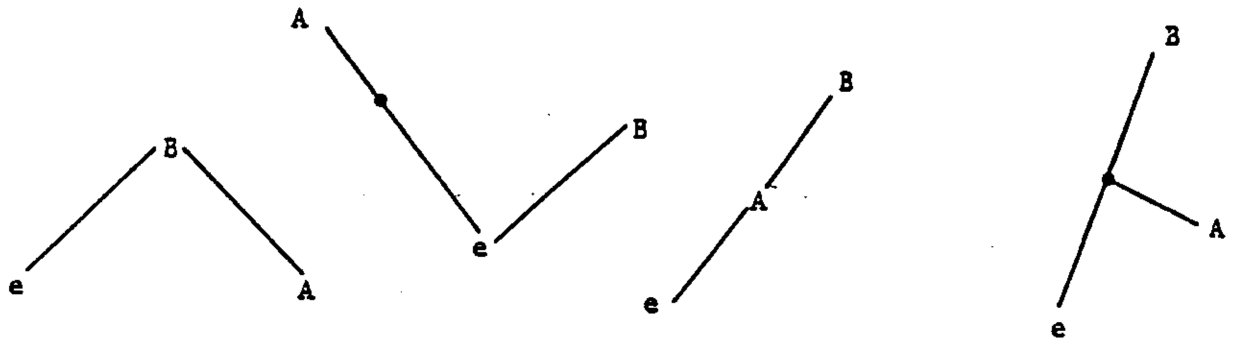
* Based on distributed schemes of belief-propagation (Pearl, Proc. AAAI Conf., pp. 133-136, 1982).

The example of *case (c)* carries two messages: First, we demonstrated again that the semantic elements of the story are sufficient for judging whether Jeffrey's assumption, the conditional-independence $P(A \mid B_i, e) = P(A \mid B_i)$, is reasonable or not, even when we have no idea how to give a precise description for the observed evidence e. In *case c*, the conditional-independence assumption is obviously false because it implies that, no matter how skilled our agent or how bright the candlelight is, there is no way to increase his confidence in the coarseness of a green cloth beyond the initial degree of belief, *prob*$(coarse \mid green)$. The correct independence relation in this case is $P(B_i \mid A, e) = P(B_i \mid A)$, as illustrated in Figure 1(c). But Jeffrey's independence assumption is invalid not only when A lies directly on the inference path from e to B; it is enough that A branches off someplace in the middle of that path, as in Figure 1(d). In the case of the color-blind observer, for example, if A stands for the proposition that the cloth will sell tomorrow, and if customers buy cloths on the basis of their texture (see Fig. 1(d)), then Jeffrey's rule will again lead to contradictions.

The second point of this example is to show that even in cases where a full description of e is impractical, explicating just a small part of the agent inference process may yield the desired result. Although the process by which our color-blind agent updates his belief in the coarseness of the cloth will forever remain a mystery, knowing only how the agent perceives the connection between coarseness and color was sufficient for recovering his belief in the former from the reported belief in the latter, using Bayesian conditioning.

A general pattern emerges from the graphic representations of the three examples (Figure 1); whenever the inference path from the evidence e to the updated proposition B shares a segment with the path from B to the proposition we wish to update, A, Jeffrey's rule yields erroneous (i.e., paradoxical) results. When these two paths are disjoint, sharing no other node besides B, Jeffrey's rule is applicable and is identical to Bayes conditioning. In simple terms, Jeffrey's Rule makes sense *only* when B *separates* A from e and, in all fairness to Jeffrey, perhaps this is what he meant by saying that A is ''not one of the propositions whose probabilities were directly affected by the passage of experience.''

The fact that simple criteria based on graph topology lead to conclusions that match our intuition (about the soundness of Jeffrey's rule) suggests that human intuition itself can be represented by a graph of relations and that intuitive judgements themselves involve mental tracing of those graphs. These suggestions motivate the study of another topic, dependency-graphs.



case (a) case (b) case (c) case (d)

Figure 1