CAN INFORMATION THEORY HELP US REMOVE OUTDATED BELIEFS?

Erik J. Olsson

Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 Lund, Sweden

E-mail Erik.Olsson@filosofi.uu.se

Abstract: It is argued that a rational agent should not be able to gain information simply by removing a previously accepted proposition, i.e. a proposition having probability 1. An example is given showing that the Gärdenfors axioms for probabilistic contraction in fact do not exclude that the agent's uncertainty is decreased as a result of contraction. The concept of entropy is borrowed from statistical information theory and used as a tool for measuring the uncertainty pertaining to a belief state. Entropy is then used to formulate a new axiom in addition to the Gärdenfors axioms. The new axiom explicitly rules out the possibility of agents getting more informed as a result of belief removal.

INTRODUCTION

Imagine that a crime has been commited and that you cannot exclude that Herman is the criminal. You read in the morning paper that in fact Herman has confessed, information that leads you to believe that it was Herman who commited the crime. Assume that the next day you learn, to your great surprise, that new evidence has been presented in the Herman case. It turns out that Herman never confessed after all; that he had confessed was only a rumour. The only evidence pointing at Herman is that his blood type matches that of the criminal.

When discussing belief changes like those in the above example, it is common to refer to a distinction between two types of basic belief change: *expansion*, i.e. the addition of a new belief consistent with the agent's prior beliefs; and *contraction*, i.e. the removal of some previous belief[1]. To return to the example above, when you read that Herman has confessed you perform an expansion. When you later read that Herman didn't confess after all, this new information calls, among other things, for a contraction of the belief that Herman is the criminal. It is often argued that *revision*, by which is usually meant the addition of a belief inconsistent with prior beliefs, can be explained as a sequence of contractions and expansions. To add the belief -A assuming that you already believe A must mean that you first have to make place for -A by removing A after which you can consistently add -A.

There are, hence, two main uses of contraction: as a stand-alone belief change operation and as a tool for making rational revisions. While standard conditionalization can be used to represent expansion in a probabilistic context, there is, unfortunately, no similary evident way to represent contraction[2]. Peter Gärdenfors has presented a collection of axioms intended to characterize probabilistic contraction[3]. Gärdenfors' axioms are, however, compatible with many quite different contraction methods. It would

[1]The literature on non-probabilistic belief revision seems to grow almost exponentially. For the basic concepts the reader should consult Gärdenfors, 1988. Philosophical and interpretational problems are discussed in Levi, 1991. In this context, see also Harman, 1986. For the state-of-the-art, see Fuhrmann & Morreau, 1989 and Gärdenfors, 1992.

[2]Probabilistic belief revision has not received the same attention as non-probabilistic revision. See for example Diaconis & Zabell (1982), Gärdenfors (1988), Lindström & Rabinowicz (1989), May (1976), May & Harper (1976), van Fraassen (1980) and Williams (1980).

[3]In Gärdenfors 1988.

Lund University Cognitive Studies – LUCS 36 1995. ISSN 1101–8453.

be desirable to further restrict the possible choices of a posterior state after contraction. The main purpose of this paper is to investigate the possibility of stronger axioms for probabilistic contraction.

A *belief state* will be represented as a probability distribution over a space of possibilities (or alternatives or atomic events or possible worlds/situations) W.[4] We denote the set of all possible probability distribution by $\prod$. A *proposition* is a subset of W. For each probability distribution P and each proposition A, the probability which is assigned to A by P is defined:

$$P(A) = \sum_{w \in A} P(w)$$

Note that we allow for the possibility that W is infinite. In all examples, however, W will be finite. We use -A to denote W-A, i.e., the set-theoretical difference between W and A.

2 THE UNIQUENESS PROBLEM FOR EXPANSION AND CONTRACTION

The problems of contraction is closely connected with expansion which motivates that we first consider expansion. Assume that P(A)>0 and that the effect of new information is to give conclusive support for A. This situation can be represented as a constraint on the new state of information. The constraint is the set of all probability distributions Q such that Q(A)=1, and the task is to choose one of the belief states in this set. As we will see there are strong reasons to believe that this case can be handled by standard conditionalisation in the sense that

$$P_A^+(B) = \frac{P(A \cap B)}{P(A)}, \text{ for all propositions B.}$$

That is, the new probability distributions after adding the information A is the old distribution conditionalised on A. The motivation is that conditionalisation is a minimal change of belief in the following sense: "unlike all other revisions of P to make A certain, it does not distort the profile of probability ratios, equalities, and inequalities among sentences that imply A" (Lewis 1976, p. 311). This is the case since conditionalisation works by assigning all propositions inconsistent with the new information a zero probability value, while all propositions consistent with the new information are scaled by the same constant, i.e. P(A).

[4]For criticism of this "Bayesian" way to represent belief states see Gärdenfors & Sahlin (1988), Shafer (1976) and Shafer (1981).

Example.2.1 Let $W=\{w_1, w_2, w_3\}$ and $P(w_1)=1/3$, $P(w_2)=1/6$ and $P(w_3)=3/6$. A compact way to represent P is as the vector P=(1/3, 1/6, 3/6). Let A be the proposition $\{w_1, w_2\}$. Then the result of incorporating A is the vector (2/3, 1/3, 0).

The proposal is, then, that we can use conditionalisation to single out a unique new belief state from the imposed constraint. This solves the uniqueness problem for expansion. Unfortunately, the uniqueness problem for contraction is much more difficult. To see where the difficulties lie, let us look at a concrete example.

Example.2.2 Suppose $W=\{w_1, w_2, w_3, w_4, w_5\}$, P=(1/4, 3/4, 0, 0, 0) and $A=\{w_1, w_2\}$. Now, assume that we receive new information to the effect that A is falsified. The constraint on the posterior belief state is $\Omega=\{Q: Q(A)<1\}$. From this set we should, ideally, be able to choose a unique distribution as the posterior belief state.

To solve the uniqueness problem for contraction, if possible, the following questions must be given definite answers: What probability should be assigned to $P^-_A(A)$? That is, what probability should A have in the posterior belief state? How should we distribute probability values over possibilities in A? In the posterior state, -A will have a positive probability, i.e. $P^-_A(-A) = \alpha > 0$. How should the probability value α be distributed over possibilities in -A? Specifically, what possibilities in -A should be assigned 0?

The rest of this paper will be devoted to the uniqueness problem for contraction.

3. GÄRDENFORS' AXIOMS FOR PROBABILISTIC CONTRACTION

The Gärdenfors axioms for contraction are given next[5]. Originally, these axioms were formulated in terms of probability distributions over sentences in a propositional language rather than over propositions formed from an outcome space. It turns out that the problems involved in contraction stand out much clearer if we use the latter "possible world" terminology. As it happens, one of the original six Gärdenfors axioms, saying that contractions using equivalent sentences should give the same result, becomes superfluous. The remaining five axioms are:

(G-1) For all probability distributions P and all propositions A, P^-_A is a probability distribution.

[5]In Gärdenfors, 1988, p. 118.-

(G-2) $P^-_A(A)<1$ iff $A \neq W$.

(G-3) If $P(A)<1$, then $P^-_A=P$.

(G-4) If $P(A)=1$, then $(P^-_A)^+_A=P$.

(G-5) If $P^-_{A\cap B}(-A)>0$, then $P^-_A(C/-A) = P^-_{A\cap B}(C/-A)$ for all C.

Axioms (G-1)-(G-3) are uncontroversial regularity conditions. The real force lies in (G-4) and, to some extent, in (G-5). Axiom (G-4) states that if we first contract with respect to -A and then conditionalize on A, we should get P back. According to (G-5), if we contract with respect to $A\cap B$ and A is given up in that process, then this contraction and the result of contraction with A simpliciter should give the same proportions of probabilites to the sentences implying -A. This is a parallel to what happens when we conditionalize. Recall that when we conditionalize with respect to A, the proportions between sentences implying A are preserved. The proposal is, then, that contraction should be seen as "backward" conditionalisation.

Example.3.1 Let us continue example 3.2. As before $W=\{w_1, w_2, w_3, w_4, w_5\}$, $P=(1/4, 3/4, 0, 0, 0)$ and $A=\{w_1, w_2, w_3\}$. It is compatible with (G-4) that the result of contracting P with respect to A is $Q=(1/8, 3/8, 0, 0, 4/8)$. For if we conditionalize Q with respect to A we get P back. However, $(1/8, 3/8, 0, 4/8, 0)$ and $(1/16, 3/16, 0, 8/16, 4/16)$ are equally permissable. Now, let $B=\{w_1, w_2, w_4\}$. A and B are both accepted in P. To illustrate (G-5) we note that a contraction rule to the effect that $P^-_A=(1/8, 3/8, 0, 1/8, 3/8)$ but $P^-_{A\cap B}=(1/8, 3/8, 0, 3/8, 1/8)$ would violate (G-5) (but not (G-4)). A possible rule could instead let $P^-_{A\cap B}$ be equal to $(1/16, 3/16, 0, 3/16, 9/16)$.

Gärdenfors proves the following useful consequences of (G-1)-(G-4):[6]

Theorem 3.1. For all B, if $P^-_A(B) = 1$, then $P(B) = 1$.

Theorem 3.2. $P^-_A(B) = \alpha P + (1-\alpha)(P^-_A)^+_{-A}(B)$ for all B (where $\alpha = P^-_A(A)$ when $P(A) = 1$ and $\alpha = 1$ otherwise).

Theorem 3.1 says that the set of propositions accepted in P^-_A is a subset of the set of propositions accepted in P. According to Theorem 3.2, the result of a contraction is always a probabilistic mixture of two distributions, one of them being P and the other $(P^-_A)^+_{-A}$. Note in connection with Theorem 3.2 that the Gärdenfors axioms do not say anything about the magnitude of $P^-_A(A)$, except that it should be strictly less than 1. As Gärdenfors remarks, this leaves open a large number of possibilities for an explicit construction of a contraction function. If $P^-_A(A) = \alpha$, then P^-_A is called an α-contraction of P with respect to A. Concerning the distribution of probabilities over possibilities in A, only axiom (G-4) is relevant. As indicated earlier, (G-4) says that the relative proportions between these probabilities should be preserved. Regarding the distribution of probabilities over possibilites in -A, we have only (G-5), according to which the relative proportions between these probabilites should be maintained along the line indicated above. Note, however, that the axioms are indifferent to what possibilities in -A should be assigned probability zero.

We have seen that the Gärdenfors axioms are rather weak. It would be quite surprising if nothing more could be said about contraction than what is expressed in these axioms. I now turn to the task of finding stronger axioms to complete Gärdenfors' account.

4. Contraction and loss of information

What intuitions can we rely on in our search for stronger contraction axioms? Whenever we decide to perform a contraction, it is intuitively clear that, given a reasonable contraction function, we should lose information. That contraction should involve loss of information is an implicit assumption in Gärdenfors' Knowledge in Flux: "Because information is valuable, it is rational to minimize the loss of information when giving up sentences in a contraction of a state of belief".[7] Falsifying a proposition that we previously believed cannot make us more opinionated. Intuitively, it should instead make us more uncertain. One way to express this idea would be to say that fewer propositions should be accepted after the contraction than before. As we have seen this is in fact provable from the Gärdenfors axioms (Theorem 3.1). But it is not clear that the number of accepted propositions indeed is a plausible information measure. We would, hence, like to have a way to measure the uncertainty of a probability distribution. Assuming that we could do this we would like to postulate (or, if possible, prove from the Gärdenfors axioms) that contraction always involves a loss of information.

However, the Gärdenfors' axioms do not assure that the agent's uncertainty increases as a result of contraction. This is shown by the following intuitive example.

Example 4.1. Let $W=\{w_1, w_2, w_3\}$, $A=\{w_1, w_2\}$, $P=(1/2, 1/2, 0)$ and $Q=(1/16, 1/16, 14/16)$. Then Q is an admissable contraction of A from P according to the Gärdenfors axioms. Note especially that we get P

[6] Appendix C in Gärdenfors 1988.

[7] Gärdenfors 1988, p 91.

back if we conditionalize on A with respect to Q. The problem is that Q is intuitively more opinionated than P. For in Q almost all of the probability mass is concentrated on one alternative, while P devides the probability equally between two alternatives. All reasonable ways to measure uncertainty should make Q less uncertain (i.e. more informed) than P, which means that the example does not rely on any particular choice of uncertainty measure.

Since the Gärdenfors axioms, contrary to intuition, do not garantee that the agent loses information in the contraction process, we would like to add an axiom to this effect. One way to do this would be to state postulates that characterize the class of all reasonable ways to measure uncertainty, and then to say that every measure in this class should make the posterior state after contraction more uncertain than the prior state. However, we will be satisfied here with the less ambitious problem of formulating an axiom in terms of a specific, and commonly used, information measure: entropy. This measure has some advantages over many other measures, one of them being that it easily generalizes to the infinite case.

5. CONTRACTION AND INFORMATION-THEORETICAL ENTROPY

The purpose of this section is to define and motivate entropy as a measure of the uncertainty pertaining to a probability distribution[8]. The most convenient way to do this is to start by defining the information in an event $A \neq \emptyset$. We want to measure the information which represents for the agent. A fundamental principle of information theory is that the quantity of information associated with A equals the reduction of uncertainty that would be the result if the agent got to know A[9].

Suppose for example that we wanted to know were on a chess table the white king is positioned. There are 64 possible alternatives which are supposed to be equally probable. We could use the number 64 to quantify the uncertainty of the initial state. Instead we choose the number log 64 = 6 bits (if we use 2 as the logarithm base). Now we are told that the white king is actually on a black square. We are now left with 32 possible outcomes, that is, the uncertainty has been reduced from 6 bits to log 32 = 5 bits. This means that the amount of information associated with the event that the white king is on a black square is 1 bit:

$$\log 64 - \log 32 = \log \frac{64}{32} = \log 2 = 1.$$

Given that all atomic events are equally probable, the information in an event A is given by:

$$\inf(A) = \log \frac{n}{\text{card}(A)} = \log n - \log \text{card}(A),$$

where n is the number of elements in W. Note that if all atomic events are equally probable the probability of A is given by P(A)=card(A)/n. Combining this fact with the above equation has the following consequence:

$$\inf(A) = \log \frac{1}{P(A)} = \log 1 - \log P(A) = -\log P(A).$$

It is natural to extend this idea to the general case in which the atomic events are not necessarily equiprobable. This completes the motivation for the following definition:

Def 5.1 $\quad \inf_P(A) = -\log P(A).$

We will drop the subscript and only write inf(A) when the associated probability distribution is determined by the context.

We now go on to defining the information, or rather lack of information (i.e. uncertainty), in a probability distribution. Obvious requirements are that the uncertainty should be at a maximum for the uniform distribution and at a minimum for a distribution which concentrates all probability mass on one single atomic event. It turns out that the expected value of the agent's information indeed has the required intuitive properties making it highly suitable for the task at hand:

Def 5.2 $\quad \text{Ent}(P) = -\sum_{w \in W} P(w) \log P(w).$

The function Ent(P), the entropy of P, measures the expected informational gain of finding out what the real outcome is. This intuitive reading gives additional support to Ent as an uncertainty measure.

We are now in position to verify that entropy conforms with our intuition in example 4.1.

[8] I would like to thank Sten Lindström for giving me access to his unpublished notes on information theory.

[9] Other attempts to connect probabilistic belief updating and information theory can be found in the following papers: May (1976), May & Harper (1976), van Fraassen (1980) and Williams (1980). Cziszár (1977) is a discussion of information measures. Rosenkrantz (1977) gives postulates that characterize entropy. Shannon (1948) and Kullbach & Leibler (1951) are basic papers on relevant information-theoretical notions. For a modern textbook, see McEliece (1977)

Example 4.1 (continued). Recall that P=(1/2, 1/2, 0) and Q=(1/16, 1/16, 14/16). Calculation using 2 as the logarithm base give that Ent(P)=1 while Ent(Q)≈0.67. So P is, not surprisingly, more uncertain according to Ent than Q.

The previous argument motivates a new axiom that prevents the agent from shifting over too much probability to possibilities that were excluded in the prior distribution. The idea is that the posterior state after contraction should always be more uncertain than the prior state before contraction, i.e:

(G-6) $Ent(P^{-}_{A}) \geq Ent(P)$, with equality iff either $P(A) < 1$ or $A = W$.

Axiom (G-6) could be formulated more generally in terms of any uncertainty measure satisfying certain reasonable conditions. Here we have chosen to sacrifice generality for definiteness and ease of calculation. It is important to observe that the example showing that Gärdenfors' axioms for probabilistic contraction are insufficient did not rely on the use of entropy as the only way to measure the information content of a distribution. The example would go through using any reasonable information measure. In future work it should be investigated what exactly could be required of a reasonable information measure, an investigation that should lead to a more general axiom than (G-6).

6. CONCLUSION

It was argued that a rational agent should not be able to gain information simply by removing a previously accepted proposition. Such a belief change should instead always result in a less committed belief state. A simple example was given showing that the Gärdenfors axioms for probabilistic contraction in fact do not exclude that the agents uncertainty is decreased as a result of belief removal. The concept of entropy was borrowed from statistical information theory and used as a tool for measuring the uncertainty pertaining to a probability distribution. This concept was then used to formulate an a new axiom in addition to the Gärdenfors axioms. The new axiom explicitly rules out the possibility of agents getting more informed as a result of contraction.

REFERENCES

Csiszár, I. (1977) 'Information measures: a critical survey', in *Transactions of the Seventh Prague Conference,* Prague: Academia, 73-86.

Diaconis, P. and Zabell, L. (1982), 'Updating subjective probability', in *Journal of the American Statistical Association*, 77, 822-30.

Fuhrmann, A., Morreau, M. (1989) (eds) *The logic of Theory Change*, Springer-Verlag.

Gärdenfors, P. (1988) *Knowledge in flux*, The MIT Press.

Gärdenfors, P. (1992) (ed) *Belief Revision*, Cambridge Tracts in Theoretical Computer Science 29.

Gärdenfors, P. and Sahlin, N.-E. (1988) (eds) *Decision, Probability and Utility*, Cambridge University Press.

Harman, G. (1986) *Change in view*, The MIT Press.

Kullback, S. and Leibler R. A. (1951) 'On information and sufficiency', in *Ann. Math. Statist.*, 22, 79-86.

Lewis, D K (1976) 'Probabilities of conditionals and conditional probabilities', *The Philosophical Review* 85:297-315.

Lindström, S. and Rabinowicz, W. (1989), 'On probabilistic representation of non-probabilistic belief revision', in *Journal of Philosophical Logic*, 18, 69-101.

McEliece, R. J. (1977) *The theory of information and coding*, Addison-Wesley Publishing Company.

May, S. (1976) 'Probability kinematics: a constraint optimization problem', in *Journal of Philosophical Logic* 5, 395-398.

May, S. and Harper W. (1976) 'Toward an optimization procedure for applying minimum change principles in probability kinematics', in Harper and Hooker (eds), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. 1,* 137-166.

Rosenkrantz, R. D. (1977) *Inference, method and decision*, R. Reidel Publishing Company.

Shafer, G. (1976) *A mathematical theory of evidence,* Princeton University Press.

Shafer, G. (1981) 'Constructive probability', in *Synthese* 48, 1-60.

Shannon, C.E (1948) 'A mathematical theory of communication', in *Bell System Technical Journal,* Vol. 27, 379-423.

van Fraassen, B. C. (1980), 'Rational belief and probability kinematics', in *Philosophy of Science,* 47, 165-187.

Williams, P.M. (1980) 'Bayesian conditionalisation and the principle of minimum information', in *British Journal for the Philosophy of Science 31*, 131-144.