Bayesian and Non-Bayesian Evidential Updating

Henry E. Kyburg, Jr.
Department of Philosophy, University of Rochester, Rochester, NY 14627, U.S.A.

Recommended by Judea Pearl

ABSTRACT

Four main results are arrived at in this paper. (1) *Closed convex sets of classical probability functions provide a representation of belief that includes the representations provided by Shafer probability mass functions as a special case.* (2) *The impact of "uncertain evidence" can be (formally) represented by Dempster conditioning, in Shafer's framework.* (3) *The impact of "uncertain evidence" can be (formally) represented in the framework of convex sets of classical probabilities by classical conditionalization.* (4) *The probability intervals that result from Dempster–Shafer updating on uncertain evidence are included in (and may be properly included in) the intervals that result from Bayesian updating on uncertain evidence.*

1. Introduction

Recent work in both vision systems [37] and in knowledge representation [1, 7, 26, 29] has employed an alternative, often referred to as Dempster–Shafer updating, to classical Bayesian updating of uncertain knowledge. Various other investigators have gone beyond classical Bayesian conditionalization (MYCIN, EMYCIN, DENDRAL, . . .) but in a less systematic manner. It is appropriate to examine the formal relations between various Bayesian and non-Bayesian approaches to what has come to be called evidence theory, in order to explore the question of whether the new techniques are really more powerful than the old, and the question of whether, if they are, this increment of power is bought at too high a price.

2. Three Departures from Classical Probability Theory

Orthodox probability theory supposes that (1) we commence with known statistical distributions, (2) these distributions are such as to give rise to real-valued probabilities, and (3) these probabilities can be modified by using Bayes' theorem to conditionalize on evidence that is taken to be certain. There are thus three ways to modify the classical theory.

Artificial Intelligence **31** (1987) 271–293

We may dispense with the supposition that we are dealing with known statistical distributions. The best known advocate of this gambit was Savage [30], who argued that probabilities represent personal, subjective, opinions, and not objective distributions of quantities in the world. This approach has given rise to Bayesian statistics, based on the fact that the opinions of most people are such that, faced with frequency data, they will converge reasonably rapidly. Furthermore, in practice, it is common to recognize that some opinions are better than others, and to use as prior distributions in statistical inference the distributions representing the opinions of knowledgeable experts. This approach has been incorporated in some expert systems, for example, PROSPECTOR. It has both virtues and limitations. A purely pragmatic virtue is that it allows us to get on with our business even when we don't have the knowledge of prior distributions we would like to have. It has the practical virtue that the considered opinions of genuinely knowledgeable experts are formed in response to, and reflect with some degree of accuracy, relative frequencies in nature. But it has two drawbacks: it does not incorporate any indication of whether the opinion is a wild guess, or a considered judgement based on long experience; and it calls for expert opinions even in the face of total, acknowledged ignorance.

This suggests the second departure from the classical picture; abandoning the assumption that our probabilities are point-valued. This has recently been hailed as a novel departure [7, p. 1; 10, p. 319; 26, p. 21; 27, p. 7; 29, p. 9; 37, p. 16]. The idea of representing probabilities by intervals is not new (cf. [11, 18–25, 35, 36]), and the notion of probabilities that constitute a field richer than that of the real numbers goes back even further: Keynes [13, pp. 38–40] offers a formal philosophical treatment of such entities; Koopman [16, 17] offers a mathematical characterization. Even the standard subjectivistic or personalist view of probability can be construed in this way; while each person has a set of real-valued probabilities defined over a given field, a group of people will reflect a *set* of probability functions defined over the field. We may quite reasonably focus our attention on the maximum and minimum of these functions evaluated at a member of the field.[1]

[1] This approach is similar to that of Smith [35]. It is also similar to the approach of Levi [20, 24], Good [11] and Kyburg [18], but as Levi points out in [24] there are important differences. Levi represents a credal state by a set of conditional probability functions $Q(x, y)$. For every y consistent with background knowledge, the set of functions $Q(x, y)$ is convex. Since distinct convex sets of conditional probability functions give rise to the same convex sets of absolute probability functions, the two representations are not equivalent. Smith and Kyburg represent a credal state by the convex closure of all probabilities consistent with a set of probability intervals. Shafer, as will be seen, implicitly offers the same characterization. Dempster [5] offers a more restricted characterization: the convex set representing the credal state is the largest that both satisfies the interval constraints, *and* can be obtained from a space of "simple joint propositions" in a certain way. Levi has shown [24, pp. 338–392] that these additional restrictions are incompatible with certain natural forms of direct inference of probabilities from known statistics.

In general the representation in terms of intervals seems superior to the representation in terms of point values. Even in the *ideal* case, in which all of our measures are based on statistical inference from suitably massive quantities of data, it is most natural to construe these measures as being constrained by intervals. In confidence interval estimation, for example, what we get from our statistics is a high confidence that a given parameter is contained in a certain interval. This translates neatly and conveniently into an interval constraint. The results of statistical inference should reflect indeterminacy or vagueness. What we can properly claim to know is not that a parameter has certain value, but (with probability or high confidence) that it lies within certain limits. This limitation of human knowledge should surely be mirrored in computer-based systems.

The third departure from the classical scheme is to consider alternatives to Bayes' theorem as a way of updating probabilities in the light of new evidence. This departure is recent, and was first stated by Dempster [4]. Dempster's novel rule of combination, subsequently adopted by Shafer [32], is often referred to as a "generalization" of Bayesian inference (e.g., by Shafer [33, p. 337]: "The theory of belief functions . . . is a thoroughgoing generalization of the Bayesian theory . . ."; by Lowrance and Garvey [27, p. 9]: "Dempster's rule can be viewed as a direct generalization of Bayes' rule . . ."; and also in [7, p. 1; 10, p. 319; 26, p. 21]). This suggests, on the one hand, that Bayes' rule can be regarded as a special or limiting case of Dempster's rule, which is true, and on the other hand that Dempster's rule can be applied where Bayes' rule cannot, which is false. Dempster himself [4, 5] recognizes that his rule results from the imposition of additional constraints on the Bayesian analysis (see footnote 6).

One criticism of the usual Bayesian approach to evidential updating is the quantity of information that may be required to specify the probability function covering the field of propositions with which we are concerned. This may be *empirical* information (if the underlying probabilities are thought of as being based on statistical knowledge), *psychological* information (if a personalistic interpretation of probability is adopted), or *logical* information (if we interpret probability as degree of confirmation, à la Carnap [3]). Suppose we consider a field of propositions based on the logically independent propositions $p_1, \ldots, p_n$; the set of what Carnap called "state descriptions" induced by this basis consists of 2^n atoms, each of which is the conjunction of the n (negated or unnegated) p_i. It is obvious that for reasonably large n this assignment of probabilities presents great difficulties. But once we have those 2^n numbers, we're done—we can calculate all conditional probabilities as well as the probability of any proposition in the field based on $p_1, \ldots, p_n$.

Is there a saving in effort if we go to a Dempster–Shafer system? Using the handy representation of Shafer [32], we take θ, the universal set, to be the set of all 2^n possibilities represented by the state descriptions, and assign a mass to each *subset* of θ. This requires $2 \exp 2^n$ assignments! As far as the number of parameters to be taken account of is concerned, we are exponentially worse off. But if we construe probabilities as intervals, or represent them by sets of simple probability functions, we are just as badly off. (For an example relating mass assignments to interval assignments, see Table 1. For the general equivalence, see Theorem A.1.) Dillard [7, p. 4] refers to "computational limitations" and Lowrance and Garvey [27] mention that with large θ, maintaining the model is "computationally infeasible."

In either case, we need to find some systematic and computationally feasible procedure for obtaining the masses or probabilities we need. Bayesian and non-Bayesian approaches are in essentially the same difficult situation in this respect, although there are often plausible ways of systematizing the parameter assignments on either view.

3. Assumptions

Whether the representation of our initial knowledge state is given by an assignment of masses to subsets of θ or by a set of classical probability distributions over the atoms of θ, it is important that these masses or probabilities be justifiable. As already suggested, a straightforward way of obtaining them is through statistical inference, which (when possible) yields interval-valued estimates of relative frequencies. But there may also be other ways to obtain masses or intervals of probability. If so, then the deep and difficult problem arises of how to combine both statistical and nonstatistical sources of information.[2]

It has been suggested that Dempster–Shafer updating relieves us of the necessity of making assumptions about the joint probabilities of the objects we are concerned about. Thus, Quinlan claims that INFERNO "makes no assumptions whatever about the joint probability distributions of pieces of knowledge . . ." [29]. Other writers have made similar claims (e.g., Wesley and Hanson [37, p. 15]: "To make independence assumptions is exactly to make assumptions about joint probability distributions.").

It is clear that the assignment of masses to subsets of θ involves just as much in the way of "assumptions" as the assignment of a priori probabilities to the corresponding propositions. In view of the reducibility of the Dempster–Shafer formalism to the formalism provided by convex sets of classical probability functions (to be shown below), moreover, we may recapture the assumptions about joint probability distributions from the convex Bayesian representation.

4. Uncertain Evidence

One important novelty claimed for the Dempster–Schafer system is its ability to handle uncertain evidence. But even this is not in itself anti-Bayesian. There

[2] In another place I shall argue that we can find all our probabilities on direct or indirect statistical inference, or on set-theoretical truths. No other source is needed.

TABLE 1

	Mass	Lower measure	Upper measure
① AB	X_1	X_1	$1 - X_2 - X_3 - X_4 - X_{23} - X_{24} - X_{34} - X_{234}$
② $A\bar{B}$	X_2	X_2	$1 - X_2 - X_3 - X_4 - X_{13} - X_{14} - X_{34} - X_{134}$
③ $\bar{A}B$	X_3	X_3	$1 - X_1 - X_2 - X_4 - X_{12} - X_{14} - X_{24} - X_{124}$
④ $\overline{AB}$	X_4	X_4	$1 - X_1 - X_2 - X_3 - X_{12} - X_{13} - X_{23} - X_{123}$
①∪②	X_{12}	$X_1 + X_2 + X_{12}$	$1 - X_3 - X_4 - X_{34}$
①∪③	X_{13}	$X_1 + X_3 + X_{13}$	$1 - X_2 - X_4 - X_{24}$
①∪④	X_{14}	$X_1 + X_4 + X_{14}$	$1 - X_2 - X_3 - X_{23}$
②∪③	X_{23}	$X_2 + X_3 + X_{23}$	$1 - X_1 - X_4 - X_{14}$
②∪④	X_{24}	$X_2 + X_4 + X_{24}$	$1 - X_3 - X_1 - X_{13}$
③∪④	X_{34}	$X_3 + X_4 + X_{34}$	$1 - X_1 - X_2 - X_{12}$
①∪②∪③	X_{123}	$X_1 + X_2 + X_3 + X_{12} + X_{13} + X_{23} + X_{123}$	$1 - X_4$
①∪②∪④	X_{124}	$X_1 + X_2 + X_4 + X_{12} + X_{14} + X_{24} + X_{124}$	$1 - X_3$
①∪③∪④	X_{134}	$X_1 + X_3 + X_4 + X_{13} + X_{14} + X_{34} + X_{134}$	$1 - X_2$
②∪③∪④	X_{234}	$X_2 + X_3 + X_4 + X_{23} + X_{24} + X_{34} + X_{234}$	$1 - X_1$
θ	$X_\theta = 1 - \Sigma X_i$	1	1

are also Bayesian methods for handling uncertain evidence. One of these, used in PROSPECTOR and mentioned by Lowrance [26, p. 17] is known in the philosophical world as Jeffrey's rule. (It is presented and discussed in [12].) It follows from the addition and multiplication axioms that

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) .$$

If you adopt a new (coherent) probability function P', there are essentially no constraints on $P'(A)$. But one often confronts situations where if a shift in probability *originates* in the assignment of a new probability to B, that should *not* affect the *conditional* probability of A *given* B: $P(A|B) = P'(A|B)$. We have learned something new about B, but we haven't learned anything new about the bearing of the truth of B on the truth of A.

Given such a situation the response of a shift in the probability of B from $P(B)$ to $P'(B)$, resulting from new evidence, should propagate itself according to:

$$P(A) = P(A|B)P'(B) + P(A|\bar{B})P'(\bar{B}) .$$

When new evidence leads us to shift our credence in B from $P(B)$ to $P'(B)$, a corresponding shift in probability is induced for every other proposition in the field: the new probability of a proposition A is the weighted average of the probability of A, given B, and the probability of A given not-B, weighted by the new probabilities of B and not-B.

Lowrance [26] worries about the problem of iterating this move. Having made it, should we then update the probability of B in the light of the new probability $P'(A)$? Wesley and Hanson [37, p. 15] worry about a potential "violation of Bayes' law." But what is offered is not a relaxation method; it is a method of evaluating the impact of evidence which warrants a shift in the support for B. It makes no sense to consider updating $P'(B)$ in the light of the new value of $P(A)$; $P'(B)$ is the *source* of the updating. No contradiction lurks here; the notion of a *source* is clear to us, and can even be represented in artificial systems [14].

Other Bayesian updating procedures are possible (cf. [6, 7]), but it is hard to think of one so simple and often so natural. This is particularly true in the epistemological framework considered by Shafer; the weights of the masses assigned to subsets of θ reflect our a priori intuitions; there is no way in which the values of these masses, *given* our observations, can be changed without changing the model entirely. What impact *given* evidence has should not also change according to the evidence we *happen* to have. Shafer himself [33] has explored the relation between Jeffrey's rule and his own updating recommendations.

5. Belief Functions and Sets of Probabilities

In order to investigate more closely the relations between the Bayesian and Dempster–Shafer strategies for updating, it will be helpful to have several formal results. In this section we establish the partial equivalence between the assignment of masses to subsets of θ (the space of possibilities) and the assignment of a convex set of simple classical probability functions defined over the atoms of θ. The equivalence is only partial, since some plausible situations do not have a natural representation in terms of mass functions.

Shafer's belief functions are defined relative to a *frame of discernment* θ, and are given by either a belief function or a mass function defined over the subsets of θ. The atoms of θ are the most specific states of affairs that concern us in a given context. The belief function Bel and the mass function m are related by:

$$\mathrm{Bel}(X) = \sum_{A \subset X} m(A) .$$

(Throughout, "$\subset$" is to be understood as allowing improper inclusion. Proofs have been relegated to Appendix A.)

Our first observation is that to every belief function defined over a frame of discernment, there corresponds a closed set of classical probability functions S_P defined over the atoms of θ such that for any $X \subset \theta$,

$$\mathrm{Bel}(X) = \min_{P \in S_P} P(X) .$$

This result is Theorem A.1 which is stated and proved in Appendix A. The proof gives a way of constructing members of the set of classical probability functions, but the intuitive idea is simply this: Consider a set X, to which is assigned mass $m(X)$. That mass may be construed as probability mass that may be assigned in any way (subject to other constraints) to the atoms of X. We obtain the set of classical probability functions that corresponds to the mass function m by considering all the ways in which the mass that is not assigned to atoms by m can be assigned to atoms while maintaining the constraints imposed by the assignment of mass to *sets* of atoms. Tables 1 and 2 show both the general and a specific computation for a simple four-atom frame of discernment.

An example that shows the converse does not hold is the following[3]: Consider a compound experiment consisting of either (1) tossing a fair coin twice, or (2) drawing a coin from a bag containing 40% two-headed and 60% two-tailed coins and tossing it twice. The two parts (1) and (2) are performed in some unknown ratio p, so that, for example the probability that the first toss

[3] This example was suggested in conversation by Teddy Seidenfeld.

TABLE 2. A: white; B: magnetic

	Set	Mass	Frequency
① AB	X_1	0.2	[0.2, 0.4]
② $A\bar{B}$	X_2	0.2	[0.2, 0.4]
③ $\bar{A}B$	X_3	0.1	[0.1, 0.1]
④ $\overline{AB}$	X_4	0.2	[0.2, 0.4]
①∪②	X_{12}	0.1	[0.5, 0.7]
①∪③	X_{13}	0.0	[0.3, 0.5]
①∪④	X_{14}	0.1	[0.5, 0.7]
②∪③	X_{23}	0.0	[0.3, 0.5]
②∪④	X_{24}	0.1	[0.5, 0.7]
③∪④	X_{34}	0.0	[0.3, 0.5]
①∪②∪③	X_{123}	0.0	[0.6, 0.8]
①∪②∪④	X_{124}	0.0	[0.9, 0.9]
①∪③∪④	X_{134}	0.0	[0.6, 0.8]
②∪③∪④	X_{234}	0.0	[0.6, 0.8]
θ	$X_\theta = 1 - \Sigma X_i$	0.0	[1.0, 1.0]

lands heads is $p \cdot \frac{1}{2} + (1-p) \cdot 0.4$, $0 < p < 1$. Let A be the event that the first toss lands heads, and B the event that the second toss lands tails. The representation by a convex set of probability functions is straightforward. Since the frequency of the two kinds of experiment is completely unknown, the sample space may be taken as TT, TH, HT, HH. Probabilities are assigned as follows:

$$\begin{aligned}
&\text{TT:} \quad P(\bar{A} \cap B) = \tfrac{1}{4}p + 0.6(1-p) = p \cdot (\tfrac{1}{2})^2 + (1-p) \cdot 0.6(1)^2 \,, \\
&\text{TH:} \quad P(\bar{A} \cap \bar{B}) = \tfrac{1}{4}p + 0 = p \cdot (\tfrac{1}{2})^2 + (1-p) \cdot 0 \,, \\
&\text{HT:} \quad P(A \cap B) = \tfrac{1}{4}p + 0 = p \cdot (\tfrac{1}{2})^2 + (1-p) \cdot 0 \,, \\
&\text{HH:} \quad P(A \cap \bar{B}) = \tfrac{1}{4} + 0.4(1-p) = p \cdot (\tfrac{1}{2})^2 + (1-p) \cdot 0.4(1)^2 \,.
\end{aligned}$$

The convex set of probability functions P over the sample space is just the set of these measures for $p \in [0, 1]$:

$$S_P = \{\langle \tfrac{1}{4}p + 0.6(1-p), \tfrac{1}{4}p, \tfrac{1}{4}p, \tfrac{1}{4}p + 0.4(1-p)\rangle : p \in [0, 1]\} \,.$$

But if

$$P_* = \min_{p \in [0,1]} P \,,$$

then

$$\begin{aligned}
P_*(A \cup B) = 0.75 < 0.9 &= P_*(A) + P_*(B) - P_*(A \cap B) \\
&= 0.4 + 0.5 - 0 \,.
\end{aligned}$$

By Shafer's Theorem 2.1 [32]

$$\mathrm{Bel}(A \cup B) \geqslant \mathrm{Bel}(A) + \mathrm{Bel}(B) - \mathrm{Bel}(A \cap B) ,$$

P_* is therefore not a belief function. It is possible to compute a mass function, but the masses assigned to the union of any three atoms must be negative.

Subject to the condition, however, that for every $A_1, \ldots, A_n \subset \theta$,

$$P_*(A_1 \cup \cdots \cup A_n) \geqslant \sum_{I \subset \{1, \ldots, n\}} (-1)^{|I|+1} P_*\Big(\bigcap_{i \in I} A_i\Big) , \tag{1}$$

we can represent any closed convex set of classical probability functions by a Shafer function. This result is presented as Theorem A.2.

Theorems A.1 and A.2 show that the representation of uncertain knowledge provided by Shafer's probability mass functions is equivalent to a representation provided by a convex set of classical probability functions, and that the representation of uncertain knowledge by a convex set of classical probability functions is equivalent to a representation provided by a probability mass function so long as the convex set of probability functions satisfies the general relation (1) for every $A_1, \ldots, A_n \subset \theta$.

6. Dempster Conditioning and Convex Conditionalization

Of more interest than the mere representation of belief is the possibility of representing the way that beliefs should change in response to new evidence. If we represent beliefs by convex sets of distributions, one natural procedure for updating might be called "convex Bayesian conditionalization." If $\boldsymbol{P}$ is our initial convex set of distributions, then $\boldsymbol{P}_e$ is to be the set of the *conditional* distributions: If $P \in \boldsymbol{P}$ and $P(e) > 0$, then $P_e = P(\cdot\,|e) = P(\cdot\,\&e)/P(e)$ belongs to $\boldsymbol{P}_e$. The result of conditionalization applied to a convex set is convex, since for any a in $[0, 1]$ we can compute α to satisfy

$$aP_e(s) + (1 - a)P'_e(s) = \frac{\alpha P(s\&e) + (1 - \alpha)P'(s\&e)}{\alpha P(e) + (1 - \alpha)P'(e)} .$$

What we propose to look at in this section is the relation between Dempster–Shafer updating, and convex Bayesian updating. We shall first look at the relation in the case of evidence that is "certain"; and then we shall look at it in the case of "uncertain evidence."

Suppose that our beliefs can be represented either by a closed convex set of classical probability functions S_P, or by a Shafer mass function. Let B be evidence assigned probability 1, or support 1. Shafer defines upper and lower conditional support functions,

$$\mathrm{Bel}(A|B) = (\mathrm{Bel}(A \cup \bar{B}) - \mathrm{Bel}(\bar{B}))/(1 - \mathrm{Bel}(\bar{B})) ,$$

$$P^*(A|B) = P^*(A \cap B)/P^*(B) ,$$

where $P^*(X) = 1 - \mathrm{Bel}\,\bar{X}$ is called the plausibility of X.

Theorem A.3 shows that the following inequalities hold:

$$\min_{P \in S_P} P(A|B) \leq \mathrm{Bel}(A|B) \leq P^*(A|B) \leq \max_{P \in S_P} P(A|B) .$$

For the case of a frame of discernment with four atoms, illustrated in Table 1, we have the following, where X_i is the mass assigned to the set i in θ, X_{ij} is the mass assigned to the union of sets i and j, etc.

$$\min P(A|B) = \frac{X_1}{(X_1 + X_3) + (X_{13} + X_{23} + X_{34}) + (X_{123} + X_{134} + X_{234}) + X_\theta} ,$$

$$\mathrm{Bel}(A|B) = \frac{X_1 + (X_{12} + X_{14} + X_{124})}{(X_1 + X_3) + (X_{13} + X_{23} + X_{34}) + (X_{123} + X_{134} + X_{234}) + X_\theta + (X_{12} + X_{14} + X_{124})} ,$$

$$P^*(A|B) = \frac{X_1 + (X_{12} + X_{13} + X_{14}) + (X_{123} + X_{124} + X_{134}) + X_\theta}{(X_1 + X_3) + (X_{12} + X_{13} + X_{14}) + (X_{123} + X_{124} + X_{134}) + X_\theta + (X_{23} + X_{34} + X_{234})} ,$$

$$\max P(A|B) = \frac{X_1 + (X_{12} + X_{13} + X_{14}) + (X_{123} + X_{124} + X_{134}) + X_\theta}{(X_1 + X_3) + (X_{12} + X_{13} + X_{14}) + (X_{123} + X_{124} + X_{134}) + X_\theta} .$$

We observe that:

$$\min P(A|B) = \mathrm{Bel}(A|B) \quad \text{iff} \quad X_{12} + X_{14} + X_{124} = 0 ,$$

$$\mathrm{Bel}(A|B) = P^*(A|B) \quad \text{iff} \quad X_{13} + X_{123} + X_{134} + X_\theta = 0 ,$$

$$\max P(A|B) = P^*(A|B) \quad \text{iff} \quad X_{23} + X_{34} + X_{234} = 0 .$$

Before turning to a discussion of the inequalities of Theorem A.3, we show that they hold in general, and are not restricted to the case of "certain" evidence. Given two lemmas, the proof of the general result, Theorem A.6, is trivial. The two lemmas themselves may not be without interest.

7. Updating on Uncertain Evidence

The first lemma, Lemma A.4, states that by expanding the frame of discernment θ, we can represent the impact of uncertain evidence as the impact of "certain" evidence.[4] This is not to say that we need to *specify* that evidence; it is that there is an algorithm by means of which the impact of the uncertain evidence can be represented as the impact of other "certain" evidence.

[4] Although Lemma A.4 is stated only for *simple* support functions—support functions whose whole impact may be represented by an assignment of positive mass to a single subset of θ—*separable* support functions (the only ones that Shafer [32] attends to) are the support functions that can be obtained as the orthogonal sum of a sequence of *simple* support functions. What we show, therefore, seems to apply to all the support functions Shafer takes seriously.

The general idea of the argument is this. Suppose that θ is the frame of discernment, and that our initial belief function is Bel_1. The impact of uncertain evidence can be represented by a simple support function Bel_C whose single focus is $C \in 2^{\theta}$, to which Bel_C attributes mass s (and therefore mass $1-s$ to θ). To give a representation by "certain" evidence, we split every atom of θ into two new atoms to obtain θ'. We define a new belief function on θ', Bel_1', which is such that

(a) $\mathrm{Bel}_1'(X) = \mathrm{Bel}_1(X)$, if $X \subset \theta$,

(b) $(\mathrm{Bel}_1 \oplus \mathrm{Bel}_C)(X) = \mathrm{Bel}_1(X|E)$, if $X \subset \theta$, where E is a subset of θ' such that the evidence partially supporting C provides total support for E.

Two remarks on this construction are in order. First, we have given no rule for finding the "possibility" E. But in general that should be no problem. Suppose C is the proposition that there is a squirrel on the roof of the barn. The light is bad, so Bel_C assigns a mass of only 0.8 to C, and assigns the remaining mass to θ. We take E in θ' to be the proposition that it *seems* (0.8) to be the case that there is a squirrel on the roof, for which the evidence is conclusive. The index 0.8 indicates the force of the seeming, and is reflected in our assignment of masses in θ'. In many situations it seems quite natural to replace "uncertain evidence" by the "certain" data on which it is based.

Even the case discussed by Diaconis and Zabell [6] does not seem too difficult. The case is one in which we have one degree of belief that a Shakespearean actor to be heard on a record is Gielgud (say a half), but after hearing his voice for a while, we come to have a degree of belief of 0.8 that it is Gielgud. It is quite true that we would be hard put to it to describe in language the acoustic characteristics we come to assign to that voice with probability 1 that in turn provide evidence that it is the voice of Gielgud. But we can always *refer* to those characteristics as "the characteristics I have been (consciously or unconsciously) reacting to."

Second, however, whether or not we can always do this is unimportant for the comparison of Bayesian and Dempster conditioning. We can regard the introduction of E to be merely a computational device that helps us to compare the distribution of masses in θ according to the function $\mathrm{Bel}_1 \oplus \mathrm{Bel}_C$ to the corresponding set of Bayesian conditional distributions.

Lemma A.5 proves a corresponding fact about Jeffrey's rule for uncertain evidence.[5] It, too, may be represented as the effect of (possibly articificial) "certain" evidence. The argument is similar. Suppose our original degrees of

[5] This result was stated informally by Levi [21], and is reflected in Diaconis and Zabell's [6, Theorem 2.1]. In fact, any reasonable updating function must be representable as the result of conditionalization in an expanded space. This holds for convex updating, too. Suppose our original set of classical probability functions is $\boldsymbol{P}$, something happens to us to cause us to shift our probability from $P(e)$ to $P'(e)$ for each $P \in \boldsymbol{P}$. For each $P \in \boldsymbol{P}$ our uncertain observation is represented by a shift in the probability of e. Add the observation o to the algebra; we are free to suppose that for every $P \in \boldsymbol{P}$, $P'(e) = P(e|o)$, since the values of $P(e\&o)$ and of $P(\bar{e}\&o)$ are ours for the choosing.

belief are defined over a certain field of propositions. We introduce a new elementary statement into that field, thereby dividing each atom of the original field into two new atoms. The new elementary statement stands for that statement that, if it were "certain," would have just the effect that our "uncertain" evidence does. We then show that the resulting new probabilities obtained by conditionalizing on our new statement are exactly those yielded by applying Jeffrey's rule to the shift in probability of the "uncertain" evidence.

With these two results, and our previous theorem that shows the relation of Dempster–Shafer and convex Bayesian updating in the case of "certain" evidence, it follows immediately that the inequalities of Theorem A.3 hold whether or not the updating is done on the basis of "certain" evidence. In any case, the intervals resulting from Dempster–Shafer updating will be subintervals, and may be proper subintervals, of the intervals resulting from the application of conditionalization to sets of classical probability functions.[6]

8. Grounds for Decisions?

Dempster–Shafer evidential updating, we have seen, leads to more tightly constrained representations of rational belief than does convex Bayesian updating.[7] It might be thought that this is a virtue. But whether or not this is a "good thing" is open to question.

Suppose that $D = D_1, \ldots, D_n$ are alternative decisions open to you, and that you have a utility function defined over the cross product of D and the set θ of possible states. You begin with a belief function, and you obtain some evidence. If you combine this evidence with your initial belief function according to convex Bayesian conditionalization, your new beliefs will be characterized by a set of probability functions $\boldsymbol{P}_{\mathrm{B}}$. If you perform the combination of evidence according to non-Bayesian procedures, your new beliefs will be characterized by a set of probability functions $\boldsymbol{P}_{\mathrm{N}}$ that is (in general) a proper subset of $\boldsymbol{P}_{\mathrm{B}}$.[8]

Given any probability function P in either $\boldsymbol{P}_{\mathrm{B}}$ or $\boldsymbol{P}_{\mathrm{N}}$, you can calculate the expected value of each decision: $E(D_i, P)$. Let us say that D_i is *admissible* relative to a set of probability functions just in case there is some probability function in the set according to which the expected value of D_i is at least as

[6] Dempster [4, 5] was well aware that his rule of combination led to results stronger than those that would be given by a mere generalization of orthodox Bayesian inference. His reasons for preferring the rule at which he arrives are essentially philosophical: in an orthodox Bayesian framework, unless you restrict the family of priors, you don't get useful results starting with zero information. But in expert systems, we have no desire or need to start with zero information.

[7] Quinlan's [29] subtitle suggests the opposite: "A cautious approach to uncertain inference."

[8] It is not clear that Shafer's belief functions were intended to be used in a decision-theoretic context. Even if they were, there would be serious difficulties standing in the way of such employment [22, 23, 25, 31]. For present purposes, these difficulties need not concern us.

great as the expected value of any other decision.[9] Since P_N is included in P_B, the admissible decisions we obtain if we update in a non-Bayesian way are included among those we obtain if we update in a Bayesian way.

There are three cases to consider. (1) We obtain the same set of admissible decisions by either updating procedure. In this case we have gained nothing by using the stronger procedure. (2) If P_N leads to a set of admissible decisions containing more than one member, then so does P_B, and we must in either case invoke additional constraints in order to generate a unique decision. (3) If P_N leads to a unique admissible decision and P_B does not, we appear to have accomplished something useful by means of non-Bayesian updating.

But it is open to question whether the added power should be built into the evidential updating rule, or whether it should appear as part of a decision procedure that takes us beyond the evidence. Many people feel that principles of evidence and principles of decision should be kept distinct.

Consider an urn filled with black and white iron balls, some of which are magnetized and some of which are not. It is easy to imagine that by extensive sampling, or by word of the manufacturer, our statistical knowledge about the contents of the urn may be as represented in Table 2, where the set of black balls is represented by A, and the set of magnetized balls is represented by B. Given that this is our initial state, we may ask what our attitude should be toward the proposition that a ball selected from the urn is magnetic, given that it is white.

Dempster conditioning yields the degenerate interval [0.8, 0.8].

Bayesian conditionalization yields the interval [0.33, 0.8]. Suppose you are offered a ticket for \$0.75 that returns a dollar if the ball is magnetic. On the view identified with Dempster and Shafer, it is not only permissible, but, given the usual utility function, *mandatory* to buy it. On the convex Bayesian view either accepting or rejecting the offer would be admissible. It is true that, *for all you know*, the true expectation is positive; but it is also true that, *for all you know*, the true expectation is negative. If everything you know is true, the expected loss may still be \$0.25.

On the other hand, there are cases where Dempster's rule of combination leads to intuitively appealing results, but the convex Bayesian approach does not.[10] Suppose you know that 70% of the soft berries in a certain area are good to eat, and that 60% of the red berries are good to eat. What are the chances that a soft red berry is good to eat? Dempster's rule yields $0.42/0.54 = 0.78$, which has intuitive appeal. But the set of distributions compatible with the conditions of the problem as they have been stated leaves the probability of a soft red berry being good to eat completely undetermined: it is the entire

[9] This corresponds to Levi's notion [24] of E-admissibility.

[10] This elegant and simple example was proposed by Jerry Feldman.

interval [0, 1]! It is possible that 100% of the soft red berries are good, and it is possible that 0% of the soft red berries are good.

It is clear that, in applying the rule of combination, we are implicitly constraining the set of (joint) distributions we regard as possible. This is suggested by Shafer's requirement that the items of evidence to be combined be "distinct" or "independent." The most natural sufficient condition that leads to the same result as Dempster's rule of combination is that all the probability functions in our convex set satisfy the three conditions:

$$P(G) = \tfrac{1}{2}\,, \tag{2}$$

$$P(S|G\&R) = P(S|G)\,, \tag{3}$$

$$P(S|\bar{G}\&R) = P(S|\bar{G})\,. \tag{4}$$

Condition (2), of course, is our old friend, the principle of indifference. Conditions (3) and (4) represent conditional independence, and it is not hard to imagine that we have warrant for supposing they are satisfied.

The exact necessary and sufficient conditions for agreement between the two methods are that each of our set of probability functions satisfy the condition

$$P(\bar{G}\&R\&S)P(G\&R)P(G\&S) = P(G\&R\&S)P(\bar{G}\&R)P(\bar{G}\&S)\,. \tag{5}$$

Conditional independence of S and R, expressed in (3) and (4) is *not* sufficient for agreement, as this sample shows:

$$P(G\&R\&S) = P(G\&R\&\bar{S}) = P(G\&\bar{R}\&S) = P(G\&\bar{R}\&\bar{S}) = 0.075\,,$$
$$P(\bar{G}\&R\&S) = P(\bar{G}\&R\&\bar{S}) = P(\bar{G}\&\bar{R}\&S) = P(\bar{G}\&\bar{R}\&\bar{S}) = 0.175\,.$$

Conditional independence, as expressed in (3) and (4) is not necessary either, as the distribution

$$P(G\&R\&S) = P(\bar{G}\&R\&S) = P(G\&R\&\bar{S}) = P(\bar{G}\&\bar{R}\&S) = \tfrac{1}{8}\,,$$
$$P(G\&\bar{R}\&S) = P(\bar{G}\&R\&\bar{S}) = P(\bar{G}\&\bar{R}\&\bar{S}) = \tfrac{1}{16}\,,$$
$$P(G\&\bar{R}\&\bar{S}) = \tfrac{5}{16}$$

shows.

If we are taking our probabilities to be based on known or assumed general statistical relationships, then we should make this knowledge or these assumptions explicit so that they are open to critical assessment. This is not at all to say that conditional independence, say, can only be justified by *direct* statistical evidence. It is far more likely to be justified by our general knowledge of the causal connections, or the lack of them, between events belonging to two general classes. Nor is it to say that probabilities are to be *identified* with frequencies or chances. But that opens a quite different topic.

It should be strongly emphasized that the present arguments are not intended as arguments in favor of the general applicability of convex Bayesian conditionalization. Rather, what I have shown is (i) that the representation of belief states by distributions of masses over subsets of a set θ of possibilities is a special case of the convex Bayesian representation in terms of simple classical probabilities over the atoms of θ, (ii) that the treatments of uncertain evidence in both Bayesian and non-Bayesian updating are reducible to the corresponding treatments of certain evidence, and (iii) that non-Bayesian updating yields more determinate belief states as outcomes, but that the benefits afforded by non-Bayesian updating are limited and questionable.

Appendix A

Theorem A.1. *Let m be a probability mass function defined over a frame of discernment θ. Let* Bel *be the corresponding belief function,* $\mathrm{Bel}(X) = \Sigma_{A \subset X} m(A)$. *Then there is a closed, convex set of classical probability functions S_P defined over the atoms of θ such that for every subset X of θ,*

$$\mathrm{Bel}(X) = \min_{P \in S_P} P(X) .$$

Proof. Let S_P be the set of classical probability functions P defined on the atoms of θ such that for every $X \subset \theta$, $\mathrm{Bel}(X) \leqslant P(X) \leqslant 1 - \mathrm{Bel}(\bar{X})$. S_P is closed, since $P(X) = \mathrm{Bel}(X)$, $P(\bar{X}) = 1 - \mathrm{Bel}(X)$ is a classical probability function. S_P is convex, since for $0 < a < 1$, $aP_1(X) + (1 - a)P_2(X)$ lies between $\mathrm{Bel}(X)$ and $1 - \mathrm{Bel}(X)$ whenever $P_1(X)$ and $P_2(X)$ do. Since for any given X there is a $P \in S_P$ such that $P(X) = \mathrm{Bel}(X)$, $\mathrm{Bel}(X) \geqslant \min_{P \in S_P} P(X)$. And $\min_{P \in S_P} P(X) \geqslant \mathrm{Bel}(X)$ since this inequality holds for every $P \in S_P$.

To show that S_P is nonempty, it suffices to show that there is a $P \in S_P$ such that for every $X \subseteq \theta$, $\mathrm{Bel}(X) \leqslant P(X)$, since if this is so, then $\mathrm{Bel}(\bar{X}) \leqslant P(\bar{X})$ and $1 - \mathrm{Bel}(\bar{X}) \geqslant 1 - P(\bar{X}) = P(X)$.

Suppose the atoms of θ are ordered lexicographically. For every set X, $X \subset \theta$, add the mass assigned to X, $m(X)$, to the mass assigned to $\{a_i\}$, where $\{a_i\}$ is the lexicographically earliest atom in X. Let the new mass function be m'. Define $P(X) = \Sigma_{a \in X} m'(\{a\})$. $P(\emptyset) = 0$; $P(\theta) = 1$, since all the original mass ends up on the atoms, and $P(X) \geqslant \mathrm{Bel}(X)$, since the mass assigned to any subset of X ends up on the atoms of X. □

Theorem A.2. *If S_P is a closed convex set of classical probability functions defined over the atoms of θ, and for every $A_1, \ldots, A_n \subset \theta$,*

$$\min P(A_1 \cup \cdots \cup A_n) \geqslant \sum_{I \subset \{1, \ldots n\}} (-1)^{|I|+1} \min P\Big(\bigcap_{i \in I} A_i\Big),$$

then there is a mass function m defined over the subsets of θ such that for every X in θ, the corresponding Bel *function satisfies*

$$\mathrm{Bel}(X) = \min_{P \in S_P} P(X) .$$

Proof. Since S_P is closed and convex, for every $X \subset \theta$ there is a $P \in S_P$ such that $P(X) = \min_{P \in S_P} P(X)$. For every $X \subset \theta$, define

$$P_*(X) = \min_{P \in S_P} P(X) .$$

By Shafer's Theorem 2.1 [32] if θ is a frame of discernment the function $\mathrm{Bel}\, 2^\theta \to [0, 1]$ is a belief function if and only if

(1) $\mathrm{Bel}(\emptyset) = 0$, $P_*(\emptyset) = 0$,

(2) $\mathrm{Bel}(\theta) = 1$, $P_*(\theta) = 1$,

(3) for every positive integer n and every collection $A_1, \ldots, A_n$ of subsets of θ,

$$\mathrm{Bel}(A_1 \cup \cdots \cup A_n) \geqslant \sum_{I \subset \{1, \ldots, n\}} (-1)^{|I|+1} \mathrm{Bel}\Big(\bigcap_{i \in I} A_i\Big) .$$

Since Shafer's [32, Theorem 2.2] gives an algorithm to recapture the mass function from the belief function, we need merely establish (3) for our function P_*.

But (3) is just the transcription of the second condition of the theorem into Shafer's notation. □

Theorem A.3. *Let θ be a frame of discernment,* Bel *a belief function, and S_P the corresponding set of Bayesian probability functions. Let B be evidence assigned probability* 1, *or support* 1, *and suppose $P(B) > 0$ for every $P \in S_P$. Then for every $A \subset \theta$,*

$$\min_{P \in S_P} P(A|B) \leqslant \mathrm{Bel}(A|B) \leqslant P^*(A|B) \leqslant \max_{P \in S_P} P(A|B) ,$$

where $P^*(A|B) = 1 - \mathrm{Bel}(\bar{A}|B)$ is Shafer's plausibility function.

Proof. (All maxima and minima are taken over $P \in S_P$.) For $X \subset \theta$,

$$\mathrm{Bel}(X) = \min P(X) \quad \text{and} \quad P^*(X) = \max P(X)\,.$$

$$\mathrm{Bel}(A|B) = \frac{\mathrm{Bel}(A \cup \bar{B}) - \mathrm{Bel}(\bar{B})}{1 - \mathrm{Bel}(\bar{B})} = \frac{\min P(A \cup \bar{B}) - \min(\bar{B})}{1 - \min(\bar{B})}\,,$$

$$\min P(A|B) = \min\left[\frac{P(A \cap B)}{P(B)}\right] = \min\left[\frac{P(A \cup \bar{B}) - P(\bar{B})}{1 - P(\bar{B})}\right].$$

Let $Q \in S_P$ be such that $Q(A \cup \bar{B}) = \min P(A \cup \bar{B})$. Then

$$\min P(A|B) \leqslant \frac{Q(A \cup \bar{B}) - Q(\bar{B})}{1 - Q(\bar{B})} \leqslant \frac{Q(A \cup \bar{B}) - \min P(\bar{B})}{1 - \min P(\bar{B})}$$
$$= \mathrm{Bel}(A|B)\,,$$

$$\max P(A|B) = \max\left[\frac{P(A \cap B)}{P(B)}\right].$$

Let $R \in S_P$ be such that $R(A \cap B) = \max P(A \cap B)$. Then

$$\max P(A|B) \geqslant \frac{R(A \cap B)}{R(B)} \geqslant \frac{\max P(A \cap B)}{\max P(B)} = P^*(A|B)\,. \qquad \square$$

Lemma A.4. *Let θ be a frame of discernment. Let our initial belief function be* Bel_1. *We obtain new evidence whose impact on the frame of discernment θ can be represented by a simple support function* [32, p. 7] Bel_C *whose single focus is* $C \in 2^\theta$. Bel_C *attributes mass s to C and mass $(1 - s)$ to θ.*

Let the foci of Bel_1*—the subsets A of θ receiving mass $m_1(A) > 0$—be A_1, $A_2, \ldots, A_n$. We can construct a new frame of discernment θ' and a new belief function* Bel'_1, *such that*

(a) *for every $X \subset \theta$,* $\mathrm{Bel}'_1(X) = \mathrm{Bel}_1(X)$;

(b) *for every $X \subset \theta$,*

$$(\mathrm{Bel}_1 \oplus \mathrm{Bel}_C)(X) = \mathrm{Bel}'_1(X|E)\,,$$

where $E \in 2^{\theta'}$, and the evidence partially supporting C provides "total" support for E. "$\oplus$" represents the application of Dempster's rule of combination to Bel_1 *and* Bel_C; $\mathrm{Bel}'_1(X|E)$ *represents Dempster's rule of conditioning on E—the analog of Bayesian conditionalization* [32, p. 67].

Proof. Let e be new to θ, and for every "possible world" w in θ generate two new "possibilities" we and $w\bar{e}$. Let $\theta' = \{w' \colon \exists w \in \theta\ (w' = we \vee w' = w\bar{e})\}$.

Let $E = \{w' : \exists w \in \theta\ (w' = we)\}$. For every subset X of θ, let us write X' for the corresponding subset of θ'—i.e., $X' = \{w' \in \theta' : \exists w \in X\ (w' = we \vee w' = w\bar{e})\}$. Since the evidence that supports C is to render E certain, we have $C' \subset E$; i.e. $C' = \{w' : \exists w \in C\ (w' = we)\}$.

We define a new belief function Bel_1' on θ' as follows: Bel_1' has n foci A_i' corresponding to the foci of Bel_1, each with mass $(1-s) \cdot m_1(A_i)$.

For every $i \leqslant n$ such that $A_i' \cap C' = \emptyset$, $A_i' \cap \bar{E}$ is to be a focus with mass $s \cdot m_1(A_i)$. For convenience we take the first p of the A_i to be those for which $A_i' \cap C' = \emptyset$. Note that p may be 0, but cannot be n else $\mathrm{Bel}_1 \oplus \mathrm{Bel}_C$ would be undefined.

The remaining i give rise to the remaining foci. These are of the form $(A_i' \cap C') \cup (A_i' \cap \bar{E})$, and receive the remaining mass. Since

$$(A_i' \cap C') \cup (A_i' \cap \bar{E}) = (A_j' \cap C) \cup (A_j' \cap \bar{E})$$

is a possibility for $i \neq j$, we write

$$m_1'((A_i' \cap C') \cup (A_i' \cap \bar{E})) = \sum_{j=J} s \cdot m_1(A_j)\,,$$

where $J = \{j : (A_j' \cap C') \cup (A_j' \cap \bar{E}) = (A_i' \cap C') \cup (A_i' \cap \bar{E})\}$. Note that

$$\sum m_1'((A_i' \cap C') \cup (A_i' \cap \bar{E})) = \sum_{i=p+1}^{n} s \cdot m_1(A_i)\,,$$

since these sets have positive mass only if $A_i' \cap C' \neq \emptyset$.

We first show that Bel_1' is a belief function. Obviously its mass function m' is nonnegative for every $X' \subset \theta'$, so we need only show that $\Sigma_{X' \subset \theta'} m'(X') = 1$. Summing over the three kinds of foci, we have:

$$\begin{aligned}\sum_{X' \subset \theta'} m'(X') &= \sum_{i=1}^{n} (1-s) \cdot m_1(A_i) + \sum_{i=1}^{p} s \cdot m_1(A_i) + \sum_{i=p+1}^{n} s \cdot m_1(A_i) \\ &= 1\,.\end{aligned}$$

We next show that Bel_1' is equivalent to Bel_1—i.e. that for any $X \subset \theta$, $\mathrm{Bel}_1'(X') = \mathrm{Bel}_1(X)$.

$$\begin{aligned}\mathrm{Bel}_1'(X') &= \sum_{A' \subset X'} m'(A') \\ &= \sum_{A_i' \subset X'} m'(A_i') + \sum_{A_i' \cap E \subset X'} m'(A_i' \cap E) \\ &\quad + \sum_{(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X'} m'((A_i' \cap C') \cup (A_i' \cap \bar{E}))\,.\end{aligned}$$

The first term yields

$$\sum_{A_i'\subset X'} (1-s)\cdot m_1(A_i) = (1-s) \sum_{A_i\subset X} m_1(A_i)\,.$$

Since $X' = (X' \cap E) \cup (X' \cap \bar{E})$, $A_i' \cap \bar{E} \subset X' \cap \bar{E}$ if and only if $A_i \subset X$, in view of the fact that $p\bar{e} \in A_i' \cap \bar{E}$ if and only if $p \in A_i$, and the same holds for X. Thus the second term yields

$$\sum_{\substack{A_i\subset X\\ 1\leq i\leq p}} s\cdot m_1(A_i)\,.$$

To evaluate the third term, we claim that $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X'$ if and only if $A_i \subset X$. If $A_i \subset X$, then $A_i \cap C \subset X$ and $A_i' \cap \bar{E} \subset X'$ and so $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X'$. Suppose $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X'$. Then $A_i' \cap \bar{E} \subset X'$, $A_i' \cap \bar{E} \subset X' \cap \bar{E}$, and by the preceding argument $A_i \subset X$. Thus the third term yields

$$\sum_{\substack{A_i\subset X\\ p<i<n}} s\cdot m_1(A_i)\,.$$

Putting the three parts together, we have $\mathrm{Bel}_1'(X') = \mathrm{Bel}_1(X)$.

We now show that conditioning on E in the frame of discernment θ' is equivalent to combining uncertain evidence C with Bel_1 in the frame of discernment θ according to Dempster's rule of combination:

For every $X \subset \theta$, $(\mathrm{Bel}_1 \oplus \mathrm{Bel}_C)(X) = \mathrm{Bel}_1'(X'|E)$,

$$(\mathrm{Bel}_1 \oplus \mathrm{Bel}_C)(X) = \frac{\displaystyle\sum_{\emptyset\neq A_i\cap C\subset X} s\cdot m_1(A_i) + \sum_{A_i\subset X} (1-s)\cdot m_1(A_i)}{\displaystyle 1-\sum_{A_i\cap C=\emptyset} s\cdot m_1(A_i)}\,. \tag{A.1}$$

(The numerator comprises two sums, since Bel_C has two foci: C and θ with masses s and $(1-s)$ respectively.)

$$\mathrm{Bel}_1'(X'|E) = \frac{\displaystyle\sum_{A'\subset X'\cup\bar{E}} m_1'(A') - \sum_{A'\subset\bar{E}} m_1'(A')}{\displaystyle 1-\sum_{A'\subset\bar{E}} m_1'(A')}\,, \tag{A.2}$$

where

$$\sum_{A' \subset \bar{E}} m_1'(A') = \sum_{i=1}^{p} s \cdot m_1(A_i)\,,$$

since only the foci of the form $A_i' \cap \bar{E}$ are included in $\bar{E}$: $A_i' = (A_i' \cap E) \cup (A_i' \cap \bar{E})$ is not included in $\bar{E}$, and since $C' \subset E$, $(A_i' \cap C') \cup (A_i' \cap \bar{E})$ is included in $\bar{E}$ only if $A_i' \cap C' = \emptyset$, in which case it has no mass.

$$\sum_{A_i' \cap C' = \emptyset} s \cdot m_1(A_i) = \sum_{i=1}^{p} s \cdot m_1(A_i)\,.$$

Hence the denominators of (A.1) and (A.2) are the same.

It remains to evaluate $\Sigma_{A' \subset X' \cap \bar{E}}\, m_1'(A')$. Consider foci of the form A_i'. $A_i' \subset X' \cup \bar{E}$ if and only if $A_i \subset X$, so these foci yield mass

$$\sum_{A_i' \subset X'} m'(A_i') = \sum_{A_i \subset X} (1-s) \cdot m_1(A_i)$$

corresponding to the right-hand term in the numerator of (A.1). Consider foci of the form $A_i' \cap \bar{E}$. All of these are included in $X' \cup \bar{E}$; they yield

$$\sum_{i=1}^{p} m'(A_i') = \sum_{i=1}^{p} s \cdot m_1(A_i) = \sum_{A' \subset \bar{E}} m_1'(A')\,,$$

so they drop out of the numerator of (A.2).

Finally, consider foci of the form $(A_i' \cap C') \cup (A_i' \cap \bar{E})$. We first show that $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X' \cup \bar{E}$ if and only if $A_i' \cap C' \subset X'$. Suppose $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X' \cup \bar{E}$. Then $A_i' \cap C' \subset (X' \cup \bar{E})$. But $C' \subset E$, so $A_i' \cap C' = A_i' \cap C' \cap E \subset X' \cup \bar{E}$ only if $A_i' \cap C' \subset X'$. Suppose $A_i' \cap C' \subset X'$. Then since $A_i' \cap \bar{E} \subset \bar{E} \subset X' \cup \bar{E}$, $(A_i' \cap C') \cup (A_i' \cap \bar{E}) \subset X' \cup \bar{E}$.

We compute the mass in the numerator of (A.2) due to foci of this sort. They have mass only when $A_i' \cap C' \neq 0$. And then they have mass

$$\sum_{j \in J} s \cdot m_1(A_i)\,,$$

where $J = \{j\colon (A_j' \cap C') \cup (A_j' \cap \bar{E}) = (A_i' \cap C') \cup (A_i' \cap \bar{E})\}$.

Each A_i such that $A_i' \cap C' \subset X'$ contributes $s \cdot m_1(A_i)$. Their total mass is therefore

$$\sum_{\emptyset \neq A_i' \cap C' \subset X} s \cdot m_1(A_i)\,,$$

corresponding to the first term of the numerator of (A.1). We have therefore shown that $(\mathrm{Bel}_1 \oplus \mathrm{Bel}_C)(X) = \mathrm{Bel}_1'(X'|E)$. $\square$

Lemma A.5. *Suppose that P_0 is an assignment of probabilities to the field of propositions whose basis is $a_1, a_2, a_3, \ldots, a_n$. Let P_1 be generated by a shift in the probability assigned to A; this shift is the source of our new probability P_1. By Jeffrey's rule, for all X,*

$$P_1(X) = P_0(X|A) \cdot P_1(A) + P_0(X|A) \cdot P_1(A) .$$

Then there exists a new field of propositions $\mathscr{F}'$, and a proposition E, and a new probability function P_0' defined on $\mathscr{F}'$ such that for every proposition X in the old field $\mathscr{F}$,

$$P_0'(X) = P_0(X) \quad \textit{and} \quad P_0'(X|E) = P_1(X) .$$

Proof. Add a new basic proposition e to the basis of $\mathscr{F}$ to obtain the field $\mathscr{F}'$, and represent it by E. We impose the constraint $P_0'(A|E) = P_1(A)$; $P_0'(E)$ may have any value that strikes our fancy.

We extend P_0' so that for any $X \in \mathscr{F}$, $P_0'(X) = P_0(X)$; P_0' is fully equivalent to P_0, so far as $\mathscr{F}$ is concerned, before we obtain information about A. Specifically, set

$$k = \frac{P_1(A)}{P_0(A)} ;$$

$$k' = \frac{P_1(\bar{A})}{P_0(\bar{A})} = \frac{1 - P_1(A)}{1 - P_0(A)} = \frac{1 - k \cdot P_0(A)}{1 - P_0(A)} .$$

For $X \in \mathscr{F}'$, set

$$P_0'(X \wedge E) = P_0'(E)[k \cdot P_0(X \wedge A) + k' \cdot P_0(X \wedge \bar{A})] ,$$

$$P_0'(X \wedge \bar{E}) = P_0(X) - P_0'(X \wedge E) .$$

Clearly, for $X \in \mathscr{F}$,

$$P_0'(X) = P_0'(X \wedge E) + P_0'(X \wedge \bar{E}) = P_0(X) .$$

We now show that for $X \in \mathscr{F}$, probabilities conditional on E are equal to the probabilities given by Jeffrey's rule: $P_1(X) = P_0'(X|E)$.

For $X \in \mathscr{F}$,

$$P_0'(X|E) = \frac{P_0(X \wedge E)}{P_0(E)} = \frac{P_0'(E)[k \cdot P_0(X \wedge A) + k' \cdot P_0(X \wedge \bar{A})]}{P_0'(E)}$$

$$= \frac{P_0(X \wedge A)}{P_0(A)} P_1(A) + \frac{P_0(X \wedge \bar{A})}{P_0(\bar{A})} P_1(\bar{A}) = P_1(X) . \qquad \square$$

Theorem A.6. *Let a distribution of beliefs be given both by the function* Bel_1 *and by the prior set of probability distributions* S_P. *Suppose new evidence is obtained whose impact is given by a simple support function* Bel_A *assigning positive mass to* A *and* θ, *or, alternatively, by a shift in the probability of* A *on each of the distributions in* S_{P_1}; *let* S_{P_2} *be the result of propagating this shift by Jeffrey's rule, and let* Bel_2 *be the result of applying Dempster's rule of combination. Then*

$$\min_{P \in S_{P_2}} P(X) \leqslant \mathrm{Bel}_2(X) \leqslant 1 - \mathrm{Bel}_2(\bar{X}) \leqslant \max_{P \in S_{P_2}} P(X)$$

for all subsets X *of* θ.

Proof. Immediate from Lemmas A.4 and A.5, and Theorem A.3. □

ACKNOWLEDGMENT

Research for this paper was supported in part by the U.S. Army Signals Warfare Laboratory, and was stimulated in large part by conversations with Jerry Feldman and Ron Loui of the Department of Computer Science at the University of Rochester. Judea Pearl and Norman Dalkey carried out their duties as referees with exemplary efficiency and offered much good advice. I hope I have succeeded in following it. An anonymous referee pointed out an error in the original proof of Theorem A.3, and provided a suggestion to correct it.

REFERENCES

1. Barnett, J.A., Computational methods for a mathetmatical theory of evidence, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 868–875.
2. Buchanan, B.G., Sutherland, G.L. and Feigenbaum, E.A., Heuristic Dendral: A program for generating explanatory hypotheses in organic chemistry, in: B. Meltzer and D. Michie (Eds.), *Machine Intelligence* (Edinburgh University Press, Edinburgh, 1969).
3. Carnap, R., *The Logical Foundations of Probability* (University of Chicago Press, Chicago, IL, 1950).
4. Dempster, A.P., Upper and lower probabilities induced by a multi-valued mapping, *Ann. Math. Stat.* **38** (1967) 325–339.
5. Dempster, A.P., A generalization of Bayesian inference, *J. Roy. Stat. Soc. B* **30** (1968) 205–247.
6. Diaconis, P. and Zabell, S., Updating subjective probabilities, *J. Am. Stat. Assoc.* **77** (1982) 822–830.
7. Dillard, R.A., The Dempster–Shafer theory applied to tactical fusion in an inference system, in: *Proceedings Fifth MIT/ONR Workshop*, Cambridge, MA, 1982.
8. Duda, R.O., Hart, P.E. and Nilsson, N.J., Subjective Bayesian methods for rule-based inference systems, in: *Proceedings 1976 National Computer Conference*, New York (1976) 1075–1082.
9. Field, H., A note on Jeffrey conditionalization, *Philos. Sci.* **45** (1978) 361–367.
10. Garvey, R.D., Lowrance, J.D. and Fishler, M.A., An inference technique for integrating knowledge from disparate sources, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 319–325.
11. Good, I.J., Subjective probability as a measure of a non-measurable set, in: H.H. Nagel, Suppes and Tarski (Eds.), *Logic, Methodology and Philosophy of Science* (Stanford University Press, Stanford, CA, 1962).

12. Jeffrey, R., *The Logic of Decision* (McGraw-Hill, New York, 1965).
13. Keynes, J.M., *A Treatise on Probability* (MacMillan, London, 1921).
14. Kim, J. and Pearl, J., A computational model for causal and diagnostic reasoning in inference systems, in: *Proceedings IJCAI-83*, Karlsruhe, F.R.G. (1983) 190–193.
15. Koopman, B.O., The bases of probability, *Bull. Am. Math. Soc.* **46** (1940) 763–774.
16. Koopman, B.O., The axioms and algebra of intuitive probability, *Ann. Math.* **41** (1940) 269–292.
17. Koopman, B.O., Intuitive probabilities and sequences, *Ann. Math.* **42** (1941) 169–187.
18. Kyburg, H.E., Jr., *Probability and the Logic of Rational Belief* (Wesleyan University Press, Middletown, CT, 1961).
19. Kyburg, H.E., Jr., *The Logical Foundations of Statistical Inference* (Reidel, Dordrecht, The Netherlands, 1974).
20. Levi, I., On indeterminate probabilities, *J. Philos.* **71** (1974) 391–418.
21. Levi, I., Probability kinematics, *British J. Philos. Sci.* **18** (1967) 197–209.
22. Levi, I., Dissonance and consistency according to Shackle and Shafer, in: P. Asquith and I. Hacking (Eds.), *PSA 1978* (Philosophy of Science Association, 1979) 466–477.
23. Levi, I., Potential surprise: Its role in inference and decision making, in: L.J. Cohen and M. Hesse (Eds.), *Applications of Inductive Logic* (Oxford University Press, Oxford, 1980) 1–27.
24. Levi, I., *The Enterprise of Knowledge* (MIT Press, Cambridge, MA, 1981).
25. Levi, I., Consonance, dissonance and evidentiary mechanism, in: Gardenfors, Hansson and Sahlin (Eds.), *Evidentiary Value*, Library of Theoria **15** (1983) 27–43.
26. Lowrance, E., Dependency-graph models of evidential support, Tech. Rept., University of Massachusetts, Amherst, MA, 1982.
27. Lowrance, J.D. and Garvey, T.D., Evidential reasoning: A developing concept, in: *IEEE 1982 Proceedings International Conference on Cybernetics and Society*, Seattle, WA (1982) 6–9.
28. Mackie, J.S., The relevance criterion of confirmation, *British J. Philos. Sci.* **20** (1969) 27–40.
29. Quinlan, J.R., *Inferno: A cautious approach to uncertain inference*, Rand Note, CA, 1982.
30. Savage, L.J., *The Foundations of Statistics* (Wiley, New York, 1954).
31. Seidenfeld, T., Statistical evidence and belief functions, in: P. Asquith and I. Hacking (Eds.), *PSA 1978* (Philosophy of Science Association, 1979) 451–465.
32. Shafer, G., *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ, 1976).
33. Shafer, G., Jeffrey's rule of conditioning, *Philos. Sci.* **48** (1981) 337–362.
34. Shortliffe, E.H., *Computer-Based Medical Consultations: MYCIN* (American Elsevier, New York, 1976).
35. Smith, C.A.B., Consistency in statistical inference and decision, *J. Roy. Stat. Soc. B* **23** (1961) 1–37.
36. Smith, C.A.B., Personal probability and statistical analysis, *J. Roy. Stat. Soc. A* **128** (1965) 469–499.
37. Wesley, L.P. and Hanson, A.R., The use of an evidential based model for representing knowledge and reasoning about images in the vision system, *IEEE 1982 Proceedings International Conference on Cybernetics and Society*, Seattle, WA (1982) 14–25.

Received May 1984; revised version received May 1986