Two views of belief: Belief as generalized probability and belief as evidence

Joseph Y. Halpern
Ronald Fagin

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

email: halpern@ibm.com, fagin@ibm.com

This is essentially the same as a paper which appears in *Artificial Intelligence* **54**, 1992, pp. 275–317.

Abstract: Belief functions are mathematical objects defined to satisfy three axioms that look somewhat similar to the Kolmogorov axioms defining probability functions. We argue that there are (at least) two useful and quite different ways of understanding belief functions. The first is as a generalized probability function (which technically corresponds to the inner measure induced by a probability function). The second is as a way of representing *evidence*. Evidence, in turn, can be understood as a mapping from probability functions to probability functions. It makes sense to think of *updating* a belief if we think of it as a generalized probability. On the other hand, it makes sense to *combine* two beliefs (using, say, Dempster's *rule of combination*) only if we think of the belief functions as representing evidence. Many previous papers have pointed out problems with the belief function approach; the claim of this paper is that these problems can be explained as a consequence of confounding these two views of belief functions.

1 Introduction

A *belief function* is a function that assigns to every subset of a given set S a number between 0 and 1. Intuitively, the belief in a set (or event) A is meant to describe a lower bound on the degree of belief of an agent that A is actually the case. The corresponding upper bound is provided by a *plausibility function*. The idea of a belief function was introduced by Dempster [Dem67, Dem68] (he uses the terms *lower probability* for belief and *upper probability* for plausibility), and then put forth as a framework for reasoning about uncertainty in Shafer's seminal work *A Mathematical Theory of Evidence* [Sha76a]. Since then belief functions have become a standard tool in expert systems applications (see, for example, [Abe88, Fal88, LU88, LG83]).

While belief functions have an attractive mathematical theory and many intuitively appealing properties, there has been a constant barrage of criticism directed against them, going back to when they were first introduced by Dempster (see the discussion papers that appear after [Dem68], particularly the comments of Smith, Aitchison, and Thompson). The fundamental concern seems to be how we should interpret belief functions. This point is made in a particularly sharp way by Diaconis and Zabell [Dia78, DZ86]. They consider the *three prisoners problem*, and show that applying the belief function approach to this problem, particularly Dempster's *rule of combination* (which is a rule for combining two belief functions to produce a new belief functions) leads to counterintuitive results. Other authors have shown that the belief function approach leads to counterintuitive or incorrect answers in a number of other situations (see, for example, [Ait68, Bla87, Hun87, Lem86, Pea88, Pea89, Zad84]).

In this paper, we argue that essentially all these problems stem from a confounding of two different views of belief functions: the first is as a generalized probability function, while the second is as a representation of evidence. In the remainder of this introduction, we briefly sketch these two views.

Formally, a belief function can be defined as a function satisfying three axioms (just as a *group* is a mathematical object satisfying a certain set of axioms). These axioms can be viewed as a weakening of the Kolmogorov axioms that characterize probability functions. From that point of view, it seems reasonable to try to understand a belief function as a generalized probability function. A number of authors have in fact tried to find characterizations of belief functions in terms of probability functions (e.g., [Dem67, Dem68, FH91b, FH91a, Kyb87, Rus87, Sha79]). We focus here on the approach of [FH91b, FH91a].

A probability function is a function that assigns a number between 0 and 1 to some (*but not necessarily all*) of the subsets of a set. The sets to which a probability is assigned are called *measurable* sets. Note the contrast here with belief functions, which do assign a number to all subsets of a set. There are two standard ways of extending a probability function Pr so that it is defined on all subsets: namely, by considering the *inner measure* Pr_* and *outer measure* Pr^* induced by Pr. Intuitively, the inner measure of a set A is the best approximation we can make to its probability from below, while the upper measure is the best approximation from above. Thus, the inner and outer measure of a set A define an interval, just as do the belief and plausibility of A. This analogy is more than a superficial one. It is straightforward to show if we are given a probability function Pr, then Pr_* is a belief function (in that it satisfies the three axioms characterizing belief functions) and Pr^* is the corresponding plausibility function. Moreover, the converse essentially holds; every belief function can essentially be viewed as being the inner measure induced by some probability function [FH91b].

Thus, we have a natural way of viewing a belief function as a generalized probability function: it is just an inner measure induced by a probability function.

This view of a belief function as a generalized probability function is quite different from the view taken by Shafer in [Sha76a]. Here, belief is viewed as a representation of *evidence*. The more evidence we have to support a particular proposition, the greater our belief in that proposition.

Now the question arises as to what exactly evidence is, and how it relates to probability (if at all). Notice that if we start with a probability function and then we get some evidence, then we can *update* our original probability function to take this evidence into account. If the evidence comes in the form of an observation of some event B, then this updating is typically done by moving to the conditional probability. Starting with a probability function Pr, we update it to get the (conditional) probability function $Pr(\cdot|B)$. This suggests that evidence can be represented by a function that takes as an argument a probability function and returns an updated probability function. By using ideas that already appear in [Sha76a], it can be shown that a belief function can in fact be viewed as representing evidence in this sense.

This point is perhaps best understood in terms of an example. Imagine we toss a coin that is either a fair coin or a double-headed coin. We see k heads in a row. Intuitively, that should provide strong evidence in favor of the coin being a double-headed coin. And, indeed, if we encode this evidence as a belief function following the methodology suggested in [Sha76a], we find that the larger k is, the stronger our belief that the coin is double-headed. On the other hand, we cannot compute the probability that the coin is double-headed if our only information is that we have seen k heads in a row. The actual probability depends on the prior. For example, if we knew that *a priori*, the probability of the coin being fair is .9999 and $k = 8$, then it is still quite probable that the coin is fair. Once we are given a prior probability on the coin being fair then, using conditional probability, we can compute the probability that the coin is fair given that we have observed k heads. If we use Shafer's method, then it can be shown that the conditional probability is exactly the result of using the rule of combination to combine the prior probability with the belief function that encodes our evidence (the fact that we have seen k heads). Thus, the belief function provides us a way of updating the probability function, that is, with a way of going from a prior probability to a posterior (conditional) probability.

Once we decide to view belief functions as representations of evidence, we must tackle the question of *how* to go about representing evidence using belief functions. A number of different representations have been suggested in the literature. We have already mentioned the one due to Shafer; still others have been suggested by Dempster and Smets [Dem68, Sha82]. Walley [Wal87] compares a number of representations of evidence in a general framework. We review his framework here, and present a slight strengthening of one of his results, showing that perhaps the best representation is given by a certain belief function that is also a probability function, in that it is the only representation satisfying certain reasonable properties that acts correctly under the combination of evidence.

Both of the viewpoints discussed here give us a way of understanding belief in terms of well-understood ideas of probability theory. (Indeed, it is one of the goals of this paper to explain as large a part as possible of the theory of belief functions in terms of probability theory, in the hope of getting a better understanding of belief functions.) However, as we show by example, these two viewpoints result in very different ways of modelling situations (although, if we do things right, we expect to reach the same conclusions no matter which viewpoint we take!). The major difference between the viewpoints is how they treat new evidence. If we view belief as a generalized probability, then it makes sense to *update* beliefs but not *combine* them. On the other hand, if we view beliefs as a representation of evidence, then it makes sense to combine them, but not update them. This suggests that the rule of combination is appropriate only when we view beliefs as representations of evidence. A way of updating beliefs, appropriate when we view beliefs as generalized probabilities, is described in [FH91a]. It seems that all the examples showing the counterintuitive nature of the rule of combination arise from an attempt to combine two beliefs that are really being viewed as generalized probabilities.

It is interesting to note that the claim that there is more than one interpretation of belief functions is not new. In fact, it goes back to the early work of Shafer. In commenting on Dempster's work in [Sha76b, p. 432], Shafer says "...instead of thinking of his lower probabilities as degrees of belief or degrees of support, [Dempster] preferred, at least originally, to think of his lower and upper probabilities as bounds for some true but somehow unknowable probabilities, thus retaining the identification of degrees of belief with additive probabilities." A few paragraphs later, Shafer continues "It is the new understanding of the meaning of Dempster's upper probabilities [essentially, as representations of evidence] that I offer as the primary contribution of this essay." Our results give a precise sense in which Dempster's interpretation is correct. If we view belief and plausibility as representing the inner and outer measures induced by some probability function Pr, they are indeed bounds for all the possible extensions of Pr (see Theorem 2.1 in the next section). The alternate way of viewing a belief function, namely, as a representation of evidence, can also be given a precise probabilistic interpretation.[1] Indeed, the distinction between the approaches essentially is closely

[1]We do not mean to suggest that Shafer would necessarily subscribe to our interpretation. In fact, he would almost certainly dispute the primacy given to probability theory in this paper, as well as some of our conclusions. Shafer is also quite explicit about rejecting the view of belief functions as lower envelopes (see [Sha90, p. 16] for perhaps the clearest statement of his views on this issue).

related to the well-known distinction in probability theory between *absolute beliefs* and *belief updates* (see [HH86] for discussion and further references). Viewing a belief function as a representation of evidence essentially amounts to viewing it as a *likelihood function*; we return to this point later in the text.

More recently, Smets, in a sequence of unpublished papers such as [SK89], has been a strong proponent of the fact that there are two views of belief. One view for him is what we are calling belief as generalized probability. He identifies the second view with what he calls the *transferable belief model* (TBM). Smets specifically rejects an interpretation of the TBM in terms of probability theory, and offers it as an alternative to probability theory. It is definitely not meant to be viewed as a representation of evidence; rather, it measures degree of belief. Smets attempts to justify Dempster's rule of combination in this framework by viewing it as a way of reassigning or transferring beliefs from one proposition to another in light of new evidence. It seems to us that this interpretation leads to the same counterintuitive results we have already mentioned.[2]

The rest of this paper is organized as follows. In Section 2 we review the viewpoint of belief as generalized probability; in Section 3 we consider how to best update beliefs given this viewpoint. The material in these two sections is largely drawn from [FH91b, FH91a], so is not discussed here in great detail. We include it here mainly to contrast this viewpoint with the viewpoint of belief as evidence, which is discussed in Section 4. In this section we also consider what is the best way of representing evidence as a belief function, and argue that a probability function gives the best representation. In Section 5 we consider what happens when we combine the two viewpoints, in that we try to view belief as evidence when our information as represented in terms of nonmeasurable sets. In Section 6 we illustrate our points by considering a number of examples, including a lottery example from [Hun87] and the puzzle of Mr. Jones' murderer, taken from [SK89]. We conclude in Section 7 with further discussion on the appropriateness of belief functions as a representation of uncertainty.

2 Belief as generalized probability

This section summarizes the work of [FH91b]; portions of the material in this section also appear in [FH91a].

We begin by reviewing basic definitions from probability theory. The presentation follows that of [FH91b]; the reader should consult a basic probability text such as [Fel57, Hal50] for more details.

A *probability space* $(S, \mathcal{X}, Pr)$ consists of a set S (called the *sample space*), a σ-algebra $\mathcal{X}$ of subsets of S (i.e., a set of subsets of S containing S and closed under complementation and countable union, but not necessarily consisting of all subsets of

[2]Smets, of course, disagrees. We refer the reader to his papers for more details.

S) whose elements are called *measurable sets*, and a *probability measure* $Pr: \mathcal{X} \rightarrow [0,1]$ satisfying the following properties (known as the *Kolmogorov* axioms for probability):

P1. $Pr(X) \geq 0$ for all $X \in \mathcal{X}$

P2. $Pr(S) = 1$

P3. $Pr(\cup_{i=1}^{\infty} X_i) = \Sigma_{i=1}^{\infty} Pr(X_i)$, if the X_i's are pairwise disjoint members of $\mathcal{X}$.

Property P3 is called *countable additivity*. Of course, the fact that $\mathcal{X}$ is closed under countable union guarantees that if each $X_i \in \mathcal{X}$, then so is $\cup_{i=1}^{\infty} X_i$. If we restrict to finite sample spaces, the we can replace countable additivity by *finite additivity*, namely, the property

P3′. $Pr(\cup_{i=1}^{n} X_i) = \Sigma_{i=1}^{n} Pr(X_i)$, if the X_i's are pairwise disjoint members of $\mathcal{X}$.

A subset $\mathcal{Y}$ of $\mathcal{X}$ is said to be a *basis* (of $\mathcal{X}$) if the members of $\mathcal{Y}$ are nonempty and disjoint, and if $\mathcal{X}$ consists precisely of countable unions of members of $\mathcal{Y}$. It is easy to see that if $\mathcal{X}$ is finite then it has a basis. Moreover, whenever $\mathcal{X}$ has a basis, it is unique: it consists precisely of the minimal elements of $\mathcal{X}$ (the nonempty sets in $\mathcal{X}$ none of whose proper nonempty subsets are in $\mathcal{X}$). Note that if $\mathcal{X}$ has a basis, once we know the probability of every set in the basis, we can compute the probability of every measurable set by using countable additivity.

In a probability space $(S, \mathcal{X}, Pr)$, the probability measure Pr is not necessarily defined on 2^S (the set of all subset of S), but only on $\mathcal{X}$. We can extend Pr to 2^S in two standard ways, by defining functions Pr_* and Pr^*, traditionally called the *inner measure* and *outer measure induced by* Pr [Hal50]. For an arbitrary subset $A \subseteq S$, we define

$$Pr_*(A) = \sup\{Pr(X) \mid X \subseteq A \text{ and } X \in \mathcal{X}\},$$
$$Pr^*(A) = \inf\{Pr(X) \mid X \supseteq A \text{ and } X \in \mathcal{X}\}.$$

If there are only finitely many measurable sets (in particular, if S is finite), then it is easy to see that the inner measure of A is the measure of the largest measurable set contained in A, while the outer measure of A is the measure of the smallest measurable set containing A.

It is easy to check that, for any set A, we have $Pr_*(A) \leq Pr^*(A)$; if A is measurable, then $Pr_*(A) = Pr^*(A) = Pr(A)$. The inner and outer measures of a set A can be viewed as our best estimate of the "true" measure of A, given our lack of knowledge. To make this precise, we say a probability space $(S, \mathcal{X}', Pr')$ is an *extension* of the probability space $(S, \mathcal{X}, Pr)$ if $\mathcal{X}' \supseteq \mathcal{X}$, and $Pr'(A) = Pr(A)$ for all $A \in \mathcal{X}$ (so that Pr and Pr' agree on $\mathcal{X}$, their common domain). The following result is well known (a proof can be found in [Rus87]).

Theorem 2.1: *If $(S, \mathcal{X}', Pr')$ is an extension of $(S, \mathcal{X}, Pr)$ and $A \in \mathcal{X}'$, then $Pr_*(A) \leq Pr'(A) \leq Pr^*(A)$. Moreover, there exist extensions $(S, \mathcal{X}_1, Pr_1)$, $(S, \mathcal{X}_2, Pr_2)$ of $(S, \mathcal{X}, Pr)$ such that $A \in \mathcal{X}_1$, $A \in \mathcal{X}_2$, $Pr_1(A) = Pr_*(A)$, and $Pr_2(A) = Pr^*(A)$.*

Intuitively, the first part of Theorem 2.1 tells us that if we acquire extra information enabling us to compute the probability of A, then it is bound to lie somewhere between the inner measure and outer measure of A. The second part of the theorem tells us that the inner measure and outer measure are the best estimates we can get.

Now let us consider belief functions. Like a probability function, a belief function is a function mapping subsets of a set S to the interval $[0,1]$ satisfying certain axioms. Unlike a probability function, it is defined on *all* subsets of S. Formally, a belief function Bel on S is a function $Bel : 2^S \rightarrow [0,1]$ satisfying:

B0. $Bel(\emptyset) = 0$

B1. $Bel(A) \geq 0$

B2. $Bel(\mathrm{S}) = 1$

B3. $Bel(A_1 \cup \ldots \cup A_k) \geq \sum_{I \subseteq \{1,\ldots,k\}, I \neq \emptyset} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i)$.

We can also define the plausibility of a set A, written $Pl(A)$, as $1 - Bel(\overline{A})$, where $\overline{A}$ is the complement of A. Clearly Pl is also a function that associates with each subset of S a number in the range $[0,1]$. Using B2 and B3, we can easily see that $1 = Bel(S) = Bel(A \cup \overline{A}) \geq Bel(A) + Bel(\overline{A})$, from which it immediately follows that $Bel(A) \leq 1 - Bel(\overline{A} = Pl(A)$. As we shall see, the interval defined by $Bel(A)$ and $Pl(A)$ can be viewed as defining the range in which the "true" probability of A lies. Of course, the bigger the interval, the greater our uncertainty of the true probability of A.

Other than B3, the axioms for belief functions look like what we would expect from a probability function. Properties B1 and B2 are analogues of P1 and P2. (There is also a probabilistic analogue P0 of B0, but the fact that the probability of the empty set is 0 already follows from P2 and P3.) While B3 looks quite different from P3, the differences are not as significant as they might appear at first. For one thing, probability functions satisfy B3 with the inequality replaced by an equality, at least if we restrict attention to measurable sets $A_1, \ldots, A_k$. (This is the well-known *inclusion-exclusion* rule, and can be proved for probability functions by induction on k; see [Fel57].) Moreover, if we replace P3 by B3 (with the inequality replaced by equality, and the sets $A_1, \ldots, A_k$ restricted to measurable sets), then we get an axiom equivalent to finite additivity. (This is easy to see: if the sets A_i in B3 are disjoint, from B3 we immediately get $Bel(A_1 \cup \ldots A_k) \geq \sum_{i=1}^{k} Bel(A_i)$.) Thus, we get another characterization of probability functions in finite spaces.[3]

It seems clear that in many ways Bel and Pl act like inner and outer measure. For one thing, the relationship between them is analogous: $Pl(A) = 1 - Bel(\overline{A})$ and $Pr^*(A) =$

[3]When considering belief functions on infinite spaces, another continuity axiom, which says that $lim_{i \rightarrow \infty} Bel(A_i) = Bel(\cap_i A_i)$ if $A_1 \supseteq A_2 \supseteq \ldots$, is occasionally added [Sha79]. This axiom is easily shown to be redundant in finite spaces. If we replace the inequality in B3 by an equality and restrict to measurable sets, then, together with the continuity axiom, we get an alternative characterization of probability functions in arbitrary spaces.

$1-Pr_*(A)$. Moreover, inner and outer measure, like belief and plausibility, are defined on all subsets of S. It is not hard to show that every inner measure induced by a probability function is indeed a belief function [FH91b]. That is, if $(S, \mathcal{X}, Pr)$ is a probability space, then Pr_* is a belief function on S and Pr^* is the corresponding plausibility function. The converse essentially holds as well; given a belief function Bel defined on a set of formulas (rather than on sets), we can find a probability space $(S, \mathcal{X}, Pr)$ and associate with each formula φ a subset S_φ of states of S (intuitively, S_φ is the subset of states where φ is true) such that $Bel(\varphi) = Pr_*(S_\varphi)$. These results are discussed and proved in [FH91b]. Thus, in a precise sense, a belief function is no more and no less than an inner measure; the plausibility function is the corresponding outer measure.

There is another formulation of belief functions that is perhaps more intuitive, and will be useful in our later discussion. A *mass function* is simply a function $m\colon 2^S \rightarrow [0,1]$ such that

M1. $m(\emptyset) = 0$

M2. $\sum_{A\subseteq S} m(A) = 1$.

Intuitively, $m(A)$ is the weight of evidence for A that has not already been assigned to some proper subset of A. With this interpretation of mass, we would expect that an agent's belief in A is the sum of the masses he has assigned to all the subsets of A; i.e., $Bel(A) = \sum_{B\subseteq A} m(B)$. Indeed, this intuition is correct.

Proposition 2.2: ([Sha76a, p. 39])

1. *If m is a mass function on S, then the function $Bel\colon 2^S \rightarrow [0,1]$ defined by $Bel(A) = \sum_{B\subseteq A} m(B)$ is a belief function.*
2. *If Bel is a belief function on 2^S and S is finite, then there is a unique mass function m on 2^S such that $Bel(A) = \sum_{B\subseteq A} m(B)$ for every subset A of S.*

Using mass functions, we can easily connect probability, belief, and inner measure in finite spaces (or, in fact, in a probability space with a basis). If Pr is a probability function defined on a set $\mathcal{X}$ of measurable subsets of a finite set S, and $\mathcal{Y}$ is a basis of $\mathcal{X}$, let m be the mass function such that

$$m(A) = \begin{cases} Pr(A) & \text{if } A \in \mathcal{Y} \\ 0 & \text{otherwise,} \end{cases}$$

and let Bel be the belief function corresponding to m. Then it is easy to show that $Bel(A) = Pr_*(A)$ for all $A \subseteq S$. Thus, Bel agrees with Pr on the measurable sets and, more generally, is equal to the inner measure on arbitrary subsets. We refer to Bel as the *belief function corresponding to Pr*. Notice that the mass function m has the property that its *focal elements*—those sets to which it assigns positive mass—are disjoint. It is easy to check that if we are given a belief function Bel' whose corresponding mass function m' has disjoint focal elements, then there is some probability function Pr' such that Bel' corresponds to Pr'. We say that a belief function is a *discrete probability function* if not only are its focal elements disjoint, but they are singletons. Thus, a belief function is a discrete probability function if it is a probability function with respect to which every element in the sample space is measurable. Notice that if we restrict attention to finite or countable sets (as we do in this paper), this means that every subset is measurable.

There is another way of looking at belief functions as generalized probabilities, closely associated with the one we have just discussed. Given a set $\mathcal{P}$ of probability functions all defined on a sample space S, define the *lower envelope* of $\mathcal{P}$ to be the function f such that for each $A \subseteq S$, we have $f(A) = \inf\{Pr(A) : Pr \in \mathcal{P}$ and A is measurable with respect to $Pr\}$. We have the corresponding definition of *upper envelope* of $\mathcal{P}$. Theorem 2.1 says that the inner measure induced by a probability function Pr is the lower envelope of the family of probability functions extending Pr; the outer measure is the corresponding upper envelope. Since a belief function is essentially an inner measure, this suggests that a belief function is also a lower envelope. This is true, and was already known to Dempster [Dem67]. Let Bel be a belief function defined on S, and let $(S, \mathcal{X}, Pr)$ be a probability space with sample space S. We say that Pr is *consistent with Bel* if $Bel(A) \leq Pr(A) \leq Pl(A)$ for each $A \in \mathcal{X}$. Intuitively, Pr is consistent with Bel if the probabilities assigned by Pr are consistent with the intervals $[Bel(A), Pl(A)]$ given by the belief function Bel. It is easy to see that Pr is consistent with Bel if $Bel(A) \leq Pr(A)$ for each $A \in \mathcal{X}$ (that is, it follows automatically that $Pr(A) \leq Pl(A)$ for each $A \in \mathcal{X}$). This is because $Pl(A) = 1 - Bel(\overline{A}) \geq 1 - Pr(\overline{A}) = Pr(A)$. Then Bel is the lower envelope of $\mathcal{P}$ and Pl is the upper envelope of $\mathcal{P}$.

Although every belief function is a lower envelope, the converse does not hold. It is well known that not every lower envelope is a belief function (see [Kyb87, Pea89] for counterexamples). For further discussion on lower envelopes and their relationship to belief functions, the reader is referred to [FH91a, Sha81, WF82, Wil78].

3 Updating probabilities and beliefs

Quite often we start with a probability distribution or a belief function defined on a set of events and then want to update it in the light of new evidence. Define a *probability update* function to be a partial function from probability functions to probability functions; intuitively, if τ is a probability update function and Pr is a probability function, then $\tau(Pr)$ is the probability function that arises as a result of updating Pr in the light of the new information encoded by τ. We can similarly define a *belief update* function to be a partial function from belief functions to belief functions.

The type of evidence we are most used to dealing with is an observation showing that an event B has occurred. The standard way to update a probability function Pr in this case is to move to the conditional probability function $Pr(\cdot|B)$, where $Pr(A|B)$ is defined to be $Pr(A \cap B)/Pr(B)$. The reason we consider *partial* functions can already be seen when we consider conditional probability functions. For the remainder of this section, fix a set S and a σ-algebra $\mathcal{X}$ of subsets of S. For $B \in \mathcal{X}$, we can define $cond_B$ to be the probability update function such that $cond_B(Pr) = Pr(\cdot|B)$ if Pr is a probability function on $\mathcal{X}$ with $Pr(B) > 0$, and undefined otherwise. The partiality of the update function allows it to be undefined if the evidence that it encodes is incompatible with the probability to be updated. For example, the fact that B has been observed, which is encoded in the update function $cond_B$, is incompatible with a probability function Pr such that $Pr(B) = 0$.

We can combine a sequence of probability updates by composition. Thus, the result of updating by τ_1, then τ_2, and then τ_3 is given by the update function $\tau_3 \circ \tau_2 \circ \tau_1$. Although the composition operation is associative, it is not in general commutative; the order of updating matters. However, if we update probability functions by conditioning, then the order is irrelevant. Although the following result is well known, we prove it again here both for the sake of completeness and because we know of no reference to it.

Proposition 3.1: *Let $B, C \in \mathcal{X}$. Then*

$$cond_C \circ cond_B = cond_{B \cap C} = cond_B \circ cond_C.$$

Proof: Fix a probability function Pr. First assume that $Pr(B \cap C) > 0$, and let $Pr' = Pr(\cdot|B)$. Then for all sets $A \in \mathcal{X}$, we have

$$\begin{aligned}
cond_C \circ cond_B(Pr)(A) &= cond_C(Pr')(A) \\
&= Pr'(A|C) \\
&= Pr'(A \cap C)/Pr'(C) \\
&= Pr(A \cap C|B)/Pr(C|B) \\
&= (Pr(A \cap C \cap B)/Pr(B))/(Pr(C \cap B)/Pr(B)) \\
&= Pr(A \cap C \cap B)/Pr(C \cap B) \\
&= Pr(A|B \cap C) \\
&= cond_{B \cap C}(Pr)(A).
\end{aligned}$$

Thus, $cond_C \circ cond_B(Pr) = cond_{B \cap C}(Pr)$ if $Pr(B \cap C) > 0$. If $Pr(B \cap C) = 0$, then $cond_{B \cap C}(Pr)$ is undefined; we must show that $cond_C \circ cond_B(Pr)$ is undefined. If $Pr(B) = 0$, then this is immediate. Otherwise, it is easy to check that $Pr(C|B) = 0$, so that $cond_C(cond_B(Pr))$ is undefined. This shows that $cond_C \circ cond_B = cond_{B \cap C}$ in general. A similar argument shows that $cond_B \circ cond_C = cond_{B \cap C}$, and hence that $cond_B \circ cond_C = cond_C \circ cond_B$. ∎

Notice that the conditional probability function $Pr(\cdot|B)$ is well defined only if B, the observation, is a measurable set. In [FH91a], this definition is extended to allow nonmeasurable sets, by providing a notion of inner and outer conditional probability. The definition is inspired by Theorem 2.1. Let $(S, \mathcal{X}, Pr)$ be a probability space. Define the *inner conditional probability* $Pr_*(A|B)$ and the *outer conditional probability* $Pr^*(A|B)$ of A given B as follows:

$$\begin{aligned} Pr_*(A|B) &= \inf\{Pr'(A|B) \,|\, (S, \mathcal{X}', Pr') \text{ extends } (S, \mathcal{X}, Pr) \text{ and } A, B \in \mathcal{X}'\} \\ Pr^*(A|B) &= \sup\{Pr'(A|B) \,|\, (S, \mathcal{X}', Pr') \text{ extends } (S, \mathcal{X}, Pr) \text{ and } A, B \in \mathcal{X}'\}. \end{aligned}$$

Since the infimum and supremum above are not well-defined unless $Pr_*(B) > 0$, we define $Pr_*(A|B)$ and $Pr^*(A|B)$ only if $Pr_*(B) > 0$.

The next theorem (from [FH91a]) gives elegant closed-form expressions for the inner and outer conditional probabilities. This formula appears also in [Wal81], [SK89] and [dCLM90]. Indeed, this formula even appears (lost in a welter of notation) as Equation 4.8 in [Dem67]!

Theorem 3.2: *For any probability function Pr on S and subsets $A, B \subseteq S$ such that $Pr_*(B) > 0$, we have*

$$Pr_*(A|B) = \frac{Pr_*(A \cap B)}{Pr_*(A \cap B) + Pr^*(\overline{A} \cap B)}$$
$$Pr^*(A|B) = \frac{Pr^*(A \cap B)}{Pr^*(A \cap B) + Pr_*(\overline{A} \cap B)}$$

As we discussed earlier, every belief function is a lower envelope. Let Bel be a belief function defined on S, and let $(S, \mathcal{X}, Pr)$ be a probability space with sample space S. Recall that Pr is *consistent with* Bel if $Bel(A) \leq Pr(A) \leq Pl(A)$ for each $A \in \mathcal{X}$. Let $\mathcal{P}_{Bel}$ be the set of all probability functions consistent with Bel, such that every subset of S is measurable. The next theorem tells us that the belief function Bel is the lower envelope of $\mathcal{P}_{Bel}$, and Pl is the upper envelope.

Theorem 3.3: *([FH91a]) Let Bel be a belief function on S. Then for all $A \subseteq S$, we have*

$$Bel(A) = \inf_{Pr \in \mathcal{P}_{Bel}} Pr(A)$$
$$Pl(A) = \sup_{Pr \in \mathcal{P}_{Bel}} Pr(A).$$

Theorem 3.3 suggest how we might update a belief function to a *conditional belief function*, and a plausibility function to a *conditional plausibility function*, by using the following definitions as given in [FH91a]:

$$Bel(A|B) = \inf_{Pr \in \mathcal{P}_{Bel}} Pr(A|B)$$
$$Pl(A|B) = \sup_{Pr \in \mathcal{P}_{Bel}} Pr(A|B).$$

It is not hard to see that the infimum and supremum above are not well-defined unless $Bel(B) > 0$; therefore, we define $Bel(A|B)$ and $Pl(A|B)$ only if $Bel(B) > 0$. It is straightforward to check that if Pr is a probability function, Bel is the belief function corresponding to Pr, and A and B are measurable sets with respect to Pr, then $Bel(A|B) = Pr(A|B)$. Thus, this definition of conditional belief generalizes that of conditional probability.

Because of the close analogy between the definitions of conditional inner measures and conditional belief functions, and the fact that inner measures and belief functions are essentially the same, we might suspect that a closed-form formula for the conditional belief function can be obtained by replacing inner measures in Theorem 3.2 by belief functions and outer measures by plausibility functions. The next theorem says that this is indeed the case.

Theorem 3.4: *([FH91a]) If Bel is a belief function on S such that $Bel(B) > 0$, we have*

$$Bel(A|B) = \frac{Bel(A \cap B)}{Bel(A \cap B) + Pl(\overline{A} \cap B)}$$
$$Pl(A|B) = \frac{Pl(A \cap B)}{Pl(A \cap B) + Bel(\overline{A} \cap B)}.$$

It is well known that the conditional probability function is a probability function. That is, if we start with a probability function Pr defined on a σ-algebra $\mathcal{X}$ of subsets of S and if $B \in \mathcal{X}$, then the function $Pr(\cdot|B)$ defined on $\mathcal{X}$ is a probability function. We might hope that the same situation holds with belief functions, so that the conditional belief and plausibility functions are indeed belief and plausibility functions. Given the definitions of conditional belief and plausibility as lower and upper envelopes, it is not clear that this should be so, since lower and upper envelopes of arbitrary sets of probability functions do not in general result in belief and plausibility functions. Fortunately, as the next result shows, in this case they do. Thus, we have a way of updating belief and plausibility functions to give us new belief and plausibility functions in the light of new information.

Theorem 3.5: *([FH91a]) Let Bel be a belief function defined on S, and Pl the corresponding plausibility function. Let $B \subseteq S$ be such that $Bel(B) > 0$. Then $Bel(\cdot|B)$ is a belief function, and $Pl(\cdot|B)$ is the corresponding plausibility function.*

Using these definitions, we can extend the updating function $cond_B$ so that it is defined on belief functions as well as probability functions, by taking $cond_B(Bel) = Bel(\cdot|B)$ if $Bel(B) > 0$, and undefined otherwise. Unfortunately, when we extend $cond_B$ to belief functions, Proposition 3.1 no longer holds. (See [FH91a] for further discussion of this point.)

Dempster [Dem67] defines another notion of conditional belief. He defines

$$Bel(A||B) = \frac{Bel(A \cup \overline{B}) - Bel(\overline{B})}{1 - Bel(\overline{B})}.$$

$Bel(\cdot||B)$ is indeed a belief function, and the corresponding plausibility function satisfies

$$Pl(A||B) = \frac{Pl(A \cap B)}{Pl(B)}.$$

For the remainder of this paper, we call this the *DS notion of conditioning.*

As shown in [FH91a], there is a sense in which the two notions of conditioning that we have been considering both correspond to conditional probability. Suppose that we have a probability space $(S, \mathcal{X}, Pr)$ with basis $\mathcal{Y}$, and let Bel be the belief function corresponding to Pr. Then we can consider two processes. In the first process, an agent chooses a set $X \in \mathcal{Y}$ with probability $Pr(X)$ and then chooses an element $x \in X$. We are not given the probability with which a particular $x \in X$ is chosen. Thus, given $A \subseteq S$, we cannot compute a precise probability that $x \in A$ is chosen; $Pr_*(A)$ and $Pr^*(A)$ give us the best possible lower and upper bounds. Similarly, if we fix a set $B \subseteq S$, then we cannot compute a precise probability that $x \in A$ is chosen given that x is in B. In this case, the best possible lower and upper bounds are given by $Bel(A|B)$ and $Pl(A|B)$ (i.e., $Pr_*(A|B)$ and $Pr^*(A|B)$). In the second process, we slightly change the rules so that when choosing an element $x \in X$, the agent chooses x in B whenever possible. There is a difference between the two processes only if both $X \cap B \neq \emptyset$ and $X \cap \overline{B} \neq \emptyset$ for some basis set $X \in \mathcal{Y}$. Since X is an element of a basis, this in turn can happen only if B is a nonmeasurable set (since every measurable set is the union of basis sets). In this case, the agent definitely chooses an element in $X \cap B$ (although again, we don't know the probability that a particular element will be chosen). We can then ask for the probability that an element in A will be chosen by the second process, given that an element in B is chosen. It can be shown that the bounds are provided by $Bel(A||B)$ and $Pl(A||B)$. (See [FH91a] for more details.) These observations show that the DS conditioning notion corresponds to a somewhat unusual updating process, where before we condition on B, we try to choose an element in B if possible.

Although the focus here has been on updates that arise from a conditioning process, there are clearly other ways of updating beliefs and probabilities. In general, when we make an observation, we do not observe that B is the case. More likely, the best we can say is that our observation leads us to believe that B occurred with some probability. Methods such as *Jeffrey's rule* [Jef83] have been proposed for updating probability functions given such an observation. The details are beyond the scope of this paper. The key point is that they again lead to an *update* function, which maps one probability function to another, and that they can be extended to provide an update function on beliefs in an appropriate way. (See [FH91a] for further discussion of this point.)

4 Belief as evidence

Up to now we have viewed belief as a generalized probability. This does *not* seem to be the view of belief that Shafer espouses in [Sha76a]. He talks of belief as being a representation of a body of *evidence*. To say that $Bel(A) = p$ is to say that, as a result of the evidence encoded by Bel, the agent has a degree of belief p in the proposition represented by the set A.

From this point of view, it makes sense to *combine* two belief functions Bel_1 and Bel_2. The resulting belief function Bel is meant to represent the combined evidence encoded by each of Bel_1 and Bel_2 separately. On the other hand, it is not clear what it should mean to combine two probability functions. The theory of probability provides no straightforward answer to the problem of how to combine two probability functions. For example, if one person examines a coin and says that it is fair (so that the probability of heads is 1/2), while another says that it is slightly biased and the probability of heads is .4, there seems to be no obvious way to combine these two probability distributions. Intuitively, one ought to put more weight on the person that is judged to be more reliable, but this a question of subjective judgment, not of mathematics. (The subject of combining probability distributions has inspired a great deal of research; we refer the reader to [GZ86] for an overview.)

Roughly speaking, it seems that *updating* makes sense for (generalized) probability, while *combining* makes sense for evidence.

In order to combine two or more independent[4] pieces of evidence, Shafer suggests the use of Dempster's *rule of combination*. For the remainder of this section, let us restrict attention to belief functions defined only on finite sets S. With this restriction, the rule of combination can be easily described as follows.[5]

If m_1 and m_2 are mass functions with the same domain 2^S, let $m_1 \oplus m_2$ be the mass function m where $m(A) = c \sum_{\{B_1, B_2 \,|\, B_1 \cap B_2 = A\}} m_1(B_1) m_2(B_2)$ for each nonempty $A \subseteq S$, and where c is a normalizing constant chosen so that the sum of all of the $m(A)$'s is 1. It is easy to check that $c = (\sum_{\{B_1, B_2 \,|\, B_1 \cap B_2 \neq \emptyset\}} m_1(B_1) m_2(B_2))^{-1}$. Note that if there is no pair B_1, B_2 where $B_1 \cap B_2 \neq \emptyset$ and $m_1(B_1) m_2(B_2) > 0$, then we cannot find such a normalizing constant c. In this case $m_1 \oplus m_2$ is undefined. If $m_1 \oplus m_2$ is defined, then the corresponding belief functions Bel_1 and Bel_2 are said to be *combinable*. If Bel_1 and Bel_2 are combinable belief functions with mass functions m_1 and m_2 respectively, then the belief function that is the result of combining Bel_1 and Bel_2, denoted $Bel_1 \oplus Bel_2$, is the belief function with mass function $m_1 \oplus m_2$ ($Bel_1 \oplus Bel_2$ is undefined if Bel_1 and Bel_2 are not combinable).

Shafer presents many examples of the intuitively appealing nature of the rule of combination in [Sha76a]. He also shows that in some sense we can use the rule of combination to capture the idea of updating a belief function as the result of learning new evidence. The effect of learning B can be captured by the belief function $Learn^B$ corresponding to the mass function m which puts all the mass on B; i.e., $m(B) = 1$ and

[4]For now, like Shafer, we take *independence* to be an intuitive, primitive notion. The probabilistic definition of independence—namely, that A and B are independent if $Pr(A \cap B) = Pr(A) \times Pr(B)$—is a consequence of our intuitive notion, but does not seem to us to completely capture it.

[5]These definitions can all be extended to the case where S is infinite. We restrict to finite S here for ease of exposition and because it is the case most often considered in the literature.

$m(A) = 0$ if $A \neq B$. Thus, we have

$$Learn^B(A) = \begin{cases} 1 & \text{if } A \supseteq B \\ 0 & \text{otherwise.} \end{cases}$$

It is this idea of learning that is used to define the DS notion of conditional belief. In fact, it is easy to check that $Bel(\cdot||B) = Bel \oplus Learn^B$; i.e., $Bel(\cdot||B)$ is the result of combining Bel with the belief function that corresponds to learning B.

While this definition seems very natural, the reader should recall our earlier discussion, which showed that the DS notion of conditioning corresponds to a somewhat unusual updating process. If we view Bel as a representation of evidence, then a case can be made that $Bel(\cdot||B)$ represents that body of evidence that results from combining the evidence encoded by Bel with the evidence that B is actually the case. On the other hand, if we view Bel as a generalized probability distribution, we can no longer expect that the rule of combination should correspond to a natural updating process. In fact, as was shown above, it does not. The key point here is that updating and combining are different processes; what makes sense in one context does not necessarily make sense in the other.

The discussion above suggests that, whatever evidence is, evidence and probability are different. They are related though. A probability function gets updated as a result of evidence. This suggests that one way we can represent evidence is as an update function. For the remainder of this paper, we consider this particular representation of evidence, as a function that maps probability functions to probability functions. While we believe this is the first paper that has explicitly suggested the representation of evidence as an update function, the idea is implicit in many other papers. For example, the likelihood function is often viewed as a way of representing evidence, and as an update function (see, for example, [HH86, Hec86]). The key point for us is that, as we shall see, belief functions can be viewed as representations of evidence, i.e., , as update functions. The idea is that given a belief function Bel and a prior probability Pr, we transform this to a posterior probability Pr' by using the rule of combination. That is, we can consider the mapping $Pr \mapsto Pr' = Pr \oplus Bel$. *A priori*, it is not clear that this mapping does anything interesting. Clearly, for this mapping to have the "right" properties, we need to consider how to represent evidence as a belief function.

4.1 Representing evidence

In most of the examples given in [Sha76a], subjective degrees of belief are assigned to various events in the light of evidence. Although Shafer shows that the degrees of belief seem to have reasonable qualitative behavior when the evidence is combined, there is no external notion of "reasonable" against which we can evaluate how reasonable these numbers are. The one place where there is an external of reasonableness comes in the area that Shafer terms *statistical evidence*. In this case, we have numbers provided by certain conditional probabilities. A prototypical example of this type of situation is given by the following coin-tossing situation.

Imagine a coin is chosen from a collection of coins, each of which is either biased towards heads or biased towards tails. The coins biased towards heads land heads with probability 2/3 and tails with probability 1/3, while those biased towards tails land tails with probability 2/3 and heads with probability 1/3. We start tossing the coin in order to determine its bias. We observe that the first k tosses result in heads. Intuitively, the more heads we see without seeing a tail, the more evidence we have that the coin is in fact biased towards heads. How should we represent this evidence in terms of belief functions?

Suppose that we have a space $S = \{BH, BT\}$, where BH stands for *biased towards heads*, and BT stands for *biased towards tails*. Let Bel_{heads} be the belief function on S that captures the evidence in favor of BH and BT as a result of seeing the coin land heads. We would certainly expect that $Bel_{heads}(BH) > Bel_{heads}(BT)$,[6] since seeing the coin lands heads provides more evidence in favor of the coin being biased towards heads than it does in favor of the coin being biased towards tails. But what numeric values should we assign to $Bel_{heads}(BH)$ and $Bel_{heads}(BT)$? According to a convention introduced by Shafer [Sha76a, Chap. 11] (which we discuss in more detail below), we should take $Bel_{heads}(BH) = 1/2$ and $Bel_{heads}(BT) = 0$. Thus, if m_{heads} is the corresponding mass function, we take $m_{heads}(BH) = 1/2$, $m_{heads}(S) = 1/2$, and $m_{heads}(BT) = 0$. By symmetry, the belief function Bel_{tails} representing the evidence of the coin landing tails satisfies $Bel_{tails}(BH) = 0$ and $Bel_{tails}(BT) = 1/2$.

If we assume that our observations are independent, then it seems reasonable to expect that the belief function which represents the observation of k heads should correspond in some sense to combining the evidence of observing one head k times. Let $m^k_{heads} = m_{heads} \oplus \cdots \oplus m_{heads}$ (k times); a straightforward computation shows that $m^k_{heads}(BT) = 0$, $m^k_{heads}(BH) = (2^k - 1)/2^k$, $m^k_{heads}(S) = 1/2^k$. Thus, we also have $Bel^k_{heads}(BT) = 0$ and $Bel^k_{heads}(BH) = (2^k - 1)/2^k$. This seems qualitatively reasonable. If we see k heads in a row, then it is much more likely that the coin is biased towards heads than that it is biased towards tails. It is also easy to compute that

$$(m_{heads} \oplus m_{tails})(BH) = (m_{heads} \oplus m_{tails})(BT) = 1/3.$$

Thus,

$$(Bel_{heads} \oplus Bel_{tails})(BH) = (Bel_{heads} \oplus Bel_{tails})(BT) = 1/3.$$

Again, it seems reasonable that if we see heads followed by tails, we should have no more evidence in favor of the coin being biased towards heads than it being biased towards heads (although the particular choice of 1/3 as the appropriate amount of evidence may seem somewhat mysterious).

[6] For readability, we write $Bel_{heads}(BH)$ for $Bel_{heads}(\{BH\})$, and similarly throughout the paper when singleton sets are arguments.

What do these numbers tell us about the probability that the coin is biased towards heads or biased towards tails? Without knowing something about how the coin is chosen, probability theory does not give us much guidance. For example, if the coin was chosen at random from a collection of $1,000,000$ coins only one of was biased towards heads and all the rest biased towards tails, then even after seeing 10 heads in a row, we would still say that it is extremely likely that the coin is biased towards tails.

Now suppose that we knew that the coin was chosen at random from a collection with proportion α of coins biased towards heads and $1-\alpha$ of coins biased towards tails. By definition, $Pr(BH|k \text{ heads}) = Pr(BH \wedge k \text{ heads})/Pr(k \text{ heads})$.[7] Now the probability that the coin is biased towards heads and the first k coin tosses are heads is $2^k\alpha/3^k$, while the probability that the coin is biased towards tails and the first k tosses are heads is $(1-\alpha)/3^k$. The probability of getting k heads is thus $(1+(2^k-1)\alpha)/3^k$, hence the conditional probability of the coin being biased towards heads given that k heads are observed is $2^k\alpha/(1+(2^k-1)\alpha)$. As we would expect, this probability approaches 1 as k gets larger.

Let m be the mass function that describes the initial probability; thus $m(BH)=\alpha$ and $m(BT)=1-\alpha$. If we define $m_1 = m \oplus m_{heads}$ and $m_k = m \oplus m^k_{heads}$, then a straightforward computation shows $m_1(BH) = 2\alpha/(1+\alpha)$ and $m_1(BT) = (1-\alpha)/(1+\alpha)$, while $m_k(BH) = 2^k\alpha/(1+(2^k-1)\alpha)$ and $m_k(BT) = (1-\alpha)/(1+(2^k-1)\alpha)$. The upshot of this calculation is that $Bel_1 = Pr(\cdot|heads)$ and $Bel_k = Pr(\cdot|k\ heads)$. Thus, by combining the prior with the belief function that represents the evidence, we get the posterior. The same phenomenon occurs if we combine the prior with Bel_{tails}.

Judging by this example, Shafer's definition of Bel_{heads} and Bel_{tails} has two very interesting properties. At the risk of being repetitive, we summarize them again:

- when we combine Bel_{heads} with a prior on $S=\{BH, BT\}$, we get the conditional (posterior) probability on S given that heads is observed.
- Bel^k_{heads} in some sense represents the evidence encoded observing k heads, and $Bel_{heads} \oplus Bel_{tails}$ represents the evidence encoded by observing heads and then tails, in that if we combine these belief functions with the prior, we get the appropriate conditional probability.

Obviously, we now want to know whether these properties hold not just for certain observations made in this coin-tossing example, but in general. The answer is yes, and the appropriate theorems that show this can already be found in [Sha76a]. We review and extend this material here.

[7] Strictly speaking, by using the $\wedge$ symbol, we are confounding propositions and sets. We continue to be a bit sloppy in our usage when discussing this and later examples, in the hope that the reader will not have any trouble following what is meant.

4.2 A general framework

We want to consider the question of representing statistical evidence in a general framework. Suppose that we have a set $\mathcal{H}$ consisting of *basic hypotheses* $H_1, \ldots, H_m$, and another set $\mathcal{O}$ consisting of *basic observations* $Ob_1, \ldots, Ob_n$. Intuitively, we are considering a situation (which is standard in statistical testing) where exactly one of these hypotheses holds, and we are testing which one it is. The basic observations are the data given to us by our tests. In our example above, the basic hypotheses are BH and BT, while the basic observations are *heads* and *tails*. Although there are often difficulties in deciding precisely what hypotheses one should test and what the observations are (indeed, this is one of the fundamental problems in statistics), the precise choice of basic hypotheses and basic observations is clear in many applications of interest. In any case, our goal here is to understand what are appropriate ways to represent evidence. The hope is that by analyzing this relatively simple situation, we can gain insight into more complicated situations.

We assume that for each basic hypothesis H_i, we have a probability Pr_i on $\mathcal{O}$. More formally, we have a probability space $(\mathcal{O}, 2^{\mathcal{O}}, Pr_i)$ (the set of measurable sets being $2^{\mathcal{O}}$ tells us that every subset of $\mathcal{O}$ is measurable). Intuitively, $Pr_i(Ob)$ is the probability of observing Ob given that the hypothesis H_i holds. The reason we write $Pr_i(Ob)$ rather than something like $Pr(Ob|H_i)$ is that in writing the latter expression, we implicitly assume that we have a probability function Pr on the space $\mathcal{H} \times \mathcal{O}$; this is an assumption we do not want to make at this point (although we do make it later). We shall be mainly interested in $Pr_i(Ob_j)$ for a basic observation Ob_j. Of course, once we know $Pr_i(Ob_j)$ for each basic set, we can easily extend Pr_i by additivity to all of $2^{\mathcal{O}}$. In the example above, we have $Pr_{BH}(heads) = 2/3$, $Pr_{BT}(heads) = 1/3$, and so on. The probability of seeing heads given that the coin is biased towards heads is 2/3, while the probability of seeing heads given that the coin is biased towards tails is 1/3.

We want to compute a belief function that represents the result of making a basic observation $Ob \in \mathcal{O}$, using these probabilities. The general approach to doing this goes back to the statistician R. A. Fisher, who called the expression $Pr_i(Ob)$ a *likelihood*, and viewed it as the likelihood that the hypothesis H_i was true, given the observation Ob. The hypothesis that is taken as most likely to be true is the one whose likelihood is the greatest, given the observation Ob. We would expect that observing Ob would provide more support to H_i than H_j if $Pr_i(Ob) > Pr_j(Ob)$. (See [Hac65] for further discussion of likelihoods.) Shafer's convention provides a particular way of capturing this intuition. According to Shafer's convention, the evidence Ob should be represented by the belief function Bel_{Ob} such that for each subset $A \subseteq S$, we have

$$Bel_{Ob}(A) = 1 - [\max_{H_j \in \overline{A}} Pr_j(Ob) / \max_{j=1,\ldots,m} Pr_j(Ob)].$$

In our example, the observation is *heads*. Since $Pr_{BH}(heads) = 2/3$ and $Pr_{BT}(heads) = 1/3$, it is easy to see that we have $Bel_{heads}(BH) = 1/2$ and $Bel_{heads}(BT) = 0$, just as we assumed above.

One important consequence of the general definition is that $Pl_{Ob}(H_i) = 1 - Bel_{Ob}(\overline{H_i}) = Pr_i(Ob)/c$, where $c = \max_{j=1,\ldots,m} Pr_j(Ob)$. Thus, the plausibility of the basic hypothesis H_i is proportional to the likelihood $Pr_i(Ob)$. As we now show, this property of Bel_{Ob} is enough to guarantee that it acts correctly as an update function.

Fix the functions $Pr_1, \ldots, Pr_m$. In order to show that Bel_{Ob} acts correctly as an update function, we need to show that, when combined with a prior on $\mathcal{H}$, we get the conditional probability given Ob. Thus, suppose that we have a prior probability Pr on $\mathcal{H} \times \mathcal{O}$. Since we can identify subsets of $\mathcal{H}$ and $\mathcal{O}$ with subsets of $\mathcal{H} \times \mathcal{O}$ in the obvious way (for example, we can identify $Ob \subseteq \mathcal{O}$ with the subset $\mathcal{H} \times Ob = \{(H_i, Ob_j) \mid H_i \in \mathcal{H}, Ob_j \in Ob\}$), Pr actually can be viewed as giving us a probability function on both $\mathcal{H}$ and $\mathcal{O}$: we simply identify $Pr(H_i)$ with $Pr(H_i \times \mathcal{O})$ and $Pr(Ob_j)$ with $Pr(\mathcal{H} \times Ob_j)$. In particular, this lets us view Pr as giving us a prior on $\mathcal{H}$. Moreover, we can make sense out of the conditional probability $Pr(\cdot|Ob)$; this will be important in our later discussion.

We do not want to consider arbitrary probability functions on $\mathcal{H} \times \mathcal{O}$. We want to consider only those probability functions which are consistent with the information already provided to us by the probability functions $Pr_1, \ldots, Pr_m$. Recall that $Pr_i(Ob)$ intuitively represents the conditional probability of seeing Ob given that hypothesis H_i is true. Thus, we say that Pr is *consistent with* $Pr_1, \ldots, Pr_m$ if $Pr(H_i) > 0$ implies $Pr(Ob_j|H_i) = Pr_i(Ob_j)$, for $i = 1, \ldots, m$. This means that Pr is consistent with $Pr_1, \ldots, Pr_m$ exactly if Pr_i is the probability on $\mathcal{O}$ obtained by conditioning Pr with respect to H_i. Note that for any probability function Pr' on $\mathcal{H}$, there is a unique probability function Pr on $\mathcal{H} \times \mathcal{O}$ consistent with $Pr_1, \ldots, Pr_m$ such that $Pr(H_i) = Pr'(H_i)$. We obtain Pr by simply defining $Pr(H_i \times Ob_j) = Pr'(H_i)Pr_i(Ob_j)$ and extending by additivity. For each probability Pr on $\mathcal{H} \times \mathcal{O}$, we denote the restriction of Pr to $\mathcal{H}$ by $Pr|_{\mathcal{H}}$, where of course $Pr|_{\mathcal{H}}(H) = Pr(H \times \mathcal{O})$.

Intuitively, a belief function Bel provides an appropriate representation of the evidence in the observation Ob if, by combining it with $Pr|_{\mathcal{H}}$, we get the conditional probability function $Pr(\cdot|Ob)$. Formally, we say that a belief function *Bel captures the evidence of the observation Ob* if for every probability function Pr on $\mathcal{H} \times \mathcal{O}$ consistent with $Pr_1, \ldots, Pr_m$, we have $Pr(H_i|Ob) = (Pr|_{\mathcal{H}} \oplus Bel)(H_i)$, $i = 1, \ldots, m$, provided that $Pr(Ob) > 0$. This definition is meant to capture the intuition we started with: Bel captures the evidence of Ob if, whenever we combine it with a prior, we get the conditional probability given Ob.

In the coin-tossing example above, we showed that the belief function Bel_{heads} that arises from the observation *heads* using Shafer's representation did capture the evidence of *heads*. We want to prove that Shafer's representation has this property in general. The following result follows from Theorem 9.7 in [Sha76a].

Theorem 4.1: *Let Bel be a belief function on $\mathcal{H}$, and Pl be the corresponding plausibility function. Bel captures the evidence of Ob iff $Pl(H_i) = cPr_i(Ob)$ for some constant $c > 0$.*

Theorem 4.1 essentially says that all that matters about a belief function when assessing whether it captures evidence appropriately is the relative plausibility of the basic hypotheses; these plausibilities must be in the same ratio as the likelihood of these hypotheses given the observation Ob. *Any* belief function which assigns the appropriate relative plausibilities to basic hypotheses will do. We have already observed that in Shafer's representation, the relative plausibility of hypotheses is in the right ratio. Thus, we immediately get

Corollary 4.2: *Bel_{Ob} captures the evidence of Ob.*

Now what happens when we combine observations? If we make k observations, this results in the observation set $\mathcal{O}^k$, consisting of k-tuples of elements of $\mathcal{O}$. Suppose that we have a sequence $(Ob^1, \ldots, Ob^k)$ of observations in $\mathcal{O}^k$, and that the belief function Bel_j captures the evidence of Ob^j, for $j = 1, \ldots, k$.[8] Further suppose that these observations are independent. This means that for each basic hypothesis H_i, the probability of observing a particular sequence of observations given H_i is the product of the probabilities of making each observation in the sequence. More formally, we assume that we have a probability Pr_i^k on $\mathcal{O}^k$, for $i = 1, \ldots, m$. Then the observations $Ob^1, \ldots, Ob^k$ are independent (with respect to Pr_i^k) if $Pr_i^k((Ob^1, \ldots, Ob^k)) = Pr_i(Ob^1) \times \cdots \times Pr_i(Ob^k)$.

Intuitively, since the belief function $Bel_1 \oplus \cdots \oplus Bel_k$ is intended to represent the combination of the evidence represented by making each observation individually, we might hope that the evidence of the sequence $(Ob^1, \ldots, Ob^k)$ of observations is captured by $Bel_1 \oplus \cdots \oplus Bel_k$. This is a property that held for Shafer's representation in our example. The following result, which follows from Theorem 9.8 in [Sha76a], shows that it holds in general.

Theorem 4.3: *Suppose Ob^j, for $j = 1, \ldots, k$, are independent observations and Bel_j captures the evidence of Ob^j. Then $Bel_1 \oplus \cdots \oplus Bel_k$ captures the evidence of $(Ob^1, \ldots, Ob^k)$.*

Again, we want to stress that Theorems 4.1 and 4.3 show that not only does Shafer's representation give a belief function that satisfies our criteria for appropriately capturing an observation Ob, but so would any other belief function for which the plausibilities of the basic hypotheses are in the same ratio as the likelihoods of the basic hypotheses given Ob. Another such representation is suggested in [Dem68] (see [Sha76b] for a comparison between his approach and that of Dempster). Yet another is given by Smets (see [Sha82] for a presentation and discussion of Smets' approach). We consider a fourth choice (also considered in [Sha82]), which we shall shortly argue is perhaps the most natural of all; namely, to consider the unique belief function that captures the evidence of Ob that is a (discrete) *probability function*. To emphasize the fact that it is a probability function, we call it Pr_{Ob}. By Theorem 4.1, we must take $Pr_{Ob}(H_i) = cPr_i(Ob)$, where c is a normalizing constant chosen so that $\sum_{i=1}^m Pr_{Ob}(H_i) = 1$. The following proposition is immediate from Theorem 4.1:

[8]We are using superscripts rather than subscripts so that these observations will not be confused with the basic observations $Ob_1, \ldots, Ob_n$.

Proposition 4.4: *Pr_{Ob} captures the evidence of Ob.*

The representation Pr_{Ob} is quite easy to work with. For example, in the coin-tossing example, we have $Pr_{BH}(heads) = 2/3$ and $Pr_{BT}(heads) = 1/3$. Since the ratio of these probabilities is $2 : 1$, the belief/probability function Pr_{heads} that is intended to represent the evidence of seeing heads must give mass to the hypotheses BH and BT in the ratio $2 : 1$. Thus we must have $Pr_{heads}(BH) = 2/3$, $Pr_{heads}(BT) = 1/3$. Similarly, we have $Pr_{tails}(BH) = 1/3$, $Pr_{tails}(BT) = 2/3$. Although Pr_{heads} and Pr_{tails} can be viewed as probability functions on $\mathcal{H}$, they should not be thought of as representing the probability of BH or BT in any sense corresponding to the frequentist or subjectivist interpretation of probability. Rather, these are encodings of the evidence for BH and BT given the observations *heads* and *tails* respectively. It is easy to check that, for example, we have $Pr^k_{heads}(BH) = 2^k/(2^k + 1)$ and $Pr^k_{heads}(BT) = 1/(2^k + 1)$, where $Pr^k_{heads} = Pr_{heads} \oplus \cdots \oplus Pr_{heads}$ (k times). Again, the more heads we see, the greater the evidence that the coin is biased towards heads. And if we combine this with a prior Pr such that $Pr(BH) = \alpha$, then an easy computation shows that $(Pr|_{\mathcal{H}} \oplus Pr^k_{heads})(BH) = 2^k\alpha/(1 + (2^k - 1)\alpha)$. This is the conditional probability $Pr(BH|k\ heads)$, which is just what we expect from Theorem 4.1.

Before we compare Pr_{Ob} to Bel_{Ob}, we briefly consider Shafer's motivation in choosing Bel_{Ob}. It turns out that Bel_{Ob} is the unique *consonant* belief function among the belief functions that capture the evidence of Ob, where a consonant belief function is one for which the focal elements are nested, i.e., we have that if $m_{Ob}(A) > 0$ and $m_{Ob}(B) > 0$, then either $A \subseteq B$ or $B \subseteq A$. Shafer discusses consonance in [Sha76a, Chap. 10]. He does present arguments that consonance is a reasonable assumption to consider in some cases (see also [Sha76b]); it would take us too far afield to discuss them here. Further arguments for Shafer's representation are given in [KM90] and [Was88]. Nevertheless, it seems to us that the case for this representation is not a strong one. Indeed, as we now show, there is one rather nonintuitive consequence of using Shafer's consonant belief function in this context.

4.3 Representing the combination of evidence

Suppose we make k independent observations $Ob^1, \ldots, Ob^k$. It seems that this should be equivalent to making the one joint observation $(Ob^1, \ldots, Ob^k)$. Although we showed above that $Bel_{Ob^1} \oplus \ldots \oplus Bel_{Ob^k}$ appropriately captures the evidence of $(Ob^1, \ldots, Ob^k)$, we might hope for something stronger, namely that $Bel_{Ob^1} \oplus \ldots \oplus Bel_{Ob^k} = Bel_{(Ob^1,\ldots,Ob^k)}$. This just says that the belief function that represents the joint observation is equal to the combination of the belief functions representing the individual observations. Unfortunately, as Shafer already observed [Sha76a, p. 249–250], this is *not* the case in general. Returning to our coin-tossing example, recall that $(Bel_{heads} \oplus Bel_{tails})(BH) = 1/3$. Suppose we now compute $Bel_{(heads,tails)}(BH)$. Since $Pr^2_{BH}((heads, tails)) = Pr^2_{BT}((heads, tails)) = 2/9$ (where $Pr^2_{BH} = Pr_{BH} \oplus Pr_{BH}$), it follows from Shafer's definitions that $Bel_{(heads,tails)}(BH) =$

0. Thus $Bel_{heads} \oplus Bel_{tails} \neq Bel_{(heads,tails)}$.[9]

The fact that Shafer's approach to representing evidence does not represent a joint observation in the same way that it represents the combination of the individual observations has disturbed a number of authors [Dia78, Sei81, Wil78]. In fact, in [Sha82], Shafer indicates that he is inclined to agree that this property is unacceptable. We now focus on this problem in more detail.

First observe that the problem does not arise if we use the probabilistic representation of evidence. For example, it is easy to check that $(Pr_{heads} \oplus Pr_{tails})(BH) = Pr_{(heads,tails)}(BH) = 1/2$. Intuitively, the two observations of heads and tails cancel each other out, so BH and BT are given the same relative weight as a result of these observations. The fact that this example works out right is not an accident.

Proposition 4.5: *If $Ob^1, \ldots, Ob^k$ are independent observations, then $Pr_{Ob^1} \oplus \ldots Pr_{Ob^k} = Pr_{(Ob^1,\ldots,Ob^k)}$.*

Proof: By definition, Pr_{Ob^j} is the discrete probability function on $\{H_1, \ldots, H_m\}$ where the probability of H_i is proportional to $Pr_i(Ob^j)$. So $Pr_{Ob^1} \oplus \ldots Pr_{Ob^k}$ is the discrete probability function on $\{H_1, \ldots, H_m\}$ where the probability of H_i is proportional to $Pr_i(Ob^1) \cdots Pr_i(Ob^k)$.

By definition of independence, the probability of observing the sequence $(Ob^1, \ldots, Ob^k)$, given the hypothesis H_i, is equal to the product $Pr_i(Ob^1) \cdots Pr_i(Ob^k)$. So $Pr_{(Ob^1,\ldots,Ob^k)}$ is the discrete probability function on $\{H_1, \ldots, H_m\}$ where the probability of H_i is proportional to $Pr_i(Ob^1) \cdots Pr_i(Ob^k)$. Together with the result of the previous paragraph, this proves the proposition. ∎

Proposition 4.5 shows that the probabilistic representation of evidence acts correctly under combination. Although the example above showed that Shafer's representation does not act correctly under combination, there might perhaps be other representations besides the probabilistic representation that act correctly under combination in the sense of Proposition 4.5. In the remainder of this section, we show that this is not the case. Under some reasonable assumptions, the representation of evidence using a discrete probability function is the *only* representation of evidence that acts correctly under combination in the sense of Proposition 4.5.

In order to make these ideas precise, we need to define carefully the phrase "representation of evidence". We use here the general framework defined by Walley [Wal87].[10]

[9] We mentioned earlier that, while the fact that $(Bel_{heads} \oplus Bel_{tails})(BH) = (Bel_{heads} \oplus Bel_{tails})(BT)$ seemed reasonable, the fact that $(Bel_{heads} \oplus Bel_{tails})(BH)$ should be 1/3 was a bit mysterious. The observations above suggest that not only is 1/3 mysterious, it is inappropriate. Of course, as far as getting the "right" answer when combined with a prior, all that matters is that we have equality. A similar phenomenon arises with Dempster's representation of evidence; indeed, this is precisely the core of Aitchison's criticism of this representation [Ait68].

[10] We actually developed our ideas independently of Walley; we thank Larry Wasserman for pointing out Walley's work to us.

Assume, as above, that we have a set $\mathcal{H}$ of hypotheses, with basic hypotheses $H_1, \ldots, H_m$, and a set $\mathcal{O}$ of observations, with basic observations $Ob_1, \ldots, Ob_n$. Suppose that corresponding to each basic hypothesis H_i, we have a probability Pr_i on $\mathcal{O}$. Let $BEL(\mathcal{H})$ be the set of all the belief functions on $\mathcal{H}$. We now make an observation Ob. We take a representation of evidence to be a general technique to associate with the observation Ob a belief function in $BEL(\mathcal{H})$ which captures the evidence of Ob. If $Pr_i(Ob) = a_i$, and if Pl is the plausibility function corresponding to the belief function representing Ob, then, by Theorem 4.1, we know that $Pl(H_i) = ca_i$, for some constant c. The relative plausibilities of the basic hypotheses must be in the right ratio.

Since the only information we are given regarding Ob are the likelihoods $Pr_i(Ob)$, $i = 1, \ldots, m$, we would expect the belief function which represents Ob to depend only on these likelihoods. This is consistent with what has been called the *likelihood principle* [Hac65]: only likelihoods count in assessing the evidence contained in an observation. We remark that of the representation methods mentioned above, the probabilistic representation, Shafer's representation, and Smets' representation all satisfy this assumption; however, Dempster's representation does not. In Dempster's representation, the belief assigned to hypothesis H as a result of making observation Ob might depend on the probabilities $Pr_j(Ob')$ assigned to an observation Ob' other than Ob.

To capture formally our assumption that all that matters are the likelihoods $Pr_i(Ob)$, we take a *representation of evidence on* $\mathcal{H}$ to be a function $f : ([0,1]^m - \{(0,\ldots,0)\}) \rightarrow BEL(\mathcal{H})$. (The reason we do not allow $(a_1, \ldots, a_m) = (0, \ldots, 0)$ is that if all the likelihoods are 0, then we do not have any information about the relative plausibilities we should assign to the basic hypotheses.) We refer to the belief function $f(a_1, \ldots, a_m)$ as $Bel_{(a_1,\ldots,a_m)}$. Intuitively, if we fix an observation Ob and if $Pr_i(Ob) = a_i$, $i = 1, \ldots, m$, then under the representation of evidence f, the belief function $Bel_{(a_1,\ldots,a_m)}$ is the one that represents the evidence encoded in Ob. In particular, this formalizes the assumption that the belief function representing Ob depends only on the likelihoods $Pr_i(Ob)$, $i = 1, \ldots, m$. Let $Pl_{(a_1,\ldots,a_n)}$ be the corresponding plausibility function. As we noted above, we require that $Pl_{(a_1,\ldots,a_n)}(H_i) = ca_i$, for some constant c.

Shafer's convention gives a representation of evidence on $\mathcal{H}$. Using our current notation, Shafer's representation gives

$$Bel_{(a_1,\ldots,a_m)}(A) = 1 - [\max_{H_i \in \overline{A}} a_i / \max_{i=1,\ldots,m} a_i].$$

Our probabilistic representation of evidence gives us

$$Bel_{(a_1,\ldots,a_m)}(A) = \sum_{H_i \in A} a_i / \sum_{i=1}^{m} a_i.$$

Note that if we take $(a_1, \ldots, a_m) = (0, \ldots, 0)$ in either Shafer's representation or the probabilistic representation, the resulting belief function is not well defined.

Walley [Wal87] considers various assumptions that a representation of evidence might satisfy, and shows that under quite weak assumptions, a representation of evidence results in a belief function that acts essentially like a probability function. We focus here on one assumption (also considered by Walley), that is easily seen to be satisfied by both the probabilistic representation and Shafer's representation, and seems to us very natural. It is a stronger version of the likelihood principle, namely, that all that counts are *relative likelihoods*. While this assumption is not a necessary one, it is consistent both with our use of *relative* plausibilities above (for example we observed in Theorem 4.1 that for a belief function to correctly represent an observation Ob, all that matters is that the ratio of the plausibilities of the basic sets be the same as the ratio of the likelihood functions $Pr_i(Ob)$). It is also consistent with the use of likelihoods typically made in the literature, where what is considered is the *likelihood ratio* (the ratio of the likelihood of Ob given an hypothesis H to the likelihood of Ob given $\neg H$). Here too, the intuition is that the absolute likelihood should not matter, but only the relative likelihood. Although this assumption seems quite natural, we remark that it is not satisfied by Smets' representation.

We encapsulate these ideas in the following definition. An *appropriate representation of evidence on* $\mathcal{H}$ is a function $f : ([0,1]^m - \{(0,\ldots,0)\}) \rightarrow BEL(\mathcal{H})$. We refer to $f(a_1,\ldots,a_m)$ as $Bel_{(a_1,\ldots,a_m)}$, and require that it satisfy the following properties:

R1. $Pl_{(a_1,\ldots,a_m)}(H_j) = ca_j$ for some constant $c > 0$ and for $j = 1,\ldots,m$,

R2. $Bel_{(a_1,\ldots,a_m)} = Bel_{(da_1,\ldots,da_m)}$, for all d with $0 < d < 1$.[11]

Now we need one last definition. Let f be a representation of evidence. We say that f *acts correctly under combination* if for all $(a_1,\ldots,a_m),(b_1,\ldots,b_m) \in [0,1]^m - \{(0,\ldots,0)\})$, we have

R3. $Bel_{(a_1b_1,\ldots,a_mb_m)} = Bel_{(a_1,\ldots,a_m)} \oplus Bel_{(b_1,\ldots,b_m)}$.

To understand why this definition captures our intuition that a representation acts correctly under combination, suppose that we make two independent observations, say Ob^1 and Ob^2. Further suppose that $Pr_j(Ob^1) = a_j$ and $Pr_j(Ob^2) = b_j$, $j = 1,\ldots,m$, so that $Bel_{(a_1,\ldots,a_m)}$ represents the observation Ob^1 and $Bel_{(b_1,\ldots,b_m)}$ represents the observation Ob^2. If Ob^1 and Ob^2 are independent, then $Pr_j((Ob^1,Ob^2)) = ca_jb_j$, for $j = 1,\ldots,m$ and for some appropriate normalizing constant c. Thus we expect $Bel_{(a_1b_1,\ldots,a_mb_m)}$ to represent the joint observation.

As we have observed, Shafer's representation does *not* act correctly under combination, while the probabilistic representation does. As the following theorem shows, the probabilistic representation is the only appropriate representation which acts correctly under combination. This result also essentially appears in [Wal87] (see the discussion on p. 1449). We include a proof here both because our proof is more direct than Walley's, and because we do not require a few weak regularity conditions that he imposes.

[11] By taking $0 < d < 1$, we guarantee that $Bel_{(da_1,\ldots,da_m)}$ is well defined. If we consider some $d > 1$, then it is possible that $da_i > 1$ for some i. Note that if $d > 1$ and $da_1 \leq 1$ for $i = 1,\ldots,m$, then it already follows that $Bel_{(da_1,\ldots,da_m)} = Bel_{(a_1,\ldots,a_m)}$: we can simply multiply by $1/d$, since $0 < 1/d < 1$.

Theorem 4.6: *The probabilistic representation of evidence is the only appropriate representation of evidence which acts correctly under combination.*

Proof: Let f be an appropriate representation of evidence which acts correctly under combination. As before, denote $f(a_1,\ldots,a_m)$ by $Bel_{(a_1,\ldots,a_m)}$. By assumption, f satisfies properties R1, R2, and R3.

By R3, it follows that $Bel_{(a_1,\ldots,a_m)} = Bel_{(a_1,\ldots,a_m)} \oplus Bel_{(1,\ldots,1)}$. Thus, $Bel_{(1,\ldots,1)}$ acts as the identity. Let us denote $Bel_{(1,\ldots,1)}$ by Bel_{Id}, with mass function m_{Id}. Since Bel_{Id} is the identity, we know that $m_{Id} \oplus m_{Id} = m_{Id}$. Our next goal is to prove that Bel_{Id} is a discrete probability function, with $Bel_{Id}(H_i) = 1/m$, for $i = 1,\ldots,m$. In particular, this means that we must show that $m_{Id}(H_i) = 1/m$, for $i = 1,\ldots,m$, and $m_{Id}(A) = 0$ if A is not a singleton. There are three main steps in showing this. First, we show that there are no nested focal elements of m_{Id} (recall that the focal elements of m_{Id} are those sets A where $m_{Id}(A) > 0$); that is, there are no focal elements A, B with $A \subset B$. Second, we show that no focal elements overlap; that is, there are no focal elements A, B with $A \cap B \neq \emptyset$. Third, we show the focal elements are singleton sets.

Suppose first that there are nested focal elements $A \subset B$ of m_{Id}. Let us take B as large as possible so that this is true, that is, such that there is no focal element C with $B \subset C$. Assume that $m_{Id}(A) = a > 0$ and $m_{Id}(B) = b > 0$. By maximality of B, it follows from the definition of $\oplus$ that $(m_{Id} \oplus m_{Id})(B) = c(m_{Id}(B)m_{Id}(B)) = cb^2$, where c is a normalization constant. Furthermore, $(m_{Id} \oplus m_{Id})(A) \geq c(m_{Id}(A)m_{Id}(A) + m_{Id}(A)m_{Id}(B)) \geq c(a^2 + ab)$. Since $m_{Id} \oplus m_{Id} = m_{Id}$, it follows that

$$\frac{a}{b} = \frac{m_{Id}(A)}{m_{Id}(B)} \geq \frac{a^2 + ab}{b^2}.$$

Since $b > 0$, we can multiply both the left- and right-hand sides by b^2, and simplify to obtain $0 \geq a^2$. But this is impossible, since a is strictly positive. Thus, there are no nested focal elements of m_{Id}.

We now show that no focal elements of m_{Id} overlap. Suppose that A and B are focal elements where $A \cap B \neq \emptyset$. Let $C = A \cap B$. Since C is a proper subset of either A or B, it follows from the fact that there can be no nested focal elements of m_{Id} that $m_{Id}(C) = 0$. However, $m_{Id}(C) = (m_{Id} \oplus m_{Id})(C) \geq m_{Id}(A)m_{Id}(B) > 0$, a contradiction. Thus, no focal elements of m_{Id} overlap.

We now show that every focal elements of m_{Id} is a singleton set. Let $X_1,\ldots,X_r$ be the focal elements. By the arguments above, the X_i's must be pairwise disjoint subsets of $\{H_1,\ldots,H_m\}$. Assume that some X_i is not a singleton. Without loss of generality, we can assume that H_1 and H_2 are elements of X_1. Consider the two belief functions $Bel' = Bel_{(1,\frac{1}{2},\frac{1}{2},\ldots,\frac{1}{2})}$ and $Bel'' = Bel_{(\frac{1}{2},1,1,\ldots,1)}$, with corresponding mass functions m' and m'', respectively. By R3, it follows that $Bel' \oplus Bel'' = Bel_{(\frac{1}{2},\ldots,\frac{1}{2})}$, which by R2 equals $Bel_{(1,\ldots,1)} = Bel_{Id}$. By R1, we know that $Pl'(H_1) > Pl'(H_2)$. It follows easily that there must be some focal element A of m' that contains H_1 but not H_2. Further, by R1, we know that $Pl''(H_1) > 0$, so there must be some focal element B of m'' that contains H_1. Let $C = A \cap B$. Then $m_{Id}(C) = (m' \oplus m'')(C) \geq m'(A)m''(B) > 0$. But this is a contradiction, since C contains H_1 but not H_2, and the only focal element of m_{Id} that contains H_1 is X_1, which also contains H_2.

We have shown that the only focal elements of Bel_{Id} are singleton sets; thus, Bel_{Id} is in fact a discrete probability function. Since Bel_{Id} is a discrete probability function, $Bel_{Id} = Pl_{Id}$. By property R1, it follows easily that $Bel_{Id}(H_i) = 1/m$, for $i = 1, \ldots, m$. It follows by definition of $\oplus$ that if Bel is an arbitrary belief function, then $Bel \oplus Bel_{Id}$ has the property that each focal element is a singleton set $\{H_i\}$. This means that $Bel \oplus Bel_{Id}$ is a discrete probability function. In particular, this is true when Bel is $Bel_{(a_1,\ldots,a_m)}$. But $Bel_{(a_1,\ldots,a_m)} \oplus Bel_{Id} = Bel_{(a_1,\ldots,a_m)}$; thus, $Bel_{(a_1,\ldots,a_m)}$ is a discrete probability function. Again, this means that $Bel_{(a_1,\ldots,a_m)} = Pl_{(a_1,\ldots,a_m)}$. By R1, we then find that $Bel_{(a_1,\ldots,a_m)}(H_j)$ is proportional to a_j, for $j = 1, \ldots, m$. So f is our probabilistic representation of evidence, as desired. ∎

If we accept that it is important that a representation of belief act correctly under combination, where does this result leave us? If in addition we accept the strong likelihood principle (that all that matters are relative likelihoods), then we are forced to use the probabilistic representation. We can either give up the strong likelihood principle (while still accepting the likelihood principle, that all that matters are the likelihoods of the observation we have made). In this case, there are non-probabilistic representations of evidence that can be used, such as Smets'. The consequences of giving up the strong likelihood principle are not yet clear; there may be computational advantages to assuming strong likelihood. This issue requires further investigation. Finally, if we were willing to give up the likelihood principle altogether, we might consider using Dempster's representation. We remark, that in [Sha82], Shafer considers a very special case of Dempster's representation, and shows that there is a sense in which it combines correctly. However, in this special case, Shafer uses a different technique than Dempster's rule of combination to compute the belief function that represents the joint observation. In fact, if Dempster's representation is applied in the most straightforward way to the coin-tossing example discussed above, we can show that it too does not act correctly under combination.

We conclude this section by briefly comparing the approach to encoding evidence by using a probability function as described above to the more standard probabilistic approach using the *likelihood ratio*, where the likelihood ratio $L(H_i, Ob)$ of H_i given the observation Ob is defined to be $Pr(Ob|H_i)/Pr(Ob|\neg H_i)$ Clearly, $L(\cdot, Ob)$ has some of the same spirit as Pr_{Ob}. Indeed, the computations of the conditional probability using the rule of combination and Pr_{Ob} very much resemble standard computations using Bayes' rule. However, it is not hard to see that $L(H_i, Ob)$ cannot be computed directly from $Pr_{Ob}(H_i)$ nor can $Pr_{Ob}(H_i)$ be computed directly from $L(H_i, Ob)$. It has been shown [Goo60, Hec86] that any "reasonable" notion of strength of evidence must be a function of the likelihood ratio, where a notion of strength of evidence is taken to be "reasonable" if it satisfies a number of requirements. (It would take us too far afield to discuss these requirements here, but we remark that they are similar in spirit to Cox's requirements for a "reasonable" notion of belief, and the proof has the same spirit as Cox's proof that any reasonable notion of belief must essentially be a probability function [Cox46].) Since Pr_{Ob} is not a function of the likelihood ratio, it must fail to satisfy one of the requirements of "reasonableness" given by Good and Heckerman. It turns out that the one it fails is that the result of the updating the prior probability on H_i by the observation Ob should depend only on $Pr_{Ob}(H_i)$ and $Pr(H_i)$. This is almost, but not quite, the case for Pr_{Ob}. The problem is that in order to compute the right normalization constant for $(Pr(\cdot|Ob) \oplus Pr_{Ob})(H_i)$, we need to know $Pr(H_j)$ and $Pr_{Ob}(H_j)$ for $j = 1, \ldots, m$. It is not enough to know the values just for $j = i$. However, it should be clear that the normalization constant does not play a crucial role here. We can still compute the *relative* conditional probabilities of for H_i and H_j just knowing their priors, $Pr(H_i)$ and $Pr(H_j)$, and $Pr_{Ob}(H_i)$ and $Pr_{Ob}(H_j)$.

5 Evidence and envelopes

Up to now we have assumed that for each basic hypothesis H_j we have a probability function Pr_j on $\mathcal{O}$ such that the basic observations Ob are *measurable* with respect to Pr_j. This implies that there is no uncertainty in $Pr_j(Ob)$; it is given by a single number, rather than by an interval. We have also assumed that each basic hypothesis H_j is measurable with respect to the prior Pr. This implies that there is no uncertainty in $Pr(H_j)$. In this section, we consider what happens if there is some uncertainty. As before, we model this uncertainty by assuming that there is some family of possible probability functions, rather than a single probability function.

We focus first on the case where, although we have a probability Pr_j on $\mathcal{O}$ for each basic hypothesis H_j, there is some uncertainty in $Pr_j(Ob)$, for some $j = 1, \ldots, m$. For example, if we modify our earlier example with the coin, suppose that instead of knowing that

1. if the coin is biased towards heads, then its probability of landing heads is 2/3, and

2. if the coin is biased towards tails, then its probability of landing tails is 2/3,

all we know is that

1. if the coin is biased towards heads, then its probability of landing heads is somewhere in the interval $[2/3, 1]$, and

2. if the coin is biased towards tails, then its probability of landing tails is somewhere in the interval $[2/3, 1]$.

That is, we consider the set $\mathcal{P}$ of all the probability functions Pr on $\{BH, BT\} \times \{heads, tails\}$ such that $Pr(heads|BH) = \beta_1$, for some β_1 with $2/3 \leq \beta_1 \leq 1$, and $Pr(heads|BT) = \beta_2$, for some β_2 with $0 \leq \beta_2 \leq 1/3$. Thus, $\mathcal{P}$ consists of all probability functions consistent with the information that we are given. Now suppose that we observe heads. Given the way we have modified the example, there no longer some definite probability of *heads*, so the representation techniques discussed in the previous section do not immediately apply.

Nevertheless, we might hope that we could capture this evidence by a belief function that, when combined with the prior probability function that gives the probabilities of the coin being biased towards heads or tails, gives some interval of possible posterior probabilities of the coin being biased towards heads or tails. We would further expect that this interval of possibilities would be the same as the interval of possibilities obtained by taking all possible conditional probabilities. Unfortunately, this hope cannot be attained.

To understand why, suppose that we represent the observation of heads by some belief function Bel, and that there is a prior probability Pr on the coin being biased towards heads. It is easy to see that if we start with a probability function, and combine it with any belief function (on the same space), then we get a probability function. In particular, $Pr|_{\mathcal{H}} \oplus Bel$ is a probability function. Thus, we do not get an interval $[a, b]$ of values with $a < b$, no matter what our choice of Bel.

However, there is another way that we could obtain an interval of values. Let $\mathcal{P}$ be as above. Then $\mathcal{P}$ consists of all probability functions consistent with the information that we are given. Let $\mathcal{P}_{heads}$ consist of all the conditional probability functions $Pr(\cdot|heads)$ for $Pr \in \mathcal{P}$. After the observation *heads*, we would like the belief and plausibility of BH and BT to be defined by the lower and upper envelopes of $\mathcal{P}_{heads}$. *A priori*, it is not clear that the lower and upper envelope really define belief and plausibility functions, but as we now show, they do.

Assume that $Pr \in \mathcal{P}$, that the prior probability $Pr(BH)$ that the coin is biased towards heads is α, and that $Pr(heads|BH) = \beta_1$ and $Pr(heads|BT) = \beta_2$. An easy computation using Bayes' rule shows that $Pr(BH|heads) = \alpha\beta_1/(\alpha\beta_1 + (1-\alpha)\beta_2)$, and $Pr(BT|heads) = (1-\alpha)\beta_2/(\alpha\beta_1 + (1-\alpha)\beta_2)$. Minimizing over all possible choices of β_1 and β_2, with $2/3 \leq \beta_1 \leq 1$ and $0 \leq \beta_2 \leq 1/3$, we get the function Bel such that $Bel(BH) = 2\alpha/(1+\alpha)$ and $Bel(BT) = 0$. It is easy to see that, as our notation suggests, Bel defines a belief function.

What does this tell us in terms of representation of evidence? For the purposes of this discussion, we use the probabilistic representation here, but everything we say works perfectly well for Shafer's representation as well. For fixed β_1, β_2, let $Pr_{heads,\beta_1,\beta_2}$ be the probabilistic representation of the observation of seeing heads, given that $Pr_{BH}(heads) = \beta_1$ and $Pr_{BT}(heads) = \beta_2$. An easy computation shows that we have $Pr_{heads,\beta_1,\beta_2}(BH) = \beta_1/(\beta_1+\beta_2)$ and $Pr_{heads,\beta_1,\beta_2}(BT) = \beta_2/(\beta_1+\beta_2)$. Now suppose that we are given a prior Pr on $\{BH, BT\} \times \{heads, tails\}$ such that $Pr(heads) = \alpha$. It follows from the results of the previous section that $(Pr|_{\mathcal{H}} \oplus Pr_{heads,\beta_1,\beta_2})(BH) = \alpha\beta_1/(\alpha\beta_1 + (1-\alpha)\beta_2)$, since the right-hand side is precisely $Pr(BH|heads)$, given that $Pr(heads|BH) = \beta_1$ and $Pr(heads|BT) = \beta_2$. Similarly, we get $(Pr|_{\mathcal{H}} \oplus Pr_{heads,\beta_1,\beta_2})(BT) = (1-\alpha)\beta_2/(\alpha\beta_1 + (1-\alpha)\beta_2)$. Minimizing over all possible choices of β_1 and β_2, with $2/3 \leq \beta_1 \leq 1$ and $0 \leq \beta_2 \leq 1/3$, we get our belief function Bel that we showed is the lower envelope of $\mathcal{P}_{heads}$. The upper envelope is the corresponding plausibility function.

To summarize, instead of obtaining our belief function Bel by combining the prior with the infimum of the representations $Pr_{heads,\beta_1,\beta_2}$, we instead obtain Bel by taking the infimum of the results of combining the prior with the the representation $Pr_{heads,\beta_1,\beta_2}$.

A similar situation arises if we assume that Ob is measurable with respect to Pr_i for $i = 1, \ldots, m$, but that $H_1, \ldots, H_m$ are not necessarily measurable with respect to the prior probability. Intuitively, this means that there is some uncertainty about prior probabilities of the hypotheses. Going back to our example, supposed we know that the coin has probability either 2/3 or 1/3 of landing heads, as in the original formulation of the problem, but rather than being given a precise prior α on the coin being biased towards heads, all we are given interval of possibilities. For example, suppose that all we know is that the prior probability α lies in the interval $[0, 1/2]$, and we observe heads. Again, it turns out that combining this prior with an encoding of evidence in the most straightforward way gives inappropriate results. A probability function extending the prior Pr could give BH probability anywhere between 0 and 1/2. Thus, the answer we would hope to get when we combine the prior Pr with an observation of heads is a belief function Bel such that $Bel(BH) = 0$ and $Pl(BH) = 2/3$, since this is the range defined by lower and upper envelope of the family of probability functions extending Pr.

Unfortunately, if we combine the belief function corresponding to this prior with the probabilistic representation of the evidence Pr_{heads} (recall that we have $Pr_{heads}(BH) = 2/3$ and $Pr_{heads}(BT) = 1/3$), then we get a belief function Bel such that $Bel(BH) = Pl(BH) = 1/2$ and $Bel(BT) = Pl(BT) = 1/2$. This certainly does not seem like the right answer! If instead we combine the belief function corresponding to the prior with Shafer's representation Bel_{heads} (recall $Bel_{heads}(BH) = 1/2$ and $Bel_{heads}(BT) = 0$), then we get a belief function Bel' such that $Bel'(BH) = Bel'(BT) = 1/3$, while $Pl'(BH) = Pl'(BT) = 2/3$. Although this at least allows the probability of BH to be somewhere between 1/3 and 2/3, it is still not quite the answer we want.

Just as in the previous case, we can get the lower and upper envelopes we are looking for by minimizing and maximizing the results of using the combination rule.

Notice that in the examples that we considered, the lower envelopes that gave what we felt was the appropriate answer was in fact a belief function. We have a counterexample which shows that this is not the case in general. However, we conjecture that under reasonable assumptions—namely, if our uncertainty about the prior or the conditional probability can be expressed by a belief function (i.e., if the lower envelope of the family of probability functions that describe the prior or the conditional probability is a belief function)—the lower envelope of the resulting family of conditional probability functions will also be a belief function. We remark that although we have treated separately the case where there is some uncertainty in the probability of the observation Ob, given some hypothesis H_j, and the case where there is there is some uncertainty about prior probabilities of the hypotheses H_j, we could, of course, combine these two situations. The results would be similar to what we have already seen.

6 Examples

Depending on which of the two views of belief functions we take, we will model a situation in very different ways. For example, it is typically assumed that lack of information about an event E should be modelled by the vacuous belief function Bel, so that $Bel(E) = 0$ and $Pl(E) = 1$. While this way of modelling the lack of information is consistent with the view of belief as a generalized probability (intuitively, our information is consistent with E having any probability between 0 and 1), it is *not* in general consistent with the view of belief as evidence. To take a simple example, suppose we have two fair coins, call them coin A and coin B. Someone tosses one of the two coins and announces that it lands heads. Intuitively, we now have no evidence to favor the coin being either coin A or coin B. Taking the view of belief as generalized probability, we would have $Bel(A \text{ tossed}) = 0$ and $Pl(A \text{ tossed}) = 1$. However, taking the view of belief as evidence and using the probabilistic representation of belief, we get $Bel(A \text{ tossed}) = Pl(A \text{ tossed}) = 1/2$. Lack of information is not being represented by the vacuous belief function under this viewpoint.

In general, starting with a (belief function representing a) prior, if we get new evidence, we can either update the prior, or combine it with a belief function representing the evidence. As we already saw in the coin-tossing example of Section 4, we get the same answer no matter how we do it (although the intermediate computations are quite different), providing we represent the evidence appropriately. The one thing we must be careful not to do is to represent the evidence as a generalized probability, and then combine it with the prior.

We now consider a few other examples from the literature from this point of view, showing how understanding the differences between the two viewpoints helps clarify the issues involved. We start with a slightly simplified version of a puzzle appearing in [Hun87].

Suppose that we have 100 agents, all holding a lottery ticket, numbered 00 to 99. Suppose that agent a_1 holds ticket number 17. Assume that the lottery is fair, so, *a priori*, the probability that a given agent will win is 1/100. We are then told that the first digit of the winning ticket is 1. Straightforward probability arguments show that the probability that the winning ticket is 17 given that the first digit of the winning ticket is 1 is 1/10; thus, agent 1's probability of winning in light of the new information is 1/10.

How can we represent this information using belief functions? Hunter essentially considers two belief functions on the space $S = \{a_1, \ldots, a_{100}\}$, where $Bel(a_i)$ represents the belief that a_i wins. It seems reasonable to represent the information that the lottery is fair by the belief function Bel_1 corresponding to the mass function m_1 such that $m_1(\{a_i\}) = 1/100$, $i = 1, \ldots, 100$. Now how should we represent the second piece of information? Hunter suggests representing it by the belief function Bel_2 corresponding to the mass function m_2 such that $m_2(a_1) = 1/10$ and $m_2(\{a_2, \ldots, a_{100}\}) = 9/10$. Since our belief that a_1 will win given this information is precisely $1/10$, we give the set $\{a_1\}$ mass $1/10$; since we have no further information regarding any other agent, the remaining mass is assigned to $\{a_2, \ldots, a_{100}\}$. This representation is best understood as a generalized probability: our information is consistent with the set $\mathcal{P}$ of probability functions on S that assign $\{a_1\}$ probability $1/10$. It is easy to see that Bel_2 is the lower envelope of this family of probability functions. (We consider a more refined view of Bel_2 as a lower envelope below.)

Hunter then considers the result of combining these two belief functions by using the rule of combination. In light of our previous discussion, it should not be surprising that the result does not seem to represent the combined evidence at all. In fact, an easy computation shows that the result of combining Bel_1 and Bel_2 is a probability function that places probability $1/892$ on a_1 winning, and probability $9/892$ on a_i winning, for $i = 2, \ldots, 100$. It certainly does not seem appropriate that the evidence that a_1's probability of winning is $1/10$, when combined with the information that the lottery is fair, should decrease our belief that a_1 will win and, in fact, result is a belief that any other agent is 9 times as likely to win as a_1!

There are two objections to this use of the rule of combination. The first is that, at least the way we have told the story, the fact that our belief probability that a_1 wins is $1/10$ given that the first digit of his number agrees with the winning number is not independent of our belief that the lottery is fair. In fact, it is a direct consequence of our belief that the lottery is fair. There would be no reason to assign probability $1/10$ to a_1 winning upon hearing that the first digit of his number is the same as the first digit of the winning number in the absence of an assumption of fairness. (For example, if we believed that the lottery was fixed and that 19 was bound to be the winning number, hearing that a_1's first digit agreed with the winning number would not cause us to change our belief that a_1 was sure to lose.)

This objection, while correct, does not seem to get to the heart of the problem. Consider the following (admittedly artificial) situation: Again, we assume that the lottery is fair, but now we hear from an insider that the winning number was drawn and that a_1 was the winner. Moreover, suppose from previous experience we know that this insider is not terribly truthful. In fact, he tells the truth precisely $1/10$ of the time. This information certainly seems independent of the fact that the lottery is fair. If we represent it using Bel_2, we still get the same counterintuitive answer: a piece of information that seems like it should increase our belief that a_1 is the winner in fact decreases it significantly.

As the discussion in the previous section suggests, the real problem here is that we are trying to use the rule of combination with a belief function that is meant to represent a generalized probability. The point is that Bel_2 does not represent the evidence appropriately. In order to apply the techniques discussed in the previous section for representing the evidence, we need to know the likelihood, for each agent a_i, that the first digit of the winning number is 1, given the hypothesis that a_i wins the lottery. In the case of a_1, it is easy to compute this probability: since a_1's number is 17, the probability that the first digit of the winning number is 1 given that a_1 wins is 1. In the case of the other agents, we cannot compute this probability at all, since we do not know what their lottery numbers are.

In order to deal with this problem, first consider a fixed assignment A of lottery numbers to agents, so that $A(i)$ is the lottery number of a_i. We assume (as is the case in the story) that $A(1) = 17$. With respect to this fixed assignment, it is easy to see that there are 10 agents a_i for which the probability that the first digit of the winning number is 1 given that a_i wins is 1; namely, all those agents a_i such that the first digit of $A(i)$ is 1. For every other agent a_i, the probability that the first digit of the winning number is 1 given that a_i wins is 0. Using the probabilistic representation of evidence discussed in the previous section,[12] we would thus represent the evidence that the first digit of the winning lottery number is 1 by the belief function Bel_2^A such that the mass function $m_2^A(\{a_i\}) = 1/10$ for each agent a_i such that the first digit of $A(i)$ is 1 (note that, in particular, this includes a_1), and $m_2^A(\{a_i\}) = 0$ if the first digit of $A(i)$ is not 1. It is now easy to check that $Bel_1 \oplus Bel_2^A = Bel_2^A$. Thus, independent of the assignment A of lottery numbers, we have $(Bel_1 \oplus Bel_2^A)(\{a_1\}) = 1/10$, as expected.

Notice that Bel_2^A is actually a probability function, for each choice of A. Moreover, if we take the lower envelope of the family Bel_2^A over all choices of assignment A, we get Hunter's belief function Bel_2. Thus, in this weak sense, we can say that Bel_2 represents the evidence that the first digit of the winning number is 1. However, as we have observed, combining Bel_2 with Bel_1 is not equivalent to combining each Bel_2^A separately with Bel_1. Moreover, there is information lost if we consider Bel_2 rather than the family of functions Bel_2^A, namely, that the mass is distributed evenly among precisely 10 of the agents (one of which is a_1). Although this information is contained in the family of functions Bel_2^A, it is not contained in Bel_2.

This example also points out the subtle interplay between nonprobabilistic choices (the choice of assignment of lottery numbers in this case), and probabilistic (random) choices (choosing a winner of the lottery). This issue turns out to be closely related to issues of reasoning about knowledge. It is beyond the scope of this paper to examine these issues in more detail; the interested reader is referred to [HT93] for further discussion.[13]

[12]We would get essentially the same results using the other representations discussed in the previous section.

[13]We remark that in the version of the puzzle presented by Hunter, we are not given a_1's lottery number. In order to deal with that situation, we extend the analysis above by considering pairs (A, w), where A is an assignment of lottery numbers to agents and w is the winning number, with the added constraint that the first digit of $A(1)$ is the same as the first digit of w. None of the essential details in the discussion above change.

We next consider the Puzzle of Mr. Jones' Murderer, taken from [SK89]:[14]

> Big Boss has decided that Mr. Jones must be murdered, and the murderer will be one of Peter, Paul, and Mary. Big Boss will select the sex of the killer according to the results of a coin toss: if the coin lands heads, then the killer will be a female; if the coin lands tails, then the killer will be a male. Although we know how the killer is to be chosen, we do not know the result of the coin toss, nor do we know how Big Boss would have decided between Peter and Paul if the coin had landed tails. However, we do know that if Peter is not chosen, then he will go to the police station in order to give himself an alibi.[15] The murder is committed. We also learn that Peter was indeed at the police station during the time the murder is known to have been committed. What is the probability that the killer was Paul?

Let *Peter*, *Paul*, and *Mary* be the hypotheses that Peter, Paul, and Mary, respectively, committed the murder. Let Pr be the prior probability of these hypotheses. We are told $Pr(\{Peter, Paul\}) = Pr(Mary) = 1/2$. We would like to compute the probability $Pr(Paul|\neg Peter)$. Equivalently, we must compute $Pr(Paul \wedge \neg Peter)/Pr(\neg Peter)$. Since it is implicit in the story that exactly one person commits the murder, it follows that $Paul \wedge \neg Peter$ is logically equivalent to $Paul$. Thus, we are reduced to computing $Pr(Paul)/Pr(\neg Peter)$. Unfortunately, we are not given either $Pr(Paul)$ or $Pr(\neg Peter)$; thus, we cannot immediately solve the problem using the Bayesian approach.

Note that if we were given $Pr(Paul)$, then we could compute

$$Pr(\neg Peter) = Pr(\{Paul, Mary\}) = Pr(Paul) + Pr(Mary) = Pr(Paul) + 1/2.$$

Thus, we could solve our problem if we only knew $Pr(Paul)$. At this point, what we might call a *risky Bayesian* would say that, since we know that $Pr(\{Peter, Paul\}) = 1/2$, we will apply the *maximum entropy principle* [Jay78] and assume $Pr(Peter) = Pr(Paul) = 1/4$ (this is essentially what is called the *insufficient reason principle* by Laplace). Under this assumption, it is easy to see that $Pr(Paul|\neg Peter) = 1/3$. Despite the fact that this has been referred to as a "noninformative prior", one that somehow makes the "minimum" assumptions [Che85], it actually makes quite serious assumptions, not always justified [Fin73]. In this case, these assumptions lead to a particular answer (1/3) that cannot be justified as the right answer without additional assumptions on Big Boss' method of choosing between Peter and Paul.

An alternative approach, still within the Bayesian framework, is what is called in [SK89] the *cautious Bayesian* approach. Suppose we assume that there is some prior

[14] We have slightly simplified the presentation of [SK89], but, again, the essential details remain the same.

[15] As pointed out in [SK89], this assumption is necessary in order to make "Peter is not the killer" and "Peter has an alibi" equivalent. Without it, we would know that "Peter has an alibi" implies that "Peter is not the killer", but not the converse.

probability α that Paul committed the murder. Although we do not know exactly how Big Boss chooses between Peter and Paul given that the original coin toss lands tails (so that one of them must commit the murder), suppose we assume that he chooses Paul in this case with probability α; i.e., assume $Pr(Paul|\{Peter, Paul\}) = \alpha$, where $0 \leq \alpha \leq 1$. It is easy to see that we then have $Pr(Paul) = \alpha/2$, so $Pr(Paul|\neg Peter) = (\alpha/2)/(1/2 + \alpha/2) = \alpha/(\alpha+1)$. (Notice that the particular case of $\alpha = 1/2$, which was assumed in the risky Bayesian approach, gives us $Pr(Paul|\neg Peter) = 1/3$, as we computed above.) Since $0 \leq \alpha \leq 1$, all we can say is that $Pr(Paul|\neg Peter)$ is in the interval $[0, 1/2]$. This is intuitively reasonable; if, for example, we know that, rather than randomly choosing between Peter and Paul when the original coin toss lands heads, Big Boss definitely chooses Peter (so that $\alpha = 0$), then we know Paul could not have done it, and $Pr(Paul|\neg Peter) = 0$. On the other hand, if Big Boss definitely would choose Paul if the coin landed tails, then learning that Peter did not do it gives us no additional useful information. The probability that Paul does it remains at 1/2 once we learn that Peter did not do it.

One way we can view the original statement of the problem is that $\{Peter\}$ and $\{Paul\}$ represent nonmeasurable sets. According to the problem specification, the only measurable sets are $\{Peter, Paul\}$, $\{Mary\}$, $\{Peter, Paul, Mary\}$, and $\{\}$. The first two sets each have probability 1/2, the third has probability 1, and the empty set has probability 0. It is now easy to compute that $Pr_*(Paul) = 0$, $Pr^*(Paul) = 1/2$, and $Pr_*(Mary) = Pr^*(Mary) = 1/2$. Using the definitions of inner and outer conditional probability from [FH91a] described in Section 2 and the fact that $\neg Paul \wedge \neg Peter$ is logically equivalent to $Mary$, we can compute

$$\begin{aligned} Pr_*(Paul|\neg Peter) &= Pr_*(Paul)/(Pr_*(Paul) + Pr^*(Mary)) = 0, \\ Pr^*(Paul|\neg Peter) &= Pr^*(Paul)/(Pr^*(Paul) + Pr_*(Mary)) = 1/2. \end{aligned}$$

Notice that the interval $[0, 1/2]$ defined by $Pr_*(Paul|\neg Peter)$ and $Pr^*(Paul|\neg Peter)$ is precisely that computed by the cautious Bayesian. This is not an accident, but a direct consequence of the definition of the inner and outer conditional probabilities as lower and upper envelopes.

Now let Bel be the belief function corresponding to Pr. From the definitions, it is immediate that $Bel(Paul|\neg Peter) = Pr_*(Paul|\neg Peter) = 0$ and $Pl(Paul|\neg Peter) = Pr^*(Paul|\neg Peter) = 1/2$. By way of contrast, we get

$$\begin{aligned} Bel(Paul||\neg Peter) &= (Bel(\{Paul, Peter\} - Bel(Peter))/(1 - Bel(Peter)) = 1/2, \\ Pl(Paul||\neg Peter) &= Pl(Paul)/Pl(\neg Peter) = 1/2. \end{aligned}$$

It may seem strange that using DS conditioning there is no uncertainty regarding the conditional probability; both the conditional belief and conditional plausibility are 1/2. This unintuitive result is best explained in terms of the probabilistic process described in Section 3 corresponding to $Bel(Paul||\neg Peter)$. Recall that according to this process, we first choose an element satisfying $\neg Peter$ whenever possible. This amounts to assuming that Big Boss chooses Paul whenever he has a choice between choosing Peter and Paul; i.e., $Pr(Paul|\{Peter, Paul\}) = 1$. With this additional assumption, it is clear that the probability of choosing Paul given that Peter is not chosen is precisely 1/2.

Now suppose that we try to capture the evidence encoded in the observation that Peter did not commit the murder by $Pr_{\neg Peter}$, the probabilistic representation of evidence described earlier. (We could also use Shafer's representation, $Bel_{\neg Peter}$; the results would be the same.) A straightforward computation shows $Pr_{\neg Peter}(Paul) = 1/2$ (and $Pr_{\neg Peter}(Peter) = 0$, $Pr_{\neg Peter}(Mary) = 1/2$). However, in order to now compute $Pr(Paul|\neg Peter)$, we need to combine $Pr_{\neg Peter}(Paul)$ with $Pr_{prior}(Paul)$. Unfortunately, the problem statement does not give us this prior. This uncertainty in the prior can be modelled by considering a family of probability functions, just as we did in the previous section. Suppose we again assume that $Pr(Paul|\{Peter, Paul\}) = \alpha$. Then we get $Pr_{prior}(Paul) = \alpha/2$, $Pr_{prior}(Peter) = (1-\alpha)/2$, $Pr_{prior}(Mary) = 1/2$. As α ranges from 0 to 1, the prior probability that Paul is chosen ranges from 0 to 1/2. This is consistent with the information that we were given, since we know that the probability that one of Peter or Paul is chosen is 1/2, so the probability that Paul is chosen can be at most 1/2. Now by Proposition 4.4, we can compute the posterior probability by combining Pr_{prior} and $Pr_{\neg Peter}$. Sure enough, an easy computation shows that we get that $(Pr_{prior} \oplus Pr_{\neg Peter})(Paul) = \alpha/(\alpha+1)$. Again, this is the same answer as obtained by the cautious Bayesian.

Once more, we see from this example that different representations can lead to the same conclusions. However, we must be careful in our representation. If we view belief functions as generalized probabilities, then using DS conditioning leads us to inappropriate answers. If we view belief as evidence, we still have to take into account the conditional probability that Big Boss chooses Paul when he has to choose between Peter and Paul in order to even be able to use our techniques.

We remark that techniques similar to those used for the previous puzzle can also be used to analyze the *three prisoner puzzle* mentioned in the introduction. An analysis using the viewpoint of beliefs as generalized probabilities is carried out in [FH91a]; the interested reader is referred there for further details.

7 Discussion and conclusions

There are a number of ways that belief functions can be viewed, all of which give rise to a collection of mathematical objects that satisfy the same axioms (see Shafer's recent [Sha90] for a summary of most of the leading viewpoints). Many of these ways are essentially equivalent but, as we have seen, not all of them are. Different viewpoints may suggest strikingly different approaches to notions like updating and combining. The two viewpoints that we have discussed here, although quite distinct, both allow belief functions to be understood in terms of probability theory. Rather than being mysterious objects, belief functions now fit into a well-understood framework.

Of the two viewpoints that we have suggested, the idea of beliefs as generalized probabilities, although explicitly disavowed by Shafer [Sha90], is quite prevalent in the literature. The idea of beliefs as representations of evidence is also quite common, although perhaps not always in terms of the formulation we have presented here. As we have shown in our examples, either viewpoint can be used, provided we represent the evidence appropriately. Our key point is that confusing these viewpoints can lead to problems.

As we have shown, it is important to carefully distinguish these two views of belief functions. Indeed, the examples in [Bla87, Hun87, DZ86, Lem86, Pea89] regarding the counterintuitive nature of belief functions can all be explained in terms of a confusion of these two views. The confusion between the two views seems prevalent throughout the literature. For example, in [LG83], belief functions are used to represent the evidence of sensors; yet, they are introduced as generalized probabilities. That is, it is argued that a belief function which assigns $Bel(A) = 1/3$ and $Pl(A) = 2/3$ is appropriate to represent the fact that a reading on a sensor gives us uncertain information about the true probability of A, and all that can be said about the probability of A is that it is between $1/3$ and $2/3$. Yet these belief functions which are viewed as generalized probabilities are combined using the rule of combination. As our results suggest, this may lead to inappropriate representations of evidence.

While our framework does allow us to dismiss one type of criticism that has been directed at belief functions, there is another criticism, perhaps best formulated in [Pea89], that deserves close attention: namely, how useful are belief functions? To what extent can they serve as a basis for evidential reasoning?

In order to look at this issue more carefully, we need to consider each of the two views of belief functions separately. If we view belief functions as generalized probabilities, then there clearly is a useful role that they can serve. Kyburg [Kyb88] and others have argued forcefully in terms of looking at intervals rather than at point-valued probabilities. We subscribe to this point of view as well. Belief and plausibility functions do determine an interval that can be well understood in terms of probability theory (cf. Theorem 2.1). On the other hand, it is not clear that even if we subscribe to intervals then belief and plausibility functions are always the best representations. An alternative is just to work directly with a family of probability functions, and consider lower and upper envelopes of this family. As some of our results suggest, this might be a more useful representation. If, instead, we view belief functions as representations of evidence, then our results suggest that although the rule of combination does have a central role to play here, we do not need belief functions; probability functions will do. Moreover, the rule of combination breaks down in this context too if there is uncertainty in the probability of the evidence. It would be of interest to know if there is a variant of the rule of combination that can deal with this case.

It may perhaps be argued that our comments on and criticisms of belief functions are an artifact of our goal of trying to understand belief functions in a probabilistic framework. We agree with critics of the Bayesian approach who argue that it is not always appropriate to assign a probability to every event. Nevertheless, it does seem that there are situations when it is appropriate to assign probabilities. In this case, we feel that the results obtained from the belief function approach should agree with those obtained by using probabilities. Moreover, we feel that a thorough understanding of what happens in the purely probabilistic case can lead us to appropriate extrapolations in situations when precise probabilities are not available.

Ultimately, it seems to us that in order to use belief functions with any degree of confidence, we need to understand how beliefs are to be interpreted in practice. We have suggested two interpretations here, both firmly rooted in probability theory. Because probability theory is familiar, with a large body of results on both theory and practice, we feel that these interpretations are more useful than those of, say, Shafer and Tversky in terms of *canonical examples* [ST85, Sha82]. Indeed, in the commentary by the discussants which appears in [Sha82], there are numerous concerns expressed about the connection between the canonical examples and the way belief functions are applied in practice. A further advantage of the two particular interpretations we have taken is that they suggest the sources of the nonintuitive results that can arise from using belief functions. In particular, our results show that in situations where precise probabilities are not available, great care must be taken not to confound the two views of belief functions.

It is possible that in our effort to put belief functions in a probabilistic framework, we may have overlooked some important aspects of belief functions. There may be some features of belief functions that cannot be explained in terms of probability, but are nevertheless important in representing evidence. However, we feel that it is up to the advocates of the belief function approach to spell out clearly what these features are, and argue their importance.

Acknowledgements: This paper benefitted from discussions with numerous people, including Fahiem Bacchus, Alan Bundy, Dan Hunter, Ben Grosof, Ric Horvitz, Judea Pearl, Philippe Smets, Tom Strat, Moshe Vardi, and Larry Wasserman. Larry in particular pointed out numerous references and gave us background to the work in the statistical literature.

References

[Abe88] S. Abel. The sum-and-lattice-points method based on an evidential-reasoning system applied to the real-time vehicle guidance problem. In *Uncertainty in Artificial Intelligence 2*, pages 365–370, 1988.

[Ait68] J. Aitchison. Discussion on Professor Dempster's paper. *Journal of the Royal Statistical Society, Series B*, 30:234–237, 1968.

[Bla87] P. Black. Is Shafer general Bayes? In *Proc. Third AAAI Uncertainty in Artificial Intelligence Workshop*, pages 2–9, 1987.

[Che85] P. Cheeseman. In defense of probability. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI '85)*, pages 1002–1009, 1985.

[Cox46] R. Cox. Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.

[dCLM90] L. M. de Campos, M. T. Lamata, and S. Moral. The concept of conditional fuzzy measure, to appear. *International Journal of Intelligent Systems*, 1990.

[Dem67] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[Dem68] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30:205–247, 1968.

[Dia78] P. Diaconis. Review of "A Mathematical Theory of Evidence". *Journal of the American Statistical Society*, 73(363):677–678, 1978.

[DZ86] P. Diaconis and S. L. Zabell. Some alternatives to Bayes's rule. In B. Grofman and G. Owen, editors, *Proc. Second University of California, Irvine, Conference on Political Economy*, pages 25–38, 1986.

[Fal88] B. Falkenhainer. Towards a general-purpose belief maintenance system. In *Uncertainty in Artificial Intelligence 2*, pages 125–131, 1988.

[Fel57] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, New York, 2nd edition, 1957.

[FH91a] R. Fagin and J. Y. Halpern. A new approach to updating beliefs. In P.P. Bonissone, M. Henrion, L.N. Kanal, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence: Volume VI*, pages 347–374. Elsevier Science Publishers, Amsterdam, 1991.

[FH91b] R. Fagin and J. Y. Halpern. Uncertainty, belief, and probability. *Computational Intelligence*, 7(3):160–173, 1991.

[Fin73] T. L. Fine. *Theories of Probability.* Academic Press, New York, 1973.

[Goo60] I. J. Good. Weights of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society, Series B*, 22:319–331, 1960.

[GZ86] C. Genest and J. V. Zidek. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.

[Hac65] I. Hacking. *Logic of Statistical Inference.* Cambridge University Press, Cambridge, U.K., 1965.

[Hal50] P. Halmos. *Measure Theory*. Van Nostrand, 1950.

[Hec86] D. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. North-Holland, Amsterdam, 1986.

[HH86] E. Horvitz and D. Heckerman. The inconsistent use of measures of certainty in artificial intelligence research. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 137–151. North-Holland, Amsterdam, 1986.

[HT93] J. Y. Halpern and M. R. Tuttle. Knowledge, probability, and adversaries. *Journal of the ACM*, 40(4):917–962, 1993.

[Hun87] D. Hunter. Dempster-Shafer vs. probabilistic logic. In *Proc. Third AAAI Uncertainty in Artificial Intelligence Workshop*, pages 22–29, 1987.

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

[Jef83] R. C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, 1983.

[KM90] D. H. Krantz and J. Miyamoto. Priors and likelihood ratios as evidence. *Journal of the American Statistical Association*, 78(382):418–423, 1990.

[Kyb87] H. E. Kyburg, Jr. Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

[Kyb88] H. E. Kyburg, Jr. Higher order probabilities and intervals. *International Journal of Approximate Reasoning*, 2:195–209, 1988.

[Lem86] J. F. Lemmer. Confidence factors, empiricism, and the Dempster-Shafer theory of evidence. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. North-Holland, Amsterdam, 1986.

[LG83] J. D. Lowrance and T. D. Garvey. Evidential reasoning: an implementation for multisensor integration. Technical Report TN 307, SRI International, Menlo Park, Calif., 1983.

[LU88] Z. Li and L. Uhr. Evidential reasoning in a computer vision system. In *Uncertainty in Artificial Intelligence 2*, pages 403–412, 1988.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, Calif., 1988.

[Pea89] J. Pearl. Reasoning with belief functions: a critical assessment. Technical Report R-136, UCLA, 1989.

[Rus87] E. H. Ruspini. The logical foundations of evidential reasoning. Research Note 408, revised version, SRI International, Menlo Park, Calif., 1987.

[Sei81] T. Seidenfeld. Statistical evidence and belief functions. *PSA 1978*, 2:478–489, 1981.

[Sha76a] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

[Sha76b] G. Shafer. A theory of statistical evidence. In W. L. Harper and C. A. Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II*. Reidel, Dordrecht, Netherlands, 1976.

[Sha79] G. Shafer. Allocations of probability. *Annals of Probability*, 7(5):827–839, 1979.

[Sha81] G. Shafer. Constructive probability. *Synthese*, 48:1–60, 1981.

[Sha82] G. Shafer. Belief functions and parametric models (with commentary). *Journal of the Royal Statistical Society, Series B*, 44:322–352, 1982.

[Sha90] G. Shafer. Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 1990.

[SK89] P. Smets and R. Kennes. The transferable belief model: comparison with Bayesian models, manuscript, IRIDIA, Université Libre de Bruxelles, 1050 Brussels, Belgium. 1989.

[ST85] G. Shafer and A. Tversky. Languages and designs for probability judgment. *Cognitive Science*, 9:309–339, 1985.

[Wal81] P. Walley. Coherent lower (and upper) probabilities. Manuscript, Dept. of Statistics, University of Warwick, 1981.

[Wal87] P. Walley. Belief function representations of statistical evidence. *Annals of Statistics*, 18(4):1439–1465, 1987.

[Was88] L. Wasserman. Belief functions and likelihood. Technical Report 420, Carnegie-Mellon University, Department of Statistics, 1988.

[WF82] P. Walley and T. L. Fine. Towards a frequentist theory of upper and lower probability. *Annals of Statistics*, 10(3):741–761, 1982.

[Wil78] P. M. Williams. On a new theory of epistemic probability (review of "A Mathematical Theory of Evidence"). *The British Journal for the Philosophy of Science*, 29:74–85, 1978.

[Zad84] L. A. Zadeh. A mathematical theory of evidence (book review). *AI Magazine*, 5(3):81–83, 1984.