A Generalization of Bayesian Inference in the Dempster-Shafer Belief Theoretic Framework

J. N. Heendeni, K. Premaratne, M. N. Murthi, J. Uscinski
University of Miami
Coral Gables, FL, USA
Email: j.anuja@umiami.edu, {kamal, mmurthi, uscinski}@miami.edu

M. Scheutz
Tufts University
Medford, MA, USA
Email: matthias.scheutz@tufts.edu

***Abstract*—In the literature, two main views of Dempster-Shafer (DS) theory are espoused: DS theory as evidence (as described in Shafer's seminal book) and DS theory as a generalization of probability. These two views are not always consistent. In this paper, we employ the generalized probability view of DS theory to arrive at results that allow one to perform Bayesian inference within the DS theoretic (DST) framework. The importance of this generalization is its capability of handling a wider variety of data imperfections, a feature inherited from the DST framework. In the process of developing these results akin to Bayesian inference, we also arrive at an evidence combination strategy which is consistent with the generalized probability view of DS theory, a feature lacking in the popular Dempster's combination rule (DCR). Finally, using the data from a political science survey, we demonstrate the application of our results on an experiment which attempts to gauge the hidden attitude of an individual from his/her observed behavior.**

I. Introduction

Updating a knowledge base conditional upon new evidence is a primitive operation in evidence fusion [1]. In a probabilistic setting, the Bayes' rule performs this conditioning operation. On the other hand, Dempster-Shafer (DS) belief theory offers a more convenient framework for working with real-world data possessing a wider variety of data imperfections and when decision-making systems are called upon to carry out their reasoning tasks with a better understanding of these underlying data imperfections [2], [3], [4], [5], [6].

Motivation. A DST generalization of the Bayes' rule and Bayesian inference would therefore be of immense value in a wide variety of application scenarios. In [7] Dempster, and in [8] Smets, have already suggested such a generalization. The work being presented in this paper is however different. In particular, our work is based on the *Fagin-Halpern (FH) conditional* and it is consistent with upper and lower probability bounds. In the generalized probability view of DS theory, belief and plausibility are regarded as lower and upper bounds respectively for an underlying probability which is unknown. While this viewpoint has been refuted by Shafer in [9], Fagin and Halpern have suggested that, by utilizing two views of belief — belief as *evidence* and belief as *generalized probability* — and addressing them separately could solve most of the difficulties associated with the DST notion of belief [10].

Challenges. When these two views of belief are studied, one realizes that the point of separation between the two views is the *Dempster's combination rule (DCR)* which, in a sense, may not be 'compatible' with probability theory. Shafer also points out that, while the DST belief and plausibility functions can be interpreted as lower and upper bounds for the underlying probability, the same is not true for the belief and plausibility functions resulting from the application of the DCR [9].

Dempster's conditional is a direct derivative of the DCR [2], and therefore, it inherits the DCR's lack of compatibility with probability. Fully aware of this, Fagin and Halpern have introduced a conditional which is consistent with probability bounds [11].

Contributions. In this paper, we develop a DST generalization of the Bayes' rule and Bayesian inference based on this FH conditional so that compatibility with probability theory is preserved. In particular, this generalization allows one to compute the posterior belief/plausibility given the likelihood beliefs/plausibilities and prior beliefs/plausibilities. We also explore an equation that allows one to compute the posterior belief/plausibility given the posterior beliefs/plausibilities of 'atomic' propositions and prior beliefs/plausibilities. In a sense, this latter equation can be thought of as a counterpart to the DCR, but compatible with probability theory. We then validate our results by applying them to an uncertain dataset extracted from a political science survey [12], [13], where we attempt to use the observed behavior to gauge the hidden attitude of an individual.

II. Preliminaries

Let $\Theta = \{\theta_1, \ldots, \theta_M\}$ represent a sample space of finite mutually exclusive and exhaustive outcomes of an experiment; θ_i are the elementary events. Let $Pr(\cdot) : 2^\Theta \mapsto [0,1]$ denote a probability mass function (p.m.f.) defined on the σ-algebra of the power set of events 2^Θ.

In DS theory, Θ and θ_i are usually referred to as the *frame of Discernment (FoD)* and the *singletons.* We say that $m(\cdot) : 2^\Theta \mapsto [0,1]$ is a *basic belief assignment (BBA)* or *mass assignment* when $\sum_{A \subseteq \Theta} m(A) = 1$ and $m(\emptyset) = 0$. Propositions that receive non-zero mass are the *focal elements;* the set of focal elements is the *core* $\mathfrak{F}$. The triple $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$ is the *body of evidence (BoE).* For arbitrary $A \subseteq \Theta$, $\overline{A}$ denotes those singletons that are not in A. For $A, B \subseteq \Theta$, we will interchangeably use $(A \cap B)$ and (A, B) to denote the set theoretic intersection operation.

In DS theory, a focal element can be any singleton or a composite (i.e., non-singleton) proposition. A mass function is called *Bayesian* if each focal element is a singleton. DS theory captures the notion of *ignorance* by allowing masses to be allocated to composite propositions. For instance, the mass $m(\theta_i, \theta_j)$ allocated to the composite doubleton proposition (θ_i, θ_j), $\theta_i, \theta_j \in \Theta$, represents ignorance or lack of evidence to differentiate between its two constituent singletons. The state of complete ignorance can be captured via the DST *vacuous BoE* 1_Θ, which has Θ as its only focal element, i.e., in the vacuous BoE, $m(\Theta) = 1$ and $m(A) = 0$, $\forall A \subset \Theta$.

Given a BoE, $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$, the *belief* function $Bl : 2^\Theta \mapsto [0, 1]$ is $Bl(A) = \sum_{B \subseteq A} m(B)$. The *plausibility* function $Pl : 2^\Theta \mapsto [0, 1]$ is $Pl(A) = 1 - Bl(\overline{A})$. The *uncertainty interval* associated with A is $[Bl(A), Pl(A)]$. The p.m.f. $Pr(\cdot)$ defined on the σ-algebra 2^Θ is said to be *consistent* with $\mathcal{E}$ if $Bl(A) \leq Pr(A) \leq Pl(A)$, for all $A \subseteq \Theta$ [11]. The *pignistic p.m.f.* defined as $BetPr(\theta_i) = \sum_{\theta_i \in A} m(A)/|A|$, $\theta_i \in \Theta$, is such a consistent p.m.f. [14].

III. Conditioning Operation in DS Theory

Among the various notions of DST conditional abound in the literature, the Dempster's conditional, which is derived from the DCR [2], appears to be the more widely used.

Definition 1. Dempster's Conditional. Given the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$ and $B \subseteq \Theta$ s.t. $Pl(B) > 0$, the *Dempster's conditional belief and plausibility of A w.r.t. B* are

$$\begin{aligned} Bl(A\|B) &= [Bl(A \cup \overline{B}) - Bl(\overline{B})]/Pl(B); \\ Pl(A\|B) &= Pl(A \cap B)/Pl(B). \end{aligned}$$

However, the FH conditional [11] possesses several attractive properties and offers a unique probabilistic interpretation and hence a natural transition to the Bayesian conditional notion [11], [15], [6]. More importantly, the FH conditional, which we employ in our work, is consistent with probabilistic lower and upper bounds.

Definition 2. FH Conditional. Given the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$ and $B \subseteq \Theta$ s.t. $Bl(B) > 0$, the *FH conditional belief and plausibility of A w.r.t. B* are

$$\begin{aligned} Bl(A|B) &= Bl(A \cap B)/[Bl(A \cap B) + Pl(\overline{A} \cap B)]; \\ Pl(A|B) &= Pl(A \cap B)/[Pl(A \cap B) + Bl(\overline{A} \cap B)]. \end{aligned}$$

A less rigorous way to demonstrate the relationship of the FH conditional with the probabilistic lower and upper bounds is the following: Consider a BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$ and a consistent p.m.f. $Pr(\cdot)$ defined on Θ. For $Pr(A \cap B) \neq 0$, express $Pr(A|B)$ as

$$\begin{aligned} Pr(A|B) &= Pr(A \cap B)/[Pr(A \cap B) + Pr(\overline{A} \cap B)] \\ &= \frac{1}{1 + Pr(\overline{A} \cap B)/Pr(A \cap B)}. \end{aligned} \tag{1}$$

Use the bounds $Bl(A \cap B) \leq Pr(A \cap B) \leq Pl(A \cap B)$ and $Bl(\overline{A} \cap B) \leq Pr(\overline{A} \cap B) \leq Pl(\overline{A} \cap B)$ to get

$$Bl(A|B) \leq Pr(A|B) \leq Pl(A|B), \tag{2}$$

where $Bl(A|B)$ and $Pl(A|B)$ are the FH conditional belief and plausibility, respectively, as they appear in Definition 1.

We first show the following result regarding the relationship between the Dempster's conditional and the FH conditional. Although it is stated in [11] (as Corollary 3.8), the authors could not locate a proof of this result.

Lemma 1: The Dempster's conditional and the FH conditional are related as

$$Bl(A|B) \leq Bl(A\|B) \leq Pl(A\|B) \leq Pl(A|B).$$

Proof. Since $Pl(B) \geq Pl(A \cap B) + Bl(\overline{A} \cap B)$, we have

$$\frac{Pl(A \cap B)}{Pl(B)} \leq \frac{Pl(A \cap B)}{Pl(A \cap B) + Bl(\overline{A} \cap B)},$$

which implies that $Pl(A\|B) \leq Pl(A|B)$. In turn, this yields $Pl(\overline{A}\|B) \leq Pl(\overline{A}|B)$. So,

$$1 - Pl(\overline{A}\|B) \geq 1 - Pl(\overline{A}|B),$$

which implies that $Bl(A\|B) \geq Bl(A|B)$. ■

IV. Bayesian Inference in the DS Framework

Probabilistic inference makes use of the following relationship:

$$Pr(A|B) = \frac{Pr(B|A)\, Pr(A)}{Pr(B|A)\, Pr(A) + Pr(B|\overline{A})\, Pr(\overline{A})}. \tag{3}$$

The corresponding DST counterpart to (3) is the following:

Lemma 2: For $Bl(A), Bl(B) \neq 0$ and $Pl(A) \neq 1$,

$$\begin{aligned} Bl(A|B) &\geq \frac{Bl(B|A)\, Bl(A)}{Bl(B|A)\, Bl(A) + Pl(B|\overline{A})\, Pl(\overline{A})}; \\ Pl(A|B) &\leq \frac{Pl(B|A)\, Pl(A)}{Pl(B|A)\, Pl(A) + Bl(B|\overline{A})\, Bl(\overline{A})}. \end{aligned}$$

Proof. To show the lower bound on $Bl(A|B)$, note that

$$\begin{aligned} Bl(A|B) &= \frac{Bl(A \cap B)}{Bl(A \cap B) + Pl(\overline{A} \cap B)} \\ &= \frac{1}{1 + Pl(\overline{A} \cap B)/Bl(A \cap B)}. \end{aligned} \tag{4}$$

Use $Bl(A) \leq Bl(A \cap B) + Pl(A \cap \overline{B})$ to get

$$Bl(B|A) \leq \frac{Bl(A \cap B)}{Bl(A)} \implies Bl(B|A)\, Bl(A) \leq Bl(A \cap B). \tag{5}$$

Substitute (5) in (4):

$$\begin{aligned} Bl(A|B) &\geq \frac{1}{1 + Pl(\overline{A} \cap B)/Bl(B|A)\, Bl(A)} \\ &= \frac{Bl(B|A)\, Bl(A)}{Bl(B|A)\, Bl(A) + Pl(\overline{A} \cap B)}. \end{aligned} \tag{6}$$

To proceed, use $Pl(\overline{A}) \geq Pl(\overline{A} \cap B) + Bl(\overline{A} \cap \overline{B})$ to get

$$Pl(B|\overline{A}) \geq \frac{Pl(\overline{A} \cap B)}{Pl(\overline{A})} \implies Pl(\overline{A} \cap B) \leq Pl(B|\overline{A})\, Pl(\overline{A}). \tag{7}$$

Substitute (7) in (6) to get the required lower bound on $Bl(A|B)$.

To show the upper bound on $Pl(A|B)$, use the lower bound on $Bl(\overline{A}|B)$:

$$Bl(\overline{A}|B) \geq \frac{Bl(B|\overline{A})\, Bl(\overline{A})}{Bl(B|\overline{A})\, Bl(\overline{A}) + Pl(B|A) Pl(A)}.$$

This yields

$$1 - Bl(\overline{A}|B) \leq 1 - \frac{Bl(B|\overline{A})\, Bl(\overline{A})}{Bl(B|\overline{A})\, Bl(\overline{A}) + Pl(B|A) Pl(A)},$$

$1 - Bl(\overline{A}|B) = Pl(A|B)$, which gives the required upper bound on $Pl(A|B)$. ■

We generalize this result as follows:

Theorem 3: Consider the partition $\{A_i\}_{i=1}^n$ of Θ, i.e., $A_i \cap A_j = \emptyset$, $\forall i \neq j$ and $i, j \in \overline{1, n}$, and $\bigcup_{i=1}^n A_i = \Theta$. Then, for $k \in \{1, \ldots, n\}$,

$$Bl(A_k|B) \geq \frac{Bl(B|A_k)\, Bl(A_k)}{Bl(B|A_k)\, Bl(A_k) + \sum_{i \neq k} Pl(B|A_i)\, Pl(A_i)};$$

$$Pl(A_k|B) \leq \frac{Pl(B|A_k)\, Pl(A_k)}{Pl(B|A_k)\, Pl(A_k) + \sum_{i \neq k} Bl(B|A_i)\, Bl(A_i)}.$$

Proof. To show the upper bound on $Bl(A_k|B)$, use (6):

$$Bl(A_k|B) \geq \frac{Bl(B|A_k)\, Bl(A_k)}{Bl(B|A_k)\, Bl(A_k) + Pl(\overline{A}_k \cap B)}.$$

Since $\{A_i\}_{i=1}^n$ is a partition, we have

$$(\overline{A}_k \cap B) = \bigcup_{i \neq k} (A_i \cap B).$$

But, for mutually exclusive sets C_j, $j \in \overline{1, m}$, for some m, we know that $Pl\left(\bigcup_j C_j\right) \leq \sum_j Pl(C_j)$. So,

$$Pl(\overline{A}_k \cap B) \leq \sum_{i \neq k} Pl(A_i \cap B).$$

From (7) we can write

$$Pl(\overline{A}_k \cap B) \leq \sum_{i \neq k} Pl(A_i \cap B) \leq \sum_{i \neq k} Pl(B|A_i)\, Pl(A_i).$$

Therefore,

$$\begin{aligned} Bl(A_k|B) &\geq \frac{Bl(B|A_k)\, Bl(A_k)}{Bl(B|A_k)\, Bl(A_k) + Pl(\overline{A}_k \cap B)} \\ &\geq \frac{Bl(B|A_k)\, Bl(A_k)}{Bl(B|A_k)\, Bl(A_k) + \sum_{i \neq k} Pl(B|A_i)\, Pl(A_i)}. \end{aligned}$$

To show the lower bound on $Pl(A_k|B)$, note that

$$\begin{aligned} Pl(A_k|B) &= \frac{Pl(A_k \cap B)}{Pl(A_k \cap B) + Bl(\overline{A}_k \cap B)} \\ &= \frac{1}{1 + Bl(\overline{A}_k \cap B)/Pl(A_k \cap B)}. \end{aligned}$$

From (7) we know that

$$Pl(A_k \cap B) \leq Pl(B|A_k)\, Pl(A_k).$$

Since the sets $\{A_i \cap B\}_{i=1}^n$ are mutually exclusive,

$$Bl(\overline{A}_k \cap B) = Bl\left(\bigcup_{i \neq k} (A_i \cap B)\right) \geq \sum_{i \neq k} Bl(A_i \cap B).$$

From (5), we have

$$Bl(\overline{A}_k \cap B) \geq \sum_{i \neq k} Bl(A_i \cap B) \geq \sum_{i \neq k} Bl(B|A_i)\, Bl(A_i).$$

Therefore,

$$\begin{aligned} &Pl(A_k|B) \\ &= \frac{1}{1 + Bl(\overline{A}_k \cap B)/Pl(A_k \cap B)} \\ &\leq \frac{1}{1 + \sum_{i \neq k} Bl(B|A_i)\, Bl(A_i)/Pl(B|A_k)\, Pl(A_k)} \\ &= \frac{Pl(B|A_k)\, Pl(A_k)}{Pl(B|A_k)\, Pl(A_k) + \sum_{i \neq k} Bl(B|A_i)\, Bl(A_i)}. \end{aligned}$$

■

V. Evidence Combination: DCR Versus the Generalized Probability Viewpoint

A. Shafer's Example

Let us start by considering an example which Shafer has used in [9] to demonstrate the application of the DCR to combine two 'independent' pieces of evidence.

Suppose that Betty tells me a tree limb fell on my car. My subjective probability that Betty is reliable is α; my subjective probability that she is unreliable is $(1 - \alpha)$. Since they are probabilities, these numbers add to 1. But Betty's statement, which must be true if she is reliable, is not necessarily false if she is unreliable. So, her testimony alone justifies an α degree of belief that a limb fell on my car, but only a zero degree of belief (*not* $(1 - \alpha)$ degree of belief) that no limb fell on my car. This zero value does not mean that one is sure that no limb fell on my car, as a zero probability would; it merely means that Betty's testimony gives me no reason to believe that no limb fell on my car.

Let T be the event that a tree limb fell on my car. Then, from a DST point-of-view,

$$Bl(T) = \alpha; \quad Bl(\overline{T}) = 0.$$

These degrees of belief were derived from my α and $(1 - \alpha)$ subjective probabilities for Betty being reliable or unreliable. Suppose these subjective probabilities were based on my knowledge of the frequency with which witnesses like Betty are reliable. Then I might think that the $(1 - \alpha)$ proportion of witnesses like Betty who are not reliable make true statements a definite (though unknown) proportion of the time and false statements the rest of the time. Were this the case, I could think in terms of a large population of statements made by witnesses like Betty. In this population, α proportion of the statements would be true statements by reliable witnesses, x would be true statements by unreliable witnesses, and $(1 - \alpha - x)$ would be false statements by unreliable witnesses, where $x \in [0, 1 - \alpha]$ is an unknown number. The total chance of getting a true statement from this population would be $(\alpha + x)$, and the

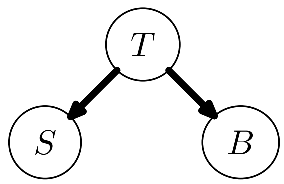

Fig. 1. Graphical model: T = A tree limb fell on my car; B = Betty tells me a tree limb fell on my car; and S = Sally tells me a tree limb fell on my car.

total chance of getting a false statement would be $(1-\alpha-x)$. My degrees of belief of α and 0 are lower bounds on these chances. Since $x \in [0, 1-\alpha]$, α and and 0 are the lower bounds for $(\alpha+x)$ and $(1-\alpha-x)$, respectively.

From this example Shafer has shown that a single belief function is always consistent with probability bounds [9].

Now consider another witness Sally. My subjective probability that she is reliable is β; my subjective probability that she is unreliable is $(1-\beta)$. We assume that these values are independent of my subjective probability of reliability of Betty. Suppose Sally also tells me that a tree limb fell on my car.

From the DCR, my combined degree of belief that a tree limb fell on my car will be [9]

$$1-(1-\alpha)(1-\beta).$$

B. Probabilistic Viewpoint of Shafer's Example

Now, let us view this example of Shafer [9] from a probabilistic viewpoint.

As described earlier let $(\alpha+x)$ be the value we assign for total chance of getting a true statement from Betty; $(1-\alpha-x)$ corresponds to a false statement from Betty. Using the same arguments, let $(\beta+y)$ be the value we assign for total chance of getting true statement from Sally; $(1-\beta-y)$ corresponds to a false statement from Sally. Consider the following events:

$T =$ A tree limb fell on my car;

$B =$ Betty tells me a tree limb fell on my car;

$S =$ Sally tells me a tree limb fell on my car.

Suppose we want to find the posterior probability $Pr(T|B,S)$, i.e., the combined probability that a tree limb fell on my car: For $Pr(T) \neq 0$ and $Pr(T) \neq 1$, and for $Pr(B,S) \neq 0$,

$$Pr(T|B,S) = \frac{Pr(B,S|T)\,Pr(T)}{Pr(B,S|T)\,Pr(T) + Pr(B,S|\overline{T})\,Pr(\overline{T})}.$$

Clearly, B and S are conditionally independent given T. See Fig. 1. Therefore,

$$Pr(T|B,S) = \frac{Pr(B|T)\,Pr(S|T)\,Pr(T)}{Pr(B|T)\,Pr(S|T)\,Pr(T) + Pr(B|\overline{T})\,Pr(S|\overline{T})\,Pr(\overline{T})} \tag{8}$$

$$= \frac{Pr(T|B)\,Pr(T|S)\,Pr(\overline{T})}{Pr(T|B)\,Pr(T|S)\,Pr(\overline{T}) + Pr(\overline{T}|B)\,Pr(\overline{T}|S)\,Pr(T)}, \tag{9}$$

where $Pr(T|B) = (\alpha+x)$ and $Pr(T|S) = (\beta+y)$. Note that, (8) is applicable for $Pr(B,S), Pr(T) \neq 0$ and $Pr(T) \neq 1$; (9) is applicable for $Pr(B,S), Pr(T) \neq 0$ and $Pr(T) \neq 1$.

To find $Pr(T|B,S)$, we need the value of $Pr(T)$. However, the minimum value that $P(T|B,S)$ can achieve is not $1-(1-\alpha)(1-\beta)$. In other words, the probabilistic lower bound is not consistent with what the DCR yields.

What is the reason for this incompatibleness of the DCR with probabilistic reasoning? Let us explore how DCR behaves when confronted with probabilistic evidence, i.e., only singleton propositions receive non-zero mass. When combining two pieces of evidence, the DCR uses multiplication of mass values, a tacit endorsement that the two pieces of evidence are independent. But then, independent events (with non-zero probability) cannot be disjoint. Therefore, the normalization operation of the DCR which removes the product values of propositions that are disjoint or incompatible, is not compatible with the independence assumption on which the DCR is based upon.

VI. Evidence Combination and Inference: A Generalized Probability Viewpoint

A. Notion of Independence

Let us consider Shafer's example again. From Fig. 1, one realizes that the two pieces of evidence B and S are not *independent,* but are only *conditionally independent* given T. Capturing the notion of independence (let alone conditional independence) within the DST framework is fraught with its own challenges. In probability, independence of A and B simply refers to $Pr(A \cap B) = Pr(A)\,Pr(B)$ where $Pr(\cdot)$ is the underlying p.m.f. Within the DST framework, should we use $Bl(A \cap B) = Bl(A)\,Bl(B)$, or $Pl(A \cap B) = Pl(A)\,Pl(B)$, or another similar relationship to refer to independence? Based on such relationships, different DST notions of independence have been proposed and abound in the literature e.g., see [2], [16]. However, what we can show, and all we know, is

$$\begin{aligned} &\max\{Bl(A \cap B), Bl(A)\,Bl(B)\} \\ &\qquad \leq Pr(A \cap B) = Pr(A)\,Pr(B) \\ &\qquad \leq \min\{Pl(A \cap B), Pl(A)\,Pl(B)\}. \end{aligned} \tag{10}$$

This does not imply $Bl(A \cap B) = Bl(A)\,Bl(B)$, or $Pl(A \cap B) = Pl(A)\,Pl(B)$, or any other similar relationship. Our approach, on the other hand, does not require picking any one particular DST independence notion. Rather, we look at the independence only in terms of the the underlying p.m.f.

B. Evidence Combination

Let us start with (9): for $Pr(B,S), Pr(T) \neq 0$ and and $Pr(T) \neq 1$,

$$Pr(T|B,S) = \frac{1}{1 + \dfrac{Pr(\overline{T}|B)\,Pr(\overline{T}|S)\,Pr(T)}{Pr(T|B)\,Pr(T|S)\,Pr(\overline{T})}}. \tag{11}$$

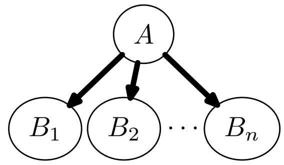

Fig. 2. A generalization of the graphical model in Fig. 1.

This then yields

$$Pr(T|B,S) \geq \frac{Bl(T|B)\,Bl(T|S)\,Bl(\overline{T})}{Bl(T|B)\,Bl(T|S)\,Bl(\overline{T}) + Pl(\overline{T}|B)\,Pl(\overline{T}|S)\,Pl(T)}, \tag{12}$$

for $Bl(B), Bl(S), Bl(T) \neq 0$ and $Pl(T) \neq 1$. Similarly,

$$Pr(T|B,S) \leq \frac{Pl(T|B)\,Pl(T|S)\,Pl(\overline{T})}{Pl(T|B)\,Pl(T|S)\,Pl(\overline{T}) + Bl(\overline{T}|B)\,Bl(\overline{T}|S)\,Bl(T)}, \tag{13}$$

for $Bl(B), Bl(S), Bl(T) \neq 0$ and $Pl(T) \neq 1$.

For $P(B_1,\ldots,B_n), Bl(B_i), Bl(A) \neq 0$ and $Pl(A) \neq 1$, we can generalize these bounds to correspond to the structure in Fig. 2 and get

$$\begin{aligned} L(A|B_1,\ldots,B_n) &\leq Pr(A|B_1,\ldots,B_n) \\ &\leq U(A|B_1,\ldots,B_n), \end{aligned} \tag{14}$$

where

$$L(A|B_1,\ldots,B_n) = \frac{\prod_{i=1}^{n} Bl(A|B_i)\,Bl(\overline{A})}{\prod_{i=1}^{n} Bl(A|B_i)\,Bl(\overline{A}) + \prod_{i=1}^{n} Pl(\overline{A}|B_i)\,Pl(A)}; \tag{15}$$

$$U(A|B_1,\ldots,B_n) = \frac{\prod_{i=1}^{n} Pl(A|B_i)\,Pl(\overline{A})}{\prod_{i=1}^{n} Pl(A|B_i)\,Pl(\overline{A}) + \prod_{i=1}^{n} Bl(\overline{A}|B_i)\,Bl(A)}. \tag{16}$$

One may interpret these lower and upper bounds $L(A|B_1,\ldots,B_n)$ and $U(A|B_1,\ldots,B_n)$, respectively, as the combined posterior (from the evidence of *all* the sources) given the individual posteriors (from evidence of *each* source). These quantities can be thought of as the results of an evidence combination strategy which is consistent with the generalized probability viewpoint of DS theory.

C. Inference

The bounds in (15) and (16) were obtained from (9). Now, let us start with (8). In a manner similar to above, we now get a generalization for Bayes' rule which allows one to get at the posterior from the likelihoods: for $Bl(A) \neq 0$ and $Pl(A) \neq 1$,

$$\begin{aligned} L_I(A|B_1,\ldots,B_n) &\leq Pr(A|B_1,\ldots,B_n) \\ &\leq U_I(A|B_1,\ldots,B_n), \end{aligned} \tag{17}$$

where we express $Pr(A|B_1,\ldots,B_n)$ in terms of likelihoods as

$$Pr(A|B_1,\ldots,B_n) = \frac{\prod_{i=1}^{n} Pr(B_i|A)\,Pr(A)}{\prod_{i=1}^{n} Pr(B_i|A)\,Pr(A) + \prod_{i=1}^{n} Pr(B_i|\overline{A})\,Pr(\overline{A})}, \tag{18}$$

for $Pr(B_1,\ldots,B_n), Pr(A) \neq 0$ and $Pr(A) \neq 1$; and

$$L_I(A|B_1,\ldots,B_n) = \frac{\prod_{i=1}^{n} Bl(B_i|A)\,Bl(A)}{\prod_{i=1}^{n} Bl(B_i|A)\,Bl(A) + \prod_{i=1}^{n} Pl(B_i|\overline{A})\,Pl(\overline{A})}; \tag{19}$$

$$U_I(A|B_1,\ldots,B_n) = \frac{\prod_{i=1}^{n} Pl(B_i|A)\,Pl(A)}{\prod_{i=1}^{n} Pl(B_i|A)\,Pl(A) + \prod_{i=1}^{n} Bl(B_i|\overline{A})\,Bl(\overline{A})}. \tag{20}$$

Note that, for $Pr(B_1,\ldots,B_n), Pr(A) \neq 0$ and $Pr(A) \neq 1$, (18) is a valid expression for the conditional probability function given that $B_1,\ldots,B_n$ are conditionally independent given A. When this is the case, for $Bl(A) \neq 0$ and $Pl(A) \neq 1$, one can show that $L_I(A|B_1,\ldots,B_n)$ in (19) is a valid conditional belief function. Furthermore, with (20), $L_I(A|B_1,\ldots,B_n) + U_I(\overline{A}|B_1,\ldots,B_n) = 1$ and so $U_I(A|B_1,\ldots,B_n)$ is the corresponding conditional plausibility function. Therefore, henceforth, we will employ the notation

$$\begin{aligned} Bl_I(A|B_1,\ldots,B_n) &\equiv L_I(A|B_1,\ldots,B_n); \\ Pl_I(A|B_1,\ldots,B_n) &\equiv U_I(A|B_1,\ldots,B_n). \end{aligned} \tag{21}$$

These equations are very useful for decision-making with uncertainty. For instance, consider the following decision criterion:

- Decide on A over $\overline{A}$ if

  $$Bl_I(A|B_1,\ldots,B_n) > Pl_I(\overline{A}|B_1,\ldots,B_n). \tag{22}$$

  We in fact used this criterion in the experiment which appears in Section VII.
- If $\{A_i\}_{i=1}^{m}$ is a partition of Θ, then we may employ an analogous criterion: Decide on A_k if

  $$Bl_I(A_k|B_1,\ldots,,B_n) > \max_i Pl_I(A_i|B_1,\ldots,B_n)\}, \tag{23}$$

where $i, k \in \{1, 2, \ldots, m\}$ and $i \neq k$.

Note that no relationship above requires one to pick one particular DST independence notion. Independence in our work refers to the underlying probability $Pr(\cdot)$, and it is irrelevant how $Bl(A \cap B)$ is related to $Bl(A)\,Bl(B)$, or how $Pl(A \cap B)$ is related to $Pl(A)\,Pl(B)$, etc.

VII. Experimental Results

TABLE I
DATA FROM THE SURVEY IN [12]: VARIABLES USED

Variable	State	Code
[Hidden Variable]	Very Liberal	1
Ideology - Yourself	Liberal	1
	Somewhat Liberal	1
	Middle of the Road	0
	Somewhat Conservative	-1
	Conservative	-1
	Very Conservative	-1
	Not Sure	0
	Skipped	0
	Not Asked	0
[Observed Behavior]	Yes	1
All things considered,	No	-1
do you think it was a	Not Sure	0
mistake to invade	Skipped	0
Afghanistan?	Not Asked	0
[Observed Behavior]	Strongly Approve	1
Institution Approval -	Somewhat Approve	1
Obama	Somewhat Disapprove	-1
	Strongly Disapprove	-1
	Not Sure	0
	Skipped	0
	Not Asked	0
[Observed Behavior]	Yes	1
All things considered,	No	-1
do you think it was a	Not Sure	0
mistake to invade	Skipped	0
Iraq?	Not Asked	0
[Observed Behavior]	Very Positive	1
What is your view of	Somewhat Positive	1
the Tea Party	Neutral	0
movement?	Somewhat Negative	-1
	Very Negative	-1
	Don't Know Enough to Say	0
	No Opinion	0
	Skipped	0
	Not Asked	0

Using the results developed above, we now explore if one may infer a person's implicit hidden attitude on a particular subject from his/her observed behavior. We assume that the relationship between this hidden attitude of a person and its related observed behavior is as depicted in Fig. 2: Hidden attitude is represented by A and multiple types of observed behavior (e.g., oral and written sentiments, facial expressions, responses to different questions, etc.) are represented by $B_1, \ldots, B_n$.

To test this idea, we use a political science dataset generated from a survey used in [13]. The survey, which was conducted online by YouGov, uses the input from a total of 1230 people (573 men and 657 women) who participated in the *2012 Cooperative Congressional Election Study (CCES)* [12].

From this dataset, we selected five variables: one as the hidden attitude and the other four as observed behavior. Table I shows these five variables and the set of answers that were allowed for each variable. The five variables were chosen among many other variables in the survey because they appeared to fit the 'hidden-observed' two-layer structure in Fig. 2. In the survey, not all the participants had answered all the questions and we employed the DST model of ignorance to capture this absence of information.

The first variable in the table, *Ideology - Yourself*, is the hidden variable chosen; the other four variables were chosen to be observed behaviors. For simplicity in this preliminary study, we converted all responses to two states, namely 1 and -1; 0 was used to represent uncertainty which captures the lack of information to discern between the two states. Note that, although some variables had clear alternate states (e.g., *Neutral*), these states were also categorized as 0.

We selected all the individuals who had selected one particular state of the hidden variable and used five-fold cross validation to train and test the system. The equations (19) and (20) were employed to calculate the lower and upper values of each of the two states $A = 1$ and $A = -1$ of the hidden variable given the observed states of the four observed variables $B_1, \ldots, B_4$. For convenience, hereafter, we will use A to denote $A = 1$ and $\overline{A}$ to denote $A = -1$.

The equations (19) and (20) need values for the likelihood parameters that appear in the right side. These parameters are calculated by the training dataset in cross validation with a frequentist approach.

For instance, using the notation B_1 to denote $B_1 = 1$ and A for $A = 1$, we used the following estimates from the records belonging to the training set:

$$Bl(B_1|A) = \frac{\text{\# of people with } A = 1 \text{ and } B_1 = 1}{\text{\# of people with } A = 1};$$
$$Pl(B_1|A) = \frac{\text{\# of people with } A = 1 \text{ and } B_1 = \{0, 1\}}{\text{\# of people with } A = 1}. \quad (24)$$

When making a decision regarding the hidden state of a data record belonging to the testing set, we used two criteria. For a given observed behavioral pattern (B_1, B_2, B_3, B_4), use the following:

- *Criterion 1:* Select A if
$$Bl_I(A|B_1, B_2, B_3, B_4) > Pl_I(\overline{A}|B_1, B_2, B_3, B_4); \quad (25)$$
select $\overline{A}$ if
$$Bl_I(\overline{A}|B_1, B_2, B_3, B_4) > Pl_I(A|B_1, B_2, B_3, B_4). \quad (26)$$
- *Criterion 2:* Select A if
$$Bl_I(A|B_1, B_2, B_3, B_4) > Bl_I(\overline{A}|B_1, B_2, B_3, B_4) \quad (27)$$
and
$$Pl_I(A|B_1, B_2, B_3, B_4) > Pl_I(\overline{A}|B_1, B_2, B_3, B_4); \quad (28)$$
select $\overline{A}$ if
$$Pl_I(\overline{A}|B_1, B_2, B_3, B_4) > Bl_I(A|B_1, B_2, B_3, B_4) \quad (29)$$

and

$$Bl_I(\overline{A}|B_1, B_2, B_3, B_4) > Pl_I(A|B_1, B_2, B_3, B_4). \tag{30}$$

Note that Criterion 1 implies Criterion 2. Therefore, Criterion 2 classifies more cases including all those classified by Criterion 1. However, Criterion 1 provides decisions with higher confidence. When taking a decision, say in favor of A over $\overline{A}$, Criterion 1 guarantees the underlying probabilities $P(A|B_1, P_2, P_3, P_4) > P(\overline{A}|B_1, P_2, P_3, P_4)$ whereas Criterion 2 does not provide such a guarantee.

We classified all the cases of the testing datasets and compared with the response that the subjects had already provided for the hidden variable. With Criterion 1, from among those cases that could be classified, a classification accuracy of 90.77% was obtained. There was 10.79% unclassified cases out of 890 total cases for which people have provided responses to the hidden variable with one of the states. With Criterion 2, a classification accuracy of 87.42% was obtained without any unclassified cases.

VIII. Conclusion

DST generalization of Bayesian inference can have many applications. The one we used in our experiment can be categorized as a classification problem of an uncertain dataset. Most methods available for classification of uncertain data are probability based and these methods have limitations in representing certain types of uncertainty [4], [17]. These limitations can be overcome by the generalized probability viewpoint of DS theory. The advantage of our method is that it can handle evidential data possessing a wider variety of data imperfections (including, probabilistic and possibilistic data [17]). While the proposed method reduces to a probabilistic method when the data are purely probabilistic, the advantages it offers become more pronounced in evidential data which call for DST models possessing non-singleton focal elements [5]. We must also mention that the absence of any DST independence notion also constitutes a significant advantage associated with our results which set them apart from previous work [7], [8].

This DST generalization can be used in dynamic applications as well. There are also several challenges to overcome. Some algorithms, e.g., the EM algorithm, are based on probabilistic properties and their generalization may result in the loss of some of the useful features. Another challenge is to extend the generalization to handle sample spaces possessing a continuum of values (which call for probability density functions in the probabilistic approach). A general criticism of DS theory is its computational complexity [6]. This computational problem was not encountered with our newly developed results.

Acknowledgment

This work is based on research supported by the U.S. Office of Naval Research (ONR) via grant #N00014-10-1-0140 and the U.S. National Science Foundation (NSF) via grant #1343430.

References

[1] S. Parsons, "Current approaches to handling imperfect information in data and knowledge bases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 3, pp. 353–372, June 1996.

[2] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press, 1976.

[3] R. R. Yager, M. Fedrizzi, and J. Kacprzyk, Eds., *Advances in the Dempster-Shafer Theory of Evidence*. New York, NY: John Wiley and Sons, 1994.

[4] S. Blackman and R. Popoli, *Design and Analyis of Modern Tracking Systems*. Norwood, MA: Artech House, 1999.

[5] K. K. R. G. K. Hewawasam, K. Premaratne, S. P. Subasingha, and M. L. Shyu, "Rule mining and classification in imperfect databases," in *International Conference on Information Fusion (FUSION)*, vol. 1, Philadelphia, PA, July 2005, pp. 661–668.

[6] T. L. Wickramarathne, K. Premaratne, and M. N. Murthi, "Towards efficient computation of the Dempster-Shafer belief theoretic conditionals," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 712–724, Apr. 2013.

[7] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no. 2, pp. 205–247, 1968.

[8] P. Smets, "Belief functions: The disjunctive rule of combination and the generalized bayesian theorem," *International Journal of Approximate Reasoning*, vol. 9, no. 1, pp. 1–35, Aug. 1993.

[9] G. Shafer, "Perspectives on the theory and practice of belief functions," *International Journal of Approximate Reasoning*, vol. 4, no. 5/6, pp. 323–362, Nov. 1990.

[10] J. Y. Halpern and R. Fagin, "Two views of belief: Belief as generalized probability and belief as evidence," *Journal Artificial Intelligence*, vol. 54, no. 3, pp. 275–317, Apr. 1992.

[11] R. Fagin and J. Y. Halpern, "A new approach to updating beliefs," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, Eds. New York, NY: Elsevier Science, 1991, pp. 347–374.

[12] S. Ansolabehere, "Cooperative congressional election study, 2006," 2012. [Online]. Available: http://doi.org/10.3886/ICPSR30141.v1

[13] J. E. Uscinski, C. Klofstad, and M. D. Atkinson, "What drives conspiratorial beliefs? the role of informational cues and predispositions," *Political Research Quarterly*, vol. 69, no. 1, pp. 57–71, Mar. 2016.

[14] P. Smets, "Practical uses of belief functions," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, K. B. Laskey and H. Prade, Eds. San Francisco, CA: Morgan Kaufmann, 1999, pp. 612–621.

[15] D. Denneberg, "Conditioning (updating) non-additive measures," *Annals of Operations Research*, vol. 52, no. 1, pp. 21–42, 1994.

[16] R. Jiroušek, "Local computations in Dempster-Shafer theory of evidence," *International Journal of Approximate Reasoning*, vol. 53, no. 8, pp. 1155–1167, Nov. 2012.

[17] T. L. Wickramarathne, K. Premaratne, M. Kubat, and D. T. Jayaweera, "CoFiDS: A belief-theoretic approach for automated collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 175–189, Feb. 2011.