Value of Information in the Framework of Evidence Theory

Radu Balan

Department of Mathematics and AMSC
University of Maryland, College Park, MD

February 5, 2025
American University, Washington DC

Acknowledgments

”This material is based upon work partially supported by the National Science Foundation under grant no. DMS-2108900 and Simons Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.”

Table of Contents:

1. DS Evidence Theory
2. Combination Rules, Conditionals and Marginalization for BPAs
3. Entropies
4. Next Step

Introduction to the Evidence Theory

Consider a finite state space $X = \{1, 2, \cdots, n\}$ of n possible outcomes (or results, or possible worlds). In a probability framework, one assigns a *probability mass function* (pmf) to each outcome:

$$p : X \to [0, 1] \ , \ \sum_{x \in X} p(x) = 1$$

so that, for any subset $A \subset X$, $P(A) = \sum_{i \in A} p(i)$, represents the *probability* that the outcome of an experiment is included in the set A.

The *Dempster-Shafer (DS) Evidence Theory*, known also as the theory of belief functions, was created to quantify and deal with *uncertainties*. Each experiment provides evidence that supports more or less certain possible outcomes (or results). The question is, how to quantify such cases?

Introduction to the Evidence Theory (2)

Example: A tossing die has six possible outcomes $X = \{1, 2, 3, 4, 5, 6\}$. Alice tossed the die and then says with 60% probability, the die landed on an even number. Bob got a quick glimpse of the die and was able to see that it had at least four dots showing. How to quantify these statements?
In classical probability one can formulate a pmf p so that
$P(\{2, 4, 6\}) = 0.6$ and $P(\{4, 5, 6\}) = 1$. What can we say about individual (atomic) probabilities? First we obtain: $p(1) = p(2) = p(3) = 0$. Thus $P(\{4, 6\}) = 0.6$ and hence $p(5) = 0.4$.
If the prior belief is the uniform probability, $p_0 = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ we can argue that $p(4) = p(6) = 0.3$.
But isn't this curious? Alice would believe it must be one of the three possible outcomes: 2,4 or 6. If Bob tells her that the die did not land on 2, Alice would conclude that the outcome must be either 4 or 6. In any case, not 5.

Introduction to the Evidence Theory (3)

For Bob, absent any information, he would believe that 4,5 and 6 would occur with equal probability. If Charlie (a third party) hears Alice believing that the outcome must have been an even number, and then Bob saying that is must be greater than or equal to 4, Charlie would not choose 5 at all.

In reality, the die may have landed on 4, and this does not produce a contradiction with any of Alice's or Bob's statements.

Notice that I used Bayes' rule to conclude the posterior distribution p satisfies $p(4) = p(6)$. But in reality, the die may have been biased. Maybe the true distribution is $p(1) = p(2) = p(3) = 0$, $p(4) = 0.5$, $p(5) = 0.4$ and $p(6) = 0.1$. These numbers would be consistent with Alice's and Bob's statements.

Basic Probability Assignments (BPA)

The DS Evidence Theory starts with a different concept: the *basic probability assignment* (bpa), or *mass function*. A BPA is a map m,

$$m : \mathcal{P}(X) \to [0,1]\ ,\ \ \sum_{A \subset X} m(A) = 1\ ,\ \ m(\emptyset) = 0$$

where $\mathcal{P}(X) = 2^X = \{A \,|\, A \subset X\}$ represents the set of all subsets of X (the power set of X).

Terminology:

- $A \subset X$ is called a *focal* element if $m(A) > 0$.
- m is said *deterministic* if m has a single focal element. We denote by e_A such a deterministic bpa with focal element A.
- m is said *Bayesian* if all focal elements are singleton subsets. This means that m is equivalent to a pmf p, $p(x) = m(\{x\})$ for all $x \in X$.
- The deterministic bpa e_X is called the *vacuous BPA*.

Belief, Plausibility and Commonality functions

Given a BPA $m : \mathcal{P}(X) \to [0,1]$ one constructs the following three functions:

Belief function *Bel*,

$$Bel_m : \mathcal{P}(X) \to [0,1] \ , \ Bel_m(A) = \sum_{B \subset A} m(B)$$

Plausibility function *Plaus*,

$$Plaus_m : \mathcal{P}(X) \to [0,1] \ , \ Plaus_m(A) = \sum_{B \in \mathcal{P}(X) : B \cap A \neq \emptyset} m(B)$$

Commonality function *Q*,

$$Q_m : \mathcal{P}(X) \to [0,1] \ , \ Q_m(A) = \sum_{B \in \mathcal{P}(X) : A \subset B} m(B)$$

BPA m, Bel, Plaus and Q are equivalent representations: given one we can always transform into another. We focus on the transformation $Bel \to m$.

Möbius Transform

Given m a BPA, the Belief function is defined by $Bel_m(A) = \sum_{B \subset A} m(B)$. The inverse transform is called the *Möbius transform*: Given $Bel : \mathcal{P}(X) \rightarrow [0,1]$ its BPA m is given by

$$m(\emptyset) = Bel(\emptyset) = 0 \ , \ m(\{x\}) = Bel(\{x\}) \ , \ \forall x \in X$$

and then recursively

$$m(A) = Bel(A) - \sum_{B \subsetneq A} m(B)$$

$m(A)$ is "the amount of belief committed to A that has not already been committed to its subsets" [Halpern 2003]. The closed form formula is

$$m(A) = \sum_{B \subset A} (-1)^{|A|-|B|} Bel(B) \ , \ \forall B \in \mathcal{P}(X)$$

Belief Functions

Definition. A *Belief Function Bel* is a map $Bel : \mathcal{P}(X) \to [0,1]$ that satisfies the following:

1. $Bel(\emptyset) = 0$.
2. $Bel(X) = 1$,
3. For $A_1, A_2, \cdots, A_k \in \mathcal{P}(X)$,

$$Bel(\cup_{i=1}^k A_i) \geq \sum_{i=1}^{k} \sum_{I \in [k], |I|=i} (-1)^{i+1} Bel(\cap_{j\in I} A_j)$$

Conditions 1 and 2: positivity and normalization. Condition 3: stronger form of super-additivity. In particular, a belief function satisfies:

$$A \subset B \Rightarrow Bel(A) \leq Bel(B)$$

$$Bel(A \cup B) \geq Bel(A) + Bel(B) - Bel(A \cap B)$$

Example revisited

$X = \{1,2,3,4,5,6\}$. Alice tossed the die and then says with 60% probability, the die landed on an even number. Bob got a quick glimpse of the die and was able to see that it had at least four dots showing. The Evidence Theory framework would produce:

$$Bel(\{2,4,6\}) = 0.6 \quad , \quad Bel(\{4,5,6\}) = 1$$

Beliefs is assigned to other sets compatible with definition:

- If $A \cap \{1,2,3\} \neq \emptyset$ then $m(A) = 0$
- $m(\{4\}) + m(\{6\}) + m(\{4,6\}) = 0.6$
- $m(\{5\}) + m(\{4,5\}) + m(\{5,6\}) + m(\{4,5,6\}) = 0.4$

Two extreme cases: all sets have mass 0 except for:
(1) the class of Bayesian BPAs (PMFs): $m(\{4\}) + m(\{6\}) = 0.6$ and $m(\{5\}) = 0.4$;
(2) the class of non-PMFs: $m(\{4,6\}) = 0.6$ and $m(\{4,5,6\}) = 0.4$.

Bayes' Rule of Combination

Asume $x \in X$ and $y \in Y$ are two random variables characterized by their joint distribution $p_{X,Y}$. The goal of this secton is to describe how to obtain the conditional probability distributions $p_{X|Y}$ and $p_{Y|X}$, marginals, p_X and p_Y, and converse relationships.

The conditional probabilities and marginals are related via

$$p_{X|Y=y}(x)p_Y(y) = p_{X,Y}(x,y) = p_{Y|X=x}(y)p_X(x)$$

We can say $p_{X|Y=y} \propto p_{X,Y}(x,y)$ where the proportionality constant (independent of x) must be chose to normalize the left hand-side. What can we say about the distribution of x if we only know $y \in U \subset Y$? We denote this by $p_{X|U}$

$$p_{X|U}(x) \propto \sum_{y \in U} p_{X,Y}(x,y) \Rightarrow p_{X|U}(x) = \frac{1}{C} \sum_{y \in Y} p_{X,Y}(x,y) 1_U(y)$$

where $C = P_Y(U) = \sum_{y \in U} p_Y(y)$.

Bayes' rule of combination (2)

The marginal p_X must obey:

$$p_X(x) \propto \sum_{y \in Y} p_{X,Y}(x,y) = \sum_{y \in Y} p_{X,Y}(x,y) 1_Y(y)$$

Since the right hand-side is already normalized we conclude

$$p_X(x) = \sum_{Y \in Y} p_{X,Y}(x,y)$$

Conditionals:

Type 1: When $A \subset X$ condition the pmf p_X with respect to $x \in A$:
$p_{X|A}(x) = \frac{1}{C} p_X(x) 1_A(x)$ with $C = p_X(A) = \sum_{x' \in A} p_X(x')$.

Type 2: For $U \subset Y$, the conditioning of joint pmf $p_{X,Y}$ to $y \in U$ is given by:

$$p_{X|U}(x) \propto \sum_{y \in Y} p_{X,Y}(x,y) 1_U(y)$$

the right hand-side needs to be normalized to obtain:

$$p_{X|U}(x) = \frac{1}{C} \sum_{y \in Y} p_{X,Y}(x,y) 1_U(y) \ , \ C = P_Y(U) = \sum_{y \in U} p_Y(y).$$

Bayes' rule of combination (2)

The marginal p_X must obey:

$$p_X(x) \propto \sum_{y\in Y} p_{X,Y}(x,y) = \sum_{y\in Y} p_{X,Y}(x,y)1_Y(y)$$

Since the right hand-side is already normalized we conclude

$$p_X(x) = \sum_{Y\in Y} p_{X,Y}(x,y)$$

Conditionals:

Type 1: When $A \subset X$ condition the pmf p_X with respect to $x \in A$:
$p_{X|A}(x) = \frac{1}{C} p_X(x) 1_A(x)$ with $C = p_X(A) = \sum_{x'\in A} p_X(x')$.

Type 2: For $U \subset Y$, the conditioning of joint pmf $p_{X,Y}$ to $y \in U$ is given by:

$$p_{X|U}(x) \propto \sum_{y\in Y} p_{X,Y}(x,y)1_U(y)$$

the right hand-side needs to be normalized to obtain:

$$p_{X|U}(x) = \frac{1}{C} \sum_{y\in Y} p_{X,Y}(x,y)1_U(y) \ , \ C = P_Y(U) = \sum_{y\in U} p_Y(y).$$

Dempster's Rule of Combination

Dempster's rule of combination says how to combine two BPAs, say m_1 and m_2 referring to *same* single variable x. The combined BPA is denoted $m_1 \oplus m_2$ and is defined by:

$$m_1 \oplus m_2(A) = \frac{1}{K} \sum_{\substack{B_1, B_2 \in \mathcal{P}(X) \\ B_1 \cap B_2 = A}} m_1(B_1) m_2(B_2)$$

where the normalization constant K is

$$K = 1 - \sum_{\substack{B_1, B_2 \in \mathcal{P}(X) \\ B_1 \cap B_2 = \emptyset}} m_1(B_1) m_2(B_2)$$

The normalizaion constant is a measure of "conflict" between the two belief assignments: if $K = 0$ m_1 and m_2 are said to be in *total conflict* and cannot be combined. If $K = 1$ then m_1 and m_2 are said *non-conflicting*.

Dempster's Rule of Combination (2)

A key observation:For the commonality function:

$$Q_{m_1 \oplus m_2} = \frac{1}{K} Q_{m_1} Q_{m_2}$$

Why:

Recall $Q_m(A) = \sum_{A \subset B \subset X} m(B)$. Fix $A \subset X$,

$$\sum_{A \subset B \subset X} \sum_{\substack{B_1, B_2 \in \mathcal{P}(X) \\ B_1 \cap B_2 = A}} m_1(B_1) m_2(B_2) = \sum_{A \subset B_1 \subset X} \sum_{A \subset B_2 \subset X} m_1(B_1) m_2(B_2)$$

$$= \left(\sum_{A \subset B_1 \subset X} m_1(B_1) \right) \left(\sum_{A \subset B_2 \subset X} m_2(B_2) \right)$$

And the constant K remains the same.

Dempster's Rule of Combination (3)

Suppose that m_X and m_Y are BPAs of two different variables $x \in X$ and $y \in Y$. In this case, the combination $m_1 \oplus m_2$ is defined on $\mathcal{P}(X \times Y) = 2^{X \times Y}$ via

$$m_X \oplus m_Y(A \times B) = m_X(A) m_Y(B) \quad , \quad A \in \mathcal{P}(X), B \in \mathcal{P}(Y)$$

No additional normalization needed since this is already normalized.
The above relation assumes some sort of "independence". However it may be more general. It tries to mimick the Bayes' rule $p_{X|Y} p_Y = p_{X,Y}$.

Marginalization in DS Theory

Consider m a BPA for $X \times Y$. How to get BPAs on X and Y spaces only? We define the "projection" operators $\downarrow X : \mathcal{P}(X \times Y) \to \mathcal{P}(X)$ by

$$A \in \mathcal{P}(X \times Y) \mapsto A^{\downarrow X} = \{x \in X \ , \ \exists y \in Y \ s.t. \ (x,y) \in A\}$$

$A^{\downarrow X}$ represents the "footprint" of set A when projected to X. Similar definition for $A^{\downarrow Y}$, the projection onto Y space.
Then for m a BPA on $X \times Y$, its *marginal* $m^{\downarrow X} : \mathcal{P}(X) \to [0,1]$ is

$$m^{\downarrow X}(B) = \sum_{\substack{A \in \mathcal{X} \times \mathcal{Y} \\ A^{\downarrow X} = B}} m(A)$$

Relationship: If m_X and m_Y are BPAs on X and Y respectively and $m_X \oplus m_Y$ is the BPA on $X \times Y$ defined by Dempster's rule, then:

$$(m_X \oplus m_Y)^{\downarrow X} = m_X \quad , \quad (m_X \oplus m_Y)^{\downarrow Y} = m_Y .$$

Shannon's Entropy, Hartley's Entropy

Assume x is a random variable with finite state (sampling) space $X = \{1, 2, \cdots, n\}$ and pmf p_X.
Shannon defines the *information content* of state x as

$$I(x) = \log\left(\frac{1}{p_X(x)}\right)$$

The *Shannon entropy* is then the expected value of its information content:

$$H(x) = \sum_{x \in X} p_X(x) \log\left(\frac{1}{p_X(x)}\right) = -\sum_{x \in X} p_X(x) log(p_X(x))$$

For Hartley, the *measure of uncertainty* is given by the size of X, specifically by its logarithm:

$$H_0(x) = log(n)$$

Note the two measures (entropies) coincide when x has a uniform distribution over X, i.e., $p_X(x) = \frac{1}{n}$.

Höhle, Smets, Yager and Nguyen Entropies

Consider m a BPA on X, and Bel_m, $Plaus_m$ and Q_m its associated belief, plausability and commonality function respectively. Then

$$\text{(Höhle)} \quad H_o(m) = \sum_{A \in \mathcal{P}(X)} m(A) \log\left(\frac{1}{Bel_m(A)}\right)$$

$$\text{(Smets)} \quad H_s(m) = \sum_{A \in \mathcal{P}(X)} \log\left(\frac{1}{Q_m(A)}\right)$$

$$\text{(Yager)} \quad H_{Yager}(m) = \sum_{A \in \mathcal{P}(X)} m(A) \log\left(\frac{1}{Plaus_m(A)}\right)$$

$$\text{(Nguyen)} \quad H_{Nguyen}(m) = \sum_{A \in \mathcal{P}(X)} m(A) \log\left(\frac{1}{m(A)}\right)$$

They all capture the conflict portion of uncertainty. For Bayesian BPAs, $H_0 = H_y = H_n = H$, they reduce to Shannon's information.

Transform based Entropies

A different type of entropies are obtained by first transforming the BPA m into a PMF p, and then computing the Shannon's sntropy of p. There are three transforms:

1. The *pignistic transform*:

$$BetP_m(x) = \sum_{A\in\mathcal{P}(X):x\in A} \frac{m(A)}{|A|} \quad , \quad x \in X$$

Jousselme et al entropy:

$$H_j(m) = H(BetP_m) = \sum_{x\in X} BetP_m(x) \log\left(\frac{1}{BetP_m(x)}\right)$$

2. The *credal set* is defined by

$$\Pi_m = \{p \in [0,1]^n : \sum_x p(x) = 1 \ , \ \forall A \in \mathcal{P}(X) \, , \sum_{x\in A} p(x) \geq Bel_m(A)\}$$

Belief is the lower probability; Plausibility is the upper probability.

Transform based Entropies (2)

2. (cont'd) The *Max Entropy of the credal set* introduced by Harmanec and Klir:

$$\text{(Harmanec and Klir)} \quad H_{HK}(m) = \max_{p \in \Pi_m} H(p)$$

The Maximum Entropy Credal Set transform CrP_m is the maximizer above.

3. The *plausibility transform*:

$$PlausP_m(x) = \frac{1}{K} Plaus_m(\{x\}) = \frac{1}{K} \sum_{A \in \mathcal{P}(X): x \in A} m(A) \quad , \quad x \in X$$

where K is a nomalizing constant,
$K = \sum_{x \in X} Plaus_m(\{x\}) = \sum_{A \in \mathcal{P}(X)} |A| m(A)$.
Jirousek and Shenoy considered the Shannon entropy associated to $PlausP_m$, $H(PlausP_m)$. Their full entropy adds a secod term to it.

Expanded entropies

To satisfy additional properties, more recent families of entropies were introduced by taking the sum of two terms, one that captures the conflict portion of the uncertainty and the other that is a measure of non-specificity. In most cases, the non-specificity part is given by the Dubois and Prade entropy:

$$H_{DP}(m) = \sum_{A \in \mathcal{P}(X)} m(A) log(|A|).$$

$$\text{(Lamata and Moral)} \quad H_{LM}(m) = H_{Yager}(m) + H_{DP}(m)$$

$$\text{(Maeda and Ichihasi)} \quad H_{MI}(m) = H(CrP_m) + H_{DP}(m) = H_{HK}(m) + H_{DP}(m)$$

$$\text{(Deng)} \quad H_D(m) = H_{Nguyen}(m) + \sum_{A \in \mathcal{P}(X)} m(A) log(2^{|A|} - 1)$$

$$\text{(Jirousek and Shenoy)}^1 \quad H_{JS}(m) = H(PlausP_m) + H_{DP}(m)$$

[1] Acknowledgment: Most of the results and notation taken from their paper, Int.J.Approx.Reason. 2018

Desired properties for an "ideal" entropy

1. (Existence and Continuity) The "ideal" entropy $m \mapsto S(m)$ should be continuous.
2. (Probabilistic Consistency) If m is Bayesian BPA with PMF p, then $S(m) = H(p)$.
3. (Non-negativity) $S(m) \geq 0$ and $S(m) = 0$ iff m is Bayesian and deterministic, i.e., $m(\{x\}) = 1$ for some $x \in X$.
4. (Maximum Entropy) The vacuous BPA e_X has maximum entropy: $S(m) \leq S(e_X)$
5. (Monotonicity) If $|X| < |Y|$ then $S(e_X) < S(e_Y)$
6. (Additivity) For m_X and m_Y on two distinct X and Y so $m_X \oplus m_Y$ is defined on $\mathcal{P}(X \times Y)$, $S(m_X \oplus m_Y) = S(m_X) + S(m_Y)$.
7. (Subadditivity) If m is a BPA on $X \times Y$ then $H(m) \leq H(m^{\downarrow X}) + H(m^{\downarrow Y})$.
8. (Consistency with DS)[2] If a component of the entropy is defined via a PMF $m \mapsto p_m$, then the PMF transform must satisfy $p_{m_1 \oplus m_2} = p_{m_1} \otimes p_{m_2}$.

[2] from Jirousek and Shenoy. H_{JS} satisfies 1-8 except 7. H_{HK} and H_{DP} satisfy 7.

What is the Value of Information in this context?

In certain works, the *value of information* is defined as the decrease in Shannon's entropy between the state without information and the state with information. Specifically, assume X is the variable of interest, and Y denotes the information. Then the *value of information* Y is quantified by

$$Value_X(Y) = H(X) - H(X|Y)$$

where the conditional entropy is defined as

$$H(X|Y) = \sum_{y \in Y} p_Y(y) H(p_{X|Y=y}).$$

We posit that the value of information in the context of Dempster-Shafer theory of belief functions, should be measured by a similar decrease in entropy. The challenge will be to see which of the 10+ entropies is the most appropriate!

THANK YOU!

QUESTIONS?