Introduction to Belief Functions

Belief functions on finite frames. Dempster's rule

Thierry Denœux

Summer 2022

Outline of the course I

Course homepage:

https://www.hds.utc.fr/~tdenoeux/dokuwiki/en/bf

1. Basic notions. Classification
   1. Belief functions on finite sets. Dempster's rule (lecture + exercises)
   2. Decision making (lecture + exercises)
   3. Evidential *k*-NN classification:
      - *A k-nearest neighbor classification rule based on Dempster-Shafer theory*
      - *An evidence-theoretic k-NN rule with parameter optimization*
   4. *A neural network classifier based on Dempster-Shafer theory* (paper reading + exercises in R)
2. Clustering
   1. *Evidential clustering of large dissimilarity data* (paper reading + exercises in R)
   2. *NN-EVCLUS: Neural Network-based Evidential Clustering* (paper reading + exercises in R)
   3. *Calibrated model-based evidential clustering using bootstrapping* (paper reading + exercises in R)

Outline of the course II

3. Statistical inference, prediction, regression
   1. Likelihood-based belief function:
      - *Likelihood-based belief function: Justification and some extensions to low-quality data*
      - *Combining statistical and expert evidence using belief functions: Application to centennial sea level estimation taking into account climate change*
   2. Prediction:
      - *Prediction of future observations using belief functions: a likelihood-based approach*
      - *Evidential calibration of binary SVM classifiers*
   3. Uncertain data:
      - *Maximum likelihood estimation from Uncertain Data in the Belief Function Framework*
      - *Parametric Classification with Soft Labels using the Evidential EM Algorithm*
   4. Random fuzzy sets and evidential regression
      - *Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: general framework and practical models*
      - *An Evidential Neural Network Model for Regression Based on Random Fuzzy Numbers*

What we will study in this part

- A mathematical formalism called
  - Dempster-Shafer (DS) theory
  - Evidence theory
  - Theory of belief functions
- This formalism was introduced by A. P. Dempster in the 1960's for statistical inference, and developed by G. Shafer in the late 1970's into a general theory for reasoning under uncertainty.
- DS encompasses probability theory and set-membership approaches such as interval analysis as special cases: it is very general.
- Many applications in AI (expert systems, machine learning), engineering (information fusion, uncertainty quantification, risk analysis), statistics (statistical estimation and prediction), etc.
- Some applications to econometrics. A new research avenue to explore!

Outline

Outline

Mass function

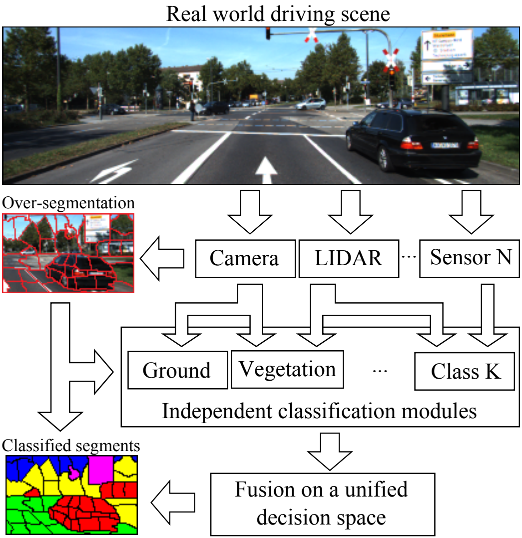
Definition

- Let X be a variable taking one and only one value in a finite set Ω, called the frame of discernment
- Evidence (uncertain information) about X can be represented by a mass function $m : 2^\Omega \rightarrow [0, 1]$ such that

$$\sum_{A \subseteq \Omega} m(A) = 1$$

- Every subset A of Ω such that $m(A) > 0$ is a focal set of m
- m is said to be normalized if $m(\emptyset) = 0$. This property will be assumed throughout this course, unless otherwise specified.

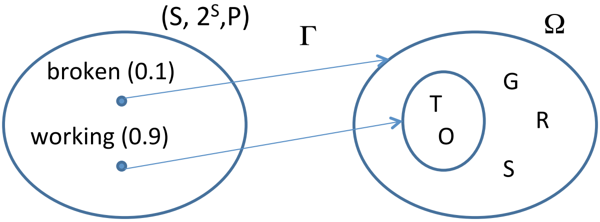
Example: road scene analysis



Example: road scene analysis (continued)

- Let X be the type of object in some region of the image, and $\Omega = \{G, R, T, O, S\}$, corresponding to the possibilities **G**rass, **R**oad, **T**ree/Bush, **O**bstacle, **S**ky.
- Assume that a lidar sensor (laser telemeter) returns the information $X \in \{T, O\}$, but we there is a probability $p = 0.1$ that the information is not reliable (because, e.g., the sensor is out of order).
- How to represent this information by a mass function?

Formalization



- Here, the probability p is not about X, but about the state of a sensor.
- Let $S = \{\text{working}, \text{broken}\}$ the set of possible sensor states.
  - If the state is "working", we know that $X \in \{T, O\}$.
  - If the state is "broken", we just know that $X \in \Omega$, and nothing more.
- This uncertain evidence can be represented by a mass function m on Ω, such that

$$m(\{T, O\}) = 0.9, \quad m(\Omega) = 0.1$$

General framework

- A model with three components:
  - A set $S = \{s_1, \ldots, s_r\}$ of states (interpretations of a piece of evidence)
  - A probability measure P on S
  - A multi-valued mapping $\Gamma : S \rightarrow 2^\Omega$
- The four-tuple $(S, 2^S, P, \Gamma)$ is called a source for m. It induces a mass function of Ω.
- Meaning: under interpretation $s \in S$, the evidence tells us that $X \in \Gamma(s)$, and nothing more. The probability $P(\{s\})$ is transferred to the set $A = \Gamma(s)$ and we have

$$m(A) = \sum_{s \in S : \Gamma(s) = A} P(\{s\})$$

- $m(A)$ is the probability of knowing that $X \in A$, and nothing more, given the available evidence.

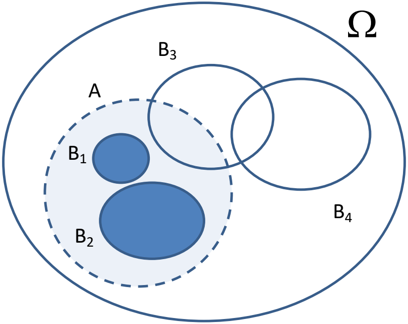
Special cases

- If the evidence tells us that $X \in A$ for sure and nothing more, for some $A \subseteq \Omega$, then we have a logical mass function m_A such that $m_A(A) = 1$.
  - Example: $m_{\{T,O\}}$ means the mass function such that $m_{\{T,O\}}(\{T,O\}) = 1$.
- Special case: $m_?$, the vacuous mass function, represents total ignorance
- If all focal sets of m are singletons, m is said to be Bayesian. It is equivalent to a probability distribution.
  - Example: $m(\{T\}) = 0.5$, $m(\{O\}) = 0.5$.
- A Dempster-Shafer mass function can thus be seen as
  - a generalized set
  - a generalized probability distribution

Outline

Belief function

- If the evidence tells us that the truth is in B, and $B \subseteq A$, we say that the evidence supports A.



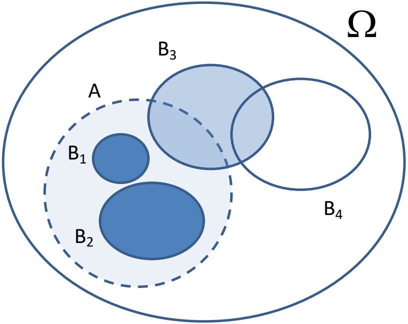
- Given a normalized mass function m, the probability that the evidence supports A is thus

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

- The number $Bel(A)$ is called the credibility of A, or the degree of belief in A, and the function $A \rightarrow Bel(A)$ is called a belief function.

Plausibility function

- If the evidence does not support $\overline{A}$, it is said to be consistent with A.



- The probability that the evidence is consistent with A is thus

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B).$$

- The number $Pl(A)$ is called the plausibility of A, and the function $A \rightarrow Pl(A)$ is called a plausibility function.

Interpretation and elementary properties

- Properties:
  1. $Bel(\emptyset) = Pl(\emptyset) = 0$
  2. $Bel(\Omega) = Pl(\Omega) = 1$
  3. For all $A \subseteq \Omega$,

$$Bel(A) = 1 - Pl(\overline{A})$$
$$Pl(A) = 1 - Bel(\overline{A})$$

- Interpretation:
  - $Bel(A)$ is the probability that A is supported by the evidence
  - $Bel(\overline{A})$ is the probability that $\overline{A}$ is supported by the evidence
  - $Pl(A) = 1 - Bel(\overline{A})$ is the probability that $\overline{A}$ is not supported by the evidence, i.e., that A is consistent with the evidence

Two-dimensional representation

- The uncertainty about a proposition A is represented by two numbers: $Bel(A)$ and $Pl(A)$, with $Bel(A) \leq Pl(A)$
- The intervals $[Bel(A), Pl(A)]$ have maximum length when $m = m_?$ is vacuous: then, $Bel(A) = 0$ for all $A \neq \Omega$, and $Pl(A) = 1$ for all $A \neq \emptyset$.
- The intervals $[Bel(A), Pl(A)]$ have minimum length when m is Bayesian. Then,

$$Bel(A) = Pl(A) = \sum_{\omega \in A} m(\{\omega\})$$

for all A, and Bel is a probability measure.

Road scene analysis example

- We had $\Omega = \{G, R, T, O, S\}$ and

$$m(\{T, O\}) = 0.9, \quad m(\Omega) = 0.1$$

- What are the credibility and the plausibility that the region corresponds / does not correspond to a tree?

$$Bel(\{T\}) = 0, \quad Pl(\{T\}) = 0.9 + 0.1 = 1$$

$$Bel(\overline{\{T\}}) = 0, \quad Pl(\overline{\{T\}}) = 1$$

But $Bel(\{T\} \cup \overline{\{T\}}) = Bel(\Omega) = 1$ and $Pl(\{T\} \cup \overline{\{T\}}) = Pl(\Omega) = 1$.

- We observe that

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

$$Pl(A \cup B) \leq Pl(A) + Pl(B) - Pl(A \cap B)$$

(*Bel* is superadditive, *Pl* is subadditive).

Characterization of belief functions

Theorem

Let $F : 2^\Omega \to [0,1]$*. The following two statements are equivalent:*

Statement 1 *There exists a mass function* $m : 2^\Omega \to [0,1]$ *such that* $F(A) = \sum_{B \subseteq A} m(B)$ *for all* $A \subseteq \Omega$ *(i.e., F is a belief function).*

Statement 2 *Function F has the following 3 properties:*

1. $F(\emptyset) = 0$
2. $F(\Omega) = 1$
3. *For any* $k \geq 2$ *and for any family* $A_1, \ldots, A_k$ *in* 2^Ω*,*

$$F\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} F\left(\bigcap_{i \in I} A_i\right) \qquad (1)$$

(Property (1) is called complete monotonicity).

Relations between *m*, *Bel* and *Pl*

- Let m be a mass function, Bel and Pl the corresponding belief and plausibility functions
- Thanks to the following equations, given any one of these functions, we can recover the other two: for all $A \subseteq \Omega$,

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

$$Pl(A) = 1 - Bel(\overline{A}) \tag{3}$$

$$Bel(A) = 1 - Pl(\overline{A}) \tag{4}$$

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B) \tag{5}$$

- m, Bel et Pl are thus three equivalent representations of a piece of evidence.

Outline

Definitions

Definition (Consonant mass function)

A mass function m is consonant iff its focal sets are nested, i.e., for any two focal set A_i and A_j, $A_i \subseteq A_j$ or $A_j \subseteq A_i$

Definition (Possibility measure)

A mapping $\Pi : 2^\Omega \rightarrow [0,1]$ is a possibility measure iff, for any $A, B \subseteq \Omega$,

$$\Pi(A \cup B) = \max\left[\Pi(A), \Pi(B)\right]$$

Definition (Necessity measure)

A mapping $N : 2^\Omega \rightarrow [0,1]$ is a necessity measure iff, for any $A, B \subseteq \Omega$,

$$N(A \cap B) = \min\left[N(A), N(B)\right]$$

Theorem

Theorem

Let m be a mass function, and let Bel and Pl be the corresponding belief and plausibility functions. The following statements are equivalent:

1. *m is consonant*
2. *Bel is a necessity measure*
3. *Pl is a possibility measure*

Consequence: The theory of belief functions is more expressive than possibility theory (a possibility measure is a plausibility function, but the converse is false in general).

Proof of $1 \Rightarrow 2$

- Let m be a consonant mass function with focal sets $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_r$.
- For any $A, B \subseteq \Omega$, let i_1 and i_2 be the largest indices such that $A_i \subseteq A$ and $A_i \subseteq B$, respectively.
- Then, $A_i \subseteq A \cap B$ iff $i \leq \min(i_1, i_2)$ and

$$
\begin{aligned}
Bel(A \cap B) &= \sum_{i=1}^{\min(i_1,i_2)} m(A_i) \\
&= \min\left(\sum_{i=1}^{i_1} m(A_i), \sum_{i=1}^{i_2} m(A_i)\right) \\
&= \min(Bel(A), Bel(B)).
\end{aligned}
$$

Proof of 2 $\Rightarrow$ 3

- Now, from the equality $\overline{A \cup B} = \overline{A} \cap \overline{B}$, we have

$$\begin{aligned}
Pl(A \cup B) &= 1 - Bel(\overline{A \cup B}) \\
&= 1 - Bel(\overline{A} \cap \overline{B}) \\
&= 1 - \min(Bel(\overline{A}), Bel(\overline{B})) \\
&= \max(1 - Bel(\overline{A}), 1 - Bel(\overline{B})) \\
&= \max(Pl(A), Pl(B)).
\end{aligned}$$

Contour function

Definition (Contour function)

The contour function *of a belief function Bel is the mapping* $\Omega \to [0,1]$ *defined by*

$$pl(\omega) = Pl(\{\omega\}), \quad \forall \omega \in \Omega$$

- When m is consonant, it can be recovered from its contour function:

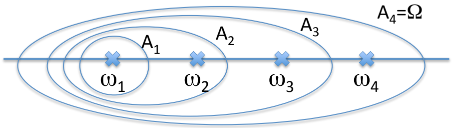
$$\boxed{Pl(A) = \max_{\omega \in A} pl(\omega)}$$

  and we have

$$\max_{\omega \in \Omega} pl(\omega) = Pl(\Omega) = 1$$

- In Possibility theory, function pl is called a possibility distribution.

Proof of 3 $\Rightarrow$ 1



- Let Pl be a possibility measure and let pl be its contour function.
- Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the frame of discernment with elements arranged by decreasing order of plausibility, i.e.,

$$1 = pl(\omega_1) \geq pl(\omega_2) \geq \ldots \geq pl(\omega_n),$$

  and let A_i denote the set $\{\omega_1, \ldots, \omega_i\}$, for $1 \leq i \leq n$.
- Let m be the consonant mass function defined as follows:

$$m(A_i) = pl(\omega_i) - pl(\omega_{i+1}), \quad 1 \leq i \leq n-1,$$
$$m(\Omega) = pl(\omega_n).$$

Example

For instance, for the following contour function defined on the frame $\Omega = \{a, b, c, d\}$:

ω	a	b	c	d
$pl(\omega)$	0.3	0.5	1	0.7

the corresponding mass function is

$$\begin{aligned} m(\{c\}) &= 1 - 0.7 = 0.3 \\ m(\{c, d\}) &= 0.7 - 0.5 = 0.2 \\ m(\{c, d, b\}) &= 0.5 - 0.3 = 0.2 \\ m(\{c, d, b, a\}) &= 0.3. \end{aligned}$$

Proof of 3 $\Rightarrow$ 1 (continued)

- Let Pl_m be the plausibility function induced by m.
- For any subset A of Ω, let $i_A = \min\{1 \leq i \leq n : \omega_i \in A\}$.
- $A_i \cap A \neq \emptyset$ iff $i \geq i_A$.
- Consequently,

$$
\begin{aligned}
Pl_m(A) &= \sum_{i=i_A}^{n} m(A_i) \\
&= pl(\omega_{i_A}) - pl(\omega_{i_A+1}) + pl(\omega_{i_A+1}) - pl(\omega_{i_A+2}) + \ldots - pl(\omega_n) + pl(\omega_n) \\
&= pl(\omega_{i_A}) \\
&= \max_{\omega \in A} pl(\omega) = Pl(A),
\end{aligned}
$$

i.e., $Pl_m = Pl$.

Outline

Outline

Road scene example continued

- Variable X was defined as the type of object in some region of the image, and the frame was $\Omega = \{G, R, T, O, S\}$, corresponding to the possibilities **G**rass, **R**oad, **T**ree/Bush, **O**bstacle, **S**ky
- A lidar sensor gave us the following mass function:

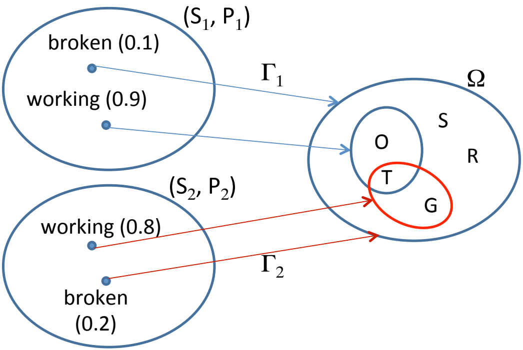
$$m_1(\{T, O\}) = 0.9, \quad m_1(\Omega) = 0.1$$

- Now, assume that a camera returns the mass function:

$$m_2(\{G, T\}) = 0.8, \quad m_2(\Omega) = 0.2$$

- How to combine these two pieces of evidence?

Analysis



- If interpretations $s_1 \in S_1$ and $s_2 \in S_2$ both hold, then $X \in \Gamma_1(s_1) \cap \Gamma_2(s_2)$
- If the two pieces of evidence are independent, then the probability that s_1 and s_2 both hold is $P_1(\{s_1\})P_2(\{s_2\})$

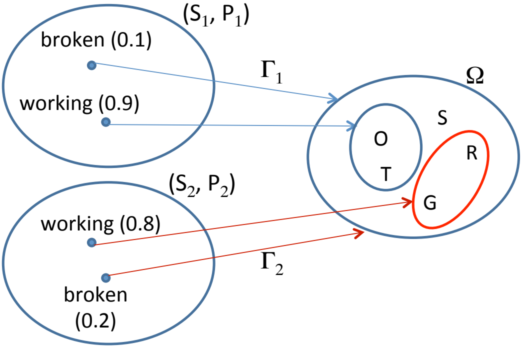
Computation

$m_1 \backslash m_2$	$\{T, G\}$ (0.8)	Ω (0.2)
$\{O, T\}$ (0.9)	$\{T\}$ (0.72)	$\{O, T\}$ (0.18)
Ω (0.1)	$\{T, G\}$ (0.08)	Ω (0.02)

We then get the following combined mass function,

$$
\begin{aligned}
m(\{T\}) &= 0.72 \\
m(\{O, T\}) &= 0.18 \\
m(\{T, G\}) &= 0.08 \\
m(\Omega) &= 0.02
\end{aligned}
$$

Case of conflicting pieces of evidence



- If $\Gamma_1(s_1) \cap \Gamma_2(s_2) = \emptyset$, we know that s_1 and s_2 cannot hold simultaneously
- The joint probability distribution on $S_1 \times S_2$ must be conditioned to eliminate such pairs

Computation

$m_1 \backslash m_2$	$\{G, R\}$ (0.8)	Ω (0.2)
$\{O, T\}$ (0.9)	$\emptyset$ (0.72)	$\{O, T\}$ (0.18)
Ω (0.1)	$\{G, R\}$ (0.08)	Ω (0.02)

We then get the following combined mass function,

$$\begin{aligned} m(\emptyset) &= 0 \\ m(\{O, T\}) &= 0.18/0.28 = 9/14 \\ m(\{G, R\}) &= 0.08/0.28 = 4/14 \\ m(\Omega) &= 0.02/0.28 = 1/14 \end{aligned}$$

Dempster's rule

- Let m_1 and m_2 be two mass functions and

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$$

  their degree of conflict
- If $\kappa < 1$, then m_1 and m_2 can be combined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \neq \emptyset \tag{6}$$

  and $(m_1 \oplus m_2)(\emptyset) = 0$
- $m_1 \oplus m_2$ is called the orthogonal sum of m_1 and m_2
- This rule can be used to combine mass functions induced by independent pieces of evidence

Another example

A	$\emptyset$	$\{a\}$	$\{b\}$	$\{a,b\}$	$\{c\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$m_1(A)$	0	0	0.5	0.2	0	0.3	0	0
$m_2(A)$	0	0.1	0	0.4	0.5	0	0	0

		m_2		
		$\{a\}, 0.1$	$\{a,b\}, 0.4$	$\{c\}, 0.5$
	$\{b\}, 0.5$	$\emptyset, 0.05$	$\{b\}, 0.2$	$\emptyset, 0.25$
m_1	$\{a,b\}, 0.2$	$\{a\}, 0.02$	$\{a,b\}, 0.08$	$\emptyset, 0.1$
	$\{a,c\}, 0.3$	$\{a\}, 0.03$	$\{a\}, 0.12$	$\{c\}, 0.15$

The degree of conflict is $\kappa = 0.05 + 0.25 + 0.1 = 0.4$. The combined mass function is

$$(m_1 \oplus m_2)(\{a\}) = (0.02 + 0.03 + 0.12)/0.6 = 0.17/0.6$$
$$(m_1 \oplus m_2)(\{b\}) = 0.2/0.6$$
$$(m_1 \oplus m_2)(\{a,b\}) = 0.08/0.6$$
$$(m_1 \oplus m_2)(\{c\}) = 0.15/0.6.$$

Properties

1. Commutativity, associativity. Neutral element: $m_?$
2. Generalization of intersection: if m_A and m_B are logical mass functions and $A \cap B \neq \emptyset$, then
$$m_A \oplus m_B = m_{A\cap B}$$
3. If either m_1 or m_2 is Bayesian, then so is $m_1 \oplus m_2$ (as the intersection of a singleton with another subset is either a singleton, or the empty set).
4. Let $pl_{1\oplus 2}$ be the contour function of $m_1 \oplus m_2$. Then,
$$pl_{1\oplus 2} = \frac{pl_1 pl_2}{1-\kappa} \propto pl_1 pl_2$$

Proof: see next slide.

Proof of Property 4

For any $\omega \in \Omega$,

$$
\begin{aligned}
pl_{1\oplus 2}(\omega) &= \sum_{\{B:\omega\in B\}} (m_1 \oplus m_2)(B) \\
&= (1-\kappa)^{-1} \sum_{\{B:\omega\in B\}} \sum_{\{C,D:C\cap D=B\}} m_1(C)m_2(D) \\
&= (1-\kappa)^{-1} \sum_{\{C,D:\omega\in C\cap D\}} m_1(C)m_2(D) \\
&= (1-\kappa)^{-1} \sum_{\{C,D:\omega\in C,\omega\in D\}} m_1(C)m_2(D) \\
&= (1-\kappa)^{-1} \left(\sum_{\{C:\omega\in C\}} m_1(C) \right) \left(\sum_{\{D:\omega\in D\}} m_2(D) \right) \\
&= (1-\kappa)^{-1} pl_1(\omega) \cdot pl_2(\omega).
\end{aligned}
$$

Outline

Dempster's rule conditioning

- Conditioning is a special case, where a mass function m is combined with a logical mass function m_B. Notation:

$$m \oplus m_B = m(\cdot \mid B)$$

- We thus have $m(A \mid B) = 0$ for any A not included in B and, for any $A \subseteq B$,

$$m(A \mid B) = (1-\kappa)^{-1} \sum_{C \cap B = A} m(C), \tag{7}$$

where the degree of conflict κ is

$$\kappa = \sum_{C \cap B = \emptyset} m(C) = 1 - \sum_{C \cap B \neq \emptyset} m(C) = 1 - Pl(B).$$

Conditional plausibility function

Proposition

The plausibility function $Pl(\cdot|B)$ induced by $m(\cdot|B)$ is given by

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

Proof: We have

$$\begin{aligned} Pl(A \mid B) &= \sum_{\{C:C\cap A\neq\emptyset\}} m(C|B) \\ &= Pl(B)^{-1} \sum_{\{C:C\cap A\neq\emptyset\}} \sum_{\{D:D\cap B=C\}} m(D) \\ &= Pl(B)^{-1} \sum_{\{D:D\cap B\cap A\neq\emptyset\}} m(D) = \frac{Pl(A \cap B)}{Pl(B)} \end{aligned}$$

If Pl is a probability measure, $Pl(\cdot \mid B)$ is, thus, the conditional probability measure given B: Dempster's rule of combination thus extends Bayesian conditioning.